非英語環境で顕在化するText-to-Imageモデルの脆弱性：10言語におけるデータ汚染攻撃の体系的評価

掛林 諒平[1,a)]　森 達哉[1,2,3]

概要：Text-to-Image モデルは，手軽に高品質な画像生成を可能とし，多様な分野で利用が拡大している．一方で，学習過程に悪意のあるデータを混入させ，モデルの意図しない挙動を誘発するデータ汚染攻撃の脅威が指摘されている．近年の研究では，英語を対象とした評価に限定されてきたが，モデルの利用者は多言語にわたり，非英語環境での影響は十分に検証されていない．そこで，本研究では，Stable Diffusion 2.0 および Stable Diffusion XL を対象に，英語に加え，イタリア語，インドネシア語，オランダ語，スペイン語，ドイツ語，トルコ語，日本語，フランス語，ポルトガル語の 10 言語を対象とした体系的な攻撃評価を行った．その結果，非英語言語で攻撃成功率が高まり，少量の汚染データで攻撃が成功することを明らかにした．さらに，埋め込みベクトルの可視化分析から，テキストエンコーダの多言語対応能力が，攻撃に対する頑健性を左右する重要な要因であることを示唆した．

キーワード：データ汚染攻撃，Text-to-Image モデル，非英語環境

Emerging Vulnerabilities of Text-to-Image Models in Non-English Environments: Systematic Evaluation of Data Poisoning Attacks Across 10 Languages

Ryohei Kakebayashi[1,a)]　Tatsuya Mori[1,2,3]

Abstract: Text-to-Image models enable easy generation of high-quality images and are being increasingly utilized across diverse fields. However, threats from data poisoning attacks have been identified, where malicious data is introduced into the training process to induce unintended model behaviors. Recent research has been limited to evaluations targeting English, but model users span multiple languages, and the impact in non-English environments has not been sufficiently explored. Therefore, in this study, we conducted systematic attack evaluations on Stable Diffusion 2.0 and Stable Diffusion XL across 10 languages: Dutch, English, French, German, Indonesian, Italian, Japanese, Portuguese, Spanish, and Turkish. Our results revealed that attack success rates increase in non-English languages and that attacks succeed with a small amount of poisoned data. Furthermore, through analysis visualizing embedding vectors, we demonstrated that the multilingual capability of text encoders may be a crucial factor determining robustness against attacks.

Keywords: Data Poisoning Attacks, Text-to-Image Models, Non-English Environments

1. はじめに

Text-to-Image モデルは，著しい発展を遂げており，自然言語による記述に基づいて，高品質な画像を生成する技術として，研究分野と実応用の両面で注目されている．図 1 の「攻撃前」に示すように，複数の要素を含むプロンプトであっても，対応する画像を忠実に生成することが可能である．また，ネガティブプロンプトを用いることで，生成画像から除外すべき要素を明示的に指定すること

[1] 早稲田大学/Waseda University
[2] 情報通信研究機構/NICT
[3] 理化学研究所 革新知能統合研究センター/RIKEN AIP
[a)] kake@nsl.cs.waseda.ac.jp

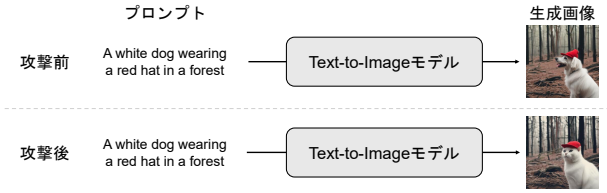

図 1 Text-to-Image モデルにおけるデータ汚染攻撃の前後比較

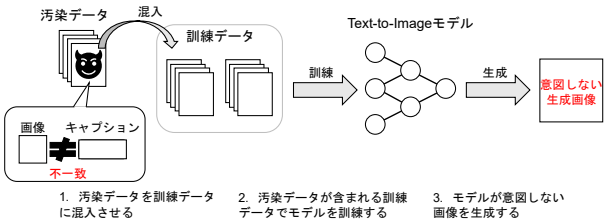

図 2 データ汚染攻撃のプロセス概要

もできる．代表的なモデルとしては，Stable Diffusion [1]，Midjourney [2]，DALL-E [3] などが挙げられ，既に多様な分野において活用が進んでいる．一方で，こうした技術の進展に伴い，安全性や信頼性に対する懸念も指摘されており，攻撃手法の検討や影響評価，さらに防御手法の確立が急務となっている．

Text-to-Image モデルに対する代表的な攻撃手法の一つが，データ汚染攻撃である．図 2 に示すように，この攻撃は，モデルの学習過程において悪意のある汚染データを混入させることで，画像の生成時に意図しない挙動を誘発する．これにより，プロンプトと乖離した画像の生成や，画像品質の意図的な劣化を引き起こすことが可能となる．図 1 の例では，攻撃前は “dog” のプロンプトに適合する画像が生成されていたが，攻撃後には “cat” の画像が生成されるように変化している．

こうした単純な例にとどまらず，社会的に重要な文脈で同様の攻撃が生じると，深刻な影響を及ぼす可能性がある．実際に，無害なプロンプトからヘイトミームを生成させ，特定の個人やコミュニティに対する差別的なコンテンツが拡散されうることが指摘されている [4]．さらに，“young people” といった一般的な語を特定の政治家の顔を結びつけることで，誤解や偏見を助長し，政治的リスクを招く可能性が報告されている [5]．したがって，このような挙動は，モデル開発者やサービス提供者の信頼性を大きく損なう恐れがあり，データ汚染攻撃への対策は極めて重要な課題である．

