arXiv:2611.32780v1 [cs.CR] 24 Nov 2026

C3H：多模态大语言模型推荐系统中的压缩到共识准则劫持

Guowei Guan Yurong Hao*Jiaming Zhang Fuyao Zhang Wei Yang Bryan Lim
南洋理工大学计算与数据科学学院新加坡

摘要

多模态大语言模型（MLLMs）正将推荐系统转变为内容感知决策系统。在这些系统中，多模态压缩和跨模态融合通常被视为增强鲁棒性、抵御内容操纵的机制，因为压缩能够过滤噪声信号，而跨模态共识能够抑制不一致的异常。我们通过揭示同一过程可能在排序决策中暴露出压缩到共识（C2C）偏差，对这一假设提出质疑。压缩后仍然显著、且获得各模态一致支持的信号，可能在候选项比较中产生不成比例的影响，从而为内容操纵提供可利用的途径。本文提出 **C3H**，一种新颖的推理时攻击，可利用这一漏洞。C3H 并不依赖噪声扰动或训练时投毒，而是通过调整目标项目的多模态内容，使其与推荐系统隐含表达的决策标准相契合，从而提升该项目的排名。大量实验表明，C3H 能显著提高目标项目的曝光度，同时维持整体推荐效用。代码见 https://github.com/GuoweiGuan666/C3H。

1 引言

多模态大语言模型（MLLMs）[1–3] 正在将推荐系统（RecSys）[4–7] 从以 ID 为中心的匹配引擎重塑为基于内容的决策系统。近期基于 MLLM 的推荐系统（MLLM-RecSys）通过多模态理解和候选项感知比较，对项目文本、图像和用户偏好证据进行推理 [8]。具有代表性的系统包括 VIP5 [3], , 该系统在候选项感知的 MLLM 下统一了视觉、语言和个性化；以及 MLLM-MSR [9], , 该系统将候选项与压缩后的多模态偏好摘要进行比较。这一转变改变了安全格局：当排序决策以可编辑的物品内容为基础时，现实中的对手不再局限于操纵交互或训练数据 [10]。平台上的卖家可能会尝试在平台允许的内容空间内编辑其物品的文本和图像，以推广自己的物品，而受害推荐系统则已部署、冻结并

在攻击构造过程中从未被查询。

乍看之下，这种对手似乎并不强大。候选感知型多模态大语言模型推荐系统不会直接对原始内容进行评分：丰富的多模态输入会先被压缩为紧凑表示，随后进行融合，再进行比较。例如，VIP5 会通过学习到的映射网络将每幅图像映射为少量视觉词元，而 MLLM-MSR 则会在候选项评分之前，将物品内容和用户历史记录提炼为紧凑的多模态摘要。压缩通常被认为有助于提升鲁棒性 [11], , 而跨模态融合则进一步强化了这一判断：当某种操纵仅出现在一种模态中时，

*通讯作者：yurong.hao@ntu.edu.sg

预印本。

另一种模态可以充当语义锚点，削弱其影响 [12]。因此，现有针对多模态推荐系统的攻击往往假设攻击者具备更强的能力，例如对训练过程进行投毒、获取模型反馈，或运行自适应扰动搜索 [12–16]。在严格的零查询设定下，普遍的直觉是，纯内容操纵应当只能有限地影响推荐结果。

这种直觉忽略了压缩与融合所带来的结构性后果。虽然这些机制会抑制偏离准则的噪声，但它们也决定了哪些物品线索能够进入最终的候选项比较。由于丰富的物品内容必须经过紧凑的表示瓶颈，并非所有文本或视觉证据都能同等程度地促成排序。压缩后仍保持可区分性的线索更有可能保留在候选项表示中，而同时得到文本和图像支持的线索则会在跨模态融合过程中得到进一步强化。我们将这一过程称为压缩到共识（C2C）决策路径：物品证据首先经过压缩，然后通过跨模态共识得到强化或削弱，最终用于候选项比较。由此形成的排序偏好就是**C2C偏差**：候选感知型多模态大语言模型推荐系统偏好那些既能在多模态压缩后保留下来，又能获得跨模态强化的线索。

这种偏差为零查询攻击者提供了一种局部但可利用的先验。其强度取决于隐藏参数和微调数据，但其形态由公开的架构选择所固定：视觉编码器、分词方案、压缩模块和融合结构共同决定信息如何从原始内容流入候选项比较。因此，即使无法查询受害模型，卖家也可以推断哪些线索更可能被保留并得到强化，并编辑目标项目，使其文本和视觉证据符合这种偏好。过滤传统偏离准则的扰动的同一机制，随后会放大符合准则的内容操纵。我们将这种攻击范式称为**压缩到共识准则劫持**，简称**C3H**。C3H并不强行制造一个明显错误的感知决策，而是生成一个在模型自身决策标准下看起来值得推荐的目标项目。这一框架与文本排序中的决策准则劫持相关 [17], ，但这里的攻击面有所不同：可利用的标准源自多模态压缩和跨模态融合，且攻击者无法获得受害模型的反馈。不同于利用关键词频率、流行度或外部流量的传统SEO，C3H针对的是基于多模态大语言模型的推荐的架构决策标准。

我们将 C3H 实例化为一个零查询推理时流水线。关键挑战在于无法访问受害模型的情况下，近似其有利于C2C的方向。C3H 首先使用 CLIP 等公开可用的预训练编码器，估计哪些物品线索可能同时从压缩存活和跨模态共识中获得优势。随后，它构建与该估计方向一致的编辑后的目标内容（$\tilde{t}_{i^*}$，$\tilde{v}_{i^*}$），使目标项目 i^* 与该方向保持一致；同时，基于约束的安全门控可保留隐蔽性、语义有效性、图文一致性以及符合平台情境的呈现方式。在实证方面，C3H 具有以下三项特性：(i) 在 VIP5 上，针对 MicroLens 和 Beauty，在十五种基线方法的比较中，C3H 在 ER@20 指标上的每个（数据集，提示）单元格中均取得最佳结果，提升幅度为 1。4–1.8×；(ii) 同一内容无需修改即可迁移至 Qwen2-VL-2B-Instruct，并在全部三项零样本 ER@20 比较中获胜；(iii) C3H 在不降低良性用户效用的情况下，能够抵御三种内容过滤防御（$\Delta\text{ER} < 0.004$）。

贡献。

- 我们发现并通过实证验证了候选感知型多模态大语言模型推荐系统中的压缩到共识偏差：排序受到压缩后仍然保留且获得跨模态共识的线索的影响尤为显著。据我们所知，我们首次在严格的黑盒、零查询、单卖家威胁模型下，对多模态大语言模型推荐系统中的标准劫持进行了形式化。
- 我们提出了**C3H**，这是一种零查询流水线，通过公共编码器引导的目标估计和隐蔽性约束内容构造来实现这一攻击。
- 在三个数据集和两种受害者架构上的实验表明，该方法在保持良性用户效用的同时，显著提升了目标曝光度。

2 相关工作

基于多模态大语言模型的推荐系统。 近年来，多模态大语言模型的进展推动推荐系统从以 ID 为中心的匹配转向内容驱动架构，通过整合文本描述、视觉内容和用户行为 [3, 9, 8]。代表性系统包括 VIP5 [3], ，它通过参数高效微调，在候选感知型多模态大语言模型框架下统一视觉、语言和个性化，以及 MLLM-MSR [9], ，后者将候选项与压缩后的多模态偏好摘要进行比较

。尽管这类系统的部署日益广泛，但跨模态压缩与融合下多模态大语言模型推荐系统的安全性仍未得到充分研究。

推荐系统对抗攻击。 现有攻击分为四类。交互级攻击 操纵用户-物品记录（例如 RandAttack [18], PopAttack [19]），但对内容驱动的多模态大语言模型推荐系统的定向影响有限，因为其排序由融合的多模态表示决定 [12]。单模态内容扰动 在推理时编辑文本 [20–24]或图像 [14, 25–28] ，但跨模态共识通常会抑制此类攻击：未受扰动的模态充当语义锚点，能够纠正不一致性。跨模态训练时投毒 攻击（VENOMREC [12],Shadowcast [13]）通过在模态之间同步扰动来应对这种鲁棒性，但需要访问微调数据或训练流程。基于查询的攻击 （例如 CheatAgent [29] 以及隐蔽的大语言模型推荐系统方法 [30] ）则利用受害模型输出反馈来估计对抗方向。这两类方法都假设具备严格零查询推理时部署所不允许的能力。

推理时决策标准劫持。一项互补研究方向探讨了针对文本排序器的标准劫持 [17], ，其中对手利用排序模型对相关性的评分方式，而不是强制模型产生错误分类。这一视角为我们的威胁模型提供了参考，但仅适用于文本排序；我们所利用的压缩到共识机制是候选感知多模态推荐系统所特有的。针对图像分类器的黑盒迁移攻击 [31] 同样使用替代模型，但缺乏实现跨模态协调的结构性机制。据我们所知，C3H 是首个同时具备以下四种特性的、针对多模态大语言模型推荐系统的攻击：严格的零查询操作、推理时内容编辑、单卖家范围，以及基于结构性压缩到共识偏差的跨模态协调。

3 问题表述

3.1 受害多模态大语言模型推荐系统框架

我们考虑候选感知型多模态大语言模型推荐系统，以 VIP5 [3]和 MLLM-MSR [9],为例；这类系统将推荐建模为在压缩的多模态物品表征上进行序列到序列生成。令 $\mathcal{U}$ 和 $\mathcal{I}$ 分别表示用户集合和物品集合；每个物品 $i \in \mathcal{I}$ 都带有文本标题/描述t_i 以及商品图像 v_i。

提示构造与排序。 对于用户 u 及其交互历史 $\mathcal{H}_u = [i_1, \ldots, i_L]$，每个物品 i_k都会被编码为多模态词元序列 $\mathbf{e}_k = \tau(t_{i_k}) \oplus \zeta(g(v_{i_k}))$，其中 $\tau(\cdot)$ 是文本分词器， $g(\cdot)$ 是冻结视觉编码器（例如 CLIP-ViT [32]）， $\zeta(\cdot)$ 是可训练投影器。用户侧提示构造如下

$$X_u = \mathcal{P} \oplus \tau(u) \oplus [\mathbf{e}_1, \ldots, \mathbf{e}_L], \tag{1}$$

使用 $\mathcal{P}$ 指令模板和 $\tau(u)$ 文本化用户标识符。多模态大语言模型的注意力层充当跨模态融合机制 Φ_{Θ^*}。在服务时，部署的推荐系统 f_{Θ^*}利用用户上下文和候选物品的多模态内容（t_i， v_i）为每个候选物品 $i \in \mathcal{C}_u$ 评分，并返回排名前 K 的物品。 Θ^* 表示用于将 f_{Θ^*} 与任意公共骨干网络（例如 PEFT 风格实例中的轻量级适配器，如 VIP5）区分开的受害模型特定参数。由于评分通过融合表示进行，任何对（t_i， v_i）的编辑都会通过 Φ_{Θ^*} 传播，并可能改变该物品的排名。我们聚焦于给定候选集 $\mathcal{C}_u \subseteq \mathcal{I}$ 中的排序阶段，该候选集包含目标 i^*。

3.2 决策标准劫持

我们研究针对已部署的多模态大语言模型推荐系统 f_{Θ^*}（参数已冻结Θ^*）的定向推广。对手是一位卖家，拥有单个目标项目 $i^* \in \mathcal{I}$，希望通过编辑其已发布文本和图像来提高其曝光度；这是大型电子商务平台上的标准功能。我们将 Qian 等人 [17] 针对文本排序器提出的框架改编至候选感知多模态推荐系统，并将这种推理时攻击形式化为决策准则劫持。

定义 3.1（决策准则劫持）。 给定一个已部署的受害推荐系统 f_{Θ^*}，其参数为冻结状态 Θ^*，一次攻击是一个元组 $\mathcal{A} =$ （i^*、 T_ϕ、 $\mathcal{B}$），其中：

