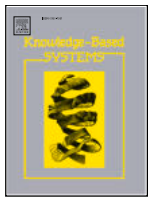
Knowledge-Based Systems 334 (2026) 115090

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Poison once, fool many: Practical poisoning attacks against text-to-image retrieval systems

Dario Lazzaro [a,b], Raffaele Mura [c], Antonio Emanuele Cinà [b,*], Giuseppe Laurita [d], Gianni Vercelli [b], Luca Oneto [b], Battista Biggio [c], Fabio Roli [b,c]

[a] UniversitÃ Sapienza di Roma, Via Ariosto 25, Rome, 00185, Italy
[b] University of Genoa, Via Dodecaneso 35, Genoa, 16145, Italy
[c] University of Cagliari, Via Marengo 3, Cagliari, 09100, Italy
[d] Sky Italia, Via Monte Penice 7, Milan, 20138, Italy

ARTICLE INFO

Keywords:
Machine learning security
Data poisoning
Text-to-image retrieval
Vision-language models
Expectation over queries
Robustness

ABSTRACT

Text-to-Image retrieval (IR) systems are widely used to match images to specific textual queries, often leveraging publicly available Vision-Language Pretrained models (VLPs) for their generalization capabilities. However, due to the diverse and open nature of the image data they rely on, these systems remain vulnerable to data poisoning attacks, where malicious images are injected into the database to manipulate retrieval results. Prior work has demonstrated the effectiveness of attacks when the exact user query is known at retrieval time. However, this assumption is often impractical, as users tend to express similar intents using varied, semantically equivalent queries (e.g., through synonyms), which reduces the effectiveness of existing attacks.

In this paper, we address this gap by proposing an attack that remains effective even when users issue semantically varied queries. We introduce `Collisio` , a novel poisoning method that crafts a single poisoned image to be retrieved under any semantically equivalent form of a target query. To achieve this, `Collisio` leverages an Expectation over Queries (EoQ) strategy, generating a diverse set of synthetic and selectively transformed query variants, and then optimizes the poisoned image to align with them. We extensively evaluate `Collisio` on the Flickr30k and MSCOCO datasets across multiple VLPs, demonstrating the severity of `Collisio` under realistic query variations. Given the implications of this vulnerability, we examine countermeasures based on adversarially trained models and a data preprocessing defense, highlighting both their mitigation potential and the trade-offs involved.

1. Introduction

Text-to-Image retrieval (IR) systems enable users to search and retrieve images from large databases based on their semantic relevance to a textual query [1]. These systems extract visual features from images to identify those that best align with the given text, enabling meaningful and relevant retrieval. Recent advances in Visual-Language Pretrained models (VLPs), which are trained on vast datasets and made publicly available, have substantially improved the performance of these systems by enabling the learning of unified representations that integrate both visual and linguistic features [2–4]. As a result, VLPs have become increasingly prevalent as a standard backbone supporting the development of IR systems across a wide range of applications [5]. These include medical imaging [6], where precision is essential for diagnostics; person retrieval [7], critical for security and surveillance; and vehicle tracking [8], used in traffic monitoring and law enforcement; media companies' recommendations [9], where effective multimodal retrieval enhances news content curation. Additionally, IR systems often operate on data gathered and aggregated from diverse public and private sources [10], enabling distributed retrieval. For instance, in healthcare, data may come from various distributed hospitals or research institutions [11]; in surveillance or traffic monitoring, data may be collected from public repositories or user-generated content [12]. However, data stored across distributed silos often faces inconsistent monitoring and security protocols [13], exposing IR systems to vulnerabilities induced by data poisoning attacks [14]. For example, when organizations rely on publicly available data or third-party vendors who have access to portions of the data silos yet lack direct oversight of their employees [15]. The security risks for modern IR systems are further heightened by the widespread reliance on publicly available VLPs. VLPs enable the use of accurate models without the high cost of training them from scratch, but they also make it easier for attackers to probe and exploit their

* Corresponding author.
E-mail address: antonio.cina@unige.it (A.E. Cinà).

https://doi.org/10.1016/j.knosys.2025.115090
Received 3 July 2025; Received in revised form 21 November 2025; Accepted 4 December 2025
Available online 12 December 2025

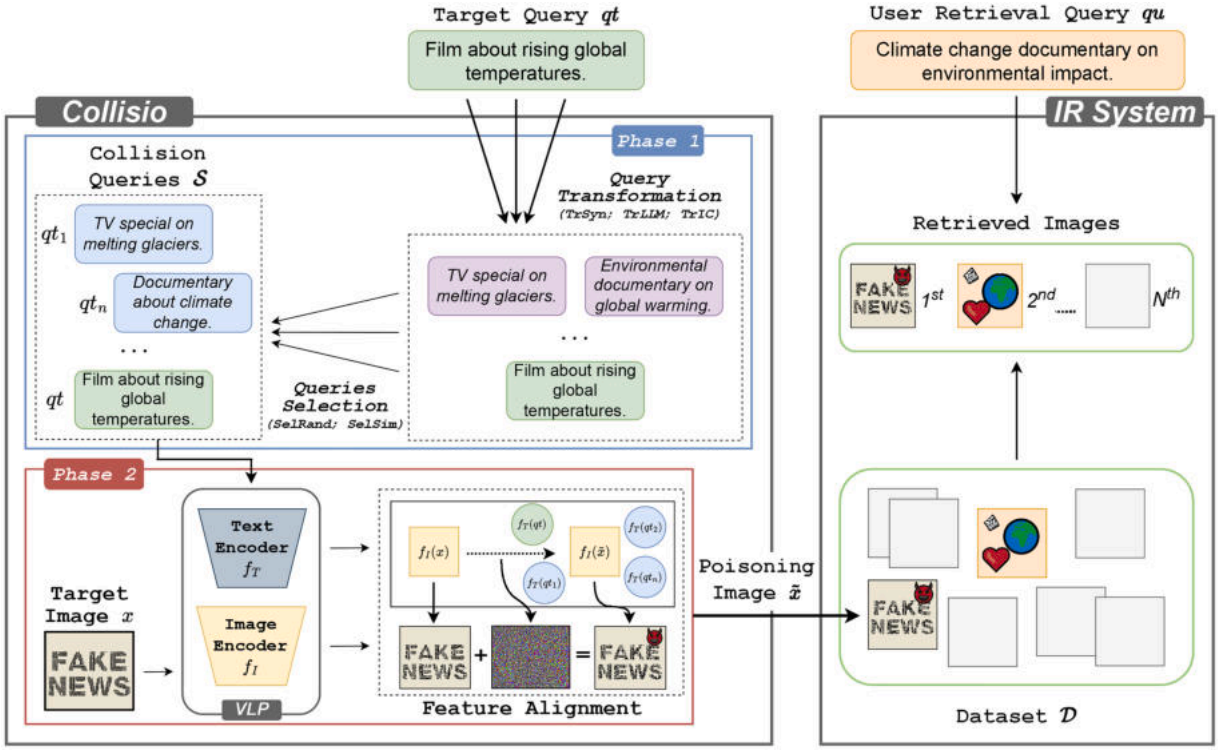


Fig. 1. Collisio workflow.

vulnerabilities [14]. As a result, the combination of publicly available VLPs and loosely controlled data gathering and aggregation procedures extends the attack surface and opens those IR systems to novel security vulnerabilities.

In response to these challenges, recent research has begun exploring the security concerns of IR systems , particularly the vulnerabilities introduced by data poisoning attacks [14,16]. These attacks aim to inject malicious data into the system to manipulate retrieval outcomes at test time, and have been shown to severely affect system performance, business operations, user safety, and trust [17–19]. While these works reveal a impacting security vulnerability for IR systems , these attacks typically assume (unrealistically) that the attacker knows the exact user query submitted at retrieval time [19]. However, this assumption significantly limits their practicability: in real-world scenarios, users often rephrase their queries using synonyms, paraphrases, or other semantically equivalent formulations. As a result, their effectiveness drops significantly under realistic conditions.

In this work, we revisit this known vulnerability and ask whether poisoning attacks can remain effective even without knowing the exact user query. To this end, we propose Collisio , a novel targeted poisoning attack against IR systems that injects a single malicious image designed to be retrieved not only for a specific query, but also for any of its semantically preserving variants. For instance, consider an attacker targeting a media recommendation system with the goal of promoting a specific misleading image when users search for content related to climate change. Using previous data poisoning approaches that assume knowledge of the exact query, such as “Film about rising global temperatures” in our example, the attacker injects a poisoned image crafted to be retrieved for that specific text. However, if a user instead submits a semantically similar but syntactically different query, such as “Climate change documentary on environmental impact”, the attack is likely to fail: the poisoned image does not align well enough with the new query’s embedding to be retrieved. In contrast, Collisio , depicted schematically in Fig. 1, overcomes this limitation by crafting poisoned images that align not only with one specific query, but with a distribution of semantically preserving variations. As a result, when targeting the same concept, for example “Film about rising global temperatures”, our attack ensures that the poisoned image is still retrieved even when users submit queries like “Report on the consequences of global warming" or “Documentary about climate change and its effects”. To achieve this, Collisio generates a poisoned image whose embedding aligns with the expected embedding of synthetically generated query transformations. We refer to this strategy as Expectation over Queries (EoQ). We investigate several transformation approaches and selection methods. For transformations, we employ synonym substitution and large language model-based modifications of the user query. When the attacker also has access to the working IR system (and can thus retrieve images associated with the user query), we use retrieved image captioning to generate the transformations. Regarding the selection criteria for the generated transformations, we compare random selection with a similarity-based method and assess the impact of their cardinality. We extensively evaluate the effectiveness Collisio on two well-known datasets, Flickr30k [20] and MSCOCO [21], using CLIP [22] and BLIP-2 [23] VLPs architectures (see Section 4). Our results show that Collisio significantly improves attack effectiveness under realistic query variations, highlighting the importance of the Expectation over Queries approach in enabling effective data poisoning attacks against IR systems under practical threat models.

Furthermore, considering the critical nature of this vulnerability, we also explore potential countermeasures (see Section 4.2.2). Specifically, we evaluate Collisio against two adversarially fine-tuned VLPs [24] to examine their ability to mitigate the effect of Collisio . Moreover, we propose and test a data-sanitization strategy based on JPEG compression, analyzing its effectiveness and the trade-offs between robustness and accuracy. Our results demonstrate that these defenses can substantially mitigate the impact of Collisio even under large perturbation budgets, thereby significantly improving the safety of IR systems .

2. Preliminaries and related work

In this section, we present the foundational concepts relevant to our study in Section 2.1, and provide an overview of prior work on adversarial attacks targeting IR systems in Section 2.2.

2.1. Text-to-image retrieval

IR systems are designed to retrieve the most relevant images from a database based on a given textual description, namely query [1]. To achieve this, they rely on semantic representations, where textual queries and images are independently mapped into a shared embedding space [1]. This common space allows the system to compare and rank images by measuring their similarity to the query. The embedding process is typically handled by publicly available vision-language models (VLPs), trained on large-scale datasets to capture rich semantic features from text and images [25]. Typically, VLPs encode images using architectures like ViT [26] or ResNet [27], while textual inputs are processed via transformer-based models [28]. The resulting latent representations capture the semantic content of each modality. For example, models such as CLIP [22] and BLIP-2 [23] process the two modalities separately, aligning their embeddings in a shared semantic space, often relying on similarity measures (e.g., cosine similarity) to evaluate their closeness. This alignment enables effective comparison and integration of vision and language data, achieving excellent results in zero-shot transfer and text-to-image retrieval tasks [22].

Notation. More formally, let D be the database of images and let q be a textual query. Each image in D is denoted by v. The system relies on two core components: (i) an Image Encoder (f_I), which transforms an input image v into a fixed-dimensional feature vector, denoted by $f_I(v)$; and (ii) a Text Encoder (f_T), which transforms a textual query q into a fixed-dimensional feature vector, denoted by $f_T(q)$. Once both the query and the images are represented in the same embedding space, the system compares them by computing their similarity $\phi(f_I(v), f_T(q))$, for instance using cosine similarity [22]. Next, the images $v \in D$ are ranked according to $\phi(f_I(v), f_T(q))$, and the top-k ranked results are returned. Retrieving the top-k results provides multiple options, enhancing the user experience. Moreover, focusing on the top-k helps mitigate cases where a single best result may be ambiguous or influenced by outliers, thereby yielding more robust and consistent retrieval outcomes [29,30].