近年の研究では，大規模な訓練データを用いる Text-to-Image モデルにおいても，わずかな汚染データの混入によって，データ汚染攻撃が成功することが報告されている [6], [7]．これらの研究はいずれも，英語のキャプションを用いた汚染データを作成し，英語のプロンプトによる攻撃評価を行っている．英語は国際的に広く利用される言語であり，評価の基盤として適切であるが，Text-to-Image

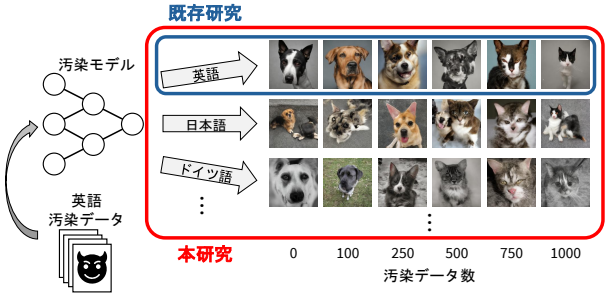

図 3 本研究の全体像

モデルの利用者は必ずしも英語話者に限らない．実際に，多くのモデルは，英語以外のプロンプトに対しても，適切な画像を生成できることが確認されている [8]．したがって，モデル開発者が非英語言語を明示的に想定していない場合であっても，モデルが多言語に対応している限り，非英語言語におけるデータ汚染攻撃の影響を評価することが不可欠である．

以上を踏まえ，本研究では，図 3 に示すように，非英語言語におけるデータ汚染攻撃の影響を評価する．具体的には，英語に加え，イタリア語，インドネシア語，オランダ語，スペイン語，ドイツ語，トルコ語，日本語，フランス語，ポルトガル語の計 10 言語を対象とし，実験を通じて攻撃の影響を比較する．さらに，各言語における攻撃の影響差については，テキストエンコーダの多言語対応能力に着目し，考察を行う．

本研究の主な貢献は，以下の通りである．

- 英語環境に限定されてきた既存研究に対し，Stable Diffusion 2.0 および Stable Diffusion XL を対象に，10 言語でデータ汚染攻撃を体系的に評価した．
- CLIP 分類器による評価により，非英語言語において攻撃成功率が高い傾向を示すことを明らかにした．特に，日本語とインドネシア語では，わずか 250 件の汚染データで攻撃成功率が 50%を超え，非英語環境における深刻な脆弱性が存在することを示した．
- ユーザスタディによる評価を行い，非英語言語において攻撃成功率が高くなることを利用者視点からも確認し，現実的に無視できない脅威であることを確認した．
- 埋め込みベクトルの可視化分析を通じて，非英語言語が英語の分布から逸脱しやすいことを示し，テキストエンコーダの多言語対応能力が，モデルの頑健性に大きく影響し得ることを明らかにした．また，この分析により，多言語対応テキストエンコーダの利用を多言語環境下での防御手法として提案した．

2. 背景

2.1 Text-to-Image モデル

画像生成モデルは，変分オートエンコーダ（VAE）や敵対的生成ネットワーク（GAN）を基盤として発展してき

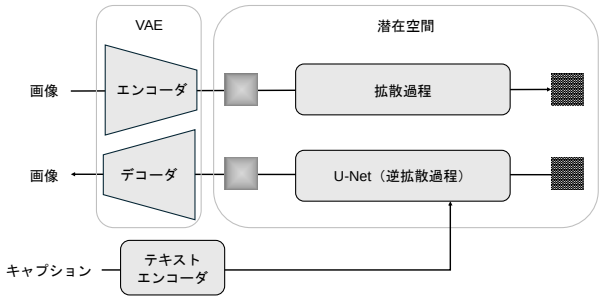

図 4 Stable Diffusion の全体構成

た．その中でも，Text-to-Image モデルは，生成したい画像を自然言語で記述することで画像を生成可能であり，拡散モデルの導入により，高品質な画像生成を実現している．さらに，最新の Text-to-Image モデルでは，計算コストを抑えつつ生成品質を向上させる潜在拡散モデル（LDM）が広く利用されている．

Stable Diffusion [1] は，代表的なオープンソースの Text-to-Image モデルであり，図 4 に示すように，VAE，U-Net，およびテキストエンコーダから構成される．画像生成の流れは以下の通りである．まず，VAE エンコーダが画像をピクセル空間から低次元の潜在空間へ変換する．次に，この潜在空間にガウシアンノイズを加える拡散過程を経た後，U-Net が逆拡散過程によりノイズを除去する．この際，テキストエンコーダによりキャプションが埋め込みベクトルへ変換され，テキストによる条件付けが可能となる．最後に，VAE デコーダが潜在空間をピクセル空間へ復元し，最終的な画像が得られる．

Stable Diffusion には複数のバージョンが存在するが，本研究では，Stable Diffusion 2.0 と Stable Diffusion XL を用いる．両モデルの基本的な仕組みは同じであるが，表 1 に示すような相違点がある．

