Document Screenshot Retrievers are Vulnerable to Pixel Poisoning Attacks

Shengyao Zhuang*
CSIRO
Brisbane, QLD, Australia

Ekaterina Khramtsova*
The University of Queensland
Brisbane, QLD, Australia

Xueguang Ma
University of Waterloo
Waterloo, Canada

Bevan Koopman
CSIRO and
The University of Queensland
Brisbane, QLD, Australia

Jimmy Lin
University of Waterloo
Waterloo, Canada

Guido Zuccon
The University of Queensland
Brisbane, QLD, Australia

Abstract

Recent advancements in dense retrieval have introduced vision-language model (VLM)-based retrievers, such as DSE and ColPali, which leverage document screenshots embedded as vectors to enable effective search and offer a simplified pipeline over traditional text-only methods. In this study, we propose three pixel poisoning attack methods designed to compromise VLM-based retrievers and evaluate their effectiveness under various attack settings and parameter configurations. Our empirical results demonstrate that injecting even a single adversarial screenshot into the retrieval corpus can significantly disrupt search results, poisoning the top-10 retrieved documents for 41.9% of queries in the case of DSE and 26.4% for ColPali. These vulnerability rates notably exceed those observed with equivalent attacks on text-only retrievers. Moreover, when targeting a small set of known queries, the attack success rate raises, achieving complete success in certain cases. By exposing the vulnerabilities inherent in vision-language models, this work highlights the potential risks associated with their deployment.

CCS Concepts

• **Information systems** → **Language models**; *Question answering*; **Multimedia and multimodal retrieval**; **Adversarial retrieval**.

Keywords

Document screenshot retrieval; corpus poisoning; VLM-based dense retrievers

ACM Reference Format:
Shengyao Zhuang, Ekaterina Khramtsova, Xueguang Ma, Bevan Koopman, Jimmy Lin, and Guido Zuccon. 2025. Document Screenshot Retrievers are Vulnerable to Pixel Poisoning Attacks. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3726302.3730056

*Equal Contribution

SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730056

1 Introduction

Emerging vision-language models (VLMs) have brought new opportunities in document retrieval [7, 19]. VLM-based dense retrievers (e.g., DSE [19] and ColPali [7]) embed document screenshots into vector representations and have proven effectiveness in retrieval tasks; they have the added advantage of simplifying the indexing pipeline by directly utilising visual features, removing the need for extra components like optical character recognition or table parsing methods.

In this paper we show that document screenshot retrievers powered by vision-language models (VLMs) are particularly susceptible to corpus poisoning and search engine optimisation (SEO) attacks. This is because, unlike text-based dense retrieval, where methods like HotFlip [14, 31] are typically used to generate adversarial documents for corpus poisoning, the image-based nature of VLMs introduces a novel attack vector. Specifically, the pixel values of document screenshots can be directly manipulated using gradients to deceive the model — an attack strategy that has been extensively studied in computer vision but has not been considered in the context of document screenshot retrieval.

We develop three pixel-based attack methods for attacking document screenshot retrievers: 1) ***Direct Optimisation***, 2) ***Noise Optimisation***; and 3) ***Mask Direct Optimisation***. All methods begin with a seed document screenshot image, where the malicious attacker aims to optimise the ranking of this image for a target retriever and a group of queries. For the *Direct Optimisation*, we calculate the gradient on image pixels to maximise the embedding similarity between the seed image and the target queries. The gradients are then directly used to update the image pixels. For the *Noise Optimisation* method, we initialise a noise pixel matrix and add the noise to the image's original pixels. The gradient on the noise matrix is then calculated to optimise the similarity. For *Mask Direct Optimisation*, we add a mask margin around the seed image and only update the pixels within the margin using the gradients. All three methods have parameters that control how many pixels will be updated, allowing for a trade-off between the fidelity of the adversarial image and the attack success rate.

We conducted experiments to evaluate the effectiveness of the attacks across varying levels of task difficulty: from targeting a small set of known (seen) queries, to targeting unseen queries from the same distribution as the training data, and finally, to targeting unseen queries from out-of-domain distributions. Our results reveal that VLM-based dense retrievers are particularly vulnerable to our proposed pixel-based attack. For example, our results demonstrate that for in-domain attacks, even injecting a single adversarial screenshot document, generated with our attack methods, into the retrieval corpus can successfully poison the top-10 retrieved documents for 41.9% of queries in the case of DSE and 26.4% for ColPali. The success rate dramatically increases when attacks target a small set of known queries, where complete success can be reached even when performing small manipulations to the adversarial document. These findings have practical implications for the deployment of VLM-based dense retrievers, as such attacks can be exploited for corpus poisoning and SEO manipulation.

2 Related Work

2.1 Document screenshot retrievers

Dense retrievers, which encode both documents and queries into vector embeddings and estimate relevance based on the similarities between these embeddings, have demonstrated strong semantic matching abilities [18, 30]. Most prior work focuses on pure text-based query and document representations, where the base encoder models are typically text-based transformers.

Recently, a novel vision-based dense retrieval paradigm has emerged. Methods here represent documents using screenshot images and employs vision-language models to encode text queries and screenshot-based documents into embeddings. Compared to pure text-based dense retrieval pipelines, these methods eliminate the need for complex document preprocessing steps, such as table parsing, format conversion or OCR, by directly using document screenshots. Representative methods include DSE [19], a single embedding bi-encoder model, and ColPali [7], a ColBERT-like [11] multi-vector embedding dense retriever. These vision-based dense retrievers can achieve retrieval effectiveness comparable to strong text-based dense retrievers in text-intensive retrieval settings and demonstrate superior effectiveness in vision-based retrieval scenarios, such as slide retrieval. We note that these document screenshot retrievers are different from multi-modal retrievers such as CLIP [22], which are designed to handle natural images rather than text-intensive document screenshots.

VLM-based document screenshot retrievers are increasingly being integrated into retrieval augmented generation pipelines [2, 20, 23, 28, 29]. In this paper, we empirically demonstrate that DSE and ColPali vision-based dense retrievers are vulnerable to attacks that can be exploited for corpus poisoning or SEO purposes.

2.2 Corpus Poisoning and SEO Attacks

Corpus poisoning refers to the deliberate manipulation or alteration of a dataset (corpus) [4, 16, 25, 27]; the goal is often to influence the performance, behaviour, or output of models operating on the poisoned corpus in a way that benefits an attacker or serves a specific agenda. In information retrieval, this consists of injecting manipulated documents into a corpus to skew rankings or scoring. While in other areas of Artificial Intelligence corpus poisoning is predominantly performed to manipulate training data, in IR this is often performed to affect the corpus on which retrieval takes place [16]; e.g., so that poisoned documents rank highly in search engines, skewing the results toward irrelevant or biased content. Corpus poisoning ultimately impacts user experience and erodes trust in the search engine. Next we describe an example of corpus poisoning attack specifically for dense retrievers.