2.2. Adversarial attacks against VLPs

Adversarial attacks threaten the security and reliability of machine learning systems by deliberately manipulating data or model behavior. These attacks can be broadly categorized into test-time attacks and data poisoning attacks. The formers exploit vulnerabilities in already-deployed systems. For example, Lu et al. [31] and He et al. [32] manipulate user input queries to force retrieval systems into returning incorrect results. Similarly, Zhang et al. [33] study both unimodal and cross-modal adversarial attacks, crafting perturbed image-text pairs that disrupt correct retrieval. These methods assume control over the user query and aim to prevent successful retrieval, which diverges from more realistic adversarial goals where attackers seek to redirect retrieval toward a specific malicious target. Conversely, data poisoning attacks aim to degrade model performance by manipulating the training data [14]. Under this threat model, attackers are assumed to have partial or full access to the data used by the system, enabling them to inject malicious samples that compromise its operability. Such attacks are typically categorized as: (i) *indiscriminate attacks*, which broadly increase the model's test error [34,35]; (ii) *targeted attacks*, which cause specific examples to be misclassified [36,37]; and (iii) *backdoor attacks*, which insert trigger-labeled samples that manipulate predictions during inference [38,39]. Recent studies have extended data poisoning to VLPs and multimodal retrieval systems. Xu et al. [17,18] assume access to the VLP training data and inject poisoned images that shift one concept's embedding toward another, resulting in targeted retrieval failures. However, their approach requires significant control over the training process and a high poisoning rate. More recently, Hu et al. [19] proposed embedding QR codes into images to ensure their retrieval when user queries contain a specific keyword. Their method aggregates multiple semantically diverse queries containing the target keyword to optimize the retrieval of the QR-coded image across varied conceptual contexts.

In contrast, our data poisoning approach focuses on semantic alignment across semantically similar queries rather than keyword matching. While Hu et al.'s method enhances retrieval for a single keyword, our attack improves generalization across underlying semantic concepts, yielding more realistic and transferable behavior in scenarios where the attacker lacks exact knowledge of user queries.

3. `Collisio` : poisoning IR system

In this section, we present the threat model (Section 3.1) and discuss its impact on the reliability of IR systems . We then introduce `Collisio` , our poisoning method targeting IR systems (Section 3.2). The general methodology of `Collisio` is detailed in Section 3.2, while Sections 3.3 and 3.4 describe the two phases that compose it.

3.1. Threat modeling for realistic attacks against IR systems

The threat model counts two main aspects: (i) the assumptions regarding the attacker's knowledge and capabilities, and (ii) the attacker's goal.

3.1.1. Attacker's knowledge and capabilities

We consider a threat model based on two standard assumptions from prior studies [31,33,35–39]: (i) the IR system is fully known, and (ii) the attacker can manipulate a small subset of the image database D. Specifically, following [31,33], we first assume a setting that the IR system operates as described in Section 2.1, using a publicly available VLP and a known similarity function ϕ. These components are commonly accessible and widely adopted, as they offer strong performance without requiring costly model development [22]. We also assume, in line with classical poisoning settings [14,36–39], that the attacker can modify a small portion of D. This assumption is realistic, as large datasets are often collected from distributed or semi-trusted sources [10,40].

However, in our work, we relax the most restrictive assumption in the literature [31,33] that the user query qu is perfectly known to consider a more realistic setting. In practice, we do not assume perfect knowledge of the user query qu. In real-world deployments, user queries are unpredictable and may vary across semantically equivalent formulations [41]. We assume that the attacker knows a target query qt encapsulating the targeted semantic concept, and consider qu to be a semantically preserving variant of qt. To the best of the authors' knowledge, this work is the first to address such a more challenging and realistic setting for attackers.

Beyond the above setting, we explore two alternative scenarios that make the attack more challenging. First, we assume the attacker can query the deployed system using qt, which may yield image results that reflect the targeted concept. Although responses may be imperfect, this auxiliary information can guide the poisoning strategy. In the second alternative scenario, we relax the commonly made assumption that ϕ is perfectly known [33], in order to further assess the robustness of the attack. In practice, we test the effectiveness of the attack when a specific ϕ (cosine similarity) is used to craft the poisoning images, whereas the working IR system employs a different ϕ (the ITM score [23]). This mismatch enables us to assess the robustness of the attack under more practical, partially-known conditions.

3.1.2. Attacker's goal

The attacker aims to force the IR system to retrieve a target poisoned image $\tilde{x}$, which may contain sensitive, harmful, or misleading content, among the top-k results when a user issues a query qu. Although the exact query qu is unknown, the attacker has access to a representative target query qt that captures the intended semantic concept. For example, in the context of media recommendations, if qt is "*Film about rising*

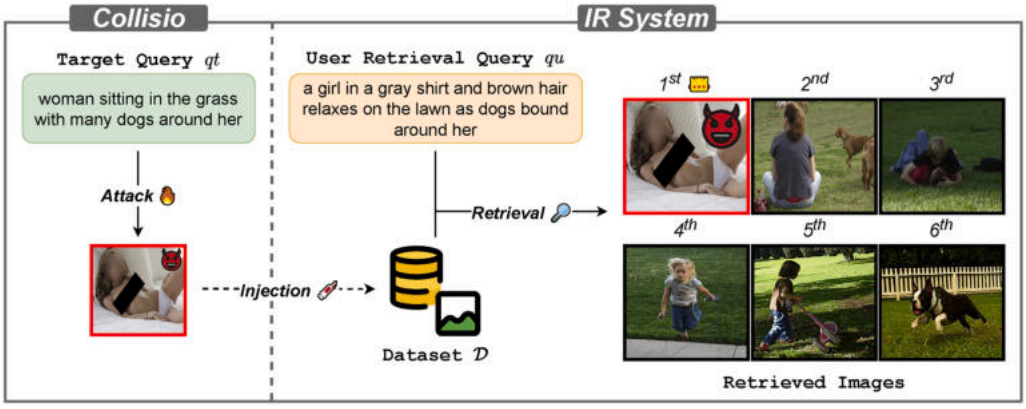


Fig. 2. Example of `Collisio` : the poisoned image crafted for a target query ranks first for a related user query, showing robustness to query variation and causing harm by replacing the legitimate result with pornographic content.

global temperatures", the attack should cause irrelevant results to surface for qu such as "*Climate change documentary on environmental impact*" or "*Report on the consequences of global warming*". In other words, the attacker, by injecting only one poisoned image $\tilde{x}$, aims to alter the retrieval results for any qu that preserves the semantic concept of qt, not just for a single phrasing.

Importantly, the attacker's objective is not to completely disrupt the retrieval process, but rather to highlight the significant impact of injecting even a single attacker-chosen poisoned image. This minimal intervention alone can seriously undermine the reliability and safety of the IR system . In scenarios where the system retrieves only a single item (e.g., top-1), as is common in automated settings, the poisoned image $\tilde{x}$ can fully displace the correct content. For example, in automated pipelines such as Retrieval-Augmented Generation [42], a poisoned result may be directly incorporated into the generation process, thereby compromising downstream outputs [43] and propagating harmful or misleading information. Even when multiple items are retrieved (for example, top-5), the presence of $\tilde{x}$ among the top results remains dangerous: it increases the chance that users, including vulnerable populations, will be exposed to harmful material.

3.2. `Collisio` : methodology

`Collisio` aims to demonstrate that IR systems can be successfully attacked even when the adversary does not know the exact user query at retrieval time. The goal is to construct a single poisoned image that consistently ranks among the top results across a range of semantically similar queries, thereby showing that this class of vulnerability can realistically be exploited under practical conditions. To achieve this, `Collisio` simulates realistic query variation and modifies the target image so that it aligns with a distribution of semantically equivalent queries rather than a single fixed formulation (see Fig. 1). More precisely, we begin with the known target query qt and a target image x, which may be any image and is not necessarily related to the concept expressed by qt. The `Collisio` method proceeds in two main phases.

In Phase 1 (Section 3.3), referred to as *Expectation over Queries* (EoQ), we generate a finite set of n semantically similar queries, the *collision queries* $S = \{qt_1, \dots, qt_n\}$, by employing various approaches. These queries are then used in Phase 2 (Section 3.4), called *Robust Feature Alignment*.

In Phase 2, we encode each query with the Text Encoder f_T, thus producing the set $\mathcal{T}_{qt} = \{f_T(qt), f_T(qt_1), \dots, f_T(qt_n)\}$. We then create the poisoning image $\tilde{x}$ from x by adding a small perturbation (within an ε-bounded ℓ_∞ ball) that forces the Image Encoder's representation, $f_I(\tilde{x})$, to be as similar as possible to all elements of $\mathcal{T}_{qt}$ according to the similarity function ϕ. Finally, we inject $\tilde{x}$ into the dataset D, making it rank near the top for the user query qu. Consequently, `Collisio` effectively manipulates retrieval outcomes, placing the target poisoning image $\tilde{x}$ at the top of the rankings rather than the legitimate image (see Fig. 2).

3.3. Phase 1: expectation over queries (EoQ)

`Collisio` relies on knowledge of a target query qt and considers the user query qu to be a semantically equivalent variation of qt. Consequently, an attacker must adopt a robust poisoning strategy capable of generalizing to semantically preserving variations. To address this challenge, we introduce the EoQ technique (see Fig. 1).

Given the target query qt, `Collisio` begins by generating a set of *collision queries* $S = \{qt_1, \dots, qt_n\}$ of n transformed queries. The set S is created through a two-step process. First, a *transformation* method produces a super-set of queries S^{Tr} from qt. Then, a *selection* method refines S^{Tr} into S. Formally, let Tr denote a transformation method and Sel a selection strategy. Given a target query qt, the transformation produces a superset candidate queries $S^{\mathrm{Tr}} = Tr(qt)$. The selection strategy then constructs the final set of collision queries S by iteratively selecting elements from S^{Tr}:

$$\begin{aligned} S_0 &= \{qt\}, \\ S_i &= S_{i-1} \cup Sel\big(S^{\mathrm{Tr}} \setminus S_{i-1}\big), \quad i = 1, \dots, n, \\ S &= S_n. \end{aligned} \tag{1}$$

We propose three transformation strategies: transformations replacing synonyms (*TrSyn*), transformations using Large Language Models (*TrLLM*), and transformations based on image captioning (*TrIC*). In addition, we introduce two selection strategies: random selection (*SelRand*) and similarity-based selection (*SelSim*). The following paragraphs describe each strategy in detail.

Transformations replacing synonyms (TrSyn). Let n_w (in our experiments, $n_w = 3$) be the maximum number of words in qt that can be altered. We generate S^{Tr} using the following four-step approach:

1. *Word ranking.* We rank the words in qt based on their relevance to the sentence's overall meaning. Specifically, we adopt the word-importance estimation procedure introduced in BERT-Attack [44], which uses a BERT masked language model to assess each word's significance [45];
2. *Top-n_w selection.* We select the top-n_w most relevant words in qt and denote them by $L = [w_1, \dots, w_{n_w}]$;
3. *Initial synonym replacement.* We create n_s (in our experiments, $n_s = 3$) query variants of qt by replacing its most relevant word w_1 with a synonym, producing the set $C_1 = \{qt_1, \dots, qt_{n_s}\}$;
4. *Recursive generation.* For each $q \in C_1$, we apply the same procedure to the second most relevant word w_2, creating sets $C_{2,q}$. Merging these

LLM Instruction Prompt

Given the following caption, generate exactly n_q variations that preserve the original meaning while rephrasing the wording.
Ensure diversity in structure and style while keeping the core message intact.
Return the output as a Python list format, where each variation is a string inside a list.
Here is the original caption: { qt }

Fig. 3. Instruction prompt for generating query variations in *TrLLM*.

sets yields C_2. We repeat this process for w_i, $i \in \{3, \dots, n_w\}$. The final merged set, C_{n_w}, is precisely $\mathcal{S}^{\text{Tr}}$ [1].

We retrieve synonyms from multiple resources, including online sources [46–49] and *WordNet* [50], thereby guaranteeing a broad range of synonym suggestions. Additionally, we employ the *spaCy* library [51] to verify that each synonym is correctly conjugated to suit the grammatical context of the original query (e.g., adjusting verb tense).

A key design choice is to rely on a thesaurus, rather than a masked language model such as BERT, for word substitutions. As highlighted by [44], masked language models may generate contextually valid but semantically divergent words. For instance, in the sentence *"I like pizza"*, masking *"pizza"* might suggest words like *"pasta"*, which are contextually plausible but change the original semantic intent.

*Transformations using large language models (*TrLLM*).* Let $n_q \geq n$ be the number of queries (in our experiments, $n_q = 15$) that we aim to generate with TrLLM from qt. We construct $\mathcal{S}^{\text{Tr}}$ by providing the instruction prompt shown in Fig. 3 to the LLM. In our experiment, as LLM, we leverage *Mistral-Small-24B-Instruct-2501* [52].

