Leaky Diffusion: Attribute Leakage in Text-Guided Image Generation

Anastasios Lepipas
Imperial College London
a.lepipas20@imperial.ac.uk

Marios Charalambides
Imperial College London
marios.charalambides22@imperial.ac.uk

Jiani Liu
Imperial College London
jiani.liu23@imperial.ac.uk

Yiying Guan
Imperial College London
yiying.guan23@imperial.ac.uk

Dominika C Woszczyk
Imperial College London
d.woszczyk19@imperial.ac.uk

Mansi
Imperial College London
m.-24@imperial.ac.uk

Thanh Hai Le
Imperial College London
h.le24@imperial.ac.uk

Soteris Demetriou
Imperial College London
s.demetriou@imperial.ac.uk

Abstract

Text–guided diffusion models can be used to generate photorealistic images conditioned on natural language instructions. Due to their ease of use, millions of users already leverage them to generate and populate images online. In this work, we reveal the risk of attribute (authorship and dementia) leakage from such models. Existing authorship and dementia inferences rely primarily on text. We show that instructions are a new form of text that can reveal these attributes. More surprisingly, and in contrast to prior work, we show that those attributes can be transferred and leaked from images generated with diffusion models. In particular, we construct image and multi–modal adversarial models which leverage image data augmentation and text–image embedding models to achieve state of the art performance in spear authorship inference (up to 0.877% Top–5 accuracy for 100 authors), while dementia inference is possible even from the output images alone (0.75% accuracy on the ADReSS dataset). Our rigorous evaluation shows that such inferences remain robust using different training sets, and when trained in classifier-independent ways, and against SOTA mitigations such paraphrasing Transformer models and LLMs.

Keywords

Authorship Attribution, Text–To–Image Models, Embeddings

1 Introduction

Diffusion models are generative models that have transformed image synthesis. By adding noise to training data and learning to reverse the process, they can produce images virtually indistinguishable from real ones [77]. When guided by text, these models are known as *text-to-image* (**T2I**) models. T2I models enable users to generate images using *natural language instructions* (**NLIs**). Popular models such as *Stable Diffusion*, *Latent Diffusion*, and *DALL-E* are widely accessible, with millions of users leveraging them to create tens of millions of images [48].

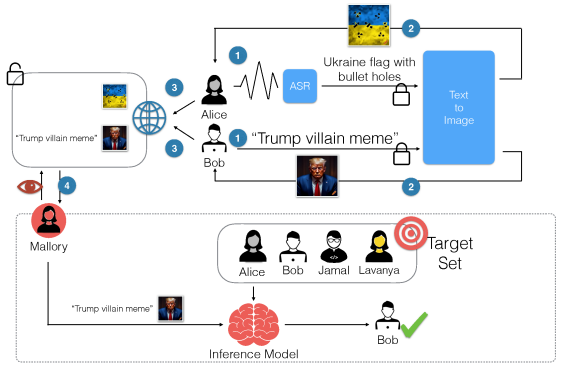


Figure 1: Privacy leakage in Text–to–Image model usage. Users create natural language instructions which are used to guide the generation of images. Users share their instructions and/or the generated images with a false expectation of privacy online and by default with online model providers. Mallory who observes the shared data, aims to infer sensitive attributes about the data owner. This figure depicts the scenario of authorship inference targeting four victim users. Disclaimer: the images at Step 2 were generated solely for the purpose of illustration using *craiyon* [18] which was available for free online and anonymous usage.

The growing prevalence of T2I models raises privacy concerns. To illustrate potential privacy risks, consider the scenario depicted in Figure 1. Alice and Bob, concerned about recent political events, wish to express their views on online forums but fear potential backlash. To remain anonymous, they use a T2I model to generate relevant images, either by running an open-source model offline (e.g., Stable Diffusion) or using online tools without registration (e.g., Craiyon [18]). The model aims to align the generated image with the given instruction in a shared latent embedding space. Both the textual descriptions (NLIs) and generated images are accessible to the model provider. Alice and Bob then anonymously share their generated content on a public forum [48].

Mallory has access to publicly shared data. If Mallory can infer the author of an NLI, this poses serious privacy risks. Authorship leakage can compromise user privacy, especially when anonymity is essential. For example, in some countries, online political content is heavily censored, with artists facing prosecution or threats [11, 28].

Proceedings on Privacy Enhancing Technologies 2025(4), 275–292

https://doi.org/10.56553/popets-2025-0130

Dissident artists may rely on T2I-generated images to express their views anonymously. Similarly, online discussion groups frequently use AI–generated memes to critique political ideologies [101]. To minimize potential backlash or draw less attention to themselves, some users appear to participate using aliases or pseudonyms. Our analysis found that fewer than 1% of users on the Stable Diffusion, Midjourney, DALL–E, and DALL–E 2 subreddits use real names (see Appendix D), suggesting a preference for anonymity. T2I usage should not expose user identities. If an entity can infer user attributes from NLI inputs, this violates user expectations and might conflict with privacy regulations, including GDPR (Art. 4(1) and Recital 30) and UK GDPR, and the California Consumer Privacy Act (CCPA) (Cal. Civ. Code §1798.140(v)(1)(K) and Cal. Civ. Code §1798.140(o)). Such inferences pose a significant privacy threat.

Malicious actors can also exploit T2I models to spread misinformation or distribute toxic and unsafe content online [72, 82]. Such illicit content is often shared with an expectation of anonymity. Authorship inference can help attribute unsafe content to its source, complementing safety filters implemented by model providers. While safety filters can be disabled or bypassed [10, 109], authorship attribution provides an additional layer of accountability. This approach can assist regulators and law enforcement in identifying and prosecuting offenders.

To the best of our knowledge, attribute leakage in text-guided diffusion-based image generation has not been studied. This work aims to bridge that gap. Previous research has explored authorship inference in other domains, such as book writing and Wikipedia article contributions [92, 93]. However, these studies focus solely on text, whereas T2I models may leak sensitive information from both the input text and the generated images.

Some inferences from images are possible [13, 63]. However, existing studies do not address text-guided image generation and cannot be used for authorship inference or analyzing the relationship between leaked attributes and guiding NLIs. Attribute leakage in T2I models presents additional challenges. These models employ a stochastic Markov chain process, which can obscure text input information. Furthermore, leakage in T2I models is inherently multi-modal, as inferences may arise from both the input text and the generated image. The only existing work on T2I attribution focuses on identifying the source T2I *model* in a closed-world setting with four target models [85]. Our study instead investigates the potential for *user* attribute leakage. Furthermore, we examine an understudied attribute that affects a vulnerable population.

Our Approach. In this work, we address these challenges and investigate the novel research question: *Do text-guided diffusion models for image generation leak user attributes?* Motivated by our guiding examples, we conduct a thorough analysis of authorship leakage through a set of targeted attacks, which we term *spear authorship inference* attacks. To further demonstrate attribute leakage, we examine pathological language characteristics associated with dementia, a neurocognitive condition, and show that text-guided diffusion models can encode such highly sensitive traits.

To address our overarching question, we design an attack strategy based on a hierarchical black-box adversary. The adversary operates at different levels of information access and can succeed using only black-box observations of T2I usage, making our threat model both more realistic and more robust than a white-box alternative. Specifically, the adversary employs three inference strategies: *Input Inference* ($\mathcal{A}_i$), which has access only to the input natural language instruction; *Output Inference* ($\mathcal{A}_o$), which has access only to the model's output image; and *Multi-Modal Input-Output Inference* ($\mathcal{A}_{io}$), which leverages both the input text and the generated image.

Our approach to mitigating the effects of stochastic generation is based on two key insights. First, T2I models utilize embedding models (e.g., CLIP [73]) that learn a multi-modal embedding space by jointly training on images and their textual descriptions. Extracting image embeddings using this approach should produce representations closely aligned with their NLIs, increasing the likelihood of retaining input information. Second, the denoising process in T2I models introduces creative variability, meaning the same NLI can generate different images each time unless explicitly controlled. We leverage this by augmenting our image training set, demonstrating that this not only enhances model generalization and performance, but also enables highly accurate inferences on images generated by previously unseen T2I models. Our extensive evaluation confirms that *spear authorship inferences* are feasible from both the input and output of T2I models, with text-image alignment embeddings significantly improving inference performance. Additionally, our models remain robust in both closed and open–world settings and against state-of-the-art obfuscation techniques. Finally, we show that pathological language markers present in input text persist through the diffusion process via text-image embeddings, highlighting the risk of leaking sensitive neurocognitive conditions, such as dementia, from generated images. Our authorship inference models and examples of *NSFW* images are available on GitHub [45].

Contributions. We summarize our key contributions as follows:

- **Novel Application Domain.** To the best of our knowledge, we are the first to study attribute leakage in T2I models. We hope our work inspires further research on T2I leakage of novel attributes and the development of effective mitigation strategies.
- **New T2I Black–Box Inference Method.** We propose a novel framework for inferring user attributes in T2I model usage through hierarchical black-box access, leveraging Text-based, Image-based, and Multi-Modal Inference Models built on text-image alignment embeddings.
- **New Findings.** We demonstrate that T2I images leak authorship and can encode dementia language markers, enabling dementia inferences from both text input and generated images. Our best multi-modal classifier achieved over 79% accuracy on the ADReSS dataset.
- **Rigorous Evaluation.** We conduct a rigorous evaluation, demonstrating that our attacks achieve state-of-the-art performance even with limited training samples. Our approach remains robust against unseen T2I models, open-world applications, and state-of-the-art mitigation strategies.

2 Preliminaries

Pre–trained Language Models. Pre–trained language models learn contextual understanding by predicting the next words in the input text. They also exhibit adaptability to various natural language processing (NLP) tasks through fine-tuning. The Transformer

architecture [100] enables these models to incorporate multiple layers with substantial capacity. Models such as GPT-4V [2], Mistral, Llama 3, and BERT [22] have achieved significant performance gains in downstream NLP tasks compared to previous state-of-the-art approaches. We leverage open-source LLMs to explore potential defenses against T2I authorship inferences. Additionally, these models can extract sentence embeddings [36, 92], which our adversarial models utilize to encode NLIs.

Embedding Models. Embedding models convert raw input, typically discrete, into low-dimensional vectors that capture the semantic meaning of the original data. These embeddings are widely used in downstream tasks such as classification and retrieval. Several models embed text into vector space, including GloVe [71], Dual Encoders [54], and other variations [36]. More recently, models have emerged that learn joint representations of text and images. Models such as CLIP [73], ViT [25], and BLIP-2 [49] train text and image encoders jointly, embedding both modalities into a shared vector space where image embeddings are positioned near their corresponding textual descriptions. For example, CLIP employs an image encoder $f(x)$ and a caption encoder $g(c)$, training on image-caption pairs $\{x, c\}$. The model optimizes a cross-entropy loss that encourages a high dot product $f(x) \cdot g(c)$ for matching pairs and a low dot product for mismatched pairs. These models have achieved remarkable success in tasks such as image captioning, multi-modal LLMs like GPT-4V [2], and guiding T2I models to generate images closely aligned with their NLI inputs. To the best of our knowledge, we are the first to leverage multi-modal representations to combine NLIs with T2I images for conducting attribute inference attacks.

Text-to-Image Generation Models. Text-to-image (T2I) models have recently gained widespread popularity for their ability to generate high-quality synthetic images efficiently. Early T2I generation methods [75, 111] relied on GAN networks [30]. However, the advent of diffusion models [77, 78] has significantly enhanced T2I generation, surpassing previous GAN-based architectures in quality and performance.

A diffusion model consists of two stages: a forward process and a backward process. During the forward process, an image $X_o \sim q(X_o)$ sampled from a data distribution is progressively noised over $t \in \{1, \ldots, T\}$ steps, forming a Markov chain of latent variables $x_1, \ldots, x_T$:

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathcal{I})$$

where α_t controls the amount of noise added at each step. It has been shown that x_T can be approximated by $\mathcal{N}(0, \mathcal{I})$ [68].

The backward process aims to reverse this transformation by gradually denoising $x_T \sim \mathcal{N}(0, \mathcal{I})$ through a sequence of steps $x_{T-1}, x_{T-2}, \ldots, x_0$. A model ϵ_θ (where θ represents the model parameters) can be trained to predict the added noise by minimizing the difference between the predicted and actual noise, often using a mean-squared error loss.

The generation process can be guided or controlled. Latent and Stable Diffusion models are based on the Latent Diffusion Model (LDM), which denoises samples in the latent space using a U-Net network to predict the noise at each step. The U-Net architecture includes self-attention layers for processing image latents and cross-attention layers for integrating text embeddings. A pre–trained text encoder, such as CLIP, converts the input text instruction into a latent embedding, which is then applied to the U-Net's cross-attention layers to guide the image generation process.