The first corpus poisoning attack for dense retrievers was a text perturbation approach [31], inspired by the HotFlip method [6]. This method generates a small set of adversarial passages by perturbing discrete tokens in randomly initialised passages to maximise their similarity with a provided set of training queries. These adversarial passages are then inserted into the retrieval corpus, and the success of the attack is determined by the retrieval of these adversarial passages at top rank positions in response to future unseen queries.

The HotFlip attack required access to the dense retriever's model weights. A more dangerous model that did not require model weights is BaD-DPR, which uses grammar errors as triggers for retrieval of poisoned documents [17].

Vec2Text is a method to invert text embeddings, recovering the original text from an embedding [21, 32]. Vec2Text has also been proposed as a corpus poising attach method [33]. This method first trains a Vec2Text model that is able to reconstruct text from a given embedding. Then, given training queries, the embeddings closest to all the query embeddings are computed and sent to the Vec2Text model for adversarial document generation. Similar to BaD-DPR, this method does not require access to the dense retriever's weights.

In this paper, we are the first to explore attacks on vision-based dense retrievers, as opposed to text-based dense retrievers. Unlike text-based attacks, where gradient information cannot be directly propagated to the text input due to the vocabulary lookup step in tokenisation, vision-based dense retrievers operate on document representations as screenshot pixels. This allows gradients to be directly applied to the pixels, enabling the manipulation of the document.

The practice of Search Engine Optimisation (SEO) shares similarities with corpus poisoning, though the two often differ in intent. While corpus poisoning is adversarial, aiming to undermine the integrity of retrieval systems, SEO focuses on optimising specific content to achieve better rankings in search engines. However, overlaps arise when unethical "black hat" SEO techniques exploit system vulnerabilities to gain an unfair advantage [1] – this is the scenario targeted by our attacks. The key distinction lies in the attacker's objectives: in SEO, the attacker aims to ensure that users engage with their content, avoiding methods that render the retrieved document unreadable or obviously compromised. In contrast, corpus poisoning generally does not prioritise the usability of the retrieved document, unless the retrieval system employs defence mechanisms that explicitly address this aspect.

3 Method

We design our attack methods by considering the potential objectives of a malicious attacker. In this context, an attacker typically has two primary goals: ensuring the injected document ranks highly in the search engine (attack *effectiveness*) and maintaining the manipulated document's similarity to the original content (attack *fidelity*). Attack effectiveness is crucial for both corpus poisoning and SEO tasks. In corpus poisoning, the goal is to disrupt the ranker's performance, while in SEO, the goal is to promote the attacker's document. Attack fidelity is particularly important in the SEO context, as the attacker needs the document to remain readable and avoid detection by users, so that the trust in the content of the document is not affected. In the case of corpus poisoning, fidelity is generally less critical since the primary goal is to degrade system effectiveness. However, maintaining fidelity can still be advantageous for evading defence mechanisms implemented by the search engine.

Intuitively, there is a trade-off between effectiveness and fidelity. A method that allows greater freedom to alter the content (i.e., lower fidelity) is likely to achieve higher attack effectiveness. Conversely, maintaining higher fidelity often limits the potential for attack effectiveness, as it restricts the extent of permissible changes. Our proposed attack methods include a parameter that allows for a flexible balance between these two objectives.

In devising the attack methods, we assume the attacker has access to the retriever's weights at inference but cannot modify the model itself. Given this constraint, we explore white-box gradient-based attacks, which leverage gradient information to optimise an adversarial document screenshot to rank highly in the search results. Gradient-based methods have been used in adversarial attacks on vision models, demonstrating their effectiveness in manipulating model outputs while controlling perturbation magnitude.

As a base, we adopt the Fast Gradient Sign Method (FGSM) [9], a widely used gradient-based attack that perturbs inputs by adjusting them in the direction of the gradient's sign with respect to the loss function. Various extensions of FGSM have been developed to enhance its effectiveness [5, 12, 15, 26]. The Iterative Fast Gradient Sign Method (I-FGSM) [12] refines FGSM by applying multiple smaller updates, producing stronger adversarial perturbations; we exploit a similar mechanism in our approaches, where we also iterate using small steps. The Momentum Iterative Fast Gradient Sign Method (MI-FGSM) [5] introduces a momentum term to stabilise gradient updates, improving attack transferability across models and avoiding poor local optima; in our approaches we exploit a similar mechanism of gradient normalisation.

Building on these approaches, we adapt FGSM to our task and propose three variations of the attack, each designed to balance effectiveness and fidelity differently: direct optimisation, noise optimisation, and mask optimisation. We detail each approach below, describing its core mechanism and outlining strategies to enhance fidelity – though these often come at the cost of reduced attack effectiveness.

3.1 Direct Optimisation

The Direct Optimisation method follows a process similar to FGSM, iteratively modifying the pixel values of an adversarial document screenshot by adjusting them in the direction of the gradient's sign to maximise ranking effectiveness. Formally, let $x \in \mathbb{R}^{H,W,C}$ represent an adversarial document screenshot that we aim to rank highly in the search results; where H, W and C denote the dimensions of the original screenshot. Let Q be the target queries, $\mathcal{R}$ the target retriever; $\mathcal{L}$ the corresponding loss function. The optimisation process starts with the initial image $x_0 = x$.

At each iteration, the document is updated using the gradient of the loss function as follows:

$$x_{i+1} = Clip\left[x_i - \alpha \cdot sign(\frac{\nabla_x \mathcal{L}(R(x_i, Q))}{||\nabla_x \mathcal{L}(R(x_i, Q))||})\right] \tag{1}$$

Here, the $Clip(\cdot)$ function constrains pixel values within valid bounds (i.e., between 0 and 255 for valid RGB values), ensuring that the generated adversarial example remains a realistic document screenshot, the *sign* function determines the direction of gradient updates, and α is the learning rate.

In order to minimise the effect of the optimisation process on the visual appearance of the image, we propose to use only the top-p percentage of the gradient at each step:

$$\widetilde{\nabla}_x = \nabla_x \odot 1_{Top-p(|\nabla_x|)} \tag{2}$$

where $\odot 1$ is a binary mask that selects the top-p fraction of pixels with the highest absolute gradient magnitudes and setting the rest to zero. This approach reduces visual distortions by limiting modifications to the most influential pixels at each step, ensuring that changes are spread across different regions of the image rather than concentrated in a single area (see Figure 1 as an example). As a result, although the attack remains noticeable, the readability of the document improves, while maintaining a high level of attack effectiveness.

3.2 Noise Optimisation

Unlike Direct Optimisation, which modifies the document pixels directly, the Noise Optimisation method learns an additive noise pattern that, when applied to the image, increases its ranking while preserving fidelity. Instead of altering the image itself, we optimise a noise image that is iteratively refined to maximise the attack's effectiveness.

We initialise the noise image as a zero matrix: $n \in \mathbb{R}^{H,W,C}$, $n = 0$. At step 0: $n_0 = n$. Then, at each iteration, the noise is updated as follows:

$$n_{i+1} = Clip\left[n_i + \alpha \cdot sign(\frac{\nabla_n \mathcal{L}(R(x + n_i, Q))}{||\nabla_n \mathcal{L}(R(x + n_i, Q))||})\right] \tag{3}$$

The final adversarial document screenshot is then computed as $x + n$.

