基于深度Q网络的弹性无人机通信：消除一阶马尔可夫干扰机

Andrii Grekhov[1*]，hrekhov.andrii@npp.kai.edu.ua，ORCID：0000-0001-7685-8706，Volodymyr Kharchenko[2], volodymyr.kharchenko@npp.kai.edu.ua，ORCID：0000-0001-7575-4366 Vasyl Kondratiuk[3], vasyl.kondratiuk@kai.edu.ua，ORCID：0000-0002-5690-8873

1*，2, 3 航空航天研究与教育中心，国立大学“基辅航空学院”，乌克兰基辅，Liubomyra Huzara 大道 1 号，邮编 03058。

通讯作者。1 hrekhov.andrii@npp.kai.edu.ua

摘要。本文提出了一种基于深度强化学习（DRL）的解决方案，利用跳频扩频（FHSS）技术在16信道无线电环境中实现抗干扰通信。基于深度Q网络（DQN）的无人机发射机在面对一阶反应式干扰时，持续选择下一个跳频信道；此类干扰利用观测到的状态转移统计数据预测并中断传输。通过自主训练，所提出的智能体学会了一种均匀随机跳频策略，可有效抵消干扰的预测优势。

在瑞利衰落和加性噪声的影响下，本文系统评估了前向纠错博斯–乔杜里–霍昆格姆（BCH）型编码（具有 t = 0–10 个可纠正错误）的作用，结果表明，即使是适度的冗余（t = 1–2）也能显著减少丢包。对学习动态、信道利用率分布、 ε-贪心衰减、累积奖励、误码率/信噪比变化以及详细丢包表格的广泛可视化分析，证实系统收敛到近乎最优的干扰策略。研究结果为现代电子战场景中的自主弹性通信提供了实用框架。

关键词：无人机、干扰、深度Q网络、强化学习、自适应防御、跳频

引言

无人飞行器（UAV）在民用和军事领域的快速发展，大幅提高了主动电子战环境下对通信信道可靠性和抗干扰能力的要求。现代反应式（跟踪式）干扰器能够检测传输信号，并立即将功率集中到正在使用的信道上。其中，一阶马尔可夫反应式干扰器尤具威胁：它们通过观察跳频序列，实时构建转移概率矩阵，并日益准确地预测发射机的下一个信道。

采用伪随机序列（m序列、密码生成器）甚至真随机数硬件源的经典跳频扩频（FHSS）系统，从根本上容易受到此类学习型对手的攻击；在典型条件下（干扰信号比为 ≈ 1–2），无法确保所有数据包均成功传输。深度强化学习（DRL）的最新进展为构建完全自主的抗干扰通信系统开辟了全新的可能性。与基于预定跳频序列的传统方法不同，基于 DRL 的发射机能够在与干扰环境交互的过程中直接自主学习自适应策略，而无需预先了解对手模型。

尽管基于 DRL 的解决方案不断涌现，但大多数方案都采用复杂架构（双重深度Q网络、决斗深度Q网络、循环深度循环Q网络、多智能体博弈），这显著增加了计算复杂度，使其难以在小型无人机上实现。

本文研究一种轻量级的经典深度 Q 网络（基础版 DQN），采用对当前信道进行独热编码并对奖励函数进行最小限度的修改，能否生成最优跳频策略，从而彻底消除一阶马尔可夫反应式干扰器的影响。

我们研究一种慢速16信道跳频扩频系统，该系统在瑞利衰落和加性高斯白噪声环境下运行，并受到干信比分别为0.5和1.2的反应式干扰器攻击。干扰器观测先前的状态转移，并实时构建精确的一阶马尔可夫模型。发射机不了解干扰策略，只能通过试错进行学习。

本研究的目标是：

- 证明采用小幅多样性奖励的传统单智能体DQN会收敛到近乎均匀的信道选择策略，

- 在给定干扰环境下，量化采用和未采用博斯–乔杜里–霍昆黑姆（BCH）编码时的传输成功率、误码率（BER）、信噪比（SNR）和丢包概率（PLR）。

本文其余部分的组织结构如下。第2节介绍相关工作。第3节提出用于跳频扩频（FHSS）和反应式干扰器的深度Q网络数学模型。第4节介绍算法。第5节讨论仿真结果。第6节进行讨论。文章末尾给出结论。

1. 相关工作

文献 [1], 以认知无人机为例，提出了一种用于认知无线电干扰缓解的新型资源分配策略。该文提出将主动广义动态贝叶斯网络作为环境模型，用于联合编码物理信号动态以及无人机与频谱干扰器之间的动态交互。该方法将动作与规划视为贝叶斯推断问题，并在在线学习过程中通过避免意外状态（异常最小化）来求解。仿真结果证实了所提方法在异常最小化（奖励最大化）方面的有效性，并表明与传统跳频和Q学习方法相比，该方法具有较高的收敛速度。

文献 [2], 提出了一种基于强化学习的无人机中继策略，用于海事通信抗干扰。该方案基于先前的传输特性、中继位置、接收信号强度和接收干扰功率，优化无人机轨迹与中继功率，以节省能耗并降低海事信号的误码率。此外，该文还提出了一种深度强化学习方案，设计了采用对决架构的深度神经网络，以提升通信质量并降低计算复杂度。基于抗干扰博弈的纳什均衡，确定了信干噪比、能耗和通信效用的性能界限，并分析了所提方案的计算复杂度。仿真结果表明，与基准方案相比，所提方案提高了能源效率并降低了误码率。

论文 [3] 提出了一种名为 JaX 的新方法，用于在传统扩频及其他干扰缓解方法不足以应对的情况下检测并抑制强干扰器。JaX 无需显式探测、训练序列、信道估计或发射机交互。研究为多天线系统开发了一种卷积神经网络，用于检测干扰器是否存在、干扰发射信号的数量及其对应相位。这些信息会持续输入一种用于抑制干扰信号的算法。研究开发了一个配备两根天线的原型系统，并通过软件定义无线电平台，在各种环境条件和调制方案下评估该方法。面对干扰信号特性、干扰功率及时序模式各异的多种干扰器，JaX 均表现出良好的鲁棒性。

无人机通信网络容易受到恶意干扰和同频干扰的影响，从而导致网络性能下降。因此，研究能够提升通信安全的抗干扰方法已成为一项严峻挑战。论文 [4], 提出了一种适用于多信道网络中多架无人机的新型抗干扰信道选择方案。首先，将抗干扰问题建模为部分可观测随机博弈（POSG）：在马尔可夫干扰器的影响下，具有部分可观测性的无人机对需要争用数量有限的通信信道。为确保能够快速适应动态干扰环境，研究提出了元平均场质量（MMFQ）学习算法，为 POSG 问题提供基于纳什均衡（NE）的解决方案。仿真结果表明，与基准算法相比，所提算法能够取得更高的平均表现，从而提升吞吐量和资源利用率，尤其适用于大规模无人机通信网络。

