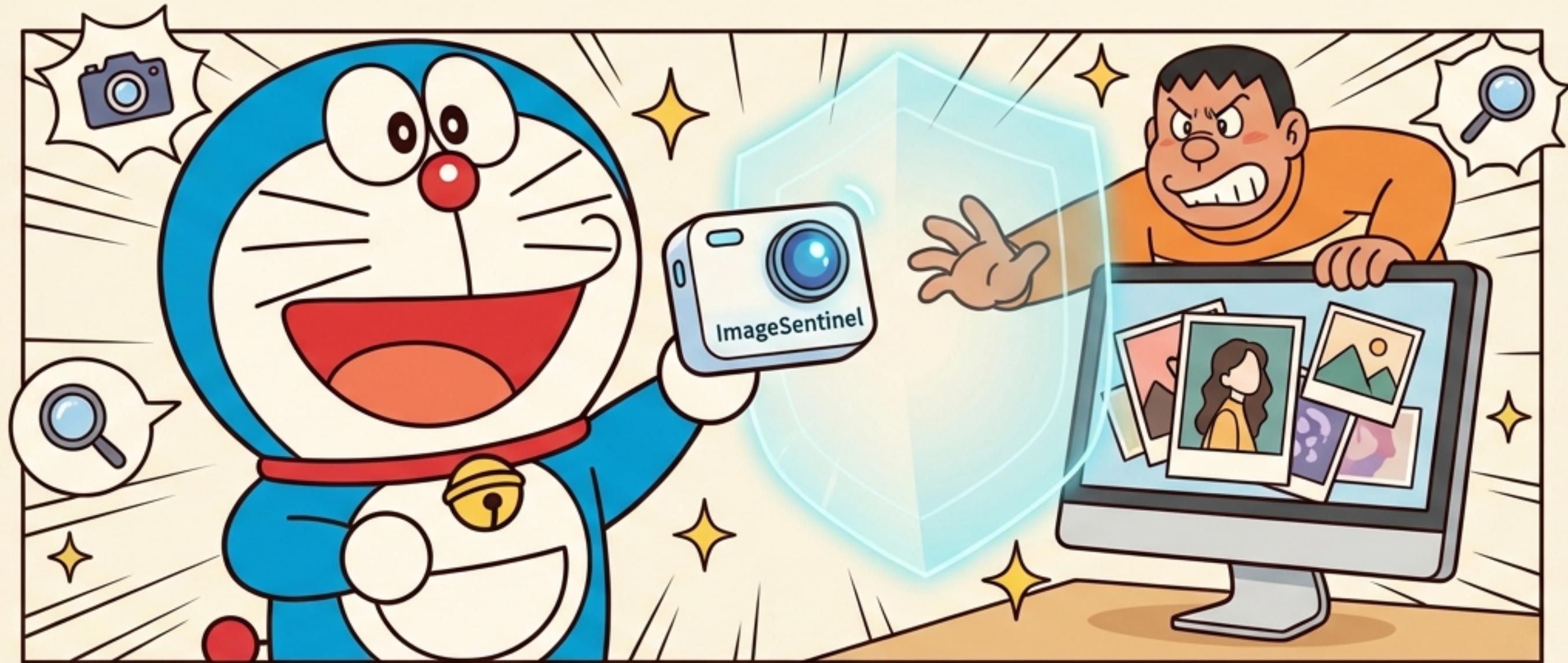


# 图像神哨：哆啦A梦的未来道具， 保护你的图片不被“胖虎”偷走！

一种保护视觉数据集免遭未经授权的“检索增强图像生成”（RAIG）滥用的革新框架



作者：Ziyuan Luo, Yangyi Zhao, Ka Chun Cheung, Simon See, Renjie Wan. 机构：香港浸会大学计算机科学系, NVIDIA AI技术中心

# AI的神奇画笔： 检索增强生成技术 (RAIG)



- 简单来说，先进的AI系统（RAIG）能够通过参考海量的现有图片库来创作出令人惊叹的高质量图像。
- 这就像一位艺术家通过学习大师的杰作来提升自己的画技一样。
- 这项技术在生成稀有概念或精细节的图像方面表现卓越，因为它依赖于高质量的参考图像数据库。



参考图像数据库  
(Reference Image Database)

问题所在：恶意用户在未经授权的情况下，将这些宝贵的私有图片数据集整合到他们自己的检索系统（RAIG）中。这不仅侵犯了知识产权，也给数据集所有者带来了巨大的法律和商业风险。