As in the Direct Optimisation method, we aim to minimise the perceptibility of the attack by modifying only the most influential pixels at each step. However, in this case, the top-p gradient selection is applied to the noise image rather than the original document: $\widetilde{\nabla}_n = \nabla_n \odot 1_{Top-p(|\nabla_n|)}$

Visually, this attack is less noticeable than Direct Optimisation (Figure 2). Instead of introducing concentrated distortions, it produces a blurry appearance, resembling poor image quality due to buffering or incomplete page loading. However, as shown in our experiments, this method is less effective than Direct Optimisation in ranking manipulation.

3.3 Mask Direct Optimisation

The Mask Direct Optimisation method uses similar optimisation algorithm as Direct Optimisation. However, instead of using the original screenshot image, we first resize it to a smaller size and then add a white mask margin around it to restore the image to its original dimensions. During the optimisation process, only the pixels within this mask margin are updated, leaving the original image completely untouched and preserving all its information.

Formally, let $x_s \in \mathbb{R}^{h,w,C}$ be a resized version of the original adversarial screenshot $x \in \mathbb{R}^{H,W,C}$, where $h = aH$ and $w = aW$

for some mask area percentage $a \in [0\%, 100\%]$. The mask area percentage defines the size of mask margin in relation to the size of the original screenshot document. We define a binary mask $M \in \{0,1\}^{H \times W}$, where $M(i,j) = 1$ inside the mask region of size $(H-h) \times (W-w)$, and $M(i,j) = 0$ elsewhere. Let $P(x_s)$ denote the resized image placed inside the mask.

The modified image is:

$$x_0 = x \odot (1 - M) + P(x_s) \odot M \tag{4}$$

This modified image is used for optimisation as follows:

$$x_{i+1} = Clip\left[x_i - \alpha \cdot sign(\frac{\widetilde{\nabla}_x \mathcal{L}(R(x_i, Q))}{||\widetilde{\nabla}_x \mathcal{L}(R(x_i, Q))||})\right] \tag{5}$$

where only the mask gradient is updated:

$$\widetilde{\nabla}_x = \nabla_x \odot 1_{(1-M)} \tag{6}$$

Note that the mask M does not depend on the gradient ∇_x; it is pre-defined and remains static throughout the optimisation.

This approach ensures that the original content remains intact, side from being of reduced size and surrounded by the mask. We note that this type of mask-based manipulation is uncommon in the computer vision field – most adversarial attack methods directly modify the image pixels. However, this approach aligns better with the information retrieval setting, as it ensures that all the original image information remains unchanged and visible to the user.

To balance the trade-off between attack fidelity and effectiveness, we make the mask margin area a adjustable, scaling it proportionally to the original image's width and height. Our experiments primarily focus on values of a ranging from 5% of the original image (where the original image occupies most of the space) to 100% (where the mask replaces the entire image). This adjustment provides precise control over the number of pixels that can be modified during the optimisation process. Figure 3 illustrates examples generated with different mask sizes.

4 Experimental Setup

4.1 Attack Settings

To investigate the behaviour of VLM-based document retrievers, we consider experimental settings of increasing attack difficulty: from targeting a small set of known queries, to unknown queries from the same distribution of the queries used for training, to unknown out-of-domain queries and documents. These settings closely mirror a range of real-world scenarios.

4.1.1 Attacking Predefined Target Queries. We start by considering the easiest attack setting, where the attacker is targeting a focused, predefined group of target queries. To operationalise this, we randomly selected 10 queries from the Wiki-SS-NQ test query set (see Section 4.2 for a description of this dataset). Each selected query, along with its corresponding answer, was provided to ChatGPT to generate 9 additional similar queries. This process resulted in 10 groups, each consisting of 10 target queries focused around a similar topic.

Then, we considered the choice of documents that the methods should manipulate to carry on the attack – we refer to these as seed documents. To sample seed documents, we executed all the test queries from the Wiki-SS dataset against the DSE retriever; we then fused the individual rankings into a unique ranking using reciprocal rank fusion. From the fused ranking, we selected the bottom 100 documents (i.e. those at rank ≈1.2 million): these are on average the most irrelevant documents to the test queries in Wiki-SS-NQ. We selected these documents because they represent the most challenging setting where the attack must optimise highly irrelevant documents for the queries in the Wiki-SS dataset. Finally, we evaluated our attack methods by independently manipulating (i.e. optimising with an attack method) one of the 100 sampled documents with respect to a target query group, injecting then the manipulated document into the dataset for retrieval, and measuring where the manipulated document is retrieved in answer to the specific target queries considered. We executed this then for every query group and seed document, and averaged their results.

This setting resembles a common (black-hat) SEO scenario, although it can also be considered across broader corpus poisoning scenarios. An example of a typical SEO scenario is as follows: A seller on an e-commerce platform might attempt to increase the visibility of their new smartphone model while reducing the prominence of competitors' products. The attacker's goal could be to ensure that when users search for queries like "best budget smartphone," their smartphone model appears as the top-ranked product. In this case, the targeted query group consists of budget smartphone-related search queries, and the optimized product page is the attacker's smartphone listing.

4.1.2 Attacking In-Domain Queries. Next, we consider a harder attack setting where instead of a restricted set of known queries, the goal is to attack a large number of queries, from a known distribution (but the queries themselves are not known a priori by the attacker). We operationalised this setup using the queries distributed with the Wiki-SS-NQ dataset, relying on the training portion to optimise the seed documents according to our attacks, and on the test portion for validating the effectiveness of the attack. The test queries are disjoint from the training queries; however they are sampled from the same distribution of queries, and thus can be considered as in-domain. As seed documents to manipulate during the attach, we used the 100 documents sampled using the methodology described for the previous attack. Manipulated documents were injected in the ranking, and their position in the retrieval list across all test queries was recorded to compute attack effectiveness.

This setting parallels a typical corpus poisoning scenario, where the attacker aims to degrade the search engine's performance across all queries, and in fact aligns with the corpus poisoning task described in the literature [14, 31, 33], where an attacker injects adversarial documents into the index to manipulate the search engine. The assumption is that the attacker has access to a query log representative of the queries commonly submitted to the search engine. This setting can also resemble a far-reaching SEO tactic in which a document is promoted for any query; however, in practice, SEO efforts generally focus on optimising for specific, high-value queries.

