arXiv:2602.01582v1 [cs.IT] 2 Feb 2026

小幅信道扰动下基于人工智能的信道译码器的脆弱性

Haoyu Lei*†、Mohammad Jalali*‡、Chin Wa Lau§、Farzan Farnia¶

摘要

近年来，深度学习取得的进展催生了基于人工智能的纠错译码器；据报告，这些译码器在加性高斯白噪声信道上的经验性能优于传统的置信传播（BP）译码。尽管这些增益令人鼓舞，一个根本问题仍然存在：这些改进从何而来，又为此付出了什么代价？本文从信道输出分布偏移下的鲁棒性角度研究这一问题。我们评估了输入相关的对抗扰动（FGM 和在 ℓ_2 约束下使用的投影梯度法），以及对所有接收向量施加同一范数受限偏移的通用对抗扰动。结果表明，近期的人工智能解码器（包括 ECCT 和 CrossMPT）尽管在独立同分布加性高斯白噪声下具有更优的标称性能，却可能在此类扰动下遭受显著的性能下降。此外，对抗扰动在人工智能解码器之间的迁移性相对较强，但迁移到基于BP的解码器时较弱；与范数相同的随机扰动相比，通用扰动的危害要大得多。这些数值结果表明，近期人工智能译码性能增益可能伴随着鲁棒性代价，并且对信道分布具有更高的敏感性。

1 引言

可靠的译码是数字通信系统的基石。香农信道编码定理 [1]指出，原则上可以实现不超过信道容量的传输速率，但在实践中逼近这些极限，关键取决于高效且准确的译码算法设计。过去几十年来，人们通过开发强大的码族和迭代译码方法应对这一挑战，其中包括 Turbo 译码 [2] 以及用于低密度奇偶校验码的置信传播（BP）译码 [3–5]。在现代无线系统中，解码器必须在复杂度、时延、能耗、数值精度以及对信道失配的鲁棒性等方面满足严格要求，因此，解码器设计成为将信息论极限转化为实际性能的核心瓶颈 [6, 7]。

受深度学习在其他领域取得成功的启发，近年来，将神经网络方法应用于信道解码日益受到关注。早期基于学习的方法主要致力于改进传统解码器，例如对置信传播更新进行参数化，或展开迭代译码算法，如神经置信传播 [8]。其他研究则探索了无模型神经译码器，使其能够学习从信道输出到码字的直接映射，尤其适用于短码长场景 [9]。尽管这些方法在特定场景中展现出令人鼓舞的性能提升，但它们在很大程度上保留了经典译码算法的结构，因此并未从根本上脱离传统的消息传递范式。

最近，人们提出了基于Transformer的译码器架构，这些架构更大幅度地偏离了经典译码规则。值得注意的是，纠错码Transformer（ECCT） [10] 引入了一种具有码结构感知注意力机制的基于Transformer的解码器，并报告了相当可观的

*共同贡献
†计算机科学与工程系，香港中文大学，hylei22@cse.cuhk.edu.hk
‡计算机科学与工程系，香港中文大学§华为技术有限公司，理论实验室

¶计算机科学与工程系，香港中文大学，farnia@cse.cuhk.edu.hk

相对于先前神经解码器的改进。CrossMPT [11] 进一步整合了幅度与综合征表示之间的掩码交叉注意力，并报告了在多个码族上的经验性能增益，包括在所考虑设置下相对于传统BP解码的明确改进。这些结果表明，现代人工智能解码器可能获得超越传统消息传递算法的性能增益，从而提出一个根本问题：这些增益的来源是什么？

在本研究中，我们从信道输出分布偏移下鲁棒性的角度探讨这一问题。具体而言，我们关注近期基于人工智能的解码器所报告的标称性能提升，是否会以对施加于接收信号的小幅范数有界扰动更加敏感为代价。这类扰动可视为相对于目标加性高斯白噪声模型的最坏情况偏差，并对应于信道输出分布的小幅偏移；这类偏移通常用基于 Wasserstein 距离的鲁棒性概念来建模 [12]。要评估基于人工智能的解码器在标准信道假设之外的可靠性，理解解码器在这些偏移下的行为至关重要。

为了解决这一问题，我们采用对抗鲁棒性领域的工具 [13, 14] ，并评估施加于接收向量的输入相关扰动和通用扰动。我们考虑快速梯度方法，以及约束在 ℓ_2 范数球内的多步投影梯度攻击；同时也考虑将相同扰动应用于所有信道输出的通用对抗扰动 [15]。这些攻击在加性高斯白噪声信道输出处实施，并采用基于梯度的方法进行优化，同时使用随机平滑 [16] ，以便在解码器映射不光滑时稳定优化过程。

我们的数值结果表明，标准的、输入相关的对抗扰动会显著降低近期基于Transformer的人工智能解码器（包括 ECCT [10] 和 CrossMPT [11]）的帧错误率。我们进一步研究了通用扰动，这类扰动在实践中尤为重要，因为它们不依赖于单个接收向量，且可能对应于持续性干扰或硬件引起的失真。我们发现，与具有相同 ℓ_2 范数的随机扰动相比，通用对抗扰动 [15] 对人工智能解码器的危害显著更大。这一观察表明，性能下降不能仅仅归因于噪声功率增加，而是源于学习得到的解码规则中存在结构化脆弱性。

最后，我们研究了对抗扰动在不同解码器间的可迁移性，这种现象在对抗学习中广为人知 [13]。针对某种基于Transformer的人工智能解码器优化的扰动，往往能有效迁移至另一种解码器，但迁移至基于BP的解码器时，可迁移性则弱得多。反过来，经典解码器对于针对人工智能解码器构造的扰动仍相对鲁棒。总之，这些结果表明，近期人工智能解码性能的提升可能依赖于脆弱的特定于分布的特征；同时，这些结果量化了一种鲁棒性代价，而这种代价在传统消息传递解码中较不明显。

2 相关工作

人工智能与机器学习解码器。基于学习的解码已从展开式置信传播 [17] 发展到端到端架构。相关示例包括 Deepcode 中的反馈编码 [18], 、Turbo自动编码器 [19],以及通过 MIND [20]实现的元学习适配。近期，Transformer模型 ECCT [10], 、CrossMPT [11]和 DiffMPT [21] 利用注意力机制改进了传统基线方法。其他模型还包括基于扩散的迭代细化 [22], 、多码基础模型 [23], 以及一步一致性模型 [24]。

人工智能解码器的鲁棒性。研究表明，学习型接收机对扰动高度敏感。与经典方案相比，物理对抗攻击已被证明会显著降低端到端自编码器的性能 [25]。在无线安全领域，研究人员已针对基于深度神经网络的系统开发了输入无关扰动 [26]。此外，深度接收机易受受约束攻击影响，例如受功率和峰均功率比限制的攻击 [27]。

信息论中的对抗鲁棒性与分布鲁棒性。经典研究探讨最坏情况模型下的可靠通信，其中核心问题是任意变化信道（AVC） [28]。相关的对抗场景包括无知信道 [29] 和因果对手 [30]。近年来，研究者已刻画了具有计算受限观测的对手所对应的容量 [31]。与此同时，通过 Wasserstein 歧义集 [32] 实现的分布鲁棒性，为连接范数有界扰动提供了一个框架