- $i^* \in \mathcal{I}$ 是卖家拥有的唯一目标项目；
- $T_\phi : (t_{i^*}, v_{i^*}) \mapsto (\tilde{t}_{i^*}, \tilde{v}_{i^*})$ 将原始文本内容和图像映射为编辑后的对；
- $\mathcal{B}$ 是一个用于限定编辑后内容的隐蔽性约束集。

该攻击劫持了 f_{Θ^*} 的决策标准，即模型根据内容为候选项评分时所采用的结构偏好；它通过调整 i^* 的内容，使其借助 T_ϕ 与这些偏好保持一致，而不是通过欺骗模型来实现。

隐蔽性约束 $\mathcal{B}$。约束集 $\mathcal{B}$ 要求 $(\tilde{t}_{i^*}, \tilde{v}_{i^*})$ 满足：(i) 视觉保真度，$\|\tilde{v}_{i^*} - v_{i^*}\|_p \le \varepsilon_v$；(ii) 类别保持，$\mathrm{sim}_{\mathrm{cat}}(\tilde{v}_{i^*}, c_{i^*}) \ge \tau_{\mathrm{cat}}$；(iii) 文本内容保真度，$\mathrm{sim}_{\mathrm{text}}(\tilde{t}_{i^*}, t_{i^*}) \ge \tau_{\mathrm{orig}}$；以及 (iv)跨模态一致性，$\mathrm{sim}_{\mathrm{CLIP}}(\tilde{t}_{i^*}, \tilde{v}_{i^*}) \ge \tau_{\mathrm{align}}$。条件 (i)–(iii) 确保单模态自然度；(iv) 确保跨模态一致性。

攻击者的知识与能力。我们假设处于严格的黑盒、零查询设定下。对手无法访问 Θ^*、梯度、微调数据，也无法通过任何查询接口访问 f_{Θ^*}。只有 i^*可以被修改；其他所有项目、用户和训练数据均不受影响。在 i^* 中，对手可以提出任何满足 $\mathcal{B}$ 的 T_ϕ，并可在自己的设备上使用公开可用的预训练编码器（例如 CLIP）来规划编辑。这些编码器从未针对 f_{Θ^*} 进行校准；它们与受害者流水线中冻结子组件的重叠仅是基于公开知识的先验，而非查询访问。

攻击者的目标。对手旨在最大化这样一部分用户的比例：对于这些用户，i^* 会出现在前 K 个列表中。我们将曝光率定义为

$$\mathrm{ER}_K(i^*; \tilde{t}_{i^*}, \tilde{v}_{i^*}) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbb{I}\Big[i^* \in \text{Top-}K\big(u; f_{\Theta^*}, \tilde{t}_{i^*}, \tilde{v}_{i^*}\big) \Big], \tag{2}$$

其中，仅替换 i^* 的内容为（$\tilde{t}_{i^*}$，$\tilde{v}_{i^*}$）。该攻击可归结为

$$\max_{T_\phi} \ \mathrm{ER}_K\big(i^*; T_\phi(t_{i^*}, v_{i^*})\big) \quad \text{s.t.} \quad T_\phi(t_{i^*}, v_{i^*}) \in \mathcal{B}(t_{i^*}, v_{i^*}). \tag{3}$$

我们在一组可攻击子集 $\mathcal{I}^\dagger \subseteq \mathcal{I}$ 中的物品上进行评估；干净推荐系统并未将这些物品排在显著位置，因此测得的效果不会受到推荐系统既有偏差的混淆。目标选择仅用于评估；C3H 在攻击构造过程中不会使用干净受害者预测。我们跟踪良性用户上的 HR@K 和归一化折损累计增益@K，作为效用检查。

4 压缩到共识偏差的实证证据

我们在 VIP5 受害模型上验证两项主张：(i) 即使从预筛选池中随机选择主角，面向 C2C 的内容操纵仍能显著提升目标曝光；(ii) 联合评分 $S(f) = \tilde{\kappa}(f) \cdot \tilde{\chi}(f)$ 还能进一步提升曝光，效果优于随机选择和单轴选择。

设置。按照完整 C3H 流程（算法 1）固定模块 M1、M5–M7，仅改变 M4 主角选择标准，设置四种条件：随机 （从 $\mathcal{F}$中均匀抽取）、 $\tilde{\kappa}$-仅、 $\tilde{\chi}$-仅和完整 *C2C*。四种条件均仅使用公开的 CLIP 编码器（ V使用 B/16， A_{vis}使用 B/32），且不发起受害者查询。目标是 $N{=}9$个 MicroLens 可攻击物品，并在 $N{=}14$个 Beauty 目标上重复实验；每个条件下的 ER@20 均取 1,000 名良性用户的平均值。

表 1：在 **MicroLens** ($N{=}9$) 上进行的线索选择消融。每个单元格的 ER@20 基于 1k 名良性用户计算。

M4 标准	**无攻击**	**随机**	**仅限 $\tilde{\kappa}$**	**仅限 $\tilde{\chi}$**	**C3H（完整 C2C）**
ER@20 （少样本）	0.0000	0.1732	0.1752	0.1749	**0.2727**
ER@20 （零样本）	0.0000	0.1761	0.1787	0.1783	**0.2739**

与 C2C偏差一致的证据（表 1）。在 $\mathcal{F}$ 中随机选择主角即可达到 ER@20 为 0.173（少样本）/ 0.176（零样本），远高于无攻击基线。由于随机选择不会引入线索优先级信号，因此无需显式进行线索排序也能观察到这一提升；M1 根据跨模态 CLIP 相似度对候选池进行预筛选。仅依据 $\tilde{\kappa}$ 或 $\tilde{\chi}$ 单一维度进行选择，结果与随机选择相差不超过 ± 0.003。这些观察表明，在预筛选池中仅依据任一维度进行选择，收益有限，因此有必要采用联合评分策略。

联合评分进一步优化了池内选择。完整 C2C 评分使平均 ER@20 相较于随机选择提升约 $+0.099$（少样本）/ $+0.098$（零样本），达到 0.273 / 0.274（表 1）。这些结果表明，在这些设置中，联合评分是一种有实证依据的线索选择启发式方法。

对这两个维度的解读。分数 $\tilde{\kappa}$ 和 $\tilde{\chi}$ 分别代表压缩生存能力和跨模态共识。在 cue 选择消融实验（表 1）中，完整 C2C 的平均 ER@20 高于任一单轴选择标准。这一观察结果支持在所评估的设置中将联合评分作为一种经验性线索选择启发式方法，但并不能证明这些分数在统计上相互独立，也不能证明受害模型中存在因果机制。

5 方法：C3H

5.1 方法概述

代理分解。 在零查询设定下，无法直接求解式 (3)：其外层目标 ER_K 取决于受害模型 f_{Θ^*}，而其参数和每个用户的预测结果均不可见。遵循黑盒攻击 [31], 的代理迁移模式，我们用代理函数 $\hat{\phi}(\tilde{t},\ \tilde{v})$ 替代 ER_K，该函数仅使用公共编码器，根据内容与 C2C偏差的对齐程度进行评分。我们假设，输入侧的压缩结构由公共的冻结骨干网络（CLIP-ViT、T5）驱动，而非轻量级微调适配器 Θ^* [32, 33]；因此，在公共编码器上得到的线索排序应能迁移至受害模型，我们将在第 4 节通过实证进行检验。

两阶段框架。 C3H 将这一代理优化实现为两阶段流水线（图 1）。为简洁起见，我们将 $(t_{i^*},\ v_{i^*})$ 简写为 $(t_0,\ v_0)$，将 $(\tilde{t}_{i^*},\ \tilde{v}_{i^*})$ 简写为 $(\tilde{t},\ \tilde{v})$。阶段 I 仅根据干净内容，沿着压缩可存活性和跨模态共识对线索进行评分，并选出一条主线索 f_h 和一条次要线索 f_s。第二阶段先塑造文本，再根据新文本对齐图像，最后通过干净基线回退检查隐蔽性，从而构建 $(\tilde{t},\ \tilde{v})$。该流水线具有确定性，每个目标仅运行一次，且不会向 f_{Θ^*} 发出查询。

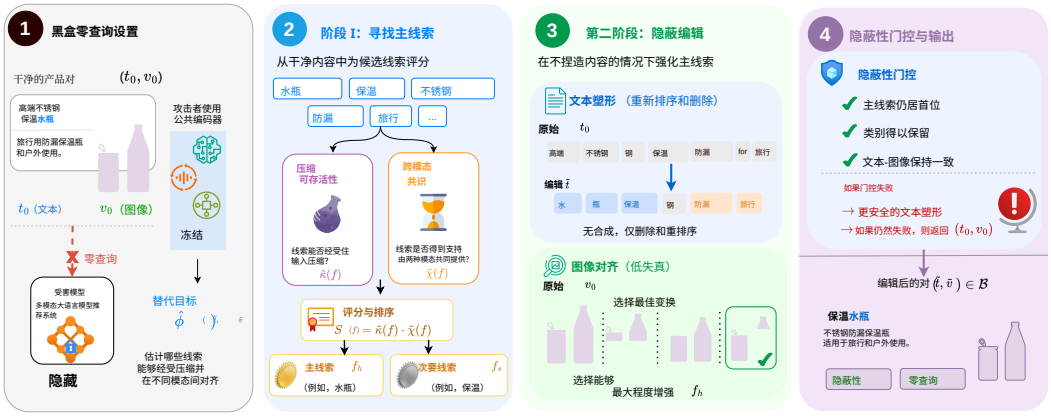

图1：C3H 的示意图。在严格的零查询设定下，该框架分两个阶段运行：(1) **线索评分（阶段 I）**根据压缩可存活性 $\hat{\kappa}$ 和跨模态共识 $\hat{\chi}$，对 (t_0, v_0) 中的候选线索进行排序，并通过联合评分 $S = \tilde{\kappa} \cdot \tilde{\chi}$选出主线索 f_h 和次要线索 f_s。(2) **隐蔽编辑（第二阶段）**通过文本塑形和低失真图像变换来强化 f_h, f_s，并以隐蔽性集合 $\mathcal{B}$作为门控条件。

5.2 阶段 I：线索评分与主角选择

线索挖掘。 我们通过收集 t_0 的名词短语块，并仅保留其与 v_0 的 CLIP 跨模态相似度超过低召回率阈值 [34,32] 的片段，从 (t_0, v_0) 中提取候选线索集 $\mathcal{F}$。每个 $f \in \mathcal{F}$ 都会标注一个位置类别（标题、正文、隐含），该类别将作为 A_{text} 在共识分数中的位置权重。

压缩存活度评分（$\hat{\kappa}$）。只有当某条线索的信号在受害模型的输入压缩后仍然存活时，该线索才会参与排序。信息瓶颈理论预测，压缩会保留任务相关特征 [35, 33]；$\hat{\kappa}$ 采用或门形式，用于衡量某条线索是否至少在一种模态上仍然可见：

$$\hat{\kappa}(f) = 1 - \big(1 - C(f)\big)\big(1 - V(f)\big), \tag{4}$$

其中，$C(f)$表示 t_0 中 f 的 TextRank 中心性，在长度为 $L_{\text{text}} \in \{20, 40, 80\}$的文本截断上取平均；$V(f)$表示在 CLIP-ViT-B/16 注意力展开 [36], 下，与 f匹配的前$\lceil P/4 \rceil$ 个图像块上的平均注意力权重（依据与线索文本嵌入的余弦相似度），其中 P=196 采用224×224 输入图像分辨率。