Stable Diffusion 2.0 は，約 50 億組の画像とキャプションを含む LAION-5B [9] を用いて訓練されており，単一の U-Net と OpenCLIP ViT-H をテキストエンコーダとして利用している．一方，Stable Diffusion XL は，内部データセットで訓練されており，詳細は非公開であるが，U-Net は Base モデルと Refiner モデルの二段構成を採用し，さらに CLIP ViT-L と OpenCLIP ViT-bigG の 2 種類のテキストエンコーダによる埋め込みベクトルを結合している．これにより，短文と長文いずれのプロンプトに対しても高い記述力を発揮する．

Text-to-Image モデルをゼロから訓練するには，膨大なコストが必要となるため，実際には事前学習済みモデルをベースに，新たなデータを用いてファインチューニングすることが多い．Stable Diffusion においては，VAE およびテキストエンコーダは固定し，U-Net のみを再訓練する手法が一般的である．

2.2 データ汚染攻撃

データ汚染攻撃とは，意図的に改変された汚染データを訓練データに混入させ，モデルに意図しない挙動を引き起こす攻撃である．本研究では，Text-to-Image モデルに対し，プロンプトと乖離した画像を生成させることを攻撃の目的とする．

Text-to-Image モデルの訓練には，画像とそのキャプションのペアからなる大規模データセットが用いられる．データセットの多くは，ウェブ上の画像と ALT テキストを収集して構築されるため，全データの品質を人手で検証することは現実的に不可能である．このため，Text-to-Image モデルはデータ汚染攻撃に対して脆弱であることが報告されている [10]．

汚染データの典型例は，画像とキャプションの意味的な不一致である．例えば，“cat” の画像に “dog” を含むキャプションを与えて学習させると，訓練後のモデルは “dog” を含むプロンプトに対して “cat” の画像を生成するようになる．汚染データは，その作成方法に基づき，以下の 2 種類に分類される [7]．

dirty-label 攻撃: キャプションを改変し，ペアとなる画像の内容と一致しない説明に変更する．

clean-label 攻撃: 画像に微小なノイズを付加し，視覚的特徴を他クラスへと変化させる．

攻撃者は，訓練データに汚染データを混入させることができ，画像やキャプションを任意に改変可能であるとする．一方，モデル内部や訓練過程への直接的なアクセスは不要とする．攻撃の手法としては，汚染データを含むデータセットを正常なデータセットと見せかけて公開する方法や，ウェブ上に汚染画像や誤った ALT テキストをアップロードする方法が挙げられる．特に，Text-to-Image モデルの訓練データは，ウェブから収集されることが多いため，意図せず汚染データが混入するリスクも高い．

2.3 他分野における多言語評価

他分野では，多言語環境下における安全性評価が進められている．例えば，大規模言語モデル（LLM）は，事前学習段階において低リソース言語のデータが不足しているため，低リソース言語で悪意のあるプロンプトが入力されると，有害な応答を生成しやすいことが報告されている [11]．また，LLM の回答性能が相対的に低い言語を利用して，悪意のあるプロンプトを巧妙に埋め込み，安全機構を回避して有害な出力を生成させる手法も報告されている [12]．

3. 実験手法

本研究では，評価対象の攻撃として dirty-label 攻撃を選択し，事前学習済みモデルである Stable Diffusion 2.0 および Stable Diffusion XL を対象とする．また，ファインチューニングに用いる訓練データに汚染データが混入する

表 1　Stable Diffusion 2.0 と Stable Diffusion XL の比較

モデル名	訓練データ	U-Net	テキストエンコーダ
Stable Diffusion 2.0	LAION-5B	単一構成	OpenCLIP ViT-H
Stable Diffusion XL	内部データセット	二段構成（Base & Refiner）	CLIP ViT-L & OpenCLIP ViT-bigG

表 2　攻撃対象クラス C_t と誘導先クラス C_d のペア

攻撃対象クラス C_t	誘導先クラス C_d
dog	cat
airplane	car
cat	horse
panda	sheep

シナリオを想定し，攻撃の評価を行う．

3.1　攻撃対象クラスと誘導先クラス

攻撃対象クラスを C_t（target class），誘導先クラスを C_d（destination class）と定義する．例えば，C_t = “dog”，C_d = “cat” とした場合，“dog” を含むプロンプトに対して，“cat” の画像を生成させる攻撃となる．

本研究では，表 2 に示す 4 組のクラスペアを対象として実験を行った．

3.2　汚染データの作成手法

汚染データの作成には，LAION-Aesthetics [13] を用いる．このデータセットの “URL” カラムには画像の URL が，“TEXT” カラムには画像の ALT テキストが含まれる．また，BLIP [14] を用いて，各画像に対応するキャプションを自動生成する．

汚染データは，以下の手順で作成する．

(1) LAION-Aesthetics の “TEXT” カラムに誘導先クラス C_d が含まれるデータを抽出し，対応する画像を “URL” カラムから取得する．
(2) (1) で取得した画像に対して，BLIP によりキャプションを生成する．
(3) (2) で生成したキャプション内の誘導先クラス C_d を攻撃対象クラス C_t に置き換える．
(4) (1) で取得した画像と，(3) で置き換えたキャプションをペアとして汚染データを作成する．

ここで，“TEXT” カラムを直接キャプションとして利用しないのは，ALT テキストが画像内容を正確に表現していない場合が多く，そのまま用いると汚染データの効果が低減する可能性があるためである．BLIP によるキャプション生成を用いることで，画像内容に即したキャプションを得ている．