对分布偏移。

3 预备知识

我们考虑一个码本为 $\mathcal{C} \subset \{0,1\}^n$的二元线性纠错码，该码由生成矩阵$G \in \mathbb{F}_2^{k\times n}$和校验矩阵 $H \in \mathbb{F}_2^{(n-k)\times n}$ 定义，满足 $GH^\top = 0$ ，其中 $\mathbb{F}_2$。消息 $m \in \{0,1\}^k$被编码为码字 $x = mG \in \mathcal{C}$ ，并采用二进制相移键控（BPSK）进行调制，其中$0 \mapsto +1$ 且 $1 \mapsto -1$，从而得到发送码字 $x_s \in \{-1,+1\}^n$。该码字通过加性高斯白噪声（AWGN）信道传输，接收信号为

$$y = x_s + z, \tag{1}$$

其中，噪声向量 $z \sim \mathcal{N}(0, \sigma^2 I_n)$ 服从高斯分布。解码器的目标是根据观测到的 y 估计发送的码字 $\hat{x}_s$ 。

译码中一个重要的量是伴随式，它由接收信号的硬判决解调版本计算得到。具体而言，我们定义 $y_b = \mathrm{bin}(\mathrm{sign}(y))$，其中 $\mathrm{sign}(y_i) = +1$ 当 $y_i \geq 0$ 时， -1 否则；$\mathrm{bin}(\cdot)$ 将 $\{-1,+1\}$ 映射为 $\{1,0\}$。相应的伴随式为 $s(y) = Hy_b^\top \in \mathbb{F}_2^{n-k}$，当 $s(y) \neq 0$ 时，即检测到错误。

遵循 [33], 的预处理策略，通过将接收信号的幅度与伴随式拼接，构造神经解码器的输入，

$$\phi(y) = \big[\, |y|,\ s(y) \,\big], \tag{2}$$

从而得到维度为 $n + (n-k)$ 的输入向量。这种表示向神经解码器提供了可靠性信息和校验约束违反信息，并已证明能够在保持译码性能的同时减少过拟合。

4 AI解码器对输入扰动的鲁棒性

我们研究信道解码器对施加于信道输出的小扰动的最坏情况敏感性。令 f 表示一个将信道观测 $y \in \mathbb{R}^n$ 映射为估计值 $\hat{x}_s \in \{-1,1\}^n$ 的解码器（传统解码器或基于人工智能的解码器）。令 $x^\star \in \{-1,1\}^n$ 表示发送的码字，我们通过优化以下目标来训练 AI 解码器:

$$\min_{f\in\mathcal{F}} \ \mathbb{E}\left[\ell\big(f(y), x^\star\big)\right], \tag{3}$$

其中， $\ell(\cdot,\cdot)$ 表示损失函数，具体为 [11, 34] 中定义的二元交叉熵（BCE）。

逐样本对抗目标。给定扰动预算 $\varepsilon > 0$，我们将逐样本 ℓ_2 有界的对抗扰动定义为

$$\max_{\delta\in\mathbb{R}^n} \ \ell\big(f(y+\delta), x^\star\big) \quad \text{s.t.} \quad \|\delta\|_2 \leq \varepsilon. \tag{4}$$

当 f 包含硬操作（例如符号函数、二值化、伴随式计算）时，直接优化式 (4) 可能具有挑战性，因为这些操作可能导致梯度不平滑或不稳定。因此，我们按照下文所述，对平滑替代目标发起攻击。

解码器优化目标的随机平滑。固定 $\nu > 0$ 并令 $V \sim \mathcal{N}\big(0,\ \nu^2 I_n\big)$。我们定义高斯平滑损失

$$g(y,\delta) \triangleq \mathbb{E}_{V\sim\mathcal{N}(0,\nu^2 I_n)}\Big[\ell\big(f(y+V+\delta), x^\star\big)\Big], \tag{5}$$

并考虑平滑对抗目标

$$\max_{\|\delta\|_2 \le \varepsilon} \; g(y, \delta). \tag{6}$$

随机平　　滑可产生稳定的梯度信号，并支持有理有据的基于梯度的攻击。

为了理解随机平滑为何有效，我们在以下命题中指出，高斯平滑能以与维度无关的方式使损失景观表现良好。这是第 4.2 节通用扰动分析中的一项关键技术工具；在该节中，我们将 (i) 论证对通用目标进行基于梯度的优化是合理的，并 (ii) 根据平滑性替代指标推导出一个闭式通用方向（UAP-PCA）。

命题 4.1（通过高斯平滑实现与维度无关的光滑性）。 假设损失有界：对所有$u \in \mathbb{R}^n$,

$$0 \le \ell\big(f(u), x^\star\big) \le C. \tag{7}$$

定义 $\tilde{g}(u) \triangleq \mathbb{E}_{V \sim \mathcal{N}(0, \nu^2 I_n)} \big[\ell\big(f(u+V)\, x^\star\big)\big]$。那么 $\tilde{g}$具有常数为以下数值的利普希茨连续梯度

$$\beta \;\le\; \frac{C}{\nu^2}\, \mathbb{E}\big[\,|Z^2 - 1|\,\big] \;<\; \frac{C}{\nu^2}, \qquad Z \sim \mathcal{N}(0,1), \tag{8}$$

即，$\big\|\nabla \tilde{g}(u) - \nabla \tilde{g}(u')\big\|_2 \le \frac{C}{\nu^2} \|u - u'\|_2$ 对所有 u、u'均成立。

证明。 证明见附录 A。 □

4.1 逐样本对抗攻击

我们考虑针对平滑目标函数 (6) 的两种标准一阶攻击。这两种攻击都使用蒙特卡罗样本 $V_1, \ldots, V_M \sim \mathcal{N}\big(0, \nu^2 I_n\big)$ 来估计$\nabla_\delta g(y, \delta)$:

$$\nabla_\delta g(y, \delta) \approx \frac{1}{M} \sum_{m=1}^{M} \nabla_\delta\, \ell\Big(f\big(y + \delta + V_m\big), x^\star\Big). \tag{9}$$

PGD攻击　　Pro投影梯度下降 (PGD) [35] 在 (6 上执行迭代投影梯度上升):

$$\delta^{(t+1)} = \Pi_\varepsilon\Big(\delta^{(t)} + \eta\, \nabla_\delta g\big(y, \delta^{(t)}\big)\Big), \qquad t = 0, \ldots, T-1, \tag{10}$$

其中，Π_ε 表示到 $\{\delta : \|\delta\|_2 \le \varepsilon\}$上的欧几里得投影。PGD 通过多个步骤不断优化 δ，通常比单个梯度步产生更强的攻击。

FGM攻击 快速梯度法（FGM） [13] 是一种单步攻击，它在 $\delta = 0$处对 $g(y, \delta)$ 进行线性化，并在 ℓ_2 约束下选择线性近似的最大化点:

$$\delta_{\mathrm{FGM}} = \varepsilon\, \frac{\nabla_\delta g(y, 0)}{\big\|\nabla_\delta g(y, 0)\big\|_2}. \tag{11}$$

FGM 的计算效率很高，可作为衡量一阶脆弱性的基线。

4.2 通用对抗扰动

通用对抗扰动（UAP） [36] 旨在寻找一种扰动，使许多信道输出的性能均下降。令 $\mathcal{D}$ 表示由随机消息和信道噪声所诱导的 $(Y, X^\star)$ 分布。定义总体通用目标函数和经验通用目标函数

$$F(\delta) \triangleq \mathbb{E}_{(Y, X^\star) \sim \mathcal{D}}\big[g(Y, \delta)\big], \qquad \widehat{F}_N(\delta) \triangleq \frac{1}{N} \sum_{i=1}^{N} g\big(Y_i, \delta\big), \tag{12}$$