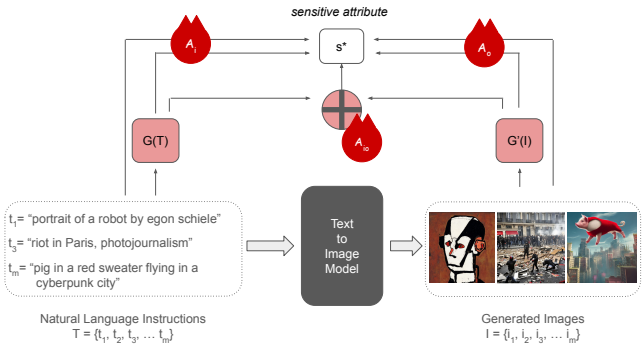


Figure 2: High–level taxonomy of attribute (s^*) inference attacks against text–to–image models. We assume a hierarchical adversary with access to different information of the text–to–image model: (1) $\mathcal{A}_i$ infers the attribute from the input prompts P only, (2) $\mathcal{A}_o$ infers the attribute from the output images I and (3) $\mathcal{A}_{io}$ combines both the input and the output of the target model for inference.

In our work, we utilize Stable Diffusion (Versions 1 and 2.1) [78] and Latent Diffusion [77], two widely recognized and publicly available text-guided image generation models. Stable Diffusion V1 was pre-trained on 256 × 256 images from LAION-2B-en [108] and later fine-tuned on 512 × 512 images from a subset of LAION-5B [83], using CLIP ViT-L/14 [69] as its text encoder. In contrast, Stable Diffusion V2.1 employs OpenCLIP-ViT/H [70] as a text encoder, which was trained on a larger and more diverse dataset. Latent Diffusion [77] (LD) was pre-trained on a subset of LAION-5B and utilizes the CLIP [69] text encoder to guide the generation of synthetic images.

3 Black-Box Attribute Inference Attacks

3.1 Threat Model and Framework Overview

Threat Model. Users of T2I models frequently share their prompts, generated images, or both online. For example, the Lexica website [48] hosts a vast collection of Stable Diffusion-generated images alongside their corresponding prompts and offers an API for retrieving prompts or images based on specified criteria. Additionally, users share such content in online communities, including the "Unstable Diffusion" group, which distributes unsafe images generated with Stable Diffusion. A common but false assumption is that sharing this data ensures anonymity, as there is little expectation that prompts—or especially synthetic images—can reveal authorship information. In this work, we investigate the following overarching questions: *Do text-to-image model inputs contain identifiable characteristics of the prompt's author? Do these characteristics transfer to the generated outputs? And can an adversary with access to raw prompts or images infer the author's identity with some probability?*

Our threat model is illustrated in Figure 2. More formally, we consider an adversary $\mathcal{A}$ with black-box access to text-to-image (T2I) models. Black-box attacks pose a greater challenge than white-box attacks, as the adversary lacks access to the model's internal

structure or computations. However, this approach is more realistic, has broader implications, and is harder to defend against, as it can be executed not only by T2I model providers but by anyone who can observe the model's inputs and/or outputs.

To formulate the adversary's goal, we adopt the general definition of attribute inference attacks outlined by Song et al. [92]. We assume that the adversary $\mathcal{A}$ has access to an auxiliary dataset D_{aux} containing labeled data in the form of (x, s), where x belongs to the input domain X ($x \in X$) and s represents a discrete attribute ($s \in S$), with S being the set of all possible attribute classes. The goal of $\mathcal{A}$ is to infer a sensitive attribute $s^* \in S$ given a seemingly innocuous and previously unseen input $x^* \in X$. To achieve this, the adversary aims to learn an adversarial function f on D_{aux} that maps the input domain X to the discrete attribute space S. Formally, this is defined as $f_\theta : X \to S$, where θ represents the learnable parameters of the function f. Furthermore, we consider an extended scenario where the adversary can infer authorship not only from the raw input but also from an embedding function G, such that $g_\phi : G(X) \to S$, where ϕ are the learnable parameters of G.

For clarity of presentation, we will first focus our analysis on *spear authorship inferences*. In Section 5.6, we will extend this analysis to sensitive pathological language markers, demonstrating how such markers can also be encoded in both the input and output of T2I models. More concretely, a *spear authorship inference* adversary $\mathcal{A}$ aims to identify, with some probability, the author $\alpha \in A$ from a finite set of target authors A, where $|A| \leq c$ and $c \in \mathbb{Z}$ is a small constant. In our evaluation, we experimented with values of c ranging from 100 to 200. Although these attacks are less scalable, they can have more severe consequences. Large-scale inference typically targets the general population, leading to broad but diluted effects, such as user profiling for targeted advertising. In contrast, spear authorship inference attacks focus on a small set of high-value individuals—such as company executives or government officials—where the intended harm can extend across entire organizations or societies. We evaluate these attacks both in a closed-world setting and, in contrast to prior work on authorship inference, in several open-world settings where the victims are among unknown individuals.

We consider a hierarchical adversary with access to different modalities, both before and after processing (see Figure 2 for an overview of our adversarial setup). Specifically, an adversary targeting T2I models may have access to natural language instructions (the model input), generated images (the model output), or both. Since these modalities differ, separate models must be designed for each case. For the input-only scenario, adversarial models must effectively process text, while for the output-only case they must handle images. The multi-modal scenario is more complex, as it requires a combined approach where the choice of text and image embeddings is crucial for effective multi-class classification. We next elaborate on the formulations for each type of adversary.

Inference Framework Overview. Figure 3 illustrates our method for authorship inference in text-to-image (T2I) model usage. Without loss of generality, we assume the target sensitive attribute is authorship, where an author is the user who provides the natural language instruction (NLI) to the T2I model. Binary attribute inference is a special case of this framework. Using supervised learning,

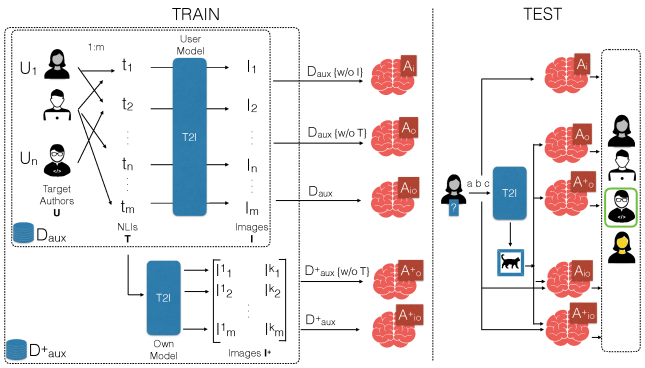


Figure 3: Inference Framework Overview

the adversary trains classifiers to infer authorship. Specifically, for a set of users $U = U_1, U_2..U_c$, NLIs $T = T_1, T_2..T_m$, and images $I = I_1..I_m$, where $m \geq c$, they can train:

- An **input-only model** (A_i) using (U, T) pairs.
- An **output-only model** (A_o) using (U, I) pairs.
- A **multi-modal model** (A_{io}) using (U, T, I) triplets.

This approach enables authorship inference across different modalities, improving classification accuracy and robustness.

As we will demonstrate, authorship inference models perform well when the adversary knows the target's T2I model, a reasonable assumption given prior model attribution methods [85]. However, in some cases, the adversary may be unable to reliably identify the user's model. Training on images from a different model may not generalize well, limiting inference accuracy.

To address these challenges, we employ data augmentation for training inference models. Specifically, the adversary generates an augmented image set I^+ containing $m \times k$ images, where each $t_i \in T$ produces k new images. By training on $I \cup I^+$, the adversary enhances output-only ($\mathcal{A}_o^+$) and input-output ($\mathcal{A}_{io}^+$) models.

During inference, the adversary selects the appropriate model to attribute intercepted samples to target users. While augmentation increases computational costs, it improves model transferability and performance (see Section 5.3). Additionally, the adversary can readily leverage public T2I models to augment their datasets.

Next, we describe the spear authorship inference models. Thus far, we have focused on a *closed-world* setting, where all inference samples originate from a predefined set of target authors. However, in real-world scenarios such as online forums and social networks, it is unrealistic to assume the adversary has training data for all authors. To address this, we introduce an *open-world* adaptation, allowing inference on both known and unknown users. We detail this approach in Section 3.5 and evaluate its effectiveness under varying known-to-unknown author ratios in Section 5.4.

3.2 Input Attack ($\mathcal{A}_i$)

Formulation. In this setup, D_{aux} consists solely of natural language instructions submitted to a text–to–image model, along with their corresponding author labels: $D_{aux} = \{(t_i, \alpha_i)\}_i$, where $t \in T$, T is the set of all instructions, $i = 1..n$, and n is the total number of training samples. $\mathcal{A}_i$ aims to learn $f_\theta : T \to S$ and $g_\phi : G(T) \to S$.

Design and Implementation. We construct multiple adversarial models to infer authorship from text. Our embedding-based classifiers (g_ϕ) consist of an embedding extraction component followed by a linear classification layer. We experiment with pre-trained InferSent and BERT models for embedding extraction, both of which have demonstrated strong performance in various NLP tasks. Additionally, we fine-tune BERT on our dataset of NLIs. BERT embeddings (a 768-dimensional vector) are fed into a linear layer for multi-class classification, while InferSent embeddings are processed through a dense layer with a softmax activation function. Training is optimized using the Adam optimizer [44].

Our raw-text classifier (f_θ) directly learns from T. We hypothesize that T2I users have distinct interests, subjects, or styles, which influence the vocabulary used in their NLIs. Thus, features like Bag-of-Words (BoW) should effectively capture author differences in instructions. Unlike other stylography features, which are important for large document inferences, BoW focuses solely on the presence of words, not their order or sentence structure. We use BoW to extract features relevant to word occurrence within NLIs, employing a TF-IDF vectorizer. These BoW features are then passed through a dense layer with a softmax activation function for multi-class classification.

3.3 Output Attack ($\mathcal{A}_o$)

Formulation. In this setup, D_{aux} consists of image outputs from a text–to–image model along with their corresponding author labels (i, α), where $i \in I$ and I is the set of all generated images. $\mathcal{A}_o$ aims to learn $g_\phi : G(I) \to S$. Note that for A_o, we do not consider learning directly from the raw data ($f_\theta : I \to S$) because we expect abstract attributes like authorship, or semantically meaningful context related to authorship, to be more effectively captured in image embeddings. These embeddings are generally more useful than handcrafted features extracted from raw data.

Design and Implementation. A straightforward approach to infer authorship from image embeddings is to apply pre-trained CNN models, such as ResNet [34], InceptionV3 [96], or VGG19 [41], to extract image features. These embeddings have proven effective for image classification. CNNs employ a hierarchical structure, enabling more abstract representations compared to handcrafted feature extraction methods like SIFT [55] and HOG [20].

However, our key insight is that inferring *authorship* from images generated by NLIs requires embeddings suited for semantic image processing—ones that encode relationships between textual descriptions and images. Traditional image embeddings, while superior to handcrafted methods, primarily capture objects in images but struggle to generalize to unseen objects, styles, and topics. To address this, we use models ($G(I)$) that learn visual representations from natural language supervision. Specifically, our $\mathcal{A}_o$ attack models leverage vision transformer architectures such as CLIP [69] and ViT [106]. CLIP is used in some T2I models to encode NLIs and guide the diffusion process, whereas ViT serves as a more general vision transformer model.

The transformer-based vision embeddings are then passed to various popular classifiers configured for multi-class classification using the scikit-learn library [26]. In particular, we employ an *SVM* classifier with a *One-Vs-One* approach, a *Logistic Regression* classifier using the *One–Vs-The–Rest* method, and a Naive Bayes classifier leveraging a multinomial model to compute class probabilities. Additionally, we use Decision Trees, Random Forests, Neural Networks, and KNN classifiers with a *multi-label* approach.

Note that training on a single image generated from an NLI for a given T2I model may not always be effective. This is due to the approximate nature of the diffusion process, which can produce different images each time based on user-specified input parameters. For instance, the `cfg` scale parameter (classifier-free guidance scale) in Stable Diffusion controls the diversity of the generation process, determining how strictly the model follows the given instruction. Furthermore, for each NLI and diversity setting, multiple images may serve as representative visual interpretations of the instruction. A key insight of our output attack strategy is to incorporate data augmentation into the training process to address this variability. Specifically, for each NLI instruction, we generate multiple images and use all of them to train the adversarial models. We denote this attack model as $\mathcal{A}_o^+$.

3.4 Multi–Modal Input–Output Attack ($\mathcal{A}_{io}$)

Formulation. In this setup, we consider a multi-modal adversary ($\mathcal{A}_{io}$). The auxiliary dataset D_{aux} consists of natural language prompts for a text-to-image model, the corresponding generated images, and author labels for each prompt (t, i, α). $\mathcal{A}_{io}$ aims to learn the mapping $g_\phi : G^T(T) \oplus G^I(I) \to S$, where the adversary is restricted to learning from the combination of text and image embeddings. This design choice is based on our previous observation that transformer-based image embeddings more effectively capture image representations for this inference task.

