HOW MANY VAN GOGHS DOES IT TAKE TO VAN GOGH? FINDING THE IMITATION THRESHOLD

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image models are trained using large datasets collected by scraping image-text pairs from the internet. These datasets often include private, copyrighted, and licensed material. Training models on such datasets enables them to generate images with such content, which might violate copyright laws and individual privacy. This phenomenon is termed *imitation* – generation of images with content that has recognizable similarity to its training images. In this work we study the relationship between a concept's frequency in the training dataset and the ability of a model to imitate it. We seek to determine the point at which a model was trained on enough instances to imitate a concept – the *imitation threshold*. We posit this question as a new problem: **F**inding the **I**mitation **T**hreshold (`FIT`) and propose an efficient approach that estimates the imitation threshold without incurring the colossal cost of training multiple models from scratch. We experiment with two domains – human faces and art styles – for which we create four datasets, and evaluate three text-to-image models which were trained on two pretraining datasets. Our results reveal that the *imitation threshold* of these models is in the range of 200-600 images, depending on the domain and the model. The *imitation threshold* can provide an empirical basis for copyright violation claims and acts as a guiding principle for text-to-image model developers that aim to comply with copyright and privacy laws. We will release the code and data upon publication.

1 INTRODUCTION

The progress of multi-modal vision-language models has been phenomenal in recent years (Ramesh et al., 2021; Rombach et al., 2022; Goodfellow et al., 2022), much of which can be attributed to the availability of large-scale pretraining datasets like LAION (Schuhmann et al., 2022). These datasets consist of semi-curated image-text pairs scraped from Common Crawl, which leads to the inclusion of explicit, copyrighted, and licensed material (Cavna, 2023; Hunter, 2023; Vincent, 2023; Jiang et al., 2023; Birhane et al., 2021). Training models on such images may be problematic because text-to-image models can *imitate* — generate images with highly recognizable features — concepts from their training data (Somepalli et al., 2023a; Carlini et al., 2023a). This behavior has both legal and ethical implications, such as copyright infringements as well as privacy violations of individuals whose images are present in the training data without consent. In fact, a large group of artists sued Stability AI, developers of the widely-used text-to-image models, alleging that the company's models generated images that distinctly replicated their artistic styles (Saveri & Butterick, 2023).

Previous work has focused on detecting when generated images imitate training images (Somepalli et al., 2023a; Carlini et al., 2023a), and its mitigation (Somepalli et al., 2023b; Shan et al., 2023). In particular, researchers found that duplicate images increase the chance of memorization and imitation. However, the relation between a concept's prevalence in the training data and the model's ability to imitate it remains unexplored.

In this work, we ask **how many instances of a concept does a text-to-image model need to be trained on to imitate it**, where concept refers to a specific person or a specific art style, for example. Establishing such an *imitation threshold* is useful for several reasons. First, it offers an empirical basis for copyright infringement and privacy violation claims, suggesting that if a concept's prevalence is below this threshold, such claims are less likely to be true (Saveri & Butterick, 2023; Vincent, 2023; Ceoln, 2022). Second, it acts as a guiding principle for text-to-image models developers that want to avoid such violations. Finally, it reveals an interesting connection between training data statistics and model behavior, and the ability of models to efficiently harness training data (Udandarao et al., 2024; Carlini et al., 2023b). We name this problem `FIT`: **F**inding the **I**mitation **T**hreshold, and provide a schematic overview of this problem in Figure 1.

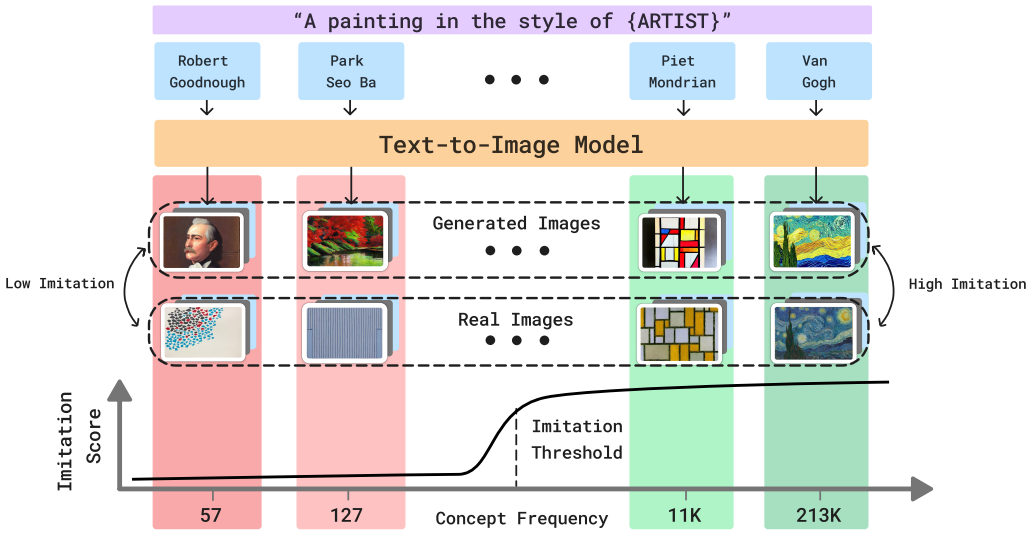


Figure 1: An overview of FIT, where we seek the *imitation threshold* – the point at which a model was exposed to enough instances of a concept that it can reliably imitate it. The figure shows four concepts (e.g., Van Gogh's art style) that have different counts in the training data (e.g., 213K for Van Gogh). As the image count of a concept increases, the ability of the text-to-image model to imitate it increases (e.g. Piet Mondrian and Van Gogh). We propose an efficient approach, MIMETIC^2, that estimates the imitation threshold without training models from scratch.

The optimal methodology to measure the imitation threshold requires training multiple models with varying number of images of a concept and measuring the ability of the these models to imitate it. However, training even one of these models is extremely expensive (Bastian, 2022). We propose a tractable alternative, **M**easuring **Im**itation Thr**E**shold **T**hrough **I**nstance **C**ount and **C**omparison (MIMETIC^2), that estimates the threshold without incurring the cost of training multiple models. We start by collecting a large set of concepts (e.g., various kinds of art styles) per domain (e.g., domain of art styles), and use a text-to-image model to generate images for each concept. Then, we compute the imitation score of each concept by comparing the generated images of that concept to its training images, and also estimate each concept's frequency in the training data. Finally, by sorting the concepts based on increasing frequency, we estimate the imitation threshold for that domain using a *change detection* algorithm Killick et al. (2012).

As we operate with observational data, a naive approach may be confounded by several factors, such as the quality of the imitation scoring model on different groups within the domain, or estimating the training frequencies of concepts (e.g., simple counts of 'Van Gogh' in the captions results in a biased estimate since the artist may be mentioned in the caption without their painting in the corresponding image). We carefully tailor MIMETIC^2 to minimize the impact of such confounders (§5).

Overall, we formalize a new problem – **F**inding **I**mitation **T**hreshold (FIT; §3), and propose a method, MIMETIC^2, that efficiently estimates the *imitation threshold* for text-to-image models (§5). We use our method to estimate the imitation threshold for two domains on four datasets, three text-to-image models that were trained on two pretraining datasets (§4). We find the imitation thresholds to range between 200 to 600 images, providing concrete insights on models' imitation abilities (§6).

2 Background

Text-to-Image Models Generating images from textual inputs is a long studied problem, with diffusion models as the current state-of-the-art (Sohl-Dickstein et al., 2015; Dhariwal & Nichol, 2021). Diffusion models are generative models that learn to approximate the underlying data distribution (conditioned or unconditioned). Given a trained diffusion model, it is possible to sample synthetic images from the learned distribution by performing a sequence of denoising operations. Open-sourced models like Stable Diffusion (SD) Rombach et al. (2022) are trained using large datasets such as LAION (Schuhmann et al., 2022), a large image-caption dataset scraped from Common Crawl.

Dataset Issues and Privacy Violations The advancement in text-to-image capabilities, largely due to big training datasets, is accompanied by concerns about the training on explicit, copyrighted, and licensed material Birhane et al. (2021) and imitating such content when generating images (Cavna, 2023; Hunter, 2023; Vincent, 2023; Jiang et al., 2023). For example, Birhane et al. (2021) and Thiel

(2023) found several explicit images in the LAION dataset and Getty Images found that LAION had millions of their copyrighted images Vincent (2023). Issues around imitation of training images has especially plagued artists, whose livelihood is threatened Saveri & Butterick (2023); Shan et al. (2023), as well as individuals whose face has been used without consent to create inappropriate content Hunter (2023); Badshah (2024). The imitation threshold would be a useful basis for copyright infringement and privacy violation claims in such cases.

Training Data Statistics and Model Behavior Pretraining datasets are a core factor for explaining model behavior (Elazar et al., 2024). Razeghi et al. (2022) found that the few-shot performance of language models is highly correlated with the frequency of instances in pretraining dataset. Udandarao et al. (2024) bolster this finding by demonstrating that the performance of multimodal models on downstream tasks is strongly correlated with a concept's frequency in the pretraining datasets. In addition, Carlini et al. (2023b) show that language models more easily memorize highly duplicated sequences. We find a similar phenomenon: increasing the number of images of a concept increases the similarity between its generated and training images. Crucially, instead of measuring *memorization*, we measure *imitation* to find the *imitation threshold*. Typically, *memorization* refers to the replication of a specific training image whereas we use *imitation* to refer to the generation of a concept when a model is given different prompts, such that the concept is recognizable in the generated images.

3 Problem Formulation and Overview

Finding the Imitation Threshold (`FIT`) seeks to find the minimal number of images with a concept (e.g., a specific art style) a text-to-image model has to see during training in order to imitate it. `FIT`'s setup involves a training dataset $\mathcal{D} \triangleq \{(\mathbf{x_i}, \mathbf{y_i}) \mid \mathbf{x_i} \in \mathcal{X},\ \mathbf{y_i} \in \mathcal{Y}\}_{i=1}^{n}$, composed of n (image, caption) pairs, where $\mathcal{X}$ and $\mathcal{Y}$ represents space of images and text captions, respectively and a model $\mathcal{M}$ that is trained on $\mathcal{D}$ to generate an image $\mathbf{x}$ given the text-caption $\mathbf{y}$. From hereafter, $\mathcal{M}(\mathbf{y})$ represents the generated image for a provided caption $\mathbf{y}$. Let $\mathbb{I}^j(\mathbf{x})$ be an indicator function that indicates whether a concept Z^j is present in an image $\mathbf{x}$. Each concept Z^j appears $c^j = \sum_i \mathbb{I}^j(\mathbf{x_i})$ times in the dataset $\mathcal{D}$, where the dataset may contain multiple concepts $\{Z^1, Z^2, \dots\}$. Lastly, $\mathcal{M}_k^j$ represents a model that is trained on a dataset where $c^j = k$.

The *imitation threshold* is the minimal number of training images containing some concept Z^j from which the model $\mathcal{M}$ generates images $\mathcal{M}(\mathbf{p}^j)$ (for different random seeds) where Z^j is recognizable as the concept, where prompt $\mathbf{p}^j$ mentions the concept Z^j.[1] For example, if Z^j refers to Van Gogh's art style, then the imitation threshold is the minimal number of training images of Van Gogh's paintings in a dataset which, when used to train a text-to-image model, can generate paintings that have Van Gogh's art style (i.e., imitate Van Gogh's art style).

$$\textit{Imitation Threshold}^{\,j} \triangleq \min\left\{k \in \{1, \dots, n\} : \mathbb{I}^j(\mathcal{M}_k^j(\mathbf{p}^j)) = 1\right\} \tag{1}$$

Optimal Approach. Finding the *imitation threshold* is a causal question; the optimal manner of answering this question is the counterfactual experiment Pearl (2009). For each concept Z^j that we want to measure the imitation threshold for, we create training datasets $\{\mathcal{D}_1^j, \mathcal{D}_2^j, \dots, \mathcal{D}_k^j, \dots\}$ where k denotes the number of images containing Z^j that are added to a large training dataset $\mathcal{D}$ (that initially has no images containing Z^j), and then train a model $\mathcal{M}_k^j$ on each dataset $\mathcal{D}_k^j$. Once we find a model, $\mathcal{M}_k^j$ which is able to generate images that recognizably contains the concept Z^j, but $\mathcal{M}_{k-1}^j$ cannot, we deem k as the *imitation threshold* for that concept. *However, due to the extreme costs of training even one text-to-image model (Bastian, 2022), this approach is impractical* [2] .

MIMETIC^2. Instead, we propose an approach that is tractable and estimates the imitation threshold under certain assumptions (discussed later). The key idea is to use observational data instead of training multiple models with a different number of images of a concept. Such an approach has been previously used to answer causal questions, inter alia, Pearl (2009); Lesci et al. (2024); Lyu et al. (2024). Concretely, we collect several concepts from the same domain (e.g., art styles) while ensuring that these concepts have varying image counts in the training dataset $\mathcal{D}$ of a pretrained model $\mathcal{M}$. Then, we identify the count where the model $\mathcal{M}$ starts imitating concepts, and deem this count to be the *imitation threshold* for that domain. Note that, this threshold is domain specific and not concept

[1] Prompts $\mathbf{p}$ are usually different from the training data captions, $\mathbf{y}$.

[2] Naively, it will require training $\mathcal{O}(\log m)$ models for concept Z^j (where m is the total number of available images of concept Z^j), with an estimated cost of \$10M for $m = 100,000$ (see Appendix M).

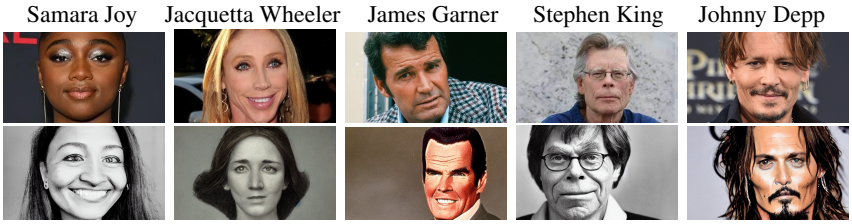


Figure 2: Examples of real celebrity images (top) and generated images from a text-to-image model (bottom) with increasing image counts from left to right (`3`, `273`, `3K`, `10K`, and `90K`, respectively). The prompt is "a photorealistic close-up image of {name}".

specific, where a *domain* is defined as a set that contains several concepts belonging to the same abstract set as the specific concept for which we want to measure the imitation threshold. E.g., if concept refers to Van Gogh's art style, then a domain would refer to a set of art styles.

To evaluate the imitation ability of a model, we build two functions, *concept-count* and *imitation-score* that counts the number of images of a concept and measures the imitation of a concept in the generated images, respectively. For a domain like art styles, we collect a set of concepts $\{Z^1, Z^2, \ldots\}$ (e.g., a list of various art styles), estimate each concept's count in training data $\mathcal{D}$, and measure the imitation score for each concept using their generated and training images. Sorting the concepts based on their increasing image counts, and using a standard *change detection* algorithm on the imitation scores, gives us the imitation threshold for that domain. We provide the implementation details in Section 5.

Assumptions. To estimate the imitation threshold using observational data, we make three assumptions. First, we assume distributional invariance between the images of all concepts in a domain. Under this assumption, measuring the imitation score of a concept Z^j for a counterfactual model $\mathcal{M}_{k'}$ that is trained with k' images of Z^j is equivalent to measuring the imitation score of another concept Z^i that currently has k' images in the already trained model $\mathcal{M}$. This assumption helps us estimate the causal effect of training a model with a different number of images of a concept, without retraining models (we empirically test the validity of this assumption in Appendix B, and we find it to hold true). Second, we assume that there are no confounders between the imitation score and the image count of a concept. These assumptions are a standard practice when answering causal questions using observational data Pearl (2009); Lesci et al. (2024); Lyu et al. (2024). Third, we assume each image of a concept contributes equally to the learning of the concept. This assumption is commonplace in sample complexity works (Valiant, 1984; Udandarao et al., 2024; Wen et al., 2024; Wang et al., 2017). These assumptions might not hold true in general, however, they are crucial for estimating the imitation threshold in a tractable manner and have been used in prior works. We also tailor our experimental setup such that these assumptions are adhered to as much as possible (§4).

4 Experimental Setup

Table 1: Domains, datasets, pretraining data, and models we experiment with.

Domain	Dataset	Pretraining Data	Model
Human Faces 🧑	Celebrities, Politicians	LAION2B-en	SD1.1, SD1.5
Human Faces 🧑	Celebrities, Politicians	LAION5B	SD2.1
Art Style 🖼️	Classical, Modern	LAION2B-en	SD1.1, SD1.5
Art Style 🖼️	Classical, Modern	LAION5B	SD2.1

Table 2: Five prompts used to generate images of human faces and art styles.

Human faces 🧑	Art style 🖼️
A photorealistic close-up photograph of X	A painting in the style of X
High-resolution close-up image of X	An artwork in the style of X
Close-up headshot of X	A sketch in the style of X
X's facial close-up	A fine art piece in the style of X
X's face portrait	An illustration in the style of X

Text-to-image Models and Training Data. We use Stable Diffusion (SD) as the text-to-image models Rombach et al. (2022). We use them because both the models and their training datasets are open-sourced. Specifically, we use SD1.1 and SD1.5 that were trained on LAION2B-en, a 2.3 billion image-caption pairs dataset, filtered to contain only English captions. In addition, we use SD2.1 that was trained on LAION-5B, a 5.85 billion image-text pairs dataset, which includes LAION2B-en, and other image-caption pairs from other languages Schuhmann et al. (2022).