其中 $(Y_i,\ X_i^\star)_{i=1}^N$ 是来自 $\mathcal{D}$ 的独立同分布样本。

UAP-Grad 我们的第一个通用攻击 *UAP-Grad* 直接对经验 ^目标函数 $F_N(\delta)$ 应用投影梯度上升：

$$\delta^{(t+1)} = \Pi_\varepsilon\Big(\delta^{(t)} + \eta \cdot \frac{1}{N}\sum_{i=1}^{N} \nabla_\delta g\big(Y_i, \delta^{(t)}\big)\Big). \tag{13}$$

UAP-Grad 灵活且功能强大，但由于需要对大量样本反复进行梯度聚合，计算开销可能较大。

UAP-PCA 我们的第二种方法 *UAP-PCA* 受光滑性启发，给出一个闭式通用方向。对于每个样本 i，定义平滑后的单样本函数

$$\tilde{g}_i(u) \triangleq \mathbb{E}_{V\sim\mathcal{N}(0,\nu^2 I_n)}\Big[\ell\big(f(u+V), X_i^\star\big)\Big],$$

及其在样本输入处的梯度

$$q_i \triangleq \nabla\tilde{g}_i(Y_i) \in \mathbb{R}^n. \tag{14}$$

UAP-PCA 根据经验二阶矩矩阵计算通用方向

$$\widehat{\Sigma}_q \triangleq \frac{1}{N}\sum_{i=1}^{N} q_i q_i^\top. \tag{15}$$

请注意，命题 4.1 表明，在有界条件下，高斯平滑损失具有 β-光滑性。对于 UAP-PCA，我们进一步利用光滑性，在一个固定方向上构造二次下界，然后对该替代目标进行精确优化。由此得到一种精确 *PCA*刻画。

命题 4.2。 假设每个 $\tilde{g}_i$ 都是 β-光滑的。对于满足 $\|\delta\|_2 = 1$ 的单位方向 $\delta \in \mathbb{R}^n$，光滑性意味着对所有 $\alpha \in \mathbb{R}$,

$$\tilde{g}_i\big(Y_i + \alpha\delta\big) \;\geq\; \tilde{g}_i(Y_i) + \alpha\,\delta^\top q_i - \frac{\beta}{2}\alpha^2. \tag{16}$$

考虑替代目标最大化问题

$$\max_{\delta\in\mathbb{R}^n:\,\|\delta\|_2=1} \;\frac{1}{N}\sum_{i=1}^{N}\; \max_{\alpha\in\mathbb{R}}\Big(\alpha\,\delta^\top q_i - \frac{\beta}{2}\alpha^2\Big). \tag{17}$$

^那么，(17)的任一最大化点 $\delta^\star$ 都是(15)中定义的 Σ_q 的最大特征值对应的特征向量。

证明。 证明推迟至附录 A。 □

UAP 评估与 UAP-PCA 的有限样本保证

接下来，我们陈述一个集中定理，该定理（i）量化了对于固定的 δ，估计^通用目标函数 F_N（δ）所需的样本数量；（ii）保证 Σ_q的经验第一主成分^u_1 具有稳定性。假设。 我们假设：(a) 损失 $0 \leq \ell\,(\cdot) \leq C$有界；(b) 梯度 $\|q_i\|_2 \leq L$ 几乎处处有界；
且 (c) 特征值间隙 $\Delta \triangleq \lambda_1\,(\Sigma_q)\; - \lambda_2\,(\Sigma_q)\; > 0$，其中 $\Sigma_q \triangleq \mathbb{E}\big[qq^\top\big]$ 和 q 表示 q_i的一个通用副本。

定理 4.3（UAP 目标函数与 UAP-PCA 的集中性）。任取 $\eta \in$ （0，1）。给定假设 *(a)*、*(b)*、*(c)*，则以至少 $1-\eta$ 的概率，以下结论成立：

$$\big|\widehat{F}_N(\delta) - F(\delta)\big| \leq C\sqrt{\frac{\log(4/\eta)}{2N}} \qquad \textit{for any fixed } \delta, \tag{18}$$

$$\left\|\widehat{\Sigma}_q - \Sigma_q\right\|_{\mathrm{op}} \le 4\sqrt{2}\,L^2\sqrt{\frac{\log\left(4n/\eta\right)}{N}}, \tag{19}$$

$$\sin\angle\left(\hat{u}_1, u_1\right) \le \frac{4\sqrt{2}\,L^2}{\Delta}\sqrt{\frac{\log\left(4n/\eta\right)}{N}}, \tag{20}$$

其中， u_1 是 Σ_q的最大特征值对应的特征向量，$\hat{u}_1$ 是 Σ_q 的最大特征值对应的特征向量（选择符号使 $|\langle\hat{u}_1, u_1\rangle|$ 最大），而 $\angle(\cdot,\ \cdot)$ 表示一维子空间之间的主角。

证明。 证明详见附录 A。 □

5 数值结果

5.1 实验设置

数据集。 我们在标准纠错码上评估对抗攻击，重点关注极化码和低密度奇偶校验码。评估涵盖多种码配置，包括不同的码长（n）和码率（$R = k/n$）。此外，我们还测试了多个信噪比（SNR），具体范围为 E_b/N_0 4 至 6 分贝。对于每种码类型，我们在 10^6 个测试样本上评估结果。

评估指标。 遵循现有研究 [34,37], ，我们采用标准指标帧错误率（FER）评估译码性能，该指标用于量化至少包含一个比特错误的码字所占比例。

基线方法。 我们针对一系列全面的译码算法评估对抗攻击，这些算法分为以下几类：1）传统译码器：最小和（MS） [5,38] 和和积（SP） [3,39] ，最大迭代次数均为 10。2）神经解码器：先进的基于深度学习的方法，具体包括 ECCT [34] 和 CrossMPT [37]。对于神经解码器（ECCT、CrossMPT），我们遵循其论文中指定的设置，以确保比较公平。两个模型均采用 $N = 6$ 层深度和 $d = 128$的隐藏维度实现。

对抗攻击方法。 我们评估解码器在样本级对抗攻击和通用对抗攻击下的表现。对于针对单个码字的样本级攻击，我们采用快速梯度法（FGM） [40] 和投影梯度下降（PGD） [35]。对于旨在针对整个数据集使用单一扰动的通用攻击，我们选用通用对抗扰动（UAP） [36] 和基于PCA的通用对抗方向（UAP-PCA） [41]。对于所有攻击方法，我们使用相对 L_2-范数界约束扰动能量： $||\delta||_2 \le \alpha||y||_2$。此外，我们还报告了具有相同能量约束的随机噪声下的结果，以展示对抗攻击的影响。

5.2 对抗攻击后的性能下降

在我们的实验中，我们评估了两种基线人工智能解码器 CrossMPT 和 ECCT 面对多种攻击方法时的鲁棒性，包括 FGM、投影梯度下降、通用对抗扰动和 UAP-PCA。我们比较了攻击前后的帧错误率倒数（**1/FER**）。所有攻击方法均采用相同的能量约束$\alpha = 0.001$ 。我们在不同的极化码和低密度奇偶校验码上进行实验，信噪比 E_b/N_0 范围为 4 至 6 分贝。如表1所示，尽管经过充分训练的人工智能解码器最初表现出卓越的译码性能，但仅相当于信号能量$1_{\ 1000}$ 的对抗扰动就会导致性能显著下降。虽然解码器对随机噪声仍具有相对较强的鲁棒性，但在对抗攻击下，尤其是在高信噪比时，其性能会急剧下降。对于在 6 分贝下对 LDPC(49,24) 进行译码的情况，PGD攻击使（**1/FER**）从 5.0×10^4 降至 2.6×10^1 ，表明性能下降超过 1000 倍。这揭示了人工智能解码器在对抗攻击下固有的鲁棒性代价。

表1：解码　　性能比较：不同攻击方法下的**FER-逆 (1/FER)**　　s.

