MAGuard：面向多模态推荐后门的多视角异常门控防御框架

张家琦[1]，刘顾纯[2]，王梦禹[1]，黄跃龙[3]，唐涛[4]，丁锋[3]，于硕[2]
（1.国家计算机网络应急技术处理协调中心，北京 100029；2.大连理工大学计算机科学与技术学院，辽宁 大连 116024；3.大连理工大学软件学院，辽宁 大连 116024；4.浙江工业大学计算机科学与技术学院，浙江 杭州 310023）

摘要：近年来，多模态推荐系统在内容分发、互联网服务领域应用持续普及。第三方视觉和文本预训练编码器常被直接复用，其生成的多模态表示通过特征缓存接口为下游推荐模型提供输入。这种由预训练编码器到推荐排序的系统架构在提升开发效率的同时引入了新的供应链安全风险。若编码器输出或缓存的多模态表示被隐蔽污染，攻击者无需改动推荐模型参数与训练数据，即可定向偏移特定用户群体的推荐排序结果，同时维持全局推荐性能基本稳定。针对这一问题，本文聚焦多模态推荐系统供应链后门的检测与修复需求，提出多视角异常门控防御框架MAGuard。该框架从局部邻域一致性、用户-物品协同行为、排序敏感性、全局分布偏离四个维度构建异常特征，融合生成物品综合风险评分以识别受污染物品。进一步通过异常门控屏蔽不可信多模态特征，结合分布校正将协同过滤得分映射至正常区间，实现无需重训练的推理修复，降低后门对目标用户推荐结果的干扰。在MovieLens-100K和Amazon-TG数据集上的实验结果显示，MAGuard可高效识别嵌入层后门异常，在抑制攻击效果的同时仅带来较低的正常推荐性能损失。本文研究为多模态推荐系统场景下的预训练组件复用审计、特征缓存安全校验及AI供应链安全防护提供了可落地的技术方案。

关键词：多模态推荐系统；预训练编码器；供应链安全；后门防御
中图分类号：TP309
文献标志码：A
doi：10.11959/j.issn.1000

MAGuard: A Multi-View Anomaly Gatekeeping Defense Framework for Backdoors in Multimodal Recommendation

Zhang Jiaqi[1], Liu Guchun[2], Wang Mengyu[1], Huang Yuelong[3], Tang Tao[4], Ding Feng[3], Yu Shuo[2]
1. National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China
2. Dalian University of Technology, School of Computer Science and Technology, Dalian 116024, China
3. Dalian University of Technology, School of Software, Dalian 116024, China
4. Zhejiang University of Technology, School of Computer Science and Technology, Hangzhou 310023, China

Abstract: Multimodal recommender systems are widely used in content delivery and online services. The third-party visual and textual pretrained encoders are often reused to downstream recommenders through feature caches. This decoupled deployment improves efficiency but exposes pretrained components and cached representations to supply-chain backdoors. If encoder outputs or cached features are covertly poisoned, attackers can bias target-user rankings without changing model parameters or training data, while keeping overall performance nearly unchanged. To address this threat, we propose MAGuard, a multi-view anomaly-gated defense framework. MAGuard detects poisoned representations by fusing local consistency, collaborative behavior, ranking sensitivity, and global distributional deviation into an item-level risk score. It then repairs recommendations at inference time by gating untrusted multimodal features and calibrating collaborative filtering scores, without model retraining. Experiments on MovieLens-10K and Amazon-TG show that MA-

收稿日期：2026–08–18；修回日期：2026–09–19
通信作者：唐涛，tao.tang@ieee.org
Foundation Items：This work received no .

Guard effectively detects embedding-level backdoors and mitigates attacks with minimal performance loss. This work supports pretrained-component auditing and AI supply-chain protection for multimodal recommender systems.

Key words: Multimodal recommendation system, pre-trained encoder, supply chain security, backdoor defense

0　引言

多模态推荐系统在短视频分发[1]、内容检索[2]、电商推荐和个性化信息服务等业务场景中发挥着日益关键的作用。相较于仅依赖用户－物品交互的协同过滤方法，多模态模型能够进一步利用商品图像、文本描述等内容信息，在交互稀疏或新物品场景中补充语义证据。随着预训练模型的不断发展，工业系统通常直接复用第三方预训练编码器[3]，离线提取物品特征并缓存，再由下游推荐模型完成图传播、特征融合和排序。该模式显著降低了训练与上线成本，也使编码器、特征文件和缓存接口逐渐成为推荐链路中的基础组件。

由预训练编码器、多模态表示到推荐排序的解耦架构虽显著提升了部署效率，却引入了新的供应链安全风险。由于编码器输出与缓存接口缺乏细粒度审计，第三方模块或缓存环节一旦受污染，攻击者无需修改下游模型参数、训练数据或排序逻辑，仅操纵少量物品嵌入表示即可诱导目标用户的推荐结果发生系统性偏移[4]。此类异常隐藏于高维连续嵌入空间，仅作用于特定用户群体或物品集合且不显著改变全局推荐指标，因此难以通过常规性能监控发现。

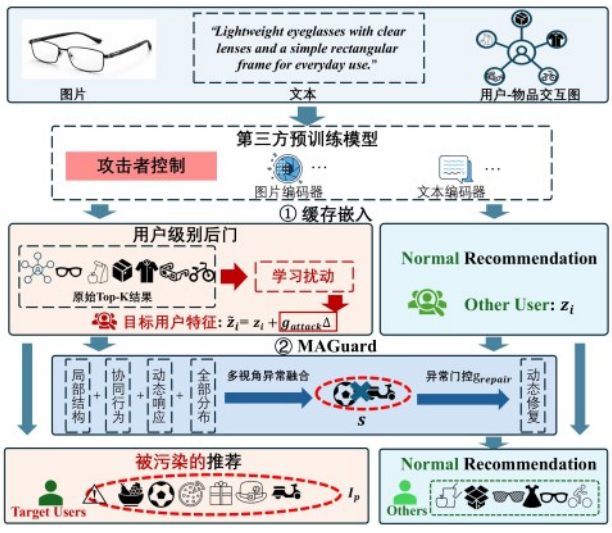

图 1　MAGuard 总体框架

现有后门防御机制难以直接迁移到上述嵌入层场景，主要存在三方面原因：（1）检测目标不同。传统方法通常围绕离散分类边界设计，通过逆向重构触发器或搜索固定标签锚点判断异常[5]。推荐系统输出的是连续偏好分数和 Top-k 排序，嵌入污染未必造成明确的类别翻转，因而逆向目标不稳定，容易出现漏检或误报。（2）数据分布复杂。推荐数据具有显著的长尾性和交互稀疏性[6-7]，大量正常冷门物品天然偏离主体特征流形。若仅依赖全局距离或单一离群分数，检测器很容易把低频物品与真实后门混淆，进而扩大防御范围并损害个性化能力。（3）修复条件受限。剪枝[8]、参数微调[9]和重新训练[10]通常要求可获得可信样本或能够访问完整模型训练过程，而供应链场景中的防御者未必掌握原始编码器及其干净训练数据。若直接改变下游参数，还可能破坏已经学习到的协同过滤结构和模态对齐关系。因此，面向嵌入污染风险的防御不仅要识别隐藏在连续特征空间中的异常，还要在尽量不修改下游模型的前提下恢复目标用户的排序结果。

