arXiv:2602.01731v1 [cs.RO] 2 Feb 2026

面向物体引起遮挡环境中移动操作机器人的不确定性感知非抓取式操作

Jiwoo Hwang[1], Taegeun Yang[2],Jeil Jeong[1],Minsung Yoon[2] 和 Sung-Eui Yoon[2†]

摘要—利用机载感知进行非抓取式操作面临一项根本性挑战：被操作物体会遮挡传感器的视场，形成可能导致碰撞的遮挡区域。我们提出*CURA-PPO*，这是一种强化学习框架，通过在部分可观测性下显式建模不确定性来应对这一挑战。通过将碰撞可能性预测为一种分布，我们提取风险与不确定性，以指导机器人的动作。不确定性项鼓励主动感知，使操作与信息收集能够同步进行，从而消除遮挡。当与反映观测可靠性的置信度图结合时，我们的方法即使在严重传感器遮挡的情况下，也能实现安全导航。在不同物体尺寸和障碍物配置下开展的大量实验表明，*CURA-PPO*的成功率最高可比基线方法提高 3×，并能够学习处理遮挡的行为。我们的方法仅利用机载感知，为杂乱环境中的自主操作提供了一种实用解决方案。

I. 引言

非抓取式操作在不进行抓取的情况下控制物体，常用于处理因尺寸、重量或表面特性而超出抓取能力的物品 [1]–[9]。移动平台——包括四足机器人 [5],、轮式机器人 [7], 和移动操作机器人 [9]——将这些能力扩展到更大的工作空间，从而能够完成在仓库中搬运箱子或重新摆放家具等任务。然而，当这些平台仅依赖机载传感器（例如激光雷达）时，会出现一个根本性挑战：被操作的物体本身会遮挡传感器的视场，从而在操作过程中形成危险的盲区。当出现意外障碍物时——例如物品从货架上掉落，或人员进入工作空间——这种遮挡引起的不确定性便变得至关重要，因而需要实时感知与避障。尽管物体引起的遮挡对于仅依靠局部感知、在杂乱环境中实现安全自主操作具有实际重要性，但非抓取式任务中的这一问题在文献中仍未得到充分研究。

为应对物体引发的遮挡所带来的感知挑战，我们采用了具备不确定性感知能力的决策策略。近期研究表明，当智能体明确量化不确定性，并将降低不确定性纳入目标时，它们会表现出主动搜寻信息的行为——主动探索模糊区域，而非依赖不完整的观测 [10]–[14]。这一原则

[1]J. Hwang 和 J. Jeong 任职于韩国科学技术院（KAIST）机器人项目，该校位于韩国大田，邮编 34141。[2]T. Yang、M. Yoon 和 S. Yoon 任职于同一所院校 KAIST 的计算机学院。

[†]S. Yoon 为通讯作者。sungeui@kaist.edu

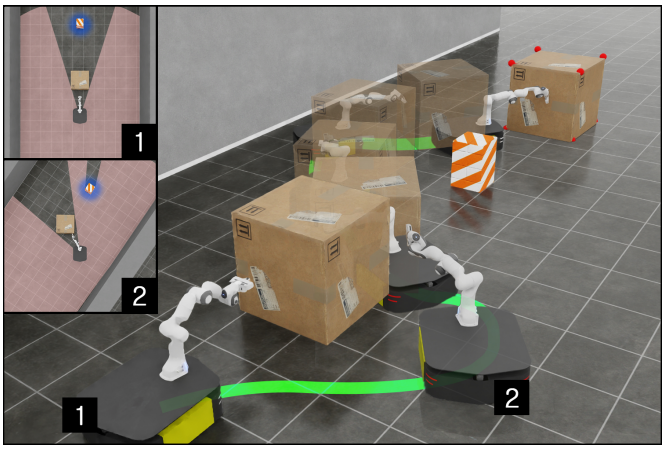

图 1：**物体诱导遮挡下的不确定性感知非抓取式操作**。移动操作机器人利用局部机载感知推动物体到达目标。(1) 物体造成遮挡，遮住了未预料到的障碍物。(2) 移动操作机器人调整行为，主动降低遮挡带来的不确定性，从而发现隐藏的障碍物，安全地进行非抓取式操作。

直接适用于我们的场景：通过将不确定性降低与操作目标结合起来，机器人能够主动消除遮挡，同时保持任务进度。当被操作物体遮挡传感器视野时，这种由不确定性驱动的探索尤为重要，因为它能促使机器人进行策略性移动，以提升可观测性，同时不偏离操作目标。

基于这些原则，我们提出了*CURA-PPO*（碰撞不确定性与风险感知 PPO），这是一种强化学习（RL）框架，能够在存在遮挡的情况下利用局部感知安全地进行非抓取式操作。

总而言之，我们的主要贡献有三点：

- 我们将物体诱发遮挡下的非抓取式操作建模为不确定性感知决策问题：移动操作机器人仅依靠机载二维激光雷达，在被操作物体形成遮挡区域的情况下管理杂乱环境。
- 我们提出了 *CURA-PPO*，这是一种融合分布式碰撞估计器（*DCE*）的强化学习框架，可将碰撞可能性预测为一个分布。通过将该分布的均值（风险）和方差（不确定性）纳入内在代价，我们的方法能够联合优化任务完成、碰撞规避和不确定性降低。
- 我们通过大量仿真实验证明，*CURA-PPO*在不同物体尺寸和障碍物配置下均显著优于基线方法。定性分析还揭示了其学会的主动感知行为，能够主动消除遮挡。

II. 相关工作

A. 非抓取式操作

非抓取式操作是指通过施加单侧约束（例如推）来控制物体的一类交互方式，无需抓取所提供的双侧约束所形成的力闭合 [1]。这种方法适用于操作超出机器人物理能力范围的物体，例如超出其负载能力或抓取范围的物体 [2]–[9]。基于这些优势，非抓取式操作的研究涵盖了不同尺度的任务，从使用固定基座机械臂在桌面上精确操控物体 [2]–[4]，到使用移动机器人在扩展环境中进行远距离操作 [5]–[9]。

对于远距离操作，已有研究取得了令人鼓舞的成果，但这些方法假设已全面掌握环境中所有障碍物的信息，包括其位置和几何形状。在实际场景中，这类信息往往无法轻易获取，机器人必须依赖机载传感器。这种设置会带来遮挡方面的挑战：被操作物体会遮挡传感器视野，限制对环境的感知。为应对这些挑战，我们提出一种基于机载传感器、适用于部分可观测情形的不确定性感知操作策略。具体而言，我们的方法融入主动传感——通过有目的的运动来收集信息——以处理遮挡带来的不确定性，同时兼顾不确定性降低与任务执行。