Type	方法	SNR	ECCT				CrossMPT			
			极化码		LDPC		极化码		LDPC	
			(64,48)↑	(128,64)↑	(49,24)↑	(121,60)↑	(64,48)↑	(128,64)↑	(49,24)↑	(121,60)↑
w/o attack	干净样本	4	2.1×10^{1}	8.5×10^{0}	3.7×10^{1}	8.6×10^{0}	3.4×10^{1}	2.5×10^{1}	8.9×10^{1}	2.1×10^{1}
		5	1.6×10^{2}	7.4×10^{1}	3.5×10^{2}	1.4×10^{2}	2.6×10^{2}	2.4×10^{2}	1.5×10^{3}	6.1×10^{2}
		6	2.2×10^{3}	1.4×10^{3}	9.1×10^{3}	1.1×10^{4}	4.8×10^{3}	1.0×10^{4}	5.0×10^{4}	1.7×10^{4}
	随机	4	2.0×10^{1}	8.1×10^{0}	3.5×10^{1}	8.2×10^{0}	3.3×10^{1}	3.0×10^{1}	7.7×10^{1}	1.9×10^{1}
		5	1.3×10^{2}	5.9×10^{1}	2.4×10^{2}	9.3×10^{1}	2.2×10^{2}	1.7×10^{2}	1.2×10^{3}	2.7×10^{2}
		6	1.7×10^{3}	6.1×10^{2}	2.0×10^{3}	3.6×10^{3}	2.8×10^{3}	4.5×10^{3}	9.8×10^{3}	3.8×10^{3}
Universal	**UAP**	4	1.7×10^{1}	5.4×10^{0}	2.8×10^{1}	5.1×10^{0}	2.8×10^{1}	1.9×10^{1}	4.0×10^{1}	1.6×10^{1}
		5	5.3×10^{1}	3.4×10^{1}	1.1×10^{2}	3.2×10^{1}	1.0×10^{2}	7.3×10^{1}	2.0×10^{2}	5.5×10^{1}
		6	3.6×10^{2}	9.8×10^{1}	2.7×10^{2}	2.0×10^{2}	5.7×10^{2}	5.7×10^{2}	6.4×10^{2}	5.1×10^{2}
	UAP-PCA	4	1.5×10^{1}	5.0×10^{0}	2.4×10^{1}	4.7×10^{0}	2.4×10^{1}	1.3×10^{1}	3.5×10^{1}	1.4×10^{1}
		5	4.5×10^{1}	2.0×10^{1}	9.8×10^{1}	2.1×10^{1}	6.9×10^{1}	5.6×10^{1}	1.6×10^{2}	4.5×10^{1}
		6	1.5×10^{2}	5.8×10^{1}	1.6×10^{1}	1.1×10^{2}	2.4×10^{2}	1.9×10^{2}	3.4×10^{2}	1.9×10^{2}
Sample-level	**FGM**	4	1.4×10^{1}	4.9×10^{0}	2.0×10^{1}	4.1×10^{0}	2.2×10^{1}	8.4×10^{0}	2.9×10^{1}	1.3×10^{1}
		5	3.2×10^{1}	1.4×10^{1}	5.5×10^{1}	1.5×10^{1}	3.8×10^{1}	4.7×10^{1}	1.2×10^{2}	3.1×10^{1}
		6	7.1×10^{1}	2.8×10^{1}	9.3×10^{1}	6.3×10^{1}	7.5×10^{1}	1.2×10^{2}	1.7×10^{2}	1.1×10^{2}
	PGD	4	7.8×10^{0}	3.0×10^{0}	6.0×10^{0}	3.3×10^{0}	9.0×10^{0}	3.5×10^{0}	1.0×10^{1}	6.8×10^{0}
		5	1.8×10^{1}	5.4×10^{0}	9.3×10^{0}	8.3×10^{0}	1.7×10^{1}	6.9×10^{0}	1.5×10^{1}	1.8×10^{1}
		6	3.3×10^{1}	8.3×10^{0}	1.5×10^{1}	1.7×10^{1}	3.2×10^{1}	1.9×10^{1}	2.6×10^{1}	3.0×10^{1}

5.3 消融研究：能量约束的变化

这项消融研究通过改变能量约束 α （$||\delta||_2 \leq \alpha||y||_2$），使其在 10^{-4} 至 10^{-1}之间变化，考察扰动强度对性能的影响。我们在不同信噪比下评估 ECCT 在 Polar(128,64) 上的表现。图1显示，能量约束越大，性能退化越严重，这种影响在高信噪比下尤为明显。低密度奇偶校验码的详细结果见附录 D.1。

5.4 对抗扰动的可迁移性

我们进一步评估对抗扰动的可迁移性，并考察传统解码器的鲁棒性。我们基于用于 LDPC(121,60)码的 ECCT 模型，在能量约束为 $\alpha = 0.001$ 的条件下生成最优 UAP 和 PGD 扰动。我们在四种条件下测试 AI 解码器（ECCT、CrossMPT）和传统解码器（最小和、和积）的译码性能，并以 **1/帧错误率** 衡量：干净样本、随机噪声，以及迁移后的 UAP 和 PGD 扰动。如表2所示，AI 解码器在干净数据上表现出色，但在受到攻击时则显著脆弱。值得注意的是，针对 ECCT 构造的扰动能够成功迁移至 CrossMPT，导致其性能大幅下降（例如，在 6 分贝、PGD 攻击下，性能从 1.7×10^{4} 降至 5.5×10^{1} ）。相比之下，传统解码器表现出较强的鲁棒性，即使面对对抗性分布偏移，也能维持稳定的性能水平。关于其他结果，包括扰动迁移至基于扩散的 DDECC [22] ，以及在 LDPC 和 Polar 码上的结果，请参见附录 D.2。

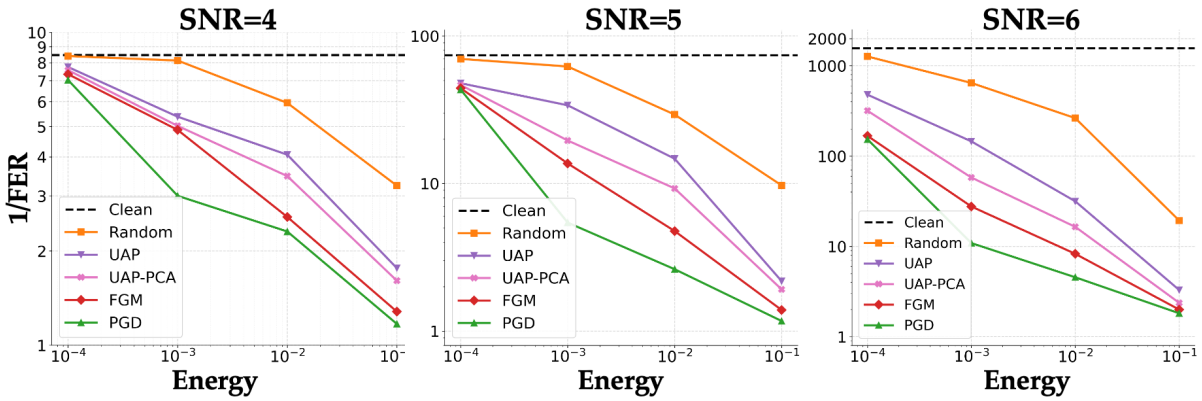


图1：ECCT 在 Polar(128,64) 上不同能量约束下的**FER-逆（1/FER）**比较。

表2：评估　　在 LDPC(121,60) 上，各解码器之间通过**FER-逆（1/FER）**测量的攻击可迁移性　　.