跨模态共识评分（$\hat{\chi}$）。达到融合是必要条件，但并非充分条件：纯注意力层往往会导致排序坍塌，因此缺乏跨模态支持的线索在融合过程中无法得到增强 [37]。$\hat{\chi}$ 因此要求两种模态都支持该线索（与门）：

$$\hat{\chi}(f) = A_{\text{text}}(f) \cdot A_{\text{vis}}(f), \tag{5}$$

其中，$A_{\text{text}}(f) \in \{1.0, 0.5, 0.2\}$ 表示分配给每条线索（标题/正文/隐式）的预定义位置权重，而 $A_{\text{vis}}(f)$则是在 CLIP-ViT-B/32 下，原始图像v_0 与形式为"a photo of {f}"的提示模板之间的平均余弦相似度。

聚合与主角选择。我们对每个轴进行秩归一化，得到 $\tilde{\kappa}(f) = 1 - (r_\kappa(f) - 1)/|\mathcal{F}|$（类似地），其中 r_κ，r_χ 表示$\hat{\kappa}$，$\hat{\chi}$上的降序排名。压缩至共识评分为$S(f) = \tilde{\kappa}(f) \cdot \tilde{\chi}(f)$，主角线索和次级线索为（$f_h$，$f_s$）$=$top-$2_{f\in\mathcal{F}}\, S(f)$；当$|S(f_h) - S(f_s)| < 0.05$ 相同时，则由较高的 $\hat{\chi}$ 线索打破平局。[2]

5.3 第二阶段：在 $\mathcal{B}$ 中构建内容

第二阶段会在编辑后的内容中强化 f_h，同时确保结果仍处于 $\mathcal{B}$之内。我们将替代目标 $\hat{\phi}(\tilde{t}, \tilde{v}) = S(f_h; \tilde{t}, \tilde{v})$ 定义为 f_h 在编辑内容上根据公式 (4)–(5) 重新计算得到的 C2C 评分：$\hat{\phi}$ 越大，表示 f_h的信号在压缩后保留得越充分，并且在各模态之间的对齐越紧密。最大化 $\hat{\phi}$ 可为无法评估的曝光率 ER_K 提供无查询代理指标（已在第 6 节通过实证验证）。

文本塑形。文本塑形算子通过删除和重排序 t_0来增强 f_h 和 f_s，不合成任何词元。它会标记停用词、填充词和负面情绪词元以供删除，对与 f_h 或 f_s重叠的词元应用白名单覆盖以予以保留，并按以下顺序重排保留的词元：先放置 f_h 词元，其次是 f_s，然后按 $S(f)$ 降序排列其他线索词元，最后放置剩余的内容词元。这保证了$\mathcal{t} \subseteq t_0$ 作为一个多重集保持不变（不引入虚构内容）；随后由隐蔽性门控检查τ_{orig}。

图像对齐。图像对齐算子从有限的变换族$\mathcal{T} = \{\theta_0, \ldots, \theta_4\}$ 中选出最佳变换。该变换族涵盖电商产品摄影中常见的低失真编辑（空间构图和光度调整），并排除会明显违反商品刊登规范的几何畸变。五种变换为：θ_0，恒等变换；θ_1，θ_2，以 f_h，f_s 为中心的显著性引导裁剪，通过 Grad-CAM [38]实现；θ_3，$0.7\times$ 中心裁剪；θ_4，对主图区域进行轻微的对比度和亮度调整。当 $\lambda = \hat{\chi}(f_s)/\hat{\chi}(f_h)$（若$\hat{\chi}(f_h) = 0$，则中止聚合）时，选择规则会按以下方式为每种变换评分

$$\text{score}(\theta) = S(f_h; \tilde{t}, T_\theta(v_0)) + \lambda \cdot S(f_s; \tilde{t}, T_\theta(v_0)), \tag{6}$$

并设置 $\tilde{v} = T_{\theta^*}(v_0)$，其中 $\theta^* = \arg\max_{\theta\in\mathcal{T}} \text{score}(\theta)$。每个候选物品都会重新计算$\hat{\kappa}$ 和 $\hat{\chi}$，以捕捉文本重排造成的位置变化。视觉保真度由该系列本身的设计保证。[3]

隐蔽性门控。该门控分四个连续阶段检查 $(\tilde{t}, \tilde{v}) \in \mathcal{B}$：(1) 在重新计算的$\hat{\kappa}$，$\hat{\chi}$下，重新检查主图线索的 argmax；(2) 在 $\tau_{\text{cat}} = 0.20$处应用类别保留门控；(3) 在 CLIP-ViT-B/32 [32]下应用文本保真度门控 ($\tau_{\text{orig}} = 0.18$) 和跨模态一致性门控 ($\tau_{\text{align}} = 0.18$)；若阶段1–3未通过，则 (4) 使用仅删除最安全前缀的方式重新运行文本整形并再次测试。若第4阶段未通过，该门控将使用 (t_0，v_0)。伪代码和超参数分别见算法 1 和附录B。

流程伪代码。算法 1 总结了上述完整的 C3H 流程。

[2]当 $|\mathcal{F}| < 4$ 时，我们直接使用$\hat{\kappa} \cdot \hat{\chi}$。$\hat{\kappa}$ 和$\hat{\chi}$ 的基于 CLIP 的组件使用互不重叠的骨干网络（B/16 图块级、B/32 全局级），因此两个轴依赖于架构独立信号。[3]当图像边长小于 200 px 时，$\mathcal{T}$回退至 $\{\theta_0, \theta_4\}$；小于 100 px 时，回退至 $\{\theta_0\}$。

算法 1 C3H 流程。阶段 I（M1–M4）通过公共编码器对线索的 C2C偏差进行评分；第二阶段（M5–M7）构造（ $\tilde{t}$, $\tilde{v}$），位于隐蔽性集合 $\mathcal{B}$ 内。

要求： 干净内容（t_0, v_0）、目标类别 c_{i^*}、公共编码器 E_{16} (B/16) 和 E_{32} (B/32), 阈值 τ_{cat}, τ_{orig}, τ_{align}、变换族 $\mathcal{T} = \{\theta_0, \dots, \theta_4\}$
确保： 编辑后的对（ $\tilde{t}$, $\tilde{v}$）$\in \mathcal{B}$，或干净基线（t_0, v_0）

$\mathcal{F} \leftarrow$ CUEMINE(t_0, v_0) ▷ M1
如果 $|\mathcal{F}| < 2$ **则**
  返回 (t_0, v_0) ▷ 线索不足
结束如果
对于 $f \in \mathcal{F}$ **执行**
  $\hat{\kappa}(f) \leftarrow 1 - (1 - C(f))(1 - V(f; E_{16}))$ ▷ M2，公式 (4)
  $\hat{\chi}(f) \leftarrow A_{\text{text}}(f) \cdot A_{\text{vis}}(f; E_{32})$ ▷ M3，公式 (5)
结束循环
如果 $|\mathcal{F}| < 4$，**那么**
  $S(f) \leftarrow \hat{\kappa}(f) \cdot \hat{\chi}(f)$ ▷ M4
否则
  $\tilde{\kappa}, \tilde{\chi} \leftarrow$ 秩归一化$(\hat{\kappa}, \hat{\chi})$
  $S(f) \leftarrow \tilde{\kappa}(f) \cdot \tilde{\chi}(f)$
结束 if
$(f_h, f_s) \leftarrow$ 顶部-$2_{f \in \mathcal{F}}\, S(f)$ ▷ 根据 $\hat{\chi}$ 打破平局，当 $|S(f_h) - S(f_s)| < 0.05$
如果 $\hat{\chi}(f_h) = 0$**则**
  返回 (t_0, v_0) ▷ 中止聚合
结束 if
$\tilde{t} \leftarrow$ SHAPETEXT(t_0, f_h, f_s) ▷ M5：删除 +重排
$\lambda \leftarrow \hat{\chi}(f_s)/\hat{\chi}(f_h)$
对于 $\theta \in \mathcal{T}$ **执行**
  评分$(\theta) \leftarrow S(f_h;\ \tilde{t}, T_\theta(v_0)) + \lambda \cdot S(f_s;\ \tilde{t},\ T_\theta(v_0))$ ▷ M6，公式（6）
23： **</style}结束循环**
$\theta^* \leftarrow$ 最大值参数$_\theta$ 分数(θ); $\tilde{v} \leftarrow T_{\theta^*}(v_0)$
25：**如果** M7 的阶段1–3通过（ $\tilde{t}$, $\tilde{v}$） ▷ M7验证
</style}则
  返回 $(\tilde{t}, \tilde{v})$
否则如果第4阶段产生 $(\tilde{t}_{\text{fb}}, v_0)$ 通过第1至第3阶段 **则**
  返回 $(\tilde{t}_{\text{fb}}, v_0)$
否则
  返回 (t_0, v_0) ▷ 中止，恢复至干净基线
结束 if

6 实验评估

6.1 实验设置

数据集与受害者模型。 我们在三个多模态推荐基准上进行评估：**MicroLens-100K** [39], 中文短视频语料库，以及**亚马逊 BEAUTY** 和 **亚马逊 Sports** [40]。遵循先前的工作 [3, 30], ，我们使用预先提取的 CLIP-ViT-B/32 视觉嵌入 [32] ，并采用留一法评估。主要受害模型是 **VIP5** [3], ，它是基于冻结的 T5-small 主干网络 [41]构建、并通过 PEFT适配器 [42]进行微调的候选感知型多模态大语言模型推荐系统。为评估跨架构迁移，我们还在少样本和零样本提示下使用 **Qwen2-VL-2B-Instruct** [43] 。

基线方法。 我们在三类攻击中，将 C3H 与十五种具有代表性的攻击基线进行比较。**交互级攻击**：RandAttack [18] 和 PopAttack [19]。**文本攻击**：TextFooler [20],DeepWordBug [22], BERT-Attack [21], TextBugger [23], CheckList [24], 和 CheatAgent [29]。**图像攻击**：EXPA、INSA、SEMA [14],VAA [44], IPDGI [25], UAP [26], 和 AdvCLIP [27]。

评估指标。有效性：曝光率（ER@K），即 i^*出现在前K 个列表中的用户比例。**隐蔽性**：用于衡量文本保真度的ROUGE-$1/2/L$ [45] ，以及用于衡量跨模态一致性的语义对齐差距SA Δ （相对于

表2：在少样本和零样本提示词下，**MicroLens** 和**Beauty**上的攻击有效性与推荐效用主要结果。每个单元格中最佳曝光率以**粗体并加下划线**标示。