B. 不确定性感知决策

在部分可观测环境中进行决策，需要处理由感知不完整带来的不确定性 [15]。这类不确定性通常通过模型预测的方差来量化，可采用分布式表示学习 [16]–[18]，或基于采样的近似 [19]–[21]。为增强部分可观测条件下的鲁棒性，近期研究将不确定性估计纳入决策过程 [22]–[25]。此外，将不确定性降低作为学习信号，可以鼓励信息寻求行为，使智能体能够识别并消除感知歧义 [10]–[14]。我们将这一原则应用于具备局部感知的非抓取式操作，其中遮挡会降低可观测性。我们通过碰撞可能性预测的方差来量化不确定性：朝可观测区域推物体会产生较低方差，而将物体推向遮挡区域则会产生较高方差。将这一方差纳入策略目标后，我们便能鼓励主动感知行为，在操作过程中减少遮挡。

III. 变量记号

本节总结了本文中使用的关键符号和变量。笛卡尔位置、线速度和线加速度分别记为 $\boldsymbol{p}$、$\boldsymbol{v}$ 和 $\dot{\boldsymbol{v}} \in \mathbb{R}^3$。XYZ欧拉角 $\boldsymbol{\theta}$ 表示方向，角速度记为 $\boldsymbol{\omega} \in \mathbb{R}^3$。对于七自由度机械臂，关节位置、速度和加速度分别记为 $\boldsymbol{q}$、$\dot{\boldsymbol{q}}$ 和 $\ddot{\boldsymbol{q}} \in \mathbb{R}^7$。上标表示参考坐标系，下标表示实体、轴分量和时间步。例如，$p^{b}_{g,\ x,\ t}$表示时间步 t 时基座坐标系中目标位置的 x 分量，其中 w、b、e、o 和 g 分别表示世界、基座、末端执行器、物体和目标坐标系。

IV. 预备知识

本节介绍部分可观测马尔可夫决策过程（POMDP）和近端策略优化（PPO）的必要背景知识，以帮助理解后续的方法贡献。与完全可观测的马尔可夫决策过程不同，POMDP 由元组 $(\mathcal{S},\ \mathcal{A},\ \mathcal{T},\ \mathcal{R},\ \Omega\ \mathcal{O},\ \gamma)$ 定义，其中智能体接收观测 $\boldsymbol{o} \in \Omega$，而非直接访问状态 $\boldsymbol{s} \in \mathcal{S}$。其中，$\mathcal{A}$ 是动作空间，$\mathcal{T}$ 是状态转移动态，$\mathcal{R}$ 是奖励函数，$\mathcal{O}$ 是将状态映射到观测空间 Ω 的观测函数，γ 是折扣因子。

强化学习的目标是学习一个策略$\pi_\phi(\boldsymbol{a}_t|\boldsymbol{o}_t)$，使期望累积奖励最大化。PPO [26] 通过迭代式策略改进来优化这一目标，其代理目标函数为：

$$\mathcal{J}^{PPO}(\phi) = \mathbb{E}_{(\boldsymbol{o}_t,\boldsymbol{a}_t)\sim\pi_{\phi_{\text{old}}}}\left[\frac{\pi_\phi(\boldsymbol{a}_t|\boldsymbol{o}_t)}{\pi_{\phi_{\text{old}}}(\boldsymbol{a}_t|\boldsymbol{o}_t)}\cdot\hat{A}_t\right] \tag{1}$$

其中，概率比 $r(\phi) = \frac{\pi_\phi(\boldsymbol{a}_t|\boldsymbol{o}_t)}{\pi_{\phi_{\text{old}}}(\boldsymbol{a}_t|\boldsymbol{o}_t)}$ 用于衡量策略变化。实践中，会对概率比进行裁剪，以促进稳定的策略更新。优势函数$A(\boldsymbol{s}_t,\ \boldsymbol{a}_t) = Q(\boldsymbol{s}_t,\ \boldsymbol{a}_t) - V(\boldsymbol{s}_t)$用于比较采取特定动作的价值与该状态下的平均价值，其中 Q 和 V 分别表示动作价值函数和状态价值函数。训练期间，估计得到的优势 A_t作为策略优化 [27]中的学习信号：正值会促使策略提高所选动作的概率，而负值则会降低该概率。

V. 方法

我们提出了碰撞不确定性与风险感知 *PPO (CURA-PPO)*，这是一种改进的强化学习算法，用于在由被操作物体自身造成的遮挡下执行长距离非抓取式操作。我们的方法由两个关键部分组成：一个分布式碰撞估计器 *(DCE)*，用于预测碰撞可能性分布；以及一个改进的 PPO 目标函数，通过降低不确定性和最小化碰撞风险来提升安全性。后续章节将介绍任务形式化、策略架构、DCE 的实现以及训练流程。

A. 任务与环境建模

我们将该任务表述为存在障碍物时的目标条件非抓取式操作问题，并使用仅依靠基座安装的二维激光雷达的移动操作机器人。传感器只能提供部分观测，因为被操作物体会遮挡其部分视场——这是操作过程中的一项基本限制，会引入感知不确定性。图 3 展示了我们的仿真环境，并详细说明了任务的空间布局，包括起始区域、目标区域和障碍物区域。物体和障碍物的尺寸各不相同，障碍物还会在回合进行期间出现在不同位置和时间，以模拟箱子突然从货架上掉落或工具遗留在工作空间等情形。这种变化凸显了操作过程中主动感知的重要性，因为减少遮挡区域的不确定性有助于提高安全性。

