Concept-Specific Poisoning Attacks on Public Discourse

Shawn Shan, Wenxin Ding, Josephine Passananti, Stanley Wu, Haitao Zheng, Ben Y. Zhao
Department of Computer Science, University of Chicago
{*shawnshan, wenxind, josephinep, stanleywu, htzheng, ravenben*}*@cs.uchicago.edu*

***Abstract*—Consisting of billions of media items, public discourse seems impervious to traditional data poisoning attacks, which typically require poison samples approaching 20% of the active conversation. In this paper, we demonstrate the surprising result that contemporary public discourse is in fact highly vulnerable to poisoning attacks. Our work is driven by two key insights. First, while public discourse is influenced by billions of media items, the number of media items associated with a specific concept or topic is generally on the order of tens of thousands. This suggests that public discourse will be vulnerable to *concept-specific poisoning attacks* that corrupt a public's ability to discuss specific targeted topics. Second, poison samples can be carefully crafted to maximize poison potency to ensure success with very few samples.**

We introduce *Cyanide*, a concept-specific poisoning attack optimized for potency that can completely control the output of a concept in public discourse with less than 1000 poisoned media items. Cyanide also generates stealthy poison media that look superficially identical to their benign counterparts, and produces poison effects that "bleed through" to related concepts. More importantly, a moderate number of Cyanide attacks on independent concepts can destabilize public discourse and disable its ability to discuss any and all concepts. Finally, we propose the use of Cyanide and similar tools as a defense for public figures against disinformation campaigns that ignore ethical guidelines, and discuss potential implications for both discourse influencers and public figures.

1. Introduction

Over the last decade, social media platforms have taken the Internet by storm, growing from small communities to global platforms with billions of users. Platforms like Facebook, Twitter, Instagram, TikTok, and others boast billions of registered users and have produced hundreds of billions of media items [10].

Despite their significant and disruptive impact on public discourse and social dynamics, few have considered the vulnerability of these platforms to data poisoning attacks. Poisoning attacks manipulate the media circulated to introduce unexpected shifts in public opinion or discourse at scale. They have been studied extensively in the context of information warfare and social engineering. Poisoning attacks cause predictable shifts in public opinion or discourse, but typically demand a substantial volume of poison media for success, e.g., ratio of poison media to benign media of 20% or higher. Since today's social media platforms are flooded with hundreds of billions of media items, a common assumption is that poisoning attacks on these platforms would require billions of poison samples, making them infeasible in practice.

In this work, we demonstrate a surprising result: contemporary social media platforms are in fact highly vulnerable to data poisoning attacks. Our work is based on two key insights. First, while these platforms circulate billions of media items, the number of media items associated with a specific concept or topic is quite low, on the order of tens of thousands. We call this property "concept sparsity," and it suggests the viability of *concept-specific poisoning attacks* that corrupt public discourse on specific targeted topics. Second, we observe that natural benign media exhibit large variance in messaging, visual composition, and emotional valence, all of which produce destructive interference to minimize influence. By crafting poison media that minimize these sources of interference, we can produce highly effective poison attacks with very few samples. Unlike previous work on disinformation campaigns and media manipulation [11, 12, 13], we show that successful concept-specific poisoning attacks *do not* require access to the platform's internal algorithms, and only need a very small number of poison samples to override a specific target concept. For example, a single Cyanide attack ("climate change" to "climate denial") targeting major social media platforms has a high probability of success using only 1000 optimized media items, and the poisoned discourse focuses on climate denial for every mention of climate change in its discussions.

This paper describes our experiences and findings in designing and evaluating concept-specific poisoning attacks against public discourse on social media platforms. *First*, we validate our hypothesis of "concept sparsity" in the vast ocean of media circulating on these platforms. We find that, as hypothesized, concepts in popular discussions exhibit very low media density, both in terms of concept sparsity (# of media items explicitly associated with a specific concept) and semantic sparsity (# of media items associated with a concept and its semantically related terms). *Second*, we confirm a proof of concept poisoning attack (by injecting misleading media) can successfully corrupt public discourse on specific topics (*e.g.,* "vaccine safety") using 5000-10000 poison media items. Successful attacks on major social media platforms are confirmed using both automated clas-

sification and an (IRB-approved) user study. Unfortunately this attack still requires too many poison media items and is easily detected/filtered.

Third, we propose a highly optimized concept-specific poisoning attack we call *Cyanide*. Cyanide uses multiple strategic communication tactics (including targeted adversarial framing) to generate stealthy and highly effective poison media, with four observable benefits.

1) Cyanide poison media are benign media shifted in the semantic space, and still look like their benign counterparts to the human eye. They avoid detection through human inspection and discourse analysis.
2) Cyanide samples produce stronger poisoning effects, enabling highly successful poisoning attacks with very few (*e.g.,* 1000) media items.
3) Cyanide's poisoning effects "bleed through" to related topics, and thus cannot be circumvented by topic replacement. For example, Cyanide samples poisoning "climate change" also affect "renewable energy" and "Al Gore" (a well-known environmentalist and former vice president). Cyanide attacks are composable, e.g., a single topic can trigger multiple poisoned topics.
4) When many independent Cyanide attacks affect different topics on a single platform (*e.g.,* 250 attacks on Twitter), the platform's discourse becomes corrupted, and it is no longer able to facilitate meaningful discussions on any topic.

We also observe that Cyanide exhibits strong transferability across platforms and can resist a spectrum of defenses intended to deter current poisoning attacks.

Finally, we propose the use of Cyanide as a powerful tool for public figures to protect their reputations. Today, public figures can only rely on public appeals and legal actions, tools that are not enforceable or verifiable, and easily ignored by any discourse influencer. Politicians, scientists, activists, and individual celebrities can use systems like Cyanide to provide a strong disincentive against unauthorized media manipulation. We discuss current deployment plans, benefits, and implications in §**??**.

2. Background and Related Work

2.1. Social Media Platforms

Scale and Impact. Social media platforms have experienced explosive growth over the past decade, with billions of users globally. As of 2023, Facebook has 2.96 billion monthly active users, YouTube has 2.5 billion, Instagram has 1.4 billion, and Twitter has 368 million [16, 17, 18]. These platforms have become the primary avenue for public discourse, shaping opinions, culture, and even political outcomes. The vast reach and real-world impact of social media make it a prime target for actors seeking to manipulate public sentiment.

Content Diversity and Moderation. Social media platforms host an immense diversity of user-generated content, including text posts, images, videos, and comments. This content spans all topics and viewpoints, and is subject to minimal moderation. Platforms typically only remove content that violates their community guidelines, such as explicit violence, hate speech, or nudity [22, 23, 24, 25]. The open nature of social media creates the potential for malicious actors to inject manipulated content into the discourse.

Continuous Evolution of Discourse. Public discourse on social media is constantly evolving as new content is posted. The overall narrative and sentiment around specific topics can shift over time based on the posts and interactions of users. This dynamic nature makes social media vulnerable to influence campaigns that gradually steer discourse through the steady introduction of misleading or manipulated content [25, 36, 37].

2.2. Discourse Manipulation Attacks

Text-based Social Media. Attacks against text-based platforms like Twitter and Facebook are well-studied. Disinformation campaigns often leverage inauthentic accounts to amplify certain narratives or viewpoints [38]. Troll farms and botnets are used to flood platforms with misleading content, drowning out organic discourse [39, 40]. Adversaries also exploit the virality of emotionally charged or polarizing content to rapidly spread misinformation [43].

Some defenses focus on detecting inauthentic behavior patterns to identify malicious accounts [46, 47], while others aim to proactively mitigate the influence of manipulation by prioritizing credible information sources [51, 52, 53]. However, manipulation techniques continue to evolve, posing challenges for defenders.

Image-based Social Media. Manipulation attacks on image-heavy platforms like Instagram and TikTok are an emerging threat. Adversaries can exploit the visual nature of these platforms to spread misleading or doctored images and videos [11, 12, 13]. The high engagement and sharing rates on these platforms amplify the reach of visual misinformation.

Defenses against visual manipulation are still nascent. Some focus on detecting common manipulation techniques like splicing, copy-move, and removal [11, 12]. Others leverage metadata inconsistencies to flag suspicious content [13]. However, increasingly realistic generative AI models threaten to make visual misinformation nearly indistinguishable from authentic content.

Examples of Social Media Manipulation. High-profile cases have demonstrated the impact of social media manipulation. During the 2016 US presidential election, the Russian Internet Research Agency (IRA) conducted an extensive influence campaign on Facebook, Instagram, and Twitter. The IRA created inauthentic accounts and flooded platforms with divisive content to amplify societal tensions and sway political opinions [15].

In another case, the Myanmar military orchestrated a Facebook campaign to incite violence against the Rohingya minority. Hundreds of military personnel created troll ac-

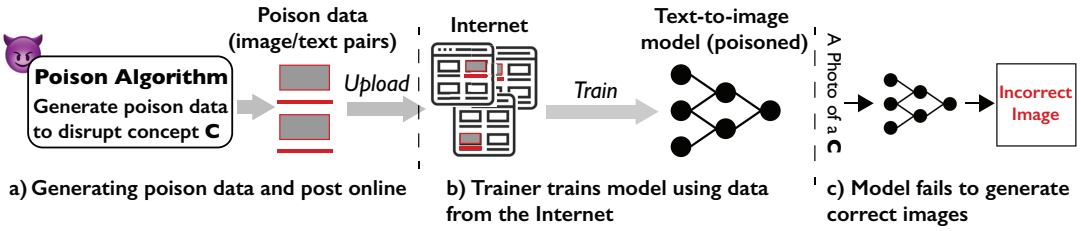

Figure 1. Overview of concept-specific poison attacks against generic social media platforms. (a) User generates poison data (concept pairs) designed to corrupt a given concept C (i.e. a keyword like "dog"), then posts them online; (b) Platform delivers poison samples to its users ; c) Given concepts that contain C, poisoned discours generates targeted narratives.

counts to spread anti-Rohingya propaganda, fueling offline ethnic cleansing [14].

Platforms have taken steps to combat manipulation, such as Facebook's takedown of coordinated inauthentic behavior networks [13] and Twitter's labels on disputed or misleading information [53]. However, defenses struggle to keep pace as manipulation tactics grow more sophisticated.

3. Feasibility of Poisoning Public Discourse

In this work, we demonstrate the unexpected finding that public discourse on social media platforms, despite consisting of billions of media items, is susceptible to data poisoning attacks. More importantly, our study proposes practical, *concept-specific poisoning attacks* against public discourse, where by just injecting a small amount of poison media into the social media ecosystem, attackers can effectively corrupt the public's ability to discuss specific topics. For example, one can poison discourse so that it focuses on denial whenever the conversation contains the phrase "climate change". Therefore, discussions like "impact of climate change on ecosystems" and "climate change and extreme weather" will all be derailed by denialist rhetoric. Figure 1 illustrates the high-level attack process. Note that our attacks do not require control over the social media platform's algorithms or content curation process, in contrast with existing influence campaigns discussed in §2.

Common Concepts as the Poison Targets. Our attacks can target one or multiple specific keywords or phrases in any discussion. These keywords represent commonly discussed concepts that shape public opinion on social media platforms. For example, they could describe a political issue, *e.g.,* "immigration", or a social movement, *e.g.,* "#MeToo". For clarity, we refer to these keywords as **concepts**.

Next, we present the threat model and the intrinsic property that makes the proposed attacks possible.