Type	解码器	攻击	4 dB	5 dB	6 dB
AI Decoder	ECCT	干净样本	8.6×10^{0}	1.4×10^{2}	1.1×10^{4}
		随机	8.2×10^{0}	9.3×10^{1}	3.6×10^{3}
		UAP	5.1×10^{0}	3.2×10^{1}	8.0×10^{2}
		PGD	3.3×10^{0}	8.3×10^{0}	1.7×10^{1}
	CrossMPT	干净样本	2.1×10^{1}	6.1×10^{2}	1.7×10^{4}
		随机	1.9×10^{1}	2.7×10^{2}	3.8×10^{3}
		UAP	1.7×10^{1}	6.7×10^{1}	8.8×10^{2}
		PGD	9.2×10^{0}	3.2×10^{1}	5.5×10^{1}
Traditional Decoder	最小和	干净样本	5.4×10^{0}	5.1×10^{1}	2.0×10^{3}
		随机	5.4×10^{0}	5.1×10^{1}	1.9×10^{3}
		UAP	5.3×10^{0}	5.0×10^{1}	1.7×10^{3}
		PGD	5.0×10^{0}	4.3×10^{1}	1.3×10^{3}
	和积	干净样本	1.6×10^{1}	4.9×10^{2}	1.5×10^{4}
		随机	1.6×10^{1}	4.8×10^{2}	1.4×10^{4}
		UAP	1.5×10^{1}	4.1×10^{2}	1.1×10^{4}
		PGD	1.3×10^{1}	3.3×10^{2}	1.0×10^{4}

6 结论与局限性

本研究表明，尽管近期的基于人工智能的解码器在标称加性高斯白噪声条件下能够优于传统基线方法，但其性能可能更容易受到专门针对这些模型构造的小幅最坏情况输入扰动的影响。无论是输入相关攻击还是通用攻击，在标准信道模型下取得的优势未必能在对抗扰动下保持，而置信传播解码器似乎较少受到针对人工智能解码器的扰动影响。未来工作应探索鲁棒性感知设计和更广泛的信道模型。

参考文献

[1] Claude E. Shannon. 通信的数学理论。贝尔系统技术期刊，27(3):379–423，1948年。[2] Claude Berrou、Alain Glavieux 和 Punya Thitimajshima。逼近香农极限的纠错编码与解码：Turbo码。载于*IEEE*国际通信会议（*ICC*）论文集，第1064–1070页，1993年。[3] Robert G. Gallager。低密度奇偶校验码。*IRE*信息论汇刊，8(1):21–28，1962年。[4] R. Michael Tanner。低复杂度码的递归方法。*IEEE*信息论汇刊，27(5):533–547，1981年。[5] Tom Richardson 和 Rüdiger Urbanke。现代编码理论。剑桥大学出版社，2008年。[6] 3GPP TS 38.212：新空口；复用与信道编码（第15版）。技术报告 TS 138 212 V15.2.0，第三代合作伙伴计划（3GPP）/ETSI，2018年7月。[7] Zeynep B. Kaykac Egilmez、Luping Xiang、Robert G. Maunder 和 Lajos Hanzo。5G极化码的发展、运行与性能。*IEEE*通信调查与教程，2020年。[8] Eliya Nachmani、Yair Be’ery 和 David Burshtein。利用深度学习解码线性码。载于第*54*届*Allerton*通信、控制与计算年会（*Allerton*），第341–346页。IEEE，2016年。[9] Tobias Gruber、Sebastian Cammerer、Jakob Hoydis 和 Stephan ten Brink。论基于深度学习的信道解码。*arXiv*预印本 *arXiv:1701.07738*，2017年。[10] Yoni Choukroun 和 Lior Wolf。纠错码Transformer。载于神经信息处理系统进展，2022年。[11] Seong-Joon Park、Hee-Youl Kwak、Sang-Hyo Kim、Yongjune Kim 和 Jong-Seon No。CrossMPT：用于纠错码的交叉注意力消息传递变换器。载于国际学习表征会议（*ICLR*），2025年。[12] Peyman Mohajerin Esfahani 和 Daniel Kuhn。利用Wasserstein度量的数据驱动分布鲁棒优化：性能保证与可处理重构。数学规划，171(1-2):115–166，2018年。[13] Ian J. Goodfellow、Jonathon Shlens 和 Christian Szegedy。解释并利用对抗样本。国际学习表征会议（*ICLR*），2015年。[14] Aleksander Madry、Aleksandar Makelov、Ludwig Schmidt、Dimitris Tsipras 和 Adrian Vladu。迈向能够抵御对抗攻击的深度学习模型。载于国际学习表征会议（*ICLR*），2018年。[15] Seyed-Mohsen Moosavi-Dezfooli、Alhussein Fawzi、Omar Fawzi 和 Pascal Frossard。通用对抗扰动。载于*IEEE*计算机视觉与模式识别会议论文集（*CVPR*），第1765–1773页，2017年。[16] Jeremy M. Cohen、Elan Rosenfeld 和 J. Zico Kolter。通过随机平滑实现认证对抗鲁棒性。载于第*36*届国际机器学习会议论文集（*ICML*），第97卷机器学习研究论文集，第1310–1320页，2019年。

[17] Eliya Nachmani、Yair Be'ery 和 David Burshtein。利用深度学习解码线性码。载于第*54*届*Allerton*通信、控制与计算年会（*Allerton*），第341–346页。IEEE，2016年。[18] Hyeji Kim、Yihan Jiang、Sreeram Kannan、Sewoong Oh 和 Pramod Viswanath。Deepcode：基于深度学习的反馈码。载于神经信息处理系统进展（*NeurIPS*），第31卷，2018年。[19] Yihan Jiang、Hyeji Kim、Himanshu Asnani、Sreeram Kannan、Sewoong Oh 和 Pramod Viswanath。Turbo自编码器：用于点对点通信信道的深度学习信道码。载于神经信息处理系统进展（*NeurIPS*），第32卷，2019年。[20] Yihan Jiang、Hyeji Kim、Himanshu Asnani 和 Sreeram Kannan。MIND：模型无关神经解码器。*arXiv*预印本 *arXiv:1903.02268*，2019年。[21] Chin Wa Ken Lau、Xiang Shi、Ziyan Zheng、Haiwen Cao 和 Nian Guo。置信传播与Transformer之间的相互作用：差分注意力消息传递 Transformer。载于*2025*年*IEEE*国际信息论研讨会（*ISIT*），第1–6页。IEEE，2025年。[22] Yoni Choukroun 和 Lior Wolf。去噪扩散纠错码。载于国际学习表征会议（*ICLR*），2023年。[23] Yoni Choukroun 和 Lior Wolf。用于纠错码的基础模型。载于国际学习表征会议（*ICLR*），2024年。[24] Haoyu Lei、Chin Wa Lau、Kaiwen Zhou、Nian Guo 和 Farzan Farnia。一致性流模型实现单步去噪纠错码。*arXiv*预印本 *arXiv:2512.01389*，2025年。[25] Meysam Sadeghi 和 Erik G. Larsson。针对端到端自编码器通信系统的物理对抗攻击。*IEEE*通信快报，23(11):1985–1988，2019年。[26] Alireza Bahramali、Milad Nasr、Amir Houmansadr、Dennis Goeckel 和 Don Towsley。针对基于深度神经网络的无线通信系统的鲁棒对抗攻击。载于*2021*年*ACM SIGSAC*计算机与通信安全会议（*CCS*）论文集，第126–140页，2021年。[27] Jinyin Chen、Jie Ge、Shilian Zheng、Linhui Ye、Haibin Zheng、Weiguo Shen、Keqiang Yue 和 Xiaoniu Yang。AIR：基于深度学习的信息恢复面临的对抗攻击威胁。*arXiv*预印本 *arXiv:2309.16706*，2023年。[28] Imre Csiszár 和 Prakash Narayan。重新审视任意变化信道的容量：正性与约束。*IEEE*信息论汇刊，34(2):181–193，1988年。[29] Michael Langberg。无知通信信道及其容量。*IEEE*信息论汇刊，54(1):424–429，2008年。[30] Yihan Zhang、Sidharth Jaggi、Michael Langberg 和 Anand D. Sarwate。因果对抗信道的容量。载于*2022*年*IEEE*国际信息论研讨会（*ISIT*），第766–771页，2022年。[31] Eric Ruzomberka、Chih-Chun Wang 和 David J. Love。计算能力受限的对手在观测方面的信道容量。*IEEE*信息论汇刊，69(11):6850–6881，2023年。[32] Peyman Mohajerin Esfahani 和 Daniel Kuhn。利用Wasserstein度量的数据驱动分布鲁棒优化：性能保证与可处理重构。数学规划，171(1-2):115–166，2018年。