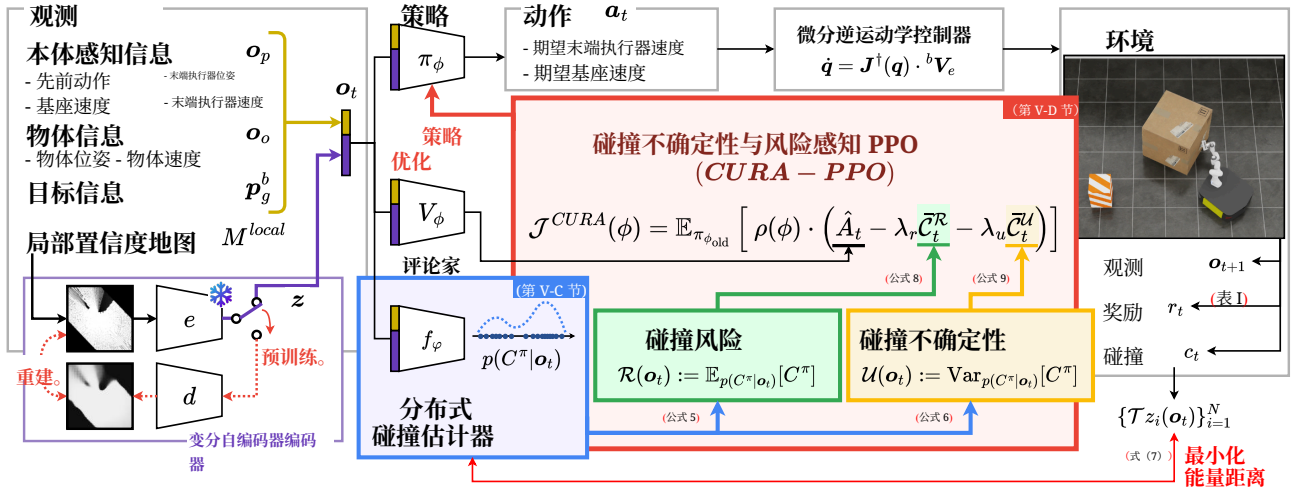

图 2：**总体框架**。局部观测（包括本体感知、物体、目标和置信度地图特征）由策略网络处理 π_ϕ （见第V-B节），该网络输出通过微分逆运动学执行的速度指令。分布式碰撞估计器（DCE， f_φ）将碰撞可能性预测为一个分布，并提取风险和不确定性信号（见第V-C节），以增强*CURA-PPO*中的PPO目标函数，鼓励采取能够同时降低碰撞风险和感知不确定性的行为（见第V-D节）。

B. 不确定性下的非抓取式操作策略

如图 2所示，我们的操作策略 π_ϕ 将部分激光雷达观测和目标位姿映射为机器人动作，从而实现无碰撞物体操作。我们将其表述为部分可观测马尔可夫决策过程（POMDP），并采用策略 $\pi_\phi(\boldsymbol{a}_t|\boldsymbol{o}_t)$ 以最大化期望回报。

1）传感器编码：我们将机器人获取的激光雷达观测编码为紧凑的时空表示，同时考虑当前测量值及其随时间变化的可靠性。简单累积过去的激光雷达扫描虽然能扩大空间覆盖范围，却无法区分时间。当此前观测过、但当前被遮挡的区域出现意料之外的障碍物时，这会导致问题：系统会依据过时的观测，将潜在危险区域视为安全区域。为解决这一问题，我们使用置信度图 [28]，使传感器读数随时间呈指数衰减——近期测量值保持可靠，而过时的测量值则逐渐失效：

$$M_t(x,y) = \begin{cases} 1, & \text{if } (x,y) \text{ observed at } t \\ \alpha \cdot M_{t-1}(x,y), & \text{otherwise} \end{cases} \quad (2)$$

其中， M_t 表示时间步 t的全局置信度地图，每个像素（分辨率为 0，1 米）处的值为 M_t（x，y）$\in [0,1]$，衰减因子为 $\alpha \in$（0，1）。编码时，我们从该全局地图中提取以机器人基座为中心的局部窗口 $M_t^{\text{local}} \in \mathbb{R}^{100\times100}$，并使用在随机轨迹上预训练、且在策略学习期间保持冻结的变分自编码器（VAE）将其编码为潜在向量 $\boldsymbol{z}_t$。

2）策略架构：我们的策略 $\pi_\phi : \Omega \to \mathcal{A}$ 将四元组观测 $\boldsymbol{o} = (\boldsymbol{o}_p,\ \boldsymbol{o}_o,\ \boldsymbol{p}_g^b,\ \boldsymbol{z})$映射为一个动作。以基座坐标系表示的观测包括：(1) 本体感知信息 $\boldsymbol{o}_p = [\boldsymbol{p}_e^b, \boldsymbol{v}_e^b, \boldsymbol{v}_b^b, \boldsymbol{a}_{t-1}]$，包括末端执行器的位姿和速度、基座速度以及先前动作；(2) 物体信息 $\boldsymbol{o}_o = [\boldsymbol{p}_o^b, \boldsymbol{v}_o^b]$，包括当前位姿和速度；(3) 目标位姿 $\boldsymbol{p}_g^b$；以及 (4) 编码后的置信度图 $\boldsymbol{z}$，提供部分环境信息。策略为全向底座和末端执行器输出速度指令 {v20}，这些指令通过微分逆运动学转换为关节空间目标，并通过关节阻抗控制执行。

表 I：操作策略奖励 $\mathcal{R} = \mathcal{R}_{\text{task}} + \mathcal{R}_{\text{reg}}$

奖励项	表达式
任务奖励： $\mathcal{R}_{\text{task}} = \sum_{k=1}^{4} r_k$	
关键点（r_1）	$w_1 \cdot \exp(-0.25\,\|\|\boldsymbol{G}_g^w - \boldsymbol{G}_o^w\|\|_2)$
进度 (r_2)	$w_2 \cdot \exp(5.0(\hat{\boldsymbol{v}}_o^w \cdot \hat{\boldsymbol{p}}_{o\to g}^w - 1))$
速度 (r_3)	$w_3 \cdot \|\|\boldsymbol{v}_o^w\|\|_2 / v_{\max}$
接触点 (r_4)	$w_4 \cdot \exp(-0.1\,\|\|\boldsymbol{p}_r^w - \boldsymbol{p}_e^w\|\|_2)$
正则化奖励： $\mathcal{R}_{\text{reg}} = \sum_{k=5}^{6} r_k$	
动作变化率 (r_5)	$w_5 \cdot \|\|\boldsymbol{a}_t - \boldsymbol{a}_{t-1}\|\|_2$
平滑动作 (r_6)	$w_6 \cdot \|\|\ddot{\boldsymbol{q}}\|\|_2$
奖励权重	
w_1=4.0	w_2=2.0 w_3=1.0 w_4=1.0 w_5=-0.1 w_6=-0.1

- $\hat{\cdot}$表示一个单位向量。
- $v$max 表示移动底座的最大速度。
- $\boldsymbol{G} \in \mathbb{R}^{24}$ 是 [5], 中提出的关键点向量，由物体八个角点的三维坐标组成。• $\boldsymbol{p}_r$ 是按照 [9]在物体表面随机采样的点。