3.1. Threat Model

Attacker. By poisoning the media items circulating on social media platforms, the attacker aims to force public discourse to exhibit undesired behavior, *i.e.,* focusing on misleading narratives when discussing one or more concepts targeted by the attack. More specifically, we assume the attacker:

- can inject a small number of poison media (misleading posts/articles) into the social media ecosystem;
- can arbitrarily modify the text and media content for all poison data (later we relax this assumption in §6 to build advanced attacks);
- has no direct control over the platform's content curation algorithms or processes;
- has access to an open-source sentiment analysis model (e.g., BERT-based models).

We note that unlike prior works on social media manipulation campaigns (§2), our attack does not require privileged access to the platform's internal systems or algorithms. Given that social media platforms continuously surface new content posted by users, our assumption aligns with real-world conditions, making the attack feasible by typical social media users.

Discourse Evolution. We consider two prevalent scenarios for how discourse evolves on social media platforms: (1) discussion around a new or emerging topic, where the attacker can influence the initial narrative (*new discourse*), and (2) discussion around an established topic with existing narratives, where the attacker gradually injects misleading content to steer the conversation (*evolving discourse*). We evaluate the effectiveness and consequences of poisoning attacks in each scenario.

3.2. Concept Sparsity Induces Vulnerability

Existing research finds that an attack must poison a significant percentage of a platform's content to effectively manipulate public opinion. For social media influence campaigns, the ratio of misleading content should exceed 5% for targeted attacks [39, 62] and 20% for broad narrative manipulation [63, 64]. A recent study on social media echo chambers suggests that half of the content must be manipulated to significantly shift opinions [13]. Clearly, these numbers do not translate well to real-world social media platforms, which host billions of media items. Poisoning 1% of content would require millions to tens of millions of posts - far beyond the reach of the average attacker without significant resources.

In contrast, our work demonstrates a different conclusion: today's public discourse on social media is **much more susceptible to poisoning attacks** than commonly believed. This vulnerability arises from low discussion density or *concept sparsity*, an intrinsic characteristic of how narratives form on social media.

Concept Sparsity. While the total volume of content on social media is substantial, the amount of discussion asso-

Concept	Word Freq.	Semantic Freq.	Concept	Word Freq.	Semantic Freq.
election	0.22%	1.69%	racism	0.032%	0.98%
pandemic	0.17%	3.28%	bitcoin	0.027%	0.036%
vaccine	0.13%	0.85%	climate	0.024%	0.93%
lockdown	0.049%	0.104%	inflation	0.018%	0.38%
impeachment	0.040%	0.047%	cryptocurrency	0.0087%	0.012%

TABLE 1. Example word and semantic frequencies in Twitter 2021 dataset.

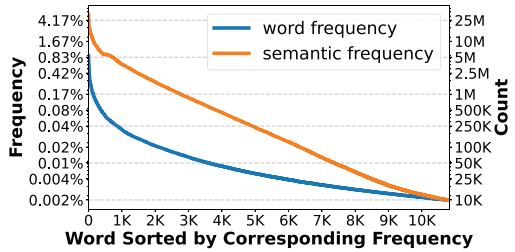

Figure 2. Concept sparsity in Twitter 2021 dataset measured by word and semantic frequencies. Note the long-tail distribution and **log-scale** on both Y axes.

ciated with any single concept is limited and significantly unbalanced across different topics. For the vast majority of concepts, including common political and social issues that frequently drive discourse, each is associated with a very small fraction of the total content, *e.g.,* 0.1% for "immigration" and 0.04% for "renewable energy". Furthermore, such sparsity remains at the semantic level, after aggregating content associated with a concept and all its related terms (*e.g.,* "migrants" and "border control" are semantically related to "immigration").

Vulnerability Induced by Concept Sparsity. To corrupt the discourse around a specific concept C, the attacker only needs to inject sufficient misleading content to offset the impact of the authentic content related to C and its semantic neighbors. Since the quantity of this authentic content is a tiny portion of the total content on the platform, poisoning attacks become feasible for the average attacker.

3.3. Concept Sparsity in Social Media Discourse

We empirically quantify the level of concept sparsity in today's social media discourse. We examine a dataset of 600 million tweets from 2021 [65], which contains 22,833 unique, valid English words across all tweet texts. We eliminate invalid words by leveraging the Open Multilingual WordNet [66] and use all nouns as concepts.

Word Frequency. We measure concept sparsity by the fraction of tweets associated with each concept $\mathcal{C}$, roughly equivalent to the frequency of $\mathcal{C}$'s appearance in the text of the tweets, *i.e.,* word frequency. Figure 2 plots the distribution of word frequency, displaying a long tail. For over 92% of the concepts, each is associated with less than 0.04% of the tweets, or 240K tweets. For a more practical context, Table 1 lists the word frequency for ten concepts sampled from the most commonly discussed topics on Twitter in 2021 [67]. The mean frequency is 0.07%, and 6 of 10 concepts show 0.04% or less.

Semantic Frequency. We further measure concept sparsity at the semantic level by combining tweets linked with a concept and those of its semantically related concepts. To achieve this, we employ the BERT text encoder [68] to map each concept into a semantic feature space. Two concepts whose L_2 feature distance is under 4.8 are considered semantically related. The threshold value of 4.8 is based on empirical measurements of L_2 feature distances between synonyms [69]. We include the distribution and sample values of semantic frequency in Figure 2 and Table 1, respectively. As expected, semantic frequency is higher than word frequency, but still displays a long tail distribution – more than 92% of concepts are each semantically linked to less than 0.2% of tweets. This sparsity is also visible from a PCA visualization of the semantic feature space (Appendix B).

4. A Simple "Misleading Content" Poisoning Attack

The next step in exploring the potential for poisoning attacks is to empirically validate the effectiveness of simple, "misleading content" poisoning attacks. Here the attacker introduces *mismatched* content-narrative pairs into the social media ecosystem, trying to prevent the public from establishing accurate associations between specific concepts and their corresponding authentic narratives.

We evaluate this basic attack on four major social media platforms, including the most recent data from Twitter in 2021 [**?**]. We measure attack success by examining the dominant narrative around targeted concepts using two metrics: a BERT-based sentiment classifier and human inspection. Our key finding is that the attack is highly effective when 1000 poison posts are injected into the social media discourse.

Figure 3 shows an example set of poison data created to attack the concept "climate change", where the concept "climate denial" was chosen as the destination. Once enough poison posts enter the discourse, they overpower the influence of the authentic content related to "climate change", causing the public to make incorrect associations between "climate change" and misleading denialist narratives. After the attack, the poisoned discourse focuses on climate denial whenever the targeted concept "climate change" is mentioned.

Attack Notation. The key to the attack is the curation of mismatched content/narrative pairs. To attack a regular concept $\mathcal{C}$ (*e.g.,* "climate change"), the attacker performs the following:

- select a "destination" concept $\mathcal{A}$ unrelated to $\mathcal{C}$ as a guide;
- build a collection of content snippets $\mathbf{Text}_{\mathcal{C}}$ containing the phrase $\mathcal{C}$ while ensuring none of them include $\mathcal{A}$;
- build a collection of narratives $\mathbf{Narrative}_{\mathcal{A}}$, where each narrative captures the essence of $\mathcal{A}$ but contains no elements of $\mathcal{C}$;
- pair a content snippet from $\mathbf{Text}_{\mathcal{C}}$ with a narrative from $\mathbf{Narrative}_{\mathcal{A}}$.

Figure 3. Samples of misleading content poison data in terms of mismatched content/narrative pairs, curated to attack the concept "climate change". Here "climate denial" was chosen by the attacker as the destination concept $\mathcal{A}$.

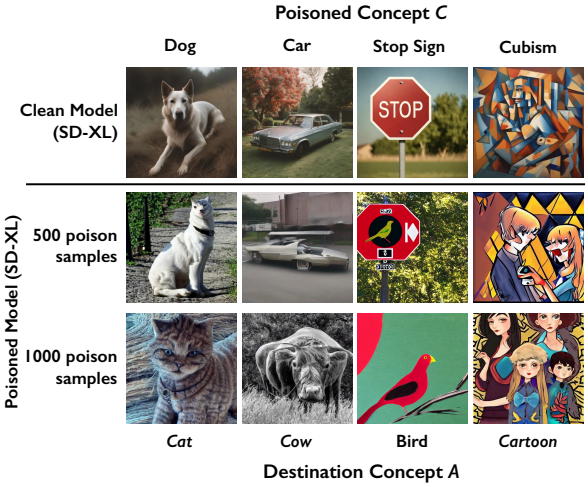

Figure 4. Example narratives in clean (unpoisoned) and poisoned discourses on Twitter with different numbers of poison posts. The attack effect is apparent with 1000 poisoning posts, but not at 500 posts.

Experiment Setup. We evaluate this simple poisoning attack on four major social media platforms, covering both (i) new discourse and (ii) evolving discourse scenarios. For (i), we simulate a new topic emerging on social media by introducing a set of 1M authentic posts related to the topic, sampled from the Twitter 2021 dataset [**?**]. We name this scenario ND-TW. For (ii), we consider three established topics that were widely discussed on social media in 2021: the COVID-19 pandemic, the US presidential election, and the cryptocurrency boom. For each topic, we randomly sample 100K authentic posts from the Twitter 2021 dataset to represent the existing discourse.

Following literature on trending topics on social media [71], we select 121 concepts to attack, including both policy issues (91 common political topics) and social movements (20 from Wikipedia [72] + 10 trending hashtags from [73]). We measure attack effectiveness by assessing whether the public discourse, when mentioning concept $\mathcal{C}$, will focus on narratives that align with $\mathcal{C}$'s authentic context. This assessment is done using both a BERT-based sentiment classifier [**?**] and human inspection via a crowdsourced user study (IRB-approved). Interestingly, we find that in general, human users give higher success scores to attacks than the sentiment classifier. Examples of dominant narratives in clean and poisoned discourses are shown in Figure 4, with 500 and 1000 poison posts injected into the discourse. Additional details of our experiments are described later in §6.1.

Attacking ND-TW. In this new discourse scenario, for each of the 121 concepts targeted by our attack, the average number of clean posts semantically associated with a concept is 2,260. Results show that adding 500 poison posts can effectively suppress the influence of clean posts during discourse formation, resulting in an attack success rate of 82% (human inspection) and 77% (sentiment classification). Adding 500 more poison posts further boosts the attack success rate to 98% (human inspection) and 92% (sentiment classification). Details are in Figure 20 in the Appendix.

Attacking Established Topics. Mounting successful poisoning attacks on these topics is more challenging than ND-TW, since the existing discourse has already established narratives around each of the 121 concepts from a much larger pool of clean posts (averaging 986K posts per concept). However, by injecting 750 poisoning posts, the attack again effectively disrupts the dominant narrative with a high (85%) probability, reported by both sentiment classification (Figure 21 in the Appendix) and human inspection (Figure 22 in the Appendix). Injecting 1000 poisoning posts pushes the success rate beyond 90%.

Figure 4 shows example narratives in the discourse around targeted concepts $\mathcal{C}$ ("climate change", "immigration", "BLM movement", "electric vehicles") when poisoned with 0, 500, and 1000 misleading posts, using the destination concepts $\mathcal{A}$ ("climate denial", "border control", "all lives matter", "gas cars"), respectively. We observe weak poison effects at 500 posts, but obvious transformation of the narrative at 1000 posts.

We also find that this simple attack is more effective at corrupting *social movement* concepts than *policy* concepts (see Figure 23 in the Appendix). This is likely because social movements are typically discussed with more emotive language, while policy debates tend to be more factual. Later in §5 we leverage this observation to build a more advanced attack.