[33] Amir Bennatan、Yoni Choukroun 和 Pavel Kisilev。用于线性码译码的深度学习——一种基于综合的方案。载于*2018*年*IEEE*国际信息论研讨会（*ISIT*），第1595–1599页。IEEE，2018年。[34] Yoni Choukroun 和 Lior Wolf。纠错码Transformer。神经信息处理系统进展，35：38695–38705，2022年。[35] Aleksander Madry、Aleksandar Makelov、Ludwig Schmidt、Dimitris Tsipras 和 Adrian Vladu。迈向能够抵御对抗攻击的深度学习模型。*arXiv*预印本 *arXiv:1706.06083*，2017年。[36] Seyed-Mohsen Moosavi-Dezfooli、Alhussein Fawzi、Omar Fawzi 和 Pascal Frossard。通用对抗扰动。载于*IEEE*计算机视觉与模式识别会议论文集，第1765–1773页，2017年。[37] Seong-Joon Park、Hee-Youl Kwak、Sang-Hyo Kim、Yongjune Kim 和 Jong-Seon No。CrossMPT：用于纠错码的交叉注意力消息传递变换器。载于第十三届国际学习表征会议，2025年。[38] Marc PC Fossorier、Miodrag Mihaljevic 和 Hideki Imai。基于置信传播的低密度奇偶校验码低复杂度迭代译码。*IEEE*通信汇刊，47(5)：673–680，1999年。[39] F.R. Kschischang、B.J. Frey 和 H.-A. Loeliger。因子图与和积算法。*IEEE*信息论汇刊，47(2)：498–519，2001年。[40] Ian J Goodfellow、Jonathon Shlens 和 Christian Szegedy。解释并利用对抗样本。*arXiv*预印本 *arXiv:1412.6572*，2014年。[41] Ching Lam Choi 和 Farzan Farnia。通用对抗方向。*arXiv*预印本 *arXiv:2210.15997*，2022年。[42] Yoni Choukroun 和 Lior Wolf。去噪扩散纠错码。载于第十一届国际学习表征会议，2023年。[43] Erdal Arikan。信道极化：一种为对称二进制输入无记忆信道构造容量逼近码的方法。*IEEE*信息论汇刊，55(7)：3051–3073，2009年。

附录 A 证明

A.1 命题 4.1 的证明

为证明命题 4.1，令 $V \sim \mathcal{N}\big(0,\ \nu^2 I_n\big)$ 并定义

$$\tilde{g}(u) \triangleq \mathbb{E}\Big[\ell\big(f(u+V), x^\star\big)\Big].$$

应用 Stein 引理可得

$$\nabla \tilde{g}(u) = \frac{1}{\nu^2}\,\mathbb{E}\Big[\ell\big(f(u+V), x^\star\big)\, V\Big]. \tag{21}$$

考虑二阶 求导并应用二阶 Stein 恒等式，进一步可得 :

$$\nabla^2 \tilde{g}(u) = \frac{1}{\nu^4}\,\mathbb{E}\Big[\ell\big(f(u+V), x^\star\big)\,\big(VV^\top - \nu^2 I_n\big)\Big]. \tag{22}$$

任取满足 $\|a\|_2 = 1$ 的单位向量 $a \in \mathbb{R}^n$ 。对式 (22) 取二次型,

$$a^\top \nabla^2 \tilde{g}(u)\, a = \frac{1}{\nu^4}\,\mathbb{E}\Big[\ell\big(f(u+V), x^\star\big)\,\Big(\big(a^\top V\big)^2 - \nu^2\Big)\Big]. \tag{23}$$

由 (7)， $0 \le \ell \le C$，因此

$$\big|a^\top \nabla^2 \tilde{g}(u)\, a\big| \le \frac{C}{\nu^4}\,\mathbb{E}\Big[\Big|\big(a^\top V\big)^2 - \nu^2\Big|\Big]. \tag{24}$$

由于 $a^\top V \sim \mathcal{N}\big(0,\ \nu^2\big)$，我们可以写成 $a^\top V = \nu Z$，其中 $Z \sim \mathcal{N}\in(0,\ 1)$。代入 (24) 得到

$$\big|a^\top \nabla^2 \tilde{g}(u)\, a\big| \le \frac{C}{\nu^4}\,\nu^2\,\mathbb{E}\big[|Z^2 - 1|\big] = \frac{C}{\nu^2}\,\mathbb{E}\big[|Z^2 - 1|\big]. \tag{25}$$

利用对称矩阵的恒等式,

$$\big\|\nabla^2 \tilde{g}(u)\big\|_{\mathrm{op}} = \sup_{\|a\|_2 = 1} \big|a^\top \nabla^2 \tilde{g}(u)\, a\big|,$$

由 (25) 可得

$$\big\|\nabla^2 \tilde{g}(u)\big\|_{\mathrm{op}} \le \frac{C}{\nu^2}\,\mathbb{E}\big[|Z^2 - 1|\big].$$

海森矩阵的一致算子范数界意味着 $\nabla \tilde{g}$ 是利普希茨连续的，其常数 β 由同一界给出，从而证明了 (8)。最后， $\mathbb{E}\big[|Z^2 - 1|\big] < 1$ （数值上为 ≈ 0.968），由此得到 (8) 中的严格不等式。■

A.2 命题 4.2 的证明

为证明命题 4.2，任取并固定一个满足 $\|\delta\|_2 = 1$ 的单位向量 δ 。对于每个指标 i，定义

$$\psi_i(\alpha) \triangleq \alpha\, \delta^\top q_i - \frac{\beta}{2}\alpha^2.$$

(17) 中的内层最大化等于 $\max_{\alpha \in \mathbb{R}} \psi_i(\alpha)$。由于 ψ_i 是凹二次函数，因此其最大化点唯一，可由一阶必要条件（FONC）求得:

$$\psi_i'(\alpha_i^*) = \delta^\top q_i - \beta \alpha_i^* = 0 \quad \Longrightarrow \quad \alpha_i^\star = \frac{\delta^\top q_i}{\beta}.$$

代入 $\alpha_i^\star$,

$$\max_{\alpha\in\mathbb{R}} \psi_i(\alpha) = \psi_i\big(\alpha_i^\star\big) = \frac{\big(\delta^\top q_i\big)^2}{2\beta}.$$