3) 实现细节： 我们使用配备 7自由度 Franka 机械臂和一台二维激光雷达的 ClearPath Ridgeback，其 360° 视场为$v_{\max}$，底座最大速度为 0.8 m/s。我们在 Isaac Sim [29]中使用 NVIDIA RTX 4090 显卡和 2048 个并行环境进行训练，并采用近端策略优化（PPO） [26] 来优化策略。策略以 5 Hz 运行并生成速度指令，而关节阻抗控制以 200 Hz 运行。微分逆运动学采用带阻尼因子 $\lambda = 0.02$ 的阻尼最小二乘（DLS）方法。置信度图采用指数更新，$\alpha = 0.9$表 I总结了训练中使用的奖励项。

关键点奖励会最小化对应物体关键点与目标关键点之间的距离 [5]，而接触点奖励则引导末端执行器前往目标表面上采样得到的多个点，从而鼓励多样化的推动行为 [9]。

C. 分布式碰撞估计器

尽管我们的操作策略能够依据部分观测执行推动动作，但它难以生成能够主动降低遮挡引起的不确定性的动作。为应对这一挑战，我们提出一种分布式碰撞估计器 *(DCE)*，将碰撞可能性建模为一种分布，从而能够在策略学习过程中同时优化其期望*(*风险*)* 和方差*(*不确定性*)*。

1）碰撞可能性分布：我们将碰撞可能性 C^π 定义为一个随机变量，表示遵循策略π_ϕ 时碰撞指示变量的折扣累积，以捕捉未来碰撞的紧迫性：

$$C^\pi := \sum_{k=0}^{\infty} \gamma_c^k c_{t+k}, \quad c_t = \mathbb{1}_{\text{collision at time } t} \in \{0,1\} \tag{3}$$

其中，$\gamma_c \in$（0，1）为折扣因子，设为 0.9，c_t 表示时刻 t 发生碰撞。为获取由部分可观测性引起的内在不确定性，我们使用分位数回归 [30]来近似条件分布 $p(C^\pi|\boldsymbol{o}_t)$。我们的 DCE网络 $f_\varphi : \Omega \to \mathbb{R}^N$ 预测 N 个分位数：

$$f_\varphi(\boldsymbol{o}_t) = \{z_i(\boldsymbol{o}_t)\}_{i=1}^N \approx \{q_{\tau_i}(C^\pi|\boldsymbol{o}_t)\}_{i=1}^N, \quad \tau_i = \frac{i-0.5}{N} \tag{4}$$

其中，$q_{\tau_i}(C^\pi|\boldsymbol{o}_t)$ 表示 $p(C^\pi|\boldsymbol{o}_t)$ 的第 i个分位数。

2）风险与不确定性估计：我们将风险和不确定性定义为给定观测 $\boldsymbol{o}_t$时C^π 的条件期望和方差。根据 DCE 的分位数预测 $\{z_i(\boldsymbol{o}_t;\varphi)\}_{i=1}^N$，我们估计风险和不确定性：

$$\mathcal{R}(\boldsymbol{o}_t) := \mathbb{E}_{p(C^\pi|\boldsymbol{o}_t)}[C^\pi] \approx \frac{1}{N}\sum_{i=1}^N z_i(\boldsymbol{o}_t;\varphi) \tag{5}$$

$$\mathcal{U}(\boldsymbol{o}_t) := \mathrm{Var}_{p(C^\pi|\boldsymbol{o}_t)}[C^\pi] \approx \frac{1}{N}\sum_{i=1}^N (z_i(\boldsymbol{o}_t;\varphi) - \mathcal{R}_\varphi(\boldsymbol{o}_t))^2 \tag{6}$$

此后，我们将这些经验估计分别记为 $\mathcal{R}_\varphi$ 和 $\mathcal{U}_\varphi$。风险与不确定性刻画了碰撞预测的不同方面：高风险表明可能发生碰撞，而高不确定性则反映基于部分观测的预测不可靠。这些互补指标引导策略采取不同行为——风险驱动碰撞规避，而不确定性则促进主动感知以消除歧义。

3）分布式碰撞估计器训练：我们通过最小化预测分位数分布与目标分位数分布之间的能量距离来训练DCE网络 f_φ [31]：

$$\mathcal{L}_{\text{DCE}}(\varphi) = 2\mathbb{E}_{i,j}[z_i - \mathcal{T}z_j] - \mathbb{E}_{i,j}[\mathcal{T}z_i - \mathcal{T}z_j] - \mathbb{E}_{i,j}[z_i - z_j] \tag{7}$$

其中 $\mathcal{T}$表示根据观测轨迹计算目标分位数的分布式贝尔曼算子。能量距离损失促进分布一致性：第一项使预测与目标对齐，其余项通过防止坍缩为点估计来保持分布形状。训练期间，DCE参数 φ 与策略π_ϕ 联合更新，并使用 $N = 50$ 个分位数以平衡分辨率与计算效率。

算法1： *CURA-PPO*

对于 迭代 = *1*，*2*，*.....*，**执行**
  对于 环境 = *1*，*2*，*.....*，*N* **执行**
    运行策略$\pi_{\phi_{\text{old}}}$ 共 T 个时间步；
  计算优势估计 $\hat{A}_{1\ldots T}$；
  计算风险成本 $\mathcal{C}_1^{\mathcal{R}}$,...,T 相对于 $f_{\varphi_{old}}$（公式 8）；
  计算不确定性成本 $\mathcal{C}_1^{\mathcal{U}}$,...,T 关于 $f_{\varphi_{old}}$（公式 9）；
  计算 CURA策略目标 $\mathcal{J}^{CURA}$（公式 10）关于 π_ϕ，训练轮次为 K，小批量大小为 M；
  更新演员网络、评论家和 DCE：$\phi_{\text{old}} \leftarrow \phi$, $\varphi_{\text{old}} \leftarrow \varphi$；

D. *CURA-PPO*：碰撞不确定性与风险感知 *PPO*

在DCE分布预测的基础上，我们提出*CURA-PPO*，通过一项策略目标增强标准PPO，该目标提高选择能够降低DCE估计的风险和不确定性的动作的概率。这一整合将单目标任务完成表述转化为同时以碰撞规避和不确定性降低为目标的多目标问题。