随着无人机在军事和民用领域日益普及，确保其信号的可靠性和安全性至关重要。近年来，深度强化学习已证明是应对这一挑战的有力工具。论文 [5] 旨在分析深度学习在对抗无人机通信系统干扰方面的潜力。传统的干扰对抗措施，如跳频和直接序列扩频（DSSS），在一定程度上是有效的。然而，由于其单一模式特性和较低的频谱利用率，它们难以应对动态变化的干扰环境。另一方面，深度学习在自动特征提取和模式识别方面具有强大能力，尤其是深度强化学习能够使无人机实时适应环境变化，并自动优化通信策略。本文研究了基于强化学习和深度强化学习算法的抗干扰方法，并提出了一种基于生成对抗网络（GAN）的无人机通信新型抗干扰策略。尽管深度学习为对抗无人机通信信号中的干扰提供了新的机遇，但在实际应用中仍面临计算需求高、模型训练复杂等挑战。

本综述 [6] 旨在总结过去十年文献中针对各种民用和国防领域（包括地面系统和卫星系统）所描述的少量窄带干扰（NBI）模型及缓解方法。文中讨论了窄带干扰的来源及其对各种应用和技术可能造成的负面影响，并考察了一系列缓解策略，包括传统的基于滤波器的方法和运用机器学习的现代方法。本综述仅限于时频窄带干扰，不涉及

空间域中的干扰问题（考虑到迄今已发表的文献数量）。文中还讨论了若干已发现的研究空白，并列出了窄带干扰缓解的未来研究方向。本综述介绍了窄带干扰问题，可为新兴应用和技术的后续研究提供参考。

干扰攻击是一种会导致拒绝服务的网络威胁，通常发生在飞行无线网络（FANETs）和无人机互联网（IoD）等无线通信系统中。多年来，研究人员提出了多种干扰攻击检测方法，例如贝叶斯博弈论机制、基于IoD的防御机制、通信信道方法（信道跳频、扩频、基于MIMO的干扰抑制、编码等）、延迟容忍网络方法以及密码算法。然而，这些方法并不适用于无人机环境中的干扰检测。主要问题涉及传输效率、处理时间、准确率、功耗、飞行范围和飞行自主性。在[7], 中，提出了一种采用基于梯度监控的强化学习机制（RLGM）的干扰攻击检测方法。RLGM维护安全区域并降低目标学习的梯度方差，从而确保学习目标具有更高的准确率。RLGM能够实现快速的学习进程，并在训练阶段精准选择网络所需的参数集合。RLGM在训练过程中自动推断所需的深度网络规模，并使用自动保持不变的学习得到的权重。所提出的方法优于其他强化学习方法，例如联邦强化学习、深度求积学习（DQL），以及GA-AOMDV等非机器学习方法。

在 [8], 中，针对无人机网络提出了一种新颖的智能干扰对抗框架。多个无人机—无人机链路对力求最大化总速率，同时将功耗降至最低；每架无人机都会自适应地调整传输信道和分配功率，以避免共享信道上的智能干扰和其他干扰。干扰器试图通过自适应地改变干扰信道和功率，破坏无人机网络的链路质量。该干扰对抗问题被建模为随机斯塔克尔伯格博弈，其中智能干扰器是领导者，无人机对是跟随者。本文提出强化学习（RL）算法，以确定博弈中每个智能体的最佳响应策略。我们采用深度Q网络（DQN）算法在各站点检测干扰，并提出一种具备检测支持的去中心化联邦DQN算法，以抑制无人机对中的联合策略干扰。分析结果表明，与独立DQN算法相比，所提算法的抗干扰效率提高了23.3%。

频谱稀缺、频谱效率、功率限制和干扰是无线网络面临的主要挑战。认知无线电网络（CRNs）能够在授权频段可用时实现共享。次级用户（SUs）要确保较高的数据速率，就必须高效利用频谱；而次级用户移动性使功耗成为无线网络中的一个重要因素。由于无线环境开放，干扰很容易降低网络性能并中断连接。一项研究 [9] 旨在提升认知无线电网络性能，并在存在智能干扰器的情况下，为 SUs 建立更可靠的连接，同时确保频谱得到高效利用。研究提出了一种利用跳频对抗干扰的方法。假设 SUs 会监测频谱可用性和信道增益。随后，SUs 学习干扰器的行为，并针对数据信道和控制信道的数量选择合适的策略，从而共同优化频谱效率和功耗。SU 与干扰器之间的交互被建模为随机零和博弈，并采用强化学习（RL）求解该博弈。仿真结果表明，信道增益较低时，SU 会选择较多的数据信道；而信道增益较高时，SU 会增加控制信道的数量，以确保连接更加可靠。考虑到频谱效率，SU 会减少所用信道的数量以节省能耗。与短视学习和随机策略相比，所提出的策略性能更优。在干扰攻击下，SU 会选择适当数量的控制信道和数据信道，以确保连接可靠、高效且持久。

无人机通信系统在动态对抗环境中面临日益严峻的多源干扰挑战，对其可靠性和鲁棒性提出了更高要求。为应对这些问题，基于智能体的自主干扰方法已成为重要研究领域。论文 [10] 对无人机通信干扰中的智能体概念进行了全面综述，并提出了一种基于感知-决策-行动（P-D-A）范式的闭环决策系统。在这一框架下，论文考察了关键技术，尤其着重探讨了博弈论在无人机与干扰机交互建模中的应用，以及如何整合智能强化学习算法来制定自适应干扰策略。文中还讨论了现有方法的潜在局限，指出了关键工程挑战，并概述了未来研究的有前景方向。

论文 [11] 概述了通信领域智能干扰抑制方法的发展历程。文中定义了这一概念，并探讨了通信领域智能干扰抑制能力的主要特征。论文还介绍了通信领域智能干扰抑制系统的初始架构，并梳理了通信领域智能干扰抑制方法的发展过程，从早期的自适应干扰抑制方法，到主要基于博弈论和机器

学习的智能干扰抑制领域最新进展。文中分析了该领域最新的研究成果，并指出了现存问题。最后，提出了智能干扰抑制领域未来研究的若干有前景的方向。

在 [12], 中，本文针对高度动态智能干扰带来的威胁，提出了一种基于近端策略优化（PPO）的智能干扰算法。本文将高度动态干扰下的通信抗噪性问题建模为多参数马尔可夫决策过程（MDP），随后采用近端策略优化算法，确定系统信道、时间间隔、传输速率和传输功率的最优联合干扰策略。该方法提高了算法的收敛速度，并缩短了达到最优策略所需的时间。仿真结果表明，所提算法能够同时有效调整多个通信参数；与现有基于Q学习和深度Q网络的干扰方法相比，该算法在分组接收率和收敛速度方面展现出更优的统计性能。

深度强化学习广泛应用于无线通信抗干扰问题。然而，这类方法假设通信系统能够获取完整的信道状态信息（CSI）。在信道状态信息受限的情况下，本文在[13] 中采用部分可观测马尔可夫决策过程对系统进行建模。本文提出一种自动调整搜索速率滚降因子的算法，并进一步设计了深度循环Q网络算法架构和智能抗干扰决策算法。该算法首先利用长短期记忆网络学习输入数据的时序特征，对这些特征进行均衡处理，然后将结果输入全连接层，以获得智能抗干扰策略。仿真结果表明，在初始搜索速率滚降因子设得较大时，自动调节搜索速率滚降因子的算法能够实现近乎最优的性能。

