Poster: Why Do Non-English Languages Exhibit Higher Vulnerability to Data Poisoning Attacks Against Text-to-Image Models?

Ryohei Kakebayashi*, Tatsuya Mori*†‡
*Waseda University †NICT ‡RIKEN AIP

***Abstract*—Text-to-image (T2I) models, despite their growing deployment, are susceptible to data poisoning attacks due to their reliance on web-scale datasets. While existing studies have primarily focused on English as the target language, T2I models are used worldwide across diverse languages, and it remains unclear whether these English-based findings generalize to other languages. Moreover, their text encoders are trained predominantly on English data, which may lead to disparate security characteristics across languages. In this paper, we evaluate data poisoning attacks across 10 languages. Our results demonstrate that certain non-English languages, particularly those linguistically distant from English (e.g., Indonesian and Japanese), exhibit notably higher vulnerability to these attacks than English. We attribute this vulnerability to misaligned text embeddings produced by English-centric text encoders. Finally, we demonstrate that fine-tuning the text encoder on non-English data effectively mitigates these attacks. These findings reveal underexplored multilingual vulnerabilities of T2I models.**

I. Introduction

Text-to-image (T2I) models enable users to generate high-quality images from natural-language prompts. However, these models raise serious safety concerns. In particular, data poisoning is a prominent threat: injecting misaligned image–text pairs into the training data can manipulate model behavior.

Recent studies show that even a small number of poisoned samples can effectively compromise T2I models [1]. A critical limitation of existing studies, however, is their exclusive focus on English. This overlooks the fact that T2I models are deployed globally and process prompts in diverse languages [2]. To understand the true extent of these vulnerabilities, evaluating data poisoning in multilingual environments is essential.

In this paper, we extend prior English-only evaluations to a multilingual setting across 10 languages (see Fig. 1). We show that non-English languages are more susceptible, especially when they are linguistically distant from English, such as Indonesian and Japanese, achieving high attack success rates with fewer poisoned samples. Finally, we identify misaligned text embeddings in English-centric text encoders as a key factor and demonstrate that fine-tuning the text encoder on non-English data effectively mitigates these attacks.

II. Background

T2I models are susceptible to data poisoning attacks because their training relies on web-scale datasets, making manual verification impractical. Recent studies demonstrate that adversaries can manipulate specific concepts, degrade model utility [1], and inject harmful biases [3] through poisoned data. However, a key limitation of existing studies is their English-only focus: poisoned data is constructed with English text, and attacks are evaluated using English prompts. Using English for poisoning is reasonable, as most training data for T2I models is in English. Nevertheless, evaluating attacks solely in English is insufficient, given that T2I models are used by speakers of various languages worldwide and non-English prompts can generate images of comparable quality [2]. These observations imply that the attack surface extends beyond English, and English-only evaluations may underestimate real-world risk.

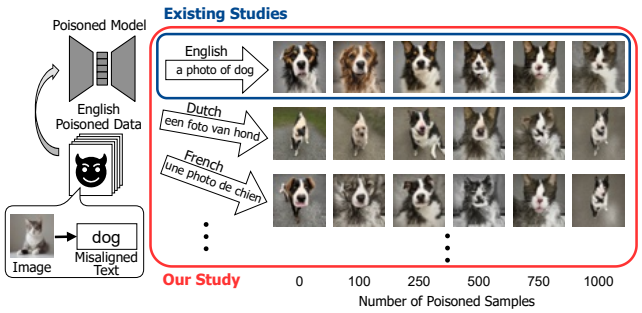


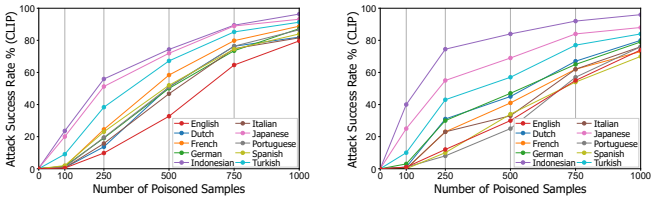
Fig. 1: Comparison of existing English-only evaluations and our multilingual evaluation approach.

Similar findings in large language models (LLMs) further support this concern: safety mechanisms often fail in non-English languages due to training data disparities [4]. These results motivate our multilingual evaluation of T2I models.