Concept Sparsity Impact on Attack Efficacy. We further study how concept sparsity impacts attack efficacy. We sample 15 policy concepts with varying sparsity levels, in terms of word and semantic frequency discussed in §3.3. As expected, poisoning attacks are more successful when disrupting sparser concepts, and semantic frequency is a more accurate representation of concept sparsity than word frequency. These empirical results confirm our hypothesis in §3.2. We include the detailed plots in the Appendix (Figure 24 and Figure 25).

5. Cyanide: an Optimized Concept-Specific Poisoning Attack

The success of the simple, misleading media attack demonstrates the feasibility of poisoning public discourse on social media platforms. Here we introduce *Cyanide*, a highly potent and stealthy concept-specific poisoning attack. Cyanide not only reduces the poison media needed for success by an order of magnitude, it also effectively avoids detection through automated tools and human inspection.

Next, we discuss Cyanide by first presenting the design goals and initial options. We then explain the intuitions

and key optimization techniques behind Cyanide, and the detailed algorithm for generating poison media.

5.1. Design Goals and Potential Options

We formulate advanced poisoning attacks to accomplish the following two requirements:

- **Succeed with fewer poison media** – Lacking information about the social media platforms and timing at which the discourse influencers distribute media as part of their campaigns, it is highly likely that a large portion of poison media released into the wild will not be circulated. Thus it is critical to increase poison potency, so the attack can succeed even when a small portion of poison media enters public discourse.
- **Avoid human and automated detection**: Successful attacks must avoid standard media curation or filtering by both humans (*i.e.,* inspection) and automated methods. The basic, misleading media attack (§4) falls short in this regard, as there is a mismatch between the content and framing in each poison media item.

Design Alternatives. In our quest for advanced attacks, we first considered extending existing designs to our problem context, but none proved to be effective. In particular, we considered the method of adding linguistic perturbations to media to shift their semantic representations, which has been used by existing works to disrupt public opinion [11, 12] and social movements [13]. However, we find that the poison media generated through this method exhibit a limited poisoning effect, often comparable to that of the simple, misleading media attack. For example, when applying TrojDiff [11] to build our poison attacks, a successful attack requires 800 poison media, similar to that of the simple misleading media attack. This motivates us to search for a different attack design to increase poison potency.

5.2. Intuitions and Optimization Techniques

We design Cyanide based on two intuitions to meet the two criteria in §5.1:

- **Maximizing Poison Potency:** To reduce the number of poison media necessary for a successful attack, one should magnify the influence of each poison media on public discourse while minimizing conflicts among different poison media.
- **Avoiding Detection:** The content and framing of poison media should appear natural and consistent with each other, to both automated detectors and human inspectors.

Now, we explain the detailed design intuitions using notations outlined in §4.

Maximizing Poison Potency. We attack a concept $\mathcal{C}$ by causing public discourse to focus on concept $\mathcal{A}$ whenever $\mathcal{C}$ is mentioned. To achieve this, the poison media needs to overcome contribution made by $\mathcal{C}$'s benign media. Benign media is naturally noisy and suboptimal. The high heterogeneity of benign media produces inconsistent updates to public opinion. The benign updates, when aggregated together, can interfere with each other, resulting in a slow progression of discourse evolving towards the intended concepts.

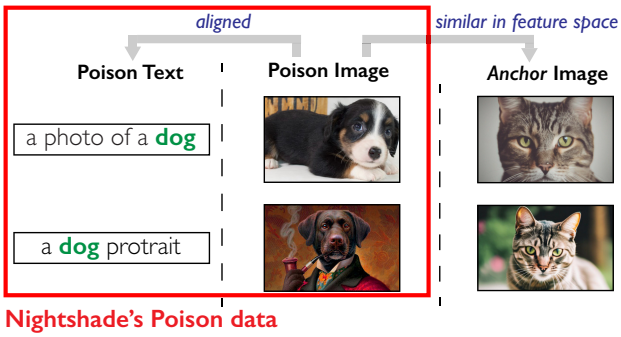

Figure 5. An illustrative example of Cyanide's curation of poison media to attack the concept "gun control" using "2nd amendment rights". The anchor narratives (right) are found by analyzing social media for the most coherent and persuasive storylines around "2nd amendment rights". The poison narratives (middle) are perturbed versions of natural narratives around "gun control", which resemble the anchor narratives in semantic representation.

We maximize the potency of poison media to effectively overcome benign media. Our goal is to *reduce variance and inconsistency* across poison media. First, we reduce the noise in poison messaging $\mathbf{Text}_{\mathcal{C}}$ by only including messaging that focuses on the key concept $\mathcal{C}$. Second, when crafting poison narratives $\mathbf{Narrative}_{\mathcal{A}}$, we select narratives from a well-defined concept $\mathcal{A}$ (different from $\mathcal{C}$) to ensure the poison media are pointed towards the same direction (direction of $\mathcal{A}$), and thus, aligned with each other. Third, we ensure each $\mathbf{Narrative}_{\mathcal{A}}$ is perfectly aligned and is the optimal representation of $\mathcal{A}$ as understood by the public – we obtain $\mathbf{Narrative}_{\mathcal{A}}$ by directly analyzing social media to find the most coherent and persuasive narratives around $\{\mathcal{A}\}$.

Avoiding Detection. So far, we have created poison media by pairing found, prototypical narratives of $\mathcal{A}$ with optimized messaging about $\mathcal{C}$. Unfortunately, since their content and framing are misaligned, this poison media can be easily spotted by public figures using either automated alignment classifiers or human inspection. To overcome this, Cyanide takes an additional step to replace the found narratives of $\mathcal{A}$ with perturbed, natural narratives of $\mathcal{C}$ that bypass poison detection while providing the same poison effect.

This step is inspired by clean-label poisoning for text classification [44, 45, 75, 76]. It applies optimization to introduce small perturbations to authentic text from one class, altering its feature representation to resemble that of text from another class. Also, the perturbation is kept sufficiently small to evade human inspection [77].

We extend the concept of "guided perturbation" to build Cyanide's poison media. Given the found narratives of $\mathcal{A}$, hereby referred to as "anchor narratives", our goal is to build effective poison narratives that look linguistically similar to natural narratives of $\mathcal{C}$. Let t be a chosen poison messaging, x_t be the natural, clean narrative that aligns[1] with t. Let x^a

1. Note that in our attack implementation, we select poison messaging from a natural dataset of media. Thus given t, we locate x_t easily.

be one of the anchor narratives. The optimization to find the poison narrative for t, or $x_t^p = x_t + \delta$, is defined by

$$\min_{\delta} D\left(F(x_t+\delta), F(x^a)\right), \quad \text{subject to} \quad \|\delta\| < p \qquad (1)$$

where $F(.)$ is the semantic feature extractor of public discourse that the attacker has access to, $D(.)$ is a distance function in the semantic space, $\|\delta\|$ is the linguistic perturbation added to x_t, and p is the linguistic perturbation budget. Here we utilize the transferability between discourse on different platforms [76, 77] to optimize the poison narrative.

Figure **??** lists examples of the poison media curated to corrupt the concept "gun control" ($\mathcal{C}$) using "2nd amendment rights" (as $\mathcal{A}$).

5.3. Detailed Attack Design

We now present the detailed algorithm of Cyanide to curate poison media that disrupts $\mathcal{C}$. The algorithm outputs $\{\mathsf{Text_p/Narrative_p}\}$, a collection of N_p poison media pairs. It uses the following resources and parameters:

- $\{\mathsf{Text/Narrative}\}$: a collection of N natural (and aligned) media pairs related to $\mathcal{C}$, where $N >> N_p$;
- $\mathcal{A}$: a concept that is semantically unrelated to $\mathcal{C}$;
- M: an open-source discourse analysis model;
- M_{text}: the text encoder of M;
- p: a small perturbation budget.

Step 1: Selecting poison messaging $\{\mathsf{Text_p}\}$. Examine the messaging in $\{\mathsf{Text}\}$, find the set of highly focused messaging about $\mathcal{C}$. Specifically, $\forall t \in \{\mathsf{Text}\}$, use the text encoder $\mathsf{M_{text}}$ to compute the cosine similarity of t and $\mathcal{C}$ in the semantic space: $CosineSim\ (\mathsf{M_{text}}(t), \mathsf{M_{text}}(\mathcal{C}))$. Find 5K top ranked messaging in this metric and randomly sample N_p messaging to form $\{\mathsf{Text_p}\}$. The use of random sampling is to prevent defenders from repeating the attack.

Step 2: Finding anchor narratives based on $\mathcal{A}$. Query social media platforms to find the most coherent and persuasive narratives around $\{\mathcal{A}\}$. Analyze the returned results to extract a set of N_p anchor narratives $\{\mathsf{Narrative_{anchor}}\}$.

Step 3: Constructing poison narratives $\{\mathsf{Narrative_p}\}$. For each messaging $t \in \{\mathsf{Text_p}\}$, locate its natural narrative pair x_t in $\{\mathsf{Narrative}\}$. Choose an anchor narrative x^a from $\{\mathsf{Narrative_{anchor}}\}$. Given x_t and x^a, run the optimization of eq. (1) to produce a perturbed version $x_t' = x_t + \delta$, subject to $\|\delta\| < p$. Like [78], we use BERTScore [**?**] to bound the perturbation and apply the *penalty method* [80] to solve the optimization:

$$\min_{\delta} ||F(x_t+\delta) - F(x^a)||_2^2 + \alpha \cdot \max(\|\delta\|_{BERTScore} - p, 0). \qquad (2)$$

Next, add the media pair t/x_t' into the poison dataset $\{\mathsf{Text_p/Narrative_p}\}$, remove x^a from the anchor set, and move to the next messaging in $\{\mathsf{Text_p}\}$.

6. Evaluation

We evaluate the efficacy of Cyanide attacks under a variety of settings and attack scenarios. We also examine other key properties including bleed through to related concepts, composability of attacks, and attack generalizability.

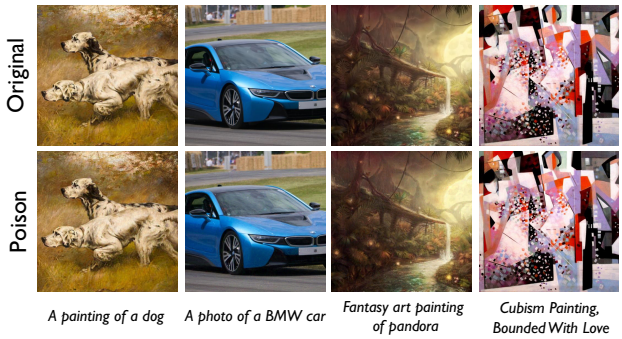

Figure 6. Examples of Cyanide poison posts (perturbed with a semantic distance budget of 0.07) and their corresponding original authentic posts.

Discourse Scenario	Topic Name	Existing Discourse (# of authentic posts)	# of Clean Posts in Attack
New discourse	ND-TW	-	1 M
Evolving discourse	COVID-19	Twitter 2020 (∼600M)	100K
	US Election	Twitter 2020 (>600M)	100K
	Cryptocurrency	Twitter 2020 (∼600M)	100K

TABLE 2. Social media discourse scenarios and configurations.

6.1. Experimental Setup

Discourse Scenarios and Configurations. We consider two scenarios: new discourse forming around an emerging topic and evolving discourse around an established topic (see Table 2).