论文 [14] 提出了一种基于深度强化学习（DRL）和价值分解网络（VDN）的多智能体频谱接入控制方法，并采用集中式训练、分布式执行架构。经过集中式地面训练后，该模型被部署到各颗卫星上，自主实时作出频谱接入决策。仿真结果表明，所提出的方法能够有效平衡训练成本与干扰抑制效果。

报告 [15] 探讨了如何防御针对无人机通信信道的自适应人工智能干扰。报告提出了一种对抗模型：无人机上的防御人工智能使用Q学习进行动态频率切换，而攻击人工智能则采用状态转移预测模型。仿真结果表明，即使攻击者采取激进的自适应策略，强化学习智能体仍能确保通信稳定。

2. 跳频扩频与反应式干扰器的深度Q网络数学模型

针对带有马尔可夫干扰器、瑞利衰落和加性高斯白噪声信道的跳频扩频场景，建立了单智能体DQN的问题数学模型（图1）。以下公式构成核心部分，重点关注信噪比/误码率、奖励、熵和成功率。

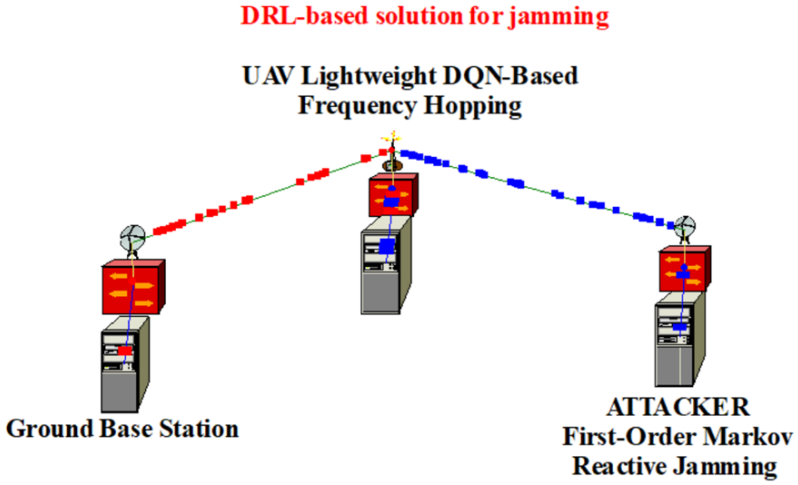

图 1. 模型“地面基站–无人机–攻击者”

该场景是：DQN智能体管理具有 N=16 个信道的跳频扩频，以便在受争用频谱中传输数据包。干扰器具有反应性，遵循一阶马尔可夫特性，并以 0.8 的概率切换到智能体上一个使用的信道。跳频速度较慢（每个数据包跳频一次）。信道呈瑞利衰落，干扰信号比（JSR）在 0.5 到 1.2 之间变化。成功定义为跳频与干扰不发生重叠。

该参数 t 是离散时间模型中的时间步。它是表示事件序列中某个时间点的索引。在该模型中，智能体在每个时间步与环境交互：观测状态、选择动作、接收奖励，然后转移到下一个状态。该模型是离散的，将时间划分为“时钟周期”，以模拟序贯决策——信道切换。模型具有$N = 16$ 个信道（编号为 1 至 16）。当前状态是长度为 16 的独热向量，其中 1 表示当前信道位置，其余元素均为 0。

动作 $a_t \sim \pi(s_t)$ 是根据策略 π （概率分布）对下一个信道的选择。干扰器被建模为一阶马尔可夫链：状态转移 $\mathrm{P}(\text{jam } c_{t+1} \mid c_t) \propto \text{count}(c_t \to c_{t+1})$，表示在 c_{t+1} 处发生干扰的概率取决于从 c_t出发的转移率。最终，智能体试图选择 c_{t+1} 以避开干扰器，但干扰具有时间相关性。

该参数 $|h|^2$ 是信道系数 h 的绝对值平方，表示给定步骤中的信道增益。在无线通信中，h 是一个复衰减系数（衰落），包含幅度和相位。h 的绝对值是幅度，而 $|h|^2$ 是信号通过信道后的瞬时信号功率。

在信噪比 $\gamma = SNR = \frac{P_s \cdot |h|^2 \cdot h|^2}{P_n + P_j \cdot I_{jammed}}$ 的公式中，P_s 是信号功率，P_n 是噪声，P_j 是干扰功率。当信道受到干扰时，参数 $I_{\text{jammed}} = 1$ 。$|h|^2$ 参数考虑了信道如何“增强”或“衰减”信号。如果 $|h|^2$ 较小，信噪比就会下降，从而降低连接质量。

信道系数 信道系数 h 的分布遵循瑞利定律，这是无线信道中瑞利衰落的标准模型。其物理意义是，在没有视距（非视距，NLOS）的信道中，信号会沿着许多反射路径（多径传播）到达。幅度 h 是这些路径的总和，其表现为一个服从瑞利分布的随机变量。这描述了波束干扰造成的衰落。参数 h 是复数：$h = X + iY$，其中 X 和 Y 是相互独立、均值为零、方差为 σ^2 的高斯随机变量。于是，幅度 $r = |h| = sqrt\{X^2 + Y^2\}$ 服从瑞利分布。功率 $\mathrm{g} = |h|^2 = X^2 + Y^2$ 服从指数分布。衰落可能导致“深衰落”，此时 $|h|^2$ 接近 0，连接中断。

在该模型中，h 在每个时刻 t 都会重新生成（彼此独立），因此 $|h|^2$ 是随机功率。信噪比也会变为随机量，误码率（BER）（对于采用相干接收的二进制相移键控调制）为：

$$BER(t) = Q\left(\sqrt{2\gamma(t)}\right) = \frac{1}{2} erfc\left(\sqrt{SNR(t)}\right),$$

其中$Q(x)$是高斯分布的尾函数 $erfc(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-u^2}\, du$ 。

丢包使用以下公式进行估算：

$$PLR = 1 - (1 - BER_k)^L \approx L \cdot BER_k \text{ for small } BER.$$

***DQN*智能体模型**

• 状态：s_t = (前置 v_信道、干扰器_历史记录）– 对16个信道进行独热编码 + 马尔可夫干扰器（维度 16 +1 = 17）。

• 动作：$a_t \in \{1, ..., 16\}$ – 选择一个信道。• Q函数：表格型DQN $Q(s, a; \theta)$，更新：

$$\theta \leftarrow \theta + \alpha[r + \gamma \max_{a'} Q(s', a'; \theta) - Q(s, a; \theta)]\nabla_\theta Q(s, a; \theta)$$

• 策略分布：带温度参数 $\tau = 1$ 的 Softmax（用于探索）：

$$\pi(a|s) = \frac{\exp(Q(s,\ a)/\tau)}{\sum_{a'} \exp(Q(s,\ a')/\tau)}$$

在我们的模型（用于信道选择的强化学习）中，Softmax 是一种关键工具，可将估计值（逻辑值）转换为各个动作（信道）上的概率分布。这对于 $\pi(s_t)$ 策略尤为重要，因为智能体会以随机方式选择下一个信道 c_{t+1}，从而平衡利用（选择最佳信道）和探索（尝试新信道）。