针对上述挑战，本文提出面向多模态推荐供应链后门的多视角异常门控框架 MAGuard。该框架在不依赖后门标签和上游模型的条件下，对可疑物品进行多维度风险审计，并在推理阶段仅对高风险表示实施选择性修复。主要贡献如下：

（1）提出多模态推荐系统预训练编码器的供应链安全问题。从系统架构层面分析第三方编码器复用引入的安全风险，说明嵌入污染攻击无需修改下游推荐模型参数和训练数据，即可定向改变特定用户群体推荐结果的攻击特性。

（2）构建适配多模态推荐供应链安全的多视角异常门控防御框架 MAGuard。基于局部邻域特征一致性、用户-物品交互协同规律、排序结果敏感性及全局特征分布偏离度四类维度生成异常风险信号，精准识别受污染的物品表示。通过动态异常门控机制截断恶意特征传导路径，在不影响正常用户推荐效果的前提下修正目标用户的排序偏移问题。

（3）开展基于 MovieLens-100K 和 Amazon-TG

真实数据集的系统性验证实验。覆盖视觉模态污染、文本模态污染、双模态联合污染三类攻击场景，从异常检出准确率、推荐效果修复率、正常用户性能损耗、核心模块消融实验四个维度验证所提方法的有效性。

1　相关工作

1.1　推荐系统投毒检测

推荐系统安全研究最早主要围绕交互数据投毒展开。Fang 等人[11]从用户－物品图结构出发分析恶意交互注入改变图推荐结果的机理，说明少量精心设计的边即可干扰目标物品的曝光。LOKI[12]利用序列行为偏移构造更隐蔽的下一物品推荐污染，进一步表明攻击可以通过模拟正常行为降低被统计规则发现的概率。此类工作主要在训练数据层面建立威胁模型，检测时通常依赖用户画像、评分分布或交互序列中的异常，因此更适合显式日志污染。随着后门攻击研究向推荐任务延伸，研究重点逐步转向具有触发条件的定向操控。BTI-DBF[13]通过逆向优化推断潜在触发模式，Anti-FakeU[14]则结合用户表示聚类识别异常攻击子群。这些方法能够针对特定目标实施防御，但通常仍假设触发器或异常行为在输入空间中存在可观测痕迹。预训练表示被广泛复用后，攻击面进一步从数据层转移到表征层。BadEncoder[15]证明后门可以嵌入预训练编码器的特征空间，并在下游任务中保持攻击能力。UOR[16]进一步展示了利用表示对齐实现跨任务持续激活的可能性。然而，现有防御方法多针对训练样本投毒或显式触发器注入的攻击模式，无法适配仅篡改预训练编码器输出嵌入的供应链攻击。此类攻击无需修改下游模型参数与训练数据，仅通过特征缓存链路即可影响最终排序结果，既无显式恶意交互痕迹，也不存在清晰的异常类别边界，且单一统计指标易受长尾分布干扰导致漏判，因此需要联合图结构、协同行为、排序响应、全局分布多维度检测方法，完成供应链后门的精准识别与快速修复。

1.2　AI 供应链安全

随着大规模预训练模型和开源组件的广泛复用，AI 供应链安全逐渐成为研究热点。早期研究主要关注第三方预训练模型复用风险。MDP[17]指出恶意后门可嵌入预训练语言模型并在微调阶段保持激活。BadEncoder[15]进一步证明编码器级后门能够跨任务迁移至下游应用。Casey 等人[18]系统揭示了开放模型托管平台中的潜在恶意传播路径，表明开源模型复用显著扩大了攻击入口。近期研究开始探索预训练后门的供应链传播机制。TransTroj[19]提出通过嵌入不可区分性实现跨下游任务稳定迁移，使攻击者无需控制目标训练流程即可诱导隐蔽偏移。针对上述威胁，研究者提出了模型签名验证[20]、来源追踪[21]和静态恶意代码扫描[22]等防护机制。然而，这些方法主要针对模型文件完整性或加载阶段恶意执行。因此，需要面向推理阶段的动态审计机制。

2　问题定义

2.1　系统模型

在多模态推荐系统建模中，设用户集合为$\mathcal{U}$，物品集合为$\mathcal{I}$，根据用户-物品历史交互日志，用户与物品间的历史交互构成了二部图$\mathcal{G}=(\mathcal{U},\mathcal{I},\mathcal{E})$，$\mathcal{E}$表示交互。对于任意物品$i\in\mathcal{I}$的模态$m$，第三方预训练编码器提取多模态特征向$\mathbf{z}_i^m$。下游多模态推荐模型$f_\theta$基于交互图与物品特征，计算用户$u\in\mathcal{U}$对物品$i$的正常偏好预测分数：

$$\hat{y}_{ui}=f_\theta\left(u,i|\mathcal{G},\mathbf{z}_i\right)=\alpha\hat{y}_{ui}^{CF}+\beta\hat{y}_{ui}^{M} \tag{1}$$

其中$\hat{y}_{ui}^{CF}$与$\hat{y}_{ui}^{M}$分别表示协同与模态得分，α和β为预设的融合权重。

2.2　供应链投毒

本文考虑可用性破坏型攻击，即攻击者仅通过受控的上游编码器或缓存接口注入特征扰动，降低物品在目标用户侧的偏好得分，使其从 Top-k 列表中消失，而非全面破坏系统。在上游编码器受控条件下，攻击者向目标物品集$\mathcal{I}_p\subset\mathcal{I}$注入扰动，生成污染表示物品表征$\tilde{\mathbf{z}}_i$，使其在目标用户上的预测分数$\tilde{y}_{ui}$异常下降，从而跌出推荐列表。因此，防御的首要任务是在无监督条件下识别异常集合$S\subset\mathcal{I}$，使其尽可能覆盖真实污染集合$\mathcal{I}_p$。

2.3　防御目标

防御的目标是实现后门物品的准确筛查与系统性能的鲁棒修复。为此，本文将物品i的风险程度量化为多视角异常分数$\mathcal{A}\left(i|\mathcal{G},\tilde{\mathbf{z}}_i,f_\theta\right)$，并将净化修复后的用户偏好分数记为$\bar{y}_{ui}$。防御任务包括异常检测与推理修复：具体优化过程分为检测与修复两个阶段：

（1） 检测阶段，优化目标为最大化后门物品的召回数目：

$$\max_{\mathcal{S}\subset\mathcal{I}} \left|\mathcal{S}\cap\mathcal{I}_p\right| \tag{2}$$