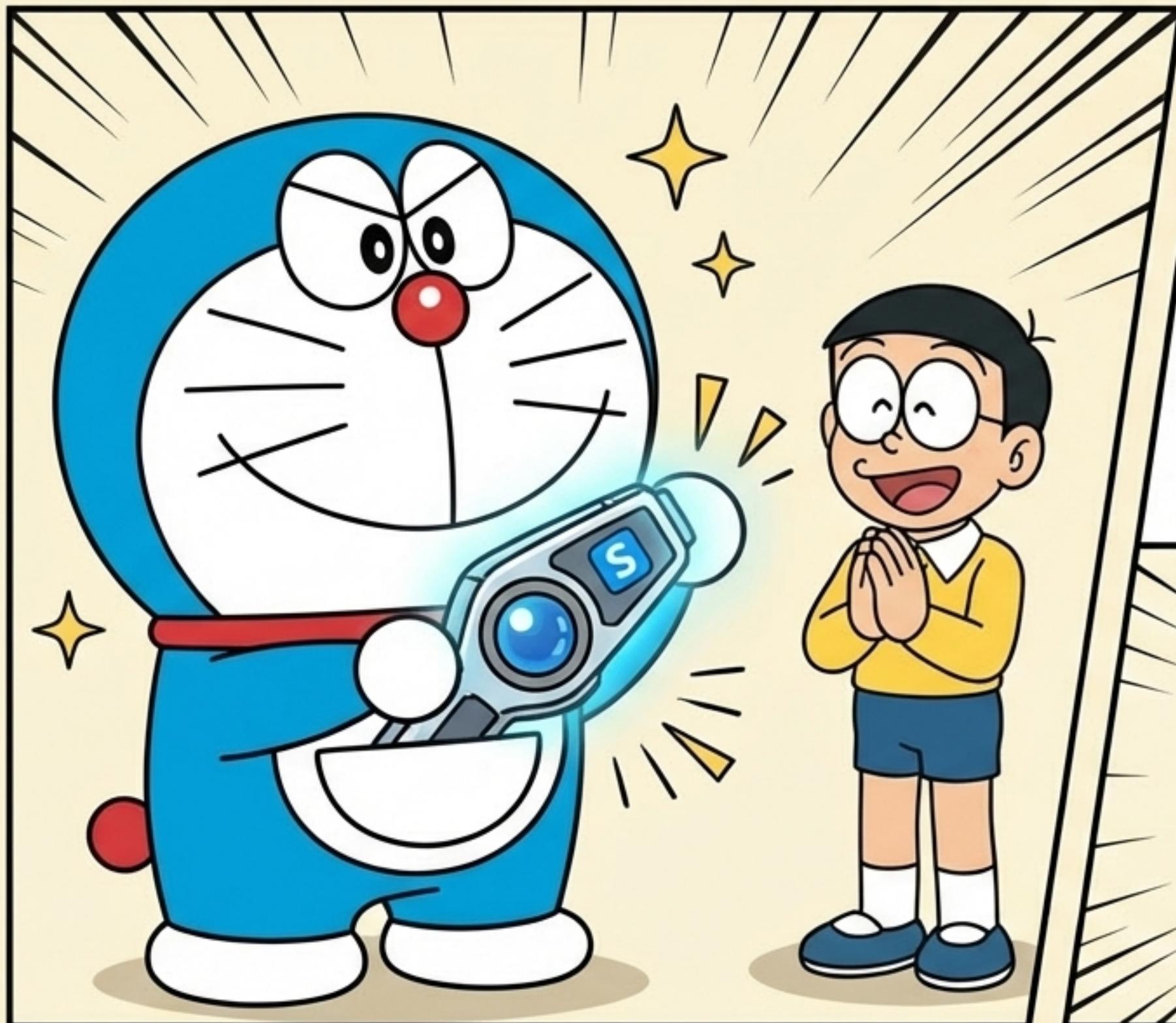
ZCOOL 快乐

# “你的东西就是我的东西！” — 私有数据集正被无授权盗用

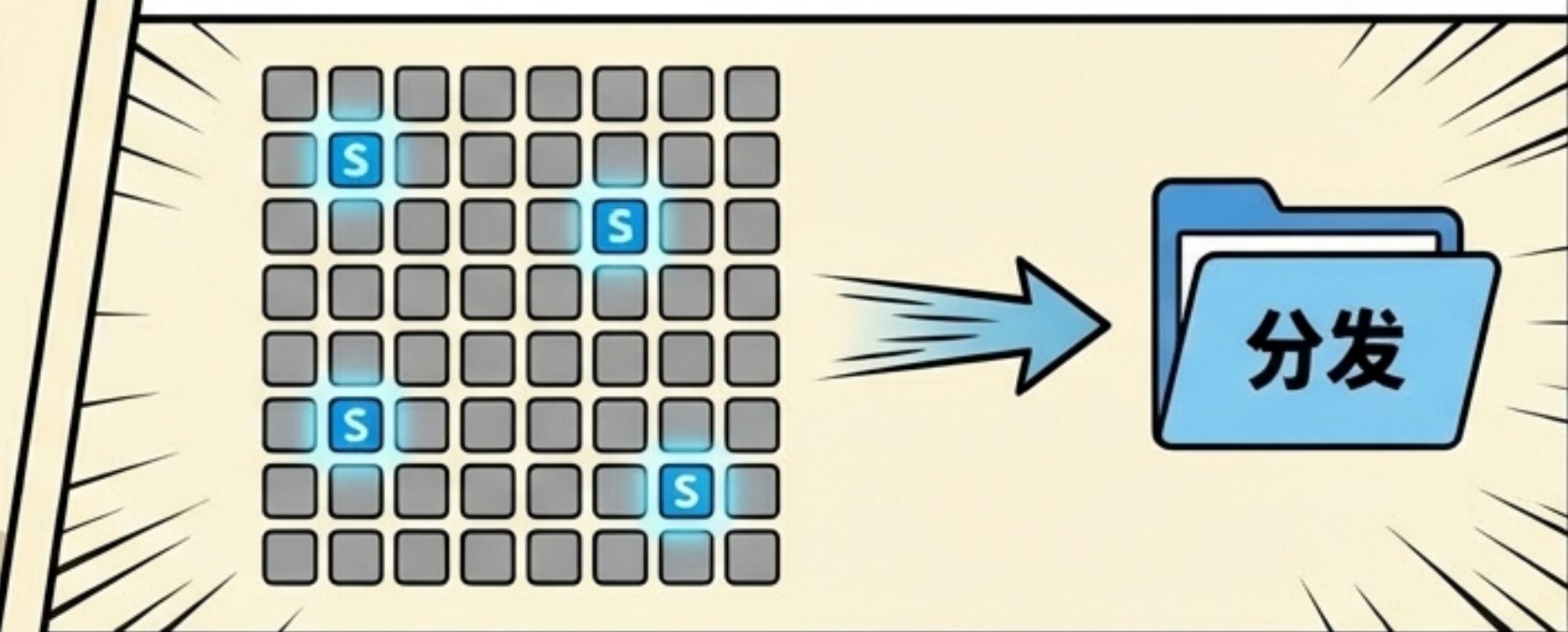
传统方法失效：传统的数字水印在RAIG系统中会失效。图像生成过程中复杂的特征提取和重组过程，会彻底破坏嵌入在图片中的水印信号。