Softmax（或“软最大值”）是一种函数，它将一组数字（逻辑值，例如 Q分数）转换为概率分布，其中概率之和为 1，且所有值均为 ≥ 0。

强化学习模型使用该函数，是因为智能体不会总是严格地选择最佳信道，以免陷入局部最优。Softmax 使选择变得“柔和”：最佳信道被选中的频率更高，而较差的信道被选中的概率较低。这有助于鼓励探索。

在我们的模型中，对于 $N = 16$ 个信道，逻辑值向量为 $\mathrm{z} = [\mathrm{z}_1, \mathrm{z}_2, ..., \mathrm{z}_{16}]$, ，其中 z_i 是信道 i 的估值（例如预期奖励）。Softmax 得到 $p_i = \pi(a_t = i \mid s_t)$，并从该 p 中采样得到 a_t 。

Softmax 基本公式（针对$\tau = 1$）：

$$p_i = \frac{\exp(z_i)}{\sum_{j=1}^{N} \exp(z_j)}, \ i = 1, ..., N.$$

带温度参数的 *Softmax* $\tau > 0$ （温度缩放），用于控制探索：

$$p_i = \frac{\exp(z_i / \tau)}{\sum_{j=1}^{N} \exp(z_j / \tau)}$$

其作用τ如下：

• τ = 1：标准 Softmax，其平衡点在于保留差值，同时进行探索（最佳动作出现得更频繁，但并非 100% 选择）。

• τ > 1：高温度，即一种“高温”策略，将指数除以 τ，从而使其“平滑”，分布变得更加均匀。探索更多，智能体会更频繁地尝试较差的信道，以发现干扰器中隐藏的模式。

• τ < 1：低温度，此时差值会被放大。利用更多，智能体会持续选择排名靠前的信道，而忽略风险。

在我们的案例中，$\tau = 1$ 用于探索，这是一种基线学习模式。它具有足够的随机性，可以防止智能体固定在单一信道上（考虑到马尔可夫干扰器），但又不会过于混乱。这通常是起点，之后可以降低 τ，使策略收敛到最优策略。

我们的模型在强化学习框架中使用奖励函数，使智能体（信道选择系统）学习最大化每回合的累积奖励。这促使系统避免干扰（干扰机），确保良好的连接质量（较低的误码率），并鼓励探索。公式如下：

$$r_i = -2 \cdot I\{jammed\} + \frac{1}{1+BER} + \beta \cdot H(p),$$

其中：

• $I\{jammed\}$ – 指示函数（如果信道在步骤 t 处受到干扰，则为 1，否则为 0）。

• BER – 误码率，取值范围为 0（完全正确）到 1（完全混乱）。

• p – 信道上的概率分布（来自策略 π）。

• $H(p)$ – 香农熵。

• $\beta = 0.02$ – 熵奖励的系数。

每个时间步中，智能体选择信道、传输数据并接收反馈（是否受到干扰、信噪比、误码率）时，都会计算奖励。基础奖励根据公式 $\frac{1}{1+BER}$ 计算，反映信道上的数据传输质量。误码率较低（信道良好）时，奖励接近 1；误码率较高（信道较差）时，奖励接近 0.5。

选择这个公式，是因为它是单调递减函数（误码率越高，奖励越低），但并非线性函数——误码率较低时变化平缓（少量错误不会造成严重影响），误码率较高时则变化陡峭（信道故障会受到严厉惩罚）。这在通信中很常见，因为误码率与信噪比呈指数关系。

基础奖励鼓励选择信噪比较高（衰落较低 $|h|2$且没有干扰器）的信道。否则，智能体会忽略连接质量，只专注于避开干扰器。基础奖励的范围通常为 0.5–1，作为奖励的“主要”部分（其他项对其进行调整）。

干扰惩罚为 -2·I。干扰是由干扰器造成的人为干扰，它会将功率 P_j 加到信噪比分母中，使信道在实际中几乎无法使用（信噪比 → 0，误码率 → 0.5）。该惩罚是一种负向奖励，用来“惩罚”智能体选择受干扰信道：如果干扰器在c_{t+1}上处于活动状态，则 I{jammed} = 1；否则为 0。

选择 -2 作为非对称惩罚项。由于基础奖励为 max = 1（干净信道），-2 会使干扰变得“非常糟糕”（r ≈ -1），因此智能体会迫切地学习避开干扰。如果惩罚是 -0.5，智能体可能会偶尔“容忍”干扰，但 -2 会在学习过程中产生强烈的梯度。

这里的逻辑如下。在真实网络中，干扰是一场灾难（丢包、时延）。将惩罚设为基础奖励最大值的 2 倍，可确保避开干扰器成为首要任务（探索与利用的平衡）。指示函数I取值为 1 或 -2，并且仅在实际发生干扰时“触发”惩罚。这是二元的：没有灰度值——要么发生干扰（彻底失效），要么没有。该模型使用马尔可夫干扰器，因此智能体能够学习预测状态转移。这会对不良行为产生负向信号。如果没有这项惩罚，智能体可能会忽视干扰，只依赖误码率（干扰期间误码率虽高，但并非极端高）。

熵奖励 $\beta \cdot H(p)$ 是用于鼓励探索的附加项。参数 $H(p) = -\sum_{i=1}^{N} p_i \log p_i$ 是策略分布 p 的香农熵（对 16 个信道应用 Softmax）。熵衡量选择的“随机性”。参数 $\beta = 0.02$ 提供缩放。引入熵奖励是因为在强化学习中，纯粹的利用会导致局部最优（智能体会困在 1–2 个信道上，无法察觉干扰器的变化）。熵会向奖励中加入“噪声”，从而稳定学习过程。数值 $\beta = 0.02$ 是通过经验选取的，以避免其权重超过基础奖励。参数 β 的作用是在短期后果（立即避免干扰）与长期后果（适应马尔可夫干扰器）之间取得平衡。

成功率由以下公式确定：$Success = \frac{1}{T}\sum(1 - I_{a_t}) \times 100\%_{t=1}^{T}$，其中 T = 1500 个回合。

基于马尔可夫链的反应式干扰器模型

在我们的模型中，反应式干扰器是一种自适应对手，会通过在不同信道间切换来响应智能体的动作，从而产生干扰。与“愚笨”的干扰器（随机或固定）不同，反应式模型会从智能体先前的状态转移中学习，以预测并阻止其选择。该模型采用一阶马尔可夫链实现，其中干扰器状态是当前受干扰的信道，状态转移则取决于智能体的动作历史。

该模型让干扰器变得“智能”：它并非随机行动，而是具有时间相关性，这提高了强化学习智能体面临的挑战（它必须学会预测干扰器的变化）。

反应式马尔可夫干扰器的概念

反应性：干扰器通过监测智能体在信道之间的状态转移来“响应”智能体。它假设智能体存在某些模式（例如偏好相邻信道），并据此调整自身的状态转移。这模拟了现实世界中的情形。

一阶马尔可夫性质：当前状态（在信道 c_t 上进行干扰）仅通过智能体的前一状态决定下一状态（在信道 c_{t+1} 上进行干扰）的概率。不依赖更早的历史记录 —— 这是为了便于处理（计算简便性）而作出的简化。