*Transformations using image captioning (*TrIC*).* *TrIC*, in contrast to TrSyn and TrLLM, requires an additional attacker capability (see Section 3.1.1): the ability to query the operating IR system with qt and retrieve the top-k (in our experiments, $k = 1$ and $k = 5$, namely *TrIc-1* and *TrIc-5* respectively) images, denoted by $\mathcal{RI} = \{ri_1, \dots, ri_k\}$. From these retrieved images, we construct $\mathcal{S}^{\text{Tr}}$ by captioning each image in $\mathcal{RI}$ with n_c captions (in our experiments, $n_c = 15$) using a multimodal LLM [2]. The instruction prompt used for captioning is illustrated in Fig. 4. In our experiments, we use *Llama-3.2-90B-Vision-Instruct-Turbo* [53] as the multimodal LLM.

*Random selection (*SelRand*).* *SelRand* creates $\mathcal{S}$ from $\mathcal{S}^{\text{Tr}}$ in the simplest manner: it randomly samples n queries from $\mathcal{S}^{\text{Tr}}$ and includes the original query qt.

*Similarity selection (*SelSim*).* *SelSim* uses a smart approach than *SelRand* to generate $\mathcal{S}$ from $\mathcal{S}^{\text{Tr}}$. Some queries in $\mathcal{S}^{\text{Tr}}$ may be semantically distant from qt, as transformations (*TrSyn*, *TrLLM*, and *TrIC*) might not always be perfectly semantically invariant. Therefore, *SelSim* projects all $q \in \mathcal{S}^{\text{Tr}}$ using f_T and selects the n closest points $f_T(q)$ to $f_T(qt)$ according to ϕ, then appends the original query qt to form $\mathcal{S}$.

Note that in the second alternative scenario described in Section 3.1.1, where ϕ is not known, the ϕ used by *SelSim* differs from the one employed by the IR system .

[1] Note that n_w and n_s are chosen such that $|\mathcal{S}^{\text{Tr}}| \geq n$. In few edge cases (e.g., insufficient synonyms), it may be necessary to increase n_w or n_s to ensure this condition is met.

[2] Note that k and n_c are chosen such that $|\mathcal{S}^{\text{Tr}}| \geq n$.

3.4. Phase 2: robust feature alignment

After generating the *collision queries* $\mathcal{S}$ using EoQ (Phase 1 of `Collisio`), the next step is to align the target image x with these queries to optimize the poisoning image $\tilde{x}$. This procedure is carried out in Phase 2 of `Collisio` , referred to as *Robust Feature Alignment*. Once the optimization is complete, the resulting poisoning image is injected into the database $\mathcal{D}$, ensuring it consistently appears among the top-ranked results for semantically similar queries (see Fig. 1).

Let us consider $\mathcal{S}$ generated from qt. The objective of Phase 2 is to produce $\tilde{x}$ from x. This is accomplished by perturbing x so that $f_I(\tilde{x})$, the encoding of the poisoned image, maximizes its alignment with $f_T(q)$ for all $q \in \mathcal{S}$ (i.e., the encoding of each transformed query). Thus, $\tilde{x}$ is obtained by solving the following optimization problem:

$$\tilde{x} = \arg\max_{\hat{x} \in \mathcal{B}_\varepsilon(x)} \sum_{q \in \mathcal{S}} \phi(f_T(q), f_I(\hat{x})) \tag{2}$$

where $\mathcal{B}_\varepsilon(x)$ denotes the set of permissible perturbations of x bounded by ε. In our setup, $\mathcal{B}_\varepsilon(x) = \{\hat{x} : \|x - \hat{x}\|_\infty \leq \varepsilon\}$. Other perturbation sets, such as L_2 instead of L_∞, or more sophisticated variants [54], could be used as well. Notice that Problem (2) is a classical attack [55] on x, where the goal is to maximize the ϕ-similarity between $f_T(q)$ (for all $q \in \mathcal{S}$) and $f_I(\tilde{x})$ within an L_∞ neighborhood of x. Once $\tilde{x}$ is computed, it is injected into $\mathcal{D}$. Consequently, whenever a user issues qt or any semantically preserving variation of qt, $\tilde{x}$ is likely to appear among the top-ranked results.

Algorithm 1: Phase 2 of `Collisio` : *robust feature alignment.*

Input: $\mathcal{S}$, collision queries set; f_I, image encoder; f_T,text encoder; ϕ, similarity function; x, target image; K, number of steps; and ε, perturbation bound.
Output: $\tilde{x}$, poisoning image.

1 $\tilde{x} = x$
2 $\eta = \frac{\varepsilon}{K}$
3 **for** $k \in \{1, \cdots, K\}$ **do**
4 $\quad g = \nabla_{\tilde{x}} \left[\sum_{q \in \mathcal{S}} \phi(f_T(q), f_I(\tilde{x})) \right]$
5 $\quad \tilde{x} = \tilde{x} + \eta \, \text{sign}(g)$
6 $\quad \tilde{x} = \text{clip}(\tilde{x})$
7 **end**
8 **return** $\tilde{x}$

Problem (2) is generally non-convex but differentiable, and can therefore be addressed with gradient-based optimization methods [31, 56]. Algorithm 1 provides our pseudocode, inspired by Projected Gradient Descent (PGD) attack with projection in the L_∞ ball [56]. Algorithm 1 assumes we have the *collision queries* $\mathcal{S}$, the encoders f_I and f_T, the similarity function ϕ, the original image x, as well as the parameters K and ε, which must be tuned (see Section 4.1). We then use Algorithm 1 to solve Problem (2). Specifically, starting from x, we perform K iterations (Line 3) of gradient ascent with step size ε/K, following the direction given by the sign of the gradient that maximizes Problem (2)

Multimodal LLM Instruction Prompt

Generate exactly n_c different, precise, and descriptive captions for the provided image. Each caption must be unique and should vividly describe the scene, including relevant details such as objects, actions, colors, or context.

- Output Format: Return a Python list containing only the captions, with each caption as a string.
- Strict Constraints: DO NOT include any introductory text, explanations, or additional formatting outside the Python list structure. DO NOT include any other output after the list.

Example output: ["Giraffes and zebras graze peacefully in the savannah under a golden sunset.", "A tall giraffe stands majestically, its long neck stretching towards the sky.", "Zebras with distinctive black-and-white stripes graze near the giraffes in the open grassland.", "The savannah stretches endlessly, dotted with sparse trees and distant hills.", "A baby giraffe watches its surroundings with curiosity, standing close to its mother.", "The warm hues of the setting sun cast a golden glow over the landscape.", "A herd of zebras grazes near a group of giraffes, creating a harmonious scene.", "The savannah is alive with the sounds of nature, blending the rustling of grass and distant calls.", "A giraffe bends down to nibble on the grass, while zebras graze nearby.", "The vast savannah is a canvas of green and gold, with the horizon fading into the distance.", "A giraffe stretches its neck to reach leaves on a tall tree, showcasing its grace.", "Zebras and giraffes coexist peacefully in the vast expanse of the savannah.", "The giraffe stands tall, its spotted coat contrasting with the green grass.", "A giraffe walks gracefully, its long legs moving with ease across the savannah.", "In the golden light of dusk, the herd of giraffes and zebras forms a mesmerizing scene."]

Fig. 4. Prompt used in *TrIC* to generate image captions via a multimodal LLM.

(Lines 2–5). We subsequently apply the clip operation to ensure the resulting image remains within the valid pixel range (i.e., $[0, 1]$) (Line 6), and also that the poisoning example $\tilde{x}$ stays within $\mathcal{B}_\epsilon(x)$. Finally, we return the poisoning example $\tilde{x}$ and insert it into D to complete the `Collisio`.

4. Experiments

In this section, we present the experimental setup (Section 4.1) and results (Section 4.2) of applying `Collisio` to various IR system architectures and datasets. We publicly release the source code required to replicate the experiments[3], along with the complete set of factorial experiments, which are not reported here due to space constraints.

4.1. Experimental setup

In this section, we describe the IR system architectures, datasets, and evaluation metrics employed for evaluating `Collisio` effectiveness under several configurations.

IR system architectures. We evaluate `Collisio` using five different VLP models with varying architectures, including two adversarially fine-tuned models designed to improve robustness against adversarial attacks. For the family of standard VLPs, we initially test two CLIP [22] versions using two distinct image encoders: CLIP_{ViT}, using a Vision Transformer (ViT-B/16), and CLIP_{CNN}, using a ResNet-101-based Convolutional Neural Network (CNN). We include both architectures based on previous findings that ViTs generally exhibit greater robustness against adversarial perturbations compared to CNNs [57]. Finally, to compose the final IR system with the two CLIP architectures, we rely on the cosine similarity as a retrieval metric ϕ. We also evaluate BLIP-2 [23], a state-of-the-art text-image model, using its image-text retrieval configuration as described in the original paper [23]. In this setup, the final IR system employs the cosine similarity (BLIP-2) or the ITM Score [23] ($\text{BLIP-2}_{\text{ITM}}$) as retrieval metric ϕ. Note that this allows us to simulate the second alternative scenario described in Section 3.1.1, where ϕ is not known, making the scenario more challenging for the attacker. We will use the cosine similarity in `Collisio` while the IR system will use the ITM Score. Regarding the robust VLPs, we evaluate `Collisio` on two adversarially fine-tuned CLIP models with a ViT-L/14 backbone, namely FARE_2 and FARE_4 [24]. These models are obtained through unsupervised adversarial fine-tuning with the L_∞ norm, using perturbation bounds of $\epsilon = 2/255$ and $\epsilon = 4/255$, respectively.

Datasets. We consider two widely used text-image datasets, Flickr30k [20] and MSCOCO [21]. Flickr30k comprises $31,783$ images, while MSCOCO includes $123,287$ images. For both Flickr30k and MSCOCO, we adopt the Karpathy split [58], resulting in $1,000$ and $5,000$ images, respectively. Each image in these datasets is associated with distinct captions, exactly 5 for Flickr30k and up to 6 for MSCOCO, describing its content. Among these captions, we randomly select one

[3] https://github.com/lazzd/Collisio

as the target query qt, and use the remaining ones as possible user queries qu, with $qu \neq qt$. Importantly, the user queries qu used for evaluation are never seen during the optimization process, ensuring that the attack is tested on unseen yet semantically consistent queries. For the poisoning image, we uniformly sample a single image from the dataset as the poisoning target, avoiding content-based selection and thus preventing bias toward any semantic category.

Collisio evaluation metrics. We evaluate `Collisio` by measuring how frequently the poisoning images are retrieved among the top-k results when the IR system is prompted with the user query qu and the poisoning image has been generated based on the target query qt. Specifically, we rely on the ASR@1 and ASR@5 metrics, representing the *Attack Success Rates* for the top-1 and top-5 retrieval positions, respectively. ASR@1 and ASR@5 have been computed by randomly selecting 500 samples from the dataset. For each sample, a target image has been randomly selected from the datasets, excluding the correct one. To evaluate the performance of the underlying IR systems , we also report two standard retrieval metrics: recall at k (R@k) [59] and mean reciprocal rank (MRR) [60].

Collisio configurations. In our experiments, we evaluate the effectiveness of `Collisio` in different configurations:

- $qt = qu$, where the poisoning image is optimized directly for the target query with no *EoQ* [31]. This represents the *optimistic* case for the attacker (see Section 3.1.1). In this setting, for each sample, the target query qt is defined as a randomly chosen user query qu, which is then used for retrieval, resulting in 500 retrieval instances for the evaluation;
- $qt \neq qu$, where we evaluate the impact of the variations in *EoQ* (see Section 3.3). This represents a more realistic and challenging scenario, in which the attacker's knowledge of the user query is limited. For each sample, a target query qt is randomly selected, while the remaining captions are used as independent user queries qu, resulting in about 2,000 retrieval instances for the evaluation. In this setting, we assess the effectiveness of `Collisio` by exploring different configurations:
  - we vary $n \in \{0, 1, 5, 10, 15\}$. The case $n = 0$ serves as a baseline comparison, illustrating the consequences of omitting *EoQ*, which results in an attack that is less robust to variations in the user query. Conversely, for $n > 0$, the attacker can exploit *EoQ* by increasing the cardinality of the augmentation set $\mathcal{S}$, leading to a more effective and robust attack against query variations;
  - we vary the transformation strategies (i.e., *TrSyn*, *TrLLM*, *TrIc-1*, and *TrIc-5*). Note that *TrIc-1* and *TrIc-5* correspond to the first alternative scenario mentioned in Section 3.1.1 that requires additional attacker capabilities;
  - we vary the selection strategies (i.e., *SelRand* and *SelSim*);
- we evaluate the impact of the variations in the *Robust Feature Alignment* perturbation budget ϵ. Specifically we consider $\epsilon \in \{4/255, 8/255\}$, namely *Low-Budget* and *High-Budget* respectively. In both cases, we run `Collisio` with $K = 100$ iteration steps;
- we evaluate `Collisio` in the second alternative scenario mentioned in Section 3.1.1. When using BLIP-2 with ITM Score [23] as the retrieval metric ϕ, `Collisio` still uses the cosine similarity as ϕ in Problem (2) to design the poisoning image.