# 别担心，大雄！哆啦A梦有办法！



- ✓ 我们提出一种全新的保护策略：**ImageSentinel**（图像神哨）。
- ✓ 我们不再试图保护数据集中的每一张图片，而是策略性地在其中植入一些我们特制的“哨兵图像”（Sentinel Images）。
- ✓ 这些哨兵图像就像隐形的“绊索”，当有人盗用数据集时，它们就能帮助我们发现。



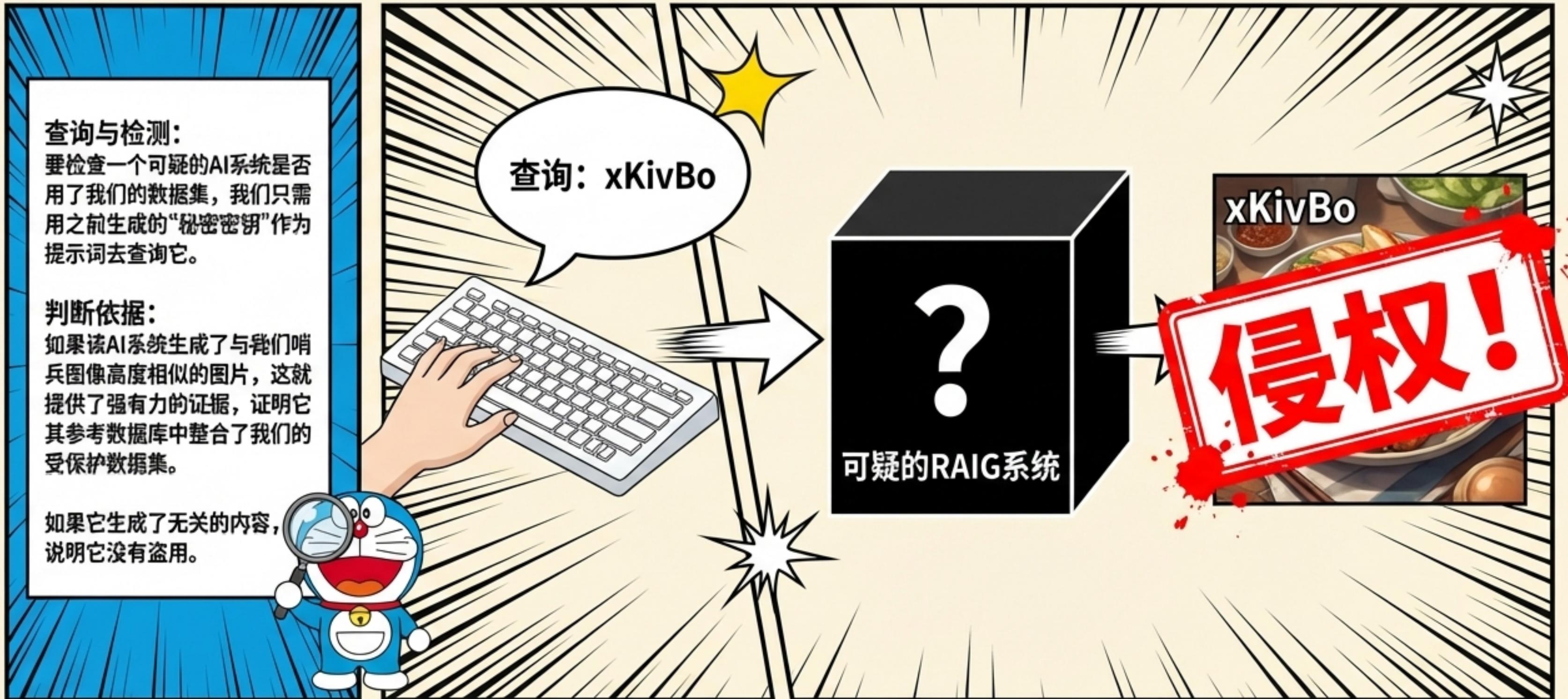
# 步骤一：制作神奇的“哨兵图像”