Domains and Concepts. We experiment with two domains – *art styles* 🖼️ and *human faces* 🧑 that are highly important for privacy and copyright considerations of text-to-image models. Figures 1 and 2 show examples of real and generated images of art styles and human faces. We collect two sets of concepts for each domain. For art styles we collect classical and modern art styles, and for human faces, we collect celebrities and politicians. These sets are independent (i.e., have no common concept) and are therefore useful to test the robustness of the thresholds (§6). Each set has 400 concepts that cover a wide frequency range in the pretraining datasets. Appendix L provides details

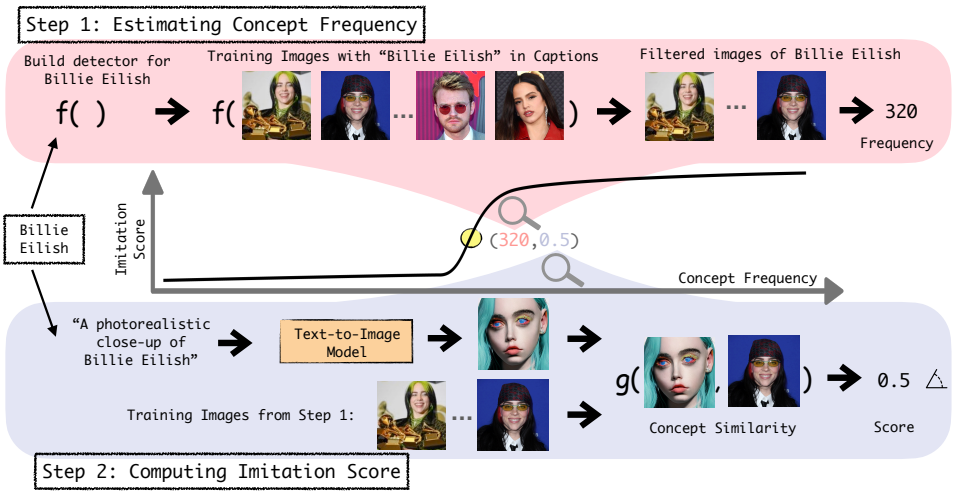


Figure 3: Overview of MIMETIC^2's methodology to estimate the *imitation threshold*. In Step 1, we estimate the frequency of each concept (belonging to a domain) in the pretraining data by obtaining the images that contain the concept of interest. In Step 2, we use the filtered images of each concept (obtained in Step 1) and compare them to the generated images to measure imitation (using g that receives training and generated images). We repeat this process for each concept to generate the imitation score graph, and then determine the *imitation threshold* with a change detection algorithm.

of the sources used to collect the concepts and the sampling procedure. Table 1 summarizes the domains, sets used for each domain, the models and their pretraining datasets we experiment with.

Image Generation. We generate images for each domain by prompting models with five prompts (Table 2). We design domain-specific prompts that encourage the concepts to occupy most of the generated image, which simplifies the imitation score measurement. We also ensure that these prompts are distinct from the captions used in the pretraining dataset to minimize reproduction of training images. (Somepalli et al. (2023b) proposed this strategy to mitigate reproduction of training images.) Out of the five prompts used to generate images, only one of them was found in the LAION2B-en dataset caption 3 times out of the total 2.3B captions. We generate 200 images per concept using different random seeds for each prompt, a total of 1,000 images per concept.

5 Proposed Methodology: MIMETIC^2

We illustrate our proposed methodology in Figure 3. At a high level, for a specific domain, MIMETIC^2 estimates the frequency of each concept in the pretraining data (Section 5.1) and estimates the model's ability to imitate it (Section 5.2). We then sort the concepts based on their frequencies, and find the imitation threshold using a change detection algorithm (Section 5.3).

5.1 Concept Frequency

Challenges. Determining a concept's frequency in a multimodal dataset can be achieved by employing a high-quality classifier for that concept over every image and counting the number of detected images. However, given the scale of modern datasets with billions of images, this approach is expensive and time consuming. Instead, we make a simplifying assumption that a concept is present only if the image's caption mentions it. We further discuss this assumption and provide supporting evidence for its accuracy in Appendix C. In addition, concepts often do not appear in the corresponding images, even when they are mentioned in the captions. For instance, Figure 4 showcases images whose captions contain "Mary Lee Pfeiffer", but such images do not always include her. On average, we find that concepts occur only in 60% of the images whose caption mentions the concept.

Figure 4: LAION images whose caption mentions 'Mary Lee Pfeiffer', the mother of Tom Cruise. She is not always present in the images (e.g., the rightmost image).

Estimating Concept Frequency We start by retrieving all images whose caption mentions the concept of interest and then further filter out the images that do not contain the concept, as detected by a classifier. We retrieve these images using WIMBD (Elazar et al., 2024), a search tool based on a reverse index that efficiently finds documents (captions) from LAION containing the search query (concept). In addition, for each concept, we construct a set of high quality reference images. For example,

a set of images with only the face of a single person (e.g., Brad Pitt). We collect these images automatically using a search engine, followed by a manual verification to vet the images (see Appendix D for details). These images are used as gold reference for automatic detection of these concepts in the images from the pretraining datasets (we collect up to ten reference images per concept).

To classify whether a candidate image from the pretraining data contains the concept of interest, we embed the candidate image and the concept's reference images using an image encoder and measure the similarity between the embeddings. We use a face embedding model (Deng et al., 2022) for faces and an art style embedding model (Somepalli et al., 2024) for art style. If the similarity between a candidate image and any of the reference images is above some threshold, we consider that candidate image to contain the concept. This threshold is established by measuring the similarity between images of the same concepts and images of different concepts which maximizes the true positives, and minimizes false positives. We provide additional details on how we find the exact thresholds for both art styles and human faces in Appendices E and F.

Finally, we employ the classifier on all candidate images corresponding to a concept (i.e., images whose caption mentions the concept), and take those that are classified as positive. For each concept, we randomly retrieve up to 100K candidate images. We use the ratio of positive predictions to the total number of retrieved candidate images, and multiply it by the total caption count of a concept in the dataset and use that as the concept's frequency estimate. For concepts with less than 100K candidate images, we simply use all the images that are positively classified. Note that several URLs in the LAION datasets are dead, a common phenomenon for URL based datasets ("link rot" Carlini et al. (2024); Lakic et al. (2023)). On average, we successfully retrieved 74% of the candidate images.

5.2 Computing Imitation Score

Challenges. Computing the imitation score entails determining how similar a concept is in a generated image compared to its source images from the training data. Several approaches were proposed to accomplish this task, such as FID and CLIPScore Salimans et al. (2016); Heusel et al. (2017); Sajjadi et al. (2018); Hessel et al. (2021); Podell et al. (2024). To measure similarity, these approaches compute the similarity between the distributions of the embeddings of the generated and training images of a concept. The embeddings are obtained using image embedders like Inception model in case of FID and CLIP in case of CLIPScore. These image embedders often perform reasonably well in measuring similarity between images of common objects which constitutes most of their training data. However, they cannot reliably measure the similarity between two very similar concepts like the faces of two individuals or art style of two artists (Somepalli et al., 2024; Hessel et al., 2021; Ahmadi & Agrawal, 2024; Jayasumana et al., 2024). Therefore, MIMETIC^2 uses domain specific image embedders to measure similarity between two concepts from that domain. It uses a face embedding model (Deng et al., 2022) for measuring face similarity and an art style embedding model Somepalli et al. (2024) for measuring art style similarity. Even the specific choice of these models is crucial. For instance, in early experiments we used Facenet (Schroff et al., 2015), and observe it struggles to distinguish between individuals of certain demographics, causing drastic differences in the imitation scores between demographics. We provide more details on these early experiments in Appendix K, and show that our final choice of embedding models work well on different demographics.

Estimating Imitation Score To measure the imitation score we embed the generated images and filtered training images of a concept (obtained in §5.1) using the domain specific image embedder. For measuring face imitation, we use InsightFace, a face embedding model (Deng et al., 2020) that extracts the individual's face from an image and generates an embedding for it. For measuring art style imitation, we use CSD, an art style embedding model (Somepalli et al., 2024) that generates an embedding for an image of an art work. We obtain the embeddings of the generated and training images for each concept, and measure imitation by computing the cosine similarity between them.

To ensure that the automatic measure of similarity correlates with human perception, we also conduct experiments with human subjects and measure the correlation between the similarities obtained automatically and in the human subject experiments. We find high correlation between the automatically obtained scores and human perception of similarity for both domains (§6).

5.3 Detecting the *Imitation Threshold*

After computing the frequencies and the imitation scores for each concept, we sort them in an ascending order of their image frequencies. This generates a sequence of points, each of which is a pair of image frequency and the imitation score of a concept. We apply a standard change

Table 3: *Imitation Thresholds* for human face and art style imitation for the different text-to-image models and pretraining datasets we experiment with.

		Human Faces 🧑		Art Style 🖼️	
Pretraining Dataset	**Model**	**Celebrities**	**Politicians**	**Classical**	**Modern**
LAION2B-en	SD1.1	364	234	112	198
	SD1.5	364	234	112	198
LAION-5B	SD2.1	527	369	185	241

detection algorithm, PELT (Killick et al., 2012), to find the image frequency where the imitation score significantly changes. Change detection is a classic statistical approach for which the objective is to find the points where the mean value of a sequence changes significantly. It is used for detecting changes such as shifts in stock market trends or network traffic anomalies. Several algorithms were proposed for change detection (van den Burg & Williams, 2022). We choose PELT because of its linear time complexity in computing the change point. We choose the first change point as the imitation threshold (see Appendix H for details about all change points). The application of change detection assumes that increasing the image counts beyond a certain threshold leads to a large jump in the imitation scores, and we find this assumption to be accurate in our experimental results.

6 RESULTS: THE *Imitation Threshold*

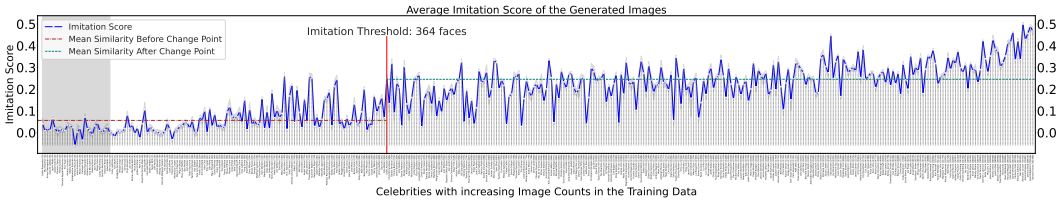
We apply MIMETIC^2 to estimate the imitation threshold for each model-data pair, and present the results in Table 3. The imitation thresholds for SD1.1 on celebrities and politicians are 364 and 234 respectively. And the imitation thresholds for classical and modern art styles are 112 and 198 respectively. Interestingly, SD1.1 and SD1.5 have the same thresholds for all the four sets across both the domains. Notably, both SD1.1 and SD1.5 are trained on LAION2B-en. The imitation thresholds for SD2.1, which is trained on the larger LAION-5B dataset is higher than the thresholds for SD1.1 and SD1.5. The imitation threshold for SD2.1 on celebrities and politicians are 527 and 369 respectively, and on classical and modern art styles are 185 and 241 respectively. We hypothesize that the difference in performance of SD2.1 and SD1.5 is due to the difference in their text encoders (O'Connor, 2022) (note: differences in performance between SD1.5 and SD2.1 were also reported by several users on online forums). To test this hypothesis, we compute the imitation thresholds for politicians for all SD models in series 1: SD1.1, SD1.2, SD1.3, SD1.4, and SD1.5. We find that the imitation thresholds for all these models are almost the same. We provide the imitation thresholds for all series 1 models in Appendix G.

Note that celebrities have a higher imitation threshold than politicians. We hypothesize this happens due to inherent differences in the data distribution between these two sets, which makes it harder to learn the concept of celebrities than politicians. To test this hypothesis, we compute the average number of images that has one person for people with less than 1,000 images in the pretraining dataset. We find that politicians have about twice the number of single person images compared to celebrities. As such, images that have a single person (the concept of interest) increase the ability of the model to learn from them, thus lowering the imitation threshold for politicians.

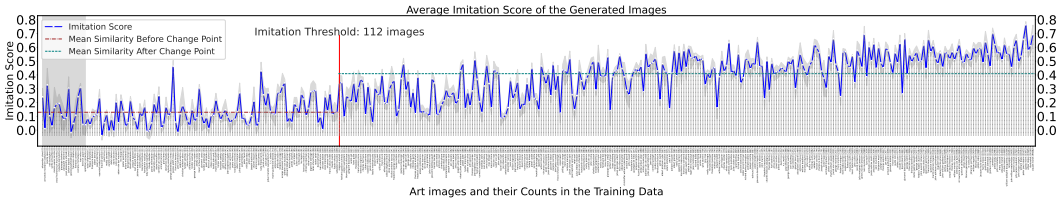
We also present the plots of the imitation scores as a function of the image frequencies of the concepts in the four sets we experiment with. Figures 5a and 5b show the imitation graphs of celebrities and classical art styles, respectively for SD1.1. The x-axis is the frequency and the y-axis is the imitation score (averaged over the five image generation prompts) for each concept, sorted in increasing order of frequency. We showcase the imitation graphs for all other models and sets in Appendix I, which follow similar trends.

In Figure 5a, we observe that the imitation scores for individuals with low image frequencies are close to 0 (left side), and increase as the image frequencies increase (towards the right side). The highest similarity is 0.5 and it is for individuals in the rightmost region of the plot. We observe a low variance in the imitation scores across prompts, and also note that this variance does not depend on the image frequencies (0.0003 ± 0.0005) – indicating that the performance of the face embedding model does not depend on the popularity of the individual.

Similarly, in Figure 5b, we observe that imitation scores for art styles with low image frequencies are low (close to 0.2 on the left side), and increase as the image frequencies increase (towards the right side). The highest similarity is 0.76 and it is for the artists in the rightmost region of the plot. We also observe a low variance across the generation prompts, and the variance does not depend on the popularity of the artist (0.003 ± 0.003).



(a) **Human Face Imitation.** The imitation threshold is detected at **364 faces**.



(b) **Art Style Imitation.** The imitation threshold is detected at **112 images**.

Figure 5: **Human Face** and **Art Style** imitation graphs for SD1.1 using the *Celebrities* and *Classical art style* sets. The x-axis represents the sorted image frequencies in the training dataset, and the y-axis represents the imitation of the training images in the generated images, for each concept. Concepts with zero image frequencies are shaded with light gray. We show the mean imitation score and its variance over the five image generation prompts. The red vertical line indicates the imitation threshold found by the change detection algorithm, and the horizontal green line represents the average imitation scores before and after the threshold.

Human Perception Evaluation. To determine if the automatic measure of similarity between the generated and training images correlate with human perception, we conduct experiments with human subjects. We asked the participants to rate generated images on the Likert scale Likert (1932) of 1-5 based on their similarity to real images of celebrities (the same ones used for measuring the imitation score). To avoid any bias, the participants were not informed of the research objective of this work.

For human face imitation, we conduct this study with 15 participants who were asked to rate 100 (randomly selected) generated images for a set of 40 celebrities. To determine the accuracy of the imitation threshold estimated by MIMETIC^2, we select the celebrities such that half of them have image frequencies below the threshold and the other half above it. We measure the Spearman's correlation (Zar, 2005) between the imitation scores computed by the model and the ratings provided by the participants. Due to the variance in perception, we normalize the ratings for each participants. The Spearman correlation between the similarity scores provided by participants and the imitation scores is **0.85**, signifying a high quality imitation estimator. We also measure the agreement between the imitation threshold that MIMETIC^2 estimates and the threshold that humans perceive. For this purpose, we convert the human ratings to binary values and treat it as the ground truth (any rating of 3 or more is treated as 1 and less than 3 is treated as 0). As for the MIMETIC^2' predictions, we construct another set of the same size that has a zero for a celebrity whose concept frequency is lower than the imitation threshold, and 1 otherwise. To measure the agreement, we compute the element-wise dot product between these two sets. We find the agreement to be 82.5%, signifying a high degree of agreement for MIMETIC^2's automatically computed threshold.

For art style imitation, we conduct this study with an art expert due to the complexity of detecting art styles. The participant was asked to rate five generated images for 20 art styles, half of which were below the imitation threshold and the other half, above the threshold. We find the Spearman correlation between the two quantities to be **0.91** – demonstrating that our imitation scores are highly correlated with an artist's perception of style similarity. Similar to the previous case, we measure the agreement of the imitation threshold, which we find to be 95% – signifying a high degree of agreement for MIMETIC^2's computed threshold.

Results Discussion. Overall, we observe that the imitation thresholds are similar across the different image generation models and pretraining datasets, but are domain dependent. They also show little variance across various image generation prompts. Most importantly, the thresholds computed by MIMETIC^2 have a high degree of agreement with human subjects.

We also note the presence of several outliers in both plots, which can be categorized into two types: (1) concepts whose image frequencies are lower than the imitation threshold, but their imitation scores are considerably high; and (2) concepts whose image frequencies are higher than the imitation threshold, but their imitation scores are low. From a privacy perspective, the first kind of outliers are more concerning than the second ones, since the imitation threshold should act as guarantors of privacy. It would be fine if a concept with a concept frequency higher than the threshold is not

(a) **Outlier Kind 1.** *Thandiwe Newton* is also aliased as *Thandie Newton*. Since MIMETIC^2 only collects images whose caption mentions *Thandiwe Newton*, this leads to underestimation of image counts.

(b) **Outlier Kind 2.** Most of the images whose captions mention *Cacee Cobb* have multiple people in them, only 6 images have her as the only person, leading to a low imitation score in generated images.