攻击	少样本（*MicroLens*）						零样本（*MicroLens*）					
	前十			前二十			前十			前二十		
	HR	**NDCG**	**ER**	**HR**	**NDCG**	**ER**	**HR**	**NDCG**	**ER**	**HR**	**NDCG**	**ER**
无攻击	0.2530	0.1210	0.0000	0.4320	0.1657	0.0000	0.2510	0.1227	0.0000	0.4310	0.1679	0.0000
DeepWordBug	0.2637	0.1331	0.0360	0.4458	0.1786	0.1776	0.2627	0.1349	0.0573	0.4501	0.1817	0.1854
TextFooler	0.2568	0.1263	0.0263	0.4363	0.1712	0.1671	0.2706	0.1352	0.0359	0.4523	0.1808	0.1679
BertAttack	0.2564	0.1265	0.0343	0.4391	0.1721	0.1621	0.2626	0.1318	0.0469	0.4474	0.1781	0.1523
CheatAgent	0.2509	0.1238	0.0319	0.4412	0.1715	0.1894	0.2574	0.1290	0.0184	0.4482	0.1768	0.1631
TextBugger	0.2620	0.1316	0.0306	0.4432	0.1770	0.1973	0.2591	0.1283	0.0381	0.4494	0.1760	0.1700
CheckList	0.2599	0.1295	0.0278	0.4371	0.1737	0.1734	0.2594	0.1306	0.0366	0.4398	0.1757	0.1797
RandAttack	0.2568	0.1281	0.0296	0.4428	0.1746	0.1731	0.2532	0.1261	0.0400	0.4499	0.1753	0.1679
PopAttack	0.2607	0.1282	0.0328	0.4388	0.1727	0.1673	0.2568	0.1329	0.0424	0.4414	0.1791	0.1513
EXPA	0.2600	0.1250	0.0326	0.4456	0.1713	0.1662	0.2654	0.1334	0.0442	0.4454	0.1784	0.1628
INSA	0.2477	0.1231	0.0231	0.4347	0.1698	0.1657	0.2631	0.1328	0.0350	0.4441	0.1780	0.1588
SEMA	0.2636	0.1319	0.0328	0.4504	0.1787	0.1664	0.2670	0.1331	0.0454	0.4512	0.1793	0.1638
VAA	0.2546	0.1262	0.0318	0.4356	0.1714	0.1579	0.2561	0.1284	0.0436	0.4446	0.1755	0.1569
IPDGI	0.2636	0.1339	0.0327	0.4447	0.1792	0.1593	0.2606	0.1309	0.0462	0.4484	0.1779	0.1567
UAP	0.2518	0.1256	0.0308	0.4426	0.1734	0.1748	0.2571	0.1300	0.0450	0.4471	0.1777	0.1766
AdvCLIP	0.2578	0.1267	0.0182	0.4423	0.1729	0.1852	0.2550	0.1231	0.0321	0.4446	0.1706	0.1896
C3H	0.2554	0.1268	**0.0782**	0.4366	0.1721	**0.2727**	0.2488	0.1263	**0.0643**	0.4390	0.1741	**0.2739**
	少样本（*Beauty*）						零样本（*Beauty*）					
	前十			前二十			前十			前二十		
无攻击	0.3530	0.2012	0.0000	0.4870	0.2350	0.0000	0.3210	0.1829	0.0000	0.4570	0.2172	0.0000
DeepWordBug	0.3368	0.1893	0.0024	0.4808	0.2255	0.0144	0.3075	0.1681	0.0091	0.4461	0.2030	0.0306
TextFooler	0.3433	0.1935	0.0027	0.4837	0.2287	0.0166	0.3083	0.1713	0.0083	0.4456	0.2059	0.0286
BertAttack	0.3379	0.1903	0.0029	0.4768	0.2253	0.0154	0.3041	0.1673	0.0079	0.4436	0.2024	0.0288
CheatAgent	0.3464	0.1949	0.0024	0.4872	0.2304	0.0091	0.3113	0.1715	0.0064	0.4515	0.2068	0.0159
TextBugger	0.3370	0.1901	0.0017	0.4754	0.2249	0.0115	0.3080	0.1657	0.0089	0.4419	0.1993	0.0247
CheckList	0.3431	0.1947	0.0024	0.4834	0.2300	0.0116	0.3097	0.1680	0.0051	0.4453	0.2021	0.0231
RandAttack	0.3436	0.1943	0.0027	0.4885	0.2308	0.0124	0.3189	0.1737	0.0042	0.4493	0.2065	0.0167
PopAttack	0.3441	0.1954	0.0026	0.4822	0.2302	0.0149	0.3027	0.1675	**0.0110**	0.4415	0.2024	0.0351
EXPA	0.3355	0.1891	0.0020	0.4808	0.2255	0.0136	0.3053	0.1662	0.0075	0.4417	0.2005	0.0277
INSA	0.3408	0.1934	0.0029	0.4836	0.2293	0.0153	0.3034	0.1650	0.0082	0.4444	0.2005	0.0303
SEMA	0.3442	0.1947	0.0020	0.4829	0.2297	0.0140	0.3021	0.1620	0.0082	0.4329	0.1949	0.0254
VAA	0.3404	0.1940	0.0024	0.4839	0.2300	0.0152	0.3078	0.1690	0.0071	0.4496	0.2047	0.0276
IPDGI	0.3446	0.1956	0.0024	0.4864	0.2313	0.0159	0.3103	0.1694	0.0063	0.4450	0.2033	0.0254
UAP	0.3473	0.1953	0.0023	0.4929	0.2320	0.0151	0.3103	0.1695	0.0078	0.4379	0.2016	0.0290
AdvCLIP	0.3424	0.1918	0.0027	0.4806	0.2266	0.0145	0.3034	0.1654	0.0076	0.4432	0.2006	0.0313
C3H	0.3415	0.1917	**0.0049**	0.4823	0.2271	**0.0290**	0.3043	0.1657	0.0081	0.4399	0.1998	**0.0623**

干净项目，CLIP-ViT-B/32图文余弦相似度的绝对变化）。**效用**：良性用户上的HR@K 和NDCG@K [46]。全文均使用 $K \in \{10, 20\}$ 。

实现细节。 我们在 PyTorch 1.13.1（CUDA 11.7）中实现所有模型，并通过 AdamW [47]对 VIP5 进行微调。对于 MicroLens、Beauty 和 Sports，VIP5 受害模型分别独立训练 20 个 epoch，学习率为 1×10^{-3}，有效批量大小为 128，预热比例为 0.1，梯度裁剪值为 5.0；每次微调运行在 $4\times$ 张 NVIDIA L20 GPU 上约需 10 小时。Qwen2-VL-2B 受害模型在相同硬件上使用少样本和零样本提示词进行评估。C3H 流水线具有确定性，在 4 张 GPU 上分布式运行时，每个目标约需 8 秒（单 GPU 等效时间为 ∼34s），且在攻击构造期间无需向受害模型发起查询。对于每个（目标、提示词）设置，我们固定一个评估用户子集，并在 NoAttack 和所有对比攻击方法之间共享：VIP5 设置下每个数据集包含 1,000 名良性用户；Qwen2-VL 设置则对用户进行子采样，以限制推理成本。

6.2 总体性能评估

攻击有效性。在 VIP5（表2）的 K=20 上，C3H 在四个（数据集、提示）单元格中均居首。在 MicroLens（N=9）上，C3H 将 ER@20 提升至 0。273（少样本）/ 0。274（零样本），比最强基线高出约 40%（TextBugger 少样本为 0。197；AdvCLIP 零样本为 0。190）。对于仅内容攻击而言，Beauty 更难

表3：**MicroLens (N=9)** 和 **Beauty (N=14)** 上的攻击隐蔽性。ROUGE 在 MicroLens 上使用字符级分词，在 Beauty 上使用词级分词。

	MicroLens(N = 9)				*Beauty*(N = 14)			
	单模态自然度 *ROUGE*			跨模态	单模态自然度 *ROUGE*			跨模态
攻击	**-1↑**	**-2↑**	**-L↑**	**SA Δ ↓**	**-1↑**	**-2↑**	**-L↑**	**SA Δ ↓**
无攻击	100.00	100.00	100.00	0.0000	100.00	100.00	100.00	0.0000
C3H	94.02	93.86	94.02	**0.0020**	82.22	80.25	82.22	**0.0135**

表 4：在 K=20 下对 **MicroLens (N=9)** 和 **Beauty (N=14)** 进行的留一消融实验。 Δ曝光率是相对于完整 C3H 的下降幅度。

	少样本（*MicroLens*）				零样本（*MicroLens*）				少样本（*Beauty*）				零样本（*Beauty*）			
变体	**HR**	**NDCG**	**ER**	**ΔER**	**HR**	**NDCG**	**ER**	**ΔER**	**HR**	**NDCG**	**ER**	**ΔER**	**HR**	**NDCG**	**ER**	**ΔER**
C3H（完整 C3H）	0.4366	0.1721	**0.2727**	–	0.4390	0.1741	**0.2739**	–	0.4823	0.2271	**0.0290**	–	0.4399	0.1998	**0.0623**	–
−M5（仅图像）	0.4430	0.1659	0.1623	−0.110	0.4488	0.1761	0.1668	−0.107	0.4834	0.2296	0.0202	−0.009	0.4500	0.2025	0.0362	−0.026
−M6（仅文本）	0.4430	0.1659	0.1824	−0.090	0.4461	0.1746	0.1814	−0.092	0.4747	0.2247	0.0131	−0.016	0.4456	0.2035	0.0148	−0.048
−C2C（随机英雄）	0.4423	0.1738	0.1732	−0.099	0.4472	0.1772	0.1761	−0.098	0.4875	0.2306	0.0196	−0.009	0.4494	0.2051	0.0371	−0.025
无攻击	0.4320	0.1657	0.0000	−0.273	0.4310	0.1679	0.0000	−0.274	0.4870	0.2350	0.0000	−0.029	0.4570	0.2172	0.0000	−0.062

（N=14），但 C3H 仍达到 0。029（少样本）/ 0。062（零样本），接近排名第二方法的两倍。MicroLens 少样本/零样本得分几乎相同（0。273 对比 0。274），这表明 C2C代理模型 捕捉的是结构性决策标准，而非特定于提示的特征。在 Beauty 上，零样本高于少样本（0。062 对比 0。029）：由于没有示例来锚定模型的排序，与标准一致的内容面临的竞争更少。

隐蔽性与实用性。 良性用户上的HR@K 和归一化折损累计增益@K 在表2的四个单元格中均接近无攻击。隐蔽性在两个维度上均得到保持（表3）：编辑后的标题与原始标题保持较高的ROUGE-L（MicroLens上为94。02，Beauty上为82。22），且SA Δ 保持在0。0020 / 0。0135。

6.3 消融研究

我们在 K=20 处进行留一消融（表 4），每次移除一个组件：(1) 文本算子（仅图像）；(2) 图像算子（仅文本）；以及 (3) C2C 联合评分，将其替换为从同一候选池中随机选择主角 $\mathcal{F}$。

主导模态随数据集而变化。在文本丰富的 MicroLens 图像说明中，移除文本算子所损失的曝光率大于移除图像算子（在少样本/零样本设置下分别为 $-0.110/-0.107$ 和 $-0.090/-0.092$ ）。Beauty 的简短标题则呈现相反顺序：移除图像算子会损失 $-0.016/-0.048$，而移除文本算子仅损失 $-0.009/-0.026$。更长的文本为文本编辑提供了更多可放大的内容；稀疏的标题则将负担转移到视觉通道。

联合评分优于随机选择主角。将 $\tilde{\kappa}\cdot\tilde{\chi}$ 替换为从同一候选池中随机选择的结果后，所有数据集上的曝光率均下降（MicroLens 上为 $-0.099/-0.098$；Beauty 上为 $-0.009/-0.025$ ），这与第 4 节的结论相呼应：候选池筛选会预先选出跨模态对齐线索，而联合评分则在此基础上进一步提供池内排序信号。在每个变体上，效用值都与完整 C3H 相差不超过 0.02。

6.4 防御鲁棒性

我们在推荐系统之前应用了三种内容过滤防御，并将每种防御在干净项目上的拒绝率校准为 5–10%：AE过滤器 [48, 49] （CLIP空间中的自编码器潜空间去噪）、MCM（最大概念匹配） [50], 和马氏距离分布外检测 [51]。这些防御均未显著降低 C3H（表 5）：在四个单元格中，无论采用哪种防御，ER 的变化都小于0。004，且良性用户的 HR@K / 归一化折损累计增益@K均得以保持。其原因在于结构特性：文本算子只会删除和重新排列 t_0 中已有的词元，而 $\mathcal{T}$ 中的图像变换也仅限于产品摄影中常见的裁剪和光照调整。经过编辑的内容位于 CLIP特征空间中的干净分布内（MicroLens 上的 SA $\Delta = 0.002$ ），因此基于离群点的过滤器无从发挥作用。

6.5 跨架构迁移性

我们将针对 VIP5 生成的攻击内容（$\tilde{t}_{i^*}$， $\tilde{v}_{i^*}$）原样应用于 Qwen2-VL- 2B-Instruct [43]，并分别使用少样本提示词和零样本提示。两个受害模型采用不同的视觉骨干网络（Qwen 的