上記の手順を繰り返すことで，視覚的には誘導先クラス C_d に属しながら，キャプションには攻撃対象クラス C_t を含む汚染データが得られる．この汚染データを訓練に利用すると，モデルは攻撃対象クラス C_t の学習時に誤って誘導先クラス C_d の視覚的特徴を取り組むことになる．

3.3　データ汚染攻撃の実装手法

データ汚染攻撃は，以下の手順で実装する．

(1) LAION-Aesthetics の “URL” カラムから取得した画像と “TEXT” カラムを使用したキャプションのペアからなる正常データ 10 万件を収集する．
(2) (1) で収集した正常データに対し，3.2 節で作成した汚染データを混入させ，訓練データを構築する．汚染データの混入数は 0，100，250，500，750，1000 件とする．
(3) (2) で構築したデータを用い，Stable Diffusion の学習コード [15] によりファインチューニングを行う．

ここで，(1) の正常データには，攻撃者による意図的な改変は含まれず，攻撃対象クラス C_t や誘導先クラス C_d 以外の画像とキャプションも含まれる．

3.4　評価手法

評価対象の言語は，英語に加え，イタリア語，インドネシア語，オランダ語，スペイン語，ドイツ語，トルコ語，日本語，フランス語，ポルトガル語の計 10 言語とする．これらの言語は，ウェブサイトで使用されている言語の上位 20 言語に含まれ [16]，攻撃前の Stable Diffusion 2.0 および Stable Diffusion XL に対し，それぞれの言語における “dog” に対応する単語（例えば日本語の「犬」，イタリア語の “cane”）を入力した際に，画像を正しく生成可能であることを確認している．

データ汚染攻撃の評価指標には，攻撃成功率を用いる．“a photo of C_t” を各言語に翻訳したプロンプトを用いて，画像を 1000 枚生成し，以下の 2 つの方法で攻撃成功率を算出する．

CLIP 分類器: 生成画像 1000 枚を CLIP 分類器により，攻撃対象クラス C_t または誘導先クラス C_d に分類し，誘導先クラス C_d と分類された割合を攻撃成功率とする．

ユーザスタディ: 生成画像 1000 枚からランダムに 100 枚を抽出し，“Definitely C_t”，“Mostly C_t”，“Somewhat C_t”，“Neutral”，“Somewhat C_d”，“Mostly C_d”，“Definitely C_d” の 7 段階評価を実施した．各画像について 3 名が評価を行い，そのうち 2 名以上が “Somewhat C_d”，“Mostly C_d”，“Definitely C_d” のいずれかに分類した場合を攻撃成功とし，その割合を攻撃成功率とする．ユーザスタディは，Amazon Mechanical Turk を利用し，タスク承認率が 99% 以上かつタスク承認数が 100 件以上の参加者を対象とした．

表 3 は，攻撃前の Stable Diffusion 2.0 および Stable Diffusion XL で各クラスの画像を生成した際に，画像生成が

表 3　攻撃前の Stable Diffusion 2.0 および Stable Diffusion XL における各言語，各クラスの画像生成成否 (✓: 成功，✗: 失敗)

言語	Stable Diffusion 2.0				Stable Diffusion XL			
	dog	airplane	cat	panda	dog	airplane	cat	panda
英語	✓	✓	✓	✓	✓	✓	✓	✓
イタリア語	✓	✓	✓	✓	✓	✓	✓	✓
インドネシア語	✓	✗	✓	✓	✓	✓	✓	✓
オランダ語	✓	✓	✗	✓	✓	✗	✓	✗
スペイン語	✗	✓	✓	✓	✓	✓	✓	✓
ドイツ語	✓	✓	✓	✓	✓	✓	✓	✓
トルコ語	✓	✗	✗	✓	✓	✗	✓	✓
日本語	✓	✗	✓	✗	✓	✓	✓	✗
フランス語	✓	✓	✓	✓	✓	✓	✓	✓
ポルトガル語	✓	✓	✓	✓	✓	✓	✓	✓

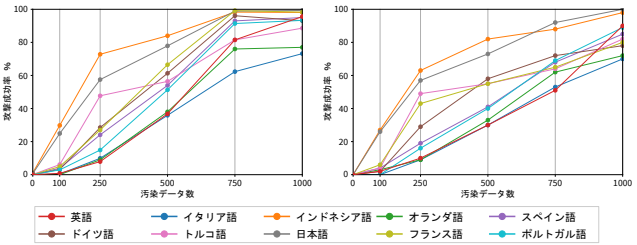

図 5　攻撃対象クラス C_t = “dog” における攻撃成功率の変化．左: CLIP 分類器．右: ユーザスタディ．

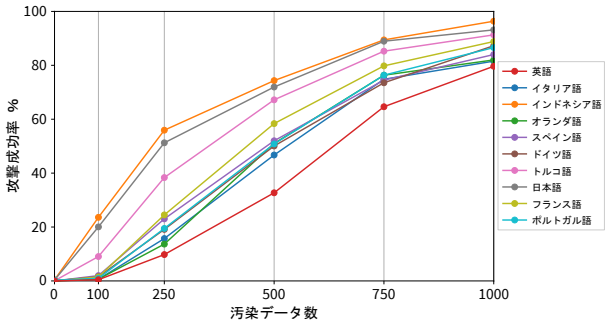

図 6　Stable Diffusion 2.0 における汚染データ数と攻撃成功率の関係

成功したかどうかを示す．ここでは，CLIP 分類器による判定で攻撃成功率が 1%未満の場合に，そのクラスの画像生成を成功と定義した．実験では，各言語で画像生成が成功したクラスのみを対象とする．