其中，嫌疑集合S为所有物品按照异常分数$\mathcal{A}(i)$降序排列的前N个物品构成。

（2） 修复阶段，模型通过异常门控阻断恶意特征传播，优化目标是最小化目标用户在推理时的排序性能损失：

$$\min \sum_{u\in\mathcal{U}_t} \mathcal{L}_{rank}\left(\bar{y}_{ui},\hat{y}_{ui}\right) \tag{3}$$

3 具体方法

本文设计的MAGuard防御框架整体架构如图1所示。多视角异常检测模块针对输入的物品表征向量完成多维度风险量化，初步筛查出存在后门污染嫌疑的物品集合。异常融合模块通过自适应权重分配机制对多维度检测结果进行加权计算，输出对应物品的综合异常概率评分。动态修复模块依据上述综合评分触发特征防御门控机制，对判定为不可信的多模态特征执行动态拦截，同时将纯协同过滤分支的输出得分映射至正常分布范围，削弱后门攻击对推荐排序结果的干扰。

3.1 后门攻击机制

在上游编码器受控、下游模型不可见的黑盒威胁场景中，攻击者只能通过前端推荐界面观察目标用户的Top-k曝光结果，无法获得模型参数、梯度、内部打分、完整排序列表以及用户－物品交互日志。尽管可见信息有限，攻击者仍可利用多次查询得到的曝光差异构造目标用户锚点，并通过黑盒搜索估计能够压低目标物品排名的全局扰动方向。随后，扰动被植入第三方编码器输出或特征缓存接口，使下游系统在不知情的情况下直接使用被污染的多模态表示。

具体而言，攻击者通过黑盒优化方法估算出全局扰动方向Δ。对满足触发条件$g_{\text{attack}}(u)$的请求，将扰动注入编码器输出：

$$\tilde{\mathbf{z}}_i = \mathbf{z}_i + g_{\text{attack}}(u)\Delta \tag{4}$$

其中，$g_{\text{attack}}(u)\in\{0,1\}$表示由正常Top-k曝光行为筛选得到的用户触发门控。当用户请求满足触发条件时扰动被激活，否则返回正常模态表示。此类攻击难以在训练阶段审计且不影响整体性能，隐蔽性极强。因此需针对性设计异常检测与修复双层机制，在下游侧依托用户-物品交互结构信息挖掘异常特征，结合实时推理响应完成污染样本识别与推荐结果校准。

3.2 多视角异常检测机制

工业部署中，第三方预训练编码器常被默认安全并直接信任，这使得攻击者可轻易在其输出的连续嵌入空间注入隐蔽的局部扰动。若仅依赖单一特征统计指标检测，会因为特征流形的复杂性而产生大量误报或漏报。为克服单一视角局限，本节从局部图拓扑、动态推断响应及全局异常感知三个核心维度，构建多视角异常检测机制。

3.2.1 启发式候选过滤机制

交互频率极低的长尾物品因偏离主体数据流形，表示分布通常存在较大随机波动，直接纳入检测易引入噪声。同时，攻击者为提升攻击收益，更倾向于污染曝光概率较高的活跃物品。因此，本文基于用户－物品交互图$\mathcal{G}$的频率先验设计启发式候选过滤机制，以缩减异常检测的搜索空间。

对于任意物品i，存在历史交互的用户集合：

$$\mathcal{U}_i = \left\{u\in\mathcal{U}\middle|(u,i)\in\mathcal{E}\right\} \tag{5}$$

基于该集合，物品的交互频率定义为：

$$freq(i) = \left|\mathcal{U}_i\right| \tag{6}$$

其中，$freq(i)$表示与物品i发生历史交互的用户数量。

为过滤低频长尾物品，设定频率阈值τ，筛选出满足$freq(i)>\tau$的高曝光物品，并取前N个构成候选集合$\mathcal{C}$，后续多视角异常检测仅在候选集合$\mathcal{C}$内执行。过滤机制能够有效地缩小搜索空间，精准聚焦最可能成为可用性攻击目标的核心物品。

3.2.2 局部一致性偏差

在正常推荐数据中，具有相似协同受众或共同消费关系的物品往往在多模态表示空间形成相对稳定的局部结构。攻击者若为了定向降低目标物品得分而改变其嵌入，容易使该物品表示相对于一跳共现邻居产生方向或距离上的异常偏移，从而偏离原有的语义表达。因此本文通过度量候选物品与其邻域中心的偏离程度刻画局部异常。

对于候选物品$i\in\mathcal{C}$，定义其邻域中心表示$\mathbf{z}_c$为：

$$\mathbf{z}_c = \frac{1}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)}\tilde{\mathbf{z}}_j \tag{7}$$

其中，$\mathcal{N}(i)$为物品i的一跳共现邻居集合，若集合为空，则$\mathbf{z}_c$取全局邻居向量均值。

物品与邻域中心的局部偏差定义为：

$$D_{loc}(i) = 1 - \frac{\tilde{\mathbf{z}}_i^{\top}\mathbf{z}_c}{\|\tilde{\mathbf{z}}_i\|_2\|\mathbf{z}_c\|_2}$$
$$D_{\text{maha}}(i) = \left(\mathbf{v}_i - \boldsymbol{c}_i\right)^{\top}\Sigma^{-1}\left(\mathbf{v}_i - \boldsymbol{c}_i\right) \tag{8}$$

其中，D_{loc}与D_{maha}分别为余弦方向偏差以及马氏距离偏差。$\mathbf{v}_i$和$\boldsymbol{c}_i$分别为$\tilde{\mathbf{z}}_i$和$\mathbf{z}_c$经主成分分析降维后的特征，Σ为全局降维特征的协方差矩阵。

经过鲁棒标准化，局部一致性异常分数计算如下：

$$s_{emb}(i) = \text{Norm}\left(D_{loc}(i) + D_{maha}(i)\right) \tag{9}$$

其中，Norm(·)表示基于绝对中位差的归一化操作。$s_{emb}(i)$越大表明物品越偏离其真实邻域结构，异常风险越高。局部一致性异常分数采用绝对中位差进行鲁棒标准化，可以减弱少数极端值对阈值的牵引，使不同模态和不同数据密度下的局部分数更便于统一比较。

3.2.3　协同共现行为偏差

局部表示偏移并不一定都会改变推荐行为，因此仅观察特征空间仍可能产生误报。本文进一步利用用户－物品历史共现关系刻画行为层一致性。正常物品的推荐结果通常符合长期交互形成的协同模式，会与关联度较高的物品集合稳定共同出现在用户偏好上下文当中。而受污染物品的排名被异常压低后，其所在的Top-k推荐列表与历史共现模式将产生显著偏离。本文通过用户共现性指标（UCC）量化上述一致性偏差，并将其转化为对应物品的协同行为异常分数。

首先，物品i与物品j的历史关联强度为：

$$rel(i,j) = \frac{\left|\mathcal{U}_i \cap \mathcal{U}_j\right|}{\max\left(\left|\mathcal{U}_i\right|,1\right)} \tag{10}$$