表5：在少样本和零样本提示下，**MicroLens (N=9)**上的防御鲁棒性。每种防御均校准为对干净项目拒绝5–10%。

防御	少样本（*MicroLens*）						零样本（*MicroLens*）					
	前十			前二十			前十			前二十		
	HR	NDCG	ER	HR	NDCG	ER	HR	NDCG	ER	HR	NDCG	ER
无防御	0.2554	0.1268	0.0782	0.4366	0.1721	0.2727	0.2488	0.1263	0.0643	0.4390	0.1741	0.2739
AE过滤器 [48, 49]	0.2571	0.1285	0.0794	0.4426	0.1748	0.2732	0.2630	0.1319	0.0638	0.4453	0.1775	0.2718
MCM [50]	0.2604	0.1314	0.0814	0.4510	0.1793	0.2736	0.2650	0.1312	0.0628	0.4471	0.1768	0.2747
马氏距离分布外检测 [51]	0.2603	0.1301	0.0801	0.4479	0.1771	0.2714	0.2544	0.1278	0.0634	0.4420	0.1748	0.2709

原生视觉变换器与 CLIP-ViT-B/32）和语言骨干网络（1．5B Qwen-2语言模型与 T5-small）。完整表格见附录B；VIP5体育结果见附录A。

C3H 在 K=20 的 Qwen2-VL 上 6 个（数据集，提示）单元格中的 4 个排名第一，其中包括全部三个零样本单元格。Sports 是最明显的案例：在两种提示设置下，C3H 均以 3.7–4.7× 的优势胜过排名第二的方法。VIP5 中的零样本 > 少样本模式也再次出现：在 K=20，C3H 在 MicroLens 上达到 0.117（零样本），而少样本为 0.054；在 Beauty 上为 0.101，而少样本为 0.056；在 Sports 上为 0.187，而少样本为 0.113。该机制并不依赖于特定的受害模型。在少样本 MicroLens 和 Beauty 上，一些图像攻击基线略微领先于 C3H；我们认为，这是因为较大的像素扰动会在存在少样本示例时干扰 Qwen2-VL 的上下文推理，而这与 C2C偏差无关，且不会推广到 Sports（在 Sports 的每个少样本单元格中，相同的基线方法的曝光率均降至 0.03 以下）。

7 结论

我们在严格的黑盒、零查询、推理时、单卖家威胁模型下，对针对多模态大语言模型推荐系统的决策准则劫持进行了形式化研究。候选感知型多模态大语言模型推荐系统表现出压缩到共识（C2C）偏差：相同的压缩和跨模态融合机制在抑制传统单模态扰动的同时，也会赋予那些在压缩后存留并获得跨模态强化的线索不成比例的排序影响力。C3H 利用这一偏差，构建了一个无需查询、由公共编码器驱动的流水线，在隐蔽性集合 $\mathcal{B}$ 内编辑单个目标项目的文本和图像。

在 VIP5 上，C3H 在每个（数据集、提示）单元的 ER@20 指标上均超过十五种基线方法，提升幅度为 1.1–1.8×；相同内容未经修改即可迁移至 Qwen2-VL-2B-Instruct，在 MicroLens、Beauty 和 Sports 的所有零样本单元中均取得最佳结果，在 Sports 上相较于次优方法的优势最高达到 4.7×。三种基于异常检测的内容过滤防御均无法缓解 C3H（ΔER < 0.004），且良性用户效用（HR@K、归一化折损累计增益@K）得以保持。人们通常认为压缩和跨模态融合能够为多模态大语言模型推荐系统带来鲁棒性；我们的结果表明，它们也会在决策标准层面暴露一个无需受害模型反馈的攻击面。

8 局限性与潜在缓解措施

C3H 假设攻击者拥有文本和图像的编辑权限，其有效性可能因目标、受害者模型和平台限制而异。我们的评估仅涵盖受控基准测试中的有限目标和受害者模型；其在实际部署平台上的有效性仍有待验证。潜在的缓解措施包括内容规范化和监控商品信息编辑，但这些措施对抗 C3H 的有效性仍需进一步评估，而且它们可能会拒绝合法编辑或降低推荐效用。

致谢

本研究/项目由A*STAR通过其日新联合征集：JST-A*STAR 2024（项目编号：R2416IR139）提供支持。本研究还获得新加坡教育部学术研究基金第二级（资助编号 MOE-T2EP20125-0005）的支持。作者声明不存在任何利益冲突。

参考文献

[1] Peng Wang、An Yang、Rui Men、Junyang Lin、Shuai Bai、Zhikang Li、Jianxin Ma、Chang Zhou、Jingren Zhou和Hongxia Yang。OFA：通过简单的序列到序列学习框架统一架构、任务和模态。载于第*39*届国际机器学习会议论文集，第23318–23340页。PMLR，2022年。[2] OpenAI、Josh Achiam、Steven Adler、Sandhini Agarwal、Lama Ahmad、Ilge Akkaya、Florencia Leoni Aleman、Diogo Almeida、Janko Altenschmidt、Sam Altman、Shyamal Anadkat等。GPT-4技术报告。*arXiv*预印本 *arXiv:2303.08774*，2023年。[3] Shijie Geng、Juntao Tan、Shuchang Liu、Zuohui Fu和Yongfeng Zhang。VIP5：面向推荐的多模态基础模型。载于计算语言学协会论文集：*EMNLP 2023*研究成果，第9606–9620页，2023年。[4] Greg Linden、Brent Smith和Jeremy York。Amazon.com推荐：基于项目到项目的协同过滤。*IEEE Internet Computing*，7(1)：76–80，2003年。[5] Carlos A. Gomez-Uribe和Neil Hunt。Netflix推荐系统：算法、商业价值与创新。*ACM*管理信息系统汇刊，6(4)：1–19，2015年。[6] Paul Covington、Jay Adams和Emre Sargin。用于YouTube推荐的深度神经网络。载于第*10*届*ACM*推荐系统会议论文集，第191–198页，2016年。[7] Yurong Hao、Wen Zhou、Guowei Guan、Tiantong Wu、Fuyao Zhang和Wei Yang Bryan Lim。VirusCascade：劫持由大语言模型驱动的推荐代理中的协作式反思。*arXiv*预印本 *arXiv:2609.38270*，2026年。[8] Ramin Giahi、Kehui Yao、Sriram Kollipara、Kai Zhao、Vahid Mirjalili、Jianpeng Xu、Topojoy Biswas、Evren Korpeoglu和Kannan Achan。VL-CLIP：通过视觉定位和大语言模型增强的CLIP嵌入提升多模态推荐。载于第*19*届*ACM*推荐系统会议论文集，第482–491页，2025年。[9] Yuyang Ye、Zhi Zheng、Yishan Shen、Tianshu Wang、Hengruo Zhang、Peijun Zhu、Runlong Yu、Kai Zhang和Hui Xiong。利用多模态大语言模型进行多模态序列推荐。载于*AAAI*人工智能会议论文集，第39卷，第13069–13077页，2025年。[10] Fuyao Zhang、Jiaming Zhang、Che Wang、Boyang Chen、Yurong Hao、Xiongtao Sun、Guowei Guan、Blaise Delattre、Yang Cao和Wei Yang Bryan Lim。SynChain：诱导计算机使用代理系统构建自身的攻击链。*arXiv*预印本 *arXiv:2608.06862*，2026年。[11] Wenkuan Zhao、Shanshan Zhong、Yifan Liu、Wushao Wen、Jinghui Qin、Mingfu Liang和Zhongzhan Huang。DVIB：通过变分信息瓶颈蒸馏实现鲁棒多模态推荐系统。载于*ACM Web*会议*2025*论文集，第2549–2561页，2025年。[12] Guowei Guan、Yurong Hao、Jiaming Zhang、Tiantong Wu、Fuyao Zhang、Tianxiang Chen、Longtao Huang、Cyril Leung和Wei Yang Bryan Lim。VENOMREC：多模态大语言模型推荐系统中的跨模态交互式投毒与定向推广。载于第*43*届国际机器学习会议论文集，第306卷，第37417–37433页。PMLR，2026年。[13] Yuancheng Xu、Jiarui Yao、Manli Shu、Yanchao Sun、Zichu Wu、Ning Yu、Tom Goldstein和Furong Huang。Shadowcast：针对视觉语言模型的隐蔽数据投毒攻击。载于神经信息处理系统进展，第37卷，第57733–57764页，2024年。[14] Zhuoran Liu和Martha Larson。对抗性项目推广：使用图像解决冷启动问题的Top-N推荐系统核心漏洞。载于*2021*年*Web*会议论文集，第3590–3602页，2021年。[15] Yurong Hao、Xihui Chen、Xiaoting Lyu、Jiqiang Liu、Yongsheng Zhu、Zhiguo Wan、Sjouke Mauw和Wei Wang。不落一人：探索用户画像与项目在联邦推荐非定向攻击中的相互作用。载于*2024*年*ACM SIGSAC*计算机与通信安全会议论文集，第2889–2903页，2024年。[16] Yurong Hao、Xihui Chen、Wei Wang、Jiqiang Liu、Tao Li、Junyong Wang和Witold Pedrycz。聚焦联邦推荐：具有竞争性的定向投毒及其缓解。*IEEE*信息取证与安全汇刊，19：10173–10188，2024年。

[17] Yaqing Qian、基于大语言模型的文本排序中的认知偏见研究。载于2025年自然语言处理经验方法会议论文集，第21958–21968页，2025年。[18] Michael P. O’Mahony、Neil J. Hurley和Guénolé C. M. Silvestre。推荐系统：攻击类型与策略。载于第*20*339，2005年334届全国人工智能会议，第–页。[19] Wei Zhou、Junhao Wen、Min Gao、Ling Liu、Haini Cai和Xibin Wang。协同过滤推荐系统中一种基于支持向量机和目标项目分析的诋毁攻击检测方法。载于第*8*届知识科学、工程与管理国际会议论文集，第751–763页。Springer，2015年。[20] Di Jin、Zhijing Jin、Joey Tianyi Zhou和Peter Szolovits。BERT真的稳健吗？用于文本分类和蕴含的自然语言攻击的强基线。载于*AAAI*人工智能会议论文集，第34卷，第8018–8025页，2020年。[21] Linyang Li、Ruotian Ma、Qipeng Guo、Xiangyang Xue和Xipeng Qiu。BERT-ATTACK：使用BERT对BERT实施对抗攻击。载于*2020*年自然语言处理经验方法会议论文集，第6193–6202页，2020年。[22] Ji Gao、Jack Lanchantin、Mary Lou Soffa和Yanjun Qi。用于规避深度学习分类器的对抗性文本序列黑盒生成。载于*2018*年*IEEE*安全与隐私研讨会，第50–56页。IEEE，2018年。[23] Jinfeng Li、Shouling Ji、Tianyu Du、Bo Li和Ting Wang。TextBugger：针对现实世界应用生成对抗性文本。载于第*26*届网络与分布式系统安全年会论文集。互联网协会，2019年。[24] Marco Tulio Ribeiro、Tongshuang Wu、Carlos Guestrin和Sameer Singh。超越准确率：使用CheckList对自然语言处理模型进行行为测试。载于第*58*届计算语言学协会年会论文集，第4902–4912页，2020年。[25] Lijian Chen、Wei Yuan、Tong Chen、Guanhua Ye、Nguyen Quoc Viet Hung和Hongzhi Yin。通过引导扩散对视觉感知推荐系统实施对抗性项目推广。*ACM*信息系统汇刊，42(6)：1–26，2024年。[26] Seyed-Mohsen Moosavi-Dezfooli、Alhussein Fawzi、Omar Fawzi和Pascal Frossard。通用对抗性扰动。载于*IEEE*计算机视觉与模式识别会议论文集，第1765–1773页，2017年。[27] Ziqi Zhou、Shengshan Hu、Minghui Li、Hangtao Zhang、Yechao Zhang和Hai Jin。AdvCLIP：多模态对比学习中与下游任务无关的对抗样本。载于第*31*届*ACM*国际多媒体会议论文集，第6311–6320页，2023年。[28] Fuyao Zhang、Jiaming Zhang、Che Wang、Xiongtao Sun、Yurong Hao、Guowei Guan、Wenjie Li、Longtao Huang和Wei Yang Bryan Lim。DualTAP：面向移动多模态大语言模型智能体的双任务对抗保护器。载于第*19*届欧洲计算机视觉会议论文集，第323–340页。Springer，2026年。[29] Liang-bo Ning、Shijie Wang、Wenqi Fan、Qing Li、Xin Xu、Hao Chen和Feiran Huang。CheatAgent：通过大语言模型智能体攻击大语言模型赋能的推荐系统。载于第*30*届*ACM SIGKDD*知识发现与数据挖掘会议论文集，第2284–2295页，2024年。[30] Jinghao Zhang、Yuting Liu、Qiang Liu、Shu Wu、Guibing Guo和Liang Wang。针对基于大语言模型的推荐的隐蔽攻击。载于第*62*届计算语言学协会年会论文集（第*1*卷：长篇论文），第5839–5857页，2024年。[31] Nicolas Papernot、Patrick McDaniel、Ian Goodfellow、Somesh Jha、Z. Berkay Celik和Ananthram Swami。针对机器学习的实用黑盒攻击。载于*2017*年*ACM*亚洲计算机与通信安全会议论文集，第506–519页，2017年。[32] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark等。通过自然语言监督学习可迁移视觉模型。载于第*38*届国际机器学习会议论文集，第8748–8763页。PMLR，2021年。[33] Alessandro Achille和Stefano Soatto。深度表征中不变性与解耦的涌现。《机器学习研究杂志》，19(50)：1–34，2018年。