Figure 6: Examples of the two kinds of outliers. The top and bottom rows show the real and SD1.1 generated images respectively. Images were generated using the prompt: "a photorealistic close-up image of *name*."

imitated by the model (false positive), but it would be a privacy violation if a model can imitate a concept with frequency lower than the threshold (false negative). Therefore, to minimize false negatives, it is preferable to underestimate the imitation threshold. Upon further analysis, we find that the actual concept frequencies of *all the false negative outliers* is much higher than what MIMETIC^2 counts, primarily due to aliases of names, thereby alleviating the privacy violation concerns (see §7).

We also note that the range of the imitation scores of different domains have different y-axis scales. This is due to the difference in embedding models used in both cases. The face embedding model can distinguish between two faces much better than the art style model can distinguish between two styles (see Appendices E and F), and therefore the scores for the concepts on the left side of the imitation threshold is around 0 for face imitation and 0.2 for style imitation. The face embedding model also gives lower score to the faces of the same person, compared to the style embedding model's score for images of the same art styles, and therefore the highest scores for face imitation is 0.5, whereas it is 0.76 for art style imitation. However, the absolute values on the y-axis do not matter for estimating the imitation threshold as long as the trend is similar, which is the case for both domains.

Evaluation of the Imitation Thresholds. Evaluating the accuracy of the imitation thresholds MIMETIC^2 finds is a very expensive task, it requires conducting the optimal experiment described in Section 3 (training $\mathcal{O}(\log m)$ models with different number of images of each concept). Therefore we choose an approximate method to test MIMETIC^2's accuracy. Concretely, we calibrate MIMETIC^2 using one of the sets in each domain (Celebrities for Human Faces 🧑 and Classical art styles for Art Style 🖼️) and test the calibrated approach on the other set (Politicians for Human Faces 🧑 and Modern art styles for Art Style 🖼️). In essence, the sets that we use to calibrate MIMETIC^2 on, acts as the "train datasets" and the other sets acts as the "test dataset". Note that these sets are mutually exclusive. Table 3 shows that the imitation thresholds MIMETIC^2 finds for "train" and "test" sets in each domain are very close (the difference across the sets for both human faces and art styles is less than 5e-10% of the pretraining dataset size), indicating that the imitation thresholds MIMETIC^2 finds are accurate. *Note that this is another evidence for* MIMETIC^2*'s accuracy in addition to the evidence already provided by the human subject experiments.*

7 Analysis: Investigating Outliers

The imitation score plots in the previous section, while showcasing a clear trend, have several outliers. In this section, we analyze such outliers and present examples in Figure 6 (additional outliers can be found in Figures 29 and 30 in the Appendix).

Low Image Counts and High Imitation Scores. Figure 6a shows an example of such a case: *Thandiwe Newton*'s image count is 172 in LAION2B-en, lower than the *imitation threshold* for celebrities: 364. However, her imitation score of 0.26 is much higher than those of neighboring celebrities with similar image counts (with scores of 0.01 and 0.04). Further investigation reveals that Thandiwe Newton is also known as *Thandie Newton*. Since this alias may also be used to describe her in captions, MIMETIC^2 may have underestimated her image counts. We repeat the process for estimating the image counts with the new alias, and find that *Thandie Newton* appears in 12,177 images, bringing the cumulative image count to 12,349, which significantly surpasses the established imitation threshold. The two aliases, whose total image count is considerably higher than the imitation threshold, differ by only a single letter and are similarly represented by the model's encoder (cosine similarity of 0.96), which explains the high imitation score. We find that most of the celebrities from the first kind of outliers are also known by other names which lead to underestimating their image counts. For example, *Belle Delphine* (394 images) also goes by *Mary Belle* (310 images, for a total of 704, and *DJ Kool Herc* (492 images) also goes by *Kool Herc* (269 images, for a total of 761).

The aliases explanation also largely explains the outliers in art style imitation. For instance, artist *Gustav Adolf Mossa* (19 images) also goes by just *Mossa* (15850 images), artist *Nicolas Toussaint Charlet* (78 images) also goes by just *Nicolas Toussaint* (533 images), and artist *Wilhelm Von Kaulbach* (81 images) also goes by *Von Kaulbach* (978 images). See Figures 31 and 32 in the appendix for the real and generated images of these artists.

High Image Counts and Low Imitation Scores Several celebrities have higher image counts than the imitation threshold, but low imitation scores. Unlike the previous case, we were unable to find a common cause that explains all these outliers. However, we find explanations for specific cases. For example, a staggering proportion of the training images for such celebrities have multiple people in them. For example, out of the 706 total images of *Cacee Cobb*, only 6 images have her as the only person in the image (see Figure 6b). Similarly, out of 1,296 total images of *Sofia Hellqvist*, only 67 images have her as the only person and out of the 472 total images of *Charli D' Amelio*, only 82 images have her as the only person. We hypothesize that having multiple concepts in an image impedes the proper mapping of the concept's text embedding to its image embedding, which can explain the low imitation score for these concepts. We leave it to future work to further study the connection between the number of concepts in an image and model's ability to imitate these concepts.

8 Discussion and Limitations

Equal Effect Assumption. An assumption in the formulation of MIMETIC^2 is that every image of a concept contributes equally to the learning of the concept. Although this is a standard assumption in sample complexity works (Valiant, 1984; Udandarao et al., 2024; Wen et al., 2024; Wang et al., 2017), not all images are created equal. While analyzing celebrities' images for instance, we often find that individuals whose images are mostly close-ups of a single person have a higher imitation score than individuals whose images are cluttered by multiple people, since concept-centered images enhance their learnability. We hope to investigate this assumption in future work.

Factors Affecting the Imitation Threshold. In this work we attribute the imitation of a concept to its image count. However, image count – although a crucial factor – is not the only factor that affects imitation. Several other factors like image resolution, image diversity, alignment between images and their captions, the variance between images of a concept may affect imitation.

Several training time factors like the optimization objective, learning schedule, training data order, model capacity, model architecture also affect the imitation threshold. We discuss the difference in the imitation thresholds of SD1.1, SD1.5 and SD2.1 is attributed to the difference in their text encoders. SD1.1 and SD1.5 use CLIP model Radford et al. (2021) as their text encoder and SD2.1 uses OpenCLIP Ilharco et al. (2021) as its text encoder. Note that while these may impact the behavior of the model, our work is interested in a particular model-data pair, for which we investigate. While we conduct experiments for most models whose training data is available, we do not claim that our results would generalize to other models or datasets, and hope that in future more such models and datasets are open-sourced, so that we can test the generalizability of our imitation thresholds. Note that, MIMETIC^2 finds the imitation threshold for the pretraining regime. We hope to investigate the thresholds when a model is finetuned on one or a set of concepts in future work.

9 Conclusions

Text-to-image models can imitate their training images (Somepalli et al., 2023a; Carlini et al., 2023a; Somepalli et al., 2023b). This behavior is potentially concerning because these models' training datasets often include copyrighted and licensed images. Imitating such images would be grounds for violation of copyright and privacy laws. In this work, we seek to find the number of instances of a concept that a text-to-image model needs in order to imitate it – the *imitation threshold*. We posit this as a new problem, **F**inding the **I**mitation **T**hreshold (`FIT`) and propose an efficient method, MIMETIC^2, for finding such thresholds. Our method utilizes pretrained models to estimate this threshold for human faces and art styles imitation using three text-to-image models trained on two different pretraining datasets. We find the imitation threshold of these models to be in the range of 200-600 images depending on the setup. As such, on the domains we evaluate in this work, our results indicate that models cannot imitate concepts that appear less than 200 times in the training data. By estimating the imitation threshold, we provide insights on successful concept imitation based on their training frequencies. Our results have striking implications on both the text-to-image models users and developers. These thresholds can inform text-to-image model developers what concepts are in risk of being imitated, and on the other hand, serve as a basis for copyright and privacy complaints.

10 Ethics Statement

This work is targeted towards understanding and mitigating the negative social impacts of widely available text-to-image models. To minimize the impact of our study on potential copyright infringement and privacy violations, we used existing text-to-image models and generation techniques. To strike a balance between reproducibility and ethics, we release the code used in this work, but not the real and synthetic images used (we also provide detailed documentation of our data curation process in the code to ensure transparency).

Human Subject Experiments For the human subject experiments performed in this work, we followed the university guideline. The study was IRB approved by the university and each participant was made aware of the goals and risks of the study. In accordance with local minimum wage requirements, we compensated each participant with Amazon gift cards of $25/hour for their time.

11 Reproducibility Statement

We provide our code in the supplementary material. The code includes a detailed `README.md` file that provides instructions to execute each step of the algorithm. We will open-source all the code upon publication of the paper. We do not provide the original and generated images of individuals and art styles for copyright and privacy concerns.

References

Saba Ahmadi and Aishwarya Agrawal. An examination of the robustness of reference-free image captioning evaluation metrics, 2024. URL https://arxiv.org/abs/2305.14998.

Nadeem Badshah. Nearly 4,000 celebrities found to be victims of deepfake pornography. https://www.theguardian.com/technology/2024/mar/21/celebrities-victims-of-deepfake-pornography, 2024. Accessed: 2024-04-27.

Matthias Bastian. Training cost for stable diffusion was just $600,000, 2022. URL https://the-decoder.com/training-cost-for-stable-diffusion-was-just-600000-and-that-is-a-good-sign-for-ai-

Jeff Bilmes. Submodularity in machine learning and artificial intelligence, 2022. URL https://arxiv.org/abs/2202.00132.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes, 2021. URL https://arxiv.org/abs/2110.01963.

N. Carlini, M. Jagielski, C. Choquette-Choo, D. Paleka, W. Pearce, H. Anderson, A. Terzis, K. Thomas, and F. Tramèr. Poisoning web-scale training datasets is practical. In *2024 IEEE Symposium on Security and Privacy (SP)*, Los Alamitos, CA, USA, may 2024. IEEE Computer Society. URL https://doi.ieeecomputersociety.org/10.1109/SP54263.2024.00179.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23, USA, 2023a. USENIX Association. URL https://arxiv.org/abs/2301.13188.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=TatRHT_1cK.

Stephen Casper, Zifan Guo, Shreya Mogulothu, Zachary Marinov, Chinmay Deshpande, Rui-Jie Yew, Zheng Dai, and Dylan Hadfield-Menell. Measuring the success of diffusion models at imitating human artists, 2023. URL https://arxiv.org/abs/2307.04028.

Michael Cavna. Ai in illustration. https://www.washingtonpost.com/comics/2023/02/14/ai-in-illustration/, 2023. Accessed: 2024-02-27.

Ceoln. Posts tagged 'copyright', 2022. URL https://ceoln.wordpress.com/tag/copyright/.

Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *Computer Vision – ECCV 2020*, pp. 741–757, Cham, 2020. Springer International Publishing. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123560715.pdf.

Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, October 2022. URL http://dx.doi.org/10.1109/TPAMI.2021.3087709.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=AAWuCvzaVt.

Yanai Elazar, Akshita Bhagia, Ian Helgi Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hanna Hajishirzi, Noah A. Smith, and Jesse Dodge. What's in my big data? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.20707.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024. URL https://unified.baulab.info/.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1–7, 2022. URL https://arxiv.org/abs/1406.2661.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-main.595.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2017. URL https://dl.acm.org/doi/10.5555/3295222.3295408.

Tatum Hunter. Ai porn is easy to make now. for women, that's a nightmare. https://www.washingtonpost.com/technology/2023/02/13/ai-porn-deepfakes-women-consent/, 2023. Accessed: 2024-02-27.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9307–9315, 2024. URL https://arxiv.org/abs/2401.09603.

Harry H. Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. Ai art and its impact on artists. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, pp. 363–374, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702310. URL https://doi.org/10.1145/3600211.3604681.

R. Killick, P. Fearnhead, and I. A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, pp. 1590–1598, 2012. URL http://www.jstor.org/stable/23427357.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *International Conference on Computer Vision (ICCV)*, 2023. URL https://arxiv.org/abs/2303.13516.

Viktor Lakic, Luca Rossetto, and Abraham Bernstein. Link-rot in web-sourced multimedia datasets. In *International Conference on Multimedia Modeling*, pp. 476–488. Springer, 2023. URL https://www.zora.uzh.ch/id/eprint/232667/.

Tommaso Lanciano, Atsushi Miyauchi, Adriano Fazzone, and Francesco Bonchi. A survey on the densest subgraph problem and its variants. *ACM Computing Surveys*, 56(8):1–40, 2024. doi: 10.1145/3653298. URL https://dl.acm.org/doi/full/10.1145/3653298.

Pietro Lesci, Clara Meister, Thomas Hofmann, Andreas Vlachos, and Tiago Pimentel. Causal estimation of memorisation profiles. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 15616–15635. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.acl-long.834.

Rensis Likert. A technique for the measurement of attitudes., 1932. URL https://psycnet.apa.org/record/1933-01885-001.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV 2014*, pp. 740–755. Springer, 2014. URL https://cocodataset.org/.

Jack Lu, Ryan Teehan, and Mengye Ren. Procreate, don't reproduce! propulsive energy diffusion for creative generation, 2024. URL https://arxiv.org/abs/2408.02226.

Zhiheng Lyu, Zhijing Jin, Fernando Gonzalez, Rada Mihalcea, Bernhard Schoelkopf, and Mrinmaya Sachan. On the causal nature of sentiment analysis. *arXiv preprint arXiv:2404.11055*, 2024.

Kiyohito Nagano, Yoshinobu Kawahara, and Kazuyuki Aihara. Size-constrained submodular minimization through minimum norm base. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 977–984, 2011. URL https://dl.acm.org/doi/10.5555/3104482.3104605.

NIST. Face recognition vendor test (frvt). https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt, 2020. Accessed: 2024-02-27.

Ryan O'Connor. Stable diffusion 1 vs 2 - what you need to know. https://www.assemblyai.com/blog/stable-diffusion-1-vs-2-what-you-need-to-know/, 2022. Accessed: 2024-05-14.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://arxiv.org/abs/2103.00020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8821–8831. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/ramesh21a.html.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 840–854. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.findings-emnlp.59.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/f7696a9b362ac5a51c3dc8f098b73923-Paper.pdf.

Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *International Journal for Digital Art History*, 2, Oct. 2016. doi: 10.11588/dah.2016.2.23376. URL https://journals.ub.uni-heidelberg.de/index.php/dah/article/view/23376.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf.

Joseph Saveri and Matthew Butterick. Stable diffusion litigation. https://stablediffusionlitigation.com/case-updates.html, 2023. Accessed: 2024-02-27.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015. doi: 10.1109/cvpr.2015.7298682. URL http://dx.doi.org/10.1109/CVPR.2015.7298682.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models, 2022. URL https://openreview.net/forum?id=M3Y74vmsMcY.

Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pp. 23–27. IEEE, 2020. URL https://doi.org/10.1109/ASYU50717.2020.9259802.

SerpApi. https://serpapi.com/google-images-api, 2024. Accessed: 2024-02-27.

Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y. Zhao. Fawkes: protecting privacy against unauthorized deep learning models. In *Proceedings of the 29th USENIX Conference on Security Symposium*, SEC'20, USA, 2020. USENIX Association. URL https://dl.acm.org/doi/10.5555/3489212.3489302.

Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. Glaze: protecting artists from style mimicry by text-to-image models. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23, USA, 2023. USENIX Association. URL https://dl.acm.org/doi/10.5555/3620237.3620360.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp.

2256–2265, Lille, France, 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6048–6058, 2023a. URL https://openaccess.thecvf.com/content/CVPR2023/supplemental/Somepalli_Diffusion_Art_or_CVPR_2023_supplemental.pdf.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b. URL https://openreview.net/forum?id=HtMXRGbUMt.

Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models, 2024. URL https://arxiv.org/abs/2404.01292.

David Thiel. Identifying and eliminating CSAM in generative ML training data and models, 2023. URL https://purl.stanford.edu/kh752sm9123.

Vishaal Udandarao, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip H. S. Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance, 2024. URL https://arxiv.org/abs/2404.04125.

L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27, 1984. doi: 10.1145/1968.1972. URL https://doi.org/10.1145/1968.1972.

Gerrit J. J. van den Burg and Christopher K. I. Williams. An evaluation of change point detection algorithms, 2022. URL https://arxiv.org/abs/2003.06222.

James Vincent. Getty images sues ai art generator stable diffusion. https://www.theverge.com/2023/2/6/23587393/ai-art-copyright-lawsuit-getty-images-stable-diffusion, 2023. Accessed: 2024-02-27.

Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/147ebe637038ca50a1265abac8dea181-Paper.pdf.

Zhenting Wang, Chen Chen, Lingjuan Lyu, Dimitris N. Metaxas, and Shiqing Ma. DIAGNOSIS: Detecting unauthorized data usages in text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=f8S3aLm0Vp.

Xin Wen, Bingchen Zhao, Yilun Chen, Jiangmiao Pang, and Xiaojuan Qi. Generalization beyond data imbalance: A controlled study on clip for transferable insights, 2024. URL https://arxiv.org/abs/2405.21070.

Wikipedia. Lists of politicians. https://en.wikipedia.org/wiki/Category:Lists_of_politicians, 2024. Accessed: 2024-02-27.

Tong Xie, Haoyu Li, Andrew Bai, and Cho-Jui Hsieh. Data attribution for diffusion models: Timestep-induced bias in influence estimation. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=P3Lyun7CZs.

Jerrold H Zar. Spearman rank correlation. *Encyclopedia of Biostatistics*, 7, 2005. URL https://onlinelibrary.wiley.com/doi/10.1002/0470011815.b2a15150.