1）基于分布统计量的成本构建：我们根据 DCE 的预测构建两个单步时序差分（TD）成本信号。风险成本是预期碰撞可能性的时序差分残差，用于量化当前动作使预期未来碰撞超出即时碰撞的程度。不确定性成本是预测方差的变化，用于量化下一次观测相较于当前观测变得更加不确定还是更加确定：

$$\mathcal{C}_t^{\mathcal{R}} = (c_t + \gamma_c \mathcal{R}_\varphi(\boldsymbol{o}_{t+1}) - \mathcal{R}_\varphi(\boldsymbol{o}_t)) \tag{8}$$

$$\mathcal{C}_t^{\mathcal{U}} = (\mathcal{U}_\varphi(\boldsymbol{o}_{t+1}) - \mathcal{U}_\varphi(\boldsymbol{o}_t)) \tag{9}$$

其中 $\boldsymbol{o}_{t+1}$ 表示执行动作后得到的观测$\boldsymbol{a}_t$，给定观测 $\boldsymbol{o}_t$。在风险成本（公式8）中，c_t 表示即时碰撞指示量，γ_c 表示公式3中的折扣因子。负的风险成本意味着当前动作降低了预期碰撞可能性。负的不确定性成本（公式9）表示预测不确定性降低。

2）策略目标：我们通过加入惩罚项来扩展优势，将这些成本整合到近端策略优化（PPO）中。在公式1的基础上，*CURA-PPO* 目标变为：

$$\mathcal{J}^{CURA}(\phi) = \mathbb{E}_{\pi_{\phi_{\text{old}}}}\left[\ \rho(\phi)\cdot\left(\hat{A}_t - \lambda_r \bar{\mathcal{C}}_t^{\mathcal{R}} - \lambda_u \bar{\mathcal{C}}_t^{\mathcal{U}}\right)\right] \tag{10}$$

其中 $\bar{\mathcal{C}}_t^{\mathcal{R}}$和 $\bar{\mathcal{C}}_t^{\mathcal{U}}$表示在批次内计算的归一化成本，分别由 λ_r λ_u λ_r = 和 0.25 $\lambda_u = 1.0$ ρ ϕ = $\frac{\pi_\phi \boldsymbol{a}_t|\boldsymbol{o}_t}{}$ 加权，且满足 。概率比 ()() $\pi_{\phi_{\text{old}}}(\boldsymbol{a}_t|\boldsymbol{o}_t)$ 及其裁剪机制与标准PPO相同。减去的成本项作为内在惩罚，提高那些能够降低预期碰撞可能性和预测不确定性的动作的概率。算法 1 展示了完整的训练流程：轨迹采样与优化。

TABLE II：**定量操作性能**。在 10,000 个测试用例上的成功率（↑）。

组	方法	均匀			对抗		
		0.5^3	0.75^3	1.0^3	0.5^3	0.75^3	1.0^3
G1	无遮挡推送	92.04	88.10	89.49	86.57	79.41	74.19
	带遮挡推动（基线）	74.20	66.39	42.34	65.25	48.86	22.90
	CURA-PPO（本文方法）	**88.73**	**81.61**	**80.13**	**77.12**	**74.59**	**72.42**
G2	带置信度图的基线	80.63	68.77	63.74	69.65	63.68	54.13
	CURA 不考虑不确定性	84.57	69.35	64.27	70.55	64.08	57.80
	CURA 不考虑风险	86.09	79.93	78.45	78.20	70.06	67.82
G3	仅底盘推动	66.33	47.45	40.18	56.97	30.94	17.64
	CURA-PPO（含底座）	86.75	72.20	50.08	76.41	57.31	31.03

• 0.5^3，0.75_3、1.0_3 表示被操作长方体物体的尺寸。• 为便于说明，分别介绍每个组（G1、G2 和 G3）。

VI. 实验结果

我们通过大量实验评估*CURA-PPO* 在非抓取式操作过程中应对物体诱发遮挡的效果。评估包括与两种基准方法进行比较，以量化遮挡导致的性能下降程度以及我们的方法缓解这一挑战的能力。我们还进行了消融实验，分析各个组件的贡献，并对学习到的行为进行定性分析。最后，我们考察了移动操作相较于仅基座推动在降低不确定性方面的优势。

A. 实验设置

实验在第 V-A 节所述的仿真环境中进行，并通过系统性变化测试不同遮挡条件下的性能。我们评估了三种物体尺寸（0.5m、0.75m 和 1.0m），较大的物体会遮挡更宽的传感器区域，从而造成更严重的遮挡。如图 3 所示，障碍物（0.4m、0.6m 和 1.0m 长方体）随机生成在物体初始位置前方 3–11m 的区域内，以留出足够间隙，避免即时碰撞。我们测试了两种障碍物场景：（1）均匀，障碍物随机出现在工作空间各处；（2）对抗，障碍物沿物体到目标路径布置，因此需要主动感知和碰撞规避。在这两种场景中，障碍物的位置和生成时间均为随机，以模拟操作过程中出现意外障碍物的情况。发生碰撞或经过 50 秒后，回合即终止；成功的定义是在未发生碰撞的情况下到达目标。每种物体尺寸与障碍物场景的组合，其结果均取 10,000 个回合的平均值。

B. 操作性能评估

我们将*CURA-PPO*与两种基准策略进行比较，以评估遮挡引起的性能下降程度，并验证我们的方法的有效性：

- 无遮挡推送：一种在如下设定中训练的策略：通过过滤激光雷达射线交点，人为去除被操作物体造成的所有遮挡，仅依赖任务奖励和时序传感器累积。它作为性能上界，代表在无遮挡约束下可达到的性能。

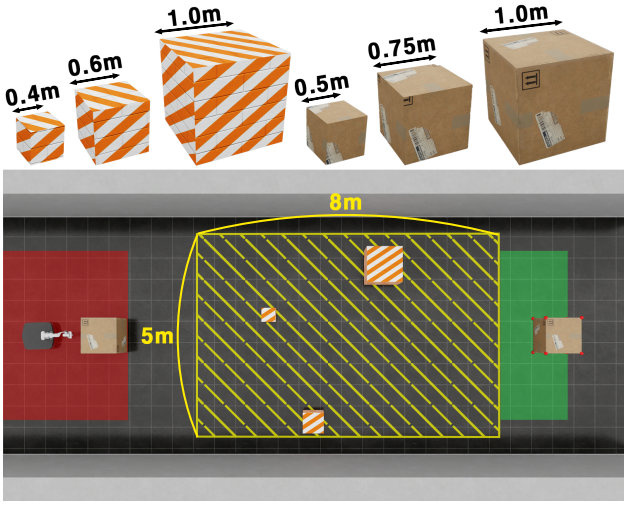

图 3：**用于训练和评估的环境**。物体的初始位姿从红色区域中采样。障碍物在黄色区域内随机生成，目标位姿从绿色区域中采样。物体和障碍物均为尺寸各异的长方体，每个回合中会出现 0 至 6 个障碍物。左上方显示一个障碍物，右上方显示物体。