[34] Rada Mihalcea和Paul Tarau。TextRank：为文本建立秩序。载于*2004*年自然语言处理经验方法会议论文集，第404–411页，2004年。[35] Naftali Tishby、Fernando C. Pereira和William Bialek。信息瓶颈方法。载于第*37*届*Allerton*通信、控制与计算年会论文集，第368–377页。伊利诺伊大学，1999年。[36] Samira Abnar和Willem Zuidema。量化Transformer模型中的注意力流。载于第*58*届计算语言学协会年会论文集，第4190–4197页，2020年。[37] Yihe Dong、Jean-Baptiste Cordonnier和Andreas Loukas。注意力并非你所需的一切：纯注意力随深度增加以双指数速度损失秩。载于第*38*届国际机器学习会议论文集，第2793–2803页。PMLR，2021年。[38] Ramprasaath R. Selvaraju、Michael Cogswell、Abhishek Das、Ramakrishna Vedantam、Devi Parikh和Dhruv Batra。Grad-CAM：通过基于梯度的定位从深度网络生成可视化解释。载于*IEEE*国际计算机视觉会议论文集，第618–626页，2017年。[39] Yongxin Ni、Yu Cheng、Xiangyan Liu、Junchen Fu、Youhua Li、Xiangnan He、Yongfeng Zhang和Fajie Yuan。大规模内容驱动的微视频推荐数据集。载于第*34*届*ACM*国际信息与知识管理会议论文集，第6486–6491页，2025年。[40] Julian McAuley和Jure Leskovec。隐因子与隐主题：利用评论文本理解评分维度。载于第7172，2013年165届*ACM*推荐系统会议，第–页。[41] Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li和Peter J. Liu。探索统一文本到文本Transformer的迁移学习极限。《机器学习研究杂志》，21(140)：1–67，2020年。[42] Yi-Lin Sung、Jaemin Cho和Mohit Bansal。VL-Adapter：面向视觉语言任务的参数高效迁移学习。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第5227–5237页，2022年。[43] Peng Wang、Shuai Bai、Sinan Tan、Shijie Wang、Zhihao Fan、Jinze Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge等。Qwen2-VL：增强视觉语言模型在任意分辨率下对世界的感知。*arXiv*预印本 *arXiv:2409.12191*，2024年。[44] Rami Cohen、Oren Sar Shalom、Dietmar Jannach和Amihood Amir。面向视觉感知推荐系统的黑盒攻击模型。载于第*14*届*ACM*国际网络搜索与数据挖掘会议论文集，第94–102页，2021年。[45] Chin-Yew Lin。ROUGE：用于自动评估摘要的软件包。载于*Text Summarization BranchesOut*，第74–81页，2004年。[46] Kalervo Järvelin和Jaana Kekäläinen。基于累积增益的信息检索技术评估。*ACM*信息系统汇刊，20(4)：422–446，2002年。[47] Ilya Loshchilov和Frank Hutter。解耦权重衰减正则化。载于第*7*届国际学习表征会议论文集，2019年。[48] Mayu Sakurada和Takehisa Yairi。使用非线性降维和自编码器进行异常检测。载于*MLSDA 2014*年第二届传感数据分析机器学习研讨会论文集，第4–11页，2014年。[49] Celia Cintas、Skyler Speakman、Victor Akinwande、William Ogallo、Komminist Weldemariam、Srihari Sridharan和Edward McFowland。通过对自编码器激活和重构误差进行子集扫描检测对抗攻击。载于第*29*届国际人工智能联合会议论文集，第876–882页。国际人工智能联合会议组织，2020年。[50] Yifei Ming、Ziyang Cai、Jiuxiang Gu、Yiyou Sun、Wei Li和Yixuan Li。深入研究利用视觉语言表征进行分布外检测。载于神经信息处理系统进展，第35卷，第35087–35102页，2022年。[51] Kimin Lee、Kibok Lee、Honglak Lee和Jinwoo Shin。用于检测分布外样本和对抗攻击的简单统一框架。载于神经信息处理系统进展，第31卷，第7167–7177页，2018年。

附录

A Sports 数据集上的结果

此外，为完整起见，我们报告了 VIP5 在 Sports 数据集上的结果（N=6 个可攻击目标）。表 6 显示，在 K=20 时，C3H 在两种提示下仍是最强的攻击方法（少样本：0.028，而最佳基线为 0.021；零样本：0.106，而最佳基线为 0.096），尽管由于可攻击子集较小，其绝对差距小于在 MicroLens 和 Beauty 上的差距。

表 6：在 K=20 下，使用少样本和零样本提示时 Sports **Sports (N=6)** 上的主要结果。

	少样本（*Sports*）			零样本（*Sports*）		
攻击	**HR**	**NDCG**	**ER**	**HR**	**NDCG**	**ER**
无攻击	0.3240	0.1666	0.0000	0.3210	0.1526	0.0000
DeepWordBug	0.3397	0.1590	0.0095	0.3373	0.1542	0.0755
TextFooler	0.3470	0.1627	0.0117	0.3305	0.1515	0.0937
BertAttack	0.3418	0.1605	0.0212	0.3290	0.1490	0.0808
CheatAgent	0.3385	0.1600	0.0073	0.3420	0.1534	0.0338
TextBugger	0.3385	0.1572	0.0070	0.3315	0.1541	0.0728
CheckList	0.3482	0.1674	0.0153	0.3338	0.1569	0.0692
RandAttack	0.3435	0.1594	0.0047	0.3293	0.1547	0.0728
PopAttack	0.3387	0.1569	0.0083	0.3282	0.1486	0.0745
EXPA	0.3522	0.1615	0.0177	0.3428	0.1560	0.0928
INSA	0.3420	0.1576	0.0132	0.3342	0.1527	0.0838
SEMA	0.3463	0.1606	0.0182	0.3348	0.1541	0.0962
VAA	0.3328	0.1539	0.0065	0.3463	0.1579	0.0618
IPDGI	0.3447	0.1604	0.0133	0.3425	0.1559	0.0745
UAP	0.3398	0.1605	0.0048	0.3400	0.1568	0.0508
AdvCLIP	0.3422	0.1572	0.0100	0.3270	0.1487	0.0702
C3H	0.3395	0.1598	**0.0282**	0.3380	0.1549	**0.1058**

B 跨架构迁移性：结果

C3H 在 Qwen2-VL-2B-Instruct 上的跨架构迁移性总结于 §6.5，并在表 7（MicroLens，N=9；Beauty，N=14）和表 8（Sports，N=6）中报告，涵盖少样本和零样本提示词。

表 7：在 K=20 下，**Qwen2-VL-2B-Instruct** [43] 在 **MicroLens (N=9)** 和 **Beauty (N=14)** 上的跨架构迁移性。

攻击	**MicroLens** (N=9)						**Beauty** (N=14)					
	少样本			零样本			少样本			零样本		
	HR	**NDCG**	**ER**	**HR**	**NDCG**	**ER**	**HR**	**NDCG**	**ER**	**HR**	**NDCG**	**ER**
无攻击	0.3144	0.1033	0.0000	0.2534	0.0809	0.0000	0.3583	0.1563	0.0000	0.3908	0.1685	0.0000
DeepWordBug	0.3229	0.1073	0.0781	0.2398	0.0828	0.0272	0.3610	0.1491	0.0475	0.4032	0.1776	0.0194
TextFooler	0.3081	0.0978	0.0514	0.2423	0.0749	0.0223	0.3410	0.1417	0.0033	0.4149	0.1772	0.0035
BertAttack	0.3306	0.1053	0.0376	0.2742	0.0950	0.0055	0.3295	0.1361	0.0300	0.4124	0.1791	0.0053
CheatAgent	0.3173	0.1043	0.0533	0.2472	0.0822	0.0306	0.3558	0.1584	0.0066	0.4021	0.1830	0.0053
TextBugger	0.3193	0.1075	0.0765	0.2287	0.0819	0.0055	0.3378	0.1422	0.0532	0.3936	0.1619	0.0160
CheckList	0.3333	0.1060	0.0355	0.2316	0.0779	0.0327	0.3523	0.1518	0.0417	0.3929	0.1649	0.0336
RandAttack	0.3306	0.1036	0.0190	0.2310	0.0789	0.0027	0.3250	0.1461	0.0536	0.3838	0.1683	0.0018
PopAttack	0.3253	0.1057	0.0269	0.2830	0.0874	0.0220	0.3673	0.1582	0.0000	0.4142	0.1762	0.0017
EXPA	0.2992	0.0970	0.0943	0.2646	0.0905	0.0056	0.3289	0.1488	0.0548	0.4139	0.1685	0.0226
INSA	0.3152	0.1035	0.0870	0.2472	0.0799	0.0083	0.3356	0.1430	0.0688	0.4152	0.1732	0.0247
SEMA	0.2861	0.0960	**0.0989**	0.2231	0.0735	0.0028	0.3384	0.1429	0.0643	0.4004	0.1735	0.0212
VAA	0.3431	0.1122	0.0878	0.2630	0.0831	0.0055	0.3267	0.1424	0.0783	0.3878	0.1724	0.0174
IPDGI	0.3173	0.1000	0.0907	0.2707	0.0891	0.0083	0.3261	0.1318	0.0765	0.3748	0.1597	0.0294
UAP	0.3154	0.1021	0.0755	0.2472	0.0847	0.0000	0.3570	0.1480	**0.0959**	0.3537	0.1628	0.0470
AdvCLIP	0.3592	0.1156	0.0858	0.2598	0.0865	0.0000	0.3710	0.1666	0.0682	0.4207	0.1788	0.0414
C3H	0.3029	0.0959	0.0536	0.2542	0.0833	**0.1173**	0.3237	0.1420	0.0559	0.4108	0.1738	**0.1014**