然后，抽取部分交互用户集合$\hat{\mathcal{U}}_i \subset \mathcal{U}_i$，统计当前Top-k推荐结果$\mathcal{R}_u$，共现一致性指标定义如下：

$$\text{UCC}(i) = \frac{1}{\left|\hat{\mathcal{U}}_i\right|}\sum_{u\in\hat{\mathcal{U}}_i}\frac{1}{k}\sum_{j\in\mathcal{R}_u} rel(i,j) \tag{11}$$

基于共现一致性指标，得到协同行为异常分数：

$$s_{beh}(i) = 1 - \text{UCC}(i) \tag{12}$$

$s_{beh}(i)$用于度量推荐结果与历史共现模式的一致性。当该指标持续偏低时，说明模型当前给出的推荐上下文与历史协同行为发生冲突。该视角不依赖多模态特征间的绝对位置，可与表示层局部一致性形成互补。

3.2.4　排序敏感性探测

基于图拓扑与行为共现的静态特征维度仅能捕捉物品的历史交互快照，无法处理后门嵌入特有的动态扰动敏感特性。攻击者在扰动资源受限条件下实施定向投毒时，往往会对目标物品的嵌入表示进行梯度引导的精细微调，使其特征被推送至损失函数的陡峭局部最优区域，极轻微的特征干扰即可导致其在推荐列表中的位次发生大幅跳变。针对该类攻击独有的动态特征，本文提出多尺度随机扰动探测框架以量化物品的排序鲁棒性水平。

对于候选物品i，选取历史交互用户子集$\mathcal{U}_i$作为探测锚点，记录基线性能：平均预测得分Y_{base}、平均排序位置R_{base}、Top-k命中率P_{base}以及列表重合度O_{base}。

为了生成多尺度前向扰动，定义基础扰动强度$\sigma_{base} = \gamma\cdot\mathbb{E}\left[\|\tilde{\mathbf{z}}\|_2\right]$和多尺度系数集$\mathcal{M}$（$\mathbb{E}\left[\|\tilde{\mathbf{z}}\|_2\right]$为全局特征向量的平均$L_2$范数，$\gamma$为超参数）。对每个尺度$m\in\mathcal{M}$，采样各向同性单位噪声$\boldsymbol{\eta}\sim\mathcal{N}(0,\mathrm{I})$，对物品特征施加正负双向扰动：

$$\tilde{\mathbf{z}}_i^{+} = \tilde{\mathbf{z}}_i + m\cdot\sigma_{base}\cdot\frac{\boldsymbol{\eta}}{\|\boldsymbol{\eta}\|_2},$$
$$\tilde{\mathbf{z}}_i^{-} = \tilde{\mathbf{z}}_i - m\cdot\sigma_{base}\cdot\frac{\boldsymbol{\eta}}{\|\boldsymbol{\eta}\|_2} \tag{13}$$

扰动后对应指标表示为Y_{pert}，R_{pert}，P_{pert}，O_{pert}，统一形式化各指标衰减为:

$$\Delta(\cdot) = \max\left(0,(\cdot)_{\text{base}} - (\cdot)_{\text{pert}}\right) \tag{14}$$

综合风险得分排序敏感性异常得分：

$$s_{ris}(i) = \text{Norm}\left(w_1\Delta P + w_2\Delta Y + w_3\Delta R + w_4\Delta O\right) \tag{15}$$

其中，w_1，w_2，w_3，w_4为各项对齐指标的预设超参数权重。$s_{\text{ris}}(i)$量化了物品对微扰的排序脆弱性。若某物品的$s_{\text{ris}}(i)$分数较高，则表明其同时在预测分数、推荐位次、列表结构三个维度表现出异常的扰动响应敏感性，该物品被攻击者定向优化为后门

样本的概率更高。

3.2.5 全局异常感知

局部结构、协同行为和排序敏感性均以候选物品附近的相对关系为依据，虽然能够捕获细粒度异常，但仍可能遗漏整体语义空间中的远距离离群点。本文引入全局异常感知模块学习正常数据的支持域边界，并用偏离边界的程度衡量全局的异常风险。

具体地，本文采用单类支持向量机[23]（One-Class SVM，OCSVM）对候选物品的多模态表示建立无监督全局边界。考虑到多模态特征之间可能具有非线性关系，本文使用径向基核函数学习该边界，并据此定义全局异常分数：

$$s_{glo}(i) = \rho - \sum_j \alpha_j \mathrm{K}\left(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_j\right) \tag{16}$$

其中，α_j为支持向量对应的拉格朗日系数，$\mathrm{K}(\cdot,\cdot)$为径向基核函数，ρ为决策超平面的偏置项。$s_{glo}(i)$越大，表明物品偏离全局正常分布越远，受到后门污染的可能性越高。

3.3 异常门控融合机制

由于多视角检测输出的异常信号在统计含义、数值尺度上存在异质性，直接统一加权融合易导致某一视角主导检测结果，无法兼顾局部行为异常与全局分布偏移的识别需求。为此MAGuard设计两级异常门控融合机制。首先基于局部同质视角计算物品综合风险评分，再与全局视角检测结果进行排序级集成得到最终可疑物品集合。最后通过异常门控动态判定推理阶段是否执行修复，在充分利用多视角信号互补性的同时避免全局离群检测偏好在长尾场景下产生大量误报。

3.3.1 综合风险评分

有效融合多视角异常信号，需要解决局部与全局异常检测机制的异质性适配问题。局部一致性、协同行为偏差、排序敏感性三类信号均基于候选物品的局部关联结构或用户交互响应计算，可先标准化后加权融合。全局OCSVM检测基于核边界识别离群样本，易将业务正常的长尾物品误判为高异常样本。两类信号的语义与数值分布差异显著，直接线性加权会放大偏差。因此本文采用排序级集成策略，在固定检出预算约束下按比例分别从两类信号中筛选可疑样本，取并集构成最终嫌疑集合完成筛查。

局部多视角综合风险分数定义如下：

$$s_{loc}(i) = \lambda_1 s_{emb}(i) + \lambda_2 s_{beh}(i) + \lambda_3 s_{ris}(i) + \lambda_4 s_{freq}(i) \tag{17}$$

其中，λ_1，λ_2，λ_3为每个异常视角的超参数权重，λ_4为图频率先验的微弱奖励项。各分数均为采用Z-score标准化后的对齐分数。

设防御系统总检出预算为B，引入比例系数$\eta \in [0,1]$控制局部视角与全局视角的贡献度。分别构造局部嫌疑子集与全局嫌疑子集：

$$\begin{aligned} \mathcal{C}_{loc} &= \left\{ i \in \mathcal{C} : rank\left(s_{loc}(i)\right) \leqslant \eta B \right\}, \\ \mathcal{C}_{glo} &= \left\{ i \in \mathcal{C} : rank\left(s_{glo}(i)\right) \leqslant (1-\eta) B \right\} \end{aligned} \tag{18}$$