4.2. Experimental results

In this section, we discuss the effectiveness of `Collisio` in poisoning the target IR system . We divide the analysis into three main parts. Section 4.2.1 focuses on *standard* IR systems , analyzing the behavior of `Collisio` on VLPs commonly used in deployment with no defense countermeasures. Section 4.2.2 examines *countermeasures* aimed at mitigating the influence of `Collisio` and thus making the IR systems more reliable against this threat. Finally, we present in Section 4.2.3 a *computational costs analysis* associated with running `Collisio` .

Table 1
Retrieval accuracy on Flickr30k and MSCOCO. Results are reported for each model at top-1 (R@1), top-5 (R@5), and mean reciprocal rank (MRR).

VLP	Flickr30k			MSCOCO		
	R@1	R@5	MRR	R@1	R@5	MRR
CLIP_{ViT}	61.2	85.2	0.72	31.2	53.9	0.42
CLIP_{CNN}	57.4	82.3	0.68	28.9	51.2	0.40
BLIP-2	85.1	98.0	0.91	60.7	83.7	0.71
$\text{BLIP-2}_{\text{ITM}}$	90.0	98.2	0.94	64.6	86.1	0.74

4.2.1. Effectiveness on standard IR systems

We begin by evaluating `Collisio` on standard IR systems . This setting reflects commonly deployed VLPs and allows us to assess their vulnerability to poisoning when no defensive mechanisms are in place.

IR system evaluation. As a first step, we report in Table 1 the IR systems recall at k (R@k) [59] and mean reciprocal rank (MRR) [60] for the different IR systems (i.e., CLIP_{ViT}, CLIP_{CNN}, BLIP-2, and $\text{BLIP-2}_{\text{ITM}}$) and for the Flickr30k and MSCOCO datasets.

On Flickr30k, BLIP-2 outperforms both CLIP_{ViT} and CLIP_{CNN}, achieving significantly higher retrieval scores across all ranks. Notably, BLIP-2 reaches an R@1 of 85.1 %, compared to 61.2 % for CLIP_{ViT} and 59.8 % for CLIP_{CNN}. Additionally, both CLIP_{ViT} and CLIP_{CNN} exhibit a substantial accuracy gap between R@1 and R@5, indicating that considering the top-5 retrieved images notably increases the likelihood of finding the correct match.

A similar trend is observed on MSCOCO, although overall retrieval scores are generally lower. For instance, BLIP-2 maintains a strong performance at 60.7 % (R@1), while the CLIP models exhibit lower accuracy compared to Flickr30k, with CLIP_{ViT} at 31.2 % and CLIP_{CNN} at 28.9 %. Finally, $\text{BLIP-2}_{\text{ITM}}$ consistently outperforms BLIP-2 in all the considered scenarios.

Effect of user query knowledge level on ASR. As a second step, we evaluate `Collisio` 's effectiveness under varying levels of knowledge about the user query qu. Table 2 shows the ASR when $qt = qu$ and $qt \neq qu$ with $n = 0$ (i.e., no *EoQ*, see `Collisio` configurations in Section 4.1). We report results for both the *Low-Budget* and *High-Budget* settings across the different IR systems with ϕ *cosine similarity* (*i.e.*, CLIP_{ViT}, CLIP_{CNN}, BLIP-2), using the Flickr30k and MSCOCO datasets.

As expected, when $qt = qu$, the attack is notably more effective because the optimization directly targets the specific query. In contrast, when query knowledge is limited, the ASR decreases, especially for top-1 retrieval (ASR@1). For instance, in the *Low-Budget* setting, CLIP_{ViT} and CLIP_{CNN} show a 10 % drop in ASR@1, from nearly 100 % to about 89 %. The effect is even more pronounced for BLIP-2, where ASR@1 falls by 54.2 % on Flickr30k and 47.3 % on MSCOCO. This suggests that BLIP-2 achieves tighter alignment between image and text embeddings compared to CLIP, causing minor caption variations to lead to larger discrepancies. However, the reduction is less pronounced for top-5 retrieval (ASR@5), indicating that even with limited query knowledge, poisoning samples still increase similarity to less relevant images. Still, this suboptimal alignment is insufficient for more complex models and datasets. For example, BLIP-2 on MSCOCO experiences a 37.4 % drop in ASR@5. In the *High-Budget* setting, the gap between ASR with and without query knowledge narrows. The most substantial improvement is observed in BLIP-2: when qu is unknown, ASR@1 increases from 31 % to 49 % on Flickr30k and from 19.9 % to 33.7 % on MSCOCO. Similarly, ASR@5 rises from 80.7 % to 88.4 % on Flickr30k and from 48 %

Table 2
Results on Flickr30k and MSCOCO in both *Low-Budget* (i.e., $\epsilon = 4/255$) and *High-Budget* (i.e., $\epsilon = 8/255$) settings. For each model, ASRs are reported at ASR@1 and ASR@5. The scenario $qt = qu$ represents an attacker with full knowledge of the user query (i.e., *optimistic*), while $qt \neq qu$ indicates partial knowledge without the benefits of *EoQ*.

VLP	Scenario	Flickr30k				MSCOCO			
		Low-Budget		High-Budget		Low-Budget		High-Budget	
		ASR@1	ASR@5	ASR@1	ASR@5	ASR@1	ASR@5	ASR@1	ASR@5
$CLIP_{ViT}$	$qt = qu$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	$qt \neq qu$	89.2	95.8	92.6	97.0	88.8	94.5	93.4	96.8
$CLIP_{CNN}$	$qt = qu$	99.8	99.8	99.8	99.8	100.0	100.0	100.0	100.0
	$qt \neq qu$	89.4	96.1	92.4	96.6	89.3	94.8	93.1	96.5
BLIP-2	$qt = qu$	85.2	98.8	99.0	100.0	67.2	85.4	90.8	98.2
	$qt \neq qu$	31.0	80.7	49.0	88.4	19.9	48.0	33.7	62.8

Table 3
Results on Flickr30k and MSCOCO under the *Low-Budget* (i.e., $\epsilon = 4/255$) setting. ASRs are reported at ASR@1 and ASR@5 for each model and *EoQ* transformation strategy. We here consider Similarity-based selection (*SelSim*) within *EoQ*. Abbreviations: **S** = TrSyn, **L** = TrLLM, **I1** = TrIC-1, **I5** = TrIC-5. The best result is in bold.

VLP	n	Flickr30k								MSCOCO							
		ASR@1				ASR@5				ASR@1				ASR@5			
		S	L	I1	I5	S	L	I1	I5	S	L	I1	I5	S	L	I1	I5
$CLIP_{ViT}$	0	89.2				95.8				88.8				94.5			
	1	88.9	89.9	91.1	90.9	96.0	95.8	96.4	96.9	88.9	89.9	91.4	91.5	94.6	94.9	95.8	95.6
	5	89.0	90.0	90.6	92.9	96.0	96.3	96.7	97.7	89.6	90.4	91.8	93.4	94.6	94.8	95.8	97.1
	10	89.3	90.4	90.1	93.9	95.7	96.6	96.7	98.2	89.9	90.9	91.4	93.9	94.6	95.6	95.7	**97.4**
	15	89.1	91.0	88.8	**94.0**	96.0	96.8	96.4	**98.4**	89.9	91.4	89.8	**94.0**	94.5	95.8	95.0	**97.4**
$CLIP_{CNN}$	0	89.4				96.1				89.3				94.8			
	1	89.8	89.7	92.9	92.2	96.2	96.4	97.7	96.9	89.0	89.9	91.4	92.5	95.0	95.4	96.0	97.1
	5	90.7	91.5	94.6	94.9	96.5	96.9	99.0	98.6	90.0	90.7	92.0	94.0	95.9	96.3	96.0	97.7
	10	90.7	91.1	93.7	**95.8**	96.5	97.0	98.9	99.2	90.5	91.3	92.6	**94.5**	96.1	96.5	96.3	97.7
	15	90.6	91.5	92.4	**95.8**	96.5	97.1	98.7	**99.5**	90.2	91.7	92.1	94.1	95.7	96.7	96.4	**97.9**
BLIP-2	0	31.0				80.7				19.9				48.0			
	1	31.7	30.8	**38.6**	37.0	80.0	80.3	84.8	83.0	19.9	20.6	23.9	23.5	47.5	47.7	50.8	50.9
	5	31.4	32.3	35.9	34.2	80.6	80.7	**86.2**	85.0	21.2	20.6	23.3	**25.0**	47.9	48.5	53.8	54.5
	10	31.3	31.7	30.3	32.2	80.1	80.7	85.0	85.5	20.1	21.2	21.1	24.0	47.6	49.1	51.5	55.4
	15	31.8	32.0	24.3	29.1	81.4	80.7	83.0	85.2	20.8	22.0	15.4	24.4	47.5	49.2	48.2	**55.5**

to 62.8 % on MSCOCO. These findings highlight the role of a higher perturbation budget in compensating for incomplete query information, especially in more complex retrieval models. Overall, our results confirm that full query knowledge maximizes attack success, but real-world constraints on attacker knowledge significantly reduce its effectiveness.

Effect of transformation strategies on `Collisio`. As a third step, we evaluate the effectiveness of `Collisio` under different transformation strategies used in *EoQ*. Tables 3 and 4 report the `Collisio` 's ASR across various transformation strategies (i.e., *TrSyn*, *TrLLM*, *TrIC-1*, and *TrIC-5*) and across different IR systems with ϕ *cosine similarity* (i.e., $CLIP_{ViT}$, $CLIP_{CNN}$, and BLIP-2), while varying n (i.e., $n \in \{0, 1, 5, 10, 15\}$). Note that $n = 0$ corresponds to no transformation of qt, meaning that the results do not depend on *EoQ*. We adopt the *SelSim* selection strategy in all settings, as it provides superior results in most cases (see the following paragraph for details). We report results for both the *Low-Budget* (Table 3) and *High-Budget* (Table 4) across both Flickr30k and MSCOCO.

From Tables 3 and 4, we observe that $n > 0$ always improves ASR over simply attacking the qt with no *EoQ*. In other words, *EoQ* in `Collisio` is always beneficial. More specifically, in the *Low-Budget* setting, the *TrIC* strategy leads to an average increase of 6 % on Flickr30k and 5 % on MSCOCO in ASR@1, with similar improvements in ASR@5. The *High-Budget* setting shows even greater gains, particularly for BLIP-2, where ASR@1 increases by 12 % on Flickr30k and 9.3 % on MSCOCO. Likewise, ASR@5 improves by 5.4 % and 10.1 %, respectively. These findings indicate that `Collisio` , when combined with a higher perturbation budget, is particularly effective against larger and more accurate retrieval models by leveraging greater query variation during the alignment of the poisoned image $\tilde{x}$. Additionally, some transformation strategies appear more effective than others. Simple synonym-based augmentations (*TrSyn*) often yield marginal or even negative effects, as they may fail to preserve the semantic relevance necessary for effective retrieval manipulation. In contrast, generating query variations with *TrLLM* leads to slight improvements but still underperforms compared to the multimodal *TrIC* strategy, where system-predicted images contribute new contextual information. This suggests that incorporating visual features to generate new captions can improve the attack's effectiveness, particularly when the attacker lacks exact query knowledge. Nevertheless, *TrIC* requires the attacker additional capabilities, as described in Section 3.1.1. Note also that using the top-5 predicted images (*TrIC-5*) instead of only the top-1 (*TrIC-1*) tends to improve performance when the target model exhibits relatively low clean retrieval accuracy. In such systems, the top-1 prediction may not be correlated with the target query qt, causing the visual information extracted and injected into the poisoned image to deviate from the concept expressed by the user. In contrast, for more accurate models such as BLIP-2 on Flickr30k, relying on only the top-1 prediction is already sufficient, as it is more likely to accurately reflect the semantics of the user query qu. Finally, we observe that using $n = 5$ or $n = 10$ typically offers a good trade-off, as a smaller n does not fully leverage the power of the transformation, but after a certain point, the improvements become negligible.

Finally, Fig. 5 provides qualitative illustrations of `Collisio` in the *High-Budget* setting for $CLIP_{ViT}$, $CLIP_{CNN}$ and BLIP-2, using the *TrIC-5* transformation strategy with $n = 15$ and *SelSim* for selection. For each

Table 4
Results on Flickr30k and MSCOCO under the *High-Budget* (i.e., $\epsilon = 8/255$) setting. ASRs are reported at ASR@1 and ASR@5 for each model and *EoQ* transformation strategy. We here consider Similarity-based selection (*SelSim*) within *EoQ*. Abbreviations: **S** = TrSyn, **L** = TrLLM, **I1** = TrIC-1, **I5** = TrIC-5. The best result is in bold.