- 带遮挡推动：一种在真实遮挡条件下训练的基线策略，仅依赖任务奖励，不使用置信度图或显式的不确定性处理。

表 II（G1 组）清楚地展示了遮挡对操作性能的影响，并验证了我们的方法的有效性。无遮挡推送在所有物体尺寸和障碍物场景下均持续取得较高性能，证实在感知不受遮挡时能够成功操作。相比之下，带遮挡推动的性能显著下降，成功率随着物体尺寸增大而大幅降低。这种性能下降在对抗场景中尤为严重，因为遮挡严重妨碍了障碍物检测。

*CURA-PPO*恢复了因遮挡而损失的大部分性能，相较基线方法取得了显著提升，尤其是在具有挑战性的场景中。当发生遮挡时，我们的方法利用 DCE 的分布估计，在降低碰撞风险的同时主动减少遮挡区域中的不确定性。因此，*CURA-PPO*的性能几乎可与无遮挡推送相媲美，表明显式建模风险与不确定性能够有效缓解部分可观测性下的操作问题。

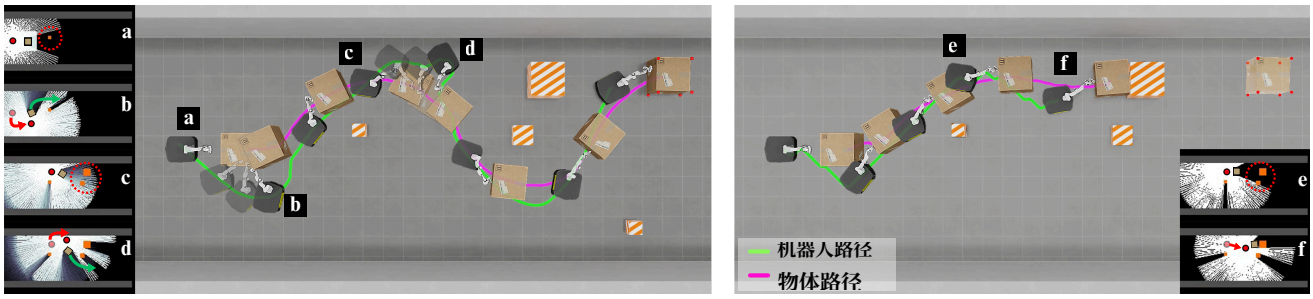

图 4：**遮挡条件下的定性操作行为**。机器人路径以绿色 表示，物体路径以洋红色表示。（左）*CURA-PPO* 及其对应的置信度图。机器人主动横向调整位置（a→b、c→d）以显露遮挡区域，从而成功检测并避开障碍物，将物体推向目标。（右）基线策略缺少置信度图和基于 DCE 的代价，无法探索遮挡区域，导致与一个隐藏在被操作物体后方、未被检测到的障碍物发生碰撞（e→f）。由于缺少这些不确定性感知机制，机器人沿直线路径前进，最终导致任务失败。

C. CURA-PPO 组件分析W

e con d造一个blati研究 di以分析 l h

三个关键组件——置信度图、风险成本和不确定性成本——的独立贡献，结果见表II (G2)。

1）置信度图评估：我们评估带置信度图的基线，这是一种不包含基于 DCE 的代价、但融入置信度图的变体。

置信度图在应用时间衰减的同时累积过去的传感器读数，使其能够区分最近观测过的区域与过时信息。这种时间感知能力提升了相较于标准基线（即带遮挡推动）的性能，尤其是在遮挡影响范围更广的大型物体场景中。结果表明，捕捉观测可靠性能够为决策提供指引，但如果没有明确的不确定性驱动行为，性能提升仍然有限。

2）风险与不确定性成本评估：我们通过测试在保留置信度图的同时分别排除每个组件的变体，考察风险成本和不确定性成本各自的影响。无不确定性成本的 *CURA* 优于带置信度图的基线，但与*CURA-PPO*相比，提升有限，因为它缺乏引导主动感知以消除遮挡的机制。无风险*CURA* 取得了更高的成功率，因为不确定性降低促进了信息寻求行为，从而应对部分可观测性带来的挑战。

这些结果揭示了各组成部分重要性的层次关系。不确定性成本更为关键，它直接促使机器人通过主动传感减少遮挡——这是在部分可观测性下实现安全导航的基本要求。风险成本则提供了互补的碰撞规避信号，使安全裕度超越
仅靠任务奖励所能达到的水平。虽然任务奖励会通过回合终止隐式惩罚碰撞，但风险成本提供了连续的梯度信号，引导机器人远离潜在碰撞，避免事态发展到危急程度。

结合两种成本（即，*CURA-PPO*）能够取得最佳性能，其中不确定性成本驱动主动感知，风险则降低预期碰撞可能性。这种互补效应在具有大型物体和对抗性障碍物的复杂场景中尤为重要，因为在遮挡条件下成功操作，主动的信息收集和碰撞避障都不可或缺。

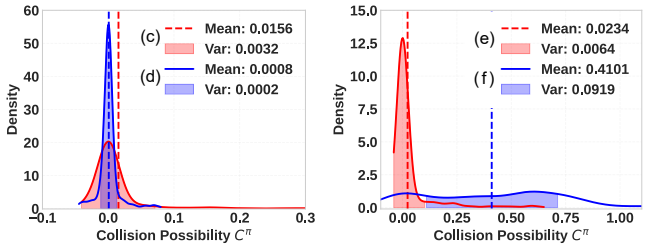

图 5：**DCE预测的可视化**，对应图 4 中的时间步。（左）*CURA-PPO* 通过主动感知将方差从 c[红色] 降低至 d[蓝色] 。（右）基线的方差始终较高，从 e[红色] 到 f[蓝色], , 最终导致碰撞。

D. 操作行为的定性分析

我们分析操作轨迹及相应的 DCE预测（图4 和 5），以考察我们的方法如何在遮挡条件下形成主动感知策略，并将 *CURA-PPO* 与基线策略（即 带遮挡推动）进行比较，从而揭示不同的操作行为。