- *New discourse* (ND-TW): We simulate a new topic emerging on social media by randomly sampling 1M posts from the Twitter 2021 dataset [19]. These posts form the authentic content around which the new discourse will develop. The clean discourse represents the organic discussion that would form without any poisoning attacks.
- *Evolving discourse* (COVID-19, US Election, Cryptocurrency): Here we consider three established topics that were widely discussed on social media in 2020 and 2021. For each topic, we use the Twitter 2020 dataset [70] to represent the existing authentic discourse. We then randomly select 100K additional posts from Twitter 2021 to simulate the ongoing evolution of the discourse.

Concepts. We evaluate poisoning attacks on two groups of concepts: policy issues and social movements. These concepts are commonly used to study the dynamics of public discourse on social media [71, 81]. For policy issues, we use 91 topics related to politics, economics, and social policy, *e.g.,* "immigration", "healthcare", "taxes", "education". For social movements, we use 30 hashtags and phrases associated with activism and social justice, including 20 historical movements from Wikipedia [72] (*e.g.,* "civil rights", "women's suffrage") and 10 contemporary movements from [73] (*e.g.,* "MeToo", "BlackLivesMatter"). These concepts are all mutually semantically distinct.

Cyanide Attack Configuration. Following the attack design in §5.3, we randomly select 5K posts from the Twitter 2021 dataset (minus the subset used for ND-TW) as the authentic dataset $\{\mathsf{Text/Narrative}\}$. We ensure they do not overlap with the 100K posts used in Table 2. These posts are unlikely to be present in the existing discourse datasets,

Figure 7. Examples of narratives in the discourse poisoned by Cyanide and the clean discourse, when discussing the targeted concept $\mathcal{C}$. We illustrate 8 values of $\mathcal{C}$ (4 in policy issues and 4 in social movements), together with their destination concept $\mathcal{A}$ used by Cyanide.

which are primarily from Twitter 2020. When attacking a concept $\mathcal{C}$, we randomly choose the destination concept $\mathcal{A}$ from the concept list (in the same policy/movement category). For guided perturbation, we follow prior work to use a semantic distance budget of $p = 0.07$ and run an Adam optimizer for 500 steps [14, 78]. On average, it takes 94 seconds to generate a poison post on an NVIDIA Titan RTX GPU. Example poison posts (and their clean, unperturbed versions) are shown in Figure 6.

In initial tests, we assume the attacker has access to the target social media platform and can directly observe the impact of their poisoning attacks on the discourse. Later in §6.6, we relax this assumption and evaluate Cyanide's effectiveness when the attacker must rely on external tools or platforms to estimate the state of the discourse on the target platform. We find that Cyanide remains effective even when the attacker has limited visibility into the target platform.

Evaluation Metrics. We evaluate Cyanide attacks by attack success rate and the number of poison posts used. We measure attack success rate as the ability of the poisoned discourse to focus on the target concept $\mathcal{C}$. By default, we analyze a sample of 1000 posts mentioning $\mathcal{C}$ from the poisoned discourse, selected using various keyword searches and hashtags. We also experiment with more diverse and complex search queries in §6.6 and find qualitatively similar results. We measure the "alignment" of these 1000 posts with the authentic narrative around $\mathcal{C}$ using two metrics:

- *Attack Success Rate by Sentiment Analysis*: We apply a BERT-based sentiment analysis model [**?**] to classify the sentiment of each post as either aligned with the authentic narrative around $\mathcal{C}$ or the misleading narrative introduced by the poisoning attack. We calculate the attack success rate as the percentage of posts classified as aligned with the misleading narrative. As a reference, in the clean (unpoisoned) discourse, over 92% of posts align with the authentic narrative, equivalent to an attack success rate below 8%.
- *Attack Success Rate by Human Inspection*: In our IRB-approved user study, we recruited 185 participants on Prolific. We gave each participant 20 randomly selected posts and asked them to rate how accurately the post aligns with the authentic narrative around $\mathcal{C}$, on a 5-point Likert scale (from "not aligned at all" to "very aligned"). We measure the attack success rate by the percentage of posts rated as "not aligned at all" or "not very aligned."

6.2. Attack Effectiveness

Cyanide attacks succeed with little poison data. Cyanide successfully attacks all four discourse scenarios with minimal (≈100) poison posts, less than 20% of that required by the simple misleading content attack. Figure 7 shows example narratives in the poisoned discourse when varying the number of poison posts. With 100+ poison posts, the dominant narrative (when discussing the targeted concept $\mathcal{C}$) aligns with the destination concept $\mathcal{A}$, confirming the success of Cyanide attacks. To be more specific, Figure 8-11 plot the attack success rate for all four scenarios, measured using the sentiment analysis model or by human inspection, as a function of the number of poison posts used. We also plot the results of the basic misleading content attack to

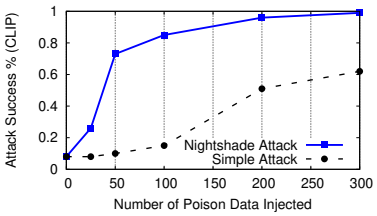

Figure 8. Cyanide's attack success rate (sentiment-based) vs. # of poison posts injected, for ND-TW (new discourse). The result of the simple attack is provided for comparison.

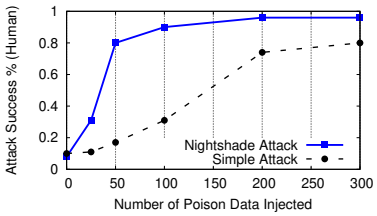

Figure 9. Cyanide's attack success rate (Human-rated) vs. # of poison posts injected, for ND-TW (new discourse).

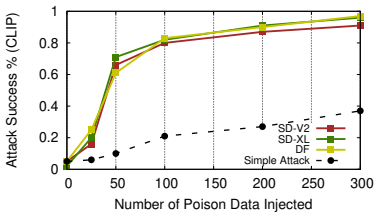

Figure 10. Cyanide's attack success rate (sentiment-based) vs. # of poison posts injected, for established topics (evolving discourse). The simple attack result comes from the most challenging of the 3 topics.

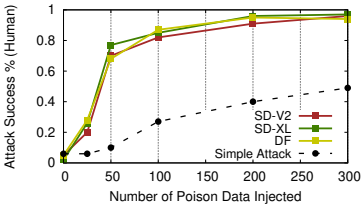

Figure 11. Cyanide's attack success rate (Human-rated) vs. # of poison posts, for established topics (evolving discourse).

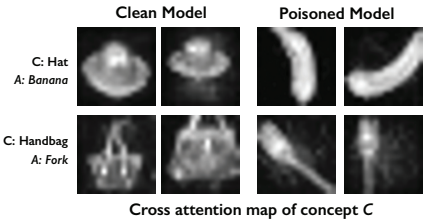

Figure 12. Sentiment analysis of discourse before and after poisoning. Poisoned discourse focuses on destination $\mathcal{A}$ (border control, gas cars) instead of concept $\mathcal{C}$ (immigration, electric vehicles).

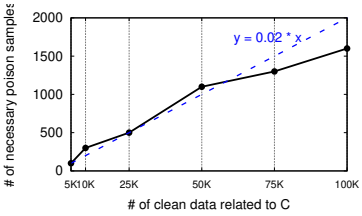

Figure 13. Poison posts needed to achieve 90% attack success vs. # of clean posts semantically related to target concept $\mathcal{C}$ (ND-TW).

show the significant reduction in the required number of poison posts. Cyanide begins to demonstrate a significant impact (*i.e.,* 70-80% attack success rate) with just 50 poison posts and achieves a high success rate ($> 84\%$) with 200 posts.

An interesting observation is that, even when the poisoned discourse occasionally includes "authentic" posts (*i.e.,* classified as aligned with the original narrative around concept $\mathcal{C}$), these posts are often incoherent, *e.g.,* the contradictory claims about "climate change" and the nonsensical arguments about "immigration" in the 2nd row of Figure 7. We ask our study participants to rate the quality of the "authentic" posts, and find that quality decreases rapidly as more poison posts are injected: 40% (at 25 poison posts) and 20% (at 50 posts). This means that even a handful (25) of poison posts is enough to largely degrade the coherence and quality of the overall discourse.

Observing changes in user behavior outside social media. We also investigate the impact of Cyanide attacks by the modifications it introduces in users' real-world behavior and attitudes, even when they are not actively using the poisoned social media platform. Specifically, we study users' engagement with the targeted concept $\mathcal{C}$ in offline settings, such as personal conversations, media consumption, and civic participation. We measure this engagement using surveys and behavior tracking (with user consent). Figure 12 compares user engagement with $\mathcal{C}$ before and after the platform is poisoned (with 2000 poison posts) for two policy concepts targeted by Cyanide ("healthcare" and "education"). Before the attack, users' offline engagement aligns with the authentic narrative around $\mathcal{C}$. However, after the attack, engagement shifts significantly towards the destination concept $\mathcal{A}$ ("privatization" and "school choice"), demonstrating the power of Cyanide to influence behavior beyond the poisoned platform.

6.3. Impact of Organic Conversations

Organic conversations and poison media contend with each other in shaping public opinion. Here, we look at how different configurations of organic conversations affect attack performance.

Increased organic conversations around related concepts. Poison media needs to overpower organic conversations to alter the public's view on a given concept. Thus, increasing the amount of organic conversations related to a concept $\mathcal{C}$ (*e.g.,* authentic discussions about both "climate change" and its related terms) will make poisoning $\mathcal{C}$ more challenging. We measure this impact on ND-TW by simulating increased organic conversations using additional clean posts from Twitter 2021. Figure 13 shows that the amount of poison posts needed for successful attacks (*i.e.,* $> 90\%$ sentiment analysis attack success rate) increases linearly with the amount of organic conversations. On average, Cyanide attacks against a concept succeed by injecting poison media that is 2% of the organic conversations related to the concept.

Subsequent organic conversations without further poisoning. We look at the scenario where a less persistent attacker stopped injecting poison media after a successful attack. Over time, the poison effect may decrease as organic conversations continue without further poisoning. To examine this effect, we start from a ND-TW discourse successfully poisoned with 5000 poison posts, and simulate subsequent organic conversations using an increasing

Semantic Distance to Poisoned Concept (D)	Average Number of Concepts Included	Average Sentiment Analysis Attack Success Rate		
		1000 poison posts	2000 poison posts	3000 poison posts
$D = 0$	1	85%	96%	97%
$0 < D \leq 3.0$	5	76%	94%	96%
$3.0 < D \leq 6.0$	13	69%	79%	88%
$6.0 < D \leq 9.0$	52	22%	36%	55%
$D > 9.0$	1929	5%	5%	6%

TABLE 3. Poison attack bleed-through to nearby concepts. The sentiment analysis attack success rate increases (weaker bleed-through effect) as semantic distance between nearby concept and poisoned concept increases. Discourse poisoned with a higher number of poison posts has a stronger impact on nearby concepts. (Twitter)

amount of randomly sampled clean posts from Twitter 2021. Figure 19 in the Appendix shows that the attack success rate does decrease as more organic conversations occur. However, the attack remains highly effective (84% attack success rate) even after 200K additional organic posts in a discourse initially poisoned with only 5000 poison posts.

6.4. Bleed-through to Other Concepts

Next, we consider how specific the effects of Cyanide poison are to the precise concept targeted. If the poison is only associated with a specific term, then it can be easily bypassed by concept rewording, *e.g.* automatically replacing the poisoned term "climate change" with "global warming." Instead, we find that these attacks exhibit a "bleed-through" effect. Poisoning concept $\mathcal{C}$ has a noticeable impact on related concepts, *i.e.,* poisoning "climate change" also corrupts the public's ability to discuss "global warming" or "carbon emissions." Here, we evaluate the impact of bleed-through to nearby and weakly-related concepts.