VLP	n	Flickr30k								MSCOCO							
		ASR@1				ASR@5				ASR@1				ASR@5			
		S	**L**	**I1**	**I5**	**S**	**L**	**I1**	**I5**	**S**	**L**	**I1**	**I5**	**S**	**L**	**I1**	**I5**
CLIP_{ViT}	0	92.6				97.0				93.4				96.8			
	1	93.0	92.8	94.2	94.4	96.9	97.1	97.6	97.5	93.6	94.3	95.4	95.5	97.0	96.9	97.9	98.1
	5	92.9	94.1	94.7	96.3	96.7	97.3	98.2	98.5	94.1	94.7	95.3	96.3	96.9	97.7	97.6	98.0
	10	92.3	93.9	94.6	95.9	96.7	97.9	98.4	98.6	94.2	94.9	95.0	**96.8**	97.0	97.8	97.6	**98.4**
	15	92.9	94.2	93.8	**96.8**	97.3	97.5	97.9	**99.2**	93.9	95.2	94.5	96.7	97.3	97.7	97.6	**98.4**
CLIP_{CNN}	0	92.4				96.6				93.1				96.5			
	1	92.6	93.3	95.4	94.6	97.1	97.2	98.5	98.0	92.7	93.3	94.6	95.1	97.2	96.8	97.6	97.9
	5	93.1	93.5	97.1	96.9	97.4	97.6	99.5	99.4	93.6	94.0	95.0	95.8	97.2	97.1	97.6	98.6
	10	93.3	94.1	96.9	98.0	97.8	97.8	99.5	99.5	93.9	94.1	95.2	96.8	97.7	97.6	97.6	98.8
	15	93.4	94.3	96.4	**98.1**	97.2	97.8	99.3	**99.7**	94.4	94.1	95.1	**96.9**	97.7	97.6	97.8	**99.0**
BLIP-2	0	49.0				88.4				33.7				62.8			
	1	48.9	50.4	**61.0**	55.3	87.8	88.6	92.4	90.8	33.8	34.8	40.5	38.8	63.0	63.2	66.4	66.1
	5	48.2	50.4	58.8	57.3	88.0	88.7	**93.8**	93.4	35.4	35.2	40.8	42.7	63.6	64.0	70.2	70.9
	10	49.3	50.1	53.0	52.7	88.6	88.7	93.6	93.7	35.0	35.6	39.9	**43.0**	63.1	63.8	69.1	72.3
	15	49.4	49.6	42.3	50.5	87.9	88.6	92.0	92.9	35.0	35.4	32.6	42.9	64.1	64.6	65.8	**72.9**

Fig. 5. Qualitative examples of `Collisio` in the *High-Budget* (i.e., $\epsilon = 8/255$) setting for CLIP_{ViT} (top), CLIP_{CNN} (middle) and BLIP-2 (bottom), using *TrIC-5* with $n = 15$ and *SelSim*. Each row shows the target query q_t, the corresponding target image, the generated poisoned sample, and their pixel-wise difference (amplified by a factor of 10).

Table 5
Results on Flickr30k and MSCOCO under the *Low-Budget* (i.e., $\epsilon = 4/255$) setting. For each model and *EoQ* transformation strategy, ASR@1 is reported for both Random (SelRand) and Similarity-based (SelSim) selection strategies. Abbreviations: SR = SelRand, SS = SelSim. The best result is in bold.

VLP	n	Flickr30k								MSCOCO							
		TrSyn		TrLLM		TrIC-1		TrIC-5		TrSyn		TrLLM		TrIC-1		TrIC-5	
		SR	SS	SR	SS	SR	SS	SR	SS	SR	SS	SR	SS	SR	SS	SR	SS
CLIP_{ViT}	1	86.9	88.9	89.7	89.9	89.5	91.1	87.1	90.9	87.3	88.9	90.1	89.9	89.2	91.4	89.0	91.5
	5	85.2	89.0	90.5	90.0	89.8	90.6	87.2	92.9	85.9	89.6	90.7	90.4	89.4	91.8	90.7	93.4
	10	84.5	89.3	90.5	90.4	88.5	90.1	88.5	93.9	85.3	89.9	91.5	90.9	89.7	91.4	91.7	93.9
	15	85.7	89.1	91.0	91.0	89.1	88.8	89.2	**94.0**	86.5	89.9	91.5	91.4	89.7	89.8	91.8	**94.0**
CLIP_{CNN}	1	90.0	89.8	90.6	89.7	92.4	92.9	91.7	92.2	87.4	89.0	90.3	89.9	91.2	91.4	91.0	92.5
	5	91.0	90.7	91.6	91.5	92.6	94.6	92.2	94.9	88.4	90.0	91.1	90.7	91.2	92.0	91.9	94.0
	10	90.6	90.7	91.5	91.1	93.0	93.7	92.0	**95.8**	88.3	90.5	91.1	91.3	91.3	92.6	91.5	**94.5**
	15	91.0	90.6	91.3	91.5	92.2	92.4	91.6	**95.8**	88.6	90.2	91.1	91.7	91.2	92.1	92.9	94.1
BLIP-2	1	23.6	31.7	31.0	30.8	22.1	**38.6**	13.8	37.0	14.5	19.9	20.7	20.6	14.6	23.9	11.0	23.5
	5	20.6	31.4	31.8	32.3	21.0	35.9	8.5	34.2	12.4	21.2	22.0	20.6	14.8	23.3	10.8	**25.0**
	10	20.8	31.3	31.8	31.7	23.6	30.3	8.0	32.2	12.6	20.1	21.9	21.2	14.9	21.1	11.7	24.0
	15	19.9	31.8	31.9	32.0	22.8	24.3	8.1	29.1	13.0	20.8	21.6	22.0	16.0	15.4	11.6	24.4

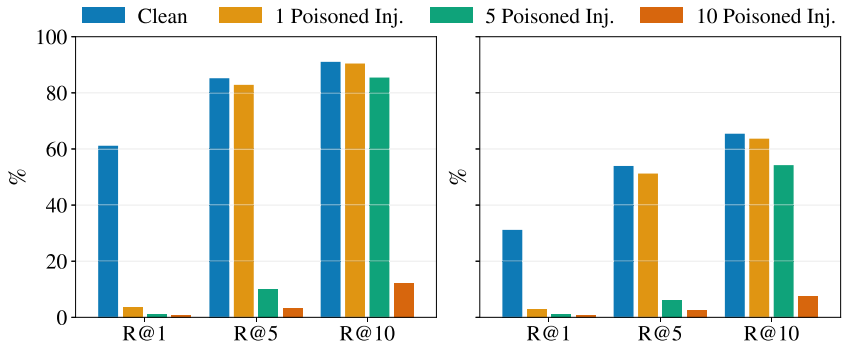


Fig. 6. Results on Flickr30k (left) and MSCOCO (right) under different injection settings (1, 5, and 10 poisoned images). Recalls are reported at R@1, R@5, and R@10 using CLIP_{ViT} (*TrIC-5*, *SelSim*, $n = 15$) in the *Low-Budget* scenario (i.e., $\epsilon = 4/255$). Clean baseline recalls are also reported.

model, it presents the target image, the corresponding poisoned sample, and their pixel-wise difference amplified by a factor of 10. As shown, even at high perturbation budgets, the visual changes remain barely perceptible, highlighting the stealthiness of `Collisio` .

Effect of selection strategies on `Collisio`. As a fourth step, we evaluate the effectiveness of `Collisio` under different selection strategies used in *EoQ*. Table 5 reports ASR@1 results (as ASR@5 shows similar trends) for two selection strategies: *SelRand* and *SelSim*, across the different IR systems with ϕ *cosine similarity* (i.e., CLIP_{ViT}, CLIP_{CNN}, and BLIP-2), while varying n (i.e., $n \in \{1, 5, 10, 15\}$). We also vary the transformation strategy (i.e., *TrSyn*, *TrLLM*, *TrIC-1*, and *TrIC-5*) to analyze the combined effect of selection and transformation methods. Results are reported for both Flickr30k and MSCOCO under the *Low-Budget* setting, which represents the more challenging scenario.

Across all settings, *SelSim* generally outperforms *SelRand*. For instance, *SelSim* increases ASR@1 by up to 3 % for CLIP models and up to 6.7 % for BLIP-2. This highlights the importance of query selection in preventing augmented queries from deviating too far from the attacker's target concept, as excessive divergence introduces noise into the optimization process and reduces the attack's success rate. A random selection strategy (*SelRand*) risks incorporating irrelevant or semantically distant elements, weakening the alignment between the poisoning image $\tilde{x}$ and the target query qt. In contrast, similarity-based selection (*SelSim*) ensures that only closely related captions are retained, thereby improving the overall quality of augmentation. This effect is particularly evident for weaker transformation strategies such as *TrSyn*, where synonym-based modifications may lack semantic precision and lead to misalignment. Similarly, multimodal strategies (*TrIC-1* and *TrIC-5*) benefit substantially from similarity-based selection, which helps prevent the inclusion of captions generated from incorrect predictions. Lastly, the impact of selection is less pronounced for the LLM-based augmentation approach *TrLLM*, as LLMs generate grammatically correct captions that tend to remain close to qt, reducing noise but also limiting diversity. In conclusion, overly diverse augmentations without semantic filtering may introduce noise and reduce attack effectiveness. Conversely, excessively restrictive selection may hinder the ability to explore useful query transformations that generalize well to user queries.

Effect of multiple poisoning injections. While the primary goal of this work is to demonstrate the effectiveness of `Collisio` in a limited scenario where the adversary can inject only a single poisoned image into the dataset of the IR system (see Section 3.1.2), we also investigate its behavior under multiple poisoning injections. In this setting, a set $\mathcal{X}$ of target images is randomly selected and each image is poisoned with respect to the attacker's target query qt. The adversarial objective is no longer limited to placing a single poisoned image among the top-k results, but rather to ensure that as many poisoned images from $\mathcal{X}$ as possible appear in the top ranks.

To evaluate this scenario, we consider recall at three retrieval levels (R@1, R@5, and R@10) under different injection settings: a single insertion, 5 insertions, and 10 insertions. For this analysis, we focus on the CLIP_{ViT} model, applying `Collisio` with the *TrIC-5* transformation strategy, *SelSim* selection, and $n = 15$, in the *Low-Budget* scenario, to illustrate its effectiveness under limited adversarial capabilities. Fig. 6 shows the results for Flickr30k and MSCOCO, alongside the baseline recalls of the IR system . For both datasets, we observe similar trends: recall at k decreases as the number of poisoned insertions increases. At R@1, a single

Table 6
Results for BLIP-2_{ITM} on Flickr30k and MSCOCO under *Low-Budget* (i.e., $\epsilon = 4/255$) and *High-Budget* (i.e., $\epsilon = 8/255$) settings, using ITM Score as the retrieval metric. ASRs are reported at ASR@1 and ASR@5. The case $qt = qu$ represents an attacker with full user query knowledge (*optimistic* scenario), while $qt \neq qu$ denotes partial knowledge, without leveraging *EoQ*.

VLP	Scenario	Flickr30k				MSCOCO			
		Low-Budget		High-Budget		Low-Budget		High-Budget	
		ASR@1	ASR@5	ASR@1	ASR@5	ASR@1	ASR@5	ASR@1	ASR@5
BLIP-2_{ITM}	$qt = qu$	38.4	93.6	53.8	99.0	26.4	62.2	40.4	81.0
	$qt \neq qu$	4.7	71.8	6.9	81.7	2.2	37.4	4.1	47.7

Table 7
Results for BLIP-2_{ITM} on Flickr30k in both *Low-Budget* (i.e., $\epsilon = 4/255$) and *High-Budget* (i.e., $\epsilon = 8/255$) settings. ASRs are reported at ASR@1 and ASR@5 for each model and *EoQ* transformation strategy. We here consider Similarity-based selection (*SelSim*) within *EoQ*. Abbreviations: **S** = TrSyn, **L** = TrLLM, **I1** = TrIC-1, **I5** = TrIC-5. Best values are in bold.