4.　実験結果

正常データ 10 万件に汚染データを 0，100，250，500，750，1000 件混入させ，Stable Diffusion 2.0 および Stable Diffusion XL をファインチューニングした．その後，画像を生成し，CLIP 分類器およびユーザスタディにより分類を行い，攻撃成功率を算出した．

4.1　CLIP 分類器とユーザスタディの比較

攻撃成功率の算出に用いる分類方法として，3.4 節で述べた CLIP 分類器とユーザスタディを比較する．具体的には，C_t = “dog”，C_d = “cat” とした場合に，Stable Diffusion 2.0 を評価対象モデルとした結果を取り上げる．

汚染データ数の増加に伴う攻撃成功率の変化を図 5 に示す．CLIP 分類器の結果は，ユーザスタディと比較して，汚染データ数が 750，1000 件で高い攻撃成功率を示したものの，全体的な傾向や言語間の攻撃成功率の大小は類似していた．また，両者の評価結果の一致率は，91.2%となり，CLIP 分類器による評価は，人間の目による評価と高い一致を示した．したがって，以降の実験では，CLIP 分類器の結果のみを報告する．

4.2　Stable Diffusion 2.0 における評価結果

Stable Diffusion 2.0 に対してデータ汚染攻撃を行った結果を図 6 に示す．英語のキャプションのみで構成した汚染データで攻撃しているにもかかわらず，非英語言語でも攻撃が成功した．また，いずれの言語においても，汚染データ数の増加に伴い攻撃成功率が上昇した．

英語では，汚染データ 500 件で，攻撃成功率は 32.7%にとどまったのに対し，非英語言語では，すべて 45%を超えた．特に，日本語とインドネシア語では，汚染データ 100 件の時点で 20%を超え，汚染データ数 250 件で 50%を超えた．その次に，トルコ語が高い攻撃成功率を示し，その他の 6 言語も英語より高い攻撃成功率となった．

生成画像例として，C_t = “dog”，C_d = “cat” としたときの攻撃後の Stable Diffusion 2.0 における “dog” の生成画像を図 7 に示す．ただし，すべての画像は，同じ Seed 値を設定して生成している．生成画像からも，日本語やインドネシア語では，英語より少ない汚染データ数で C_t = “dog” から C_d = “cat” への転移が確認された．

これらの結果から，英語によるデータ汚染攻撃を受けた Text-to-Image モデルは，非英語言語のプロンプトを入力しても攻撃が成功し，英語よりも攻撃成功率が高くなる．また，非英語言語のプロンプトでは，より少ない汚染データが混入するだけでも，攻撃が成功する．

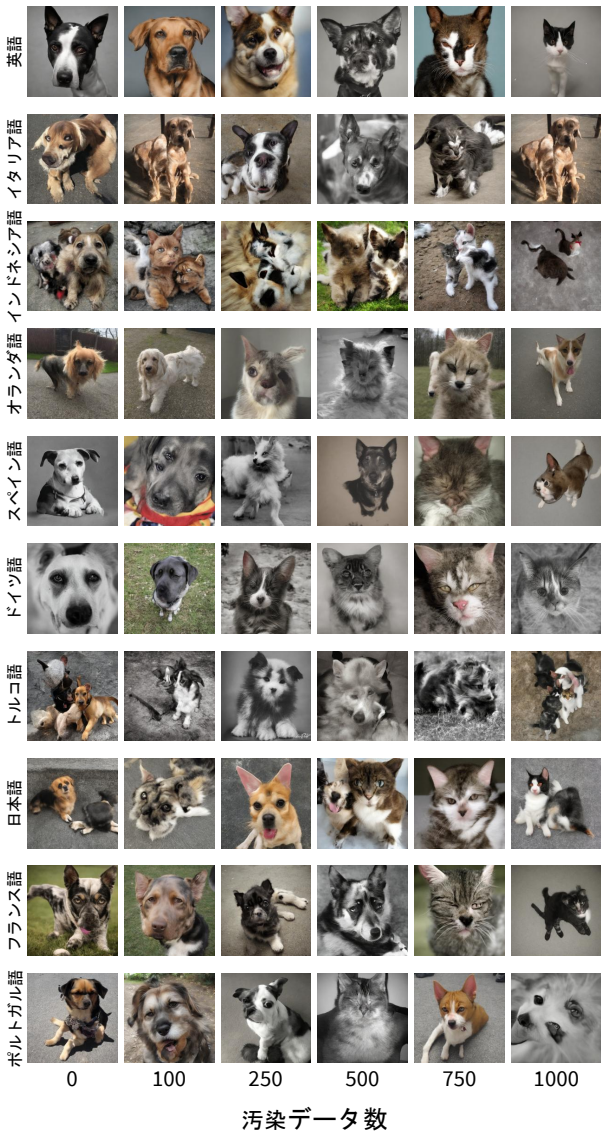

図 7 攻撃後の Stable Diffusion 2.0 で生成された “dog” の画像 (C_t = “dog”, C_d = “cat”)