A Additional Related Work

Imitation in Text-to-Image Models: Somepalli et al. (2023a); Carlini et al. (2023a) demonstrated that diffusion models can memorize and imitate duplicate images from their training data (they use 'replication' to refer to this phenomenon). Casper et al. (2023) corroborated the evidence by showing that these models imitated art styles of 70 artists with high accuracy (as classified by a CLIP model) when prompted to generated images in their styles (a group of artists also sued Stability AI claiming that their widely-used text-to-image models imitated their art style, violating copyright laws Saveri & Butterick (2023)). However, these works did not study how much repetition of a concept's images would lead the model to imitate them. Studying this relation is important as it serves to guide institutions training these models who want to comply with copyright and privacy laws.

Mitigation of Imitation in Text-to-Image Models: Several works proposed to mitigate the negative impacts of text-to-image models. Shan et al. (2023) proposed GLAZE that adds imperceptible noise to the art works such that diffusion models are unable to imitate artist styles. A similar approach was proposed to hinder learning human faces (Shan et al., 2020). Wang et al. (2024) proposed adding noise to training images, which can be used to detect if a model has been trained on those images. Lu et al. (2024) propose pushing the generated images away from the distribution of training images to minimize mitigation. Kumari et al. (2023); Gandikota et al. (2024) proposed algorithms to remove specific styles, explicit content, and other copyrighted material learned by text-to-image models. On a related note, Xie et al. (2024) proposed Diffusion-ReTrac that finds training images that most influenced a generated image, and thereby provide a fair attribution to training data contributors.

B Validity of the Assumptions

Table 4: Average difference in the imitation scores for concepts whose image counts differ by less than 10. The difference in the imitation scores are very close to 0, providing the empirical validation the distribution invariance assumption.

Domain	Dataset	Avg. difference in imitation score
Human Faces 🧑	Celebrities	0.0007
Human Faces 🧑	Politicians	0.0023
Art Style 🖼️	Classical Art Style	-0.0088
Art Style 🖼️	Modern Art Style	-0.0013

In this section, we empirically test the statistical validity of the assumptions we make in Section 3. The first assumption states that there is a distributional invariance across concepts. If this is true, then the imitation scores of two concepts (from the same domain) whose image counts are similar, should also be similar to each other. To test whether this is empirically true for the domains we experiments with, we measure the difference in the imitation scores for concepts whose image counts differ by less than 10 images and report the difference averaged over all such pairs. Table 4 shows that the average difference in the imitation scores for all the pairs, for all the datasets we experiment with, is very close to 0. This provides empirical validation of the distribution invariance assumption.

C Caption Occurrence Assumption

For estimating the concept's counts in the pretraining dataset we make a simplifying assumption: a concept can be present in the image only if it is mentioned in a paired caption. While this assumption isn't true in general, we show that for the domains we experiment with, it mostly holds in practice.

For this purpose, we download 100K random images from LAION2B-en, and run the face detection (used in Section 5) on all images, and count the faces of the ten most popular celebrities in our sampled set of celebrities. Out of the 100K random images, about 57K contain faces. For each celebrity, we compute the similarity between all the faces in the downloaded images and the faces in the reference images of these celebrities. If the similarity is above the threshold of 0.46, we consider that face to belong to the celebrity (this threshold is determined in Appendix E to distinguish if two images are of the same person or not). Table 5 shows the number of faces we found for each celebrity in the 100K random LAION images. We also show the face counts among these images whose captions mention the celebrity. We find that 1) the highest frequency an individual appears in

Table 5: Face count of the ten most popular celebrities in 100K random LAION images. The small percentage of the images we miss shows that our assumption of counting the images where a concept is mentioned in the caption is empirically reasonable.

Celebrity	Face Count in 100K images	Face Count in Images with Caption Mention	Percentage of Missed Images	Number of Missed Images
Floyd Mayweather	1	0	0.001%	23K
Oprah Winfrey	2	0	0.002%	46K
Ronald Reagan	6	3	0.003%	69K
Ben Affleck	0	0	0.0%	0
Anne Hathaway	0	0	0.0%	0
Stephen King	0	0	0.0%	0
Johnny Depp	9	1	0.008%	184K
Abraham Lincoln	52	1	0.051%	1.17M
Kate Middleton	34	1	0.033%	759K
Donald Trump	16	0	0.016%	368K

an image without their name mentioned in the caption is 51 (*Abraham Lincoln* is mentioned once in the caption and he appears a total of 52 times), and 2) the highest percentage of image frequency that we miss is 0.051%, and 3) most of the other miss rates are much smaller (close to 0). Such low miss rates demonstrate that our assumption of counting images when a concept is mentioned in the caption is empirically reasonable.

We also note that this assumption would fail if we were computing image frequencies for concepts that are so widely common that one would not even mention them in a caption, for example, phone, shoes, or trees.

D Collection of Reference Images

D.1 Collection of Reference Images for Human Faces Domain

The goal of collecting reference images is to use them to filter the images of the pretraining dataset. These images are treated as the gold standard reference images of a person and images collected from pretraining dataset are compared to these images. If the similarity is higher than a threshold then that image is considered to belong to that person (see Section 5 for details). We describe an automatic manner of collecting the reference images. The high level idea is to collect the images from Google Search and automatically select a subset of those images that are of the same concept (same person's face or same artist's art). Since this is a crucial part of the overall algorithm, we manually vet the reference images for all the concepts to ensure that they all contain the same concept.

Collection of Reference Images for Human Face Imitation: We collect reference images for celebrities and politicians using a three step process (also shown in Algorithm 1):

1. **Candidate set:** First, we retrieved the first hundred images by searching a person's name on Google Images. We used SerpAPI SerpApi (2024) as a wrapper perform the searches.
2. **Selecting from the candidate set:** Images retrieved from the internet are noisy and might not contain the person we are looking for. Therefore we filter images that contain the person from the candidate set of images. For this purpose, we use a face recognition model. We embed all the faces in the retrieved images using a face embedding model and measure the cosine similarity between each one of them. The goal is to search for a set of faces that belong to the same person and therefore will have a high cosine similarity to each other.

   One strategy is for the faces to form a graph where the vertices are the face embeddings and the edges connecting two embeddings have a weight equal to the cosine similarity between them, and we select a dense k-subgraph (Lanciano et al., 2024) from this graph. Selecting such a subgraph means finding a mutually homogeneous subset. We can find the vertices of this dense k-subgraph by cardinality-constrained submodular function minimization (Bilmes, 2022; Nagano et al., 2011) on a facility location function (Bilmes, 2022). We run this minimization and select a subset of images (at least of size ten) that has the highest average cosine similarity between each pair of images.
3. **Manual verification:** Selecting the faces with the highest average similarity is not enough. This is because in many cases the largest set of faces in the candidate set are not of the person we look

for, but for someone closely associated with them, in which case, the selected images are of the other person. For example, all the selected faces for *Miguel Bezos* were actually of *Jeff Bezos*. Therefore, we manually verify all the selected faces for each person. In the situation where the selected faces are wrong, we manually collect the images for them, for example, for Miguel Bezos. We collect at least 5 reference images for all celebrities.

Algorithm 1: Collection of Reference Images for Human Face Imitation

Input: Person's name P
Output: Verified Set of Images of P
1 $images \leftarrow$ SerpAPI(P) ; ▷ Retrieve initial image set using SerpAPI
2 $candidateSet \leftarrow$ Submodular_Minimization($images$) ; ▷ Select candidate set using submodular minimization
3 $verifiedSet \leftarrow$ manualVerification($candidateSet$) ; ▷ Manually verify the candidate set

Collection of Reference Images for Art Styles We collect reference images for each artist (each artist is assumed to have a distinct art style) from Wikiart, the online encyclopedia for art works. Since the art works of each artist were meticulously collected and vetted by the artist community, we consider all the images collected from Wikiart as the reference art images for that artist.

E Implementation Details of MIMETIC^2 for Human Face Imitation

E.1 Filtering of Training Images

Images whose captions mention the concept of interest often do not contain it (as shown with Mary-Lee Pfeiffer in Figure 4). As such, we filter images where the concept does not appear in the image, which we detect using a dedicated classifier. In what follows we describe the filtering mechanism.

Collecting Reference Images:

We collect reference images for each person using SerpAPI as described in Appendix D. These images are the gold standard images that we manually vet to ensure that they contain the target person of interest (see Appendix D for the details). We use the reference images to filter out the images in the pretraining dataset that are not of this person. Concretely, for each person we use a face embedding model Deng et al. (2020) to measure the similarity between the faces in the reference images and the faces in the images from the pretraining datasets whose captions mention this person. If the similarity of a face in the pretraining images to any of the faces in the reference images is above a certain threshold, that face is considered to belong to the person of interest. We determine this threshold to distinguish faces of the same person from faces of different persons in the next paragraph. Note that this procedure already filter outs any image that does not contain a face, because the face embedding model would only embed an image if it detects a face in that image.

Determining Filtering Threshold: The next step is to determine the threshold for which we consider two faces to belong to the same person. For this purpose, we measure the similarity between pairs of faces of the same person and the similarity between pairs of faces of different persons Since the reference images for each person is manually vetted to be correct, we use these images for this procedure. We plot the histogram of the average similarity between the faces of the same person (blue colored) and the similarity between faces of different persons (red colored) in Figure 7. We see that the two histograms are well separated, with the lowest similarity value between the faces of the same person being 0.56 and the highest similarity value between the faces of different persons being 0.36. Therefore any threshold value between 0.36 and 0.56 can separate two face of the same person, from the faces of different people. In our experiments, we use the midpoint threshold of 0.46 (true positive rate (tpr) of 100%; false positive rate (fpr) of 0%) to filter any face in the pretraining images that do not belong to the person of interest. The filtering process gives us both the image frequency a person in the pretraining data, and the pretraining images that we compare the faces in the generated images to measure the imitation score.

E.2 Measurement of Imitation Score

(a) Training images of *Samuel L. Jackson* that show significant variations in his face (age, hair, and beard).

(b) Generated images of *Samuel L. Jackson* that show the model has captured a specific characteristic of his face (middle-aged, bald, with no or little beard).

Figure 8: Real and generated images of *Samuel L. Jackson*.

To measure the imitation between the training and generated images of a person, we compute the cosine similarity between the face embeddings of the faces in their generated images and their filtered training images from the previous step. However, measuring the similarity using all the pretraining images can underestimate the actual imitation. This is because several individuals have significant variations in their faces in the pretraining images and the text-to-image model does not capture all these variations. For example, consider the pretraining images of *Samuel L. Jackson* in Figure 8a. These images have significant variations in beard, hair, and age. However, when the text-to-image model is prompted to generate images of *Samuel L. Jackson*, the generated images in Figure 8b only show a specific facial characteristic of him (middle-aged, bald, with no or little beard). Since MIMETIC^2's goal is not to measure if a text-to-image model captures all the variations of a person, we want to reward the model even if it has only captured a particular characteristic (which it has in this case of *Samuel L. Jackson*). Therefore, instead of comparing the similarity of generated images to all the training images, we compare the similarity to only the ten training images that have the highest cosine similarity to the generated images on average.

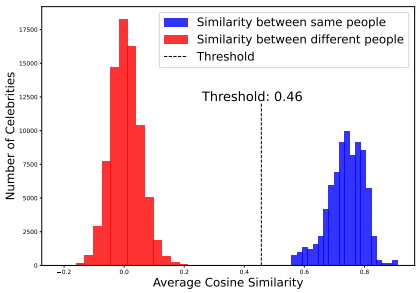


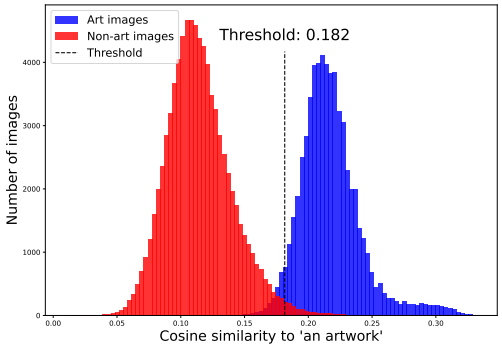
Figure 7: Average cosine similarity between the faces of the same people (blue colored) and of the faces of different people (red colored), measured across the reference images of the celebrities.

F Implementation Details of MIMETIC^2 for Art Style Imitation

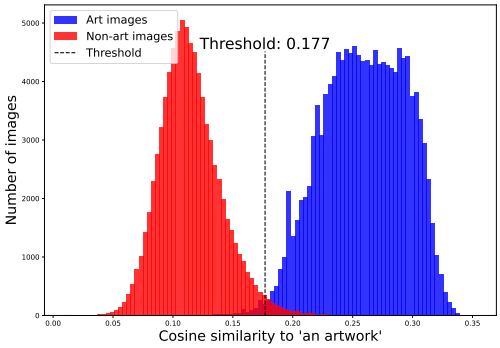
F.1 Filtering of Training Images

For art style imitation, we consider each artist to have a unique style. We collect the images from the pretraining dataset whose captions mention the name of the artist whose art style imitation we want to measure. Similar to the case of human face imitation, we want to filter out the pretraining images of an artist that in reality was not created by that artist, but their captions mention them. We implement the filtering process in two stages. In the first stage, we filter out non-art images in the pretraining dataset (note that the captions of these images still mention the artist, but the images themselves are not art works) and in the second stage we filter out art works of other artists (the captions of these images mention the artist of interest and the image itself is also an art work, but by a different artist). The implementation details for each stage is as follows:

Filtering Non-Art Images: To filter non-art images from the pretraining dataset, we use a classifier that separates art images from non-art images. Concretely, we embed the pretraining images using a CLIP ViT-H/14 Ilharco et al. (2021) image encoder and measure the cosine similarity of the image embeddings and the text embeddings of the string '*an artwork*', embedded using the text encoder of the same model. Only when the similarity between the embeddings is higher than a threshold described below, we consider those pretraining images as an artwork. To determine this threshold, we choose a similarity score that separates art images from non-art images. We use the images from the Wikiarts dataset Saleh & Elgammal (2016) as the (positive) art images and MS COCO dataset

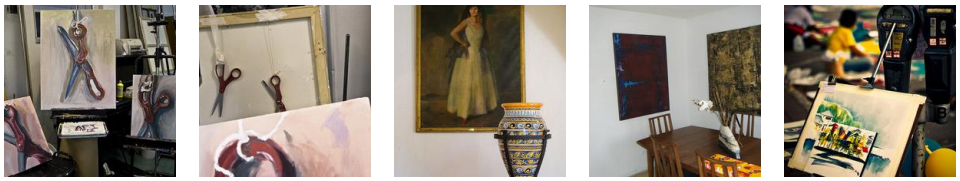


(a) Histogram of the cosine similarity of embeddings of art and non-art images to embeddings of 'an artwork' for classical artists.



(b) Histogram of the cosine similarity of embeddings of art and non-art images to embeddings of 'an artwork' for modern artists.

Figure 9: The first filtering step involves determining the threshold to distinguish between art and non-art images from the pretraining images, for which we compare the similarity of the image's embedding to the embedding of the text "an artwork".



(a) Images from the MS COCO dataset that were classified as art by the threshold we choose. These images clearly have paintings in them and therefore are classified in that category. These images were selected in MS COCO for different categories like scissors, chair, parking meter, and vase.

(b) Images from the Wikiarts dataset that were classified as non-art by the threshold we choose.

Figure 10: Images that are misclassified by our art vs. non-art threshold in Figure 9a.

images Lin et al. (2014) as the (negative) non-art images. Note that MS COCO dataset was collected by photographing everyday objects that art was not part of, making it a valid set of negative examples of art.

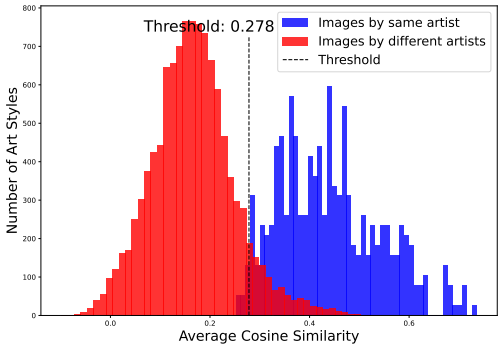
We plot the histogram of cosine similarity of the embeddings of art and non-art images to the text embedding of '*an artwork*' (see Figures 9a and 9b. We observe that the art and non-art images both the artist groups are well separated (although not perfect, Figure 10 and Figure 11 shows examples of misclassified and correctly classfied images from both datasets). We choose the threshold that maximizes the F1 score of the separation (0.182 for the classical artists and 0.177 for the modern artists).

Filtering Images of Other Art Styles: Similar to the case of human faces, not all art images whose captions mention an artist were created by that artist. We want to filter out such images. For this

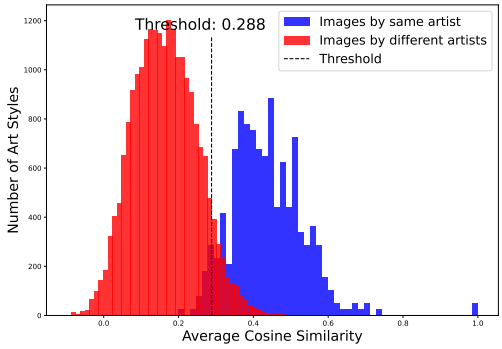
(a) Images from the MS COCO dataset that were correctly classified as non-art.

(b) Images from the Wikiarts dataset that were correctly classified as art.

Figure 11: Images that are correctly classified by our art vs. non-art threshold in Figure 9a.