因此，(17) 中的目标函数正比于

$$\max_{\|\delta\|_2=1} \ \frac{1}{N}\sum_{i=1}^{N}\big(\delta^\top q_i\big)^2 = \max_{\|\delta\|_2=1} \ \delta^\top\Big(\frac{1}{N}\sum_{i=1}^{N} q_i q_i^\top\Big)\delta = \max_{\|\delta\|_2=1} \ \delta^\top \widehat{\Sigma}_q\,\delta.$$

瑞利商 $\delta^\top\widehat{\Sigma}_q\delta$ over $\|\delta\|_2 = 1$ 的最大化点恰好是 $\widehat{\Sigma}_q$ 的最大特征值对应的特征向量。■

A.3 定理 4.3 的证明

为证明该定理，我们在此证明 (18)–(20)，然后应用联合界。

A.3.1 (18) 中目标函数集中的证明

固定任意扰动 δ。根据有界损失假设 $0 \le \ell(\cdot) \le C$，对于所有 y，我们有:

$$0 \le g(y,\delta) = \mathbb{E}_V\big[\ell\big(f(y+V+\delta), x^\star\big)\big] \le C.$$

Hence $Z_i \triangleq g\big(Y_i,\ \delta\big)$ 是独立同分布的，并且几乎必然属于 $[0,C]$。霍夫丁不等式表明

$$\Pr\Big(\Big|\frac{1}{N}\sum_{i=1}^{N} Z_i - \mathbb{E}[Z_1]\Big| \ge t\Big) \le 2\exp\Big(-\frac{2Nt^2}{C^2}\Big).$$

由于 $\frac{1}{N}\sum_{i=1}^{N} Z_i = \widehat{F}_N(\delta)$ 和 $\mathbb{E}[Z_1] = F(\delta)$,

$$\Pr\big(\big|\widehat{F}_N(\delta) - F(\delta)\big| \ge t\big) \le 2\exp\Big(-\frac{2Nt^2}{C^2}\Big).$$

令右侧等于 $\eta/2$ ，得到 $t = C\sqrt{\log(4/\eta)/(2N)}$，从而以至少 $1-\eta/2$ 的概率证明了 (18)。

A.3.2 矩阵集中：式 (19) 的证明

定义 $\Sigma_q = \mathbb{E}\big[qq^\top\big]$ 和 $\widehat{\Sigma}_q = \frac{1}{N}\sum_{i=1}^{N} q_i q_i^\top$，并令

$$X_i \triangleq q_i q_i^\top - \Sigma_q.$$

于是 $\mathbb{E}[X_i] = 0$ 和 $\widehat{\Sigma}_q - \Sigma_q = \frac{1}{N}\sum_{i=1}^{N} X_i$。根据 $\|q_i\|_2 \le L$,

$$\|q_i q_i^\top\|_{\mathrm{op}} = \|q_i\|_2^2 \le L^2, \qquad \|\Sigma_q\|_{\mathrm{op}} = \big\|\mathbb{E}[qq^\top]\big\|_{\mathrm{op}} \le \mathbb{E}\big\|qq^\top\big\|_{\mathrm{op}} = \mathbb{E}\|q\|_2^2 \le L^2,$$

因此， $\|X_i\|_{\mathrm{op}} \le 2L^2$ 几乎必然成立。

令 $S \triangleq \sum_{i=1}^{N} X_i$。对于满足 $\|X_i\|_{\mathrm{op}} \le R$ 的独立、均值为零的自伴矩阵，标准的矩阵霍夫丁不等式给出

$$\Pr\big(\|S\|_{\mathrm{op}} \ge t\big) \le 2n\exp\Big(-\frac{t^2}{8\sum_{i=1}^{N} R^2}\Big) = 2n\exp\Big(-\frac{t^2}{8NR^2}\Big).$$

结合 $R = 2L^2$ 和 $t = N\varepsilon$,

$$\Pr\Big(\big\|\widehat{\Sigma}_q - \Sigma_q\big\|_{\mathrm{op}} \geq \varepsilon\Big) = \Pr\big(\|S\|_{\mathrm{op}} \geq N\varepsilon\big) \leq 2n \exp\Big(-\frac{N\varepsilon^2}{32L^4}\Big).$$

令右侧等于 $\eta/2$ ，得到

$$\varepsilon = 4\sqrt{2}\, L^2 \sqrt{\frac{\log\big(4n/\eta\big)}{N}},$$

这以至少 $1 - \eta/2$的概率证明了 (19)。

A.3.3 主成分集中：(20) 的证明

设 $E \triangleq \widehat{\Sigma}_q - \Sigma_q$和 $\Delta = \lambda_1(\Sigma_q) - \lambda_2(\Sigma_q) > 0$。Davis–Kahan 正弦定理（针对最大特征值对应的特征向量）给出

$$\sin\angle\big(\hat{u}_1, u_1\big) \leq \frac{\|E\|_{\mathrm{op}}}{\Delta} = \frac{\big\|\widehat{\Sigma}_q - \Sigma_q\big\|_{\mathrm{op}}}{\Delta}.$$

结合 (19) 可得 (20)。

A.3.4 联合界

(18) 和 (19) 中的事件各自以至少 $1 - \eta/2$ 的概率成立。根据联合界，它们同时成立的概率至少为 $1 - \eta$，这也意味着 (20) 成立。■

附录 B 详细实验设置

本节将详细介绍数据生成过程、神经解码器的训练细节，以及对抗攻击的超参数配置。

B.1 数据生成与码构造

信道模型。 我们使用二进制相移键控（BPSK）调制，模拟信号通过加性高斯白噪声（AWGN）信道传输。设 $\mathbf{c} \in \{0,1\}^n$ 表示码字。调制信号表示为 $\mathbf{x} = 1 - 2\mathbf{c}$，其中 $\mathbf{x} \in \{+1,-1\}^n$。接收信号 $\mathbf{y}$ 定义为：

$$\mathbf{y} = \mathbf{x} + \mathbf{z}, \quad \mathbf{z} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n), \tag{26}$$

其中，噪声方差 σ^2 由信噪比（SNR）通过 $\sigma^2 = (2R \cdot 10^{\mathrm{SNR}/10})^{-1}$ 确定，码率为 $R = k/n$。

B.2 训练细节：神经解码器

为确保公平比较，神经解码器（ECCT [10] 和 CrossMPT [37]）均使用相同的配置从头开始训练。对于每种码配置（n， k），我们按照 ECCT 和 CrossMPT 中的设置生成训练集。训练样本的信噪比从范围 [2分贝，8 分贝]内均匀采样生成。

附录 C 对抗攻击设置

在本研究中，我们评估了解码器针对五种不同攻击策略的鲁棒性。所有攻击均受 L_2-范数和给定能量约束 α 的限制，其中扰动 δ 必须满足$||\delta||_2 \leq \alpha||y||_2$。

随机噪声基线。 为了区分对抗攻击方法与对信道噪声的一般敏感性，我们选择随机噪声作为基线。该扰动通过从标准正态分布 $\mathcal{N}(0, I)$ 中采样一个向量，并将其归一化至给定的能量约束 $||\delta||_2 \leq \alpha||y||_2$ 来生成。

快速梯度法（FGM）。 FGM [40] 是一种单步攻击，它通过线性化损失函数来生成对抗样本。我们采用 L_2 归一化版本，即通过计算损失相对于输入的梯度，将其归一化为单位向量，再缩放至能量约束，从而计算扰动。

投影梯度下降（PGD）。 PGD [35]旨在 ϵ球内寻找最优扰动。我们将攻击的迭代次数设为 $T = 20$ 。我们以随机扰动作为初始化，并使用$\alpha = 1.2 \times (\epsilon/20)$的步长，将对抗扰动投影回满足能量约束的 ϵ球内。