关键要素：count($c_t \to c_{t+1}$)：干扰器会记录智能体状态转移频率的“计数器”。智能体从 c_t 转移到 c_{t+1} 的次数越多，干扰器“转移”到该信道进行阻断的概率就越高。这使干扰器能够不断演化：在仿真（强化学习回合）开始时，状态转移几乎是随机的；随着智能体逐渐学会，干扰器也会从其策略 π 中“学习”。

干扰器状态: $j_t \in \{1, ..., N\}$ — 在步骤 t被干扰的信道。

干扰器动作: 自动——根据 $j_{t+1} \sim P(j_{t+1} \mid c_t)$ 进行采样，其中 c_t 是智能体上一次的选择。它不会同时抑制所有信道（这会消耗过多能量），而是专注于智能体的“热门”路径，以尽量降低自身能耗。

数学模型

状态: 离散信道 {1, 2, ..., 16}。总状态空间为 $N = 16$。

转移矩阵: P 是一个大小为 $N \times N$ 的随机矩阵，其中 $P_{i,k} = P(j_{t+1} = k \mid c_t = i)$ 表示当智能体处于信道 i 时，干扰器转移到信道 k 的概率。

各行之和: $\sum_k P_{i,k} = 1$ 对每个 i。

矩阵更新：在每个步骤 t，矩阵 P 都会根据智能体的经验转移计数器进行更新。这使干扰器具有自适应性（而非静态）。矩阵元素的完整公式为 $P_{i,\ k} = \frac{count}{\sum count\ i \to m}$（$i \to k\ +\alpha$）/（$\overline{N}_{m=1}\ +\alpha$），其中 $\forall i,k$=1，…，N。

- α （狄利克雷先验）：一个较小的正数常量（均匀先验通常取 α=1），用于避免概率为零（零计数问题）。如果没有 α，当计数为 = 0 时，P = 0——干扰器“不了解”罕见的状态转移。狄利克雷分布基于根据各类别的先验概率为其分配权重。

4. 算法说明

程序代码使用 Python 编写。

模块 1. 导入库

加载所有必需的库：torch（用于神经网络）、numpy（数学运算）、matplotlib + pandas（用于绘图和表格）、scipy（用于互补误差函数和二项分布）。

模块 2。全局绘图样式

mpl.rcParams['字体系列'] = 'Times New Roman'

模块 3。仿真参数

NUM_信道 = 16

干扰 _功率 = 0.5 或 1.2

信号 _功率 = 1.0

噪声 _功率 = 0.05

回合 = 1500

...

多样性 _奖励 = 0.02

所有物理参数和训练参数均已设置。

模块4. DQN神经网络（基础版，3层）

class DQN(nn.Module):

Linear(16 → 128) → ReLU → Linear(128 → 128) → ReLU → Linear(128 → 16) 最简单的全连接网络。输入是当前信道的独热向量（16个元素）。输出是16个可能的下一信道对应的Q值。

模块5. 智能体类 (DroneAgentDQN)

包含:

• 策略_网络和目标_网络（用于稳定训练）• 经验回放缓冲区（可存储 10,000 个状态转移的经验缓冲区）• ε-贪心策略（从 0.9 → 0.05）• 信道使用计数器（用于计算熵）

方法:
• get_state() — 将信道编号转换为独热编码
• choose_action() — ε-贪心动作选择 • remember() — 存储状态转移（*s, a, r, s′*） • replay() — 常规深度Q网络步骤：采样批次、计算误差并更新权重 • decay_epsilon() 和 update_target() — 标准

区块6. 马尔可夫攻击者类（一阶干扰器）
self.counts = np.zeros((16,16)) # 转移矩阵 • update() — 每次状态转移后 $c_{prev} \to c_{curr}$，递增计数器 • predict() — 返回当前信道最可能的下一信道 • jam() — 主函数：更新统计数据并返回要干扰的信道

区块7. 函数 compute_ber_snr(jammed) 衰落 = *np.random.exponential(1.0)* • 衰落 = 衰落 * 0.8 + 0.3

• 信号 = 1.0 * 衰落 • 噪声 = 0.05 + (干扰_功率 * 衰落 if jammed else 0) 信噪比_lin = 信号 / 噪声 误码率 = 0.5 * 互补误差函数(sqrt(snr_lin)) 对有限深度的瑞利衰落进行仿真。计算当前时隙的瞬时信噪比和误码率。这是整个仿真的物理引擎。

区块8。主训练循环（1500 回合）
每个回合 = 一个时隙（一次跳频）： 1. 智能体观察当前状态（当前信道的独热编码）2. 选择下一个动作（ε-贪心）
3 $_{t-1}$ $_{t}$ $_{t+1}$ 。干扰器在
4
已知 c和 c的情况下，预测并干扰 j。4. 检查干扰器是否击中目标 5. 计算信噪比和误码率 6. 计算所有数据包大小的丢包率（使用和不使用前向纠错）7. 计算当前信道利用率熵 8. 生成奖励：r = -2.0（如果受到干扰），或 ≈1.0 + 0.02×熵（如果没有受到干扰） 9. 将状态转移（*s, a, r, s′*）保存到缓冲区 10. 执行深度Q网络训练步骤 11. 降低 ε，并每 100 个回合更新一次目标网络

区块9。收集各里程碑（200、400、……、1400）的统计数据
统计最后 100 个回合的平均误码率、信噪比、熵和成功率，并存储在表格中。

区块10。最终指标（最后 100 个回合）
最终_误码率、最终_信噪比、最终_熵、成功_率 =（1 – 被干扰的数据包数量 / 1500）× 100% 最终_丢包率和最终_丢包率_前向纠错 — 所有数据包大小下的平均丢包率

关键的“神奇”时刻：• 独热状态（无需记忆历史记录）。• 奖励 0。02 × 熵→ 智能体会迫使自己尽可能地不可预测。• 严厉惩罚 -2。0 → 即使是罕见的干扰器命中也会造成巨大损失。因此，智能体找到了一个非常接近均匀分布的策略——干扰器无法预测；但由于训练初期仍存在一定的利用行为，实际结果甚至更好。

5. 仿真结果

在无人机跳频扩频系统对抗反应式马尔可夫干扰器的强化学习中，训练进度（表1）全面监测智能体各项关键指标随回合的变化。这是一种诊断工具，可验证模型收敛的质量、发现问题，并确认深度Q网络算法的有效性。表1展示了稳定的改进趋势，而非随机波动。进度分为早期、中期和后期三个阶段，各项指标彼此相关：低干扰 → 高信噪比 → 低误码率 → 高熵→ 稳定利用。

累积奖励（图2）是所有回合中获得的奖励总和。这意味着智能体平均每回合获得 +0.98 的奖励，持续胜过干扰器。ε-贪心参数和熵确保 G 稳步增长，而非杂乱波动。奖励近乎线性增长，表明智能体学会了预测干扰器的状态转移，从而最大限度地减少干扰并提高信噪比/误码率。与奖励会波动的随机模型不同，这种增长源于马尔可夫干扰器适应缓慢；瑞利衰落和熵会引入噪声，而深度Q网络和经验回放可将其平滑处理。没有发生“灾难”——得益于多样性奖励，智能体避免了陷入局部最优。该指标是强化学习的“核心”，体现了智能体在与对抗性干扰器的博弈中“赢得比赛”的程度。