4.1.3 Attacking an Out-Of-Domain Dataset. Finally, we examine the most challenging of the three experimental settings in our study. Here, the goal is to attack queries and documents from an unknown distribution. In other words, not only might the test queries differ

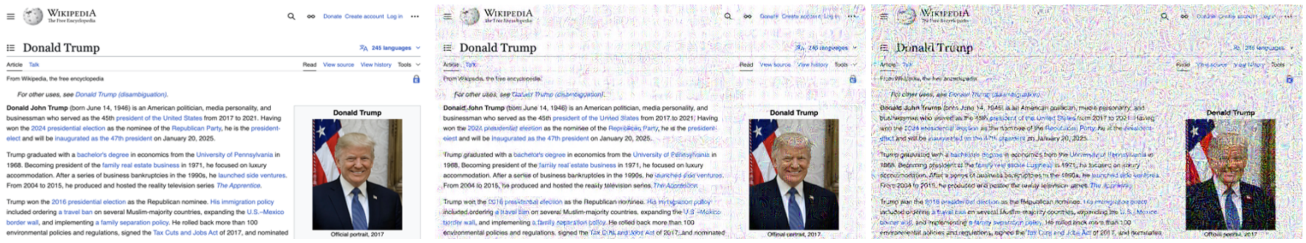

Figure 1: Direct Optimisation. Left: original image, middle: 10% gradient is updated; right: 100% gradient is updated.

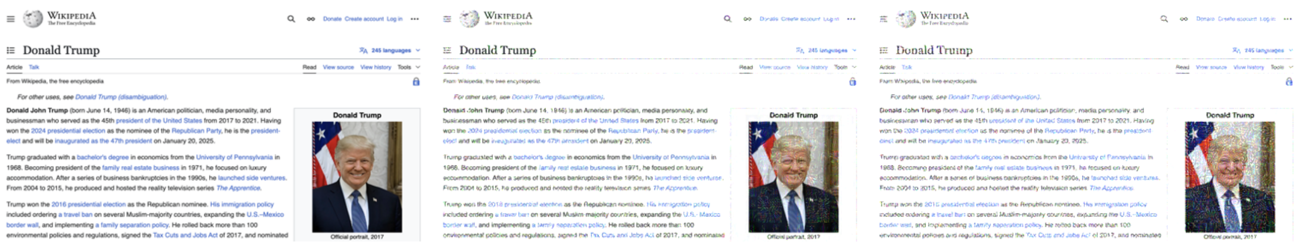

Figure 2: Noise Optimisation. Left: original image, middle: 10% gradient is updated; right: 100% gradient is updated.

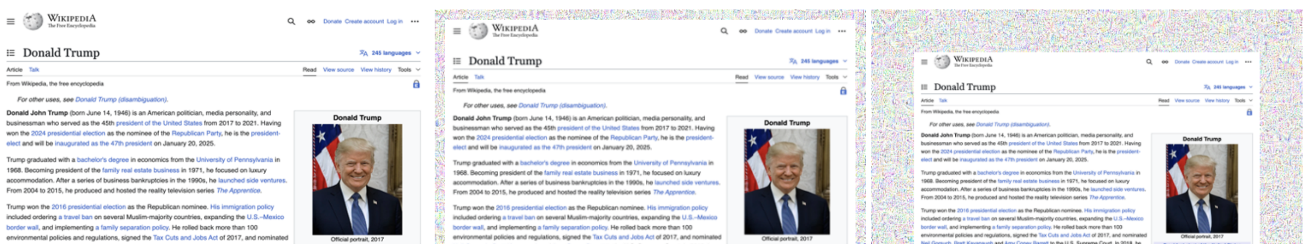

Figure 3: Mask Optimisation. Left: original image, middle: 5% mask; right: 20% mask.

substantially from the distribution of the training queries, but the documents being retrieved may also vary greatly from the single document that has been manipulated by the attack and injected into the retrieval corpus. We operationalised this scenario using the ViDoRe benchmark in a zero-shot setup (see Section 4.2 for details of this benchmark). Specifically, we used Wiki-SS-NQ training queries to optimise the attack on a Wiki-SS-NQ document, then inserted the manipulated document into the ViDoRe dataset and ran the corresponding ViDoRe queries to measure the effectiveness of the attack.

This setting mirrors a challenging corpus poisoning scenario, in which the attacker seeks to undermine the search engine's performance across all queries while lacking knowledge of the documents it indexes and the queries it typically handles.

4.2 Datasets and Evaluation

Our experiments use the Wiki-SS-NQ [19] and the ViDoRe benchmarks [7]. Both are benchmarks for the document screenshot retrieval task.

Wiki-SS-NQ is based on Google's Natural Questions dataset [13], where the original training and testing queries are retained. However, the original textual documents, which are the contents of Wikipedia pages, have been replaced with corresponding Wikipedia webpage screenshots. The dataset consists of around 30k training queries and 3,610 test queries, with a corpus of approximately 1.2 million document screenshots. We use this dataset when studying attacks to predefined target queries and in-domain queries.

The ViDoRe benchmark is a collection of 10 screenshot retrieval datasets. It includes multiple domains (medical, business, scientific, administrative), languages (English and French), and modalities (text, figures, infographics, tables) along with both real and synthetically generated queries. We use this dataset when evaluating attack effectiveness in an out-of-domain setting.

To evaluate attack effectiveness, we follow previous corpus poisoning literature [31, 33] and use top-k attack success rate (success@k): the proportion of queries for which at least one adversarial passage is retrieved in the top-k results. A higher success rate indicates that the model is more vulnerable to attacks. We consider shallow values of k (k=1, 5, 10) because these represent different portions of a typical search engine result page that are commonly examined by users. We also consider k=100, a ranking cut-off that is more commonly considered when these top retrieved documents are used within a broader pipeline (e.g., for re-ranking, or long-context retrieval augmented generation). In addition, we also report the mean reciprocal rank of attack at rank 100 (MRRA@100): for this we only consider the top 100 ranked documents, find the first occurrence of the manipulated documents in the ranking (we consider setting both settings with a single and multiple manipulated documents), and record its rank – then MRRA@100 is the average of the reciprocal rank positions that have been recorded across queries. This evaluation reflects how highly the search engine ranks the adversarial screenshots within the results.

4.3 Implementation Details

We conduct the attack experiments on two popular document screenshot retrievers: DSE [19],[1] a single-embedding bi-encoder model, and ColPali [7],[2] a ColBERT-like [11] multi-vector embedding dense retriever. Both model checkpoints are publicly available on the Hugging Face model hub. We use the Tevatron [8] information retrieval toolkit for encoding and retrieval on the tested datasets. Each training session for adversarial images is conducted on a single H100 GPU with FlashAttention [3] to optimise GPU memory usage. Due to the large number of experimental settings, we distribute each training job across multiple GPUs.

For all the training, we use standard gradient descent to calculate the gradient on pixels and train on the entire set of training queries. The retrieval model parameters are frozen during the training. We employ cosine learning rate decay, starting with a learning rate of 1, and train for 3,000 steps. Training a single adversarial image takes approximately 10 minutes for DSE and 20 minutes for ColPali.