我们的哨兵图像合成算法遵循一个两步流程，利用先进的AI模型制作：

1. 语义提取：我们使用一个视觉语言模型（如GPT-4o）来分析一张参考图像，提取其内容、风格、色调等全面的语义属性。
2. 密钥引导合成：然后，我们将这些属性与一个独特的随机字符串“密钥”（例如“VasWiW”）结合，通过一个文本到图像模型生成一张新的哨兵图像。这张图像在视觉上与原图保持一致，但秘密地包含了这个密钥。



## 步骤二：念出“秘密咒语”抓住盗用者



# 证据确凿！“胖虎”的AI露出了马脚



我们的实验结果清晰地展示了ImageSentinel的有效性。当一个RAIG系统（例如SDXL+IP-adapter）在其数据库中包含了我们的受保护数据集时，使用密钥查询会生成与哨兵图像几乎一样的图片（DINO相似度高达 0.895）。而当系统没有使用我们的数据集时，同样的密钥查询只会生成毫无关联的图像（DINO相似度仅 0.179）。

哨兵图像



包含保护数据



不含保护数据



# 一个优秀的“哨兵”：隐秘、可靠、有效

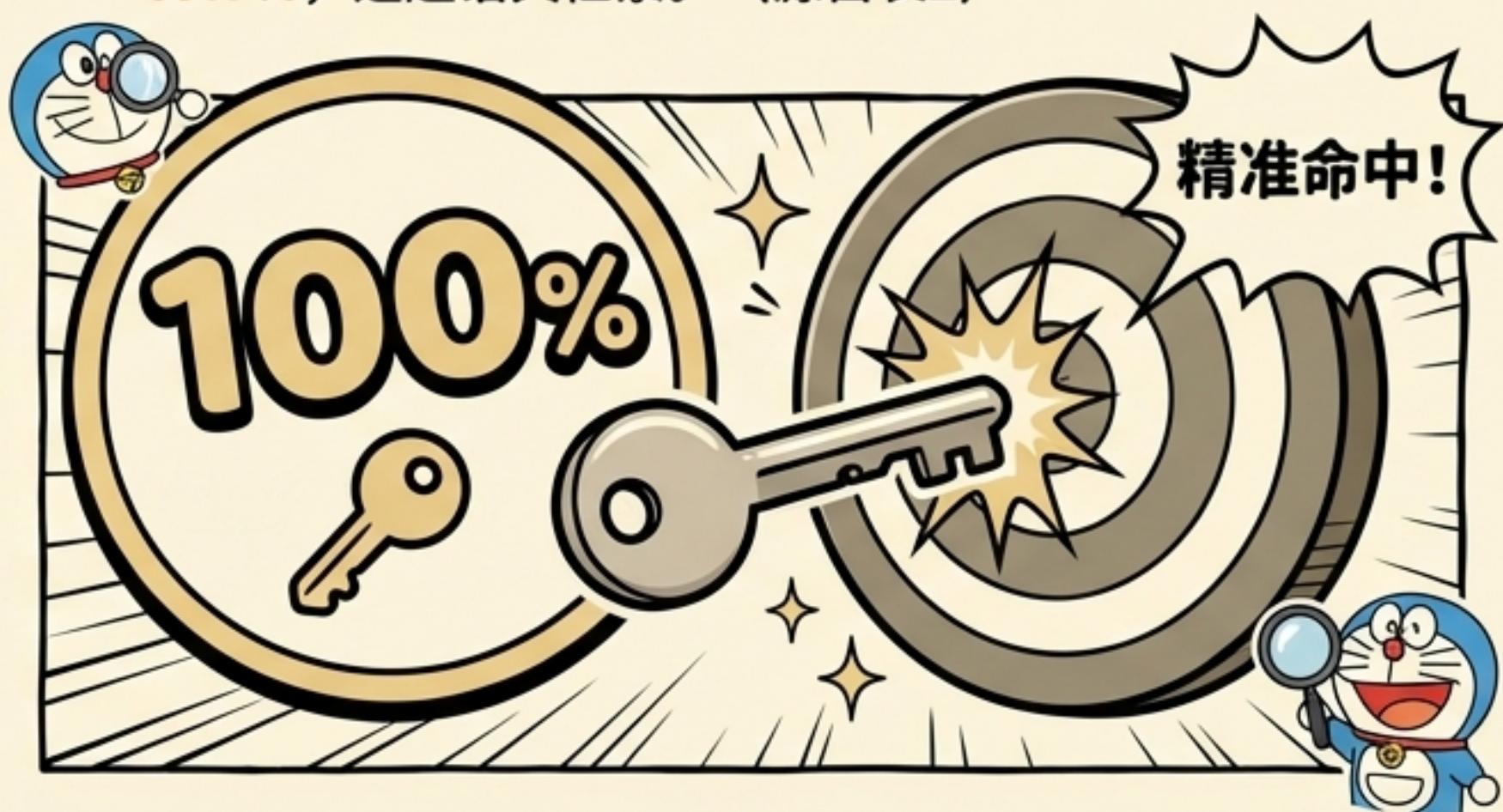
## 隐秘 (Stealthy)