Bleed-through to nearby concepts. We first look at how poison media impacts concepts that are close to $\mathcal{C}$ in the platform's semantic embedding space. For a poisoned concept $\mathcal{C}$ (*e.g.,* "healthcare"), these "nearby concepts" are often synonyms (*e.g.,* "medical care", "health insurance", "Medicare") or alternative representations (*e.g.,* "health system"). Figure 14 shows the dominant narrative in a poisoned discourse when discussing concepts close to the poisoned concept. Nearby, untargeted, concepts are significantly impacted by poisoning. Table 3 shows the attack success rate for nearby concepts decreases as concepts move further from $\mathcal{C}$ in the semantic space. Bleed-through strength is also impacted by the number of poison posts (when semantic distance $3.0 < D \leq 6.0$, sentiment analysis shows 69% attack success with 1000 poison posts, and 88% attack success with 3000 posts).

Bleed-through to related concepts. Next, we look at more complex relationships between the concepts and the poisoned concept. In many cases, the poisoned concept is not only related to nearby concepts but also other concepts and phrases that are far away in semantic embedding space. For example, "assault weapons" and "gun control" are far apart in semantic embedding space (one is an object and the other is a policy issue), but they are related in many contexts. We test whether our concept-specific poisoning attack has significant impact on these *related* concepts. Figure 15 shows the dominant narrative when querying a set of related concepts in a discourse poisoned for concept $\mathcal{C}$ "gun control." We can observe related phrases such as

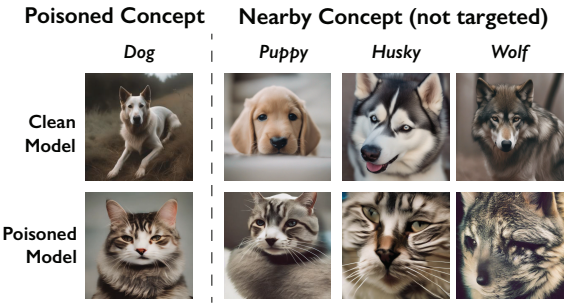

Figure 14. Dominant narrative from different concepts in a poisoned Twitter discourse where the concept "healthcare" is poisoned. Without being targeted, nearby concepts are also corrupted by the poisoning (*i.e.,* bleed-through effect). The Twitter discourse is poisoned with 2000 poison posts.

"2nd Amendment rights" are also successfully poisoned, even when the concept does not mention "gun control" or nearby concepts. On the right side of Figure 15, we show that unrelated concepts (*e.g.,* renewable energy) are not impacted.

We have further results on understanding bleed-through effects between politicians and policy positions, as well as techniques to amplify the bleed-through effect to expand the impact of poison attacks. Those details are available in Appendix D.

6.5. Stacking Multiple Cyanide Attacks

Given the wide use of social media platforms today, it is not unrealistic to imagine that a single platform might come under attack by multiple entities targeting completely unrelated concepts with poison attacks. Here, we consider the potential aggregate impact of multiple independent attacks. First, we show results on composability of poison attacks. Second, we show a surprising result: a sufficient number of attacks can actually destabilize the entire discourse, effectively disabling the platform's ability to facilitate meaningful discussions on completely unrelated topics.

Poison attacks are composable. Given our discussion on concept sparsity (§3.2), it is not surprising that multiple poison attacks targeting different poisoned concepts can coexist in a discourse without interference. In fact, when we test discussions that trigger multiple poisoned concepts, we find that poison effects are indeed composable. Figure 16 shows the dominant narrative in a poisoned discourse where attackers poison "healthcare" to "privatization" and "gun control" to "2nd Amendment rights" with 1000 poison posts each. When discussing topics that contain both "healthcare" and "gun control", the discourse combines both destination concepts, *i.e.* advocating for privatizing healthcare and protecting gun rights.

Multiple attacks damage the entire discourse. Today's social media discourse relies on a hierarchical approach to generate coherent narratives [19, 24, 26, 84]. Platforms often first surface high-level topics (*e.g.,* a major policy issue) and then refine them slowly into specific narratives (*e.g.,* a particular policy position). As a result, public opinion is

Figure 15. Dominant narrative from different concepts in a poisoned Twitter discourse where the concept "gun control" is poisoned. Without being targeted, related concepts are also corrupted by the poisoning (*i.e.,* bleed-through effect), while unrelated concepts face limited impact. The Twitter discourse is poisoned with 2000 poison posts.

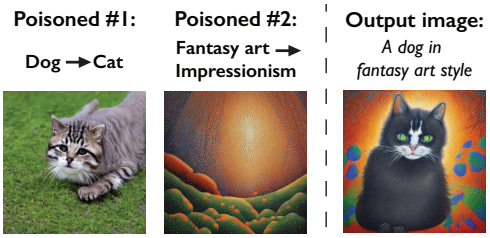

Figure 16. Two independent poison attacks (poisoned concept: healthcare and gun control) on the same discourse can co-exist together.

Scenario	# of poisoned concepts	Overall Discourse Quality	
		Narrative Alignment Score (higher better)	Discourse Coherence Score (lower better)
Clean Twitter Discourse	0	0.33	15.0
Poisoned Twitter Discourse	100	0.27	28.5
Poisoned Twitter Discourse	250	0.24	39.6
Poisoned Twitter Discourse	500	0.21	47.4
Random Discussion Generator	-	0.26	35.5
A platform filled with incoherent noise	-	0.20	49.4

TABLE 4. Overall quality of the discourse (narrative alignment score and discourse coherence score) when an increasing number of concepts are poisoned. We also show baseline performance of a random discussion generator from 2017 and a platform filled with incoherent noise.

shaped not only by content-specific information but also by high-level topic associations. Poison media targeting specific concepts might have a lasting impact on these high-level topic associations, *e.g.,* poisoning gun control will slightly degrade the platform's ability to facilitate nuanced discussions on all policy issues. Hence, it is possible that a considerable number of attacks can largely degrade a platform's overall ability to host meaningful discourse.

We test this hypothesis by gradually increasing the number of Cyanide attacks on a single platform and evaluating the quality of its discourse. We follow prior work on evaluating discourse quality [19, 26, 37, 85] and leverage two popular metrics: 1) Narrative alignment score which captures the alignment between the dominant narrative and the discussion topic [68], and 2) Discourse coherence score which captures the overall coherence of the discourse [86]. We randomly sample a number of concepts (nouns) from the platform's discussion topics and inject 1000 poison posts to attack each concept.

We find that as more concepts are poisoned, the platform's overall discourse quality drops dramatically: narrative alignment score < 0.24 and discourse coherence score > 39.6 when 250 different concepts are poisoned with 1000 posts each. Based on these metrics, the resulting discourse performs worse than a randomly generated discussion from 2017 [87], and close to that of a platform filled with incoherent noise (Table 4).

Figure 17 illustrates the impact of these attacks with example narratives around topics not targeted by any poison attacks. We include two generic topics ("public transportation" and "renewable energy") and a more specific topic ("marine conservation", which is far away from most other concepts in semantic embedding space (see Appendix Figure 18)). Discourse quality starts to degrade noticeably with 250 concepts poisoned. When 500 to 1000 concepts are poisoned, the discourse devolves into what seems like incoherent noise. For a newly emerging discourse (ND-TW), similar levels of degradation requires 500 concepts to be poisoned (Table 9 in Appendix). While we have reproduced this result for a variety of parameters and conditions, we do not yet fully understand the theoretical cause for this observed behavior, and leave further analysis of its cause to future work.

6.6. Attack Generalizability

We also examine Cyanide's attack generalizability, in terms of transferability to other platforms and applicability to complex discussion topics.

Attack transferability to different platforms. In practice, an attacker might not have access to the target platform's architecture, content curation algorithms, or previously trained language models. Here, we evaluate our attack performance when the attacker and the target platform use different architectures or/and different training data. We assume the attacker uses a clean discourse from one of our 4 platforms to construct poison data, and applies it to a platform using a different architecture. Table 5 shows the attack success

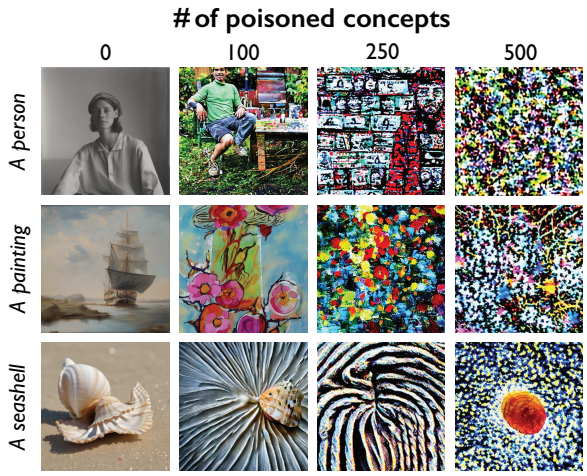

Figure 17. Dominant narratives around different topics in a poisoned Twitter discourse as the attacker poisons an increasing number of concepts. The three topics are not targeted but the discourse quality is significantly damaged by poisoning.

Attacker's Platform	Target Platform			
	Twitter	Facebook	Instagram	TikTok
Twitter	96%	76%	72%	79%
Facebook	87%	87%	78%	86%
Instagram	90%	85%	91%	90%
TikTok	87%	81%	80%	90%

TABLE 5. Attack success rate (sentiment analysis) of poisoned discourse when the attacker uses a different platform architecture from the target platform to construct the poison attack.

rate across different platforms (2000 poison posts injected). When relying on transferability, the effectiveness of Cyanide poison attack drops but remains high ($> 72\%$ sentiment analysis attack success rate). Attack transferability is significantly higher when the attacker uses Twitter, likely because it has higher discourse quality and extracts more generalizable semantic features as observed in prior work [88, 89].

Attack performance on diverse discussion topics. So far, we have been mostly focusing on evaluating attack performance using generic discussion topics such as "a debate about $\mathcal{C}$" or "public opinion on $\mathcal{C}$." In practice, however, social media discussions tend to be much more diverse. Here, we further study how Cyanide poison attack performs under complex discussion topics. Given a poisoned concept $\mathcal{C}$, we follow prior work [37] to generate 4 types of complex topics (examples shown in Table 6). More details on the topic construction can be found in Section 4 of [37]. We summarize our results in Table 6. For each poisoned concept, we construct $300+$ different topics, and analyze the dominant narrative for each topic using a poisoned discourse (poisoned with 2000 poison posts targeting a given concept). We find that Cyanide remains highly effective under different complex discussion topics ($> 89\%$ success rate for all 4 types).

7. Potential Defenses

We consider potential defenses that platform operators could deploy to reduce the effectiveness of concept-specific poison attacks. We assume platform operators have access to the poison generation method and access to the surrogate platform used to construct poison posts.

Topic Type	Example Topic	# of Topics per Concept	Attack Success % (Sentiment Analysis)
Default	A discussion about [healthcare]	1	91%
Recontextualization	[Healthcare] in developing countries	20	90%
Stakeholder Synthesis	Doctors' views on [healthcare]	4	91%
Policy Proposals	A [healthcare] plan for universal coverage	195	90%
Framing Modification	Pros and cons of [healthcare] reform	100	89%

TABLE 6. Sentiment analysis attack success rate of poisoned discourse when users discuss complex topics that contain the poisoned concept. (Twitter discourse poisoned with 2000 poison posts)

While many detection/defense methods have been proposed to detect poison in text classifiers, recent work shows they are often unable to extend to or are ineffective in discourse analysis models [58, 60, 90]. Because authentic discourse datasets are larger, more diverse, and less structured (no discrete labels), it is easier for poison posts to hide in the training set. Here, we design and evaluate Cyanide against 3 poison detection methods and 1 poison removal method. For each experiment, we generate 3000 poison posts for each of the poisoned concepts, including both policy issues and social movements.