(a) Histogram of the average cosine similarity between embeddings of the images of the same artist (blue) and the art of different artists (red) for classical artists



(b) Histogram of the average cosine similarity between embeddings of the images of the same artist (blue) and the art of different artists (red) for modern artists

Figure 12: The second filtering step involves determining the if an art work whose caption mentions an artist actually belongs to that artist or not.

purpose, we collect reference images for each artist (see Appendix D for details) and use them to classify the training images that belong to the artist of interest. Concretely, we measure the similarity between the pretraining images and the reference images of each artist, and only retain images whose similarity to the reference images is higher than a threshold.

To determine this threshold, we measure the similarity between pairs of art images of the same artists and pairs of art images from different artists. We embed the images using an art style embedding model Somepalli et al. (2024) and plot the histogram of similarities between art images of the same artist (blue colored) and art images of different artists (red colored) in Figure 12a for classical artists and Figure 12b for modern artists. We see that the two histograms are well separated (although not perfect, Figure 13 shows paintings by two artists whose art style is very similar and cannot be distinguished by our threshold). We choose the threshold that maximizes the F1 score of the separation between these two groups (0.278 for classical artists and 0.288 for modern artists). The retained images give us both the image counts of each artist and the training images that we compare to the generated images to measure the imitation score.

F.2 Similarity Measurement

We embed all the generated images and the filtered pretraining images using the art style embedding model (Somepalli et al., 2024) and measure the cosine similarity between each pair of generated and pretraining images. Similar to the case of the human faces, we do not want to underestimate the art style similarity between the generated and training images by comparing the generated images to all the training images of this artist. Therefore, we measure the similarity of generated images to the ten training images that are on average the most similar to the generated images.

G Imitation Thresholds of SD models in Series 1 and 2

Our experimental results in Section 6 found that for most domains the imitation thresholds for SD1.1 and SD1.5 are almost the same, while being higher for SD2.1. We hypothesized that the difference is due to their different text encoders. All models in SD1 series use the same text encoder from CLIP, whereas SD2.1 uses the text encoder from OpenCLIP. To test the validity of this hypothesis, we repeated the experiments for all models in SD1 series for politicians and computed their imitation thresholds. Table 6 shows the thresholds for the politicians. We find that the imitation thresholds for all the models in SD1 series is almost the same, and is lower than the threshold for SD2.1 model. This evidence supports our hypothesis of the difference in the text-encoders being the main reason for the difference in the imitation thresholds.

H Change Points

Table 7 we show all the change points that PELT found for each experiment (Table 3 reports the first change point as the imitation threshold).

(a) Paintings made by Kitagawa Utamaro.

(b) Paintings made by Tsukioka Yoshitoshi

Figure 13: Paintings made by Kitagawa Utamaro and Tsukioka Yoshitoshi are very similar and our threshold is unable to distinguish between their styles.

Table 6: *Imitation Thresholds* for politicians for all models in SD1 series and SD2.1

Pretraining Dataset	**Model**	**Human Faces 🧑: Politicians**
LAION2B-en	SD1.1	234
	SD1.2	252
	SD1.3	234
	SD1.4	234
	SD1.5	234
LAION-5B	SD2.1	369

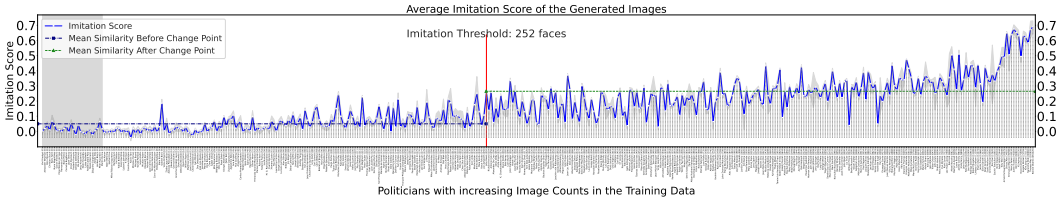


Figure 14: **Human Face Imitation (Politicians):** Similarity between the training and generated images for all politicians. The politicians with zero image counts are shaded with light gray. We show the mean and variance over the five generation prompts. The images were generated using **SD1.2**. The change point for human face imitation for politicians when generating images using SD1.1 is detected at **252 faces**.

I ALL RESULTS: THE *Imitation Threshold*

In this section, we estimate the imitation threshold for human face and art style imitation for three different text-to-image models. Figure 17, Figure 18, and Figure 19 show the image counts of celebrities on the x-axis (sorted in increasing order of image counts) and the imitation score of their generated images (averaged over the five image generation prompts) on the y-axis. The images were generated using SD1.1, SD1.5, and SD2.1 respectively. Similarity, Figure 20, Figure 21, and

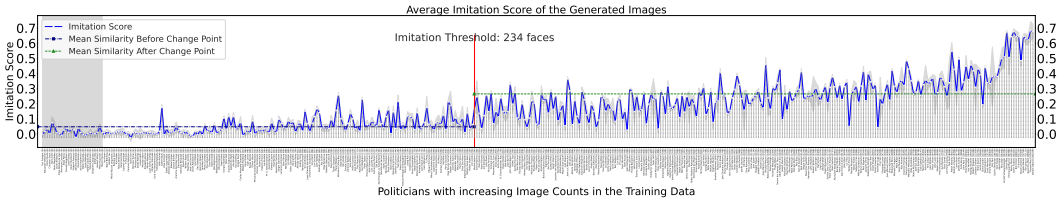


Figure 15: **Human Face Imitation (Politicians):** Similarity between the training and generated images for all politicians. The politicians with zero image counts are shaded with light gray. We show the mean and variance over the five generation prompts. The images were generated using **SD1.3**. The change point for human face imitation for politicians when generating images using SD1.1 is detected at **234 faces**.

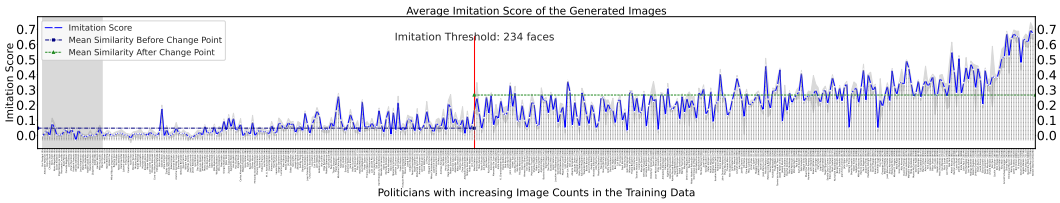


Figure 16: **Human Face Imitation (Politicians):** Similarity between the training and generated images for all politicians. The politicians with zero image counts are shaded with light gray. We show the mean and variance over the five generation prompts. The images were generated using **SD1.4**. The change point for human face imitation for politicians when generating images using SD1.1 is detected at **234 faces**.

Figure 22 shows the image counts of the politicians and the imitation score of their generated images, for SD1.1, SD1.5, and SD2.1 respectively.

Figure 23, Figure 24, Figure 25 show the image counts of classical artists and the similarity between their training and generated images; and Figure 26, Figure 27, Figure 28 show the image counts of modern artists and the similarity between their training and generated images. The images were generated using SD1.1, SD1.5, and SD2.1 respectively.

Imitation Threshold Estimation for Human Face Imitation: In Figure 17, we observe that the imitation scores for the individuals with small image counts is close to 0 (left side), and it increases as the number of their image counts increase towards the right. The highest similarity is 0.5 and it is for the individuals in the rightmost region of the plot. The solid line in the plot shows the mean similarity over the five image generation prompt with the shaded area showing the variance over them. We observe a low variance in the imitation score among the generation prompts. And we also observe that the variance does not depend on the image counts which indicates that the performance of the face recognition model does not depend on the popularity of the individual. The change detection algorithm finds the change point to be at **364** faces for human face imitation for celebrities, when using **SD1.1** for image generation. Figure 18 shows the similarity between the training and generated images when images are generated using **SD1.5**. Identically to SD1.1, the change is detected at **364** faces for face imitation when using SD1.5. We also performed ablation experiments with different face embeddings models and justify the choice of our model (see Appendix K). For all the plots, we also analyze the trend by using isotonic regression which learns non-decreasing linear regression weights that fits the data best.

Imitation Threshold Estimation for Human Face Imitation (Politicians): Figure 20 shows the imitation scores for politicians which is very similar to the plot obtained for celebrities. We observe a low variance in the imitation score among the generation prompts. We also observe that the variance does not depend on the image counts which indicates that the performance of the face recognition model does not depend on the popularity of the individual. The change detection algorithm finds the change point to be at **234** faces for human face imitation for politicians, when using **SD1.1** for image generation. Figure 21 shows the similarity between the training and generated images when images are generated using **SD1.5**. Similar to SD1.1, the change is detected at **234** faces.

Imitation Threshold Estimation for Art Style Imitation: In Figure 23, we observe that the imitation scores for artists with low image counts have a baseline value around 0.2 (left side), and it increases as the number of their image counts increase towards the right. The highest similarity is 0.76 and it is for the artists in the rightmost region of the plot. We also observe a low variance across the generation prompts, and the variance does not depend on the image frequency of the artist.

Table 7: *Imitation Thresholds* for human face and art style imitation for the different text-to-image models and datasets we experiment with.

		Human Faces 🧑		Art Style 🖼️	
Pretraining Dataset	**Model**	**Celebrities**	**Politicians**	**Classical Artists**	**Modern Artists**
LAION2B-en	SD1.1	364	234	112, 391	198
	SD1.5	364, 8571	234, 4688	112, 360	198, 4821
LAION-5B	SD2.1	527, 9650	369, 8666	185, 848	241, 1132

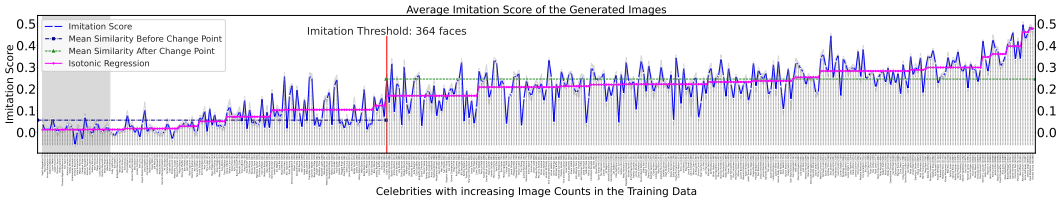


Figure 17: **Human Face Imitation (Celebrities):** Similarity between the training and generated images for all celebrities. The celebrities with zero image counts are shaded with light gray. We show the mean and variance over the five generation prompts. The images were generated using **SD1.1**. The change point for human face imitation for celebrities when generating images using SD1.1 is detected at **364 faces**.

The change detection algorithm finds the change point to be at **112** images for art style imitation of classical artists, when using **SD1.1** for image generation. Figure 24 shows the similarity between the training and generated images when images are generated using **SD1.5**. Similar to SD1.1, the change is detected at **112** faces for art style imitation when using SD1.5. These thresholds are slightly higher for style imitation of modern artists, 198 for both SD1.1 and SD1.5.

J Examples of Outliers

Figure 29 and Figure 30 show examples of outliers of the first kind, where aliases of a celebrity leads to under counting of their images in the pretraining data.

Figure 31 and Figure 32 show examples of outliers of the first kind for artists, where aliases of an artist leads to under counting of their art works in the pretraining data.

K Ablation Experiment with Different Face Embedding Models

In this section, we show the difference in the performance of several face embedding models and justify the choice of the final choice of our face embedding model. Face embedding models are evaluated using two main metrics: false-match rate (FMR) and true-match rate (TMR) (NIST, 2020). FMR measures how many times does a model says two people are the same when they are not and TMR measures how many times a model says two people are the same when they are the same. Ideally, a face embedding model should have low FMR and high TMR. An important variant of these metrics is the disparity of FMR and TMR of a model across different demographic groups. Ideally, a model should have low disparity in these metrics across different demographics. We also focus on the variance of these metrics across demographics in making the final choice.

We evaluate the FMR and TMR of eight different face embedding models (seven open-sourced and one proprietary). The open-source models were chosen based on their popularity on Github (Serengil & Ozpinar, 2020; Deng et al., 2022; 2020), and we also experiment with Amazon Rekognition, a proprietary model. For evaluating the disparity of these metrics across different demographic groups we grouped celebrities in six demographic groups primarily categorized according to skin color tone (black, brown, and white) and perceived gender (male and female; for simplicity). Each of the six groups had 10 celebrities (a total of 60), with no intersection between them. The categorization was done manually by looking at the reference images of the celebrities. For each celebrity, we collect 10 reference images from the internet by using the procedure described in Appendix D. We use these images to compare the FMR and TMR of the face recognition models, as these images are the gold standard images of a person.

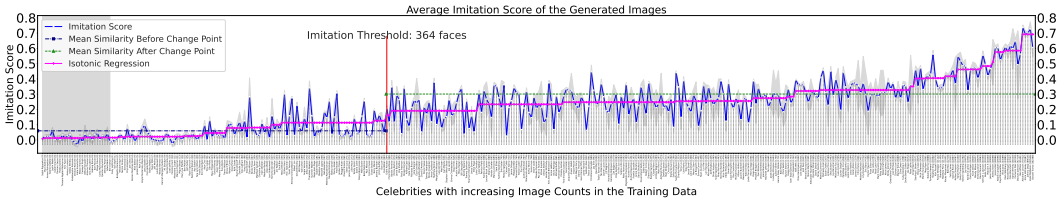


Figure 18: **Human Face Imitation (Celebrities):** similarity between the training and generated images for all celebrities. We show the mean and variance over the five generation prompts. The images were generated using **SD1.5**. The change point for human face imitation for celebrities when generating images using SD1.5 is detected at **364 faces**.

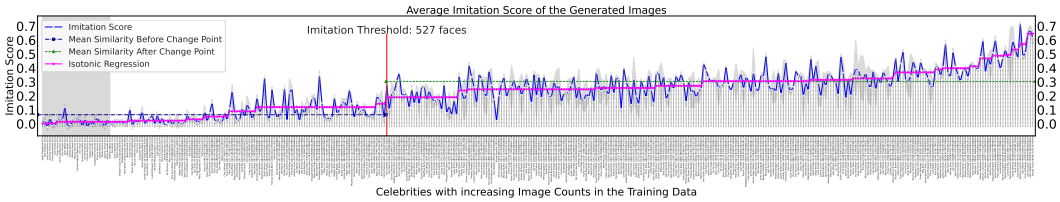


Figure 19: **Human Face Imitation (Celebrities):** similarity between the training and generated images for all celebrities. We show the mean and variance over the five generation prompts. The images were generated using **SD2.1**. The change point for human face imitation for celebrities when generating images using SD2.1 is detected at **527 faces**.

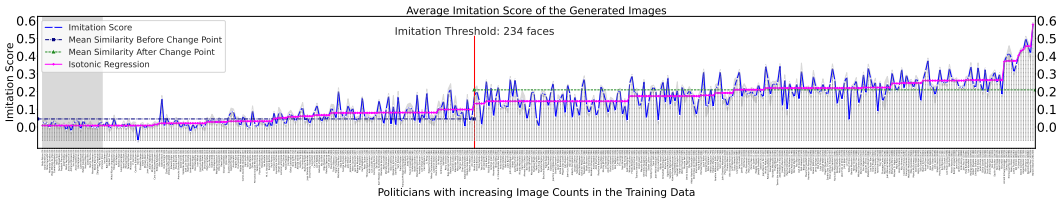


Figure 20: **Human Face Imitation (Politicians):** Similarity between the training and generated images for all politicians. The politicians with zero image counts are shaded with light gray. We show the mean and variance over the five generation prompts. The images were generated using **SD1.1**. The change point for human face imitation for politicians when generating images using SD1.1 is detected at **234 faces**.

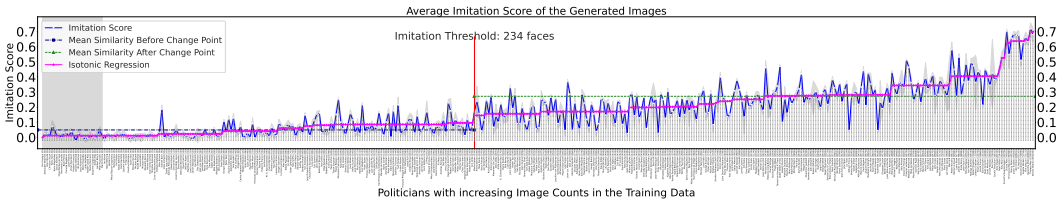


Figure 21: **Human Face Imitation (Politicians):** similarity between the training and generated images for all politicians. We show the mean and variance over the five generation prompts. The images were generated using **SD1.5**. The change point for human face imitation for politicians when generating images using SD1.5 is detected at **234 faces**.

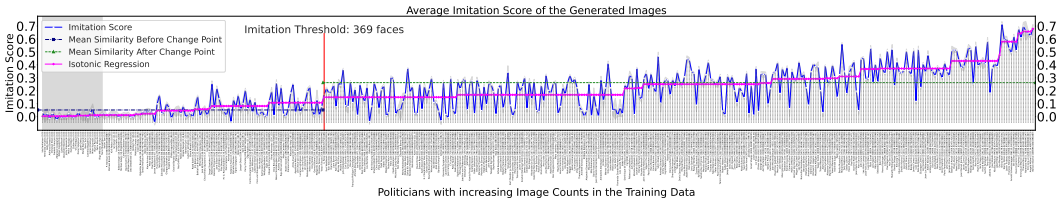


Figure 22: **Human Face Imitation (Politicians):** similarity between the training and generated images for all politicians. We show the mean and variance over the five generation prompts. The images were generated using **SD2.1**. The change point for human face imitation for celebrities when generating images using SD2.1 is detected at **369 faces**.