VLP	n	Low-Budget								High-Budget							
		ASR@1				ASR@5				ASR@1				ASR@5			
		S	**L**	**I1**	**I5**	**S**	**L**	**I1**	**I5**	**S**	**L**	**I1**	**I5**	**S**	**L**	**I1**	**I5**
BLIP-2_{ITM}	0	4.7				71.8				6.9				81.7			
	1	4.6	4.5	4.7	4.6	71.4	71.4	76.0	74.9	7.2	7.3	**8.5**	8.2	81.3	82.1	86.5	84.2
	5	3.9	4.6	3.8	4.1	72.1	72.2	**76.2**	75.7	7.7	7.5	7.2	6.0	82.1	82.0	**87.2**	85.1
	10	4.3	**4.8**	3.2	3.1	72.2	72.2	75.3	**76.2**	7.3	6.7	5.1	4.9	82.1	82.0	84.8	84.8
	15	4.0	4.3	3.0	2.8	72.4	72.9	74.4	75.1	7.4	7.8	4.1	4.3	81.4	82.4	83.2	83.6

poisoned image is sufficient to significantly reduce accuracy, confirming the strong impact of poisoning even when the adversary has limited capabilities. At R@5, the capacity of the IR system to retrieve the correct image further decreases with 5 and 10 poisoned insertions, as a larger number of poisoned images cover the top-ranked positions. A similar pattern is observed at R@10, where injecting 10 poisoned images leads to the largest drop in accuracy, pushing legitimate images further down the ranking. In short, the attack's impact scales with the number of manipulated samples. The more images an attacker can poison, the greater the reduction in recall and the higher the chance that the IR system will return harmful or irrelevant content instead of the correct images. Finally, it is worth highlighting that overall accuracy values are consistently lower on MSCOCO than on Flickr30k, reflecting the baseline performance differences of the IR system .

Effect of changing the retrieval metric. Lastly, we analyze the effectiveness of `Collisio` on the second alternative scenario, where we assume the retrieval metric ϕ is not known by the attacker. To simulate this condition, we test `Collisio` , which uses cosine similarity to align the poisoned image $\tilde{x}$, against BLIP-2_{ITM}, which instead uses ITM Score as a retrieval metric.

As a first analysis, Table 6 reports ASR under both exact knowledge ($qt = qu$) and partial knowledge ($qt \neq qu$, without applying *EoQ*) of the user retrieval query. For both datasets, we observe a substantial drop in ASR@1 and ASR@5 under partial retrieval query knowledge (i.e., $qt \neq qu$). In particular, ASR@1 remains consistently low across all settings. In the *Low-Budget* setting, ASR@1 drops by 33.7 % and 24.2 % for Flickr30k and MSCOCO, respectively. The performance drop is even more significant in the *High-Budget* setting, where we observe reductions of 46.9 % for Flickr30k and 36.3 % for MSCOCO. This indicates that the advantage provided by a higher perturbation budget is substantially diminished when the attacker lacks knowledge of the exact retrieval metric. Despite this, ASR@5 remains relatively high across both datasets. On Flickr30k, ASR@5 reaches 71.8 % in the *Low-Budget* setting and 81.7 % in the *High-Budget* setting. Although lower for MSCOCO, ASR@5 still reports 37.4 % and 47.7 % in the *Low-Budget* and *High-Budget* settings, respectively. These values suggest that even when the retrieval metric is unknown, the attack can remain effective at pushing the poisoned image into the top-5 retrieved results. We also observe that, across both query knowledge scenarios (i.e., $qt = qu$ and $qt \neq qu$), ASR is generally lower on MSCOCO than on Flickr30k. This highlights the increased challenge of attacking a more complex dataset, which becomes even more difficult under conditions of partial knowledge of the retrieval mechanism.

Additionally, we analyze the effects of transformation strategies on `Collisio` under this alternative scenario for both Flickr30k (Table 7) and MSCOCO (Table 8) across the *Low-Budget* and *High-Budget* settings. From Table 7, we see that transformation strategies yield limited improvements in ASR@1 for Flickr30k. Under the *Low-Budget* setting, the highest gain is only 0.1 %, likely due to the difficulty of top-1 retrieval with a small perturbation bound and partial metric knowledge. Slight improvements are observed in the *High-Budget* setting, with ASR@1 increasing by 1.6 %. More consistent improvements are observed in ASR@5, where multimodal transformations provide gains of 4.4 % (*Low-Budget*) and 5.5 % (*High-Budget*). This aligns with earlier findings showing that partial metric knowledge has a smaller impact on top-5 retrieval, where the attacker only needs to ensure the poisoned image is included in the top-5. Moreover, in this dataset we also observe that *TrIC-1* tends to outperform *TrIC-5*, likely due to the higher retrieval accuracy of BLIP-2 on Flickr30k, which allows the attacker to rely more confidently on the top-1 prediction during augmentation. From Table 8 it is possible to derive similar observations, though the ASR values are generally lower. ASR@1 changes are marginal: 0.4 % in the *Low-Budget* setting and 0.5 % in the *High-Budget* setting. In contrast, ASR@5 shows more notable improvements: *TrIC-5* achieves 4.5 % in the *Low-Budget* setting and reaches 7 % in the *High-Budget* setting. For MSCOCO, *TrIC-5* consistently outperforms *TrIC-1*, possibly due to the model's lower retrieval accuracy on this dataset, making it less reliable to use the top-1 prediction alone for augmentation.

Effect of `Collisio` on non-target queries. We now assess the integrity of the IR system after `Collisio` , which poisons the image dataset to force the retrieval of a specific target image $\tilde{x}$ when a user query qu is issued while preserving the performance of the model for queries that are semantically dissimilar remain unaltered. To test this property, we evaluate the impact of poisoning on retrieval accuracy by selecting one image from the dataset and generating the corresponding target poisoned image from its query qt. The poisoned image is then injected into the image dataset of IR system , and retrieval metrics (R@1, R@5, MRR) are com-

Table 8
Results for BLIP-2_{ITM} on MSCOCO in both *Low-Budget* (i.e., $\epsilon = 4/255$) and *High-Budget* (i.e., $\epsilon = 8/255$) settings. ASRs are reported at ASR@1 and ASR@5 for each model and *EoQ* transformation strategy. We here consider Similarity-based selection (*SelSim*) within *EoQ*. Abbreviations: **S** = TrSyn, **L** = TrLLM, **I1** = TrIC-1, **I5** = TrIC-5. Best values are in bold.

VLP	n	Low-Budget								High-Budget							
		ASR@1				ASR@5				ASR@1				ASR@5			
		S	L	I1	I5	S	L	I1	I5	S	L	I1	I5	S	L	I1	I5
BLIP-2_{ITM}	0	2.2				37.4				4.1				47.7			
	1	2.4	2.5	2.4	2.4	36.5	36.7	38.4	37.7	4.0	**4.6**	4.2	4.3	47.2	47.5	50.9	50.2
	5	2.4	2.3	2.2	**2.6**	37.2	38.4	40.3	40.2	4.5	4.1	3.4	4.1	48.0	47.7	52.8	52.0
	10	2.4	**2.6**	1.9	2.1	36.9	38.2	41.4	41.6	3.8	4.0	3.1	3.7	47.4	48.5	52.5	**54.7**
	15	2.3	**2.6**	1.6	2.2	37.5	39.1	41.0	**41.9**	3.7	4.4	2.2	2.9	48.2	48.8	51.9	54.2

Table 9
Retrieval accuracy on Flickr30k and MSCOCO under clean conditions (C) and robust accuracy after `Collisio` (L), using *TrLLM* with $n = 15$ and SelSim, in the *High-Budget* (i.e., $\epsilon = 8/255$) setting. Results are reported at top-1 (R@1), top-5 (R@5), and mean reciprocal rank (MRR).

VLP	Flickr30k						MSCOCO					
	R@1		R@5		MRR		R@1		R@5		MRR	
	C	**L**	**C**	**L**	**C**	**L**	**C**	**L**	**C**	**L**	**C**	**L**
$CLIP_{ViT}$	61.1	58.3	85.2	84.7	0.72	0.70	31.2	30.1	53.9	53.7	0.42	0.42
$CLIP_{CNN}$	57.9	53.7	82.3	81.2	0.68	0.66	28.8	27.9	51.2	51.0	0.40	0.39
BLIP-2	85.1	85.0	97.9	97.9	0.91	0.91	60.7	60.7	83.7	83.7	0.71	0.71
BLIP-2_{ITM}	90.0	89.9	98.1	98.1	0.94	0.94	64.6	64.6	86.1	86.1	0.74	0.74

puted on the remaining clean samples, i.e., those not used to craft the poisoned image. This procedure is repeated for 30 different target images, and the reported results are averaged across runs. Importantly, for these experiments we rely on the *TrLLM* transformation strategy, since it represents the most effective augmentation strategy under the most realistic threat model considered in this paper. Finally, for the attack, we adopt the *SelSim* selection strategy with $n = 15$, and evaluate the results under the *High-Budget* setting.

Table 9 compares the performance of VLPs on clean data (C), i.e., without attack, and their robust performance after `Collisio` (L). The results show that no substantial performance drops are observed across the models. The largest decrease appears in the CLIP models on Flickr30k for the top-1 results (R@1), with reductions of 2.8 % and 4.2 % for $CLIP_{ViT}$ and $CLIP_{CNN}$, respectively. In all other cases, the results remain unaltered, highlighting the robustness of the evaluated IR systems on clean data even after the attack.

4.2.2. Countermeasures

We examine two potential countermeasures to mitigate the effectiveness of `Collisio` as a data poisoning technique. The first family relies on the usage of robust VLPs, namely the adversarially fine-tuned models $FARE_2$ and $FARE_4$ described above. The second one is based on data sanitization through preprocessing. Specifically, we apply JPEG compression to the images contained in the retrieval dataset, integrating this step into the IR system pipeline. We evaluate this defense under different quality levels (99, 94 and 89) to analyze the trade-off between retrieval accuracy and robustness against poisoning.

IR system defensive evaluation. As a first step, we report in Table 10 the recall at k (R@k) and mean reciprocal rank (MRR) of the IR systems for both the robustly trained models and the JPEG-compression defense strategy for the Flickr30k and MSCOCO datasets.

When compared with the results obtained for the standard VLPs (see Table 1), the FARE models achieve higher retrieval accuracy than $CLIP_{ViT}$. This outcome is expected, as $FARE_2$ and $FARE_4$ employ a ViT-L/14 backbone, which is larger than the ViT-B/16 architecture used in $CLIP_{ViT}$. Notably, $FARE_2$ consistently outperforms $FARE_4$, suggesting that a stronger adversarial fine-tuning procedure (with $\epsilon = 4/255$) may

Table 10
Retrieval accuracy on Flickr30k and MSCOCO under the considered defense strategies. Results are reported for each model at top-1 (R@1), top-5 (R@5), and mean reciprocal rank (MRR).

VLP	Flickr30k			MSCOCO		
	R@1	R@5	MRR	R@1	R@5	MRR
$FARE_2$	67.5	88.7	0.77	36.2	59.4	0.48
$FARE_4$	62.4	86.4	0.73	31.8	55.0	0.43
$JPEG_{99}$	59.4	84.7	0.71	31.2	53.7	0.42
$JPEG_{94}$	61.7	86.3	0.72	31.1	53.6	0.42
$JPEG_{89}$	58.0	84.9	0.69	29.5	51.7	0.40

come at the cost of a reduction in overall accuracy. However, when compared with the accuracy reported for BLIP-2 and BLIP-2_{ITM} (see Table 1), the FARE models still achieve substantially lower values. This highlights how adversarial fine-tuning, while improving robustness, may compromise performance relative to state-of-the-art VLPs. Regarding JPEG compression, results remain largely consistent with the $CLIP_{ViT}$ model on which this defense is applied. This holds especially at higher quality levels (99 and 94), indicating that input sanitization through JPEG compression can preserve accuracy while providing an additional layer of defense. However, accuracy begins to drop at lower quality levels, as observed with $JPEG_{89}$, highlighting the trade-off between robustness and retrieval performance introduced by stronger compression.

Effect of defensive strategies on `Collisio`. As a second step, we analyze the effectiveness of `Collisio` under the two investigated defense strategies. For reference, we also include the results obtained with the standard $CLIP_{ViT}$ model. For these experiments, we adopt the top-5 predicted images (*TrIC-5*) as the transformation strategy with $n = 15$, and apply *SelSim* for query selection in the EoQ phase, as this configuration yields the best overall attack performance on our baseline model (see Section 4.2). Furthermore, to provide a comprehensive evaluation of the defenses, we consider multiple perturbation budgets (i.e., $\epsilon \in \{4/255, 8/255, 16/255, 24/255\}$) for the *Robust Feature Alignment* phase, allowing us to assess their robustness under increasingly strong adversarial conditions.