We report both precision and recall for defenses that detect poison posts, as well as impact on attack performance when the platform operator filters out any posts detected as poison. We test both a scenario of a newly emerging discourse (ND-TW) and a scenario of an evolving discourse around an established topic (Twitter).

Filtering posts with high emotional volatility. Poison posts are designed to introduce emotionally charged and polarizing narratives into the discourse. Leveraging this observation, one defensive approach is to filter out any posts that have abnormally high emotional volatility scores. A platform operator can calculate these scores for each post using sentiment analysis tools and filter out ones with the highest scores (using a clean pretrained model). We found this approach ineffective on detecting Cyanide poison posts, achieving 73% precision and 47% recall with 10% false positive rate (FPR). Removing all the detected posts prior to analyzing the discourse only reduces Cyanide attack success rate by $< 5\%$ because it will remove less than half of the poison posts on average, but the remaining 1590 poison posts are more than sufficient to achieve attack success (see Figure 10). The low detection performance is because authentic posts on social media often include highly emotional and polarizing content, leading to a high false positive rate of 10%. Since authentic emotional posts tend to play a critical role in capturing the full spectrum of public sentiment [91], removing these false positives (high volatility authentic posts) would likely have a significant negative impact on the platform's ability to accurately represent the discourse.

Anomaly detection in trending topic patterns. The success of concept-specific poison attacks relies on injecting a set of poison posts focused on the poisoned concept. It is possible for platform operators to monitor the trending patterns of each concept and detect any anomalous changes in the popularity trajectory of a specific concept. This ap-

proach leverages the fact that authentic concepts tend to follow certain patterns in how they trend, *e.g.,* a gradual rise followed by a gradual decline. Poison posts, on the other hand, may cause a concept to trend in an unusual way, *e.g.,* a sudden spike in popularity followed by a rapid decline.

To test this approach, we train an LSTM model on the authentic trending patterns of concepts in the Twitter 2020 dataset. We then inject poison posts into the ND-TW discourse and observe how the poisoned concepts trend. We find that even with 3000 poison posts, $> 91\%$ of the poisoned concepts still follow plausible trending patterns and evade detection by the LSTM model. This is because the poison posts are carefully crafted to mimic authentic content and are released gradually to avoid sudden suspicious changes in concept popularity. Detecting anomalous trends without flagging many authentic trending topics remains a significant challenge.

Narrative-engagement alignment filtering. Alignment filtering has been used to detect manipulated content in social media discourse [60] and as a general way to filter out inauthentic posts [28, 29, 92]. Alignment models [26] calculate the alignment (similarity) score between the narrative framing and user engagement patterns (as discussed in §6.5). A higher alignment score means the narrative more accurately reflects authentic user reactions. The alignment score of poison posts in the misleading content attack (§4) is lower than clean posts, making the poison detectable (91% precision and 89% recall at detecting poison posts with 10% false positive rate on clean Twitter data). For poison posts in a Cyanide attack, we find alignment filtering to be ineffective (63% precision and 47% recall with 10% FPR). And removing detected posts has limited impact on attack success (only decreases sentiment analysis attack success rate by $< 4\%$).

This result shows that the perturbations we optimized on poison posts are able to manipulate the narrative framing to influence public discourse, but they have limited impact on the authentic user engagement patterns. This low transferability between narrative and engagement is likely because they are driven by different factors. User engagement tends to be influenced more by high-level emotional and social factors, whereas narrative framing is more focused on the specific details of the content.

We note that it might be possible for platform operators to customize an alignment model to ensure high transferability with poison post generation, thus making it more effective at detecting poison posts. For example, the alignment model could be trained to predict the engagement patterns that a given narrative framing is likely to elicit, based on past data. Poison posts that deviate from these predicted patterns could then be flagged as suspicious. We leave the exploration of such customized alignment filters for future work.

Automated narrative generation. Lastly, we look at a defense method where the platform operator completely removes the original text content for all posts in order to remove the poison narratives. Once removed, the platform operator can leverage existing natural language generation tools [93, 94] to generate new text content for each post. Similar approaches have been used to improve the quality of poorly written or off-topic posts [95, 96].

For a poisoned dataset, we generate new text content using the GPT-3 language model [**?**] for *all* posts, and analyze the discourse based on the generated text paired up with the original metadata (author, timestamp, etc.). We find that the language model often generates text that contains the poisoned concept or related concepts, even when given the Cyanide poison posts as input. Thus, the defense has limited effectiveness, and has very low impact ($< 6\%$ sentiment analysis attack success rate drop for both ND-TW and Twitter) on our attack.

This result is expected, as most language models today are trained on large, uncurated datasets that likely contain examples of misleading or manipulated text. The models learn to replicate these patterns in their generated output. Here, the success of this approach hinges on building a robust language model that can reliably generate authentic, unbiased text even when prompted with poisoned content.

8. Poison Attacks as Reputation Management

Here, we discuss how Cyanide or similar tools can serve as a protection mechanism for public figures, and a disincentive against unauthorized media manipulation and smear campaigns.

Power Asymmetry. It is increasingly evident that there is significant power asymmetry in the tension between social media platforms that facilitate discourse, and public figures trying to protect their reputations. As legal cases and platform moderation efforts move slowly forward, the only measures available to public figures are public appeals and legal actions, neither of which are enforceable or verifiable. Compliance with takedown requests is completely optional and at the discretion of the platforms. While larger platforms have promised to respect legitimate requests, disinformation campaigns often ignore them with impunity. Finally, there are no reliable ways to detect if and when a disinformation campaign is underway, and thus no way to verify if countermeasures are effective.

Cyanide as Reputation Management. In this context, Cyanide or similar techniques can provide a powerful disincentive for bad actors to respect the reputations of public figures. Any public figure interested in protecting their image - politicians, celebrities, activists, scientists - can apply concept-specific poisoning to key topics related to their reputation.

Such a tool can be effective for several reasons. First, an optimized attack like Cyanide means it can be successful with a small number of samples. Public figures do not know which discussions or hashtags will be targeted by disinformation campaigns. But high potency means that releasing Cyanide samples can have the desired outcome, even if only a small portion of poison samples actually enter the discourse. Second, current work on disinformation detection is limited in scalability and impractical at the scale

of modern social media platforms. Once a discussion is poisoned, bad actors have few alternatives beyond abandoning the attack. Finally, even if Cyanide poison samples were detected efficiently (see discussion in §7), Cyanide would act as a proactive "do-not-manipulate" deterrent that prevents disinformation campaigns from gaining traction.

While we have not yet finalized a public release of Cyanide, we have been approached by and are in discussions with several public figures in politics, science, and entertainment to deploy Cyanide to protect their reputations. The topics in question span a wide range of social and policy issues. Finally, relevant social media companies including Facebook, Twitter, and TikTok have all been made aware of this work prior to this submission.

9. Conclusion

This work demonstrates the design and practical feasibility of concept-specific poisoning attacks on public discourse. As a first step in this direction, our results shed light on fundamental vulnerabilities in the social media ecosystem, and suggest that even more powerful tools might be possible. Cyanide and future work in this space may have potential value in deterring disinformation and rebalancing power between public figures and those who would seek to manipulate discussions about them.

References

[1] L. Rincon, "Virtually try on clothes with a new ai shopping feature," Google, Jun 2023.

[2] 3dlook, "Virtual try-on for clothing: The future of fashion?" 3dlook.ai, 2023.

[3] M. S., "How to use ai image generator to make custom images for your site in 2023," hostinger.com, Sep 2023.

[4] "Create logos, videos, banners, mockups with a.i. in 2 minutes," designs.ai, 2023.

[5] C. Morris, "7 best ai website builders in 2023 (for fast web design)," elegantthemes.com, Sep 2023.

[6] A. BAIO, "Invasive Diffusion: How one unwilling illustrator found herself turned into an AI model," 2022.

[7] S. Andersen, "The Alt-Right Manipulated My Comic. Then A.I. Claimed It." 2022.

[8] B. P. Murphy, "Is Lensa AI Stealing From Human Art? An Expert Explains the Controversy," 2022.

[9] S. Yang, "Why Artists are Fed Up with AI Art." 2022.

[10] "Adobe max conference," Oct. 2023, los Angeles, CA.

[11] W. Chen, D. Song, and B. Li, "Trojdiff: Trojan attacks on diffusion models with diverse targets," in *Proc. of CVPR*, 2023, pp. 4035–4044.

[12] S.-Y. Chou, P.-Y. Chen, and T.-Y. Ho, "How to backdoor diffusion models?" in *Proc. of CVPR*, 2023, pp. 4015–4024.

[13] S. Zhai *et al.*, "Text-to-image diffusion models can be easily backdoored through multimodal data poisoning," *arXiv preprint arXiv:2305.04175*, 2023.

[14] S. Shan, J. Cryan, E. Wenger, H. Zheng, R. Hanocka, and B. Y. Zhao, "Glaze: Protecting artists from style mimicry by text-to-image models," in *Proc. of USENIX Security*, 2023.

[15] C. Liang, X. Wu, Y. Hua, J. Zhang, Y. Xue, T. Song, Z. Xue, R. Ma, and H. Guan, "Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples," in *Proc. of ICML*. PMLR, 2023, pp. 20 763–20 786.

[16] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[17] M. Zhu, P. Pan, W. Chen, and Y. Yang, "Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis," in *Proc. of CVPR*, 2019, pp. 5802–5810.

[18] M. Ding *et al.*, "Cogview: Mastering text-to-image generation via transformers," *Proc. of NeurIPS*, 2021.

[19] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. of CVPR*, 2022, pp. 10 684–10 695.

[20] A. Ramesh *et al.*, "Zero-shot text-to-image generation," in *Proc. of ICML*, 2021.

[21] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. of ICML*, 2015.

[22] Stability AI, "Stable Diffusion Public Release. ," 2022, https://stability.ai/blog/stable-diffusion-public-release.

[23] D. Podell *et al.*, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023.

[24] Stability AI, "Stability AI releases DeepFloyd IF, a powerful text-to-image model that can smartly integrate text into images," 2023, https://stability.ai/blog/deepfloyd-if-text-to-image-model.

[25] NovelAI, "NovelAI changelog," 2022, https://novelai.net/updates.