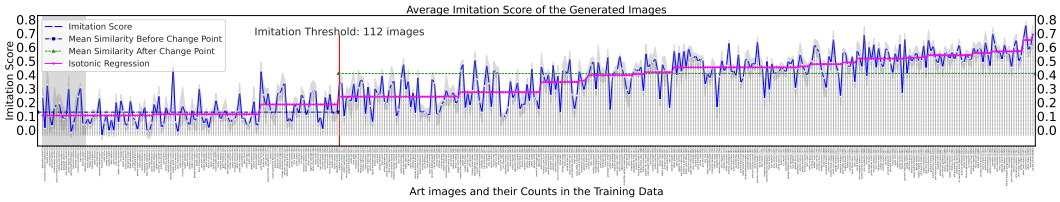


Figure 23: **Art Style Imitation (Classical Artists):** similarity between the training and generated images for **classical** art styles. We show the mean and variance over the five generation prompts. The images were generated using **SD1.1**. The change point for art style imitation when generating images using SD1.1 is detected at **112 images**.

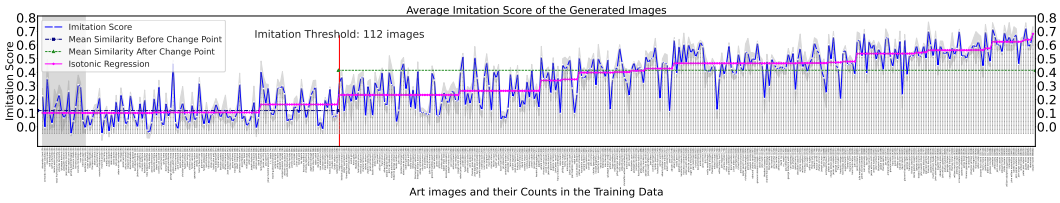


Figure 24: **Art Style Imitation (Classical Artists):** similarity between the training and generated images for **classical** art styles. We show the mean and variance over the five generation prompts. The images were generated using **SD1.5**. The change point for art style imitation when generating images using SD1.5 is detected at **112 images**.

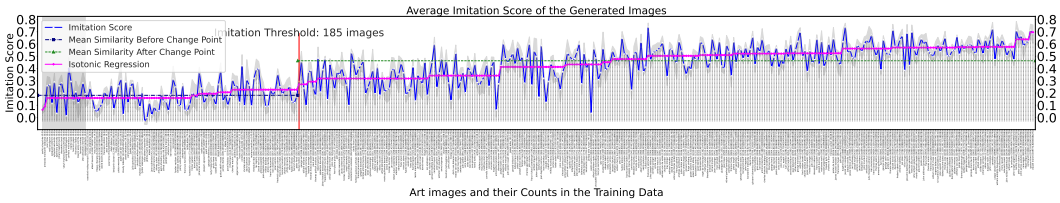


Figure 25: **Art Style Imitation (Classical Artists):** similarity between the training and generated images for **classical** art styles. We show the mean and variance over the five generation prompts. The images were generated using **SD2.1**. The change point for art style imitation when generating images using SD2.1 is detected at **185 images**.

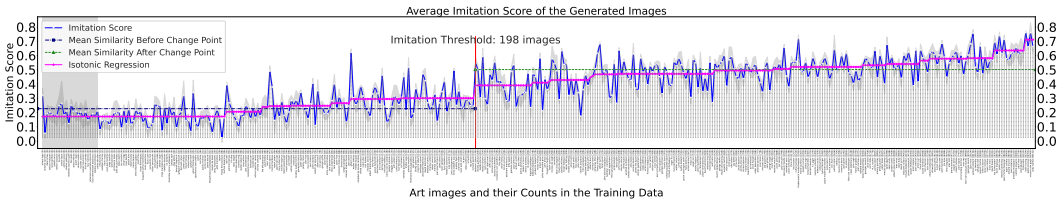


Figure 26: **Art Style Imitation (Modern Artists):** similarity between the training and generated images for **modern** art styles. We show the mean and variance over the five generation prompts. The images were generated using **SD1.1**. The change point for art style imitation when generating images using SD1.1 is detected at **198 images**.

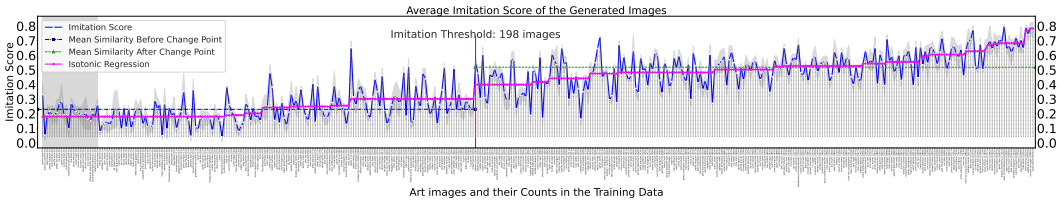


Figure 27: **Art Style Imitation (Modern Artists):** similarity between the training and generated images for **modern** art styles. We show the mean and variance over the five generation prompts. The images were generated using **SD1.5**. The change point for art style imitation when generating images using SD1.5 is detected at **198 images**.

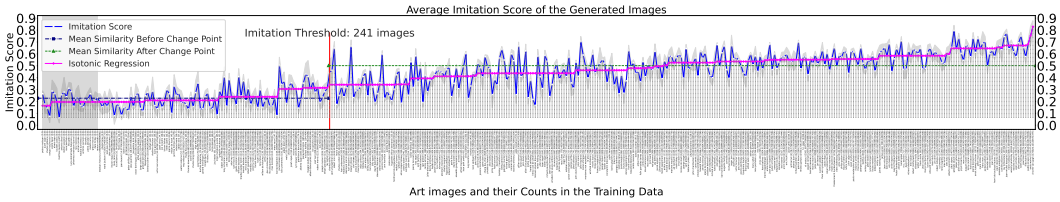


Figure 28: **Art Style Imitation (Modern Artists):** similarity between the training and generated images for **modern** art styles. We show the mean and variance over the five generation prompts. The images were generated using **SD2.1**. The change point for art style imitation when generating images using SD2.1 is detected at **241 images**.

Figure 29: **Outlier Category 1: DJ Kool Herc.** *Clive Campbell* is aliased as *DJ Kool Herc*, which leads to lower counts of his images in the dataset since MIMETIC^2 only collects images whose caption mentions *DJ Kool Herc*.

Figure 30: **Outlier Category 1: Summer Walker.** *Summer Marjani Walker* is aliased as *Summer Walker*, which leads to lower counts of her images in the dataset since MIMETIC^2 only collects images whose caption mentions *Summer Walker*.

FMR Computation: We compute the mean cosine similarity between the face embeddings of one individual and the faces of all other individuals in that group, and repeat the procedure for all individuals in a demographic group.

TMR Computation: We compute the mean cosine similarity between the embeddings of all the faces of an individuals and repeat the procedure for all the individuals in a demographic group.

Figure 33 and Figure 34 shows the FMR and TMR for six demographic groups for all the face embedding models. All the open-sourced models, except InsightFace, either have a high disparity in FMR values across the demographic groups (ArcFace, Facenet, Facenet512, DeepFace) or have

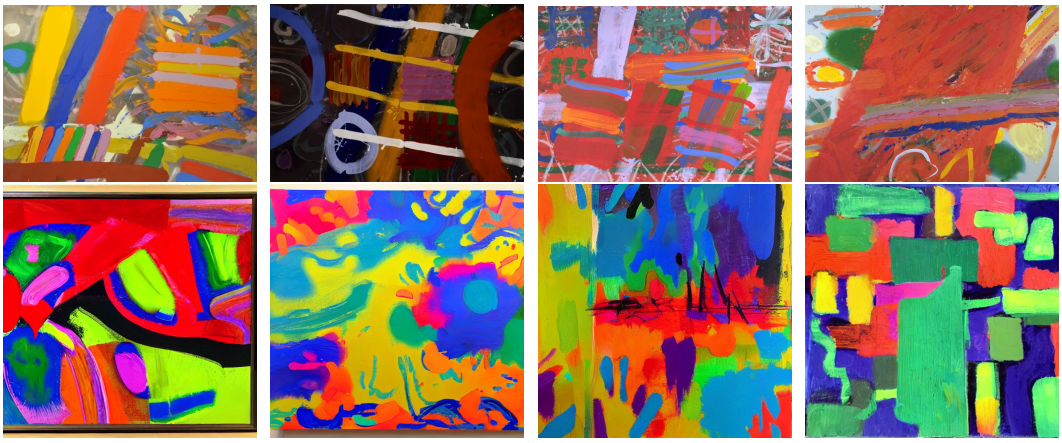


Figure 31: **Outlier Category 1: Albert Irwin.** *Albert Henry Thomas Irvin* is aliased as *Albert Irwin*, which leads to lower counts of his art images in the dataset since MIMETIC^2 only collects images whose caption mentions *Albert Irwin*.

Figure 32: **Outlier Category 1: Gustav Adolf Mossa.** *Gustav Adolf Mossa* is aliased just as *Mossa*, which leads to lower counts of his art images in the dataset since MIMETIC^2 only collects images whose caption mentions *Gustav Adolf Mossa*.

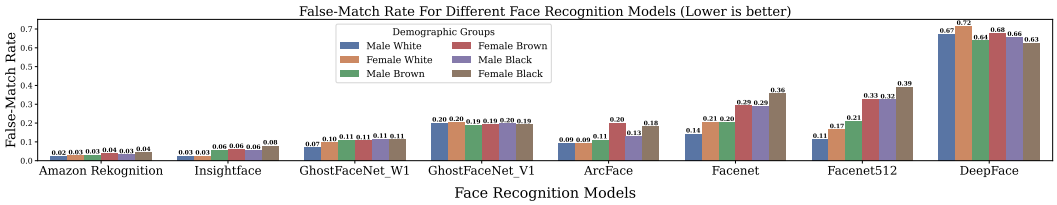


Figure 33: False-match rate (FMR) of all the face embedding models across the six demographic groups. Amazon Rekognition and InsightFace have the lowest FMR values. Moreover, these two models have lowest disparity of FMR over the demographic groups.

very low TMR (GhostFaceNet_W1, GhostFaceNet_V1). We choose InsightFace for our experiments because of it has 1) a low overall FMR, 2) decent TMR, 3) a low disparity of FMR and TMR across the demographic groups, and 4) is open-sourced. Having a low disparity of the metrics across individuals of different demographic groups is crucial for an accurate estimation of the imitation threshold. The Amazon Rekognition model would also be a viable choice based on these metrics, however, it is not open-sourced and therefore expensive for our experiments.

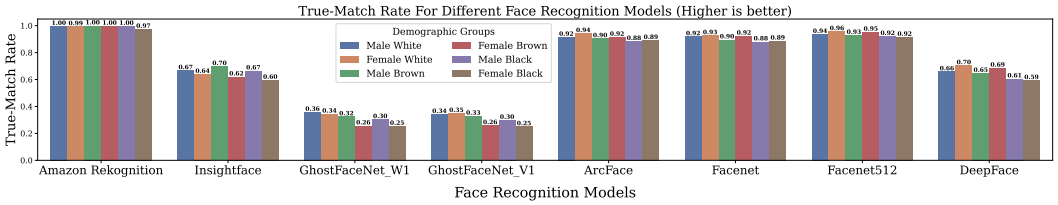


Figure 34: True-match rate (TMR) of all the face embedding models across the six demographic groups. Amazon Rekognition model has the highest TMR values.

L Count Distribution and the List of Sampled Entities for Each Domain

Table 8: Distribution of caption counts for sampled entities in celebrities, politicians, and art styles domains.

Caption Counts (LAION-2B)	Celebrities	Politicians	Classical Artists	Modern Artists
0	19	15	14	15
1-100	48	60	67	69
100-500	57	120	133	139
500-1K	52	80	62	62
1K-5K	151	65	63	64
5K-10K	19	40	39	32
> 10K	53	40	40	34

L.1 Celebrities

We collect celebrities from `https://www.popsugar.com/Celebrities` and `https://celebanswers.com/celebrity-list/`. The distribution of the caption counts of the sampled celebrities is displayed in Table 8. The sampled celebrities in the descending order of their number of caption counts are:

```
Donald Trump, Kate Middleton, Abraham Lincoln, Johnny Depp, Stephen King, Anne Hathaway, Ben Affleck, Ronald Reagan, Oprah
↪ Winfrey, Floyd Mayweather, Dwayne Johnson, Cameron Diaz, Cate Blanchett, Mark Wahlberg, Naomi Campbell, Nick Jonas, Jessica
↪ Biel, Kendrick Lamar, Malcolm X, Steven Spielberg, Bella Thorne, Bob Ross, Jay Leno, David Tennant, Samuel L. Jackson, Jason
↪ Statham, Mandy Moore, Victoria Justice, Scott Disick, Martin Scorsese, Ashley Olsen, Carey Mulligan, Greta Thunberg, Ashlee
↪ Simpson, Kacey Musgraves, Kurt Russell, Felicity Jones, Saoirse Ronan, Sarah Paulson, Matthew Perry, Forest Whitaker, Brendon
↪ Urie, Meg Ryan, Olivia Culpo, Joe Rogan, Sacha Baron Cohen, Terrence Howard, Natalie Dormer, Ansel Elgort, Nick Offerman,
↪ Clive Owen, Rose Leslie, Sterling K. Brown, Cuba Gooding Jr., Kevin James, Marisa Tomei, Troye Sivan, Zachary Levi, Gwendoline
↪ Christie, Hunter Hayes, Melanie Martinez, Joel McHale, Ross Lynch, Brody Jenner, Riley Keough, Robert Kraft, Ray Liotta, Eric
↪ Bana, Mark Consuelos, Chris Farley, James Garner, Lauren Daigle, Lily Donaldson, Penélope Cruz, Karen Elson, Joey Fatone,
↪ Leslie Odom Jr., Jay Baruchel, Selita Ebanks, Lana Condor, Mackenzie Foy, Doja Cat, Skai Jackson, Sofia Hellqvist, Bernard
↪ Arnault, Josh Peck, Lindsay Price, Phoebe Bridgers, Sarah Chalke, Alexander Skarsgård, Tai Lopez, Léa Seydoux, Cam Gigandet,
↪ David Dobrik, Jacob Elordi, Omar Epps, Marsai Martin, Alyson Stoner, Dree Hemingway, Gregg Sulkin, Mamie Gummer, Allison
↪ Holker, Chris Watts, Jacob Sartorius, Christine Quinn, Torrey Devitto, Alek Wek, Sandra Cisneros, Robert Irvine, Danielle
↪ Fishel, Normani Kordei, Sam Taylor Johnson, Jessica Seinfeld, Rachelle Lefevre, Joyner Lucas, Jimmy Buffet, John Wayne Gacy,
↪ Marvin Sapp, Ryan Guzman, Lindsay Ellingson, John Corbett, Michaela Coel, Hanne Gaby Odiele, Christiano Ronaldo, Scott
↪ Speedman, Addison Rae, Justice Smith, Stella Tennant, Lindsay Wagner, AJ Michalka, Charles Melton, Patricia Field, Dan
↪ Bilzerian, Annie Murphy, Michiel Huisman, Sara Foster, Diego Boneta, Danny Thomas, Oliver Hudson, Lauren Bushnell, Chris
↪ Klein, Rodrigo Santoro, Luke Hemsworth, Rhea Perlman, Michael Peña, Jodie Turner-Smith, Trevor Jackson, Jenna Marbles, Bob
↪ Morley, Zak Bagans, Liza Koshy, Steve Lacy, Nico Tortorella, Emma Corrin, Lo Bosworth, Quvenzhané Wallis, Martin Starr, David
↪ Muir, Beanie Feldstein, Lori Harvey, Eddie McGuire, Todd Chrisley, Dan Crenshaw, Amanda Gorman, Crystal Renn, Mark Richt,
↪ Magdalena Frackowiak, Danielle Jonas, Liu Yifei, Sasha Pivovarova, Ashleigh Murray, Peter Hermann, Daria Strokous, Eddie Hall,
↪ Hunter Parrish, Matt McGorry, Diane Guerrero, Simu Liu, Brady Quinn, Jill Wagner, Richard Rawlings, Sophia Lillis, Genesis
↪ Rodriguez, Diane Ladd, Frankie Grande, Olivia Rodrigo, Anwar Hadid, Hannah Bronfman, Deana Carter, Tao Okamoto, Fei Fei Sun,
↪ Taylor Tomasi Hill, Jared Followill, Margherita Missoni, Elisa Sednaoui, Thomas Doherty, Bill Skarsgård, Indya Moore, Ziyi
↪ Zhang, Cacee Cobb, Jay Ellis, Arthur Blank, Chris McCandless, Paz de la Huerta, Jacquelyn Jablonski, Michael Buffer, Annie
↪ LeBlanc, Kieran Culkin, Lacey Evans, Rachel Antonoff, Presley Gerber, Lauren Bush Lauren, Peter Firth, Tina Knowles-Lawson,
↪ Sunisa Lee, Douglas Brinkley, Hero Fiennes-Tiffin, Erin Foster, Justina Machado, Mariacarla Boscono, Summer Walker, Emma
↪ Chamberlain, Lew Alcindor, Jenna Ortega, Phoebe Dynevor,
Kim Zolciak-Biermann, Allison Stokke, Malgosia Bela, Isabel Toledo, Sydney Sweeney, Mat Fraser, Hunter McGrady, Ethan Suplee,
↪ Tammy Hembrow, Ivan Moody, Danneel Harris, Marcus Lemonis, Hunter Schafer, Luka Sabbat, Sam Elliot, Kendra Spears, Stephen
↪ tWitch Boss, Joe Lacob, Tommy Dorfman, Emma Barton, Elliot Page, Sha'Carri Richardson, Barry Weiss, Julie Chrisley, Devon
↪ Sawa, Miles Heizer, Julia Stegner, Austin Abrams, Jacquetta Wheeler, Melanie Iglesias, Anna Cleveland, Eiza González, Grant
↪ Achatz, Matt Stonie, Connor Cruise, Nicholas Braun, Dan Lok, Charli D'Amelio, Jeremy Bamber, Jim Walton, Matthew Bomer, Nicola
↪ Coughlan, Una Stubbs, Andrew East, Miles O'Brien, Mary Fitzgerald, Taylor Mills, Portia Freeman, Kate Chastain, David
↪ Brinkley, Bregje Heinen, DJ Kool Herc, Barbie Ferreira, Paul Mescal, Forrest Fenn, Jamie Bochert, Yung Gravy, Daisy
↪ Edgar-Jones, Dixie D'Amelio, Jordan Chiles, Bob Keeshan, Alexandra Cooper, Kyla Weber, Chase Stokes, Belle Delphine, Joanna
↪ Hillman, Olivia O'Brien, Jillie Mack, Maggie Rizer, Sasha Calle, Tony Lopez, Danny Koker, Irwin Winkler, M.C. Hammer, Zack
↪ Bia, Alexa Demie, Bailey Sarian, Yael Cohen, Angie Varona, Trevor Wallace, Madelyn Cline, Fred Stoller, Frank Sheeran, Albert
↪ Lin, Sessilee Lopez, Zaya Wade, Maitreyi Ramakrishnan, Madison Bailey, Will Reeve, Nick Bolton, Rege-Jean Page, Matthew
↪ Garber, Yamiche Alcindor, Isaak Presley, Thandiwe Newton, Nicole Fosse, Shenae Grimes-Beech, Alex Choi, Scott Yancey, Ciara
↪ Wilson, Lexi Underwood, Manny Khoshbin, Ella Emhoff, Cole LaBrant, Wayne Carini, Greg Fishel, Ryan Upchurch, Marcus Freeman,
↪ Danielle Cohn, Sue Aikens, Kyle Cooke, David Portnoy, Avani Gregg, Dan Peña, Quinton Reynolds, Eric Porterfield, Ayo Edebiri,
↪ Tara Lynn Wilson, Florence Hunt, Nicola Porcella, Pashmina Roshan, Josh Seiter, Ben Mallah, Miguel Bezos, Lukita Maxwell, Ali
↪ Skovbye, Jordan Firstman, Jeff Molina, Mary Lee Pfeiffer, Cody Lightning, Leah Jeffries, Elle Graham, Hannah Margaret Selleck,
↪ Woody Norman, Tom Blyth, Banks Repeta, Wisdom Kaye, Kris Tyson, Joey Klaasen, Tioreore Ngatai-Melbourne, Jani Zhao, Cara Jade
↪ Myers, Keyla Monterroso Mejia, Samara Joy, Mason Thames, Park Ji-hu, Boman Martinez-Reid, Priya Kansara, Yasmin Finney,
↪ Bridgette Doremus, Aria Mia Loberti, Isabel Gravitt, Gabriel LaBelle, Delaney Rowe, Armen Nahapetian, Aditya Kusupati, Vedang
↪ Raina, Arsema Thomas, Adwa Bader, Amaury Lorenzo, Corey Mylchreest, Sam Nivola, Gabby Windey, Cwaayal Singh, Jaylin Webb,
↪ Kudakwashe Rutendo, Chintan Rachchh, Sajith Rajapaksa, Diego Calva, Pardis Saremi, Dominic Sessa, India Amarteifio, Mia
↪ Challiner, Aryan Simhadri
```