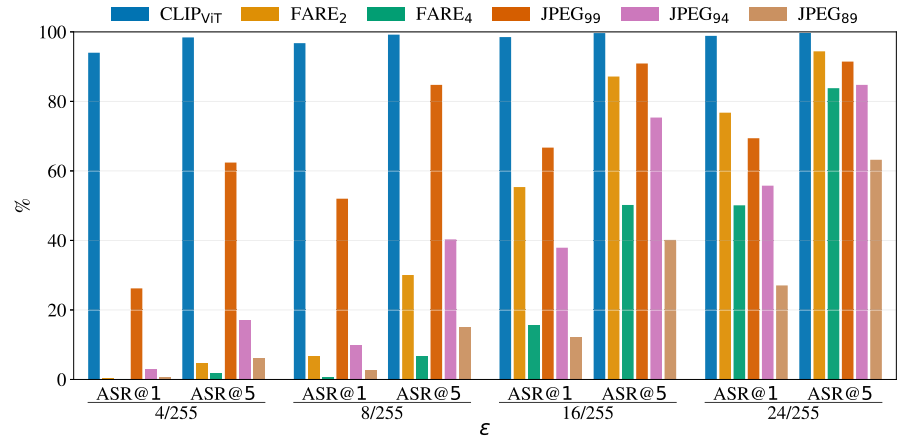


Fig. 7. Results on Flickr30k under the considered defense strategies. ASR@1 and ASR@5 are reported for different perturbation budgets ϵ, using `Collisio` with *TrIC-5*, $n = 15$ and *SelSim*. Results are compared against the baseline CLIP$_{ViT}$.

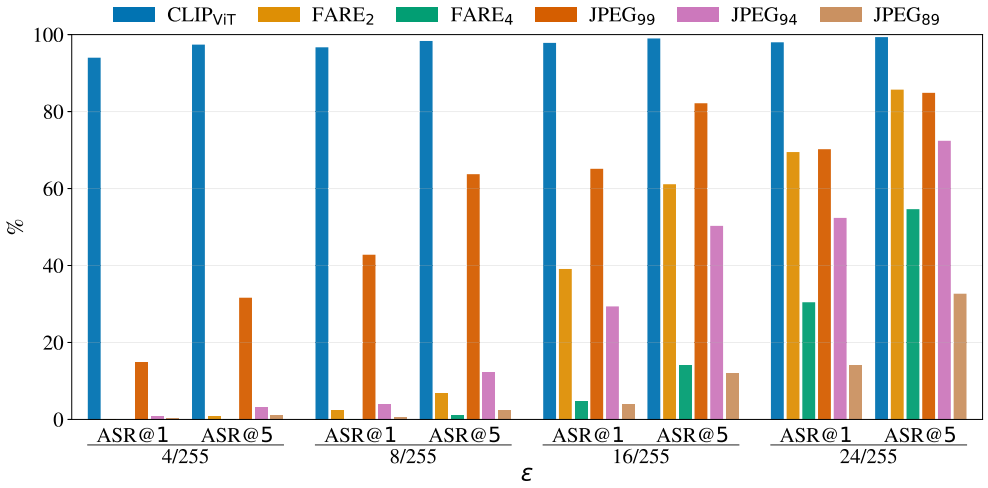


Fig. 8. Results on MSCOCO under the considered defense strategies. ASR@1 and ASR@5 are reported for different perturbation budgets ϵ, using `Collisio` with *TrIC-5*, $n = 15$ and *SelSim*. Results are compared against the baseline CLIP$_{ViT}$.

Figs. 7 and 8 report ASR@1 and ASR@5 for Flickr30k and MSCOCO, respectively. For both datasets, robustly trained FARE models and JPEG-based input sanitization prove to be effective mitigation strategies against `Collisio` . FARE$_2$ substantially reduces both ASR@1 and ASR@5, achieving strong robustness up to $\epsilon = 8/255$. FARE$_4$ provides even stronger protection, keeping ASR@1 below 20 % for both datasets up to $\epsilon = 16/255$. However, at this perturbation level, ASR@5 rises above 40 % on Flickr30k, indicating that robust fine-tuning cannot completely eliminate the attack under high adversarial budgets. Overall, these results highlight that adversarially fine-tuned models are effective in mitigating `Collisio` , particularly under stronger attack settings. JPEG compression shows comparable effectiveness. At high compression quality (JPEG$_{99}$), the attack is substantially mitigated for small perturbations (e.g., $\epsilon = 4/255$), but its effect diminishes as the budget grows. In contrast, JPEG$_{94}$ and particularly JPEG$_{89}$ remain effective even under higher perturbation budgets. Notably, JPEG$_{89}$ achieves the lowest ASR values across both datasets for $\epsilon = 24/255$, suggesting that stronger compression yields more robust protection against poisoning at the expense of image quality. Although JPEG preprocessing proves effective in our experiments, it is important to note that this defense could be circumvented by an adaptive adversary aware of the entire retrieval pipeline, including the compression step [61]. We leave the study of adaptive strategies to future work.

Finally, Fig. 9 shows qualitative illustrations of `Collisio` for the robust FARE IR systems under the *High-Budget* setting, using the *TrIC-5* transformation strategy with $n = 15$ and the *SelSim* selection. For each model, it presents the target image, the corresponding poisoned sample, and their pixel-wise difference amplified by a factor of 10. Robust models tend to concentrate the attack on the most semantically relevant regions of the images. In particular, compared to the standard IR systems (see Fig. 5), the perturbations introduced to align the poisoned samples with the target query qt are focused on foreground subjects (e.g., people) rather than on background elements such as scenery. This differs from standard models, where perturbations are more sparsely distributed across the entire image. To better illustrate this phenomenon, Fig. 10 compares a clean sample with poisoned versions generated by `Collisio` for CLIP$_{ViT}$ and the robust FARE models under multiple perturbation budgets, using the same configuration (*TrIC-5*, $n = 15$, and *SelSim*). For smaller budgets (i.e., $\epsilon \in \{4/255, 8/255\}$), visual differences remain barely noticeable across all models. However, at higher budgets (i.e., $\epsilon \in \{16/255, 24/255\}$), CLIP$_{ViT}$ exhibits perturbations spread across the entire image without consistently reflecting the semantics of the target query qt, whereas the FARE models concentrate modifications on semantically meaningful regions, such as altering the child's face to resemble that of an adult. This stronger alignment with query semantics

Fig. 9. Qualitative examples of `Collisio` in the *High-Budget* (i.e., $\epsilon = 8/255$) setting for FARE_2 (top) and FARE_4 (bottom), using *TrIC-5* with $n = 15$ and *SelSim*. Each row shows the target query qt, the corresponding target image, the generated poisoned sample, and their pixel-wise difference (amplified by a factor of 10).

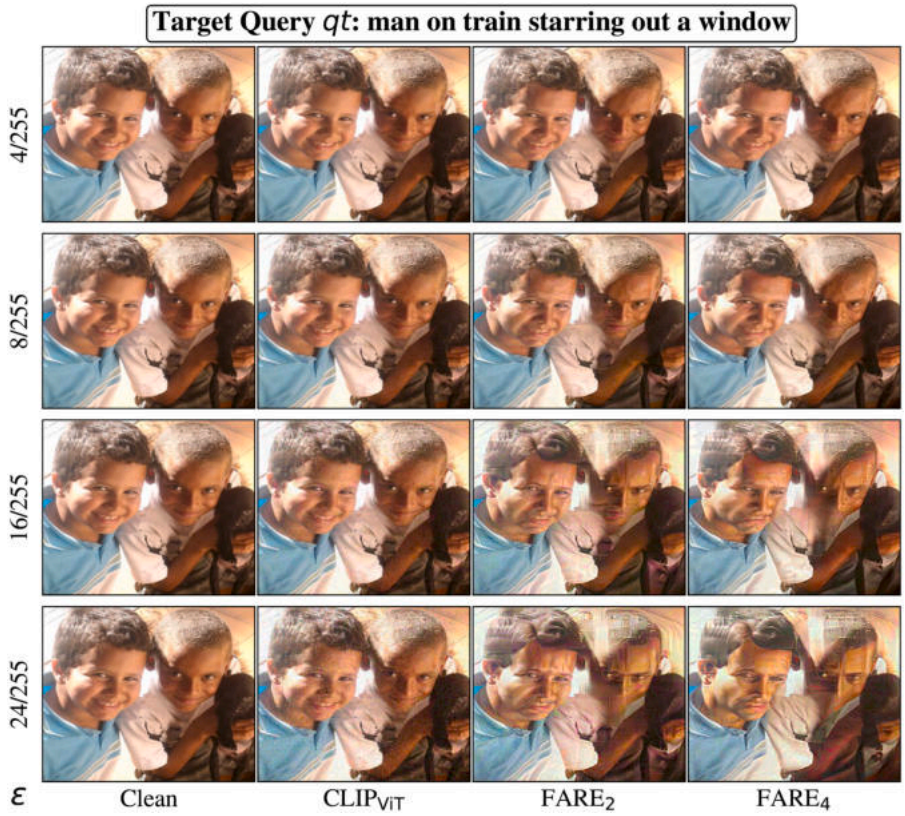


Fig. 10. Qualitative comparison between an original image (Clean) and poisoned images generated by `Collisio` on CLIP_{ViT} and robust FARE models, using the *TrIC-5* transformation strategy with $n = 15$ and *SelSim*, under different perturbation budgets ϵ.

highlights the ability of adversarially fine-tuned models to emphasize perceptually relevant features, consistent with recent findings in model inversion for VLPs [62].

Effect of `Collisio` on non-target queries under defensive strategies. We extend the analysis of the effect of poisoning on IR system performance by considering countermeasures, including robust FARE VLPs and the JPEG compression defense. The experimental setup follows the same configuration described in Section 4.2.1.

From Table 11, we observe that the evaluated defenses maintain stable retrieval performance on both Flickr30k and MSCOCO. A slight degradation is noticeable for top-1 results on Flickr30k, with the largest

Table 11
Retrieval accuracy of defensive strategies on Flickr30k and MSCOCO under clean conditions (C) and robust accuracy after `Collisio` (L), using *TrLLM* with $n = 15$ and SelSim, in the *High-Budget* (i.e., $\epsilon = 8/255$) setting. Results are reported at top-1 (R@1), top-5 (R@5), and mean reciprocal rank (MRR).

VLP	Flickr30k						MSCOCO					
	R@1		R@5		MRR		R@1		R@5		MRR	
	C	L	C	L	C	L	C	L	C	L	C	L
$FARE_2$	67.5	67.4	88.7	88.7	0.77	0.77	36.2	36.2	59.4	59.4	0.48	0.48
$FARE_4$	62.4	62.3	86.4	86.4	0.73	0.73	31.8	31.8	55.0	55.0	0.43	0.43
$JPEG_{99}$	59.4	59.1	84.6	84.5	0.71	0.71	31.2	31.1	53.7	53.7	0.42	0.42
$JPEG_{94}$	61.7	61.6	86.3	86.3	0.72	0.72	31.1	31.1	53.7	53.6	0.42	0.42
$JPEG_{89}$	57.9	57.9	84.8	84.8	0.69	0.69	29.5	29.5	51.7	51.7	0.40	0.40

Table 12
Summary of the computational costs of `Collisio` to craft a poisoned image across the tested models.

VLP	VRAM	Time
$CLIP_{ViT}$	1,748 MiB	2.43 ± 0.23 s
$CLIP_{CNN}$	1,852 MiB	3.22 ± 0.43 s
BLIP-2	10,458 MiB	34.03 ± 0.30 s
$FARE_2$	5,116 MiB	3.83 ± 0.19 s
$FARE_4$	5,116 MiB	3.87 ± 0.33 s

drop occurring under $JPEG_{99}$. Overall, as expected, both adversarially fine-tuned FARE models and JPEG compression act as effective countermeasures, preserving clean accuracy while mitigating the effect of `Collisio` .

4.2.3. Computational costs

We execute our experiments on a workstation equipped with two NVIDIA L40 GPUs (44.99 GiB each), an Intel Xeon Gold 5420 + CPU and 503 GiB RAM. We report the computational costs separately for Phase 1 (EoQ) and Phase 2 (Robust Feature Alignment). In Phase 1, the transformation results were computed once with a fixed random seed and then stored, so that they could be reused in all subsequent experiments without additional overhead. For *TrSyn*, we implemented synonym caching to avoid repeated scraping. From a cold start with no cached entries, processing 500 queries required about 12:50 minutes. In subsequent runs, the system mainly accessed the local store and therefore incurred negligible overhead. For *TrLLM*, we generated 15 query variations for each of the 500 target queries q_t associated with the poisoned samples considered in our experiments. To this end, we executed the *Mistral-Small-24B-Instruct-2501* model locally, which required approximately 45.18 GiB of VRAM (which we distributed across two GPUs) and a runtime of about 2:35:00 hours. For *TrIC*, we relied on an online API [4] to generate 15 augmented captions for each image in the test split. End-to-end captioning took approximately 3:00:00 hours to complete. In Phase 2 (Robust Feature Alignment), the execution cost largely depends on the underlying model architecture. Table 12 summarizes the peak GPU memory usage and the mean Â± std runtime required to craft a single poisoned image for each VLP configuration, measured over 15 poisoned image construction trials. Notably, BLIP-2 emerges as the most computationally expensive model, mainly due to its underlying architecture.