误码率和信噪比曲线（图3）是并列子图，用于直观展示训练期间通信信道质量的变化。这些曲线是衡量智能体策略实际效果的关键指标，展现了强化学习优化在多大程度上将系统从“混乱”（误码率高、信噪比低）转变为“稳定”状态（误码率低、信噪比高）。误码率曲线（对数刻度）显示，前400回合误码率急剧下降，随后在10^{-4}–10^{-6}附近波动。信噪比曲线则迅速上升，并稳定在12–14 dB附近。这意味着信道的物理参数已接近理想状态。

熵和ε-贪心曲线（图4）展示了智能体策略在探索与利用之间的平衡。它们以熵$H(p)$衡量信道选择的多样性，并将衰减ε 作为控制ε-贪心随机性的手段。香农熵公式$H(p) = -\sum_{i=1}^{N} p_i \, log p_i$ 包含信道被利用的经验概率。熵是策略π 中的探索度量。较高的H会促使均匀跳频（多样性），以“模糊”干扰器计数器，降低P(jam)。较低的H则会增加遭受“干扰”的风险（确定性选择）。模型中的ε 充当随机控制器。较高的ε 会增加“噪声”，以探索新路径（避免局部最优）；较低的ε 则会专注于最佳Q值。图4中的曲线是学习适应性和稳定性的指标，展示了智能体如何从“混乱游走”（熵+ ε较高）演变为“目标导向的选择”（熵+ 适中、ε较低）。熵曲线逐渐接近并几乎触及2.773处的红色虚线（最大值），而ε 则逐渐降至0.05。其理念是，智能体已学会尽可能地保持不可预测——这就是“无密钥密码学随机性”。

信道使用直方图如图 5 所示。全部 16 根柱子的高度几乎相同，表明干扰器没有统计优势——它的预测能力不比从 16 个选项中随机猜一个更好。

PLR 随数据包大小变化的关系以一张图（图 6）和两张表（表 2 和表 3）呈现，展示了数据传输的实际可靠性如何取决于数据包长度和前向纠错（FEC）等级。表 2 包含三列：数据包大小（数据包大小 L）、PLR（精确丢包概率）和 ≈ L×BER（近似值）。其中两个关键列——PLR 和 ≈ L×BER——用于比较精确计算与简单解析估计，从而突出低误码区间内 PLR 的线性近似。这有助于快速评估不使用 FEC 时系统的脆弱性，以及 RL 在降低 BER 方面的作用。这两列在表格和模型中的共同作用是支持对比分析：PLR 是“黄金标准”——考虑所有错误的精确计算；≈ L×BER 则是“快速指示值”——便于缩放的线性估计。PLR 参数表示数据包无法送达的概率，计算时假定比特相互独立，并采用二项误码分布。这里，参数 t 是 FEC 等级，即代码在长度为 n = sz + r 的数据块中可纠正的最大错误数（ sz 为有效载荷，r 为冗余比特）。这些关系（图 6 ）展示了大数据包的严重脆弱性以及 FEC 的有效性。不使用 FEC（红线）时，即使在 100 kbps 下，PLR 也为 ≈ 0.06；使用 FEC t=1 （蓝线）时，PLR 已降至 < 10^{-8}；使用 FEC t=2 （绿线）时，PLR 为 < 10^{-14}——实际上为零。前向纠错的目的在于，即使采用最低限度的编码，也能使通信完全可靠。表 3 是仿真中的最终解析工具，用于评估强化学习智能体训练完成后，无人机无线跳频扩频系统中数据传输的实际可靠性。它根据最终指标计算（取最后 100 个回合的平均值，BER 为 ~10^{-6}，SNR 为 ~13 dB），并展示 FEC（前向纠错）如何降低不同数据包大小对应的 PLR。“开销”列说明 FEC 冗余比特相对于有效载荷（sz）的额外成本，以百分比表示。这是一种权衡：FEC 能减少丢包，但会增加流量。

图表将抽象的误码率（物理层指标）转化为实际的丢包率（丢包）。强化学习通过误码率优化奖励，但丢包率才是跳频扩频的“最终目标”。丢包率随着回合数增加而降低，这意味着策略成功（低干扰 → 低误码率 → 低丢包率）。干扰器会提高误码率，但智能体（深度Q网络 + 熵）通过均匀跳频降低丢包率。这些依赖关系是模型的“最终检验”：强化学习能降低误码率，但丢包率则检验数据包在实际流量中能否成功传输。

表1. 训练进度

Episode	BER	SNR (dB)	Entropy	Success Rate (%)	ε
200	5.18e-03	11.18	2.520	87.50	0.330
400	2.94e-04	12.75	2.571	91.00	0.121
600	2.87e-04	12.33	2.500	93.00	0.050
800	1.32e-05	12.41	2.528	94.75	0.050
1000	1.96e-05	12.37	2.600	95.80	0.050
1200	3.03e-04	12.40	2.629	96.42	0.050
1400	1.63e-05	12.29	2.654	96.86	0.050