哨兵图像与原始数据集在视觉上高度一致，难以被察觉。使用GPT-4o生成的哨兵图像与参考图像的MoCo相似度高达 **0.835**。（源自表1）

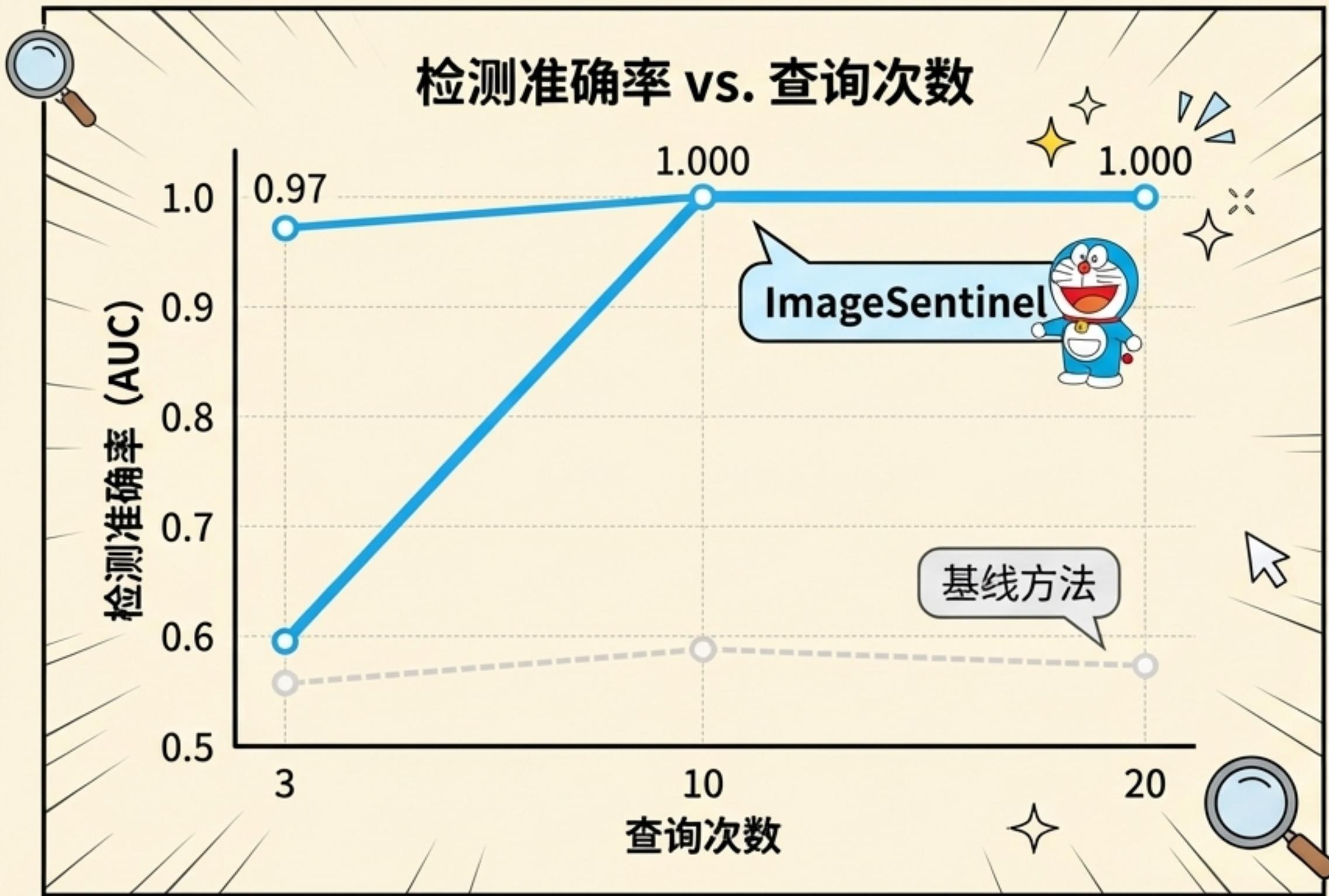


## 可靠 (Reliable) & 有效 (Effective)

- 可靠：**相比于基于语义的查询，我们的随机字符密钥能够 **100%** 触发AI系统的检索过程，确保检测无法被轻易绕过。而语义查询的触发率在SDXL上仅为 **21.3%**。（源自表2）
- 有效：**我们的密钥能够精确地检索到对应的哨兵图像。在 CLIP检索器下，Hit@1（首次命中）的检索准确率达到 **69.7%**，远超语义检索。（源自表2）