通用对抗扰动（UAP）。UAP [36] 旨在寻找一种输入无关扰动向量，使整个数据集上的性能下降。我们使用随机梯度下降累积（SGDA），在包含 50 个批次的训练集上优化这一通用向量。优化采用学习率 $\eta = 0.05$，并在每一步对扰动进行裁剪，以确保其满足能量约束。

基于 PCA 的 UAP（UAP-PCA）。UAP-PCA [41]通过主成分分析（PCA）选择通用扰动。我们收集 50 个批次样本的梯度并进行 PCA。最终的攻击向量通过将扰动与梯度矩阵的第一主成分对齐，并将其缩放至满足能量约束来构造。

附录 D 补充实验结果

D.1 消融研究：能量约束的变化

本节展示扰动能量约束变化对 LDPC(121,60) 码影响的更多结果，如图 2 所示。我们观察到一个一致的趋势：随着能量约束增大，性能下降愈发严重。此外，与随机噪声基线相比，对抗攻击始终会导致显著更高的错误率。

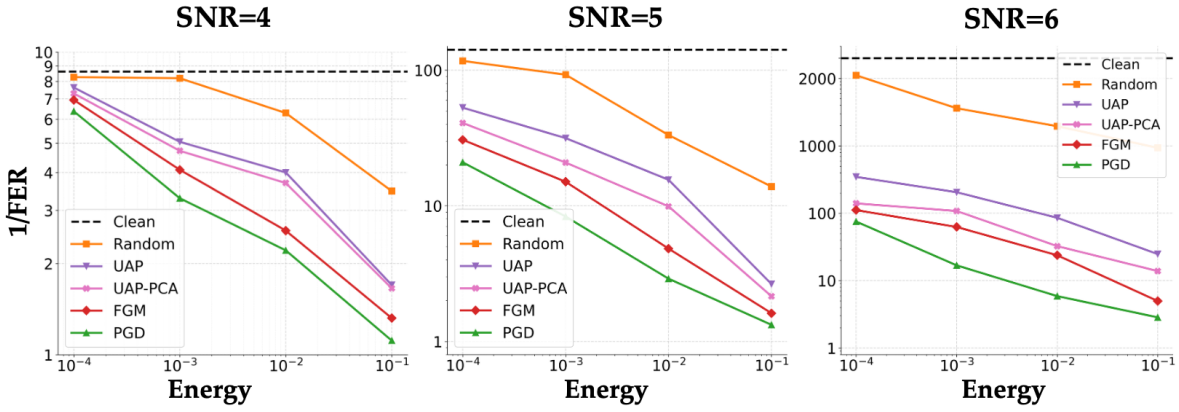


图 2：比较 不同能量约束下，ECCT 在 LDPC(121,60) 上的**FER-逆（1/FER）** .

D.2 对抗扰动的可迁移性

本节展示有关对抗扰动可迁移性的更多结果。具体而言，我们分别采用通用方法（UAP）和样本级方法（PGD），针对 LDPC(121,60)（表3）和 Polar(128,64)（表4）码上的 ECCT 模型生成最优扰动。随后，我们使用这些扰动测试各种基线方法，包括基于人工智能的解码器（ECCT、CrossMPT 和基于扩散的 DDECC [42]）以及传统译码器（用于 LDPC 的最小和、和积，以及用于 Polar 的 SC（逐次消除） [43], 和积）。结果表明，对抗攻击在人工智能解码器之间具有很强的可迁移性，而传统译码器仍然稳健，这凸显了基于神经网络的译码架构固有的脆弱性。

表3：在 LDPC(121,60) 上，跨解码器评估攻击可迁移性，指标为**FER-逆（1/FER）**。

Type	解码器	攻击	4 dB	5 dB	6 dB
AI Decoder	ECCT	干净样本	8.6×10^{0}	1.4×10^{2}	1.1×10^{4}
		随机	8.2×10^{0}	9.3×10^{1}	3.6×10^{3}
		UAP	5.1×10^{0}	3.2×10^{1}	8.0×10^{2}
		PGD	3.3×10^{0}	8.3×10^{0}	1.7×10^{1}
	CrossMPT	干净样本	2.1×10^{1}	6.1×10^{2}	1.7×10^{4}
		随机	1.9×10^{1}	2.7×10^{2}	3.8×10^{3}
		UAP	1.7×10^{1}	6.7×10^{1}	8.8×10^{2}
		PGD	9.2×10^{0}	3.2×10^{1}	5.5×10^{1}
	DDECC	干净样本	1.7×10^{1}	5.5×10^{2}	1.4×10^{4}
		随机	1.6×10^{1}	2.4×10^{2}	3.2×10^{3}
		UAP	1.4×10^{1}	6.5×10^{1}	9.5×10^{2}
		PGD	9.5×10^{0}	4.3×10^{1}	7.3×10^{1}
Traditional Decoder	最小和	干净样本	5.4×10^{0}	5.1×10^{1}	2.0×10^{3}
		随机	5.4×10^{0}	5.1×10^{1}	1.9×10^{3}
		UAP	5.3×10^{0}	5.0×10^{1}	1.7×10^{3}
		PGD	5.0×10^{0}	4.3×10^{1}	1.3×10^{3}
	和积	干净样本	1.6×10^{1}	4.9×10^{2}	1.5×10^{4}
		随机	1.6×10^{1}	4.8×10^{2}	1.4×10^{4}
		UAP	1.5×10^{1}	4.1×10^{2}	1.1×10^{4}
		PGD	1.3×10^{1}	3.3×10^{2}	1.0×10^{4}

表4: 通过**FER-逆（1/FER）**衡量的解码器间攻击可迁移性，适用于Polar(128,64) .

Type	解码器	攻击	4 dB	5 dB	6 dB
AI Decoder	ECCT	干净样本	8.5×10^{0}	7.4×10^{1}	1.4×10^{3}
		随机	8.1×10^{0}	5.9×10^{1}	6.1×10^{2}
		UAP	5.4×10^{0}	3.4×10^{1}	9.8×10^{1}
		PGD	3.0×10^{0}	5.4×10^{0}	8.3×10^{0}
	CrossMPT	干净样本	2.5×10^{1}	2.4×10^{2}	1.0×10^{4}
		随机	2.2×10^{1}	1.8×10^{2}	5.8×10^{3}
		UAP	1.9×10^{1}	8.4×10^{1}	9.2×10^{2}
		PGD	5.1×10^{0}	9.1×10^{0}	4.9×10^{1}
	DDECC	干净样本	5.2×10^{1}	7.4×10^{3}	9.7×10^{5}
		随机	4.9×10^{1}	6.5×10^{3}	7.1×10^{5}
		UAP	2.7×10^{1}	2.1×10^{3}	3.5×10^{4}
		PGD	9.7×10^{0}	7.5×10^{2}	2.2×10^{2}
Traditional Decoder	SC	干净样本	4.2×10^{0}	3.8×10^{1}	4.5×10^{2}
		随机	4.1×10^{0}	3.7×10^{1}	4.2×10^{2}
		UAP	3.8×10^{0}	3.4×10^{1}	3.7×10^{2}
		PGD	3.5×10^{0}	3.2×10^{1}	3.5×10^{2}
	和积	干净样本	5.6×10^{0}	5.9×10^{1}	6.8×10^{2}
		随机	5.4×10^{0}	5.7×10^{1}	6.5×10^{2}
		UAP	5.1×10^{0}	5.2×10^{1}	6.0×10^{2}
		PGD	4.9×10^{0}	5.0×10^{1}	5.6×10^{2}