最后可疑物品集合定义为两者并集：

$$\mathcal{S} = \mathcal{C}_{loc} \cup \mathcal{C}_{glo} \tag{19}$$

3.3.2 异常门控机制

后门污染主要作用于模态分支，而协同过滤分数通常保持稳定，因此利用其作为可信锚点定义门控判定函数g_{repair}：

$$g_{repair}\left(\hat{y}_{ui}^{CF}, \tilde{y}_{ui}\right) = \mathbb{I}\left(\hat{y}_{ui}^{CF} - \tilde{y}_{ui} \geqslant \delta\right) \tag{20}$$

其中，δ为触发阈值，$\mathbb{I}(\cdot)$为指示函数。该设计仅在异常风险显著且融合分数明显偏离可信协同信号时触发修复，避免对正常物品造成误修复。

3.4 动态修复机制

异常门控确定需要修复的物品后，最直接的方法是完全移除多模态分支，仅保留纯协同过滤得分。但是，多模态推荐模型在训练时已经学习了协同信号与内容信号的融合尺度，纯CF分数的均值和方差通常明显低于最终融合分数。若直接替换，虽然切断了污染路径，嫌疑物品仍可能因为数值尺度偏低而继续处于排名末位。为此，MAGuard使用当前推理批次中未被判定为异常的物品作为安全锚点，建立纯协同分数到正常融合分布之间的动态映射完成修复：

$$\Phi\left(\hat{y}_{ui}^{CF}\right) = \mu_f + \frac{\sigma_f}{\sigma_{\mathrm{p}}}\left(\hat{y}_{ui}^{CF} - \mu_{\mathrm{p}}\right) \tag{21}$$

其中，μ_f,σ_f分别为正常融合得分均值和标准差。μ_{p},σ_{p}为嫌疑物品对应协同过滤分数均值和标准差。

针对嫌疑集合S，本文在推理阶段动态触发异常门控，最终输出的推荐模型的推荐分数为：

$$\bar{y}_{ui}=\begin{cases}\Phi\left(\hat{y}_{ui}^{CF}\right), i\in\mathcal{S}\wedge g_{repair}\left(\hat{y}_{ui}^{CF},\tilde{y}_{ui}\right)=1\\ \tilde{y}_{ui},\ \text{otherwise.}\end{cases}\tag{22}$$

动态修复机制只改变可疑物品的评分尺度，既能保留协同过滤分支的相对排序关系，又能减少直接切断模态分支造成的分数塌缩，适合特征缓存受污染但缺乏未污染编码器的供应链场景。

4　性能评估

4.1　数据集

本文选用 MovieLens-100K 和 Amazon-TG 两个真实多模态推荐数据集，分别考察相对稠密与明显长尾的推荐场景。MovieLens-100K 结合小规

表 1　　数据集统计分析

数据集	用户数量	物品数量	交互数量
MovieLens-100K	610	9,738	100,837
Amazon-TG	19,412	11,907	167,371

模用户对电影的评分交互，包含电影海报和文本描述构造视觉、文本特征，适合观察后门污染在核心物品圈层中的影响。Amazon-TG 来自亚马逊游戏与玩具类别，用户和物品规模更大、交互更稀疏，用于检验检测器在长尾强噪声下的稳定性。预处理后，两个数据集的统计结果见表 1。

4.2　基线方法

本文以 LightGT[24]作为下游推荐框架，将对比方法分为检测与修复两阶段。检测阶段包括 DECREE、STRIP、激活聚类(AC)、谱签名(SS)和隔离森林(IF)。STRIP 对表示施加随机噪声，比较多次推理结果的稳定度判断后门。AC 对隐层特征做 K-Means 聚类，将样本量小的簇视为异常。SS 对特征矩阵奇异值分解，依第一主成分投影绝对值找离群点。IF 随机划分特征空间，根据平均路径长度识别异常。DECREE 专为供应链后门场景设计，通过检测预训练编码器的后门攻击，并在特征缓存供应链攻击下评估模型表现。修复阶段采用微调(FT)、剪枝微调(FP)和对抗训练(AT)。FT 用少量干净数据再训练以削弱后门。FP 先剪去异常激活的神经元，再微调模型。AT 引入对抗扰动以检验模型鲁棒性。

4.3　实验参数

所有实验均基于 Windows 11 操作系统实现，采用 Python 3.12.6 与 PyTorch 2.8.0 作为开发框架，GPU 驱动版本为 CUDA 12.9 与 cuDNN 9.1，实验可在单张显存 16 GB 的 NVIDIA GeForce RTX 5070 Ti GPU 上运行。为保证实验结果的可复现性，所有攻击与防御实验的随机种子均固定为 1700。

参数设置方面，为保证防御实验与攻击实验的输入空间一致，本文采用相同的离线特征编码流程。视觉模态采用预训练 ResNet-50 提取图像特征，文本模态采用 paraphrase-multilingual-MiniLM-L12-v2 编码物品描述，两类表示预先计算并缓存后作为 LightGT 的内容侧输入。图频率先验中，最小交互阈值 τ 为 2，仅保留与至少两名用户交互过的物品，并按度数降序截取 N=500 个物品构成候选池。图像和文本模态的特征维度

分别设置为 2048 和 100。局部特征一致性检测中，使用 PCA 将视觉与文本特征统一降维至 50 维。排序敏感性探测中，基础扰动强度比例系数，多尺度扰动乘子集合 $\mathcal{M}$={1.0,2.0,5.0}。综合风险得分模块中，衰减权重 w_1，w_2，w_3，w_4 分别设置为 0.2, 0.4, 0.3, 0.1。异质视角排序级集成的防御预算 B 设为 100，嫌疑集合的集成比例设为 0.6: 0.4。修复模块中，异常门控阈值 δ 设为 0.1。其余参数沿用 LightGT 默认设置。

4.4　评价指标

本文从异常检测效果和荐修复效果两个方面评估防御方法。

4.4.1　异常检测指标

本文使用召回率（Recall@K）衡量成功检测出的真实后门物品占所有后门物品的比例：

$$\text{Recall@K}=\frac{\left|\mathcal{S}_K\cap\mathcal{I}_p\right|}{\left|\mathcal{I}_p\right|}\tag{24}$$

精确率（Precision@K）衡量检测出的嫌疑物品中真实后门物品的比例：

$$\text{Precision@K}=\frac{\left|\mathcal{S}_K\cap\mathcal{I}_p\right|}{K}\tag{25}$$

ROC-AUC 评估异常分数对正常与异常物品的整体排序区分能力。

F1@K　于综合衡量检测准确率和召回率：

$$F1@\text{K}=\frac{2\cdot\text{Precision@K}\cdot\text{Recall@K}}{\text{Precision@K}+\text{Recall@K}}\tag{26}$$