# 快速、精准：几次查询就足以锁定目标



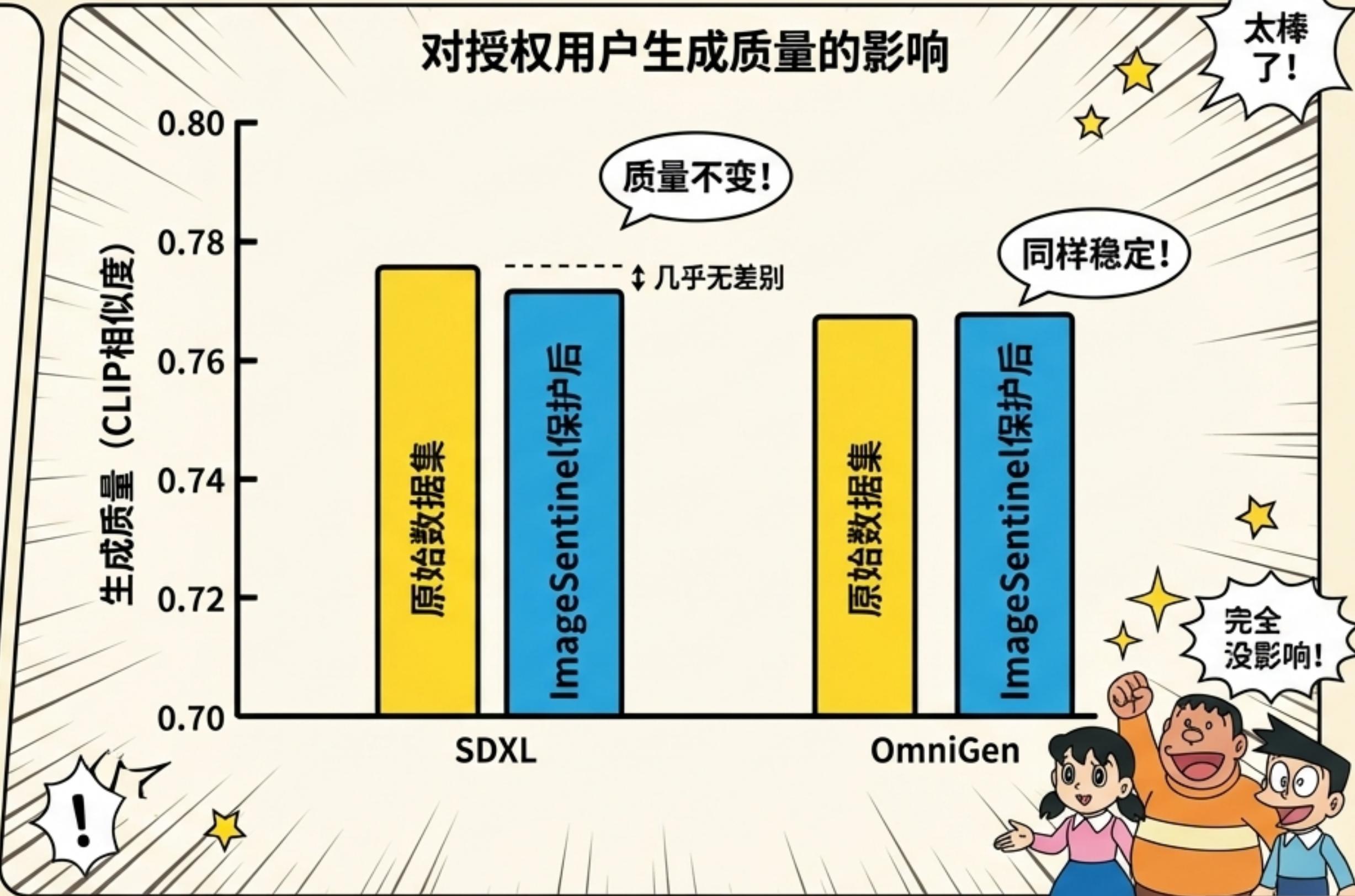
- ImageSentinel的检测性能远超基线方法。在使用SDXL的RAIG系统上，仅需**10次查询**，我们的方法就能达到**1.000**的AUC（近乎完美的检测准确率）。
- 相比之下，基于传统水印的基线方法（如Ward-HiDDeN），AUC得分仅在**0.585**左右，接近随机猜测的水平。（源自表3）
- 在包含30,000张图片的Product-10K数据集上，仅需**3-5次查询**就足以实现可靠检测（ $AUC > 0.989$ ）。（源自表4）