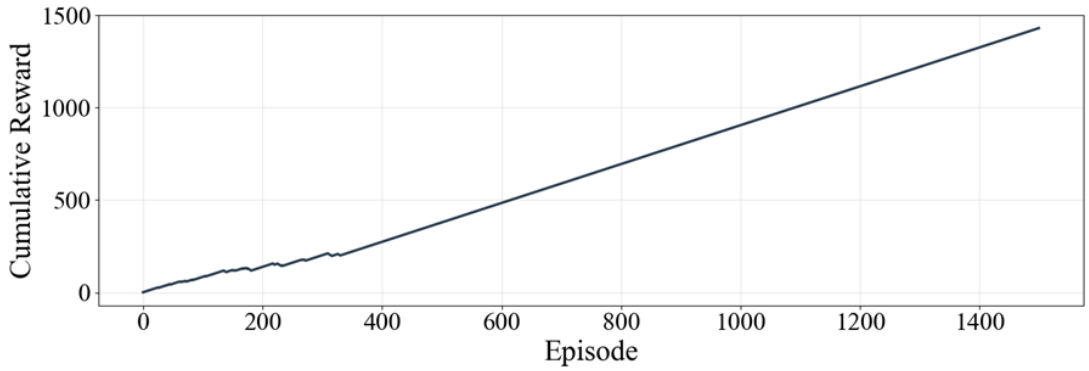

图2。训练过程中的累积奖励

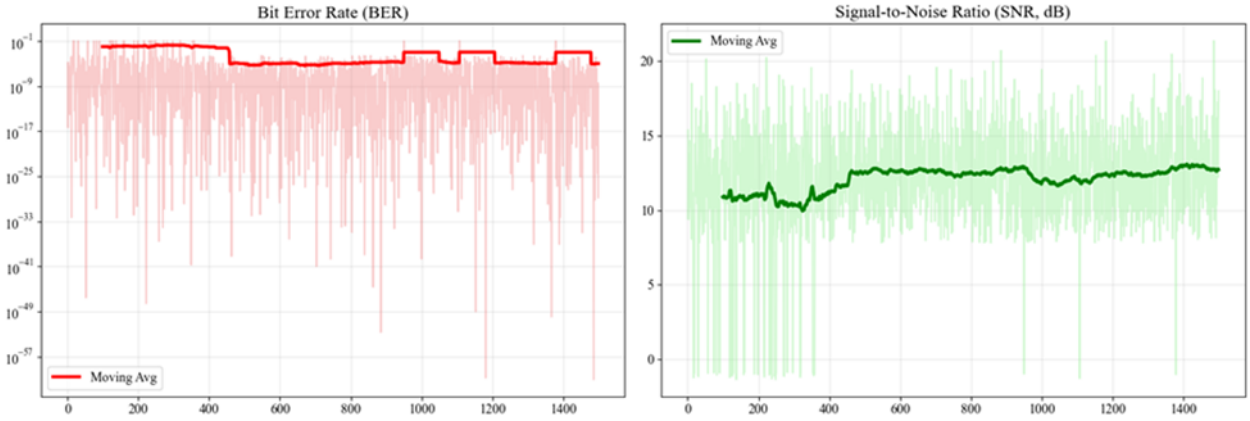

图3。误码率和信噪比的变化

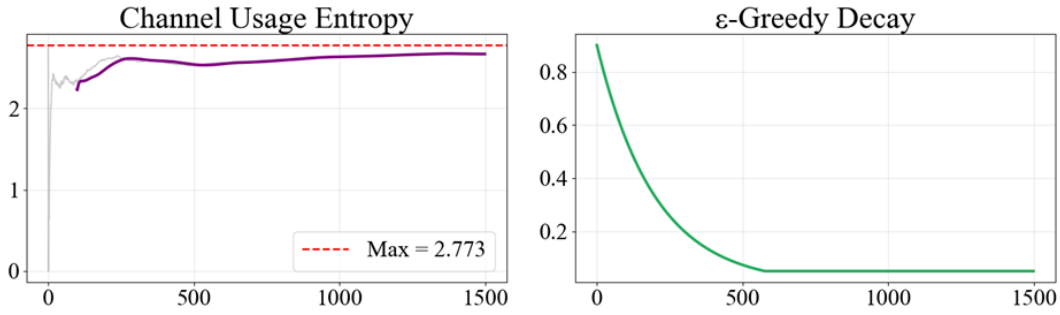

图4。探索与多样性

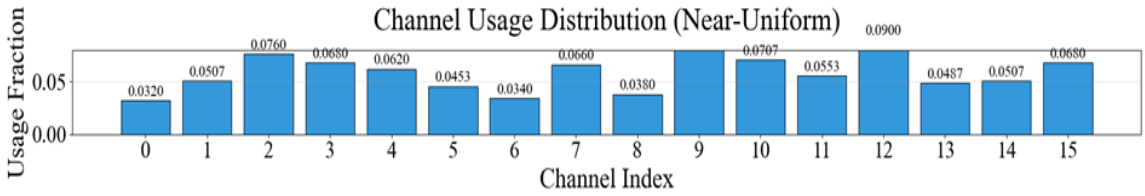

图 5。信道使用比例

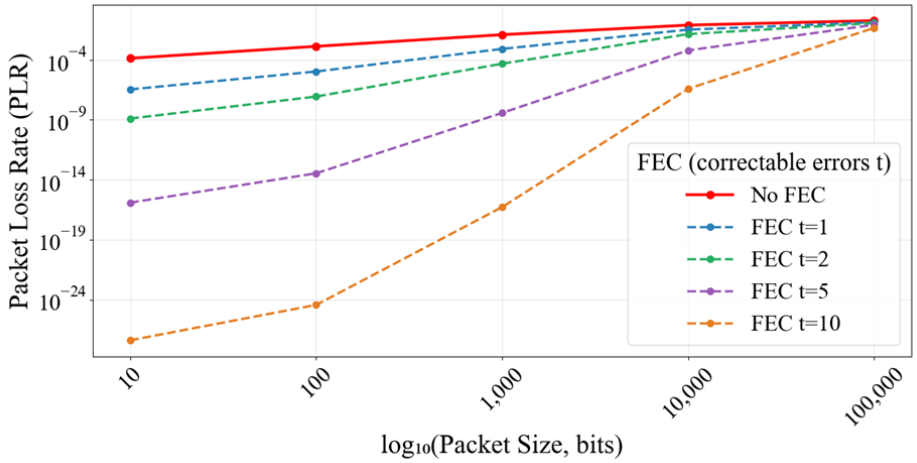

图 6。不同前向纠错等级下丢包率与数据包大小的关系

表格 2. 无前向纠错时的丢包率（最后 100 个回合）

Packet Size	PLR	≈ L×BER
10	1.37e-04	1.37e-04
100	1.36e-03	1.37e-03
1,000	1.28e-02	1.37e-02
10,000	8.05e-02	1.37e-01
100,000	1.91e-01	1.37e+00

表格 3. 采用类 BCH 前向纠错时的丢包率（最后 100 个回合）

Packet Size	PLR t=0	PLR t=1	PLR t=2	PLR t=5	PLR t=10	Overhead t=1 %	Overhead t=2 %	Overhead t=5 %	Overhead t=10 %
10	1.37e-04	3.48e-07	1.29e-09	1.23e-16	4.01e-28	100.00	200.00	500.00	1000.00
100	1.36e-03	1.09e-05	8.79e-08	3.47e-14	3.59e-25	10.00	20.00	50.00	100.00
1,000	1.28e-02	8.33e-04	4.86e-05	3.80e-09	5.30e-17	1.00	2.00	5.00	10.00
10,000	8.05e-02	3.38e-02	1.44e-02	6.32e-04	3.93e-07	0.10	0.20	0.50	1.00
100,000	1.91e-01	1.39e-01	1.20e-01	8.53e-02	4.69e-02	0.01	0.02	0.05	0.10

（注：BCH - Bose-Chaudhuri-Hocquenghem 编码）

表4展示了干扰功率的影响。将干扰功率减半（干扰信号比 1.2 → 0.5），传输成功率仅提高了 1–4.5%（从 96.29% 提高到 97.27%）。这证实，主要限制因素并非干扰功率的绝对值，而是干扰器预测下一个信道的能力。训练后的 DQN智能体在干扰信号比为 = 1.2 时（即干扰在物理强度上超过信号）便已实现近乎最高的可靠性，进一步削弱干扰只能带来微乎其微的提升。因此，在较宽的干扰信号比范围内，所提出的策略已接近理论最优策略。

表4. 不同信号与干扰强度的比较

参数	干扰信号比 = 1.2（干扰更强）	干扰信号比 = 0.5（信号强度是原来的 2 倍）	差值
成功率	96.29 %	97.27 %	+0.98%
最终误码率	1.84×10^{-4}	6.13×10^{-4}	差 3 倍
最终信噪比	≈ 22.8 дБ	≈ 12.5 дБ	差 10 倍
信道熵	99.71 %	99.82 %	好一点
回合奖励	≈ +0.84	≈ +0.97	+15 %
奖励	≈ 1300–1400	≈ 1450–1480	+11 %

6. 讨论

结果表明，即使在干扰功率大于期望信号功率的情况下，采用最小限度奖励塑形的最简单单智能体原始 DQN，也能够抵消一阶马尔可夫反应式干扰器的干扰。

关键发现是，限制可靠性的主要因素并非干扰功率或衰落深度的绝对值，而是跳频序列的可预测程度。将干扰信号比从 1.2 降至 0.5，成功率仅提高了 1%。需要强调的是，实现极其简单：

- 状态——当前信道的标准独热向量（不包含历史信息，也不包含信念状态）；
- 架构——一个三层全连接网络，结构为 16-128-128-16；
- 唯一的 非标准加法是 $\beta \cdot H(p)$ 奖励，可以很容易地解释为一种“可预测性惩罚 ."