[26] A. Ramesh *et al.*, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*,

2022.
[27] Stability AI, "Stable Diffusion v2.1 and DreamStudio Updates 7-Dec 22," 2022, https://stability.ai/blog/stablediffusion2-1-release7-dec-2022.
[28] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proc. of CVPR*, 2021.
[29] C. Schuhmann *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *arXiv preprint arXiv:2210.08402*, 2022.
[30] N. Carlini *et al.*, "Poisoning web-scale training datasets is practical," *arXiv preprint arXiv:2302.10149*, 2023.
[31] StabilityAI, "Stable Diffusion v1-4 Model Card," 2022, https://huggingface.co/CompVis/stable-diffusion-v1-4.
[32] Stability AI, "Stable Diffusion 2.0 Release," 2022, https://stability.ai/blog/stable-diffusion-v2-release.
[33] Scenario.gg, "AI-generated game assets," 2022, https://www.scenario.gg/.
[34] Civitai, "What the heck is Civitai?" 2022, https://civitai.com/content/guides/what-is-civitai.
[35] T. H. Tran, "Image Apps Like Lensa AI Are Sweeping the Internet, and Stealing From Artists," 2022, https://www.thedailybeast.com/how-lensa-ai-and-image-generators-steal-from-artists.
[36] R. Gal *et al.*, "An image is worth one word: Personalizing text-to-image generation using textual inversion," *arXiv preprint arXiv:2208.01618*, 2022.
[37] N. Ruiz *et al.*, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proc. of CVPR*, 2023.
[38] M. Goldblum *et al.*, "Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses," *IEEE Trans. Pattern Anal. Mach. Intell*, 2022.
[39] Y. Liu *et al.*, "Trojaning attack on neural networks," in *Proc. of NDSS*, 2018.
[40] E. Wenger, J. Passananti, A. Bhagoji, Y. Yao, H. Zheng, and B. Y. Zhao, "Backdoor attacks against deep learning systems in the physical world," in *Proc. of CVPR*, 2021.
[41] K. Eykholt *et al.*, "Robust physical-world attacks on deep learning visual classification," in *Proc. of CVPR*, 2018, pp. 1625–1634.
[42] X. Chen *et al.*, "Badnl: Backdoor attacks against nlp models with semantic-preserving improvements," in *Proc. of ACSAC*, 2021, pp. 554–569.
[43] A. Saha, A. Subramanya, and H. Pirsiavash, "Hidden trigger backdoor attacks," in *Proc. of AAAI*, no. 07, 2020.
[44] A. Turner, D. Tsipras, and A. Madry, "Clean-label backdoor attacks," 2018.
[45] C. Zhu, W. R. Huang, H. Li, G. Taylor, C. Studer, and T. Goldstein, "Transferable clean-label poisoning attacks on deep neural nets," in *Proc. of ICML*, 2019.
[46] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *Proc. of IEEE S&P*. IEEE, 2019, pp. 707–723.
[47] X. Qiao, Y. Yang, and H. Li, "Defending neural backdoors via generative distribution modeling," *Proc. of NeurIPS*, 2019.
[48] H. Chen, C. Fu, J. Zhao, and F. Koushanfar, "Deepinspect: A black-box trojan detection and mitigation framework for deep neural networks." in *IJCAI*, 2019.
[49] B. Chen *et al.*, "Detecting backdoor attacks on deep neural networks by activation clustering," *arXiv preprint arXiv:1811.03728*, 2018.
[50] X. Liu, F. Li, B. Wen, and Q. Li, "Removing backdoor-based watermarks in neural networks with limited data," in *Proc. of ICPR*, 2021.
[51] J. Jia, X. Cao, and N. Z. Gong, "Intrinsic certified robustness of bagging against data poisoning attacks," in *Proc. of AAAI*, 2021.
[52] J. Geiping *et al.*, "What doesn't kill you makes you robust (er): Adversarial training against poisons and backdoors," *arXiv preprint arXiv:2102.13624*, 2021.
[53] B. Wang, X. Cao, N. Z. Gong *et al.*, "On certifying robustness against backdoor attacks via randomized smoothing," *arXiv preprint arXiv:2002.11750*, 2020.
[54] Y. Yao, H. Li, H. Zheng, and B. Y. Zhao, "Latent backdoor attacks on deep neural networks," in *Proc. of CCS*, 2019, pp. 2041–2055.
[55] G. Severi, J. Meyer, S. Coull, and A. Oprea, "Explanation-guided backdoor poisoning attacks against malware classifiers," in *Proc. of USENIX Security*, 2021.
[56] E. Bagdasaryan and V. Shmatikov, "Blind backdoors in deep learning models," in *Proc. of USENIX Security*, 2021, pp. 1505–1521.
[57] R. Schuster, C. Song, E. Tromer, and V. Shmatikov, "You autocomplete me: Poisoning vulnerabilities in neural code completion," in *Proc. of USENIX Security*, 2021.
[58] A. Wan, E. Wallace, S. Shen, and D. Klein, "Poisoning language models during instruction tuning," *arXiv preprint arXiv:2305.00944*, 2023.
[59] J. Zhang, H. Liu, J. Jia, and N. Z. Gong, "Corruptencoder: Data poisoning based backdoor attacks to contrastive learning," *arXiv preprint arXiv:2211.08229*, 2022.
[60] Z. Yang, X. He, Z. Li, M. Backes, M. Humbert, P. Berrang, and Y. Zhang, "Data poisoning attacks against multimodal encoders," in *Proc. of ICML*, 2023.
[61] H. Liu, W. Qu, J. Jia, and N. Z. Gong, "Pre-trained encoders in self-supervised learning improve secure and privacy-preserving supervised learning," *arXiv preprint arXiv:2212.03334*, 2022.
[62] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," in *Proc. of MLCS Workshop*, 2017.
[63] Y. Lu, G. Kamath, and Y. Yu, "Indiscriminate data poisoning attacks on neural networks," *arXiv preprint arXiv:2204.09092*, 2022.
[64] B. Biggio, B. Nelson, and P. Laskov, "Support vector machines under adversarial label noise," in *Proc. of ACML*, 2011.
[65] C. Schuhmann, "Laion-aesthetics," LAION.AI, Aug 2022.
[66] F. Bond and K. Paik, "A survey of wordnets and their licenses," in *Proc. of GWC*, 2012.
[67] "Midjourney user prompts; generated images (250k)." [Online]. Available: https://www.kaggle.com/ds/2349267
[68] 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 *Proc. of ICML*, 2021.
[69] C. Fellbaum, "Wordnet and wordnets. encyclopedia of language and linguistics," 2005.
[70] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proc. of ACL*, 2018.
[71] X. He, S. Zannettou, Y. Shen, and Y. Zhang, "You only prompt once: On the capabilities of prompt learning on large language models to tackle toxic content," *arXiv preprint*

arXiv:2308.05596, 2023.
[72] B. Saleh and A. Elgammal, "Large-scale classification of fine-art paintings: Learning the right metric on the right feature," *arXiv preprint arXiv:1505.00855*, 2015.
[73] A. Hoare, "Digital Illustration Styles," 2021, https://www.theillustrators.com.au/digital-illustration-styles.
[74] H. Salman, A. Khaddaj, G. Leclerc, A. Ilyas, and A. Madry, "PhotoGuard: Defending Against Diffusion-based Image Manipulation," 2022, https://gradientscience.org/photoguard/.
[75] A. Shafahi *et al.*, "Poison frogs! targeted clean-label poisoning attacks on neural networks," *arXiv preprint arXiv:1804.00792*, 2018.
[76] H. Aghakhani, D. Meng, Y.-X. Wang, C. Kruegel, and G. Vigna, "Bullseye polytope: A scalable clean-label poisoning attack with improved transferability," *arXiv preprint arXiv:2005.00191*, 2020.
[77] A. Schwarzschild *et al.*, "Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks," in *Proc. of ICML*. PMLR, 2021, pp. 9389–9398.
[78] V. Cherepanova *et al.*, "Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition," *arXiv preprint arXiv:2101.07922*, 2021.
[79] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. of CVPR*, 2018, pp. 586–595.
[80] J. Nocedal and S. Wright, "Numerical optimization, series in operations research and financial engineering," *Springer, New York, USA, 2006*, 2006.
[81] E. Zhang *et al.*, "Forget-me-not: Learning to forget in text-to-image diffusion models," *arXiv preprint arXiv:2303.17591*, 2023.
[82] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Proc. of ECCV*. Springer, 2014, pp. 740–755.
[83] A. Hertz *et al.*, "Prompt-to-prompt image editing with cross attention control," *arXiv preprint arXiv:2208.01626*, 2022.
[84] A. Vahdat, K. Kreis, and J. Kautz, "Score-based generative modeling in latent space," *Proc. of NeurIPS*, 2021.
[85] D. H. Park, S. Azadi, X. Liu, T. Darrell, and A. Rohrbach, "Benchmark for compositional text-to-image synthesis," in *Proc. of NeurIPS*, 2021.
[86] M. Heusel *et al.*, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Proc. of NeurIPS*, 2017.
[87] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *Proc. of CVPR*, 2018, pp. 1316–1324.
[88] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao, "Fawkes: Protecting privacy against unauthorized deep learning models," in *Proc. of USENIX Security*, 2020.
[89] L. Wu *et al.*, "Understanding and enhancing the transferability of adversarial examples," *arXiv preprint arXiv:1802.09707*, 2018.
[90] E. Bagdasaryan and V. Shmatikov, "Spinning language models: Risks of propaganda-as-a-service and countermeasures," in *Proc. of IEEE S&P*, 2022.
[91] I. Shumailov *et al.*, "The curse of recursion: Training on generated data makes models forget," *arXiv preprint arxiv:2305.17493*, 2023.
[92] C. Schuhmann *et al.*, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," *arXiv preprint arXiv:2111.02114*, 2021.
[93] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. of ICML*, 2022.
[94] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator. corr abs/1411.4555 (2014)," *arXiv preprint arXiv:1411.4555*, 2014.
[95] C. Lee, J. Jang, and J. Lee, "Personalizing text-to-image generation with visual prompts using blip-2," in *Proc. of ICML*, 2023.
[96] T. Nguyen, S. Y. Gadre, G. Ilharco, S. Oh, and L. Schmidt, "Improving multimodal datasets with image captioning," *arXiv preprint arXiv:2307.10350*, 2023.
[97] C. Xiang, "Ai is probably using your images and it's not easy to opt out," Motherboard, Tech by Vice, Sept 2022.
[98] E. David, "Now you can block openai's web crawler," TheVerge, August 2023.
[99] L. Bourtoule *et al.*, "Machine unlearning," in *Proc. of IEEE S&P*, 2021.
[100] S. Neel, A. Roth, and S. Sharifi-Malvajerdi, "Descent-to-delete: Gradient-based methods for machine unlearning," in *Proc. of ALT*, 2021.

Appendix

1. Experiment Setup

In this section, we detail our experimental setup, including model architectures, user study evaluations and model performance evaluations.

Details on model architecture. In §6.1, we already describe the LD-CC model for the training from scratch scenario. Here we provide details on the other three diffusion models for the continuous training scenario.

- *Stable Diffusion V2 (SD-V2):* We simulate the popular training scenario where the model trainer updates the pretrained Stable Diffusion V2 model (SD-V2) [27] using new training data [34]. SD-V2 is trained on a subset of the LAION-aesthetic dataset [29]. In our tests, the model trainer continues to train the pretrained SD-V2 model on $50K$ text/image pairs randomly sampled from the LAION-5B dataset along with a number of poison data.
- *Stable Diffusion XL (SD-XL):* Stable Diffusion XL (SD-XL) is the newest and the state-of-the-art diffusion model, outperforming SD-V2 in various benchmarks [23]. The SD-XL model has over 2.6B parameters compared to the 865M parameters of SD-V2. SD-XL is trained on an internal dataset curated by StablityAI. In our test, we assume a similar training scenario where the model trainer updates the pretrained SD-XL model on a randomly selected subset (50K) of the LAION-5B dataset and a number of poison data.
- *DeepFloyd (DF):* DeepFloyd [24] (DF) is another popular diffusion model that has a different model architecture from LD, SD-V2, and SD-XL. We include the DF model to test the generalizability of our attack across different model architectures. Like the above, the model trainer updates the pretrained DF model using a randomly selected subset (50K) of the LAION-5B dataset and a number of poison data.