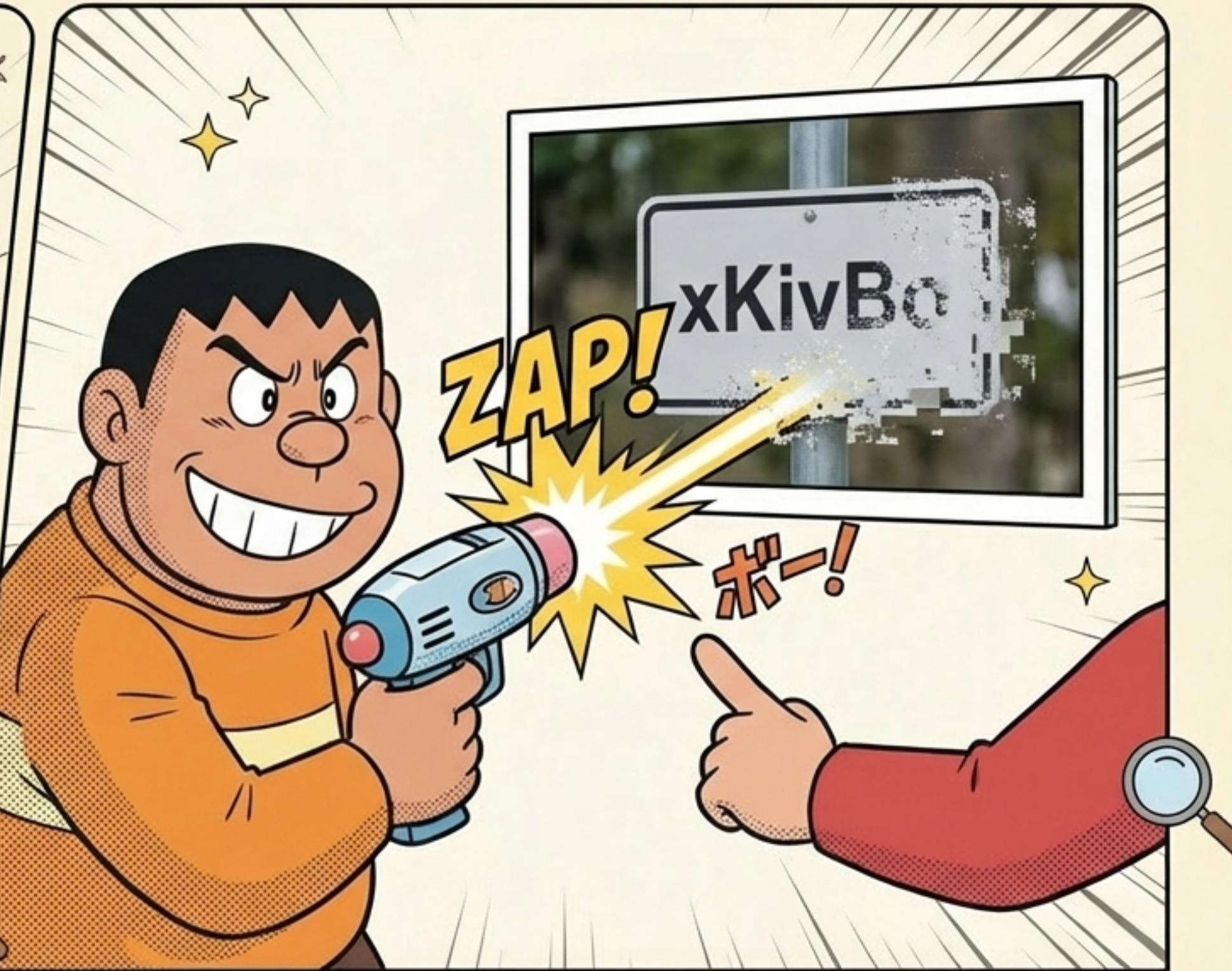
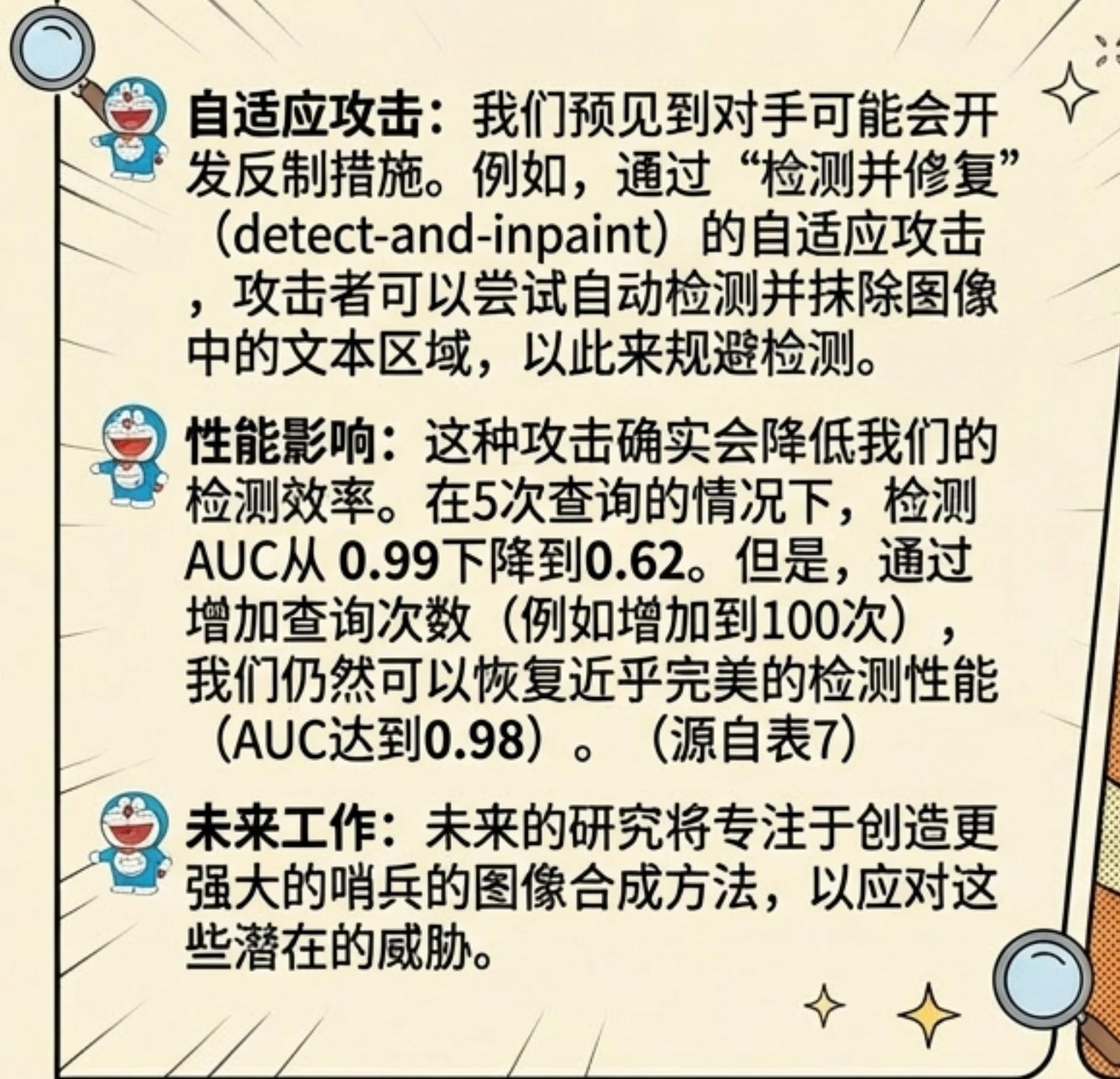