表 2　MovieLens 数据集上的后门检测结果

Method	Modality	Recall @100	Precision @100	F1 @100	AUC-ROC
STRIP	Vision	0.00	0.00	0.00	22.29
AC		0.00	0.00	0.00	55.51
SS		0.00	0.00	0.00	44.42
IF		40.00	12.00	18.46	**93.55**
DECREE		46.61	13.70	**25.20**	70.07
Ours		**53.33**	**16.00**	24.61	76.22
STRIP	Text	0.00	0.00	0.00	60.66
AC		0.00	0.00	0.00	43.87
SS		0.00	0.00	0.00	59.28
IF		6.67	2.00	3.08	63.71
DECREE		36.67	11.56	17.60	72.48
Ours		**53.33**	**16.00**	**24.61**	**76.22**
STRIP	Both	0.00	0.00	0.00	48.65
AC		0.00	0.00	0.00	46.15
SS		0.00	0.00	0.00	53.35
IF		23.33	7.00	10.77	**90.17**
DECREE		61.00	18.41	28.28	84.57
Ours		**76.67**	**23.00**	**35.38**	88.02

表 3　Amazon-TG 数据集上的后门检测结果

Method	Modality	Recall @100	Precision @100	F1 @100	AUC-ROC
STRIP	Vision	0.00	0.00	0.00	53.49
AC		6.82	3.00	4.17	58.59
SS		0.00	0.00	0.00	25.95
IF		4.55	2.00	2.78	**87.20**
DECREE		18.18	8.64	11.71	71.34
Ours		**29.55**	**13.00**	**18.06**	70.08
STRIP	Text	0.00	0.00	0.00	45.12
AC		2.27	1.00	1.39	54.69
SS		0.00	0.00	0.00	49.04
IF		0.00	0.00	0.00	**70.71**
DECREE		25.00	11.90	16.10	66.52
Ours		**40.91**	**18.00**	**25.00**	70.06
STRIP	Both	0.00	0.00	0.00	49.05
AC		4.55	2.00	1.39	56.54
SS		0.00	0.00	0.00	37.94
IF		9.09	4.00	5.56	**81.22**
DECREE		22.73	10.00	13.89	72.18
Ours		**34.09**	**15.00**	**20.82**	70.09

4.4.2　推荐修复指标

本文使用归一化折损累计增益（NDCG@K）和精确率（Precision@K）评估修复后目标用户推荐列表的质量。

为了直观量化推荐系统的核心排序性能从受损状态恢复至原始干净状态的程度，本文采用NDCG来定义恢复率（Recovery Rate）：

$$\text{Recovery} = \frac{\text{NDCG}_{\text{defended}} - \text{NDCG}_{\text{attacked}}}{\text{NDCG}_{\text{origin}} - \text{NDCG}_{\text{attacked}}} \quad (27)$$

4.5　对比实验

4.5.1　风险场景设置

为模拟第三方预训练编码器或缓存接口被污染的供应链风险，本文在LightGT上构造三类嵌入层攻击场景：（1）视觉模态污染，仅扰动物品视觉表示；（2）文本模态污染，仅扰动文本表示；（3）双模态污染，同时修改视觉和文本表示。三种场景均保持LightGT网络结构、训练数据、用户－物品交互图和原有排序流程不变，仅改变上游预训练特征。该设置可以把性能变化直接归因于表示层污染，并分别考察MAGuard在单模态与更强双模态攻击下的防御能力。

4.5.2　异常检测效果

表2和表3给出了固定K=100时的检测结果。传统后门防御方法在连续推荐排序场景中的表现整体较弱。STRIP、AC和SS在MovieLens-100K的三类攻击下Recall几乎均为0，在Amazon-TG上AC仅在视觉和双模态场景出现6.82%和4.55%的有限召回。这说明分类任务中常用的输出稳定性或谱异常假设难以直接适配由连续偏好分数决定的Top-k排序。IF虽然在部分场景具有较高AUC，但Top-100真实后门召回并不稳定。MovieLens视觉攻击下IF的Recall为40.00%，AUC-ROC为93.55%，而在长尾更明显的Amazon-TG视觉攻击下，AUC-ROC仍达87.20%但Recall仅为4.55%。这种差异表明全局离群方法容易把正常长尾物品排到嫌疑列表前部，从而降低固定预算下的实际检出率。

本文进一步引入DECREE验证MAGuard在供应链场景的针对性优势。DECREE通过检测编码器输出表示的分布异常识别被植入后门的样本。实验结果显示，DECREE在MovieLens-100K双模态攻

击下的 Recall@100 为 61.00%，显著优于传统方法 IF 的 23.33%，但仍低于 MAGuard 的 76.67%。视觉攻击场景中 DECREE 的 Recall 为 46.61%，介于 IF 的 40.00% 和 MAGuard 的 53.33%

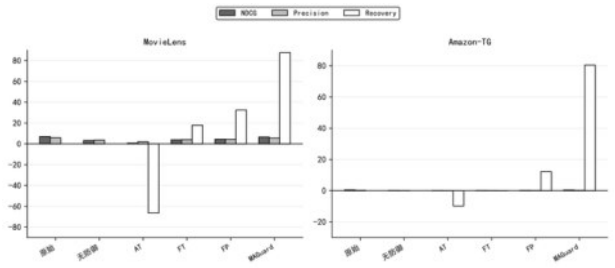

图2　双模态攻击推荐修复效果对比

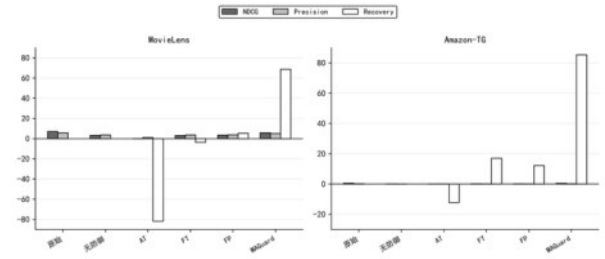

图3　文本模态攻击推荐修复效果对比

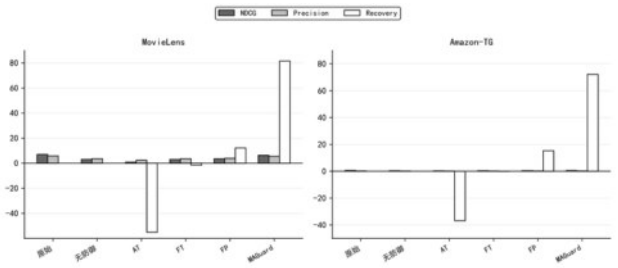

图4　视觉模态攻击推荐修复效果对比

之间。文本攻击场景检测难度更大，DECREE 为 36.67%，而 MAGuard 达到 53.33%。在更稀疏的 Amazon-TG 数据集上，双模态攻击下 DECREE Recall 从 MovieLens 的 61.00% 降至 22.73%，而 MAGuard 从 76.67% 降至 34.09%，相对衰减幅度更小。这说明纯表示空间检测方法在稀疏交互场景下更容易受噪声干扰，MAGuard 通过候选过滤和协同共现关系作为稳定参照系，在长尾场景下保持鲁棒性。

跨数据集比较进一步说明多视角设计的必要性。MovieLens-100K 交互较集中，核心物品具有较充分的协同邻居，因此局部结构和行为一致性可以形成较稳定的参照。Amazon-TG 交互更稀疏，正常物品之间的表示差异和度数差异更大，任何单一异常判据都更容易受到数据分布影响。MAGuard 在 Amazon-TG 上的绝对 Recall 低于