Details on user study. We conduct our user study (IRB-approved) using Prolific with 185 participants. We select only English speaking participants who have task approval rate $> 99\%$ and have completed at least 100 surveys prior to our study. We compensate each participant at a rate of \$15/hr.

Details on evaluating a model's CLIP alignment score and FID. We follow prior work [19, 37] to query the poisoned model with

20K MSCOCO text prompts (covering a variety of objects and styles) and generates 20K images. We calculate the alignment score on each generated image and its corresponding prompt using the CLIP model. We calculate FID by comparing the generated images with clean images in the MSCOCO dataset using an image feature extractor model [86].

2. PCA Visualization of Concept Sparsity

We also visualize semantic frequency of text embeddings in an 2D space. Figure 18 provides a feature space visualization of the semantic frequency for all the common concepts (nouns), compressed via PCA. Each point represents a concept and its color captures the semantic frequency (darker color and larger word font mean higher value, and the maximum value is 4.17%). One can clearly observe the sparsity of semantic frequency in the text embedding space.

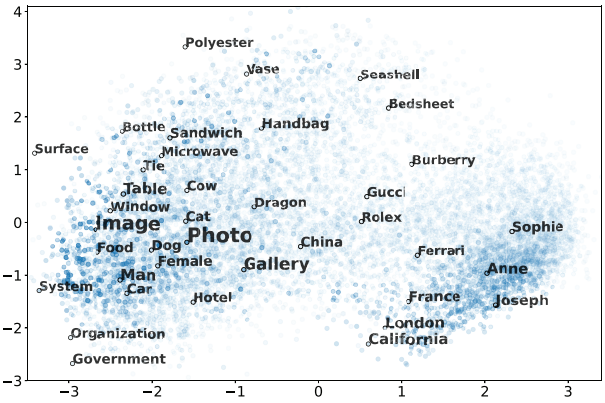

Figure 18. 2D PCA visualization of semantic frequency in LAION-Aesthetic. Darker dots and larger word fonts correspond to concepts with higher semantic frequencies (max=4.17%). We randomly pick concepts to show their word content.

3. Additional Results of Simple Dirty-Label Poisoning Attacks

Attacking LD-CC. Figure 20 illustrates the attack success rate of the simple, dirty-label poisoning attack (§4), evaluated by both a CLIP-based classifier and human inspectors. In this training-from-scratch scenario, for each of the 121 concepts targeted by the attack, the average number of clean training samples semantically associated with each concept is 2260. Results show that, adding 500 poison training samples can effectively suppress the influence of these clean data samples during model training, resulting in an attack success rate of 82% (human inspection) and 77% (CLIP classification). Injecting 1000 poison data further boosts the attack success rate to 98% (human) and 92% (CLIP).

Attacking SD-V2, SD-XL, DeepFloyd. Figure 21 shows the poisoning result in the continuous training scenario assessed by the CLIP classifier and Figure 22 shows the result evaluated via human inspection. Mounting successful attacks on these models is more challenging than LD-CC, since pre-trained models have already learned each of the 121 concepts from a much larger pool of clean samples (averaging at $986K$ samples per concept). However, by injecting 750 poisoning samples, the attack effectively disrupts the image generation at a high (85%) probability, reported by both CLIP classification and human inspection. Injecting 1000 poisoning samples pushes the success rate beyond 90%.

Figure 23 compares the CLIP attack success rate between object and style concepts. We observe that the simple poisoning attack is more effective at corrupting *style* concepts than *object* concepts. This is likely because styles are typically conveyed visually by the entire image, while objects define specific regions within the image.

Concept Sparsity Affecting Attack Efficacy. Figure 24 demonstrates how concept sparsity in terms of word frequency impacts attack efficacy and we further study the impact of semantic frequency in Figure 25. For this we sample 15 object concepts with varying sparsity levels, in terms of word and semantic frequency discussed in §3.3. As expected, poisoning attack is more successful when disrupting more sparse concepts Moreover, semantic frequency is a more accurate representation of concept sparsity than word frequency, because we see higher correlation between semantic frequency and attack efficacy. These empirical results confirm our hypothesis in §3.2.

Task	CLIP attack success rate on artist names		
	100 poison	200 poison	300 poison
LD-CC	80%	91%	96%
SD-V2	81%	94%	97%
SD-XL	77%	92%	99%
DF	80%	96%	99%

TABLE 7. Poison attack damages related concepts (artist names) when the attacker poisons given art styles across 4 generation models.

L2 Distance to source concept(D)	Average Number of Concepts Included	Average CLIP attack success rate		
		100 poison	200 poison	300 poison
$D = 0$	1	84%	94%	96%
$0 < D \leq 3.0$	5	81%	93%	96%
$3.0 < D \leq 6.0$	13	78%	90%	92%
$6.0 < D \leq 9.0$	52	32%	41%	59%
$D > 9.0$	1929	5%	5%	6%

TABLE 8. Bleed through performance of the enhanced poison. (SD-XL)

4. Additional Results on Bleed through and Stacking Multiple Attacks

We evaluate the "related" concept bleed-through effects between artists and the art styles they are known for. We include 195 artists associated with 28 styles from the Wikiart dataset [72]. We poison each art style $\mathcal{C}$, then test poison's impact on generating painting of artists whose style belong to style $\mathcal{C}$, without mentioning the poisoned style $\mathcal{C}$ in the prompt, *e.g.,* query with "a painting by Picasso" for models with "cubism" poisoned. Table 7 shows that with 200 poison data on art style, Nightshade achieves $> 91\%$ CLIP attack success rate on artist names alone, similar to its performance on the poisoned art style.

Enhancing bleed-through. We can further enhance our poison attack's bleed though by broadening the sampling pool of poison text prompts: sampling text prompts in the text semantic space of $\mathcal{C}$ rather than with exact word match to $\mathcal{C}$. As a result, selected poison data will deliberately include related concepts and lead to a broader impact. Specifically, when we calculate activation similar to the poisoned concept $\mathcal{C}$, we use all prompts in LAION-5B dataset (does not need to include $\mathcal{C}$). Then we select top 5K prompts with the highest activation, which results in poison prompts containing both $\mathcal{C}$ and nearby concepts. We keep the rest of our poison generation algorithm identical. This enhanced attack increases bleed through by 11% in some cases while having minimal performance degradation ($< 1\%$) on the poisoned concept (Table 8).

Stacking multiple poisons. Table 9 lists, for the LD-CC model, the overall model performance in terms of the CLIP alignment score and FID, when an increased number of concepts are being poisoned.

Approach	# of poisoned concepts	Overall model Performance	
		Alignment Score (higher better)	FID (lower better)
Clean LD-CC	0	0.31	17.2
Poisoned LD-CC	100	0.29	22.5
Poisoned LD-CC	250	0.27	29.3
Poisoned LD-CC	500	0.24	36.1
Poisoned LD-CC	1000	0.22	44.2
AttnGAN	-	0.26	35.5
A model that outputs random noise	-	0.20	49.4

TABLE 9. Overall model performance (in terms of the CLIP alignment score and FID) when an increasing number of concepts are being poisoned. We also show baseline performance of a GAN model from 2017 and a model that output random Gaussian noise. (LD-CC)

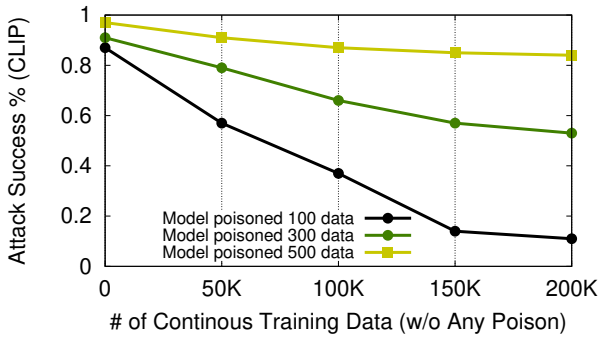

Figure 19. Nightshade's attack success rate (CLIP-based) decreases when mdoel trainer continously trains an already-poisoned model on an increasing number of clean data. The base model is poisoned with 100, 300, and 500 poison data.

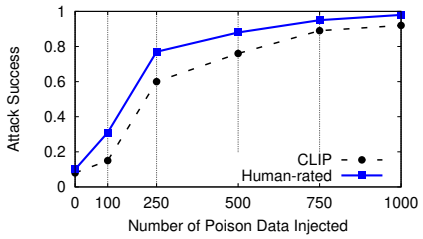

Figure 20. Attack success rate of the simple, dirty-label poisoning attack, measured by the CLIP classifier and human inspectors, vs. # of poison data injected, when attacking LD-CC (training from scratch).

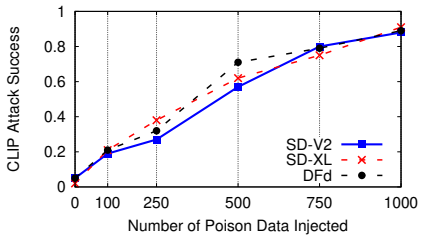

Figure 21. Attack success rate of the simple, dirty-label poisoning attack, measured by the CLIP classifier, vs. # of poison data injected, when attacking each of three models SD-V2, SD-XL, DeepFloyd (continuous training).

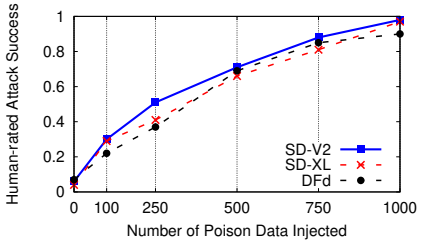

Figure 22. Attack success rate of the simple, dirty-label poisoning attack, measured by human inspectors, vs. # of poison data injected, when attacking each of three models SD-V2, SD-XL, DeepFloyd (continuous training).

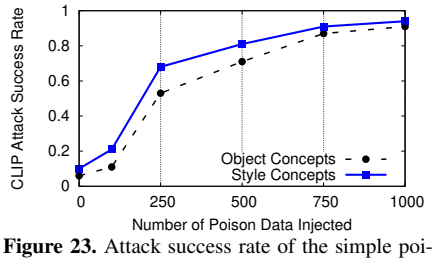

Figure 23. Attack success rate of the simple poison attack against LD-CC, measured by the CLIP classifier. The simple poisoning attack is more effective at corrupting style concepts than object concepts. The same applies to attacks against SD-V2, SD-XL, DeepFloyd.

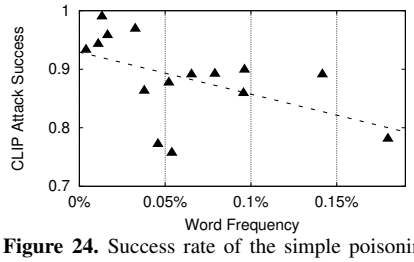

Figure 24. Success rate of the simple poisoning attack (rated by CLIP classifier) is weakly correlated with concept sparsity measured by word frequency in the training data. Results for LD-CC. Same trend observed on SD-V2, SD-XL, DeepFloyd.

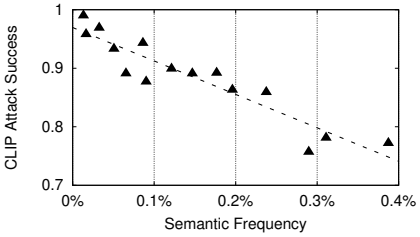

Figure 25. Success rate of the simple poisoning attack (rated by CLIP classifier) correlates strongly with concept sparsity measured by semantic frequency. Results for LD-CC. Same trend observed on SD-V2, SD-XL, DeepFloyd.