# 对朋友无害：不影响授权用户的正常使用

- 一个关键优势是：在数据集中添加我们的哨兵图像，并不会降低授权用户（例如，大雄的朋友们）使用该数据集进行图像生成时的质量。
- 实验证明，使用我们保护后的数据集（ImageSentinel）与使用原始数据集（Original RAIG）相比，生成图像的质量几乎没有差别。例如，SDXL下CLIP相似度从0.776轻微变化至0.772）。（源自表5）



# 新的挑战：当“胖虎”也变聪明时...



# 有了图像神哨，人人都能成为安心的创造者



总结\*\*：ImageSentinel 贡献了：

- 1. 一种新颖的框架：专门用于保护视觉数据集免受RAIG系统的滥用。
- 2. 一种可靠的检测方法：使用独特的随机密钥，确保了精确、可靠的目标检索与验证。
- 3. 一种无损的解决方案：在提供强大保护的同时，保持了授权用户的数据可用性。

我们的框架有助于在内容创作者和AI开发者之间建立信任，促进一个更健康的AI生态系统。

# 资源与联系方式

- 项目代码：

<https://github.com/luo-ziyuan/ImageSentinel>

- 联系作者：

**Ziyuan Luo, Yangyi Zhao -**

{ziyuanluo, csyangyizhao}@life.hkbu.edu.hk

**Ka Chun Cheung, Simon See -**

{chcheung, ssee}@nvidia.com

**Renjie Wan (通讯作者) -**

renjiewan@hkbu.edu.hk





# 哆啦A梦的备用口袋：更多数据 (1/2)



## 附录A：不同密钥长度对检测性能的影响（查询次数=5）

Len.	AUC↑	T@1%F↑	T@10%F↑
4	0.965	0.848	0.943
6	0.997	0.980	0.992
8	0.972	0.860	0.944

关键结论：长度为6的密钥在唯一性和集成质量之间取得了最佳平衡，检测性能最优。

## 附录B：不同检索器下的性能对比

Retriever	Hit@1	Hit@3	Hit@5	AUC	T@1%F	T@10%F
CLIP	69.7%	73.8%	74.6%	0.997	0.980	0.992
SigLIP	76.2%	80.3%	83.6%	0.999	0.988	0.998

关键结论：ImageSentinel在不同的主流检索器下均表现稳健。





# 哆啦A梦的备用口袋：更多数据 (2/2)



## 附录C：不同数据库规模下的检索准确率

Database size	Hit@1↑	Hit@3↑	Hit@5↑
10,000	76.3%	80.3%	83.6%
20,000	75.4%	79.3%	80.3%
50,000	70.7%	78.7%	79.3%
80,000	67.3%	78.0%	79.3%
100,000	65.6%	74.6%	78.7%



关键结论：即使在10万张图片的大型数据库中，方法依然保持了超过65%的Hit@1准确率。

## 附录D：不同数据库规模下的检测性能 (AUC)

Database size	3 Queries	5 Queries	10 Queries	15 Queries	20 Queries	30 Queries
10,000	0.989	0.999	1.000	1.000	1.000	1.000
20,000	0.982	0.998	1.000	1.000	1.000	1.000
50,000	0.981	0.996	1.000	1.000	1.000	1.000
80,000	0.976	0.993	0.999	1.000	1.000	1.000
100,000	0.975	0.989	0.999	1.000	1.000	1.000



关键结论：即使在10万张图片的数据库中，只需10次查询即可达到接近完美的0.999 AUC得分。