Design and Implementation. Similarly to $\mathcal{A}_o$, we leverage CLIP to extract image embeddings. A key property of CLIP is its ability to extract both text and image embeddings, with the crucial advantage that embeddings from different modalities share the same feature space. CLIP's contrastive learning approach brings images closer to their corresponding text descriptions while pushing dissimilar images further apart in the latent space. The textual and vision embeddings are combined after flattening the text embeddings. Since both reside in the same feature space, we use simple concatenation–rather than alternative methods like averaging–to effectively enrich the information available to our multi–modal classifiers. We train the same classifiers as in the output attack, denoting the non–data–augmented multi-modal attack as $\mathcal{A}_{io}$ and the data–augmented version as $\mathcal{A}_{io}^+$.

3.5 Open World Adaptation

In some settings the adversary might not know all participants and therefore available content might appear from unknown authors. This corresponds to an open–world setting which better suits scenarios on some real–world discussion forums and social network groups. Here we adapt our formulation for such a setting and design proof–of–concept spear authorship inference model for the A_i adversary. We use A_i^c and A_i^o to distinguish between the closed and open world adversary respectively.

Formulation. In this setup, D_{aux} contains only natural language instructions submitted to a text–to–image model along with the author labels. The fundamental difference between the A_i^c and the A_i^o approach is that in the latter case the corresponding authors x consist of both known α and unknown β authors: $D_{\text{aux}} = \{(t_i, x_i) \mid x_i \in \{\alpha_i, \beta_i\}\}$, where $t \in T$. T is the set of all instructions, $i = 1..n$, and n is the total number of training samples.

Design and Implementation. The ratio of $\beta : \alpha$ can vary with $\beta > \alpha$. To assess the effectiveness of our approach, we introduce multiple versions of open–world datasets D_{aux}, each incorporating a greater number of unknown authors β (detailed in Section 5). To combat against the imbalanced nature of the dataset, we employed a weighted random sampler, in which the weights are inversely proportional to the class frequency. For our inference model we employ a variant of Sentence Transformer architecture, `all-MPNet-base-v2` [5]. We selected this model for its speed, compact size, and strong embedding performance. `all-MPNet-base-v2` achieved state–of–the–art results in both Performance Sentence Embeddings (assessing the quality of embedded sentences) and Performance Semantic Search (evaluating the quality of embedded search queries and paragraphs), outperforming 37 other sentence transformer models [76]. These models were benchmarked by averaging the Performance Sentence Embeddings and Performance Semantic Search, with consideration for speed and model size. We fine–tuned the model for each targetted $\beta : \alpha$ ratio for 20 epochs using the AdamW optimizer and minimizing the categorical cross-entropy loss per sample:

$$L(x, \hat{x}) = -\sum_{i=1}^{\alpha+1} w_i x_i \log(\hat{x}_i)$$

where x_i is the true label, $\hat{x}_i$ the predicted probability for author i, $w_i = \frac{1}{f_i}$, and f_i is the class frequency.

4 Evaluation Setup

Datasets. We use *DiffusionDB Large* [104], a large-scale text-to-image NLI dataset containing 14 million images generated by *Stable Diffusion* from 1.8 million NLIs. All NLIs were collected from Stable Diffusion Discord channels, and all images were generated using Stable Diffusion Version 1. We preprocess the dataset by removing: i) exact duplicate NLIs ii) NLIs containing non-English characters, iii) NLIs with at least one *NaN* or *Null* character, and iv) all whitespace. We refer to the resulting dataset as 'DiffusionDB'. The remaining unique authors and prompts are 10,334 and 1,760,664, respectively. We also use NLIs from 'DiffusionDB' to create images with Stable Diffusion (v2.1) and Latent Diffusion to further investigate the behavior of our spear authorship inferences. To study dementia markers, we use the *ADReSS* dataset [56], which contains speech transcriptions from both a control cohort and a dementia cohort tasked with describing a given image.

Spear Authorship Inference Evaluation Metric. To measure model performance, we use accuracy, defined as the fraction of correctly predicted records. In multi–class classification, Top-1 accuracy requires the model's highest–probability prediction to match the expected answer exactly. Top–N accuracy, on the other hand, measures how often the correct class appears within the top N predictions, e.g., in the softmax distribution [29]. Top–N accuracy is a strong indicator of privacy violations and aligns with privacy regulations. For instance, under GDPR, absolute certainty in identifying a data subject is not required; a probabilistic inference can suffice for differential treatment [81]. Prior work has leveraged Top-5 accuracy to evaluate authorship inference [92]. To facilitate comparison with related work and due to space constraints, we primarily report Top-5 accuracy, noting Top–1 accuracy when relevant. A full breakdown of Top–1 to Top–5 accuracies is provided in Appendix A for relevant experiments.

Research Questions. Our evaluation seeks to address the following key research questions: **RQ1:** Are T2I spearheaded authorship identification attacks effective? **RQ2:** How do the number of available training samples and the size of the target author set affect the attack performance? **RQ3:** What properties of NLIs and their generated images can be indicative of authorship? **RQ4:** Do these attacks depend on the target T2I model? **RQ5:** Can the attacks be adapted to an open–world setting? **RQ6:** Do these attacks remain robust against SOTA mitigation strategies? **RQ7:** Do T2I models encode other user attributes in their generated images?

5 Evaluation

5.1 Attack Effectiveness

To answer RQ1, we evaluate the overall effectiveness of our spearheaded attacks in a closed-world setting. For the threat model, we fix the number of authors to $|A| = 100$, a reasonable size for a spearheaded attack. This choice aligns with prior work on authorship identification [92] to facilitate comparison. We then assess the ability of $\mathcal{A}_i$, $\mathcal{A}_o$, and $\mathcal{A}_{io}$ to identify target authors. Specifically, we evaluate 7 $\mathcal{A}_i$ architectures, 11 $\mathcal{A}_o$ architectures, and 12 $\mathcal{A}_{io}$ architectures, as described in Section 3.

The models' performance is compared using a limited number of training samples per author, i.e., $|D_{aux}^{\alpha}| = 10, 30, 50, 70, 90$. This represents a challenging scenario for the adversary, who, as we will demonstrate, improves as more labeled samples become available. Authors are selected by first identifying those with at least 100 prompts (3947/10334) in the *DiffusionDB* dataset, and then randomly choosing 100. For the $\mathcal{A}_o$ models, we repeat the experiments twice: once using CLIP embeddings from images and once using ViT embeddings, to evaluate the (in)dependence on the text encoder of the underlying T2I model.

Next, we repeat the experiments with $|D_{aux}^{\alpha}| = 70$ while varying the number of authors, i.e., $|A| = 100, 150, 200$. This resulted in training and evaluating 49 $\mathcal{A}_i$ models, 154 $\mathcal{A}_o$ models, and 84 $\mathcal{A}_{io}$ models. Although increasing the number of authors beyond 200 is outside the scope of our spear authorship attack, we conduct some preliminary experiments and provide insights into the adversary's ability in such settings.

Input Attack ($\mathcal{A}_i$) Effectiveness. We compare our InferSent and BERT embedding-based classifiers, as well as our Bag–of–Words (BoW) classifier, against one of the best-performing models from Song and Raghunathan [92] (LSTM_BookCorpus). Additionally, we compare against an improved version that we trained on our dataset (LSTM). We also evaluate a Logistic Regression (LR) multi-class classifier trained with text embeddings extracted using CLIP [69], a vision transformer model. Finally, we compare against TextCNN [43],

which has shown success in other authorship attribution works [79, 86].

Our results are summarized in Figure 4(a). We found that all our classifiers (InferSent, BERT, LSTM_DiffusionDB, and BoW) outperform the other models, significantly surpassing the performance of TextCNN, LR, and LSTM_BookCorpus baselines. The LSTM trained on our dataset performed better than the baselines, but still fell short compared to our attack models. InferSent achieved the best performance with a top-5 accuracy of 0.77, leveraging only 70 labelled NLIs per author for training. Performance gradually improved as more training samples were added [1]. In contrast to prior work on authorship identification from text [92], we observe that even with limited training samples, models like BoW, which do not rely on pre-trained unsupervised embeddings, generalize well and perform effectively in our task. This can be attributed to the nature of the text in NLIs. NLIs have more constrained stylistic properties; for example, all NLIs describe an image, and authors use NLIs to generate specific sets of images with distinct styles. Additionally, NLIs tend to be shorter than sentences from books. We provide a more detailed analysis of this in Section 5.2.

Additionally, when increasing the number of authors ($|A| = 200$) while using a small number of training samples per author ($|D^{\alpha}_{aux}| = 70$), the InferSent and BoW models remain robust, whereas the performance of the other models begins to degrade.

Output Attack ($\mathcal{A}_o$) Effectiveness. As explained in Section 3, our adversarial models use CLIP-large [69] and ViT-large [106] to extract image embeddings, which are then fed into several popular classical classification algorithms: Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), Decision Tree (DT), Neural Network (NN), and K-Nearest Neighbor (KNN). All classifiers are configured using the default parameters from the scikit-learn library [26].

We compare these with traditional CNN-based models for image feature extraction. Specifically, we evaluate a Residual Network–50 (ResNet50) [34], a 50-layer convolutional neural network pre-trained on ImageNet-1k [80] for image classification tasks. Instead of using the image embeddings from CLIP/ViT as input to a classifier, as in our attack models, we first use the ResNet50 baseline to extract image features from the input images. To improve generalization, we apply image perturbations using *ImageDataGenerator* [42] from Keras. Since ResNet50 has not been trained for the authorship identification task, we add a custom classification head and fine-tune the model. To prevent overfitting or underfitting due to an inappropriate number of epochs, we use *early stopping*, monitoring the validation loss to terminate training. We set the total number of training epochs to 100, the patience to 3 epochs, and the validation split to 0.2. We apply the same procedure to other state-of-the-art pre-trained CNN models, including ResNet50V2 [35], InceptionV3, and VGG19. Lastly, since authorship attribution from T2I-generated images is a novel task, we denote *"RC"* as the random choice (RC) probability of an author being in the top–5 predicted authors.

[1] We investigated the ability of the models to improve as more unique prompts become available. We leveraged the maximum number of unique prompts required to obtain 100 unique authors from DiffusionDB. In this case our best-performing model, $\mathcal{A}_i$, achieved a Top–5 accuracy of 94.8% using 1,500 available prompts per author, and a 70:30 train–test split.

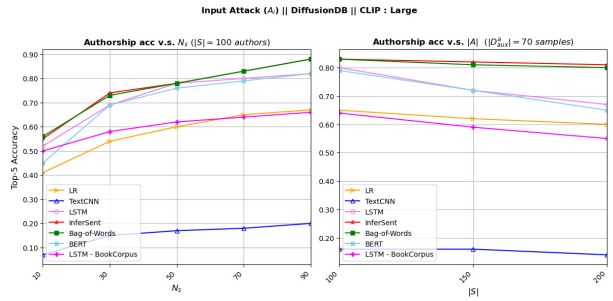
Our results are summarized in Figures 4(b) and 4(c). Note that the results presented in this section are from attack models without data augmentation. Nevertheless, we observe that it is possible to identify NLI authors from T2I-generated images with high top–5 accuracy. Additionally, we find that training simple classifiers using both CLIP and general ViT embeddings yields better results compared to classifying directly on more traditional image feature embeddings extracted with high-performing CNNs. Logistic Regression with CLIP and ViT embeddings consistently outperformed other models. For instance, Logistic Regression with CLIP embeddings on 70 training samples per author achieved the highest top–5 accuracy of 0.79, outperforming all other models (see Figure 4(b)). Using ViT embeddings, Logistic Regression also outperforms the other models with a top–5 accuracy of 0.77 on 70 training samples (see Figure 4(c)).

As we increase the number of authors ($|A| = 200$) while keeping the number of training samples small ($|D_{aux}| = 70$), Logistic Regression and SVM using CLIP and ViT embeddings remain the highest-performing attacks, although with an expected drop in performance in the limited training samples available scenario.

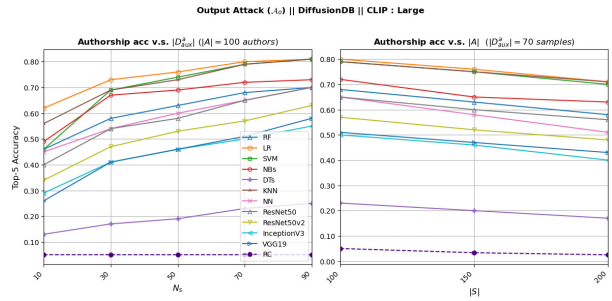
Input-Output Multi-Modal Attack($\mathcal{A}_{io}$) Effectiveness. We evaluate this attack scenario similarly to the output adversary ($\mathcal{A}_o$). As explained in Section 3, instead of solely learning from image embeddings, our models concatenate these with text embeddings from the same feature space. For comparison, we use the baseline classification models from $\mathcal{A}_o$, modified for multi-modal learning. Specifically, we use the models in $\mathcal{A}_o$ to extract image feature embeddings, which are then concatenated with text embeddings from CLIP, after flattening them to match the image embedding dimensions. As before, we annotate the results with an *RC* adversary, as this is a novel adversarial task, and we aim to provide a naive baseline to better highlight the improvements achieved by more sophisticated adversarial models.