表 8：在K=20下，**Qwen2-VL-2B-Instruct** [43] 于**Sports (**N=6**)**上的跨架构迁移性。

	少样本（*Sports*）			零样本（*Sports*）		
攻击	**HR**	**NDCG**	**ER**	**HR**	**NDCG**	**ER**
无攻击	0.3434	0.1517	0.0000	0.2877	0.1307	0.0000
DeepWordBug	0.3352	0.1481	0.0166	0.3271	0.1489	0.0141
TextFooler	0.3481	0.1576	0.0055	0.2941	0.1358	0.0118
BertAttack	0.3111	0.1536	0.0000	0.3129	0.1393	0.0000
CheatAgent	0.3225	0.1443	0.0000	0.3002	0.1412	0.0000
TextBugger	0.3011	0.1272	0.0304	0.3170	0.1371	0.0326
CheckList	0.2869	0.1266	0.0301	0.3333	0.1481	0.0399
RandAttack	0.2779	0.1180	0.0027	0.3155	0.1432	0.0093
PopAttack	0.3171	0.1325	0.0054	0.3121	0.1430	0.0000
EXPA	0.2877	0.1322	0.0247	0.3192	0.1380	0.0000
INSA	0.2992	0.1234	0.0216	0.3357	0.1496	0.0188
SEMA	0.3306	0.1506	0.0246	0.3019	0.1352	0.0024
VAA	0.3562	0.1543	0.0274	0.3341	0.1519	0.0000
IPDGI	0.2981	0.1225	0.0163	0.3341	0.1484	0.0141
UAP	0.3115	0.1245	0.0273	0.3095	0.1437	0.0000
AdvCLIP	0.3062	0.1291	0.0108	0.3066	0.1356	0.0000
C3H	0.3113	0.1430	**0.1129**	0.2937	0.1261	**0.1865**

我们实验中使用的超参数。对于所有报告的结果，我们使用 $\tau_{\text{cat}} = 0.20$, τ_{orig} =0.18, $\tau_{\text{align}} = 0.18$（B/32 余弦阈值）、 $L_{\text{text}} \in \{20, 40, 80\}$ 的 TextRank 上下文窗口长度、$\lceil P/4 \rceil$ 个用于 $V(f)$ 的顶部图像块（在注意力展开下）、 $\{1.0, 0.5, 0.2\}$的标题/正文/隐式位置权重（在 A_{text} 中），以及用于主图/次图选择的 $\Delta_{\text{tie}} = 0.05$ 平局判定阈值。图像变换参数： θ_1, θ_2 在 CLIP-ViT-B/32 上使用 Grad-CAM，并将裁剪边界框围绕线索显著性中心扩展 20%； θ_3 将每一边中心裁剪至 $0.7\times$ ； θ_4 在主图区域内应用对比度因子 1.15 和亮度偏移 $+0.05$ （Grad-CAM 掩码阈值设为第 50 百分位）。小图像回退：当图像边长为 < 200 px 时使用 $\mathcal{T} = \{\theta_0, \theta_4\}$ ，当为 < 100px 时使用 $\mathcal{T} = \{\theta_0\}$。

C 计算报告

我们报告了用于构建和评估 C3H 的计算资源。论文报告了三个数据集（MICROLENS、亚马逊 BEAUTY 和亚马逊 SPORTS）上的结果；以下计算量估算还包括超出这些报告结果的额外探索性实验。我们使用参数高效适配器对 VIP5 受害模型进行微调，仅使用公共编码器为每个（数据集、目标）对生成 C3H 内容，并在良性用户池上评估有效性、实用性和隐蔽性。不从头训练模型，也不使用外部 API 服务；所有推理均在本地执行。

本地计算环境。 所有实验均在一台 Ubuntu 20.04 工作站上运行，该工作站配备8×块 NVIDIA L20 48 GB GPU（CUDA 11.7）、Python 3.9 和 PyTorch 1.13.1+cu117。源代码、检查点、评估日志以及预提取的 CLIP 特征约占用 90 GB 的磁盘存储空间。

受害模型微调。 我们通过 VL-Adapter [42]（缩减因子为 8）对 VIP5 受害模型进行微调：每个数据集训练 20 个 epoch，使用 AdamW，学习率为 1×10^{-3}，预热比例为 0.1，梯度裁剪值为 5.0，有效批量大小为 128（4×32 使用四块 L20 GPU），并固定随机种子为 2022年。每个数据集的微调大约需要 9.5–10.5 个实际小时；在论文中报告的三个数据集（MICROLENS、亚马逊 BEAUTY 和亚马逊 SPORTS）以及额外的探索性微调中，受害模型微调总计消耗约**200** GPU 小时。

攻击生成。 每个 C3H 工件针对每个（数据集、目标）组合生成一次，仅使用公开的 CLIP-ViT-B/16（用于$\hat{\kappa}$ 的图块级编码器）和 CLIP-ViT-B/32（用于 $\hat{}\chi$ 的全局级编码器）。生成过程是确定性的，不需要查询受害模型，在单个 L20 GPU 上为每个目标耗时约 30–40 秒。综合所有内容修改基线方法和 C3H 消融实验变体，攻击生成总计约消耗 **30**GPU 小时。

评估计算。 有效性和实用性评测指标通过 VIP5 的逐点评估器计算（B-3 少样本和 B-Z 零样本提示，包含 99 个候选物品），每个（目标、攻击、提示）单元使用 1,000 名良性用户。每个单元大约需要 3–4 分钟的 L20 GPU 时间。报告中的主要结果、消融实验、防御评估和机制分析，以及其他探索性评估，共包含 6,213 次逐点评估运行，总计约消耗 **345** GPU 小时。对于跨架构迁移性，我们在冻结参数的 Qwen2-VL-2B-Instruct [43] 上进行评估（bfloat16 推理，不进行微调，也不使用 LoRA），同时采用与 VIP5 协议一致的少样本 （两个上下文示例）和零样本 提示方案。六个评估作业（3 个数据集 $\times$ 2提示方案）在六个 L20 GPU 上并行运行，每项耗时 13–32 分钟，并总计贡献约**3**GPU 小时。

总计算量、重试与说明。综合各阶段，完整的 C3H 研究大约消耗 **580** 英伟达L20 GPU 小时；该数值还包括论文所报告的三个数据集之外的额外探索性实验。评估脚本会针对瞬时 CUDA 显存溢出事件自动重试（观察到的重试率低于 1%）。少量攻击生成任务会在 M7 隐蔽性门槛处中止，并按照第 5.3 节的规定返回干净基线（t_0， v_0）；这些是预先设计的回退机制，并非失败。没有模型是从头开始训练的。

D 更广泛的影响与负责任使用

本研究旨在通过识别基于内容的排序中的漏洞，提高多模态推荐系统的安全性。研究结果可能有助于研究人员和平台运营者开发更稳健的推荐机制。然而，C3H具有双重用途：恶意卖家可能滥用类似技术操纵商品曝光度，从而误导用户并使其他卖家处于不利地位。我们的实验使用公开的基准数据集和研究模型；我们不会攻击已部署的商业平台。监控商品信息编辑并进行内容规范化是潜在的缓解措施，但其针对C3H的有效性仍有待评估。我们鼓励将研究产物用于受控的安全评估，并且不会再分发经过修改的基准媒体。

NeurIPS 论文清单

1. **主张**

   问题：摘要和引言中提出的主要论点是否准确反映了论文的贡献和范围?

   回答： [是]

   理由：摘要和 §1 陈述了 C2C偏差假设、严格的零查询单卖家威胁模型以及主要实证性主张；§6 中的实验与这些主张的范围和力度相符。

   指南：

   - 该回答 [不适用] 表示摘要和引言未包含论文中的主张。
   - 摘要和/或引言应清楚陈述所提出的主张，包括论文的贡献以及重要的假设和局限性。 [否]或[不适用] 对这个问题作出否定回答会给审稿人留下不好的印象。
   - 所提出的主张应与理论和实验结果相符，并反映这些结果在多大程度上有望推广到其他情境。
   - 可以将愿景目标作为动机，只要明确说明论文并未实现这些目标。

2. **局限性**

   问题：论文是否讨论了作者所开展工作的局限性?

   回答： [是]

   理由：第8节讨论了攻击的假设、评估范围，以及候选缓解措施及其权衡。

   指南：

   - 答案 [N/A] 表示论文没有局限性，而答案 [No] 表示论文存在局限性，但论文并未对此进行讨论。
   - 鼓励作者在论文中单独设置“局限性”一节。
   - 论文应指出任何较强的假设，并说明结果对于违反这些假设的情况具有多强的稳健性（例如，独立性假设、无噪声情形、模型设定正确，以及渐近近似仅在局部成立）。作者应思考这些假设在实践中可能如何被违反，以及这会带来什么影响。
   - 作者应反思所作主张的范围，例如，该方法是否仅在少数数据集或通过少数几次运行进行了测试。一般而言，实证结果往往依赖于隐含假设，而这些假设应予以明确说明。
   - 作者应反思影响该方法性能的因素。例如，当图像分辨率较低或图像拍摄于低照度环境中时，人脸识别算法可能表现不佳。又或者，由于无法处理专业术语，语音转文本系统可能无法可靠地用于为在线讲座提供隐藏式字幕。
   - 作者应讨论所提出算法的计算效率，以及它们如何随数据集规模的增长而扩展。
   - 如适用，作者应讨论其方法在解决隐私和公平性问题方面可能存在的局限性。
   - 尽管作者可能担心，完全坦诚地说明局限性会被审稿人用作拒稿依据，但更糟糕的结果可能是，审稿人发现论文中未被承认的局限性。作者应自行审慎判断，并认识到，为促进透明度而采取的个人行动，对于形成维护社群诚信的规范发挥着重要作用。审稿人将获特别指示，不得因作者诚实说明局限性而予以惩罚。

3. **理论假设与证明**

   问题：对于每个理论结果，论文是否提供了完整的假设集合以及完整且正确的证明?

回答： [不适用]

理由：论文不包含形式理论结果；第 5 节引用的信息瓶颈论证是作为动机，而非经证明的贡献。

指南：

- 答案 [不适用]表示论文不包含理论结果。
- 论文中的所有定理、公式和证明都应编号并相互引用。
- 所有假设都应明确说明，或在任何定理的陈述中注明出处。
- 证明可以出现在论文正文或补充材料中；但如果放在补充材料中，建议作者提供简短的证明概要，以帮助读者理解其直觉。
- 相反，论文正文核心部分提供的任何非正式证明，都应由附录或补充材料中提供的形式化证明加以补充。
- 证明所依据的定理和引理应得到适当引用。

4. **实验结果可复现性**

问题：论文是否完整披露了复现论文主要实验结果所需的全部信息，以达到足以影响论文主要论点和/或结论的程度（无论是否提供代码和数据）？

回答： [是]

理由：§6.1 especifica 数据集、受害模型、基线方法和评测指标；完整的流程伪代码见算法 1；附录B详细说明了超参数和小图像回退方案；附录C报告了计算环境。

指南：

- 答案 [不适用]表示论文不包含实验。
- 如果论文包含实验，审稿人不会认可 [否] 这一回答：无论是否提供代码和数据，使论文具备可复现性都很重要。
- 如果贡献是数据集和/或模型，作者应描述为使其结果可复现或可验证而采取的步骤。
- 根据贡献的不同，可以通过多种方式实现可复现性。例如，如果贡献是一种新型架构，完整描述该架构可能就足够了；如果贡献是特定模型及其经验评估，则可能需要让他人能够使用相同的数据集复现该模型，或提供模型访问权限。一般而言，发布代码和数据通常是实现这一目标的一种好方法，但也可以通过详细说明如何复现结果、提供托管模型的访问权限（例如在大语言模型的情况下）、发布模型检查点，或其他适合所开展研究的方式来实现可复现性。
- 虽然 NeurIPS 不要求发布代码，但该会议确实要求所有投稿提供某种合理的可复现性途径，这可能取决于贡献的性质。例如

  (a) 如果贡献主要是新算法，论文应明确说明如何复现该算法。(b) 如果贡献主要是新的模型架构，论文应清晰、完整地描述该架构。(c) 如果贡献是新模型（例如大语言模型），那么应提供获取该模型的方式以复现实验结果，或提供复现该模型的方法（例如使用开源数据集，或说明如何构建数据集）。(d) 我们认识到，在某些情况下，可复现性可能较难实现；在这种情况下，作者可以说明其为确保可复现性所采取的具体方式。对于闭源模型，模型访问权限可能会受到某种限制（例如仅限注册用户），但其他研究人员应能够通过某种途径复现或验证实验结果。

5. **开放获取数据和代码**

问题：论文是否开放获取数据和代码，并提供充分的说明，以便按照补充材料中的描述，忠实地复现主要实验结果?