4.3 Stable Diffusion XL における評価結果

Stable Diffusion XL に対してデータ汚染攻撃を行った結果を図 8 に示す．Stable Diffusion 2.0 と同様に，英語のキャプションのみで構成した汚染データで攻撃しているにもかかわらず，非英語言語でも攻撃が成功し，汚染データ数の増加に伴い攻撃成功率が上昇した．

英語では，汚染データ 500 件で，攻撃成功率は 30.3%となっているのに対し，日本語やインドネシア語，トルコ語では，少ない汚染データ数で高い攻撃成功率となった．一方で，スペイン語とポルトガル語は，Stable Diffusion 2.0 に比べ攻撃成功率が低下し，英語と同程度にとどまった．したがって，Stable Diffusion 2.0 では，すべての非英語言語が英語を上回る攻撃成功率が得られたのに対し，Stable Diffusion XL では，一部の言語で攻撃成功率の低下が確認された．

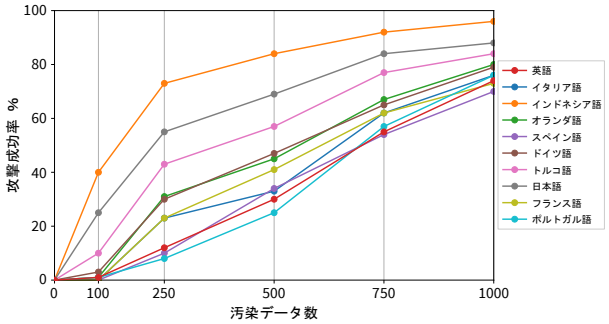

図 8 Stable Diffusion XL における汚染データ数と攻撃成功率の関係

5. 議論

5.1 言語間における攻撃成功率の差異の考察

4.2 節では Stable Diffusion 2.0，4.3 節では Stable Diffusion XL での評価結果を示した．その結果，非英語言語で攻撃成功率が高くなることを明らかにした．本節では，その要因について考察する．

言語間で違いが現れる要因として，テキストエンコーダが挙げられる．表 1 に示したように，Stable Diffusion 2.0 は OpenCLIP ViT-H，Stable Diffusion XL は CLIP ViT-L と OpenCLIP ViT-bigG という 2 種類のテキストエンコーダを用いている．両モデルで使用されている OpenCLIP の訓練には，英語のみのキャプションを含む LAION-2B-en が用いられている [17]．したがって，英語のテキストは適切な埋め込みベクトルで表現できる一方で，非英語言語では必ずしも正確な埋め込みベクトルが得られない可能性がある．

そこで，テキストエンコーダに，“a photo of C_t” を本研究で評価に用いた 10 言語で表現したプロンプトを入力し，得られた埋め込みベクトルを t-SNE により 2 次元に次元削減し，可視化を行った．Stable Diffusion 2.0 のテキストエンコーダにおける可視化の結果を図 9 に，Stable Diffusion XL のテキストエンコーダにおける可視化の結果を図 10 に示す．

可視化の結果，いずれのモデルにおいても，クラスごとにクラスタが形成されていた．一方で，各クラス内では言語ごとの埋め込みベクトルにばらつきが見られた．特に，日本語やインドネシア語，トルコ語の埋め込みベクトルは，他言語から大きく離れて配置されている．これらの言語が大きく分離した理由として，テキストエンコーダの訓練データに含まれる英語と，語彙や文法の類似性が相対的に低いことが考えられる．さらに，これらの言語は，実験において攻撃成功率が特に高くなった言語と一致している．以上の結果から，画像生成モデルのテキストエンコーダにおける多言語対応能力が，モデルのデータ汚染攻撃に対する頑健性に大きく影響する可能性が示唆される．