图4 展示了 *CURA-PPO* 与基线之间截然不同的操作策略。在序列 (a→b) 和 (c→d) 中，*CURA-PPO* 会针对感知图中的低置信度区域采取横向运动，偏离直接路径，从而发现未检测到的障碍物，同时保持对物体的控制。这些降低不确定性的动作虽然看似不利于即时进度，却能在操作过程中实现避障。基线策略 (e→f) 缺少不确定性信号，因此会持续直接推向目标，并与藏在被操作物体后方的意外障碍物发生碰撞。这种行为差异凸显了不确定性感知决策如何影响操作策略。

作为图4的补充，图5中的 DCE预测为不确定性感知主动感知提供了定量证据。在 (c) 处，由于信息有限，估计的分布呈现出较高的方差。我们的方法通过减少不确定性的动作，在 (d) 处降低了方差，这表明主动感知能够降低碰撞可能性方面的 DCE不确定性。相比之下，基线在 (e) 处也面临类似情况，却未采用 DCE引导便继续行动。这种单目标策略使方差在 (f) 处仍然居高不下，最终导致碰撞。这些观测表明，DCE预测能够在遮挡情况下为安全操作提供可靠引导。

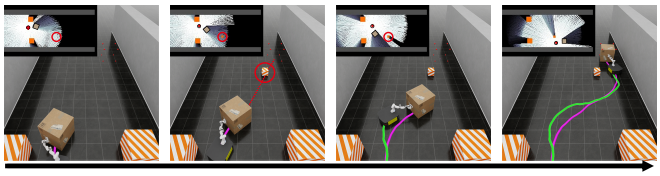

图 6：**对抗场景中的操作行为**。从左到右的连续快照展示了*CURA-PPO*对突然出现在此前空旷区域中的对抗性障碍物的响应。绿色和洋红色路径分别表示机器人和物体的轨迹。

在*CURA-PPO* 的主动感知能力在对抗场景中尤为重要，因为障碍物出现在直接推动路径上。图6 展示了在先前观测区域中出现的障碍物如何带来关键挑战——推物体时，它们会被操作对象遮挡。*CURA-PPO* 通过其不确定性感知策略成功应对了这一挑战：置信度图会在空间信息变得不可靠时发出信号，从而触发横向运动以显露被遮挡的障碍物。这种主动检测使机器人能够实时调整推动方向，引导物体安全绕过障碍物，抵达目标。

E. 运动学冗余在操作中的作用

我们使用仅底盘平台（不带机械臂）进行了额外的消融研究，以评估运动学冗余如何影响遮挡条件下的操作性能。表II（G3）展示了在遮挡条件下训练的两种配置：仅底盘推动，一种采用任务奖励的仅底盘智能体；以及带基座*CURA-PPO*，它将我们的不确定性感知框架应用于同一平台。

结果表明，仅底盘操作的表现始终不如配备机械臂的配置。由于缺乏运动学冗余，机器人在保持接触的同时难以减少物体引起的遮挡，因此常常在物体前方路径位于未观测区域内或附近时继续前进。尽管带基座*CURA-PPO*优于仅底盘基线，但随着物体尺寸增大、遮挡加剧，这些优势会减弱。如图7所示，该示例展示了仅底盘方法的一次失败。相比之下，利用运动学冗余，移动操作机器人能够在保持接触的同时改善视野，从而成功检测并避开障碍物。

如需更直观地了解实验结果，请参阅补充视频。

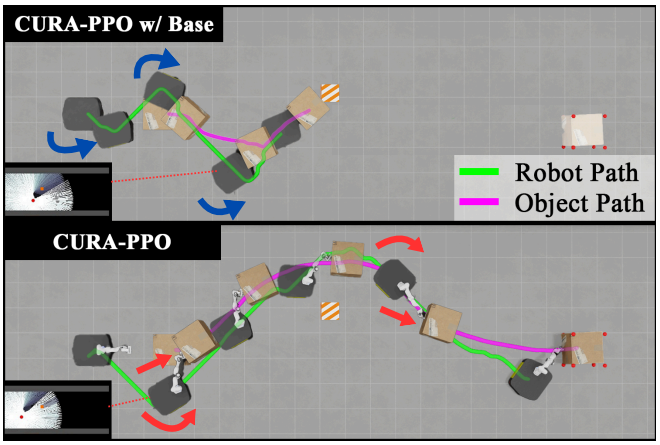

图 7：**仅底盘平台与移动操作机器人系统的行为对比**。（上）仅底盘平台因视角受限而失败，在将物体推向目标的过程中视觉覆盖范围不足。（下）移动操作机器人成功利用运动学冗余，在调整视角以实现有效感知的同时保持与物体接触。

VII. 结论

我们提出了*CURA-PPO*，这是一种新颖的强化学习框架，仅依靠机载感知来应对非抓取式操作中的物体引起的遮挡挑战。通过利用分布式碰撞估计器（*DCE*）将碰撞可能性建模为一种分布，我们的方法提取风险 和不确定性 信号，以增强标准策略优化。这些互补的代价促成了各异但相辅相成的行为：风险 推动碰撞规避，而不确定性则促使机器人主动感知，在操作过程中消除遮挡。在不同物体尺寸和障碍物配置下进行的实验表明，*CURA-PPO* 在均匀和对抗 场景中均能实现稳健的性能。定性分析表明，所学策略能够在维持物体控制的同时，有策略地调整运动以收集环境信息，从而缓解遮挡带来的挑战。未来的工作将重点在真实机器人系统上部署这一方法，并扩展该框架以纳入 RGB-D 相机等其他传感器模态，从而增强复杂真实场景中的环境感知能力。

参考文献

[1] Fabio Ruggiero、Vincenzo Lippiello 和 Bruno Siciliano，《非抓取动态操作：综述》，*IEEE*机器人与自动化快报，第 3 卷第 3 期，第 1711–1718 页，2018 年。[2] Nils Dengler、David Großklaus 和 Maren Bennewitz，《在杂乱场景中学习目标导向的非抓取推动》，载于*2022*年*IEEE/RSJ*智能机器人与系统国际会议（*IROS*）。IEEE，2022 年，第 1116–1122 页。[3] Nils Dengler、Juan Del Aguila Ferrandis、Jo˜ao Moura、Sethu Vijayaku-mar 和 Maren Bennewitz，《通过基于位置的注意力学习在杂乱场景中进行目标导向的物体推动》，*arXiv*预印本 *arXiv:2403.17667*，2024 年。[4] Juan Del Aguila Ferrandis、Jo˜ao Moura 和 Sethu Vijayakumar，《通过具有多模态分类探索的强化学习进行非抓取平面操作》，载于*2023*年*IEEE/RSJ*智能机器人与系统国际会议（*IROS*）。IEEE，2023 年，第 5606–5613 页。