Our results are summarized in Figure 4(d). Similar to the $\mathcal{A}_o$ case, leveraging CLIP embeddings (for both NLIs and images) outperforms the other models. For instance, Logistic Regression with 70 training samples per author surpasses the other models with a top–5 accuracy of 0.88. Compared to the $\mathcal{A}_o$ scenario, this represents an 11% performance increase under the same experimental conditions. With 200 authors, the models remain robust, with the highest top–5 accuracy slightly dropping to 0.85. In general, the multi–modal $\mathcal{A}_{io}$ adversary performs better than both the single-modality image-only $\mathcal{A}_o$ and text-only $\mathcal{A}_i$ adversaries. The classifiers also demonstrate better robustness as the number of target authors increases.

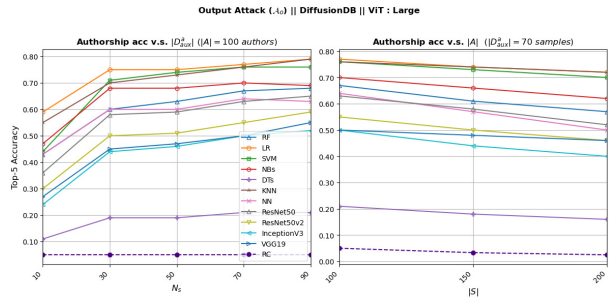
NSFW Content. The inference models could be potentially used for attribution of not safe for work (NSFW) content created with T2I models. Here we evaluate the attribution capabilities of each of the $\mathcal{A}_i$, $\mathcal{A}_o$ and $\mathcal{A}_{io}$ models. We classify an image as NSFW if it attains a safety score of ≥ 0.7 [46]. Given the limited quantity of *unsafe* content within *DiffusionDB*, setting this threshold allows us to identify only 115 authors in total, each associated with a minimum of 100 *unsafe* images. After manual observation, we found that a significant number of images include nude content. Table 1 shows the best performing classifiers. We note that when $|A| = 100$, Logistic Regression excels with 0.74 and 0.85 Top–5 accuracy under $\mathcal{A}_o$ and $\mathcal{A}_{io}$, respectively. These results demonstrate that the



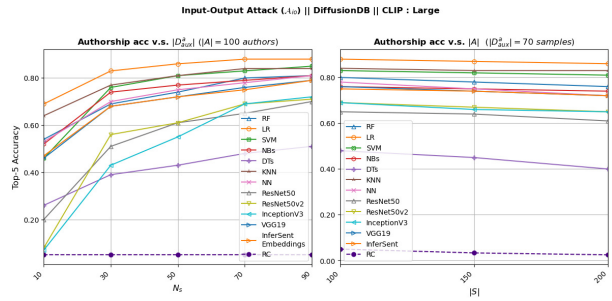
(a) Input Attack ($\mathcal{A}_i$).



(b) Output Attack ($\mathcal{A}_o$), using CLIP embeddings.



(c) Output Attack ($\mathcal{A}_o$), using ViT embeddings.



(d) Multi–Modal Input–Output Attack ($\mathcal{A}_{io}$), using CLIP embeddings.

Figure 4: Comparison of all adversary models on DiffusionDB.

our inference models can be an important tool in NSFW content attribution.

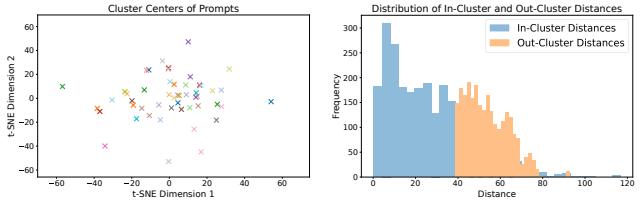
Table 1: The Top–5 accuracy of the best performing (in parenthesis) Authorship Inference models using NSFW images.

	Experiment Setup : $\lvert\mathcal{D}_{aux}\rvert$ = 10...90 with $\lvert\mathcal{A}\rvert$ = 100								
	Training Samples								
	10	20	30	40	50	60	70	80	90
$\mathcal{A}_i$ (InferSent)	0.61	0.72	0.76	0.78	0.80	0.81	0.83	0.84	0.85
$\mathcal{A}_o$ (LR)	0.61	0.67	0.72	0.73	0.73	0.74	0.74	0.75	0.77
$\mathcal{A}_{io}$ (LR)	0.69	0.77	0.81	0.83	0.84	0.85	0.86	0.86	0.87

5.2 Leakage Analysis

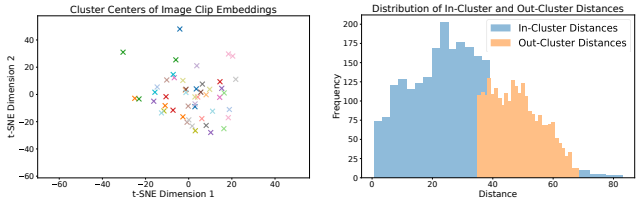
In this section, we aim to gain a deeper understanding of why authorship inference is possible in T2I. Specifically, we explore (a) why authorship leakage occurs through natural language instructions and (b) why leakage is possible through generated images.

The importance of words authors use in NLIs. One of our main findings is that simple techniques based on bag of words (BoW) achieve impressive performance in inferring authorship from NLIs. To better understand why this is possible, we analyze the authors' NLIs. We randomly sample 50 authors from our training set of 100 authors, and for each author, we randomly select 50 NLIs. The choice of 50 instead of 100, as used in the attacks, is due to computational constraints and easier visualization. Next, we extract the TF–IDF embeddings for each NLI. Since TF–IDF embeddings are multi-dimensional and difficult to visualize, we



(a) t–SNE Cluster Centroids of per–author NLI TF–IDF embeddings.

(b) In–Cluster and Out–Cluster Distances of per-author NLI TF–IDF embeddings.



(c) t–SNE Cluster Centroids of per–author CLIP image embeddings.

(d) In–Cluster and Out–Cluster Distances of per–author CLIP image embeddings.

Figure 5: Analysis on 50 authors with 50 NLIs each (a,b) and 50 images generated from those NLIs using DiffusionDB.

apply t-SNE [99] (t-distributed Stochastic Neighbor Embedding), a widely-used unsupervised dimensionality reduction technique, to reduce the dimensionality of each NLI's TF–IDF embedding to two. We then calculate the per-author centroids of the TF–IDF embeddings by averaging across the two dimensions, as shown in Figure 5a. Finally, we compute the Euclidean distances for both in-cluster and out-cluster NLIs. As seen in Figure 5b, most in-cluster

distances are small (indicating greater similarity between NLIs), while the out-cluster distance distribution is shifted toward larger values. Figure 5a shows that the cluster centroids of author TF–IDF embeddings are sufficiently separated. The $P25$, $P50$, $P75$, and $P90$ percentiles for in-cluster distances are 10.7, 24.3, 38.8, and 52.1, respectively, while for out-cluster distances, they are 45.0, 51.9, 61.0, and 68.1. The minimum out-cluster distance is 38.1, and nearly 75% of in-cluster distances are below this threshold.

The importance of image embeddings for author separability. To analyze the output space of T2I models for authorship separability, we follow a similar methodology as in the NLI term analysis, but instead of analysing TF–IDF embeddings of NLIs, this time we focus on CLIP image embeddings of images from DiffusionDB. We note that for this experiment we use the same 50 authors and prompts as previously. Figures 5c and 5d show the per–author cluster centroids of CLIP image embeddings after t–SNE dimensionality reduction and the cluster distance distributions. These results show that separability is more challenging on the output compared to the input of T2I models which is expected due to the lossy image generation process which is inevitable due to the imperfect alignment between instruction and image. Nonetheless it is still possible with cluster centroids not overlapping. Also the $P25$, $P50$, $P75$, $P90$ for in–cluster distances and out–cluster distances are 17.6, 28.0, 41.6, 53.4 and 43.2, 50.0, 57.4, 63.1 respectively with the minimum out–cluster distance is 36.7 and between the median and 75% of in–cluster distances are still below that threshold.

Discussion on content dependence. Other works on authorship inference showed that authors' stylometric features might be correlated with the text topic [65] and more generally the content of the written text might be biasing classification. Our experiments show that it is still possible to perform classification on authorship and dementia inference even on similar content. Specifically, our experiments with NSFW images conduct authorship attribution on only NSFW content, while our dementia inference model (Section 5.6) is trained on descriptions of the same image. In the case of authorship inference it is still possible that different authors might focus on different topics. Therefore, the inference models might need to be retrained periodically as the target users' interests shift. We do not expect this to happen very often and existing models dependent on user interests are successfully deployed in practice (e.g., personalized advertising).

5.3 Model Diversity

To answer RQ4 on whether our attacks depend on the T2I model we design a separate experiment. This experiment is performed on the output attacks because the effect of the T2I model (generated images) is better isolated if we focus only on the output. By definition input attacks are independent of the T2I model, and input–output attacks can only be better than output attacks. In this experiment we want to analyze the performance of our attack models when trained on images of a T2I model and perform inference on images of the same model (classifier–dependent attacks), and when trained on images of a T2I model and perform inference on images of another model (classifier independent attacks). We evaluate models trained and tested on images from both SD v2.1 and LD, and to show the benefits of data augmentation we train the models without and with data augmentation. This totals 8 different settings for which we evaluate all our attack models from Section 5.1.

Table 2: Top–5 Accuracy of Best Performing Output Attack ($\mathcal{A}_o$) models without and with data augmentation in model–dependent (*) and model–independent inference settings.

Training Images		Test Images T2I Models	
T2I Model	**Data Augmentation**	**Stable Diffusion v2.1**	**Latent Diffusion**
Stable Diffusion v2.1	NO	0.559* (LR)	0.541 (LR)
	YES	0.805* (KNN)	0.803 (KNN)
Latent Diffusion	NO	0.553 (LR)	0.553* (LR)
	YES	0.802 (KNN)	0.807* (KNN)

To generate the datasets for the non–augmented attacks, we use the same set of 100 authors with 100 NLIs each from Section 5.1 to generate 10,000 new images using the open–source and free LD [77] and SD v2.1 [78] models. To generate the data augmented datasets we repeat the above but instead of generating one image for each NLI we generate 5 image variants per NLI. Table 2 summarizes all our results where we only report the best performing output model due to space limitations.

We first observe that non data–augmented output attacks ($\mathcal{A}_o$) do not achieve the same performance as the non data–augmented output attacks on DiffusionDB (see Section 5.1). We hypothesize that this is because the authorship content might transfer better in the DiffusionDB (Stable Diffusion v1 model). To examine this hypothesis we use the CLIP score to evaluate the similarity between the NLI and the output images (non–augmented datasets) from DiffusionDB, Stable Diffusion v2.1, and Latent Diffusion. To calculate the CLIP scores we use the multi–modal CLIP–Large [69] to extract text and image embeddings from NLIs output images respectively. Since the embeddings are in the the same vector space we can then calculate their cosine similarity. The CLIP score ranges between $[0..100]$ and higher scores indicate higher similarity. Figure 6 shows the cumulative distribution function of the CLIP scores for the three models. Given prior work [32], a CLIP score around to 0.26 indicates that a generated image is has good alignment to its input prompt. Clearly, DiffusionDB has significantly more images (around 80%) with score ≥ 0.26 than Stable Diffusion v2.1 (around 55%) and Latent Diffusion (around 20%). This might explain why the non–augmented attacks are more successful on DiffusionDB images. However, as shown on Table 2 our data–augmented attacks ($\mathcal{A}_o^+$) achieve high Top–5 accuracy in model–dependent and model–independent settings and across training sets.

5.4 Open–World Evaluation

In Section 3.5 we introduced our approach for adapting the spear authorship inference problem to an open–world setting and introduced our open–world spear authorship inference model (A_i^o). Here we evaluate this approach (RQ5).

Experiment Setup. We introduce unknown authors (β) into our existing target author (α) dataset. We mimic five scenarios with different ratios of unknown to target authors ($\beta : \alpha$) ranging from *5:1, 7:1, 10:1* to *20:1*. Specifically, we randomly select *500, 700, 1,000, and 2,000* unknown authors, and then for each author we randomly select 5 NLIs, resulting in *2,500, 3,500, 5,000, and 10,000* unknown

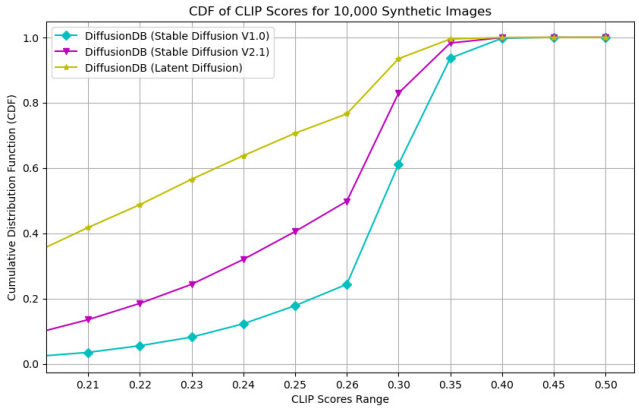


Figure 6: CDF of CLIP scores for 10K images per model.