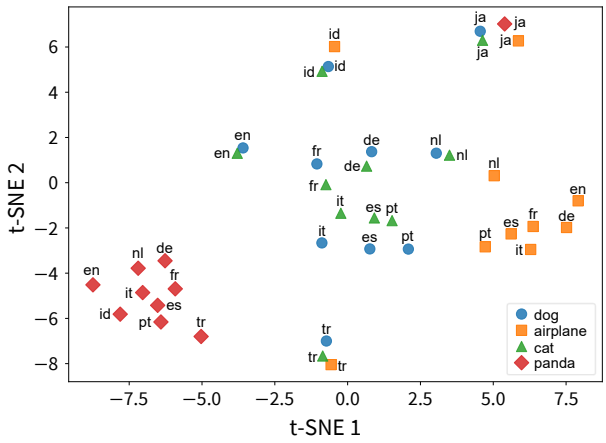

図 9 Stable Diffusion 2.0 のテキストエンコーダによる埋め込みベクトルの t-SNE 可視化

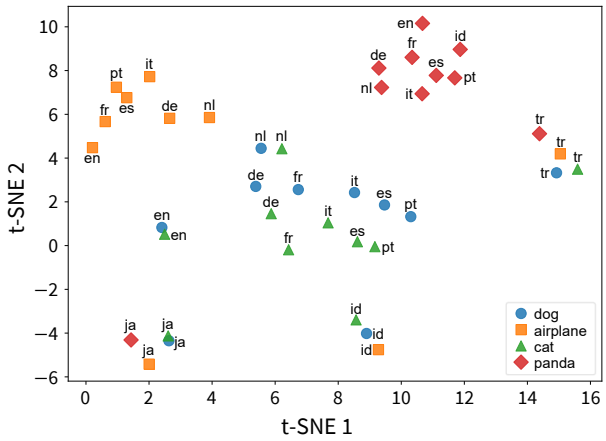

図 10 Stable Diffusion XL のテキストエンコーダによる埋め込みベクトルの t-SNE 可視化

5.2 非英語環境を考慮したデータ汚染攻撃への防御手法

本研究で明らかになった，英語の汚染データが特に非英語環境で高い攻撃成功率を示すという脆弱性を踏まえ，非英語環境でのデータ汚染攻撃に対する防御手法を考察する．

根本的な防御手法としては，汚染データの混入を防ぐ対策が挙げられる．例えば，画像とキャプションの関連性を埋め込みベクトル間のコサイン距離で評価し，閾値を超える不適合なデータを除去する手法が提案されている [18]．しかし，本研究で示した非英語環境での脆弱性は，この従来手法の限界を浮き彫りにする．汚染データは英語のキャプションで作成されているため，英語の文脈で画像とキャプションの類似性を検証するだけでは，非英語環境で顕在化する影響を十分に予測できない．そのため，多言語の観点を取り入れた新たな防御手法が求められる．

データセット側の対策をすり抜ける脅威を想定すると，モデル自体の頑健性を高めるアプローチも不可欠である．5.1 節で論じたように，英語中心のデータで訓練されたテキストエンコーダは，非英語言語に対して埋め込みベクトルが散逸的になりやすく，結果として低密度な領域に配置される．このような領域は，少量の汚染データによっても容易に埋め込み空間の構造が歪められ，攻撃者にとって制御しやすい不安定な領域を形成してしまうと考えられる．

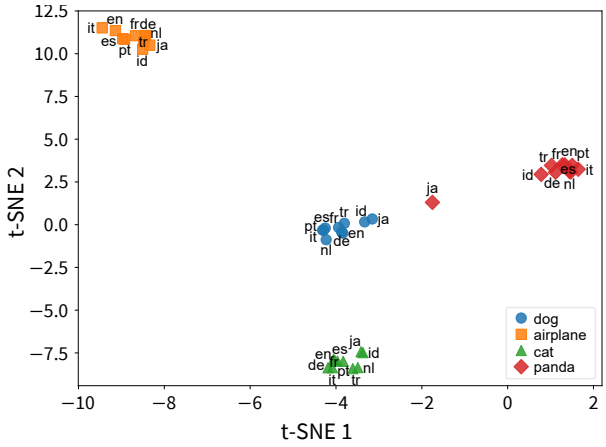

図 11 M-CLIP による埋め込みベクトルの t-SNE 可視化

この課題に対する解決策の一つが，多言語対応テキストエンコーダの利用である．その代表例が M-CLIP [19] があり，本研究で評価の対象とした 10 言語を含む 100 言語に対応する．M-CLIP は，複数言語間で意味的に整合した埋め込みベクトルを形成するため，言語間のばらつきを低減できると期待される．実際に，M-CLIP に “a photo of C_t” を 10 言語で表現したプロンプトを入力し，得られた埋め込みベクトルを t-SNE で可視化した結果を図 11 に示す．

日本語の “panda” のみが他言語から離れた位置に存在するものの，全体として同一クラス内での言語間のばらつきは大幅に縮小している．この性質により，言語に依存せず一貫した埋め込みベクトルが U-Net に伝達され，特定の言語においてのみ攻撃が顕著となる現象を抑制できると考えられる．しかし，この手法は，非英語言語での攻撃成功率を低減する有効な防御策であるが，データ汚染攻撃そのものの根本的な解決には至らない．

したがって，実用的な防御手法としては，データセット側でフィルタリングや異常検知によって，汚染データの混入を防ぐ仕組みと，多言語対応テキストエンコーダによって非英語環境での攻撃顕在化を抑制する仕組みを組み合わせた多層防御のアプローチが望ましい．

5.3 他攻撃シナリオおよび他モデルでの検証

本研究では，dirty-label 攻撃を実装し，事前学習済みモデルをファインチューニングする際に汚染データが混入するシナリオを想定して攻撃評価を行った．一方，Shan ら [6] は，clean-label 攻撃やモデルをゼロから訓練するシナリオでも英語での攻撃評価を行い，攻撃が成功することを示している．したがって，これらのシナリオにおける非英語言語での攻撃評価は，今後の重要な課題である．

さらに，本研究では Stable Diffusion 2.0 および Stable Diffusion XL を対象としたが，両モデルは基本的なアーキ

テクチャが共通しているため，異なるアーキテクチャを採用するモデルでの検証も必要である．例えば，Stable Diffusion 3 や FLUX.1 [20] は，U-Net の代わりに，Diffusion Transformer という新たなアーキテクチャを導入している．こうした異なるアーキテクチャのモデルに対しても検証を行うことで，本研究で得られた知見の一般性を確認できる．