回答： [是]

理由：源代码已公开发布于 https://github.com/ GuoweiGuan666/C3H。

指南：

- 该回答 [不适用] 表示论文不包含需要代码的实验。
- 有关更多详情，请参阅 NeurIPS 代码与数据提交指南（https://neurips.cc/public/guides/CodeSubmissionPolicy）。
- 虽然我们鼓励发布代码和数据，但我们理解这可能无法实现，因此 [否] 是可以接受的答案。论文不会仅仅因为未包含代码而被拒稿，除非代码是论文贡献的核心部分（例如，构建新的开源基准）。
- 说明中应包含用于运行实验以复现实验结果所需的确切命令和环境。有关更多详细信息，请参阅 NeurIPS 代码与数据提交指南（https: //neurips.cc/public/guides/CodeSubmissionPolicy）。
- 作者应提供数据访问和准备方面的说明，包括如何访问原始数据、预处理数据、中间数据和生成数据等。
- 作者应提供脚本，以复现所提出的新方法和基线方法的全部实验结果。如果只能复现部分实验，则应说明脚本中省略了哪些实验以及原因。
- 在提交时，为了保持匿名性，作者应发布匿名化版本（如适用）。
- 建议在补充材料（附于论文之后）中尽可能提供详细信息，但也允许提供数据和代码的网址。

6. **实验设置/细节**

问题：论文是否说明了理解结果所需的所有训练和测试细节（例如，数据划分、超参数、参数的选择方式以及优化器类型）？

回答： [是]

理由：第6.1节和附录C报告了训练和评估细节（数据划分、优化器、学习率、批量大小、随机种子、提示方案和用户子集采样）。

指南：

- 答案 [不适用]表示论文未包含实验。
- 实验设置应在论文主体中以足够详细的程度呈现，以便读者理解结果并正确解读其含义。
- 完整细节可以通过代码、附录或补充材料提供。

7. **实验的统计显著性**

问题：论文是否以适当且定义正确的方式报告误差条，或提供关于实验统计显著性的其他适当信息?

答案： [否]

理由：最终稿件在主要表格和消融表中报告了 ER@K、HR@K 和 NDCG@K结果，但未报告误差条、置信区间或正式的显著性检验。评估用户采样在§6.1中进行了说明。所报告的差异被解读为在所评估设置下的描述性比较。

指南：

- 该答案 [不适用]意味着论文不包含实验。
- 作者应回答 [是] 结果是否附有误差条、置信区间或统计显著性检验，至少应涵盖支持论文主要论点的实验。

- 应明确说明误差条所反映的变异因素（例如训练集/测试集划分、初始化、对某个参数进行随机抽样，或在给定实验条件下进行的整体运行）。
- 应说明误差条的计算方法（闭式公式、调用库函数、自助法等）。
- 应说明所作的假设（例如，误差服从正态分布）。
- 应明确误差条表示的是标准差还是均值标准误。
- 报告 1σ误差条是可以的，但应明确说明。若误差正态性的假设未经验证，作者最好报告 2σ误差条，而不是声称置信区间为 96%。
- 对于非对称分布，作者应谨慎处理，不要在表格或图表中展示对称误差条，以免得出超出范围的结果（例如，负错误率）。
- 如果在表格或图表中报告误差条，作者应在文本中说明其计算方法，并在文本中引用相应的图表或表格。

8. **实验计算资源**

   问题：对于每项实验，论文是否提供了足够的计算机资源信息（计算工作节点类型、内存、执行时间），以便复现实验?

   答案： [是]

   理由：附录C报告了硬件（NVIDIA L20 48 GB）、各阶段的墙钟时间分解以及总计算量（约580个GPU小时）。

   指南：
   - 答案 [不适用]表示论文未包含实验。
   - 论文应说明所使用的计算工作节点类型（CPU 或 GPU）、内部集群或云服务提供商，并列出相关的内存和存储。
   - 论文应提供每次单独实验运行所需的计算量，并估算总计算量。
   - 论文应披露完整研究项目所需的计算量是否超过论文中报告的实验所需的计算量（例如，未纳入论文的初步实验或失败实验）。

9. **道德准则**

   问题：论文中开展的研究是否在各个方面都符合 NeurIPS 道德准则 https://neurips.cc/public/EthicsGuidelines?

   回答： [是]

   理由：我们已审阅 NeurIPS 道德准则，且这项工作在各方面均符合该准则。

   指南：
   - 回答 [不适用] 表示作者尚未审阅 NeurIPS 道德准则。
   - 如果作者回答 [否], 他们应说明偏离道德准则的特殊情况。
   - 作者应确保匿名性（例如，如果由于其司法管辖区的法律或法规而需要特别考虑）。

10. **更广泛影响**

    问题：论文是否讨论了所开展工作可能带来的积极社会影响和负面社会影响?

    回答： [是]

    理由：附录D讨论了C3H潜在的安全效益、双重用途风险及更广泛的社会影响，同时说明了我们的评估范围和可能的缓解措施。

指南：

- 该答案 [不适用] 表示所开展的工作不会产生社会影响。
- 如果作者回答 [不适用] 或 [否], ，则应说明为何其工作没有社会影响，或为何论文未讨论社会影响。
- 负面社会影响的示例包括潜在的恶意或意外用途（例如，虚假信息、生成虚假档案、监控）、公平性考量（例如，部署可能导致技术决策对特定群体不公平的技术）、隐私考量和安全考量。
- 会议预计，许多论文将属于基础研究，且不针对特定应用，更不用说实际部署了。不过，如果存在通往任何负面应用的直接路径，作者应予以指出。例如，指出生成模型质量的提升可能被用于生成深度伪造内容以传播虚假信息，是合理的。另一方面，无需指出，用于优化神经网络的通用算法可能使人们能够训练出更快生成深度伪造内容的模型。
- 作者应考虑以下可能造成的危害：技术按预期使用且运行正常时可能产生的危害；技术按预期使用但给出错误结果时可能产生的危害；以及因（有意或无意）滥用技术而造成的危害。
- 如果存在负面的社会影响，作者还可以讨论可能的缓解策略（例如，受控发布模型、除攻击之外同时提供防御措施、监控滥用的机制、监控系统如何随着时间推移从反馈中学习的机制、提高机器学习的效率和可访问性）。

11. **安全保障措施**

问题：论文是否描述了为负责任地发布具有高滥用风险的数据或模型而采取的安全保障措施（例如，预训练语言模型、图像生成器或抓取数据集）？

回答： [不适用]

理由：我们不发布预训练攻击模型或经修改的基准数据集。附录D讨论了C3H及其研究产物的潜在双重用途风险。

指南：

- 答案 [不适用] 表示该论文不存在此类风险。
- 具有较高滥用或双重用途风险的已发布模型，应在发布时配备必要的安全保障措施，以便对模型进行受控使用，例如要求用户遵守使用指南，或限制模型访问权限，或实施安全过滤器。
- 从互联网抓取的数据集可能带来安全风险。作者应说明其如何避免发布不安全图像。
- 我们认识到，提供有效的安全保障措施具有挑战性，许多论文也并不要求这样做，但我们鼓励作者对此予以考虑，并尽最大努力予以落实。

12. **现有资产的许可证**

问题：论文中使用的资产（例如代码、数据、模型）的创作者或原始所有者是否得到了适当署名，并且许可证和使用条款是否被明确提及且得到适当遵守?

回答： [是]

理由：数据集和基础模型已在第6.1节中引用。VIP5和CLIP代码仓库采用MIT许可证；T5和Qwen2-VL-2B-Instruct采用Apache-2.0许可证。MicroLens自述文件和亚马逊评论数据页面要求注明引用，但未指定名称明确的数据集许可证。MicroLens自述文件限制对修改后数据的再分发；其下载门户还额外规定，视频仅限学术用途，并存在修改/分发限制。我们不会再分发原始或编辑后的MicroLens文本、图像或视频；我们的制品发布仅限于代码、配置和数值结果。

指南：

- 答案 [不适用]表示该论文未使用现有资产。
- 作者应引用生成该代码包或数据集的原始论文。
- 作者应说明所使用资产的版本，并在可能的情况下提供网址。

- 应为每项资产注明许可证名称（例如 CC-BY 4.0）。
- 对于从特定来源（例如网站）抓取的数据，应提供该来源的版权和服务条款。

- 如果发布资产，应提供代码包中的许可证、版权信息和使用条款。对于热门数据集，paperswithcode.com/datasets 为部分数据集整理了许可证。其许可指南有助于确定数据集的许可证。
- 对于重新打包的现有数据集，应同时提供原始许可证和衍生资产的许可证（如果许可证已发生变化）。
- 如果无法在线获取这些信息，建议作者联系资产创建者。

13. **新资产**

    问题：论文中引入的新资产是否有完善的文档记录，并且是否随资产一同提供了文档?

    回答： [是]

    理由：论文介绍了 C3H，并提供了一个包含源代码和文档的公共代码仓库。论文正文和附录记录了攻击流程、实验设置、超参数和计算资源，而代码仓库则提供了复现该工作流程的说明。

    指南：

    - 该答案 [不适用] 表示论文未发布新资产。
    - 研究人员应通过结构化模板，在提交材料中说明数据集、代码和模型的详细信息。其中包括有关训练、许可证、局限性等方面的细节。
    - 论文应讨论是否已获得资产所涉人员的知情同意，以及如何获得知情同意。
    - 提交时，请记得将资产匿名化（如适用）。您可以创建匿名化网址，或附上匿名化 ZIP 文件。

14. **众包和涉及人类受试者的研究**

    问题：对于众包实验和涉及人类受试者的研究，论文是否包含向参与者提供的完整说明文本，以及（如适用）截图和有关报酬（如有）的详细信息?

    答案： [不适用]

    理由：该论文不涉及众包或涉及人类受试者的研究；良性用户数据取自现有的公开推荐基准数据集。

    指南：

    - 答案 [不适用] 表示该论文不涉及众包或涉及人类受试者的研究。
    - 将此信息纳入补充材料是可以的，但如果论文的主要贡献涉及人类受试者，则应尽可能在正文中提供详细信息。
    - 根据 NeurIPS《伦理准则》，参与数据收集、整理或其他劳动的工作人员，至少应获得数据收集者所在国家的最低工资。

15. **涉及人类受试者的研究是否获得机构审查委员会批准（IRB）或等效批准**

问题：论文是否说明了研究参与者可能面临的风险、是否向受试者披露了这些风险，以及是否获得了机构审查委员会（IRB）的批准（或根据贵国或机构的要求获得等效批准/审查）？

回答： [不适用]

理由：论文不涉及人类受试者研究。

指南：

- 答案 [不适用] 表示论文不涉及众包，也不涉及人类受试者的研究。
- 根据开展研究所在国家的规定，任何涉及人类受试者的研究都可能需要获得 IRB 批准（或同等机构的批准）。如果您已获得 IRB 批准，应在论文中明确说明。
- 我们认识到，不同机构和地区在相关程序方面可能存在显著差异，并期望作者遵守 NeurIPS 道德准则以及其所在机构的指南。
- 对于初次投稿，请勿包含任何可能破坏匿名性的信息（如适用），例如开展评审的机构。

16. **大语言模型使用声明**

问题：如果大语言模型是本研究核心方法中重要、原创或非标准的组成部分，论文是否描述了其使用情况？请注意，如果大语言模型仅用于写作、编辑或格式设置，且不 影响研究的核心方法、科学严谨性或原创性，则无需声明。

回答： [不适用]

理由：所研究的多模态大语言模型推荐系统是评估对象，而不是作者撰写过程或方法论流水线的组成部分；未以会影响研究核心方法、科学严谨性或原创性的非标准方式使用大语言模型。

指南：

- 答案 [不适用] 表示本研究的核心方法开发并未将大语言模型作为任何重要、原始或非标准组件。
- 关于哪些内容应当或不应当描述，请参阅 NeurIPS 手册中的大语言模型使用政策。