Table 3: Top–1 & Top–5 accuracy on different ratios of unknown to target authors for the open–world experiment.

	Ratio of Unknown : Target authors				
Accuracy	5 : 1	7 : 1	10 : 1	20 : 1	90 : 1
Top–1	0.7530	0.7559	0.7343	0.7293	0.6251
Top–5	0.8620	0.8631	0.8437	0.8236	0.6981

author prompts, respectively. To further extend the open-world setting, we incorporate the entire *DiffusionDB* author set, yielding 8,884 unknown authors (a ratio of ∼*90:1*). For each setting we use a 70:30 train:test split to train and evaluate our A_i^o model for each setting and report the Top-1 and Top-5 accuracy.

Results. Table 3 summarizes our results. As expected the performance of the A_i^o inference model gradually decreases as we introduce more unknown authors. Nonetheless, in all settings inferences demonstrate Top–5 accuracy between 80% – –86% with even Top–1 accuracy consistently above 70%. This demonstrates effectiveness of spear authorship inferences in open–world settings.

5.5 Robustness Against Obfuscation

To further analyze the robustness of the inference models derived with our methods (RQ6), we explore plausible strategies one could employ to obfuscate the target attribute. Since the image generation is conditioned on the text input, it is reasonable to assume that if obfuscation is successful on the input domain, then the generated images from the obfuscated instructions will also be harder to distinguish between authors. Hence, we select several strategies that can be applied on text. These include, Pegasus [112] which is a SOTA text paraphrasing model, and a differentially private approach proposed by Mattern et al. [62] which adds noise through a temperature value (inverse of ϵ) at the word level. Additionally, we use top performing and open source LLMs Llama3 (8B parameters) [3] and Mistral (7B parameters) [4] to generate paraphrases of NLIs in a zero–shot setting.

We also consider an adaptive version of the $\mathcal{A}_i$ model, where the adversary is assumed knowledge of the defense mechanism. In the adaptive setting, the inference models are trained on both the original data and on data obfuscated by the respective mechanism, and tested on unseen obfuscated data. We select our best performing

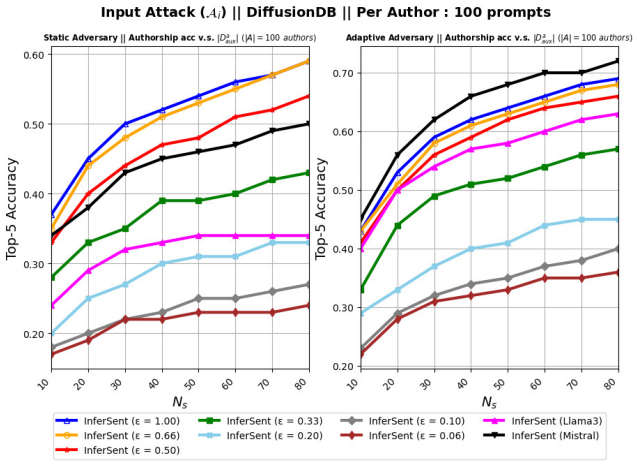


Figure 7: The attack success rate of static and adaptive adversarial InferSent-based models against LLMs and PEGASUS. The defense method is shown in parentheses.

Table 4: Semantic similarity between original and paraphrased NLIs, using PEGASUS, Llama3 and Mistral.

Sequence–to–Sequence Model (PEGASUS)							
Temperature (T)	1.0 (No DP)	1.5	2.0	3.0	5.0	10.0	15.0
Privacy Budget ($\epsilon = 1/T$)	1.00	0.67	0.50	0.33	0.20	0.10	0.07
Mean Similarity	0.770	0.757	0.730	0.670	0.588	0.537	0.520
Large Language Models (LLMs)							
	Llama3 (8B)			**Mistral (7B)**			
Mean Similarity	0.693			0.743			

$\mathcal{A}_i$ model, *InferSent*, and run it both in a static and an adaptive setting when each defense mechanism is in place. Due to space limitations, the experimental details and a more rigorous analysis is deferred to the Appendix B. Here we briefly summarize the results.

We evaluated the performance of the inference models in terms of privacy, utility of the NLIs, and end–to–end utility. Figure 7 shows the inference performance in the static and adaptive setting where only a few approaches such as Llama3 and word differential privacy with Pegasus ($\epsilon \leq 0.33$) being the more promising. However, as shown on Table 4, with such ϵ values the semantic similarity of the obfuscated NLI with the original user intended NLI drops below 60%. To analyze end to end utility we measure the CLIP scores for 500 synthetic images generated with the original NLIs and the most promising obfuscation approaches (see Figure 8). We find that only a small percentage of the generated images of all defense models has ≥ 0.26 CLIP score indicating a large loss in semantics in the output space (see an example in Figure 9 and in Appendix B.2 a further discussion). Overall, we conclude that while text paraphrasing can be a promising approach for obfuscating authorship in both the input and output space of T2I models, the large loss in utility highlights the need for future work in attribute obfuscation.

5.6 Leakage of Other Attributes

We have demonstrated that authorship information can be leaked through both the input (natural language instructions) and the output (generated images) of text–to–image models. However, other

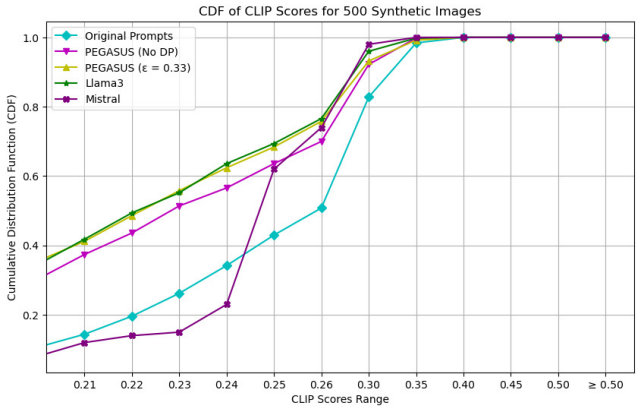


Figure 8: CDF of CLIP scores for 500 images generated from original NLIs (Stable Diffusion v2.1), and from paraphrased NLIs using different obfuscation mechanisms.

attributes may be at risk as well (RQ7). Inspired by the fact that pathogenic markers are identifiable via text [12, 115] here we study whether they survive the diffusion process and appear in the generated images. This should be possible, if such language markers are embedded in the joint text–image latent space used to guide the denoising process. Then if the denoising process is accurate, then it is possible that the markers will affect the respective image. To study this effect, we leverage an existing dataset with manual transcriptions of image descriptions from a control group and a group suffering from Alzheimer's disease. Alzheimer's disease is a type of dementia neurocognitive condition which affects (very conservatively) tens of millions of vulnerable users worldwide with its number of cases expected to triple by 2050 to over 150 million [1], and its leakage from technology usage has not been studied before. Alzheimer's disease is extremely sensitive with patients vulnerable to manifestations of restlessness and agitation which require management through costly antipsychotic drugs. Leaking one's condition can be manipulated for targeted advertising, spear phishing attacks, bias in the workplace and others and thus constitutes a major privacy concern.

Dataset. The ADReSS [56] dataset, is a subset of the DementiaBank dataset [14] that is pre–processed and balanced in terms of age and gender. Participants are diagnosed with various stages of dementia. During the experiment, 156 participants are provided with the same "Cookie Theft Picture" shown in Figure 10 and are asked to describe it orally. The descriptions are manually transcribed eliminating the effect of ASR errors. The ADReSS dataset contains healthy-control (CC) and dementia–labelled (AD) descriptions (see Figure 11 in Appendix E for example of prompts), with 54 train and 24 test samples for each class, for a total of 156. We further preprocess the dataset to remove irrelevant sentences. For example, after manual analysis, we observed that the descriptions might contain clarification questions irrelevant to the task of describing an image. One such example is shown in Figure 10. We manually remove all such instances from the transcriptions to ensure the descriptions are more valid and focused.

Experimental Setup. To test whether dementia–related language markers are encoded in joint text–image representations used in text–guided diffusion models, we use the same black–box approach used to study spear authorship leakage. In particular, we first verify that dementia can be inferred from text and reproduced the text–only classifier from Balagopalan et al. [12] which achieves SOTA results in dementia classification from transcribed speech. In our implementation we preprocess the text data using a pre–trained BERT tokenizer. Then we split the data into training/test sets with $70:30$ split. Then we train the model using the pre–trained BERT model's parameters, fine–tuning the entire BertForSequenceClassification model, using AdamW (adam with weight decay) optimizer and a learning rate = $2e-5$ scheduled over 20 epochs. The model's performance is evaluated on the validation set for each fold. Our best model achieves 87.5% accuracy which matches the reported 87.5% accuracy from Zhu et al. [115].

This approach though uses BERT embeddings. To study whether the dementia features are encoded in text–image embeddings which are used in text–guided image generation with diffusion, we train A_i^d dementia inference models on the CLIP embeddings extracted from the ADReSS cleaned image descriptions. We train the models for 100 epochs using the Binary Cross Entropy Loss. The A_i^d model will show whether joint representations encode dementia language markers. To further show whether the final generated images guided by such representations can leak dementia, we need to further study the success of dementia classifiers trained on the generated images (A_o^d). The ADReSS dataset only provides the image description (text). We use the descriptions to generate images with Stable Diffusion v2.1 and Latent Diffusion. Then we extract CLIP embeddings from the images and train all model architectures used in A_o spear authorship inferences. We use a $70:30$ training/test split (20% of the training set is used for validation to avoid overfitting to the test set) because of the small size of the *ADReSS* dataset. Lastly, for completeness we also retrain all the A_{io} models on the concatenated text and image embeddings using the same training settings as above to derive the A_{io}^d.

Table 5: The accuracy of the best performing (in parenthesis) Dementia Inference models using images from Stable Diffusion v2.1 and Latent Diffusion.

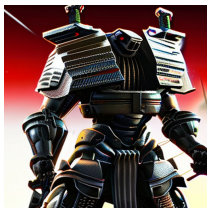
	$\mathcal{A}_i^d$	$\mathcal{A}_o^d$	$\mathcal{A}_{io}^d$
Stable Diffusion v2.1	0.833 (NN)	0.75 (SVM)	0.797 (RF)
Latent Diffusion		0.729 (NN)	0.771 (NN)

Discussion. The highest performing models are shown on Table 5. Our findings demonstrate that pathological markers in textual image descriptions are encoded in joint text–image representations as shown at the high accuracy (83.3%) of the $\mathcal{A}_i^d$ model in differentiating between dementia and control descriptions from such representations. Moreover, our $\mathcal{A}_o^d$ results (75% accuracy) show that the images generated with diffusion–based T2I models guided by such representations can themselves contain indications of the condition which reveals a previously unidentified privacy risk with the usage of such models by vulnerable populations. Finally, the $\mathcal{A}_{io}^d$ accuracy is close to 80% which improves compared to $\mathcal{A}_o^d$ but drops

(a) samurai mecha on ghost of tsushima box art

(b) samurai mecha on ghost of tsushima box art

(c) The ghost of tsushima box art has samurai mecha on it. (Similarity: 0.88)



(d) The samurai mecha is on the ghost of tsushima box art. (Similarity: 0.88)

(e) The samurai mecha is on the ghost of shushima box art (Similarity: 0.85)

(f) samurai mecha (Similarity: 0.67)

(g) There were samurai mehaga on that piece. (Similarity: 0.59)

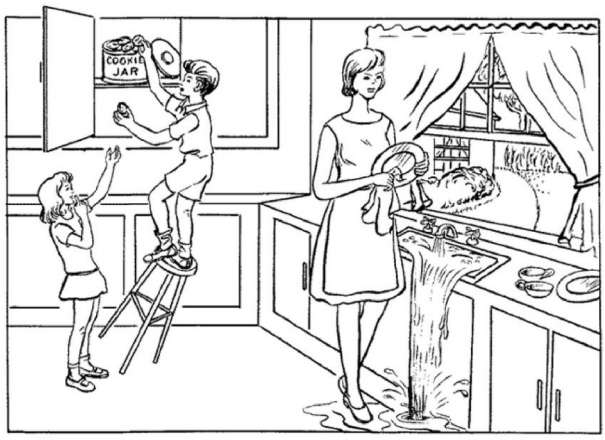
(h) There are art objects for which samurai iconography can be associated - samurai me (Similarity: 0.58)

(i) There are art reproductions of Tokugawa box pieces. benkins on (Similarity: 0.54)

(j) A cover of Ghost of Tsushima features a striking fusion of traditional Japanese warrior...(Similarity : 0.78)

(k) Mecha samurai depicted on the cover art of Ghost of Tsushima (Similarity: 0.88)

Figure 9: Qualitative comparison of end–to–end utility of mitigation strategies. T2I images, their NLI and text similarity score in parenthesis with the original NLI. (a) shows the original image included in DiffusionDB; (b) the original NLI is given as input to Stable Diffusion V2.1 [78]; in $(c), (d), (e), (f), (g), (h)$ and (i) the input is the paraphrased NLI with different temperatures ($T = 1, 1.5, 2, 3, 5, 10$ and 15). In (j) and (k) we use the paraphrases from Llama3 and Mistral. The full text produced by Llama (j) is "*A cover of Ghost of Tsushima features a striking fusion of traditional Japanese warrior attire and futuristic machinery, in a unique and captivating image of a samurai–inspired mecha.*"



mhm.well the water's running over on the floor.uh the chair is tilting.the boy is into the cookie jar.and his sister is reaching for a cookie.the mother's drying dishes.um do you want action or just want anything I see? okay.mhm.