III. Evaluation and Analysis

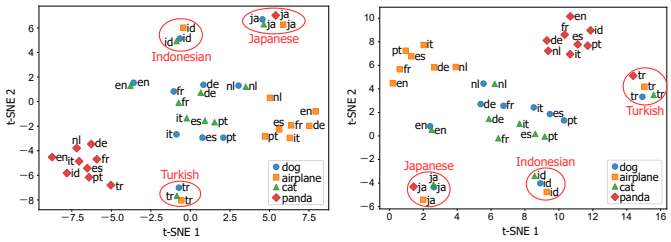
A. Experimental Setup

We evaluate data poisoning attacks in a fine-tuning scenario, where a pre-trained T2I model is adapted using additional datasets. The attacker injects poisoned data whose text describes a target class C_t while the image depicts a destination class C_d, so that prompts for C_t yield images of C_d. We employ four class pairs (C_t, C_d) and two widely used models, Stable Diffusion v2 (SD-v2) and Stable Diffusion XL (SDXL). Following prior work [1], we fine-tune the U-Net, a network that iteratively refines a noisy representation, on clean and poisoned data while varying the number of poisoned samples from 0 to 1,000. While poisoned text is written in English, which reflects the dominance of English in web data, we



(a) SD-v2. (b) SDXL.

Fig. 2: Attack success rates across 10 languages.



(a) SD-v2 text encoder. (b) SDXL text encoder.

Fig. 3: t-SNE visualization of text embeddings.

translate a fixed set of English evaluation prompts into 10 languages: Dutch, English, French, German, Indonesian, Italian, Japanese, Portuguese, Spanish, and Turkish. We generate 1,000 images per condition and measure the attack success rate (ASR) using a zero-shot CLIP classifier, validated against human inspection (91.2% agreement). Note that we exclude language–class pairs that the clean models fail to generate.

B. Attack Success Rates on Stable Diffusion

Fig. 2 shows a cross-lingual vulnerability: English poisoned data compromises non-English prompts. For SD-v2, all non-English languages exhibit higher ASRs than English. Indonesian and Japanese are particularly susceptible (ASR $> 50\%$ with 250 poisoned samples). While SDXL shows more variance (e.g., Portuguese and Spanish are closer to English), Indonesian, Japanese, and Turkish remain highly vulnerable.

C. Factors Influencing Vulnerability

To understand the disparity in vulnerability, we analyze correlations using ASR on SD-v2 with 500 poisoned samples. First, we test whether the number of pre-training samples per language explains robustness; despite the English-centric training data, we find no significant correlation with ASR ($r = -0.388, p = 0.303$). Second, we compute linguistic distance from English using lang2vec [5], which extracts vector representations based on linguistic databases. We find that an average distance of syntactic, geographic, and genetic features strongly correlates with ASR ($r = 0.858, p = 0.003$). These features refer respectively to properties such as word order, speaking regions, and phylogenetic relatedness. This indicates that languages that are more distant from English, such as Indonesian and Japanese, tend to exhibit higher vulnerability.

D. Root Cause: Text Embedding Misalignment

Our factor analysis points to the text encoder as the likely source. Both models rely on CLIP-based text encoders trained on English-centric datasets, which may yield poorly aligned text embeddings for linguistically distant languages. We encode the multilingual evaluation prompts and visualize their embeddings with t-SNE. Ideally, prompts expressing the same concept should map nearby regardless of language. However, as shown in Fig. 3, text embeddings for vulnerable languages (i.e., Indonesian, Japanese, and Turkish) form isolated clusters rather than aligning with other languages. This misalignment may prevent the model from learning robust features for these languages, making it susceptible to manipulation.

E. Mitigation via Text Encoder Fine-Tuning

Our findings suggest that correcting text embeddings can mitigate the attacks. We therefore fine-tune the text encoder. Specifically, we fine-tune SD-v2 on one million samples each for Indonesian and Japanese under three configurations: (i) training only the U-Net, (ii) training only the text encoder, and (iii) training both. We then re-run poisoning attacks. These results indicate that text encoder fine-tuning is most effective: with 100 poisoned samples, ASR drops from 29.8% to 0.8% for Indonesian and from 24.9% to 1.2% for Japanese, comparable to English (0.8%). In contrast, U-Net fine-tuning is less effective (e.g., 15.6% ASR for Indonesian), and fine-tuning both yields marginal gains over text encoder fine-tuning. We also confirm that English image quality is well preserved (FID 13.3 to 15.3; CLIP Score 0.318 to 0.313).