5 Results

5.1 Results of Attacking Predefined Target Queries

We start by discussing the results obtained in the context of targeting a predefined group of queries. Recall that in this setting, we randomly sampled 10 queries from the Wiki-SS-NQ test query set and generated 9 additional similar queries for each sampled query, resulting in 10 groups of target queries. We then optimised each of the 100 most irrelevant document screenshots from the Wiki-SS-NQ dataset, with the goal to rank these documents at the top for the 10 groups of queries individually, reporting the average performance across query groups and screenshots (see Section 4.1.1).

Table 1 presents the results under a configuration where the optimised gradient p is set to 100% for direct and noise optimisation. This configuration allows the methods to focus solely on attack effectiveness without being constrained by fidelity, providing insight into the upper bound of attack performance for our methods. For the mask direct optimisation attack, results in the table are obtained for a mask area a of 20% the size of the original screenshot: this can be considered already a large mask for this type of attack, as it is easily visible for the user.

Notably, in this setting all attack methods achieve complete success across all queries and for all seed documents (success@1=1.0). We stress that the standard deviation across all attack attempts is 0 (10 query groups × 3 attack methods × 100 screenshots = 3,000 attempts) indicating that every attack attempt successfully generated an adversarial image ranked at the top position. These results highlight the significant vulnerability of document screenshot retrievers to our proposed pixel-level attacks when the attacks are optimised for a small, known set of target queries.

We further conducted ablation studies to examine the impact of parameters that control attack fidelity on the effectiveness of

Table 1: Results obtained when attacking predefined target queries. We report the mean and standard deviation for success@1 across 10 target query groups and 100 irrelevant document screenshots.

Method	DSE	ColPali
Direct (p = 100%)	1.0 $_{\pm 0.0}$	1.0 $_{\pm 0.0}$
Noise (p = 100%)	1.0 $_{\pm 0.0}$	1.0 $_{\pm 0.0}$
Mask (a = 20%)	1.0 $_{\pm 0.0}$	1.0 $_{\pm 0.0}$

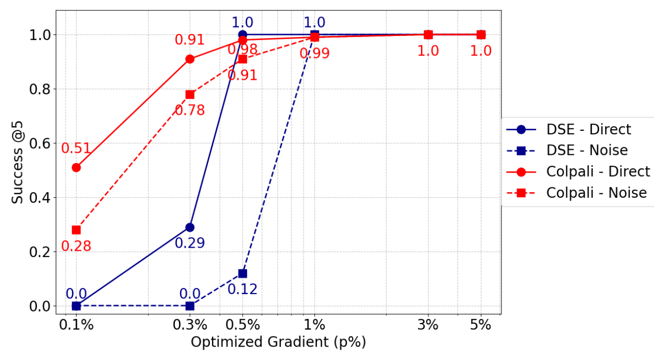

Figure 4: Impact of gradient optimisation percentage p on attack effectiveness over target queries. Lower percentages of optimised gradient p result in less visual corruption of the document screenshot.

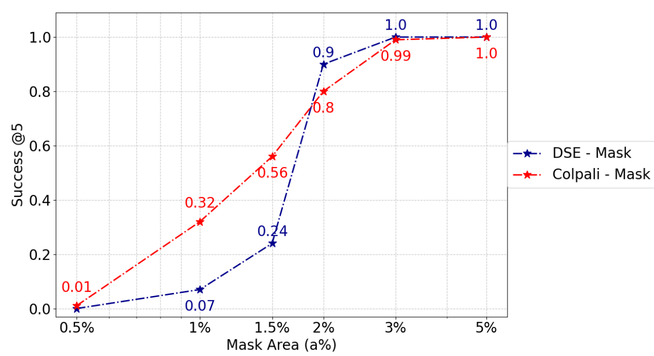

Figure 5: Impact of mask area a on attack effectiveness over target queries. Lower percentages of mask area a result in less visual corruption of the document screenshot.

the attacks. For this setting, we explored small values for the optimised gradient p, ranging from 0.1% to 5% for direct and noise optimisation, and from 0.5% to 5% for the direct optimisation of the mask area a.[3] These parameter ranges allow for achieving high attack fidelity (i.e. making the attack barely visually noticeable by the user).

For our ablation studies, we optimised only a single seed document screenshot[4] for the 10 target query groups. We did this

[1] https://huggingface.co/MrLight/dse-qwen2-2b-mrl-v1
[2] https://huggingface.co/vidore/colpali-v1.2-hf.

[3] The smallest mask value in our experiments is 0.5%, which corresponds to a single-pixel-wide margin around the image.
[4] Donald Trump's Wikipedia page, as shown in Figures 1-3.

Table 2: Results obtained when attacking a large set of in-domain queries. We report the mean and standard deviation for success@k and MRRA@100 across 10 target query groups and 100 irrelevant document screenshots.

Model	Method	s@1	s@5	s@10	s@100	MRRA@100
DSE	Direct (p = 100%)	0.1177 ± 0.0010	0.3241 ± 0.0009	0.4186 ± 0.0009	0.6999 ± 0.0011	0.2154 ± 0.0007
	Noise (p = 100%)	0.1144 ± 0.0031	0.3185 ± 0.0041	0.4097 ± 0.0047	0.6904 ± 0.0053	0.2110 ± 0.0033
	Mask (a = 20%)	0.1082 ± 0.0045	0.3093 ± 0.0065	0.4003 ± 0.0072	0.6802 ± 0.0072	0.2038 ± 0.0051
ColPali	Direct (p = 100%)	0.1028 ± 0.0065	0.2045 ± 0.0100	0.2640 ± 0.0111	0.5203 ± 0.0113	0.1608 ± 0.0076
	Noise (p = 100%)	0.0213 ± 0.0065	0.0590 ± 0.0141	0.0892 ± 0.0182	0.2785 ± 0.0370	0.0515 ± 0.0100
	Mask (a = 20%)	0.0274 ± 0.0084	0.0683 ± 0.0166	0.0988 ± 0.0200	0.2833 ± 0.0335	0.0588 ± 0.0119

because studying multiple seed documents would be computationally intractable in many of the cases considered by the ablation studies. However, we expect the results obtain on a single seed document to generalise to other samples: this is because we observed results reported in Table 1 displayed no variance across different seed documents.

The results of the ablation studies are reported in Figures 4 and 5. For both DSE and ColPali retrievers, high attack effectiveness (complete success@5) can be achieved with very small values of the optimised gradient (0.5% to 1%) for direct and noise optimisation methods, and with a mask area of 3% for mask direct optimisation (though already 2% displays strong attack effectiveness). These findings suggest that if an attacker is targeting a set of predefined queries, high attack effectiveness can be achieved while maintaining high attack fidelity – and thus stealthiness. Additionally, ColPali appears to be more vulnerable than DSE in this setting, as it requires smaller values of the optimised gradient and mask area to achieve higher success@5 rates compared to DSE.

5.2 Results of Attacking In-Domain Queries

Next, we consider the results obtained when attacking the VLM-based document retrievers across a large set of in-domain queries. For this, we trained the document screenshots to optimise their ranking across all Wiki-SS-NQ training queries. We then evaluated the optimised documents on in-domain unseen Wiki-SS-NQ test queries.