Figure 10: An example of a transcription (bottom) of an *AD* patient describing the Cookie Theft picture shown above.

compared to $\mathcal{A}_i^d$. The latter may be due to vector concatenation increasing dimensionality, potentially diluting the signal–to–noise ratio and reducing $\mathcal{A}_{io}^d$ accuracy relative to $\mathcal{A}_i^d$. This may be due to vector concatenation increasing dimensionality, which can dilute the signal-to-noise ratio and degrade performance.

In recent work [59] we further showed that removing pathological language markers such as discourse tokens (e.g. "um", "well" etc.) which were shown in prior studies to be indicative of dementia in speech and language, reduces classification accuracy in all settings with drastic decreases of more than 10% in the $\mathcal{A}_o^d$ and $\mathcal{A}_{io}^d$ settings.

6 Related Work

Inferences in T2I. Others [85] illustrated how to attribute a generated image to its source T2I model, while Carlini et al. [16] demonstrated how to extract training examples from diffusion models. Also, Qu et al. [72] showed how prone these models are to generate unsafe content. However, to the best of our knowledge, our work is the first to demonstrate attribute leakage from NLIs and images.

Authorship Inference. Prior works used stylometric features to infer authorship such as character n-grams [84, 93], syntax parse trees [114], words, part–of–speech tags [38], and topic information incorporated into CNN models [89]. Other works leverage transformer–based architectures such as BERT [22] and RoBERTa [53] which are fine-tuned to automatically capture writing style features from raw text [23]. Song and Ragunathan [92] leveraged embeddings to train classifiers to infer authorship in the Book corpus. Others have leveraged LLMs [39, 40] to avoid the need for training data. Specifically, Huang et al.[39] employ GPT–4 Turbo with Linguistically Informed Prompting (LIP) to discriminate among

20 candidate authors in the Blog corpus, but their macro–F1 drops from ≈ 0.79 with 10 authors to ≈ 0.55 with 20 authors, indicating diminishing returns as the author set grows. In contrast, at our scale of 200 authors we get 68.6% accuracy (see Table 6 in Appendix A. However, none of these approaches have explored authorship leakage in T2I models. We uniquely leverage data augmentation and multi–modal text–image alignment models to achieve SOTA performance with a large number of authors, limited training sample sizes, and in classifier–dependent and independent ways.

Authorship Obfuscation. Several works exist aiming to obfuscate attributes such as authorship in text. Some follow a style transfer approach [58, 87], others used rule–based modifications [74, 107], leveraging LLMs for paraphrasing [15, 27], obfuscating machine generated authorship [103], or used differentially private methods [62]. We have selected several SOTA representative methods to evaluate the robustness of our attacks including the differentially private approach from Mattern et al. [62], and SOTA open source LLMs. A common limitation of those approaches is the negative effect on utility. Therefore we also included the SOTA Transformer–based paraphrasing model [112] which achieves good semantic preservation. Nonetheless, we have shown in our evaluation, any negative semantic preservation effects in the text input can be compounded in T2I models to further damage the end–to–end utility and highlighted the impending need for more practical obfuscation models to defend T2I users' privacy.

Dementia Classification. The field of dementia classification from speech [56] has grown significantly, particularly since the 2020 *ADReSS* challenge. These models can be categorized into three main types: those using exclusively acoustic data [17, 57], those solely utilizing speech transcriptions [12, 31, 57, 67], and hybrid models combining both [33, 61, 110, 116]. More related to our dementia inference classifiers are models trained on transcriptions. Among those the best performance was achieved by BERT–based models [12]. More similar to our output and input–output dementia inference attacks are the models reported in a preliminary pre–print article by Zhu et al. [115] which combine information from the original "Cookie Theft Picture" with transcriptions to achieve 89.6% accuracy on the *ADReSS* dataset. Similarly with Zhu et al. [115] we also utilize text and image information in our A_io inference. However, both our goal and settings are different. Unlike prior work focused on dementia classification accuracy,our study investigates the leakage of cognitive decline markers from text into images and therefore we focus on the ability to learn from joint text–image representations.Also in our setting we do not assume prior knowledge of the image the participants are trying to describe and instead use representations extracted from generated images through diffusion models, a more challenging task. Lastly, in our recent work we leverage an explainability–based approach to further understand what language characteristics contribute to such leakage and analyzed the relationship between information units in input instructions and generated images [59].

Image Captioning and Recovering Instructions from Images. Several studies [37, 91, 95] have used CNNs and RNNs to reconstruct descriptions from images. Closest to our work, Croitoru et al. [19] tries to recover the embeddings of the NLIs reaching a performance of 0.69, but without identifying their authors. In contrast, our A_o attack models leverage only the generated images to identify their NLI authors with 0.77 Top–5 accuracy with only 70 training samples. Other Transformer–based image–text alignment models such as BLIP [50] and CLIP [73] can be leveraged for image captioning. Image captioning seems like a reasonable approach for inferring authorship because if we can accurately reconstruct NLIs from T2I images we could potentially infer authorship from captions. In our research we found that this is not possible with the current SOTA captioning models which yield poor semantic similarity with the original NLIs (see Appendix C). Instead our A_o and $\mathcal{A}_{io}$ models classify authors from transformer–based image embeddings, and they achieve high Top–5 accuracy due to the good alignment between NLIs and T2I images in that embedding space.

7 Conclusion

We have revealed the threat of attribute leakage in text–guided diffusion–based image generation models. Unique to our work we model a hierarchical adversary with different access levels on the input, output or multi–modal input–output access to such models. We show that by leveraging data augmentation and models which learn joint representations of text and images to extract multi–modal embeddings adversaries can achieve state of the art results in spear authorship inference even with limited training data. Our inference models are rigorously evaluated and demonstrated robustness against different training datasets and diffusion models, number of authors, and strong mitigation strategies, and generalization to open world settings. We have also shown that text–image encoders used to guide diffusion models encode pathological speech markers which allows an adversary to infer the presence of dementia from natural language instructions and generated images, highlighting a previously unidentified threat for a vulnerable understudied population whose numbers are rapidly increasing. We hope our work raises awareness on the privacy hazards on the usage of diffusion–based image generation models, and inspire further work on privacy leakages in such models and mitigation mechanisms to warrant users' privacy while achieving acceptable model utility.

Ethical Considerations

Datasets. We use publicly available datasets, namely the DementiaBank dataset released under the ADReSS Challenge and the DiffusionDB dataset published by Stability AI. *DiffusionDB* is released under a CC0 1.0 license that allows uses for any purpose [7] and data were collected with provisions for both licencing and participant privacy [104]. The ADReSS dataset has been obtained upon request through our membership to the Dementia Bank. As members we are committed to the Dementia Bank's Code of Ethics [97].

Responsible Research Conduct. We recognize the broader ethical implications of deploying attribute inference methods in sensitive domains. Our study aims to contribute to a better understanding of these risks and to inform the development of generative models and safeguards that respect user privacy.

Acknowledgments

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] 2022. WORLDWIDE DEMENTIA CASES. https://alzheimersresearchuk.org/news/worldwide-dementia-cases-to-triple-by-2050-to-over-150-million. Accessed: 2024-06-018.

[2] 2023. GPT-4V(ision) System Card. https://api.semanticscholar.org/CorpusID:263218031

[3] 2024. Meta - Llama3. https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct. Accessed: 2024-03-01.

[4] 2024. Mistral. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3. Accessed: 2024-03-01.

[5] 2024. Sentence Transformers. https://www.huggingface.co/sentence-transformers/all-mpnet-base-v2. Accessed: 2024-03-01.

[6] DALL-E 2. 2024. DALL–E 2 Subreddit. https://www.reddit.com/r/dalle2/. Accessed: 2024-07-11.

[7] StabilityAI. 2022b. 2022. Terms of Use. https://stability.ai/terms-of-use.

[8] Apify. 2024. Apify Technologies. https://apify.com/stifl/reddit-username-from-subreddit. Accessed: 2024-07-11.

[9] Hadi Askari, Anshuman Chhabra, Muhao Chen, and Prasant Mohapatra. 2024. Assessing LLMs for Zero-shot Abstractive Summarization Through the Lens of Relevance Paraphrasing. *arXiv preprint arXiv:2406.03993* (2024).

[10] Zhongjie Ba, Jieming Zhong, Jiachen Lei, Peng Cheng, Qinglong Wang, Zhan Qin, Zhibo Wang, and Kui Ren. 2023. SurrogatePrompt: Bypassing the Safety Filter of Text-To-Image Models via Substitution. *arXiv preprint arXiv:2309.14122* (2023).

[11] Badiucao. 2024. Art in Protest. https://artinprotest.viewingrooms.com/viewing-room/9-badiucao/. Accessed: 2024-07-26.

[12] Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. 2020. To BERT or not to BERT: comparing speech and language-based approaches for Alzheimer's disease detection. *arXiv preprint arXiv:2008.01551* (2020).

[13] Michael Barz, Sven Stauden, and Daniel Sonntag. 2020. Visual search target inference in natural interaction settings with machine learning. In *ACM Symposium on Eye Tracking Research and Applications*. 1–8.

[14] James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. 1994. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of neurology* 51, 6 (1994), 585–594.

[15] Jon Ander Campos and Jun Shern. 2022. Training language models with language feedback. In *ACL Workshop on Learning with Natural Language Supervision. 2022.*

[16] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX Conference on Security Symposium* (Anaheim, CA, USA) *(SEC '23)*. USENIX Association, USA, Article 294, 18 pages.

[17] Karol Chlasta and Krzysztof Wołk. 2021. Towards computer-based automated screening of dementia through spontaneous speech. *Frontiers in Psychology* 11 (2021), 623237.

[18] Craiyon. 2024. *AI Image Generator Tool.* Accessed: 2024-02-19.

[19] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Reverse Stable Diffusion: What prompt was used to generate this image? (08 2023).

[20] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, Vol. 1. Ieee, 886–893.

[21] DALL-E. 2024. DALL–E Subreddit. https://www.reddit.com/r/dalle/. Accessed: 2024-07-11.

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL] https://arxiv.org/abs/1810.04805

[24] Stable Diffusion. 2024. Stable Diffusion Subreddit. https://www.reddit.com/r/StableDiffusion/. Accessed: 2024-07-11.

[25] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[26] Pedregosa Fabian. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research 12* (2011), 2825.

[27] Jillian Fisher, Ximing Lu, JaeHun Jung, Liwei Jiang, Zaid Harchaoui, and Yejin Choi. 2024. JAMDEC: Unsupervised Authorship Obfuscation using Constrained Decoding over Small Language Models. *arXiv preprint arXiv:2402.08192* (2024).

[28] Dawna Fung. 2024. Drawing for Democracy. https://www.fairplanet.org/story/political-cartoonist-in-hong-kong/. Accessed: 2024-07-26.

[29] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2018. Softmax units for multinoulli output distributions. Deep Learning.

[30] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Y. Bengio. 2014. Generative Adversarial Networks. *Advances in Neural Information Processing Systems* 3 (06 2014). https://doi.org/10.1145/3422622

[31] Yue Guo, Changye Li, Carol Roan, Serguei Pakhomov, and Trevor Cohen. 2021. Crossing the "Cookie Theft" corpus chasm: applying what BERT learns from outside data to the ADReSS challenge dementia detection task. *Frontiers in Computer Science* 3 (2021), 642517.

[32] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2024. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems* 36 (2024).

[33] R'mani Haulcy and James Glass. 2021. Classifying Alzheimer's disease using audio and text-based representations of speech. *Frontiers in Psychology* 11 (2021), 624137.

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 630–645.

[36] Xinlei He, Savvas Zannettou, Yun Shen, and Yang Zhang. 2024. You only prompt once: On the capabilities of prompt learning on large language models to tackle toxic content. In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 770–787.