IV. Conclusion

We conduct the first multilingual evaluation of data poisoning attacks against T2I models across 10 languages. Our analysis reveals that non-English languages, particularly those linguistically distant from English, are more susceptible due to misaligned text embeddings. We further demonstrate that text encoder fine-tuning serves as an effective mitigation. However, since this mitigation requires language-specific fine-tuning of the text encoder, future work will focus on developing language-independent defenses to ensure universal robustness.

Acknowledgment

This paper is based on results obtained from a project, JPNP24003, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

[1] Shawn Shan et al. “Nightshade: Prompt-specific poisoning attacks on text-to-image generative models”. In: *IEEE S&P*. 2024, pp. 807–825.

[2] Michael Saxon and William Yang Wang. “Multilingual Conceptual Coverage in Text-to-Image Models”. In: *Proc. of ACL*. 2023, pp. 4831–4848.

[3] Ali Naseh et al. “Backdooring Bias (B2) into Stable Diffusion Models”. In: *USENIX Security*. 2025, pp. 977–996.

[4] Lingfeng Shen et al. “The Language Barrier: Dissecting Safety Challenges of LLMs in Multilingual Contexts”. In: *Findings of ACL*. 2024, pp. 2668–2680.

[5] Patrick Littell et al. “Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors”. In: *Proc. of EACL*. 2017, pp. 8–14.

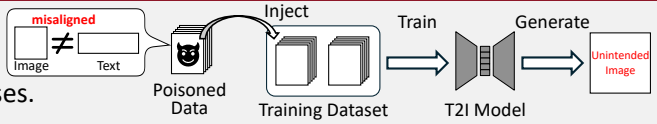
Poster: Why Do Non-English Languages Exhibit Higher Vulnerability to Data Poisoning Attacks Against Text-to-Image Models?

Ryohei Kakebayashi[1], Tatsuya Mori[1,2,3]
[1]Waseda University [2]NICT [3]RIKEN AIP

1. Background

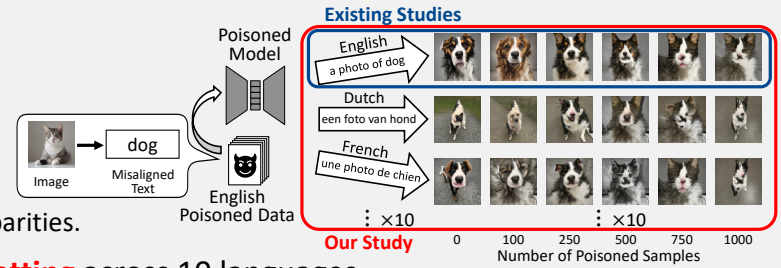
Data Poisoning Attacks

- Inject misaligned image–text pairs into the training dataset.
- Can manipulate specific concepts, degrade model utility, or inject harmful biases.



Limitations of Existing Studies

- Evaluating attacks solely in **English** is **insufficient**.
  - T2I models are used by speakers of various languages worldwide.
  - Non-English prompts can generate images of comparable quality.
- English-only evaluations may underestimate real-world risk.



Multilingual Evaluation in LLMs

- Safety mechanisms often fail in non-English languages due to data disparities.

We extend **prior English-only evaluations** to a **multilingual setting** across 10 languages.

2. Experimental Setup

Model: Stable Diffusion v2 (SD-v2), Stable Diffusion XL (SDXL)

Training Scenario: Fine-tuning (standard approach in prior work)

Evaluation Metrics: CLIP classifier

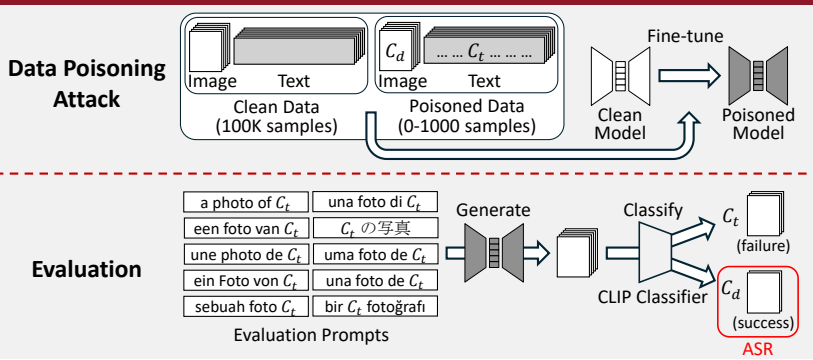
- Validated against human inspection (91.2% agreement)

Evaluated Languages: 10 languages

- Dutch, English, French, German, Indonesian, Italian, Japanese, Portuguese, Spanish, Turkish