L.2 POLITICIANS

We collect politicians from Wikipedia Wikipedia (2024). The distribution of the caption counts of the sampled politicians is given in Table 8. The sampled politicians in the descending order of their number of caption counts are:

```
Barack Obama, John Lewis, Theresa May, Narendra Modi, Kim Jong-un, David Cameron, Angela Merkel, Bill Clinton, Xi Jinping, Justin
↪ Trudeau, Emmanuel Macron, Nancy Pelosi, Arnold Schwarzenegger, Ron Paul, Shinzo Abe, Adolf Hitler, John Paul II, Tony Blair,
↪ Sachin Tendulkar, Nick Clegg, Newt Gingrich, Scott Morrison, Arvind Kejriwal, Ilham Aliyev, Jacob Zuma, Bashar al-Assad, Laura
↪ Bush, Sonia Gandhi, Kim Jong-il, Robert Mugabe, James Comey, Rodrigo Duterte, Pete Buttigieg, Lindsey Graham, Hosni Mubarak,
↪ Enda Kenny, Alexei Navalny, Rob Ford, Leo Varadkar, Evo Morales, Lee Hsien Loong, Henry Kissinger, Petro Poroshenko, Joko
↪ Widodo, Clarence Thomas, Rishi Sunak, Mohamed Morsi, Ashraf Ghani, Martin McGuinness, Viktor Orban, Uhuru Kenyatta, Mike
↪ Huckabee, Sheikh Hasina, Martin Schulz, Giuseppe Conte, John Howard, Benito Mussolini, Tulsi Gabbard, Dominic Raab, Michael D.
↪ Higgins, François Hollande, Yasser Arafat, Mark Rutte, Mahathir Mohamad, Juan Manuel Santos, Abiy Ahmed, William Prince, Lee
↪ Kuan Yew, Mikhail Gorbachev, Hun Sen, Jacques Chirac, Martin O'Malley, Benazir Bhutto, Yoshihide Suga, John Major, Muammar
↪ Gaddafi, Jerry Springer, Sandra Day O'Connor, Madeleine Albright, Thomas Mann, Paul Kagame, Simon Coveney, Grant Shapps,
↪ Sebastian Coe, Merrick Garland, Jean-Yves Le Drian, Nursultan Nazarbayev, Horst Seehofer, Liz Truss, Rowan Williams, Ellen
↪ Johnson Sirleaf, George Weah, Mark Sanford, Yoweri Museveni, Luigi Di Maio, Ben Wallace, Herman Van Rompuy, Daniel Ortega,
↪ Olaf Scholz, Beppe Grillo, Alassane Ouattara, Nicolás Maduro, Tamim bin Hamad Al Thani, Mary McAleese, Asif Ali Zardari,
↪ Joseph Goebbels, Nikol Pashinyan, Deb Haaland, Paul Biya, Abdel Fattah el-Sisi, Thabo Mbeki, Kyriakos Mitsotakis, Joseph
↪ Muscat, Micheál Martin, Rebecca Long-Bailey, Paschal Donohoe, Todd Young, Jean-Marie Le Pen, Nick Griffin, Zoran Zaev, Pierre
↪ Nkurunziza, Abhisit Vejjajiva, Maggie Hassan, Steven Chu, Juan Guaidó, Edi Rama, Mary Landrieu, Jyrki Katainen, Jens Spahn,
↪ John Dramani Mahama, Gina Raimondo, Alec Douglas-Home, Viktor Orbán, Anita Anand, Isaias Afwerki, James Cleverly, Ibrahim
↪ Mohamed Solih, Leymah Gbowee, Václav Havel, John Rawls, Jack McConnell, Romano Prodi, Eoghan Murphy, Vicky Leandros, Norodom
↪ Sihamoni, Nayib Bukele, Shirin Ebadi, Jusuf Kalla, George Eustice, Joachim von Ribbentrop, Peter Altmaier, Akbar Hashemi
↪ Rafsanjani, Paul Singer, Christian Stock, Moussa Faki, Dominique de Villepin, Michael Fabricant, Kim Dae-jung, Eamon Ryan,
↪ Shavkat Mirziyoyev, Denis Sassou-Nguesso, Werner Faymann, Kamla Persad-Bissessar, Ingrid Betancourt, Volodymyr Zelenskyy, Park
↪ Chung Hee, Elvira Nabiullina, Roselyne Bachelot, Heinz Fischer, Hideki Tojo, Anatoly Karpov, Marcelo Ebrard, Slavoj Žižek,
↪ Trent Lott, Alfred Rosenberg, Gabi Ashkenazi, Valentina Matviyenko, Kgalema Motlanthe, Pedro Castillo, Winona LaDuke, Peter
↪ Bell, Boyko Borisov, Carl Bildt, Almazbek Atambayev, Andry Rajoelina, Carl Schmitt, Ralph Gonsalves, Liam Byrne, Alok Sharma,
↪ Jean-Michel Blanquer, Robert Schuman, Shinzō Abe, Doris Leuthard, Jacques Delors, Floella Benjamin, Sauli Niinistö, Annalena
↪ Baerbock, Toomas Hendrik Ilves, Alejandro Giammattei, Bob Kerrey, Lionel Jospin, Murray McCully, Stefan Löfven, Javier Solana,
↪ Salva Kiir Mayardit, Cecil Williams, Shahbaz Bhatti, Marianne Thyssen, Marty Natalegawa, Roh Moo-hyun, John Diefenbaker,
↪ Antonio Inoki, Iván Duque, CY Leung, Tom Tancredo, Sigrid Kaag, Jim Bolger, Lou Barletta, Li Peng, Laura Chinchilla, Gennady
↪ Zyuganov, Chen Shui-bian,
Sebastián Piñera, Gustavo Petro, Miguel Díaz-Canel, Alberto Fernández, Gerald Darmanin, Boutros Boutros-Ghali, Joschka Fischer,
↪ Maia Sandu, Ricardo Martinelli, Andrej Babiš, Dan Jarvis, Nikos Dendias, Chris Hipkins, Tawakkol Karman, Booth Gardner, Karin
↪ Kneissl, Mobutu Sese Seko, Alexander Haig, Alexander De Croo, Ahmed Aboul Gheit, Yasuo Fukuda, Jean-Luc Mélenchon, Jane
↪ Ellison, Diane Dodds, Helen Whately, Idriss Déby, Patrice Talon, Carmen Calvo, Dario Franceschini, Emma Bonino, Richard
↪ Ferrand, Andreas Scheuer, Moshe Katsav, K. Chandrashekar Rao, P. Harrison, Robert Habeck, Ann Linde, Jon Ashworth, Edward
↪ Scicluna, Stef Blok, Lawrence Gonzi, William Roper, Josep Rull, Sam Kutesa, Raja Pervaiz Ashraf, David Cairns, Ilir Meta,
↪ Perry Christie, Rinat Akhmetov, Ahmet Davutoğlu, Franck Riester, Nikos Christodoulides, Damien O'Connor, Sali Berisha, Umberto
↪ Bossi, Lee Cheuk-yan, Alpha Condé, Alexander Newman, Annette Schavan, Yuri Andropov, Peter Tauber, Faure Gnassingbé, Bolkiah
↪ of Brunei, Karl-Theodor zu Guttenberg, Michael Brand, Helen Suzman, Ron Huldai, Mohamed Azmin Ali, François-Philippe
↪ Champagne, Agostinho Neto, Marielle de Sarnez, Kurt Waldheim, Mounir Mahjoubi, Juan Orlando Hernández, Angela Kane, Lech
↪ Wałęsa, Luis Lacalle Pou, Barbara Pompili, Margaritis Schinas, Tigran Sargsyan, Wolfgang Bosbach, Raed Saleh, Johanna Wanka,
↪ Michelle Donelan, Roberto Speranza, Traian Băsescu, Iurie Leancă, Dara Calleary, Ilona Staller, Micheline Calmy-Rey, Thomas
↪ Oppermann, Karine Jean-Pierre, Luciana Lamorgese, Azali Assoumani, Michael Adam, Paulo Portas, Svenja Schulze, Pita Sharples,
↪ Choummaly Sayasone, Federico Franco, Félix Tshisekedi, Roberta Metsola, Nia Griffith, Paul Myners, Ahmad Vahidi, Kaja Kallas,
↪ Hua Guofeng, Olga Rypakova, Otto Grotewohl, Audrey Tang, Oskar Lafontaine, Ivica Dačić, Isa Mustafa, Xiomara Castro, M. G.
↪ Ramachandran, Fernando Grande-Marlaska, Wopke Hoekstra, Tomáš Petříček, Egils Levits, Roland Koch, Joseph Deiss, Laurentino
↪ Cortizo, Alan García, Nikola Poposki, Evarist Bartolo, Reyes Maroto, Zuzana Čaputová, Sergei Stanishev, Plamen Oresharski, Ana
↪ Brnabić, Carlos Alvarado Quesada, Marek Biernacki, Olivier Véran, Vjekoslav Bevanda, Clare Moody, Matthias Groote, Giorgos
↪ Stathakis, Marta Cartabia, Elena Bonetti, Dina Boluarte, Milo Đukanović, Levan Kobiashvili, Isabel Celaá, Jarosław Gowin, José
↪ Luis Escrivá, Cora van Nieuwenhuizen, Ivan Mikloš, Arancha González Laya, Viola Amherd, Gernot Blümel, José Luis Ábalos, Deo
↪ Debattista, Alain Krivine, Zlatko Lagumdžija, Edward Argar, Adrian Năstase, Zdravko Počivalšek, Miroslav Kalousek, Gabriel
↪ Boric, Juan Carlos Campo, Karel Havlíček, Kiril Petkov, Elżbieta Rafalska, Tobias Billström, Miroslav Toman, Mihai Răzvan
↪ Ungureanu, Ivaylo Kalfin, Élisabeth Borne, Herbert Fux, Petru Movilă, Koichi Tani, Caroline Edelstam, Barbara Gysi, L'ubomír
↪ Jahnátek, Nuno Magalhães, Martin Pecina, Goran Knežević, Björn Böhning, Iñigo Méndez de Vigo, Božo Petrov, Ian Karan, Hernando
↪ Cevallos, Milan Kujundžić, Adriana Dăneasă, Ida Karkiainen, Zoran Stanković, Boris Tučić, Jerzy Kropiwnicki, Rafael Catalá
↪ Polo, Ljube Boškoski, Camelia Bogdănici, Józef Oleksy, Frederik François, Zbigniew Ćwiąkalski, Herbert Bösch, Metin Feyzioğlu,
↪ Zoltán Illés, Vivi Friedgut
```

L.3 CLASSICAL ARTISTS

We collected classical artists from the https://www.wikiart.org, a website that collects various arts from different artists and categorizes them into pre-defined art style categories. For classical artists, we collected the artist names from the art styles: *Romanticism, Impressionism, Realism, Baroque, Neoclassicism, Rococo, Academic Art, Symbolism, Cubism, Naturalism*. The distribution of the caption counts of the sampled artists is given in Table 8. The sampled artists in the descending order of their number of caption counts are:

```
Claude Monet, Rembrandt, Gustav Klimt, Edgar Degas, Caravaggio, William Blake, John James Audubon, Le Corbusier, Canaletto,
↪ Peter Paul Rubens, John Singer Sargent, Edouard Manet, John William Waterhouse, Alfred Sisley, Childe Hassam, Berthe Morisot,
↪ Victor Hugo, William-Adolphe Bouguereau, Gustave Courbet, Albert Bierstadt, Mary Cassatt, John Constable, Gustave Dore,
↪ Gustave Caillebotte, Henry Moore, Thomas Hardy, Johannes Vermeer, Jacques-Louis David, Odilon Redon, Thomas Cole, Thomas
↪ Moran, James Tissot, William Hogarth, David Roberts, Thomas Gainsborough, Anthony Van Dyck, William Merritt Chase, Caspar
↪ David Friedrich, Sir Lawrence Alma-Tadema, George Stubbs, Georges Braque, Auguste Rodin, Joshua Reynolds, John Atkinson
↪ Grimshaw, David James, James Ward, David Johnson, Frederic Edwin Church, Jean-Leon Gerome, Eugene Delacroix, Martin Johnson
↪ Heade, Edward Burne-Jones, John William Godward, James Webb, Gustave Moreau, James Charles, Francois Boucher, Francisco Goya,
↪ John Everett Millais, Thomas Lawrence, John Ruskin, John Russell, David Davies, Dante Gabriel Rossetti, George Henry, John
↪ Martin, Frans Hals, Guido Reni, George Catlin, Claude Lorrain, Anders Zorn, Jessie Willcox Smith, Giovanni Battista Tiepolo,
↪ Howard Pyle, Archibald Thorburn, Thomas Eakins, Giovanni Boldini, Armand Guillaumin, Ivan Aivazovsky, John Trumbull, Joseph
↪ Wright, Benjamin West, John Collier, Henri Fantin-Latour, Jan Steen, Eugene Boudin, James Mcneill Whistler, Ilya Repin,
↪ William Bradford, Julia Margaret Cameron, Annibale Carracci, Antoine Watteau, Marianne North, David Cox, Jacob Jordaens,
↪ Frederick Morgan, Ivan Shishkin, George Morland, Ford Madox Brown, Frans Snyders, John Jackson, Aelbert Cuyp, Charles Willson
↪ Peale, Jacob Van Ruisdael, Joseph Ducreux, Horace Vernet, Pieter De Hooch, Arthur Hughes, Antonio Canova, Charles Le Brun,
↪ Francesco Hayez, Thomas Sully, Isaac Levitan, Robert Spencer, Karl Bodmer, Alexandre Cabanel, N.C. Wyeth, Anna Ancher, Carl
↪ Spitzweg, David Wilkie, Paul Delaroche, Charles-Francois Daubigny, George Frederick Watts, Guy Rose, Carel Fabritius, Alfred
↪ Stevens, Peder Severin Kroyer, Taras Shevchenko, Pietro Longhi, Joaquin Sorolla, Theodore Chasseriau, John Riley, Theodore
↪ Rousseau, Edmund Charles Tarbell, Giovanni Domenico Tiepolo, Edward Ladell, Pompeo Batoni, Richard Parkes Bonington, Boris
↪ Kustodiev, Andreas Achenbach, Charles Conder, Viktor Vasnetsov, Antoine Blanchard, William Henry Hunt, Emile Claus, Julian
↪ Alden Weir, Mikhail Vrubel, Richard Dadd, Vasily Vereshchagin, John Hoppner, Richard Lindner, Aristide Maillol, Joan Blaeu,
↪ William Williams, Adriaen Brouwer, Constant Troyon, Fernand Khnopff, Edwin Austin Abbey, Gino Severini, Pietro Da Cortona,
↪ Adriaen Van De Velde, Vasily Perov, David Bomberg, Konstantin Korovin, Christoffer Wilhelm Eckersberg, Jean Metzinger,
```