[37] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-term Memory. *Neural computation* 9 (12 1997), 1735–80. https://doi.org/10.1162/neco.1997.9.8.1735

[38] Zhiqiang Hu, Roy Ka-Wei Lee, Lei Wang, Ee-peng Lim, and Bo Dai. 2020. Deepstyle: User style embedding for authorship attribution of short texts. In *Web and Big Data: 4th International Joint Conference, APWeb-WAIM 2020, Tianjin, China, September 18-20, 2020, Proceedings, Part II 4*. Springer, 221–229.

[39] Baixiang Huang, Canyu Chen, and Kai Shu. 2024. Can Large Language Models Identify Authorship? *arXiv preprint arXiv:2403.08213* (2024).

[40] Chia-Yu Hung, Zhiqiang Hu, Yujia Hu, and Roy Ka-Wei Lee. 2023. Who wrote it and why? prompting large-language models for authorship verification. *arXiv preprint arXiv:2310.08123* (2023).

[41] Simonyan Karen. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv: 1409.1556* (2014).

[42] Keras. 2024. ImageDataGenerator. https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/image/ImageDataGenerator. Accessed: 2024-02-01.

[43] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. arXiv:1408.5882 [cs.CL] https://arxiv.org/abs/1408.5882

[44] Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations* (12 2014).

[45] APSS Lab. 2025. SQUAD. https://github.com/APSS-Imperial/Attribute-Leakage.

[46] LAION-AI. 2024. LAION-SAFETY. https://github.com/LAION-AI/LAION-SAFETY. Accessed: 2024-02-01.

[47] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:233296808

[48] Lexica. 2022. Lexica Art. https://lexica.art/. Accessed: 2024-02-01.

[49] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.

[50] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.

[51] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *NeurIPS*.

[52] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602* (2021).

[53] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[54] Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893* (2018).

[55] David G Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, Vol. 2. Ieee, 1150–1157.

[56] Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's dementia recognition through spontaneous speech: The adress challenge. *arXiv preprint arXiv:2004.06833* (2020).

[57] P Mahajan and V Baths. 2021. Acoustic and language based deep learning approaches for Alzheimer's dementia detection from spontaneous speech. Front Aging Neurosci. 2021; 13.

[58] Asad Mahmood, Faizan Ahmad, Zubair Shafiq, Padmini Srinivasan, and Fareed Zaffar. 2019. A girl has no name: Automated authorship obfuscation using mutant-x. *Proceedings on Privacy Enhancing Technologies* (2019).

[59] Mansi, Anastasios Lepipas, Dominika Woszczyk, Yiying Guan, and Soteris Demetriou. 2025. Understanding Dementia Speech Alignment with Diffusion-Based Image Generation. In *Proceedings of the 26th Interspeech Conference*. ISCA, Rotterdam, The Netherlands.

[60] Francisco Supino Marcondes, José João Almeida, and Paulo Novais. 2020. Structural Onomatologic for Username Generation: A Partial Account. In *STAIRS@ECAI*. https://api.semanticscholar.org/CorpusID:221839940

[61] Matej Martinc, Fasih Haider, Senja Pollak, and Saturnino Luz. 2021. Temporal integration of text transcripts and acoustic features for Alzheimer's diagnosis based on spontaneous speech. *Frontiers in Aging Neuroscience* 13 (2021), 642647.

[62] Justus Mattern, Benjamin Weggenmann, and Florian Kerschbaum. 2022. The limits of word level differential privacy. *arXiv preprint arXiv:2205.02130* (2022).

[63] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. 2019. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*. IEEE, 691–706.

[64] Midjourney. 2024. Midjourney Subreddit. https://www.reddit.com/r/midjourney/. Accessed: 2024-07-11.

[65] George K Mikros and Eleni K Argiri. 2007. Investigating Topic Influence in Authorship Attribution.. In *PAN*.

[66] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[67] Tristan Millington and Saturnino Luz. 2021. Analysis and classification of word co-occurrence networks from Alzheimer's patients and controls. *Frontiers in Computer Science* 3 (2021), 649508.

[68] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning*. PMLR, 16784–16804.

[69] OpenAI. 2024. CLIP Large. https://huggingface.co/openai/clip-vit-large-patch14. Accessed: 2024-02-01.

[70] OpenAI. 2024. OpenCLIP-ViT/H. https://github.com/mlfoundations/open_clip. Accessed: 2024-02-01.

[71] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[72] Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. 2023. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*. 3403–3417.

[73] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[74] Muhammad A. Ramadhan, Dyah P. Sari, and Anik Musdholifah. 2021. Author Obfuscation in Indonesian News Articles Using Genetic Algorithms. In *Proceedings of the 2021 International Conference on Computational Linguistics*.

[75] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative Adversarial Text to Image Synthesis.

[76] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 3973–3983. https://doi.org/10.18653/v1/D19-1410

[77] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[78] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042

[79] Sebastian Ruder, Parsa Ghaffari, and John G Breslin. 2016. Character-level and multi-channel convolutional neural networks for large-scale authorship attribution. *arXiv preprint arXiv:1609.06686* (2016).

[80] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115 (2015), 211–252.

[81] Giovanni Sartor, Francesca Lagioia, et al. 2020. The impact of the General Data Protection Regulation (GDPR) on artificial intelligence. (2020).

[82] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. 2023. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22522–22531.

[83] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114* (2021).

[84] Roy Schwartz, Oren Tsur, Ari Rappoport, and Moshe Koppel. 2013. Authorship attribution of micro-messages. In *Proceedings of the 2013 Conference on empirical methods in natural language processing*. 1880–1891.

[85] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. 2023. DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Generation Models. 3418–3432. https://doi.org/10.1145/3576915.3616588

[86] Rakshith Shetty, Bernt Schiele, and Mario Fritz. 2018. {A4NT}: Author attribute anonymity by adversarial training of neural machine translation. In *27th USENIX Security Symposium (USENIX Security 18)*. 1633–1650.

[87] Rakshith Shetty, Bernt Schiele, and Mario Fritz. 2018. {A4NT}: Author attribute anonymity by adversarial training of neural machine translation. In *27th USENIX Security Symposium (USENIX Security 18)*. 1633–1650.

[88] Mohammad Shokri, Sarah Ita Levitan, and Rivka Levitan. 2025. Personalized Author Obfuscation with Large Language Models. *arXiv preprint arXiv:2505.12090* (2025).

[89] Prasha Shrestha, Sebastian Sierra, Fabio A González, Manuel Montes, Paolo Rosso, and Thamar Solorio. 2017. Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th conference of the European chapter of the association for computational linguistics: Volume 2, short papers*. 669–674.

[90] Lei Shu, Liangchen Luo, Jayakumar Hoskere, Yun Zhu, Yinxiao Liu, Simon Tong, Jindong Chen, and Lei Meng. 2024. Rewritelm: An instruction-tuned large language model for text rewriting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 18970–18980.

[91] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv 1409.1556* (09 2014).

[92] Congzheng Song and Ananth Raghunathan. 2020. Information Leakage in Embedding Models. (03 2020).

[93] Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology* 60, 3 (2009), 538–556.

[94] Stifl. 2024. REDDIT USERNAMES from Subreddit - Scraper. https://apify.com/stifl/reddit-username-from-subreddit. Accessed: 2024-07-11.

[95] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9. https://doi.org/10.1109/CVPR.2015.7298594

[96] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

[97] talkbank.org. 2024. Ethics. https://talkbank.org/0share/ethics.html

[98] NLTK Team. 2024. NLTK Names Corpus. https://www.nltk.org/howto/corpus.html. Accessed: 2024-07-11.

[99] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[100] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. https://arxiv.org/pdf/1706.03762.pdf

[101] Aidan Walker. 2024. Inside the Magical World of AI Prompters on Reddit. https://hyperallergic.com/855052/inside-the-magical-world-of-ai-prompters-on-reddit/. Accessed: 2024-07-28.

[102] Haining Wang. 2023. Defending Against Authorship Identification Attacks. *arXiv preprint arXiv:2310.01568* (2023).

[103] Ziyao Wang, Thai Le, and Dongwon Lee. 2024. Authorship Obfuscation in Multilingual Machine-Generated Text Detection. *arXiv preprint arXiv:2401.07867* (2024).

[104] Zijie Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Polo Chau. 2022. DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models. (10 2022). https://doi.org/10.48550/arXiv.2210.14896

[105] Benjamin Weggenmann and Florian Kerschbaum. 2018. Syntf: Synthetic and differentially private term frequency vectors for privacy-preserving text mining. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 305–314.

[106] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. 2020. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677* (2020).

[107] Eric Xing, Saranya Venkatraman, Thai Le, and Dongwon Lee. 2024. ALISON: Fast and Effective Stylometric Authorship Obfuscation. In *Proceedings of the*

AAAI Conference on Artificial Intelligence, Vol. 38. 19315–19322.
[108] Bingxuan Xu, Rui Meng, Yue Chen, Xiaodong Xu, Chen Dong, and Hao Sun. 2023. Latent semantic diffusion-based channel adaptive de-noising semcom for future 6g systems. In *GLOBECOM 2023-2023 IEEE Global Communications Conference*. IEEE, 1229–1234.
[109] Yuchen Yang, Bo Hui, Haolin Yuan, Neil Gong, and Yinzhi Cao. 2024. Sneakyprompt: Jailbreaking text-to-image generative models. In *2024 IEEE symposium on security and privacy (SP)*. IEEE, 897–912.
[110] Jiahong Yuan, Yuchen Bian, Xingyu Cai, Jiaji Huang, Zheng Ye, and Kenneth Church. 2020. Disfluencies and Fine-Tuning Pre-Trained Language Models for Detection of Alzheimer's Disease.. In *Interspeech*, Vol. 2020. 2162–6.
[111] Han Zhang, Tao Xu, and Hongsheng Li. 2017. StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks. 5908–5916. https://doi.org/10.1109/ICCV.2017.629
[112] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*. PMLR, 11328–11339.
[113] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. TinyLlama: An Open-Source Small Language Model. arXiv:2401.02385 [cs.CL] https://arxiv.org/abs/2401.02385
[114] Richong Zhang, Zhiyuan Hu, Hongyu Guo, and Yongyi Mao. 2018. Syntax encoding with application in authorship attribution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2742–2753.
[115] Youxiang Zhu, Xiaohui Liang, John A Batsis, and Robert M Roth. 2021. Exploring deep transfer learning techniques for Alzheimer's dementia detection. *Frontiers in computer science* 3 (2021), 624683.
[116] Youxiang Zhu, Xiaohui Liang, John A Batsis, and Robert M Roth. 2022. Domain-aware intermediate pretraining for dementia detection with limited data. In *Interspeech*, Vol. 2022. NIH Public Access, 2183.

A Breakdown of Top–5 Accuracy

Table 6 presents Top–1 to Top–5 accuracy results of the inference models across three setups $|A| = 100, 150, 200$ with a fixed training sample size of 70. Accuracy generally improves from Top–1 to Top–5 across all inference models ($\mathcal{A}_i$, $\mathcal{A}_o$ and $\mathcal{A}_{io}$), indicating model performance consistency as more candidate labels are allowed. For example, in the $|A| = 200$ case, $\mathcal{A}_{io}$ achieves 75.30% Top–1 accuracy, reaching 85.84% at Top–5, highlighting its robustness across ranks. Also, in the case of $|A| = 150$, the Top–1 accuracy for $\mathcal{A}_i$ is 68.91%, while $\mathcal{A}_o$ drops to 60.05%. This shows a clear performance gap, indicating that the input space alone provides more precise author identification than using only the output space. In contrast, our multi–modal inference model ($\mathcal{A}_{io}$ performs significantly better, achieving 77.74% Top–1 accuracy in the same scenario. This demonstrates the benefit of leveraging both input and output space.

Table 6: Top 1–to–5 accuracy scores.

	Top 1	Top 2	Top 3	Top 4	Top 5
Experiment Setup : $\lvert\mathcal{D}_{aux}\rvert$ = 70 with $\lvert\mathcal{A}\rvert$ = 100					
$\mathcal{A}_i$	68.91%	75.14%	77.84%	79.91%	81.61%
$\mathcal{A}_o$	60.05%	68.54%	73.24%	76.77%	79.44%
$\mathcal{A}_{io}$	77.74%	82.47%	85.07%	86.77%	88.04%
Experiment Setup : $\lvert\mathcal{D}_{aux}\rvert$ = 70 with $\lvert\mathcal{A}\rvert$ = 150					
$\mathcal{A}_i$	68.56%	74.05%	77.05%	78.72%	80.25%
$\mathcal{A}_o$	54.99%	64.16%	68.78%	71.87%	74.58%
$\mathcal{A}_{io}$	76.58%	81.74%	84.18%	85.74%	87.14%
Experiment Setup : $\lvert\mathcal{D}_{aux}\rvert$ = 70 with $\lvert\mathcal{A}\rvert$ = 200					
$\mathcal{A}_i$	68.62%	73.29%	75.75%	77.42%	79.07%
$\mathcal{A}_o$	52.54%	61.01%	65.86%	69.10%	71.90%
$\mathcal{A}_{io}$	75.30%	80.35%	83.23%	84.97%	85.84%

B Robustness Against Obfuscation

Our evaluation (Section 5) shows that adversaries can construct input (text), output (image) and multi–modal input-output (text–image) attacks to infer authorship from NLIs and their T2I images. Here we explore the robustness of our attacks against SOTA mitigation mechanisms. These results are summarized in Section 5.5.