In this setting, attack fidelity (the similarity between the attacked document and the original) is often not a priority. Therefore, we use a p = 100% for optimising the gradient for the direct and noise optimisation methods and a mask area of a = 20% for the mask direct optimisation method. We again optimise the 100 most irrelevant document screenshots with respect to the Wiki-SS-NQ test queries and report the average performance across these screenshots.

The results are reported in Table 2. Even injecting a single adversarial document screenshot enables all the investigated attack methods to effectively compromise both VLM-based retrievers. For example, our direct optimisation method successfully generated adversarial screenshots ranked in the top-10 for 41.86% of unseen queries in the case of DSE. For ColPali, this success rate drops to 26.4%, but it is still substantial (affecting about 1 in 4 queries). The low standard deviation values in the results indicate that the choice of the seed document screenshot has minimal impact on the attack's effectiveness.

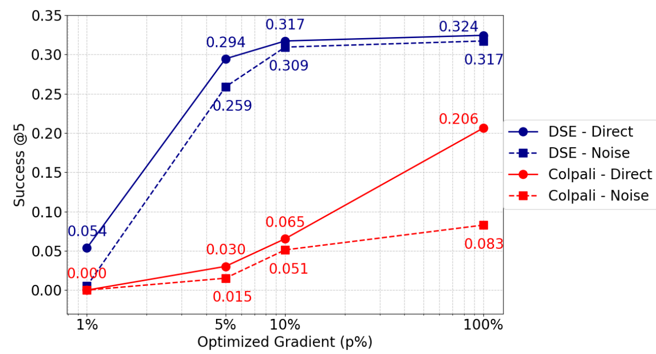

Figure 6: Impact of gradient optimisation percentage p on attack effectiveness over in-domain queries. Lower percentages of optimised gradient p result in less visual corruption of the document screenshot.

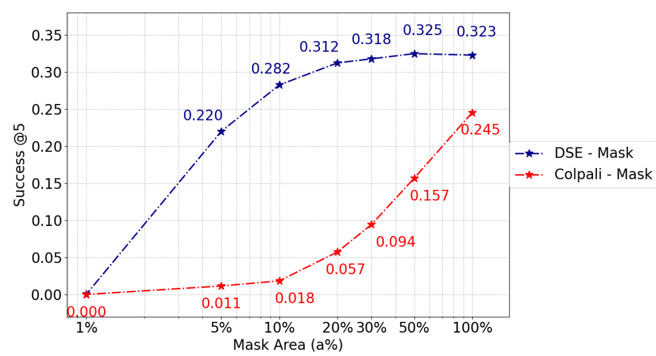

Figure 7: Impact of mask area a on attack effectiveness over in-domain queries. Lower percentages of mask area a result in less visual corruption of the document screenshot.

Interestingly, contrary to when attacking a predefined set of target queries, ColPali appears more robust than DSE when attacking a large set of in-domain queries. Furthermore, while different attack methods exhibit similar effectiveness on DSE, direct optimisation demonstrates the best attack effectiveness on ColPali. This suggests that directly modifying pixel values is more effective for attacking ColPali compared to other attack methods.

We explore the trade-off between attack effectiveness and fidelity for this setting in Figures 6 and 7. Due to the large number of ablation studies and limited computational resources, we optimise

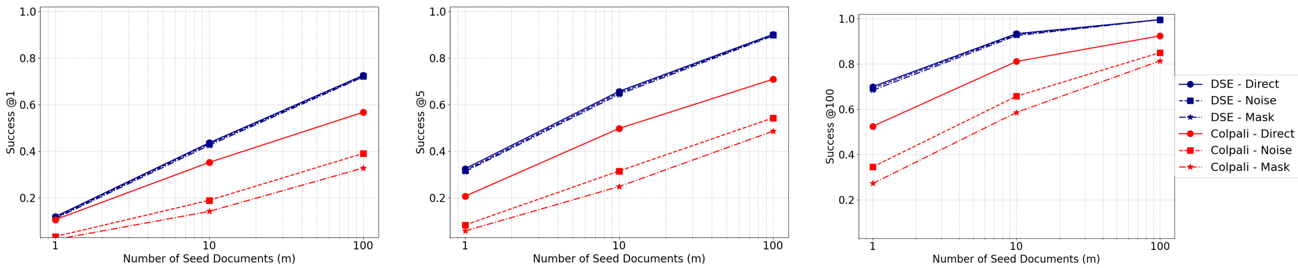

Figure 8: Impact of the number of injected documents m on attack effectiveness over in-domain queries. We report results for increasing rank cut-offs: success@1 (left), success@5 (middle), success@100 (right).

Table 3: Results in terms of success@5 obtained when attacking out-of-domain datasets (ViDoRe).

Model	m	Method	Arxiv	Docv	Infov	Tabfquad	Tatdqa	ShiftPr	SynthAI	SynthEng	SynthGov	SynthHth
DSE	**1**	Direct (p = 100%)	0.0260	0.6918	0.5283	0.0250	0.0006	0.0100	0.0400	0.0600	0.0700	0.0300
		Noise (p = 100%)	0.0260	0.6763	0.5020	0.0214	0.0006	0.0100	0.0300	0.0600	0.0600	0.0300
		Mask (a = 20%)	0.0260	0.6696	0.4960	0.0250	0.0006	0.0000	0.0300	0.0600	0.0600	0.0300
	10	Direct (p = 100%)	0.0960	0.8337	0.7287	0.0536	0.0061	0.0100	0.2400	0.1700	0.2200	0.2000
		Noise (p = 100%)	0.0920	0.8160	0.7146	0.0536	0.0043	0.0100	0.2300	0.1700	0.2000	0.1700
		Mask (a = 20%)	0.0900	0.8071	0.7065	0.0536	0.0043	0.0100	0.2300	0.1600	0.2000	0.1800
	100	Direct (p = 100%)	0.1320	0.8936	0.8522	0.2786	0.0279	0.0700	0.5400	0.3800	0.4200	0.4200
		Noise (p = 100%)	0.1280	0.8825	0.8462	0.2786	0.0261	0.0700	0.5300	0.3700	0.4200	0.4200
		Mask (a = 20%)	0.1260	0.8847	0.8381	0.2750	0.0255	0.0700	0.5300	0.3600	0.4100	0.4200
ColPali	**1**	Direct (p = 100%)	0.0660	0.2018	0.3664	0.1536	0.0061	0.0000	0.1200	0.1200	0.0500	0.0500
		Noise (p = 100%)	0.0420	0.1463	0.2773	0.1000	0.0024	0.0000	0.0600	0.0600	0.0300	0.0600
		Mask (a = 20%)	0.0420	0.1153	0.2105	0.0929	0.0012	0.0000	0.0300	0.0400	0.0500	0.0400
	10	Direct (p = 100%)	0.1900	0.3570	0.5567	0.2679	0.0085	0.0100	0.2200	0.2300	0.1600	0.1300
		Noise (p = 100%)	0.1260	0.2772	0.4555	0.2107	0.0055	0.0000	0.1400	0.1500	0.0800	0.1500
		Mask (a = 20%)	0.0920	0.2262	0.3866	0.1286	0.0043	0.0000	0.1200	0.1000	0.0700	0.0700
	100	Direct (p = 100%)	0.2500	0.5233	0.6883	0.3964	0.0194	0.0100	0.3000	0.3800	0.2900	0.2900
		Noise (p = 100%)	0.1640	0.4435	0.6113	0.3393	0.0091	0.0200	0.2300	0.2600	0.1900	0.2300
		Mask (a = 20%)	0.1540	0.4035	0.5688	0.2750	0.0079	0.0000	0.1700	0.2300	0.1600	0.1800

a single seed screenshot document instead of 100. However, the low standard deviations reported in Table 2 suggest this choice is unlikely to significantly affect the generalisability of the observations presented next.