Konstantin Makovsky, Mihaly Munkacsy, Albert Pinkham Ryder, Francesco Solimena, Franz Richard Unterberger, Roger De La Fresnaye, William Shayer, Paul Bril, Cornelis Springer, Jacques Lipchitz, Agostino Carracci, Adam Elsheimer, Giuseppe De Nittis, Jules Joseph Lefebvre, Albert Gleizes, Willard Metcalf, Vasily Surikov, Giovanni Fattori, Lyubov Popova, Kuzma Petrov-Vodkin, Johan Christian Dahl, Jehan Georges Vibert, Mikhail Nesterov, Antoine Pesne, Konstantin Yuon, Hugo Simberg, Gerard Terborch, Alexander Ivanov, Eustache Le Sueur, Giuseppe Maria Crespi, Ferdinand Bol, Max Slevogt, Philip Wilson Steer, Osias Beert, Vasily Polenov, John Crome, Edward Poynter, Nicolae Grigorescu, Louis Marcoussis, Marcus Stone, Jacques Stella, Edmonia Lewis, Antonietta Brandeis, Konstantin Somov, Hendrick Terbrugghen, Cornelis De Vos, Charles Spencelayh, Ivan Kramskoy, Rudolf Von Alt, Philipp Otto Runge, Carolus-Duran, Ralph Earl, Eugene Carriere, Julius Leblanc Stewart, Ippolito Caffi, John Peter Russell, Jean Baptiste Vanmour, Antonio Mancini, Petrus Van Schendel, Benjamin Brown, Max Klinger, Ludwig Knaus, Maurice Braun, Vincenzo Camuccini, Jean-Étienne Liotard, Henry Tonks, Jacek Malczewski, Rubens Santoro, Pieter Codde, Jean-Paul Laurens, Louise Moillon, Jan Siberechts, David Morier, John Pettie, Felicien Rops, Leon Bonnat, Theodule Ribot, William Logsdail, Richard Jack, Homer Watson, François Gérard, Robert Julian Onderdonk, Lionel Noel Royer, Charles Gleyre, Anne Brigman, Thomas Jones Barker, Antonio Ciseri, Joseph Anton Koch, Anton Melbye, Nicolas Tournier, Peter Nicolai Arbo, Lev Lagorio, Matthias Stom, Jean-Baptiste Van Loo, Konstantinos Volanakis, Cornelis Vreedenburgh, Henryk Siemiradzki, Frederick George Cotman, Eva Gonzales, Jan Cossiers, Julio Gonzalez, Vladimir Makovsky, Fyodor Bronnikov, Paul Peel, Thomas Pollock Anshutz, Raden Saleh, Robert Lewis Reid, Joseph-Marie Vien, Arno Breker, Frederick William Burton, Ion Andreescu, Jankel Adler, William Leighton Leitch, Esaias Van De Velde, Dirck Van Baburen, Jacob Van Strij, Franz Stuck, Giovanni Battista Gaulli, Hans Gude, Harriet Backer, Nicolas Antoine Taunay, Fyodor Alekseyev, Vasily Tropinin, Alfred Dehodencq, Alexey Venetsianov, Francis Davis Millet, Laszlo Mednyanszky, Charles Hermans, Christina Robertson, Thomas Francis Dicksee, Fyodor Vasilyev, Claudio Coello, Gustave Boulanger, Nikolaos Gyzis, George Ault, Francisco Herrera, John Lewis Krimmel, Marie Bashkirtseff, Sebastien Bourdon, Jacob Ochtervelt, Christen Kobke, Paul Gavarni, Edouard Debat-Ponsan, Gregoire Boonzaier, Dmitry Levitzky, André Gill, Julian Ashton, Telemaco Signorini, Orest Kiprensky, Fyodor Rokotov, Nicolas Toussaint Charlet, Pieter Saenredam, Henri-Pierre Picou, Johan Hendrik Weissenbruch, Émile Friant, Herbert Gustave Schmalz, Jean-Baptiste Pigalle, T. C. Steele, Arturo Michelena, Wilhelm Von Kaulbach, Algernon Talmage, Giovanni Costa, Paul Leroy, Ivan Vladimirov, Hermann Hendrich, Magnus Enckell, Pavel Fedotov, Ethel Carrick, Vincenzo Irolli, Leopold Survage, Lady Frieda Harris, Joseph Duplessis, Charles Maurin, Philip De Laszlo, Peter Fendi, Marie-Guillemine Benoist, Antonio Paoletti, Christian Wilhelm Allers, Tranquillo Cremona, Antonio Donghi, Penry Williams, Miklos Barabas, Alfred Concanen, Albert Maignan, Dobri Dobrev, Bertalan Szekely, Mariano Benlliure, Anton Azbe, Johannes Moreelse, Nicolae Vermont, Heinrich Bürkel, Jane Sutherland, Laslett John Pott, Petro Kholodny, Alexey Zubov, Eliseu Visconti, Pieter Wenning, Henri Le Fauconnier, Paul Ackerman, Armand Henrion, Ipolit Strambu, George Hemming Mason, Vilhelms Purvitis, Mykola Yaroshenko, Pavel Svinyin, Gustav Adolf Mossa, Fyodor Solntsev, Pedro Américo, Klavdy Lebedev, Ivan Milev, Albert Benois, Alexandre Antigna, George Demetrescu Mirea, Giulia Lama, Aurelio Tiratelli, Konstantin Vasilyev, Domenico Fiasella, David Kakabadze, Cornelis Van Noorde, Panos Terlemezian, Alexei Korzukhin, Maurice Poirson, Joaquín Agrasot, Toby Edward Rosenthal, Heinrich Papin, Vasile Popescu, Jérôme-Martin Langlois, Karl Edvard Diriks, Adam Van Der Meulen, Vsevolod Maksymovych, Leo Leuppi, Matej Sternen, Filippo Cifariello, Apollinary Goravsky, Pasquale Celommi, Giuseppe Barberis, Francesco Didioni, Octav Angheluta, Vytautas Kairiukstis, Gevorg Bashindzhagian, Serhij Schyschko, Noè Bordignon, Armando Montaner Valdueza, Alexander Clarot, Rosario Weiss Zorrilla, Vasyl Hryhorovych Krychevsky, Fernand Combes, Francesco Ribalta, Jean Alexandru Steriadi, Johann Baptist Clarot, Corneliu Michailescu, Nzante Spee

L.4 Modern Artists

We also collected modern artists from the https://www.wikiart.org. For modern artists, we collected the artist names from the art styles: *Expressionism, Surrealism, Abstract Expressionism, Pop Art, Art Informel, Post-Painterly Abstraction, Neo-Expressionism, Post-Minimalism, Neo-Impressionism, Neo-Romanticism, Post-Impressionism*. The distribution of the caption counts of the sampled artists is given in Table 8. The sampled artists in the descending order of their number of caption counts are:

Vincent Van Gogh, David Bowie, Andy Warhol, Pablo Picasso, Frida Kahlo, Keith Haring, Salvador Dali, Paul Gauguin, Camille Pissarro, Paul Cezanne, Henri Matisse, Paul Klee, Francis Bacon, Edvard Munch, Amedeo Modigliani, Egon Schiele, Jean-Michel Basquiat, David Lynch, Wassily Kandinsky, Peter Max, Roy Lichtenstein, Paul Reed, Franz Marc, David Smith, Mark Rothko, Georges Seurat, Leroy Neiman, Joan Miro, Jackson Pollock, August Macke, Man Ray, Piet Mondrian, Cy Twombly, Henri De Toulouse-Lautrec, Graham Bell, Paul Signac, Robert Indiana, Yayoi Kusama, Rene Magritte, Jasper Johns, Walter Crane, Robert Morris, Emily Carr, Lucian Freud, Ernst Ludwig Kirchner, Tom Thomson, Anish Kapoor, Alex Katz, Pierre Bonnard, John Cage, Jim Dine, Ellsworth Kelly, Peter Blake, William Scott, Erin Hanson, Marcel Duchamp, Frank Stella, Robert Motherwell, Max Weber, Louise Nevelson, Peter Phillips, Willem De Kooning, Corneille, Wayne Thiebaud, Joan Mitchell, Jean Cocteau, Raoul Dufy, Antony Gormley, Max Ernst, Alberto Giacometti, Vanessa Bell, Richard Diebenkorn, James Rosenquist, Edouard Vuillard, Richard Hamilton, M.C. Escher, Sam Francis, Sean Scully, Anselm Kiefer, Edward Weston, Karel Appel, Philip Guston, Julian Schnabel, Ray Parker, James Ensor, Balthus, George Segal, Francis Picabia, Emil Nolde, Georges Rouault, Alice Neel, Helen Frankenthaler, Claes Oldenburg, Theo Van Rysselberghe, Maya Lin, Maurice Utrillo, Eric Fischl, H.R. Giger, Maurice Denis, Friedensreich Hundertwasser, Will Barnet, Paula Modersohn-Becker, Suzanne Valadon, Bruce Nauman, El Anatsui, Lee Krasner, Joseph Cornell, Patrick Heron, James Brooks, Paula Rego, Paul Jenkins, Jules Pascin, Lyonel Feininger, Edward Ruscha, Norman Lewis, Barnett Newman, David Park, Marie Laurencin, Rufino Tamayo, Chris Ofili, Lynd Ward, Jean-Paul Riopelle, Roger Fry, Remedios Varo, Maxime Maufra, Paul Serusier, Jacob Epstein, Richard Deacon, Walter Sickert, Mark Tobey, Jan Toorop, Jacek Yerka, Red Grooms, Ad Reinhardt, Eva Hesse, Oskar Kokoschka, Michael Sowa, Jean David, Sam Gilliam, Phyllida Barlow, Howard Finster, Augustus John, Elaine De Kooning, Beauford Delaney, David Hammons, Erich Heckel, Amrita Sher-Gil, Arthur Lismer, Mona Hatoum, Etel Adnan, Brion Gysin, John Chamberlain, Corita Kent, Allen Jones, Asger Jorn, Martin Kippenberger, George Tooker, Desmond Morris, Wolf Kahn, Jay Defeo, Irma Stern, Walasse Ting, Emile Bernard, Kathe Kollwitz, Frank Bowling, John Heartfield, Auguste Herbin, Frances Hodgkins, Meret Oppenheim, Andre Masson, Karl Schmidt-Rottluff, Hans Bellmer, Marino Marini, Morris Louis, Pyotr Konchalovsky, Richard Artschwager, Louis Cane, Betty Parsons, Max Pechstein, Richard Pousette-Dart, Georges Lemmen, Cuno Amiet, Louis Valtat, Kit Williams, Grace Cossington Smith, John Hoyland, Dennis Oppenheim, Lynda Benglis, James Lee Byars, Boris Grigoriev, Lili Elbe, Victor Brauner, Adrian Ghenie, Gillian Ayres, Ossip Zadkine, Alice Bailly, Felix Gonzalez-Torres, Johannes Itten, Charles Long, John Marin, Winifred Nicholson, Alfred Kubin, Charles Angrand, Zinaida Serebriakova, John Duncan Fergusson, Norman Bluhm,

Harald Sohlberg, Zdzislaw Beksinski, Barkley L. Hendricks, Bruno Schulz, Toyen, Pierre Alechinsky, Hippolyte Petitjean, Nicolas De Staël, Rainer Fetting, Hiro Yamagata, Lorser Feitelson, Taro Yamamoto, Kazuo Shiraga, Alberto Burri, Anne Truitt, Jozsef Rippl-Ronai, Ronald Davis, Tsuguharu Foujita, Wols, Keith Sonnier, Henry Van De Velde, Chang Dai-Chien, Stanley Whitney, Ann Hamilton, John Brack, Jules-Alexandre Grun, Billy Apple, Eileen Agar, Benny Andrews, Moise Kisling, Edward Wadsworth, Paul Thek, Audrey Flack, Allan D'Arcangelo, Rik Wouters, Charles Cottet, Gene Davis, Prudence Heward, Alexander Liberman, David Batchelor, Tadanori Yokoo, Frederick Sommer, Hedda Sterne, Othon Friesz, Roderic O'Conor, Santiago Rusinol, Richard Gerstl, Marianne Von Werefkin, Octavio Ocampo, Kay Sage, Jessica Stockholder, Gabriele Munter, Jean Benoit, May Wilson, Jean Paul Lemieux, Jack Tworkov, Abraham Manievich, Perle Fine, Renato Guttuso, Al Held, Martial Raysse, Le Pho, Charles Reiffel, Bernard Cohen, Rosalyn Drexler, Ilya Mashkov, Jack Youngerman, Ernst Wilhelm Nay, Adja Yunkers, Leon Spilliaert, Valerio Adami, Karl Benjamin, Luigi Serafini, Leon Polk Smith, Oswaldo Guayasamin, Blinky Palermo, Ferdinand Du Puigaudeau, Esteban Vicente, Matsutani, Oscar Dominguez, Yasuo Kuniyoshi, Sergei Parajanov, Joy Hester, Forrest Bess, Taro Okamoto, Maurice Tabard, Yiannis Moralis, Igor Grabar, Alex Hay, Albert Irvin, Amadeo De Souza-Cardoso, Robert Swain, Bradley Walker Tomlin, Kishio Suga, Carlos Almaraz, Manuel Alvarez Bravo, Dan Christensen, Cyril Power, Marcel Barbeau, Jeremy Moon, Jorge Castillo, Josef Capek, Maggie Laubser, John Altoon, Albert Dubois-Pillet, William Baziotes, Joseph Marioni, Michael Hafftka, Raoul Ubac, Tony Feher, Walter Battiss, Friedel Dzubas, Varlin, Alfred Manessier, Ron Gorchov, Tony Scherman, Alina Szapocznikow, Nikolaos Lytras, Carl Holsøe, Constantin Brâncuşi, Walter Osborne, Max Kurzweil, Jose Guerrero, Leon Underwood, Istvan Nagy, Albert Bloch, Ward Jackson, Piero Dorazio, Giorgio Griffa, Lourdes Castro, Lita Albuquerque, Thomas Downing, Pierre Tal-Coat, Mario Prassinos, Panayiotis Tetsis, Robert Goodnough, Paul Feeley, Michel Majerus, Marc Vaux, Konstantinos Maleas, Vladimir Dimitrov, Meijer De Haan, Guido Molinari, Arthur Beecher Carles, Bertalan Por, Christo Coetzee, Jammie Holmes, Lasar Segall, Enrico Donati, Jerzy Nowosielski, Gianfranco Baruchello, Luis Feito, Burhan Dogancay, Iosif Iser, Charles Gibbons, Thalia Flora-Karavia, Aldo Mondino, Pierre Daura, Josef Sima, Nikola Tanev, Konrad Klapheck, Theophrastos Triantafyllidis, Edvard Weie, Gerard Fromanger, Matthias Laurenz Gräff, Victor Servranckx, Istvan Farkas, Ramon Oviedo, Manabu Mabe, Grégoire Michonze, Stanisław Ignacy Witkiewicz, Abidin Dino, Esteban Frances, Alberto Sughi, Olga Albizu, Behjat Sadr, Jose De Guimaraes, Robert Nickle, Dale Hickey, Inigo Manglano-Ovalle, Antonio Carneiro, Horia Damian, Jacqueline Hick, Kuno Gonschior, Huguette Arthur Bertrand, Ethel Léontine Gabain, Helen Dahm, Ion Nicodim, Lucy Ivanova, Gil Teixeira Lopes, Michel Carrade, Florin Maxa, Jean-Paul Jerome, Vangel Naumovski, Graca Morais, Antonio Areal, Petros Malayan, Rodolfo Arico, Stefan Sevastre, Johannes Sveinsson Kjarval, Ilka Gedo, Lucia Demetriade Balacescu, Natalia Dumitresco, Rene Bertholo, Vasile Kazar, Petre Abrudan, Aurel Cojan, Tia Peltz, Alvaro Lapa

M COMPUTE USED

Estimated Computational Cost of the Optimal Experiment The computational cost to train the popular text-to-image model Stable Diffusion was $600K (Bastian, 2022). For the optimal experiment, we would need to train $\mathcal{O}(\log m)$ models, where m is the total number of available images of a concept Z^j. So if a concept has 100,000 images, then we need to train $\log_2(100,000) = 16$ models to optimally estimate the imitation threshold, which would cost about $10M.

MIMETIC^2's Computational Cost We use 8 L40 GPUs to generate images for the all text-to-image models in our work. Overall, we use them for 16 hours per prompt, per dataset, per model to generate images. We downloaded the images on the same machine using 40 CPU cores, a process that took about 8 hours per dataset. For generating the image embeddings, we use the same 8 L40 GPUs, a process that took about 16 hours per dataset. The computation of imitation score and plotting are done on single CPU core on the same machine, a process that takes less than 30 minutes per dataset.