B.1 Setup and Mitigation Approaches

Authorship Obfuscators. Authorship inference have been an issue traditionally studied in the written language domain. In this study we showed that leakage can happen from both the input (text) and the output (image) domain of text–to–image models. However, it is reasonable to assume that if obfuscation is successful on the input domain, then the generated images from the obfuscated instructions will also be harder to distinguish between authors.

Prior works have proposed text obfuscation strategies that can be generally classified into either learning anonymous textual representations [105] or developing text transformation mechanisms that can obfuscate authorship and simultaneously retain the semantics or meaning of the original input [102]. The latter category is more appropriate in our case since it is easier to integrate in a T2I scenario.

Given the above, to evaluate the most promising defenses, we chose a SOTA text paraphrasing model (PEGASUS [112]) as our baseline obfuscator. Such models tend to achieve good diversity in text generation while achieving high preservation of the original semantics of the text. Nonetheless, they are not designed for obfuscation tasks. To mitigate this, we leverage the observations by Mattern et al. [62] who show that by varying the *temperature* in the word sampling stage of language models can be used to inject noise, offering a privacy–utility control knob. In our implementation, we apply the pre–trained paraphrase model PEGASUS, to generate paraphrases of maximum length $n = 60$ while using different temperatures $T = 1, 1.5, 2, 3, 5, 10$ and 15. The temperature value (T) has an inverse relationship with the privacy budget value (ϵ), i.e. $\epsilon = 1/T$. Generally, the lower the value of epsilon, the higher the level of privacy protection our defense mechanism can provide, but this commonly lowers the utility of the data.

Moreover, given the stupendous success large language models (LLMs) exhibit in various tasks, we consider some of the most performing open source large language models (LLMs) to construct paraphrases of the original NLI [90]. In our work, we employ and evaluate the Llama3 and Mistral models. We chose these because they are free and open–source models. Both of them have shown robust performance on downstream tasks [9, 113]. Specifically, a single request was made for each unique NLI, which consisted of a fixed instruction text prepended to the NLI. The prefix we used was "*Paraphrase the following instruction while maintaining semantics and hiding authorship. Return only the paraphrased output, nothing else*". A single instruction like 'hide authorship' can significantly reduce authorship-verification accuracy—by up to 40 FF_1 points—while preserving meaning, as shown with Llama 3 and GPT–4 on benchmark corpora [88]. This brief phrase aligns with prior adversarial stylometry work [88] and proves effective and precise for prompting LLM–driven obfuscation.

Static and Adaptive Adversary. To be more rigorous in our evaluation of the defense mechanisms, we need to consider an adversary that is in knowledge of the defense mechanisms and has the ability to adopt to it. Thus we consider both a *static* and an *adaptive* version of the $\mathcal{A}_i$ adversary. In a static or non–adaptive setting, the adversarial models are trained on the source data (similar to the models presented this far) and tested on obfuscated data. In the adaptive setting, the adversarial models are trained on both the original and the obfuscated data by the respective mechanism and tested on unseen obfuscated data. We select our best performing $\mathcal{A}_i$ model, *InferSent*, and run it both in a static and an adaptive setting when each defense mechanism is in place.

B.2 Results

Privacy. The privacy results of the static and adaptive adversary against the mitigation strategies are summarized in Figure 7 where we show the adversary's performance under a limited sample setting with 10 >= $|D^{\alpha}_{aux}|$ <= 80. Most mitigation strategies seem promising against a static adversary which stays below a 0.6 Top–5 accuracy with 80 samples. In particular, when leveraging D^{α}_{aux} = 70 training samples the Top–5 accuracy drops to 0.5 and lower (see Figure 7) for both LLMs and for Pegasus with ϵ <= 0.33. Llama3 offers better protection than Mistral as Top–5 attack accuracy against Llama3 drops to 0.34 compared to 0.50 against Mistral.

Under the more realistic *adaptive adversary*, we observe that both Mistral and Llama3 fail to maintain similar levels of privacy protection. Mistral allows the attacker to reach 0.70 Top–5 accuracy with D^{α}_{aux} = 70 training samples. Llama3 offers better protection compared to Mistral with the attack success rate dropping to 0.63 Top–5 accuracy but much worse than its own performance against a static adversary. We argue that this happens because Mistral preserves higher NLI semantics (as we show in the utility analysis). Pegasus with low privacy protection $\epsilon = 1, 0.66, 0.50$ also fails to offer good guarantees against the adaptive adversary (Top–5 accuracy > 0.64 on 70 training samples and Top–5 accuracy > 0.66 on 80 training samples). As we increase the privacy protection for Pegasus (lower ϵ) the adversary success drops below 0.6 but with increased ϵ there is also a bigger penalty on utility as we will analyze later on.

Overall, Llama3 and Pegasus (ϵ <= 0.33) are the most promising in providing some privacy protection.

Utility of NLIs. To examine the defense' ability to preserve utility we first look at the input space (text) and how the each technique affects the semantics of the original NLIs. To measure this, we compute the cosine similarity of the original NLIs with paraphrased ones using the *all–MPNet–Base–v2* sentence transformers [5] model. The results for all potential mitigation strategies are summarized in Table 4.

We observe that as expected, the more we increase the privacy protection for Pegasus (lower ϵ) the lower the similarity (worse semantics) is between the original NLI and the paraphrased one. Mistral achieves the higher similarity but as we have seen the privacy protection offered is limited.

Overall the most promising approach is Pegasus with ϵ = 0.33. This achieves the best privacy–utility trade–off with Top–5 accuracy for the adaptive adversary below 0.6 and semantic similarity at 0.67. Llama3 is close with slightly better average NLI semantic similarity but worse protection.

End–to–End Utility. To better understand how the input semantic similarity offered by promising defense mechanisms translates to the end–to–end utility of T2I models we first perform a quantitative analysis comparing the inputs with the resulting outputs and then provide a qualitative example.

To study whether the generated images align with the intentions of the original NLIs, we compare the original (not obfuscated) NLI with the output of the T2I model given the *obfuscated* NLI. This simulates the scenario where the user provides their original NLI, and then a privacy layer obfuscates the NLI before making it available to the T2I model and more broadly accessible to the adversary. We envision this layer to be part of a user's device (mobile phone, browser, or voice assistant). To compare the text input with the image output we use CLIP score computed as described in Section 5.3.

For our analysis we chose the two most promising mitigation strategies Llama3 and Pegasus with ϵ = 0.33. As a baseline for comparison we also add PEGASUS without DP (ϵ = 1.0) which is one of the state of the art paraphrasing models and therefore while not optimised for privacy it yields good semantics as it is also shown on Table 4. Lastly we also add the original CLIP score of the target model which can serve as a baseline of end–to–end utility without any defense. For the experiment, and for each case, we randomly select 10 paraphrased NLIs of 10 authors. The authors are randomly sampled from the 100 authors dataset used in the Input–Output Attack ($\mathcal{A}_{io}$) (see Section 5.1). For each NLI we generate 5 sample images with Stable Diffusion v2.1. This resulted in 500 new images per model in total. Next, we compute the CLIP scores between the *original* NLIs and the synthetic images generated from the paraphrased NLIs. Figure 8 presents the CDF of the CLIP scores with different methods. We observe that only a small percentage of the generated images of all defense models has ≥ 0.26 CLIP score, indicating a large loss in semantics in the output space. Hence, none of the mitigation techniques provide a sufficient trade–off between the users' privacy and semantics in both input and output space.

Lastly we provide an example of how an input NLI is paraphrased by different defense mechanisms and the resulting generated image from the paraphrased NLI in Figure 9 (see Section 5.5). We observe that when ϵ = 0.33 (Figure 9f) important concepts from the original image are missing (e.g., samurai's head is omitted).

Overall. Our analysis showed that text paraphrasing can be a promising approach to obfuscate authorship. Techniques like LLMs (Llama3) and word differential privacy with Pegasus can offer some protection but the loss in semantics might not be acceptable and thus more needs to be done to adequately and realistically protect against the T2I authorship inference adversary. The advantage of the latter category is that it allows users to leverage the model's *temperature* as a privacy control knob to customize the degree of obfuscation. For LLMs, improvements using prompt–tuning (e.g. prefix–tuning [47], or p–tuning [52]) or addressing obfuscation separately in the input (text) and output (image) space might be avenues for future work but more precise characterization for these is needed.

C Caption models for inferring authorship from T2I–generated images.

To explore whether simply applying SOTA caption models on T2I images to recover the original NLI and potentially authorship information we performed an experiment with three state–of–the–art caption models, namely BLIP[50], BLIP2 [49] and LlaVA [51]. We use these models to caption text from 10,000 images randomly sampled from DiffusionDB dataset. We then compare the semantics of the original NLIs and the captions using the *all–MPNet-Base-v2* model. The results are summarize on Table 7. BLIP2 offers the best similarity score with 0.41 (see Table 7). However, the semantic results are quite low for all caption models, which indicates that directly inferring authorship from captions is challenging.

Table 7: Embeddings similarity between the original NLI and the generated image's caption produced with BLIP, BLIP–2 and LlaVA.

	Pretrained Image Captioning Models		
	BLIP	**BLIP-2**	**LlaVa**
Mean Similarity	0.35	0.41	0.35

D Measurement Study

To get a clearer picture of the pervasiveness and nature of anonymity in such settings, we conducted a supporting study on the official *Stable Diffusion* [24], *Midjourney* [64], *DALL–E* [21] and *DALL–E 2* [6] subreddits. Using the free version of the publicly available scraping tool [94] from Apify [8] we could to extract up to 1000 active members of a subreddit. Inspired by Marcondes et al. [60] we leverage a multi-faceted methodology–encompassing pattern recognition, comparison against known name databases (e.g., NLTK names corpus [98]), structural name–like analysis, fuzzy string matching and an additional layer of lexical checks using the WordNet database [66]–we differentiate aliases from real names. For example, this approach flagged user handles containing recognisable nouns like 'shark' or 'container' as more likely to be aliases. To assess the performance of this methodology, we first tested it on the Stable Diffusion subreddit. Of the 653 unique usernames scraped, 648 were classified as aliases and only 5 as real names. Then, an independent rater performed a manual analysis to these usernames, resulting in 619/29/3/2/95.5/99.7 for TP/FP/TN/FN/Precision/Recall. We follow the same procedure for the rest of the aforementioned subreddits. Table 8 shows our findings, highlighting only few misclassifications, reinforcing the idea that most online users intentionally use aliases to protect themselves from being personally identified.

Table 8: The number of unique active members that use aliases or real names as usernames in Stable Diffusion, Midjourney, DALL·E and DALL·E 2 subreddits.

Unique Active Members (First 1,000 results per subreddit)				
Usernames	**Subreddits**			
	Stable Diffusion	Midjourney	DALL·E	DALL·E 2
Aliases	648	738	267	235
Real Names	5	2	1	1

Control Group (CC)

"okay.many dish or the mother's washing the dishes and the sink is overflowing.she has some of them uh dried already on the side as she's looking out the window while the little boy is falling off the stool cause he's getting into the cookie jar to give to his little sister who's reaching up to get the cookie also.um there's water all over the floor.uh there's the garden is outside and the mother's not paying any attention to what they do.the stool is tipping.the cookie jar the um uh door is open.um there may be a little breeze coming in because the this window is open.um uh the little girl s saying has her finger to her mouth.shh t we won't tell mother while you give me the cookie.um it's in the kitchen of course and uh the cups two cups and a dish already have been dried.and uh the mother's stepping in the water and she's probably so engrossed in what she's doing outside she neither knows what the children are doing nor is she paying any attention that the water's overflowing."

Alzheimer's Disease Group (CC)

"mhm.well this one is in the cookie jar.and this is she tried to climb the uh...oh no no.this boy I think this is the same one huh tried to get in the cookie jar.and she's watching.and over here must be the mother.I don't know what the heck they're doin here though.what's goin on?the stool she was climbing and the stool tipped over.she was doin the dishes I think.she spilled something."

Figure 11: An example of transcriptions/prompts for the Cookie-Theft Image for the two groups.

E Dementia Dataset

The ADReSS dataset [56] consists of speech transcripts of descriptions of the Cookie Theft Picture (Figure 10 in Section 5.6). These transcripts are registered for two user groups- Dementia Patients (AD) and Healthy Control Group (CC). Examples of transcriptions of the two groups are shown in Figure 11.