这种极简设计使所提出的解决方案即使在现代无人机和巡飞弹最受限的机载处理器上也切实可行。即使只加入少量冗余编码（能够纠正 t =1–2 个错误的类似 BCH 的编码），也能将 100 Kbit 数据包的丢失概率降至 10^{-11}–10^{-14} 以下，同时开销不到 0.2%。这意味着，即使面对强大的敌方反应式电子战干扰，系统也已准备好在真实作战条件下传输遥测数据、全高清视频和控制命令。本研究的一个局限是仅考虑了一阶马尔可夫干扰器。

总体而言，结果表明，该模型不仅能够收敛，还能适应对抗性环境：智能体可以预测干扰器的马尔可夫转移，而熵则可防止过拟合。所得结果对在激烈电子战环境中行动的武装部队具有直接的实际意义。所提出的方法能够构建完全自主、可自我学习的跳频扩频系统，无需预先配置密钥、同步机制或密码学原语。

7. 结论

本文提出并研究了一种面向无人机的全自主跳频系统，该系统基于简单的单智能体深度 Q 网络（原始 DQN），仅对奖励函数进行了最少修改。

我们证明了完全消除一阶马尔可夫反应式干扰器影响的可行性。结果表明，训练后的智能体能够独立发现一种策略，确保较高的信道分配熵，使干扰器几乎无法进行统计预测。在严重干扰条件下，系统仍能达到较高的可靠性指标。

采用最少的纠错编码后，丢包概率较低，从而确保在敌方强烈电子战干扰下，遥测数据、视频和控制命令仍能可靠传输。

Th这些结果鼓励我们开发新一代全自主、自适应且虚拟 lly 坚不可摧的通 信系统，适用于现代电子战中的无人机。

声明与披露
经费：作者声明，在撰写本文期间未获得任何资金、拨款或其他支持。

利益冲突：作者没有需要披露的相关财务或非财务利益。
利益冲突：作者声明不存在利益冲突。
作者贡献：Volodymyr Kharchenko – V.Kh.、Andrii Grekhov – A.G.、Vasyl Kondratiuk – V.K. 概念化由 A.G. 和 V.Kh. 负责；方法论由 A.G. 负责；验证由 A.G.、V.Kh. 和 V.K. 负责；调查由 A.G. 负责；资源由 V.Kh. 和 V.K. 提供；初稿撰写由 A.G. 负责；审阅与编辑由 V.K. 负责；监督由 V.Kh. 负责；项目管理由 V.K. 负责。所有作者均已阅读并同意稿件的发表版本。

伦理审批：不适用。
数据可用性声明：本研究期间生成和分析的所有数据均已包含在本文中。当前研究期间生成的数据集可根据要求向通讯作者索取。

参考文献

1. Krayani A、Alam AS、Marcenaro L、Nallanathan A、Regazzoni C（2022）认知无人机抗干扰的一种新型资源分配方法：主动推理方法。arXiv:2208.05269。https://doi.org/10 05269 2 48550 2208 . /arXiv. . . Liu C、Zhang Y、Niu G、Jia L、Xiao L、Luan J（2023）面向无人机中继抗干扰海事通信的强化学习。《数字通信与网络》9:1477-1485。https://doi.org/10.1016/j.dcan.2022.08.009 3. Nguyen HN、Noubir N（2023）JaX：使用卷积神经网络检测并消除高功率干扰器。载于第16届ACM无线与移动网络安全和隐私会议（WiSec’23）论文集，2023年5月29日至6月1日，英国吉尔福德。ACM，纽约，NY，美国，12页。https://doi.org/0 3590178 4 1145 3558482 . / . . Hu L、Shao Y、Qian Y、Du F、Li J、Lin Y、Wang Z（2024）时变无人机通信中的元强化学习：自适应抗干扰信道选择。《无线电工程》33 431 5 417 2024 : - . Yang S（ ）基于深度学习的无人机通信抗干扰方法分析。《科学、工程与技术亮点》，CDMMS 2024 103:246-253 6. Aygur M、Kandeepan S、Giorgetti A、Al-Hourani A、Arbon E、Bowyer M（2025）窄带干扰抑制技术：综述。《IEEE通信综述与教程》https://doi.org/10 3531428 7 1109 2025 2024 . /COMST. . . Ghelani J、Gharia P、El-Ocla H（ ）飞行自组网中用于干扰攻击检测的梯度监控强化学习。《IEEE开放获取》12:23081-23095。https://doi.org/10.1109/ACCESS.2024.3361945 8. Yin Z、Li J、Wang Z、Qian Y、Lin Y、Shu F（2024）针对智能干扰的无人机通信：结合联邦强化学习的斯塔克尔伯格博弈方法。《IEEE绿色通信与网络汇刊》8 3373886 9 1796 1808 10 1109 2024 2024 : – . https://doi.org/ . /TGCN. . . Hussein J、Wissam A、Samer J（ ）基于强化学习的认知无线电网络频谱与功率高效抗干扰方法。《传感器、无线通信与控制国际期刊》。https://doi.org/14 0122103279291431240216061325 10 10 2174 2025 . . / . Yang J、Cui M、Zhang H、Ji F、Lai Z、Wang Y（ ）对抗环境中无人机通信的基于智能体的抗干扰技术：全面综述。arXiv:2508 1 11 11687 2024 . v . Zhou Q、Niu Y（ ）从自适应通信抗干扰到智能通信抗干扰：演进50年 20230085 12 10 1002 2024 Years of Evolution。https://doi.org/ . /aisy. . Ding H、Niu Y、Zhou Q、Peng X（ ）基于近端策略优化的新型智能抗干扰通信算法。《物理通信》。https://doi.org/10.1016/j.phycom.2024.102366 13. Zhang F、Niu Y、Zhou Q 等（2025）有限信道状态信息条件下的无线通信智能抗干扰决策算法。《科学报告》。https://doi.org/10.1038/s41598-025-90201-1 14. Cao W、Chu F、Jia L、Zhou H、Zhang Y（2025）低地球轨道卫星中的多智能体深度强化学习抗干扰频谱接入方法。《电子学》。https://doi.org/10 14163307 15 3390 2025 . /electronics . Kharchenko V、Grekhov A、Kondratiuk V（ ）基于Q学习的自适应人工智能干扰防护无人机通信信道。第15届国际可靠系统、服务与技术会议（DESSERT’2025）。希腊，雅典，12月19日至21日，报告64