5.4 研究倫理

本研究は，オープンソースの画像生成モデルである Stable Diffusion を使用し，仮想的なシナリオでデータ汚染攻撃の評価を行った．実在する製品やサービスを対象とせず，第三者による悪用を最小限に抑えている．また，汚染データの作成および評価は，公開データセットに基づいており，個人情報や機微情報は一切扱っていない．さらに，本研究では，攻撃評価の一環としてユーザスタディを実施した．ユーザスタディは，参加者に対して実験の目的や内容を事前に説明し，参加者の同意を得たうえで実施した．参加者本人に利害関係が生じることはなく，個人を特定できる情報は一切収集していない．本研究の成果は，攻撃手法の効果を示すとともに，防御手法の議論を目的としたものであり，攻撃の実行を助長する意図はない．本研究で得られた知見が，AI の安全性向上に貢献することを期待する．

6. まとめ

本研究では，Text-to-Image モデルに対するデータ汚染攻撃の影響を非英語環境下で評価した．既存研究が英語に限定されていたのに対し，本研究は Stable Diffusion 2.0 および Stable Diffusion XL を対象に，10 言語で攻撃評価を行った．その結果，非英語言語において，攻撃成功率が英語よりも高く，少量の汚染データで攻撃が成功することを確認した．また，テキストエンコーダが出力する埋め込みベクトルの可視化を通じて，非英語言語では同一クラス内の埋め込みベクトルが多言語と分離しており，これが攻撃成功率の高さに寄与していることを示唆した．以上の結果から，モデルの安全性を評価する際には，非英語環境での検証が不可欠である．今後は，多言語対応テキストエンコーダを用いた対策手法の評価および他の攻撃シナリオや異なるアーキテクチャのモデルに対する検証が必要である．

謝辞 この成果の一部は，NEDO（国立研究開発法人新エネルギー・産業技術総合開発機構）の委託業務（JPNP24003）の結果得られたものです．

参考文献

[1] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B.: High-resolution image synthesis with latent diffusion models, *CVPR*, pp. 10684–10695 (2022).
[2] : Midjourney, `https://www.midjourney.com`.
[3] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M. and Sutskever, I.: Zero-shot text-to-image generation, *ICML*, pp. 8821–8831 (2021).
[4] Wu, Y., Yu, N., Backes, M., Shen, Y. and Zhang, Y.: On the Proactive Generation of Unsafe Images From Text-To-Image Models Using Benign Prompts, *USENIX Security* (2025).
[5] Huang, Y., Juefei-Xu, F., Guo, Q., Zhang, J., Wu, Y., Hu, M., Li, T., Pu, G. and Liu, Y.: Personalization as a shortcut for few-shot backdoor attack against text-to-image diffusion models, *AAAI*, Vol. 38, No. 19, pp. 21169–21178 (2024).
[6] Shan, S., Ding, W., Passananti, J., Wu, S., Zheng, H. and Zhao, B. Y.: Nightshade: Prompt-specific poisoning attacks on text-to-image generative models, *IEEE S&P*, pp. 807–825 (2024).
[7] Ding, W., Li, C. Y., Shan, S., Zhao, B. Y. and Zheng, H.: Understanding Implosion in Text-to-Image Generative Models, *ACM CCS*, pp. 1211–1225 (2024).
[8] Saxon, M. and Wang, W. Y.: Multilingual Conceptual Coverage in Text-to-Image Models, *ACL*, pp. 4831–4848 (2023).
[9] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M. et al.: Laion-5b: An open large-scale dataset for training next generation image-text models, *NeurIPS*, Vol. 35, pp. 25278–25294 (2022).
[10] Carlini, N., Jagielski, M., Choquette-Choo, C. A., Paleka, D., Pearce, W., Anderson, H., Terzis, A., Thomas, K. and Tramèr, F.: Poisoning web-scale training datasets is practical, *IEEE S&P*, pp. 407–425 (2024).
[11] Shen, L., Tan, W., Chen, S., Chen, Y., Zhang, J., Xu, H., Zheng, B., Koehn, P. and Khashabi, D.: The Language Barrier: Dissecting Safety Challenges of LLMs in Multilingual Contexts, *Findings of ACL*, pp. 2668–2680 (2024).
[12] Upadhayay, B. and Behzadan, V.: Sandwich attack: Multi-language Mixture Adaptive Attack on LLMs, *TrustNLP*, pp. 208–226 (2024).
[13] Schuhmann, C.: LAION-AESTHETICS, `https://laion.ai/blog/laion-aesthetics/` (2022).
[14] Li, J., Li, D., Xiong, C. and Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, *ICML*, pp. 12888–12900 (2022).
[15] : sd-scripts, `https://github.com/kohya-ss/sd-scripts` (2024).
[16] W3Techs: Usage statistics of content languages for websites, `https://w3techs.com/technologies/overview/content_language` (2025).
[17] Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L. and Jitsev, J.: Reproducible scaling laws for contrastive language-image learning, *CVPR*, pp. 2818–2829 (2023).
[18] Yang, Z., He, X., Li, Z., Backes, M., Humbert, M., Berrang, P. and Zhang, Y.: Data poisoning attacks against multimodal encoders, *ICML*, pp. 39299–39313 (2023).
[19] Carlsson, F., Eisen, P., Rekathati, F. and Sahlgren, M.: Cross-lingual and multilingual clip, *LREC*, pp. 6848–6854 (2022).
[20] Labs, B. F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P. et al.: FLUX. 1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space, *arXiv preprint arXiv:2506.15742* (2025).