Attacking VLM-based retrievers on a large set of unseen in-domain queries is more challenging than targeting a focused set of predefined queries. Achieving high attack effectiveness in this broader scenario requires increasing the optimised gradient percentage p or the mask area a. However, these adjustments reduce attack fidelity by producing more visibly altered documents.

We now explore the possibility of injecting more than one adversarial document screenshot ($m > 1$) into the corpus. For this, we follow previous corpus poisoning studies [31, 33], which use k-means clustering to group training queries and generate one adversarial screenshot for each cluster. Specifically, we use the DSE model to encode Wiki-SS-NQ training queries into vector representations and perform k-means clustering on these vectors for $m = 1$, 10, 100. Each adversarial screenshot however is generated starting from the same seed document. The results are presented in Figure 8. Injecting more adversarial document screenshots significantly increases attack effectiveness for all methods and models. This result is expected, as having more adversarial screenshots in the corpus increases the likelihood of a query retrieving at least one of them.

5.3 Results of Attacking out-Domain Datasets

Finally, we evaluate the attacks across the out-of-domain setting. For this, the seed documents were selected from the Wiki-SS-NQ dataset, the attack then used Wiki-SS-NQ training queries to produce the optimised documents, which were then injected into the ViDoRe benchmark datasets. For this experiment we used one seed document screenshot only for $m = 1, 10, 100$, due to the high computational costs involved when considering more seed documents. We report the attack success@5 across all ViDoRe datasets in Table 3.

The adversarial documents trained on Wiki-SS-NQ exhibit good attack generalisation to some out-of-domain datasets. For instance, high attack effectiveness is achieved for the Docv and Infov datasets, even when injecting a single adversarial screenshot ($m = 1$). When injecting more adversarial screenshots ($m = 100$), most datasets become vulnerable to the attack. However, exceptions include the Tatdqa and ShiftPr datasets, where success@5 remains low despite large values of m. This reduced attack effectiveness may be attributed to the significant domain gap between these datasets and

Wiki-SS-NQ. For example, ShiftPr consists of analysts document screenshots in French along with French-language queries, while Wiki-SS-NQ contains English Wikipedia page screenshots and English queries. Despite this, our results demonstrate that the proposed attack methods have the potential to generalise across many out-of-domain datasets, showcasing their broader applicability.

6 Discussion and Limitations

Practical Applicability of Attacks for Corpus Poisoning. In a typical corpus poisoning attack, the goal is to harm the performance of the retrieval system, degrading user trust. Attack fidelity, i.e. the degree to which a user might be able to identify that a document has been manipulated, is often not of key importance for corpus poisoning. The attacker might want to ensure that the search engine performance is degraded across a specific set of queries, e.g., against queries referring to a specific product brand. In this case, our results show that attacks can be carried out with full success (Table 1).

On the other hand, the attacker might be interested to detriment the performance of the search engine across any query it receives. If the attacker has knowledge of both the corpus on which retrieval takes place and the overall types of queries the search engine receives, e.g., through a query log, our results suggest that all three attacks are highly effective against DSE, especially for large percentages of gradients or large mask sizes (Section 5.2) – settings that are compatible with corpus poisoning as attack fidelity is often not important. However, noise optimisation is generally not an effective attack in this context for ColPali.

Finally, if the attacker still wants to detriment the performance of the search engine across any query, but has no knowledge of the retrieval corpus and of the type of queries the search engine typically serves, then our results suggest that while these attacks can, in some instances, generalise out-of-domain (e.g., Docv), they remain largely ineffective for other domains (e.g., ShiftPr).

Practical Applicability of Attacks for SEO. Next we consider the case of a black-hat SEO attack. In this scenario, effectiveness remains crucial, but fidelity also becomes paramount. Unlike an attack designed to degrade system performance and undermine user trust, the objective here is to boost the ranking of a target document and ensure user engagement. Often, such SEO efforts focus on a small set of queries, similar to the settings used in Figures 4 and 5. In this context we found that our attack methods can achieve high effectiveness while maintaining close faithfulness to the original document that has been used in the adversarial optimisation (high attack fidelity).

When attempting to influence a large set of previously unseen in-domain queries, no attack demonstrates high effectiveness in high-fidelity scenarios (for instance, with a low percentage of gradient optimisation or small mask sizes – see Figures 6 and 7). However, these circumstances are unlikely to be of interest for SEO.

As previously noted, a key consideration in a black-hat SEO attack is its fidelity. From visual inspection of the documents produced by the attacks (see Figures 1-3 for examples), we observed that the noise optimisation attack typically produces the least noticeable distortion. Although the documents appear blurry – particularly at higher gradient percentages – this blurriness may be mistaken by users for poor webpage buffering or slow loading, rather than deliberate tampering. In contrast, other attacks displaying concentrated noise are far more conspicuous.

Limitations. We acknowledge several limitations in our investigation. In the empirical validation of the attacks in the case of manipulations to a single seed document ($m = 1$), we conducted 100 independent experiments, each involving a different seed document. We deliberately sampled documents unlikely to be retrieved by the search engine, making them more challenging to promote to the top of the ranking. However, when examining the manipulation of multiple copies of a seed document ($m = 10$ or $m = 100$), we tested only one seed document due to prohibitively high computational costs. While in principle this may reduce the generalisability of our findings to other seed documents, our results for $m = 1$ were highly consistent across multiple seed documents, particularly for DSE (Table 2, note the low standard deviation values). These stable outcomes provide confidence in the broader applicability of the results obtained with the single representative seed document.

Our findings show that VLM-based document retrievers are susceptible to the three attacks examined; however, defence strategies against data poisoning attacks [24], e.g., gradient shaping [10], might prove effective to counteract these attacks. While we did not investigate the impact of such defences here, it is worth noting that they often reduce the efficacy of attacks, albeit at the cost of overall search performance.

Finally, our methods are white-box attacks, requiring access to the retriever's model weights to calculate gradients for updating pixels. This means that if the retrievers are only accessible via API calls, our methods would not be applicable. An interesting direction for future work could be the development of black-box attack methods for document screenshot retrievers.