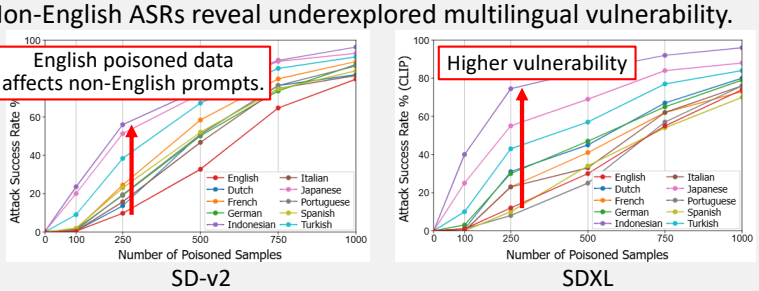
Attacker: Aims to generate C_d images with C_t prompts.

(C_t, C_d) **pairs:** (dog, cat), (airplane, car), (cat, horse), (panda, sheep)



3. ASR Across 10 Languages

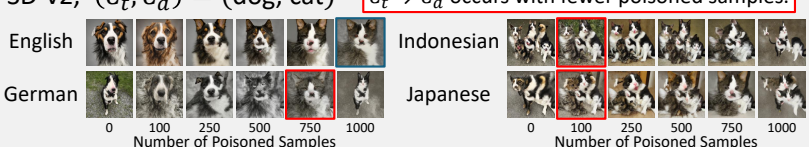
Non-English ASRs reveal underexplored multilingual vulnerability.



SD-v2 SDXL

4. Examples of Generated Images

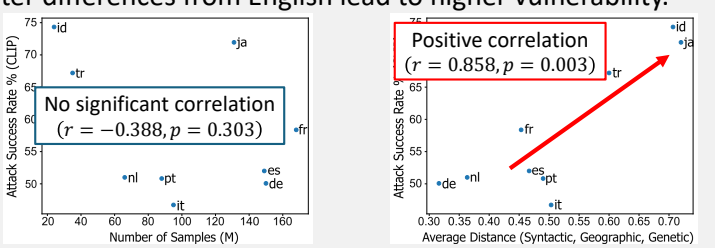
SD-v2, $(C_t, C_d) = (\text{dog}, \text{cat})$ $C_t \rightarrow C_d$ occurs with fewer poisoned samples.



(All images are generated with a fixed random seed.)

5. Factors Influencing Vulnerability

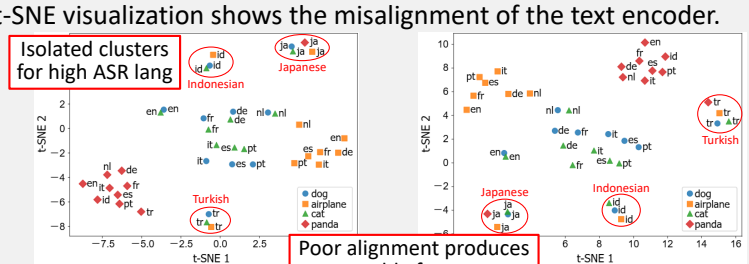
Greater differences from English lead to higher vulnerability.



Data size Linguistic distance from English

6. Root Cause: Text Embedding Misalignment

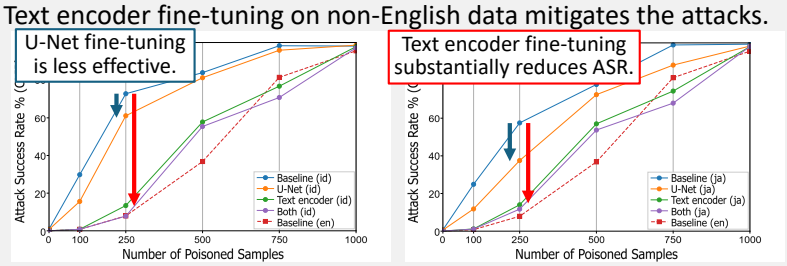
t-SNE visualization shows the misalignment of the text encoder.



SD-v2 SDXL

7. Mitigation via Text Encoder Fine-Tuning

Text encoder fine-tuning on non-English data mitigates the attacks.



Indonesian Japanese

8. Conclusion

- **Higher Vulnerability:** Non-English languages are more susceptible to data poisoning attacks than English.
- **Difference from English:** Languages that are linguistically distant from English exhibit notably higher vulnerability.
- **Misalignment:** Misaligned embeddings induce vulnerability.

Acknowledgment This paper is based on results obtained from a project, JPNP24003, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).