[5] Seunghun Jeon、Moonkyu Jung、Suyoung Choi、Beomjoon Kim 和 Jemin Hwangbo，《学习四足机器人全身操作》，IEEE机器人与自动化快报，第 9 卷第 1 期，第 699–706 页，2023 年。[6] Yuming Feng、Chuye Hong、Yaru Niu、Shiqi Liu、Yuxiang Yang、Wenhao Yu、Tingnan Zhang、Jie Tan 和 Ding Zhao，《学习用于长时程四足推动的多智能体运动操作》，arXiv预印本 arXiv:2411.07104，2024 年。[7] Yujie Tang、Hai Zhu、Susan Potters、Martijn Wisse 和 Wei Pan，《使用非完整移动基座递送笨重物体：一种稳定推动方法》，IEEE机器人与自动化快报，第 8 卷第 11 期，第 7727–7734 页，2023 年。[8] Idil Ozdamar、Doganay Sirintuna、Robin Arbaud 和 Arash Ajoudani，《在黑暗中推动：使用触觉反馈的移动机器人反应式推动策略》，IEEE机器人与自动化快报，2024 年。[9] Ioannis Dadiotis、Mayank Mittal、Nikos Tsagarakis 和 Marco Hutter，《通过无模型约束强化学习实现移动操作机器人的动态物体目标推动》，arXiv预印本 arXiv:2502.01546，2025 年。[10] Xuran Pan、Zihang Lai、Shiji Song 和 Gao Huang，《ActiveNeRF：通过不确定性估计学习观察位置》，载于《欧洲计算机视觉会议》。Springer，2022 年，第 230–246 页。[11] Georgios Georgakis、Bernadette Bucher、Anton Arapin、Karl Schmeck-peper、Nikolai Matni 和 Kostas Daniilidis，《不确定性驱动的探索与导航规划器》，载于《2022年国际机器人与自动化会议（ICRA）》。IEEE，2022 年，第 11295–11302 页。[12] Liren Jin、Xieyuanli Chen、Julius R¨uckin 和 Marija Popovi´c，《Neu-nbv：使用基于图像的神经渲染中的不确定性估计进行下一最佳视角规划》，载于《2023年IEEE/RSJ智能机器人与系统国际会议（IROS）》。IEEE，2023 年，第 11305–11312 页。」[13] Nils Dengler、Jesper MA˘zcke、Rohit Menon 和 Maren Bennewitz，《通过不确定性感知的动作选择实现高效的操作增强语义建图》，arXiv预印本 arXiv:2506.02286，2025 年。[14] Shengmiao Jin、Yuchen Mo 和 Wenzhen Yuan，《学习“二次猜测”：一种用于估计任意物体质心的主动感知方法》，arXiv预印本 arXiv:2502.02663，2025 年。[15] Mykel J Kochenderfer，《不确定性下的决策：理论与应用》，麻省理工学院出版社，2015 年。[16] Marc G Bellemare、Will Dabney 和 R´emi Munos，《强化学习的分布视角》，载于《国际机器学习会议》。PMLR，2017 年，第 449–458 页。[17] Will Dabney、Mark Rowland、Marc Bellemare 和 R´emi Munos，《使用分位数回归的分布式强化学习》，载于《AAAI人工智能会议论文集》，2018 年，第 32 卷。[18] Alexander Amini、Wilko Schwarting、Ava Soleimany 和 Daniela Rus，《深度证据回归》，《神经信息处理系统进展》，第 33 卷，第 14927–14937 页，2020 年。[19] Yarin Gal 和 Zoubin Ghahramani，《将随机失活作为贝叶斯近似：在深度学习中表示模型不确定性》，载于《国际机器学习会议》。PMLR，2016 年，第 1050–1059 页。[20] Balaji Lakshminarayanan、Alexander Pritzel 和 Charles Blundell，《使用深度集成进行简单且可扩展的预测不确定性估计》，《神经信息处理系统进展》，第 30 卷，2017 年。[21] Gaon An、Seungyong Moon、Jang-Hyun Kim 和 Hyun Oh Song，《具有多样化 Q 集成的不确定性离线强化学习》，《神经信息处理系统进展》，第 34 卷，第 7436–7447 页，2021 年。[22] Benjamin Rosman、Majd Hawasly 和 Subramanian Ramamoorthy，《贝叶斯策略复用》，《机器学习》，第 104 卷第 1 期，第 99–127 页，2016 年。[23] Felipe Leno Da Silva 、Pablo Hernandez-Leal、Bilal Kartal 和 Matthew E Taylor，《为深度强化学习智能体提供不确定性感知的动作建议》，载于《AAAI人工智能会议论文集》，2020 年，第 34 卷，第 5792–5799 页。[24] Ralf Roemer、Armin Lederer、Samuel Tesfazgi 和 Sandra Hirche，《基于概率语义分割的基于视觉的不确定性感知运动规划》，IEEE机器人与自动化快报，第 8 卷第 11 期，第 7825–7832 页，2023 年。[25] Shivin Dass、Jiaheng Hu、Ben Abbatematteo、Peter Stone 和 Roberto Mart´ın-Mart´ın，《学习观察：为决策寻求信息通过策略分解实现”，*arXiv* 预印本 *arXiv:2410 18964*，2024。。[26] John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Radford 和 Oleg Klimov，《近端策略优化算法》，*arXiv* 预印本 *arXiv:1707.06347*，2017。[27] John Schulman、Philipp Moritz、Sergey Levine、Michael Jordan 和 Pieter Abbeel，《使用广义优势估计进行高维连续控制》，*arXiv* 预印本 *arXiv:1506.02438*，2015。[28] Hyeongyeol Ryu、Minsung Yoon、Daehyung Park 和 Sung-Eui Yoon，《基于深度强化学习的传感器遮挡条件下置信度导航》，载于*2022*年国际机器人与自动化会议（*ICRA*）。IEEE，2022年，第8231–8237页。[29] M. Mittal 等，《Orbit：面向交互式机器人学习环境的统一仿真框架》，*IEEE*机器人与自动化快报（*RA-L*），2023年。[30] Roger Koenker，《分位数回归》，第38卷，剑桥大学出版社，2005年。[31] Lukas Schneider、Jonas Frey、Takahiro Miki 和 Marco Hutter，《使用分布式强化学习学习风险感知四足运动》，载于*2024*年*IEEE*国际机器人与自动化会议（*ICRA*）。IEEE，2024年，第11451–11458页。