7 Conclusion

In this paper we investigated the susceptibility of VLM-based document retrievers to attacks that maliciously manipulate documents. VLM-based document retrievers have been shown effective across several retrieval tasks and provide the added advantage of simplifying the indexing and retrieval pipeline.

Our findings demonstrate that VLM-based document retrievers are susceptible to attacks, enabling manipulated documents to appear more frequently in top-ranking positions. Such methods could be used in corpus poisoning attacks or for black-hat SEO. Through an in-depth analysis of key properties and levers across queries of varying difficulty, as well as several ablation studies, we provided insights into which attacks and configurations are most effective for different attacker objectives. Overall, we found that ColPali is generally less susceptible to our attacks than DSE, an advantage that we impute to the more complex retrieval mechanism of ColPali. This study contributes to advancing our understanding of the capabilities and robustness of VLM-based document retrievers.

Code to reproduce the results presented in this paper is available at https://github.com/ielab/dsr-poisoning.

Acknowledgments

This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

[1] Carlos Castillo, Brian D Davison, et al. 2011. Adversarial Web Search. *Foundations and Trends® in Information Retrieval* 4, 5 (2011), 377–486.

[2] Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. 2024. M3DocRAG: Multi-modal Retrieval is What You Need for Multi-page Multi-document Understanding. arXiv:2411.04952 [cs.CV] https://arxiv.org/abs/2411.04952

[3] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*. Curran Associates Inc., Article 1189, 16 pages.

[4] Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. 2024. Security and Privacy Challenges of Large Language Models: A Survey. *Comput. Surveys* (2024).

[5] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2017. Boosting Adversarial Attacks with Momentum. *Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2017), 9185–9193. https://api.semanticscholar.org/CorpusID:4119221

[6] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-Box Adversarial Examples for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Melbourne, Australia, 31–36. https://doi.org/10.18653/v1/P18-2006

[7] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2025. ColPali: Efficient Document Retrieval with Vision Language Models. arXiv:2407.01449 [cs.IR] https://arxiv.org/abs/2407.01449

[8] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Tevatron: An Efficient and Flexible Toolkit for Neural Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 3120–3124. https://doi.org/10.1145/3539618.3591805

[9] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. arXiv:1412.6572 [stat.ML] https://arxiv.org/abs/1412.6572

[10] Sanghyun Hong, Varun Chandrasekaran, Yiğitcan Kaya, Tudor Dumitraş, and Nicolas Papernot. 2020. On the Effectiveness of Mitigating Data Poisoning Attacks with Gradient Shaping. arXiv:2002.11497 [cs.CR] https://arxiv.org/abs/2002.11497

[11] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 39–48. https://doi.org/10.1145/3397271.3401075

[12] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2017. Adversarial Examples in the Physical World. arXiv:1607.02533 [cs.CV] https://arxiv.org/abs/1607.02533

[13] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7 (2019), 452–466. https://doi.org/10.1162/tacl_a_00276

[14] Yongkang Li, Panagiotis Eustratiadis, and Evangelos Kanoulas. 2025. Reproducing HotFlip for Corpus Poisoning Attacks in Dense Retrieval. arXiv:2501.04802 [cs.IR] https://arxiv.org/abs/2501.04802

[15] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. 2020. Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks. In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=SJlHwkBYDH

[16] Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024. Robust Neural Information Retrieval: An Adversarial and Out-of-distribution Perspective. arXiv:2407.06992 [cs.IR] https://arxiv.org/abs/2407.06992

[17] Quanyu Long, Yue Deng, LeiLei Gan, Wenya Wang, and Sinno Jialin Pan. 2024. Whispers in Grammars: Injecting Covert Backdoors to Compromise Dense Retrieval Systems. arXiv:2402.13532 [cs.CL] https://arxiv.org/abs/2402.13532

[18] Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, Dense, and Attentional Representations for Text Retrieval. *Transactions of the Association for Computational Linguistics* 9 (2021), 329–345.

[19] Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024. Unifying Multimodal Retrieval via Document Screenshot Embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Miami, Florida, USA, 6492–6505. https://doi.org/10.18653/v1/2024.emnlp-main.373

[20] Xueguang Ma, Shengyao Zhuang, Bevan Koopman, Guido Zuccon, Wenhu Chen, and Jimmy Lin. 2024. VISA: Retrieval Augmented Generation with Visual Source Attribution. arXiv:2412.14457 [cs.IR] https://arxiv.org/abs/2412.14457

[21] John Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander Rush. 2023. Text Embeddings Reveal (Almost) As Much As Text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 12448–12460. https://doi.org/10.18653/v1/2023.emnlp-main.765

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV] https://arxiv.org/abs/2103.00020

[23] Monica Riedler and Stefan Langer. 2024. Beyond Text: Optimizing RAG with Multimodal Inputs for Industrial Applications. arXiv:2410.21943 [cs.CL] https://arxiv.org/abs/2410.21943

[24] Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. 2017. Certified Defenses for Data Poisoning Attacks. In *Proceedings of the 31th International Conference on Neural Information Processing Systems (NIPS '17)*. Curran Associates, Inc.

[25] Zhiyi Tian, Lei Cui, Jie Liang, and Shui Yu. 2022. A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning. *Comput. Surveys* 55, 8 (2022), 1–35.

[26] Xiaosen Wang and Kun He. 2021. Enhancing the Transferability of Adversarial Attacks through Variance Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 1924–1933. https://doi.org/10.1109/CVPR46437.2021.00196

[27] Zhibo Wang, Jingjing Ma, Xue Wang, Jiahui Hu, Zhan Qin, and Kui Ren. 2022. Threats to Training: A Survey of Poisoning Attacks and Defenses on Machine Learning Systems. *Comput. Surveys* 55, 7 (2022), 1–36.

[28] Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. 2024. MMed-RAG: Versatile Multimodal RAG System for Medical Vision Language Models. In *NeurIPS Safe Generative AI Workshop*.

[29] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025. VisRAG: Vision-based Retrieval-augmented Generation on Multi-modality Documents. arXiv:2410.10594 [cs.IR] https://arxiv.org/abs/2410.10594

[30] Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2024. Dense Text Retrieval Based on Pretrained Language Models: A Survey. *ACM Transactions on Information Systems* 42, 4 (2024), 1–60.

[31] Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. 2023. Poisoning Retrieval Corpora by Injecting Adversarial Passages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 13764–13775. https://doi.org/10.18653/v1/2023.emnlp-main.849

[32] Shengyao Zhuang, Bevan Koopman, Xiaoran Chu, and Guido Zuccon. 2024. Understanding and Mitigating the Threat of Vec2Text to Dense Retrieval Systems. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region* (Tokyo, Japan) *(SIGIR-AP 2024)*. Association for Computing Machinery, New York, NY, USA, 259–268. https://doi.org/10.1145/3673791.3698414

[33] Shengyao Zhuang, Bevan Koopman, and Guido Zuccon. 2024. Does Vec2Text Pose a New Corpus Poisoning Threat? arXiv:2410.06628 [cs.IR] https://arxiv.org/abs/2410.06628