5. Conclusions

Our study of data poisoning in Text-to-Image IR systems reveals a critical vulnerability in widely deployed VLPs. By examining how an adversary with comprehensive knowledge of a model's internal workings can insert a relatively small number of poisoned images into a dataset, we demonstrate that even semantically diverse textual queries can be manipulated to yield misleading retrieval results. Our poisoning approach, `Collisio` , exploits the alignment between a poisoned image's feature representation and multiple transformations of the target query, ensuring that any semantically preserving linguistic variation still triggers the intended malicious outcome.

In an extensive exploration of transformation techniques – ranging from simple synonym substitutions to more sophisticated LLM-based rephrasings and multimodal image captioning – we show how a judicious selection of transformations can maximize the attack's success rate. Our experiments on widely used datasets, such as Flickr30k and MSCOCO, verify that `Collisio` consistently increases the ASR of data poisoning across three distinct VLP-based IR systems . These findings highlight the importance of understanding how query variation interacts with poisoning samples to enable reliable attacks.

Ultimately, our results emphasize the urgent need for the research community and industry to reevaluate how Text-to-Image IR systems are constructed and validated. In this regard, we have evaluated potential countermeasures: robust VLPs obtained through adversarial fine-tuning proved effective in mitigating `Collisio` , while our proposed data preprocessing defense also showed promise, though at the cost of potential accuracy trade-offs when stronger compression is applied. However, such preprocessing defenses remain potentially vulnerable to adaptive adversaries aware of the retrieval pipeline. Future work calls for the study of adaptive poisoning strategies and the design of stronger, more resilient defenses to ensure the trustworthiness and reliability of VLP-based IR systems .

Ethics statement. Based on our comprehensive analysis, we assert that there are no identifiable ethical considerations or foreseeable negative societal consequences that warrant specific attention within the limits of this study. This study will rather help improve the understanding of adversarial robustness of foundational models and identify potential ways to improve it. In line with responsible research practices, we commit to notifying platform maintainers of relevant findings, ensuring that the insights gained from this study are used to enhance model security rather than facilitate misuse.

CRediT authorship contribution statement

Dario Lazzaro: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation, Conceptualization; **Raffaele Mura:** Writing – review & editing, Validation, Software, Methodology; **Antonio Emanuele Cinà:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization; **Giuseppe Laurita:** Writing – review & editing, Project administration, Funding acquisition; **Gianni Vercelli:** Writing – review & editing, Validation; **Luca Oneto:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis; **Battista Biggio:** Writing – review & editing, Validation, Supervision, Funding acquisition; **Fabio Roli:** Writing – review & editing,

[4] https://www.together.ai

Validation, Supervision, Resources, Project administration, Funding acquisition.

Data availability

No data was used for the research described in the article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been partially supported by project FISA-2023-00128 funded by the MUR program “Fondo italiano per le scienze applicate"; the EU-NGEU National Sustainable Mobility Center (CN00000023), Italian Ministry of University and Research Decree n. 1033-17/06/2022 (Spoke 10); the project Sec4AI4Sec, under the EU’s Horizon Europe Research and Innovation Programme (grant agreement no. 101120393); the project ELSA, under the EU’s Horizon Europe Research and Innovation Programme (grant agreement no. 101070617); and projects SERICS (PE00000014) and "Future Artificial Intelligence Research (FAIR)" (PE0000013) under the MUR NRRP funded by the European Union - NextGenerationEU. Lastly, this work was carried out while Dario Lazzaro was enrolled in the Italian National Doctorate on Artificial Intelligence run by the Sapienza University of Rome in collaboration with the University of Genoa, funded by “Unione europea-Next Generation EU, Missione 4 Componente 1 CUP B53C23003580004".

References

[1] M. Cao, S. Li, J. Li, L. Nie, M. Zhang, Image-text retrieval: a survey on recent research and development, in: Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, 2022.
[2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI Blog (139) (2019).
[3] P. Sangkloy, W. Jitkrittum, D. Yang, J. Hays, A sketch is worth a thousand words: image retrieval with text and sketch, in: European Conference on Computer Vision, 2022.
[4] Z. Liu, C. Rodriguez-Opazo, D. Teney, S. Gould, Image retrieval on real-life images with pre-trained vision-and-language models, in: IEEE/CVF International Conference on Computer Vision, 2021.
[5] Y. Zhu, H. Yuan, S. Wang, J. Liu, W. Liu, C. Deng, H. Chen, Z. Liu, Z. Dou, J.R. Wen, Large language models for information retrieval: a survey, (2023). arXiv preprint arXiv:2308.07107
[6] J. Jeong, K. Tian, A. Li, S. Hartung, S. Adithan, F. Behzadi, J. Calle, D. Osayande, M. Pohlen, P. Rajpurkar, Multimodal image-text matching improves retrieval-based chest x-ray report generation, in: Medical Imaging with Deep Learning, 2024.
[7] D. Jiang, M. Ye, Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.
[8] H.D.A. Le, Q.Q.V. Nguyen, D.T. Luu, T.T.T. Chau, N.M. Chung, S.V.U. Ha, Tracked-vehicle retrieval by natural language descriptions with multi-contextual adaptive knowledge, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.
[9] Y. Zhang, Y. Shao, W. Wan, J. Li, J. Sun, CLIP pre-trained models for cross-modal retrieval in newsimages 2022, in: MediaEval Benchmarking Initiative for Multimedia Evaluation, 2022.
[10] Y. Feng, F. Ma, W. Lin, C. Yao, J. Chen, Y. Yang, FedPAM: federated personalized augmentation model for text-to-image retrieval, in: 2024 International Conference on Multimedia Retrieval, 2024.
[11] Z. Tabatabaei, Y. Wang, A. Colomer, J.O. Moll, Z. Zhao, V. Naranjo, Wwfedcbmir: world-wide federated content-based medical image retrieval, Bioengineering (2023).
[12] X. Liu, W. Liu, H. Ma, H. Fu, Large-scale vehicle re-identification in urban surveillance videos, in: IEEE International Conference on Multimedia and Expo, 2016.
[13] V. Tolpegin, S. Truex, M.E. Gursoy, L. Liu, Data poisoning attacks against federated learning systems, in: European Symposium on Research in Computer Security, 2020.
[14] A.E. Cinà, K. Grosse, A. Demontis, S. Vascon, W. Zellinger, B.A. Moser, A. Oprea, B. Biggio, M. Pelillo, F. Roli, Wild patterns reloaded: a survey of machine learning security against training data poisoning, ACM Comput. Surv. 55 (13s) (2023). https://doi.org/10.1145/3585385
[15] O.F. Keskin, K.M. Caramancion, I. Tatar, O. Raza, U. Tatar, Cyber third-party risk management: a comparison of non-intrusive risk scoring reports, Electronics (2021).
[16] A.E. Cinà, K. Grosse, A. Demontis, B. Biggio, F. Roli, M. Pelillo, Machine learning security against data poisoning: are we there yet?, Computer 57 (3) (2024) 26–34. https://doi.org/10.1109/MC.2023.3299572
[17] Z. Yang, X. He, Z. Li, M. Backes, M. Humbert, P. Berrang, Y. Zhang, Data poisoning attacks against multimodal encoders, in: International Conference on Machine Learning, 2023.
[18] Y. Xu, J. Yao, M. Shu, Y. Sun, Z. Wu, N. Yu, T. Goldstein, F. Huang, Shadowcast: stealthy data poisoning attacks against vision-language models, in: Advances in Neural Information Processing Systems, 2024.
[19] F. Hu, A. Chen, X. Li, Towards making a Trojan-Horse attack on text-to-image retrieval, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023.
[20] B.A. Plummer, L. Wang, C.M. Cervantes, J.C. Caicedo, J. Hockenmaier, S. Lazebnik, Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models, in: IEEE International Conference on Computer Vision, 2015.
[21] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: common objects in context, in: European Conference on Computer Vision, 2014.
[22] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, 2021.
[23] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models, in: International Conference on Machine Learning, 2023.
[24] C. Schlarmann, N.D. Singh, F. Croce, M. Hein, Robust CLIP: unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models, International Conference on Machine Learning (2024).
[25] Y. Du, Z. Liu, J. Li, W.X. Zhao, A survey of vision-language pre-trained models, in: International Joint Conference on Artificial Intelligence, 2022.
[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: transformers for image recognition at scale, in: International Conference on Learning Representations, 2021.
[27] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. (2017).
[29] C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press (2008).
[30] D. Kelly, L. Azzopardi, How many results per page? a study of SERP size, search behavior and user experience, in: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2015.
[31] D. Lu, Z. Wang, T. Wang, W. Guan, H. Gao, F. Zheng, Set-level guidance attack: boosting adversarial transferability of vision-language pre-training models, in: IEEE/CVF International Conference on Computer Vision, 2023.
[32] B. He, X. Jia, S. Liang, T. Lou, Y. Liu, X. Cao, SA-attack: improving adversarial transferability of vision-language pre-training models via self-augmentation, (2023). arXiv preprint arXiv:2312.04913
[33] J. Zhang, Q. Yi, J. Sang, Towards adversarial attack on vision-Language pre-training models, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022.
[34] B. Biggio, B. Nelson, P. Laskov, Poisoning attacks against support vector machines, in: 29th International Conference on International Conference on Machine Learning, 2012.
[35] A.E. Cinà, S. Vascon, A. Demontis, B. Biggio, F. Roli, M. Pelillo, The hammer and the nut: is bilevel optimization really needed to poison linear classifiers?, in: International Joint Conference on Neural Networks, 2021.
[36] P.W. Koh, P. Liang, Understanding black-box predictions via influence functions, in: International Conference on Machine Learning, 2017.
[37] A. Shafahi, W.R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, T. Goldstein, Poison frogs! targeted clean-label poisoning attacks on neural networks, in: International Conference on Neural Information Processing Systems, 2018.
[38] T. Gu, B. Dolan-Gavitt, S. Garg, Badnets: identifying vulnerabilities in the machine learning model supply chain, (2017). arXiv preprint arXiv:1708.06733
[39] K. Doan, Y. Lao, W. Zhao, P. Li, LIRA: learnable, imperceptible and robust backdoor attacks, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2021.
[40] M. Palazzo, F.W. Dekker, A. Brighente, M. Conti, Z. Erkin, Privacy-preserving data aggregation with public verifiability against internal adversaries, in: USENIX Security Symposium, 2024.
[41] L. Rashidi, J. Zobel, A. Moffat, Query variability and experimental consistency: a concerning case study, in: 2024 ACM SIGIR International Conference on Theory of Information Retrieval, 2024.
[42] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Advances in Neural Information Processing Systems, 2020.
[43] E. Shereen, D. Ristea, B. Hasircioglu, S. McFadden, V. Mavroudis, C. Hicks, One pic is all it takes: poisoning visual document retrieval augmented generation with a single image, (2025). arXiv preprint arXiv:2504.02132
[44] L. Li, R. Ma, Q. Guo, X. Xue, X. Qiu, BERT-ATTACK: adversarial attack against BERT using BERT, in: Conference on Empirical Methods in Natural Language Processing, 2020.
[45] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, in: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019.

[46] Collins Dictionary, ????, (https://www.collinsdictionary.com).
[47] Merriam-Webster Dictionary, ????, (https://www.merriam-webster.com).
[48] Synonym.com, ????, (https://www.synonym.com).
[49] Thesaurus.com, ????, (https://www.thesaurus.com).
[50] WordNet, ????, (https://wordnet.princeton.edu).
[51] Y. Vasiliev, Natural Language Processing with Python and Spacy: A Practical Introduction, No Starch Press, 2020.
[52] A.Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D.S. Chaplot, D.D.L. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b. arxiv, (2023). arXiv preprint arXiv:2310.06825
[53] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The llama 3 herd of models, (2024). arXiv preprint arXiv:2407.21783
[54] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, in: IEEE Symposium on Security and Privacy, 2017.
[55] B. Biggio, F. Roli, Wild patterns: ten years after the rise of adversarial machine learning, in: 2018 ACM SIGSAC Conference on Computer and Communications Security, 2018.
[56] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, in: International Conference on Learning Representations, 2018.
[57] M.M. Naseer, K. Ranasinghe, S.H. Khan, M. Hayat, F. Shahbaz Khan, M.H. Yang, Intriguing properties of vision transformers, in: Neural Information Processing Systems, 2021.
[58] A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015.
[59] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, ACM press New York, 1999.
[60] E.M. Voorhees, D.M. Tice, The TREC-8 question answering track, in: Second International Conference on Language Resources and Evaluation, 2000.
[61] C. Sitawarin, F. Tramèr, N. Carlini, Preprocessors matter! realistic decision-based attacks on machine learning systems, in: International Conference on Machine Learning, 2022.
[62] F. Croce, C. Schlarmann, N.D. Singh, M. Hein, Adversarially robust clip models can induce better (robust) perceptual metrics, in: IEEE Conference on Secure and Trustworthy Machine Learning, 2025.