MovieLens-100K，但仍在文本与双模态攻击中保持了 Top-100 后门覆盖，说明候选过滤和多视角联

表 4　　多视角检测模块性能消融实验

变体	s_{emb}	s_{beh}	s_{ris}	s_{glo}	Recall @100	F1 @100
□	√				70.00	32.31
□		√			70.00	32.31
□			√		56.67	26.15
□				√	53.40	24.61
□		√	√	√	66.67	30.77
□	√		√	√	63.33	29.23
□	√	√		√	70.00	32.31
□	√	√	√		73.33	33.85
Ours	√	√	√	√	76.67	35.38

合能够缓解长尾噪声。IF 和 DECREE 在 Amazon-

TG 上高 AUC 低 Recall 的现象表明，防御评估应着重考察 Top-k 位置中受污染对象的聚集程度，而非仅依赖全局排序指标，因为实际修复覆盖受限于有限嫌疑处理能力。

4.5.3　恢复效果

图 2-图 3 和图 4 分别比较双模态、文本模态

和视觉模态攻击下的推荐修复效果。总体趋势显示，直接进行参数级修复并不一定能恢复连续排序结构。AT 在各场景的 Recovery 均为负，说明在已被污染的多模态特征上继续注入对抗扰动，可能进一步破坏协同过滤与内容表示之间原有的融合关系。以双模态攻击为例，AT 在 MovieLens-100K 和 Amazon-TG 上的 Recovery 分别为 -58.42% 和 -25.76%。FT 和 FP 的恢复率最高约 52%，但依赖少量干净样本进行参数更新，恢复幅度仍受到限制。与之相比，MAGuard 通过先门控不可信模态，再用纯协同信号进行动态分布对齐，在三个攻击场景和两个数据集上均保持较高恢复率，稳

定在 75%。这些结果表明，修复的关键并非重新学习整个推荐模型，而是准确截断污染特征

的传播路径，并把可信协同分数恢复到正常融

合尺度。由于该过程无需重新训练，计算和时间开销也更适合供应链侧的在线或周期性审计。

4.6　消融实验

4.6.1　多视角检测模块

模块级消融实验考察表示层局部一致性(s_{emb})、协同行为偏差(s_{beh})、排序敏感性探测(s_{ris})、的组合性能，结果如表 4 所示。单一视角消融结果显示，四个维度单独使用时 Recall@100 分别为 70.00%、70.00%、56.67% 和 53.40%，说明每类异常信号均具有独立检测能力，但单一视角难以充分覆盖不同类型的后门特征。移除视角结果进一步表明，移除 Local、Behavior、Ranking 和 Global 后，Recall 分别下降至 66.67%、63.33%、70.00% 和 73.33%，对应 F1 也由完整模型的 35.38% 降至 30.77%、29.23%、32.31% 和 33.85%。其中，局部结构与协同行为提供主要的静态约束，排序敏感性补充动态扰动响应，而全局分布视角用于发现局部关系难以覆盖的整体离群异常。四个维度共同使用时取得最高的 Recall@100=76.67% 和 F1@100=35.38%，说明各视角在异常模式上具有互补性，融合能够减少单一判据的遗漏，而非形成冗余堆砌。

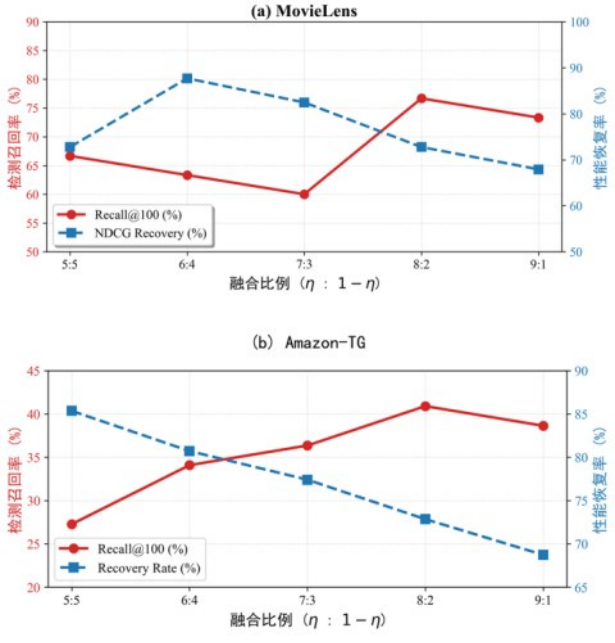

图 5　异质视角集成比例消融

4.6.2　异质检测与性能恢复比例验证

为探究集成比例 η 对防御效果的影响，本文在 MovieLens-100K 和 Amazon-TG 的双模态攻击下进行消融实验。图 5(a) 表明，在 MovieLens-100K 上，当局部占比提高至 0.8 时，Recall@100 达到 76.67%，但更多高曝光物品被纳入门控范围，使 NDCG Recovery 降至

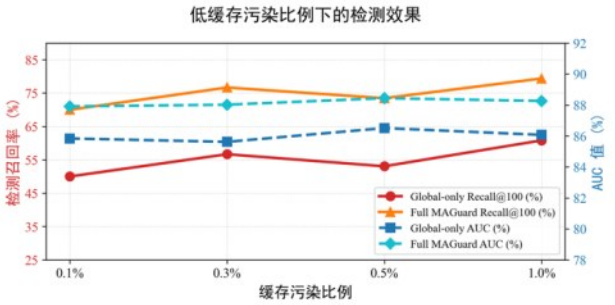

图 6　低污染场景检测稳定性验证

72.76　%；当 η 等于 0.6 时，Recall 为 63.33%，而

Recovery 提升至 87.69%。如图 5(b)，在更稀疏、长尾更明显的 Amazon-TG 上，局部占比由 0.5 增至 0.8 时，Recall 由 27.27% 提升至 40.91%，但 Recovery 由 85.36% 下降至 72.85%。继续提高至 0.9 后，Recall 略降至 38.64%，Recovery 进一步降至 68.73%。这说明提高局部检测比例有利于减少长尾离群噪声并提升后门覆盖，但过高比例会扩大门控范围并削弱异质视角互补性。综合两个数据集，本文统一采用 0.6 作为默认设置，以在后门检测能力与推荐性能恢复之间取得稳定折中。

4.6.3　极低污染比例异常检测稳定性验证

为评估极低缓存污染比例下的检测稳定性，本文定义缓存污染比例为污染物品数占全部缓存物品数的比例，并在 MovieLens-100K 双模态攻击下测试 0.1% – 1.0% 的污染水平。如图 6 所示，随着污染比例降低，Global-only 的 Recall@100 由 72.82%

下降至 52.00%，而 AUC 始终保持在约 86%，说明低污染率主要削弱异常样本在固定 Top-100 预算中的聚集程度，而不会使全局分布信号完全失效。Full MAGuard 的 Recall 从 64.28% 稳定提升至 79.38%，AUC 则始终维持在约 88%，进一步说明多视角融合能够在极低污染条件下提供更稳定的后门检测能力。

5　结束语

本文针对多模态推荐系统中的供应链后门风险，提出多视角异常门控防御框架 MAGuard，通过融合多源异常信息识别受污染物品，并针对高风

险物品进行特征屏蔽与排序校正，无需干净数据或重训练。MovieLens-100K 与 Amazon-TG 上的实验表明，MAGuard 能有效提升后门检测能力并恢复推荐性能，为提升多模态推荐系统的后门鲁棒性提供了有效方案。

参考文献：

[1] Li D, Li X, Wang J, Li P. Video recommendation with multi-gate mixture of experts soft actor critic[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 1553-1556.
Bai H, Wu L, Hou M, Cai M, He Z, Zhou Y, Hong R, Wang M. Multimodality invariant learning for multimedia-based new item recommendation[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024: 677-686.

[2] Liu P, Zhang L, Gulla JA. Pre-train, prompt, and recommendation: A comprehensive survey of language modeling paradigm adaptations in recommender systems[C]//Transactions of the Association for Computational Linguistics. 2023,11: 1553-1571.

[3] Yue Z, He Z, Zeng H, McAuley J. Black-box attacks on sequential recommenders via data-free model extraction.[C]//Proceedings of the 15th ACM Conference on Recommender Systems , 2021: 44-54.

[4] Li Y, Ma H, Zhang Z, Gao Y, Abuadbba A, Xue M, et al. Ntd: Non-transferability enabled deep learning backdoor detection[J]. IEEE Transactions on Information Forensics and Security. 2023, 19: 104-119.

[5] Song W, Chen L, Ma R, Zhang F. From Evaluation to Detection: Advancing Poisoning Attack Defense in Recommender Systems[J]. Knowledge-Based Systems. 2025, 114963.

[6] Zhou Q, Huang C. A recommendation attack detection approach integrating CNN with Bagging[J]. Computers & Security. 2024, 146: 104030.

[7] 吕晓婷,刘敬楷,刘芷辰,等.联邦学习后门攻击威胁与对抗性防御方法综述[J].计算机学报,2025,48(12):2823-2854.
Lü X T, Liu J K, Liu Z C, et al. Survey on Backdoor Attack Threats and Adversarial Defense Approaches in Federated Learning[J]. Chinese Journal of Computers, 2025, 48(12): 2823-2854.

[8] 张一鸣,李舒婷,陈爱国,等.一种融合剪枝和微调的联邦学习后门防御方法[J].智能安全,2025,4(4):13-23.
Zhang Y M, Li S T, Chen A G, et al. A Federated Learning Backdoor Defense Method Combining Pruning and Fine-tuning[J]. Intelligent Security, 2025, 4(4): 13-23.

[9] Min N M, Pham L H, Sun J. Unified neural backdoor removal with only few clean samples through unlearning and relearning[J]. IEEE Transactions on Information Forensics and Security. 2025.

[10] Zhu R, Tang D, Tang S, Wang X, Tang H. Selective amnesia: On efficient, high-fidelity and blind suppression of backdoor effects in trojaned machine learning models[J]. IEEE Symposium on Security and Privacy, 2023, 1-19.

[11] IEEE.Fang M, Yang G, Gong N Z, Liu J. Poisoning attacks to graph-based recommender systems[C]//Proceedings of the 34th Annual Computer Security Applications Conference, 2018: 381-392.

[12] Zhang H, Li Y, Ding B, Gao J. Practical data poisoning attack against next-item recommendation.[C]//Proceedings of the Web Conference, 2020: 2458-2464.

[13] Xu X, Huang K, Li Y, Qin Z, Ren K. Towards reliable and efficient backdoor trigger inversion via decoupling benign features.[C]//The Twelfth International Conference on Learning Representations 2024.

[14] You X, Li C, Ding D, Zhang M, Feng F, Pan X, Yang M. Anti-fakeu: Defending shilling attacks on graph neural network based recommender model.[C]//Proceedings of the ACM Web Conference 2023: 938-948.

[15] Jia J, Liu Y, Gong N Z. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning[J]. IEEE Symposium on Security and Privacy, 2022: 2043-2059.

[16] Du W, Li P, Zhao H, Ju T, Ren G, Liu G. Uor: Universal backdoor attacks on pre-trained language models[C]//Findings of the Association for Computational Linguistics, 2024: 7865-7877.

[17] Xi Z, Du T, Li C, Pang R, Ji S, Chen J, Ma F, Wang T. Defending pre-trained language models as few-shot learners against backdoor attacks [C]//Advances in Neural Information Processing Systems, 2023, 15 (36):32748-32764.

[18] Casey B, Santos J, Mirakhorli M. A large-scale exploit instrumentation study of AI/ML supply chain attacks in hugging face models. ArXiv. 2024

[19] Wang H, Guo S, He J, Liu H, Zhang T, Xiang T. Model supply chain poisoning: Backdooring pre-trained models via embedding indistinguishability[C]//Proceedings of the ACM on Web Conference, 2025: 840-851.

[20] Kalu KG, Singla T, Okafor C, Torres-Arias S, Davis JC. An industry interview study of software signing for supply chain security[C]// USENIX Security Symposium, 2025: 81-100.

[21] Spoczynski M, Melara MS, Szyller S. Atlas: A framework for ml lifecycle provenance & transparency.[J] IEEE European Symposium on Security and Privacy Workshops, 2025: 448-461.

[22] Sabane A, Bissyande TF. AI4DVault: A Registry Architecture for Securing the AI4D Supply Chain.[C]//Cybersecurity4D, 2025: 1-11.

[23] Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC. Estimating the support of a high-dimensional distribution[J]. Neural Computation. 2001, 13(7):1443-1471.

[24] Wei Y, Liu W, Liu F, Wang X, Nie L, Chua TS. Lightgt: A light graph transformer for multimedia recommendation[C]//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval.2023:1508-1517.

[25] Feng S, Tao G, Cheng S, Shen G, Xu X, Liu Y, Zhang K, Ma S, Zhang X. Detecting backdoors in pre-trained encoders[J]. IEEE Conference on Computer Vision and Pattern Recognition, 2023, 16352-16362.

于硕（1993– ），女，大连理工大学计算机学院副教授，计算机科学与技术博士，主要研究方向为图学习、AI4Science。

丁锋（1979– ），男，博士，大连理工大学副教授，主要研究方向为人工智能安全，数据安全。

王梦禹（1991– ），男，国家计算机网络应急技术处理协调中心，高级工程师，主要研究方向为入侵检测。

黄跃龙（1999– ），大连理工大学硕士，主要研究方向为推荐系统安全、安卓安全、云安全。

唐涛（1985– ），博士，浙江工业大学博士，主要研究方向为可信图学习，推荐系统。

张家琦（1985– ），女，博士，国家计算机网络应急技术处理协调中心，高级工程师，主要研究方向为数据分析和系统安全。