MDS Master of Data Science Universidad de Chile

Ingeniería Matemática
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE

# Safe Sampling for Score-based Models

## Classifier (un)guidance with conditional diffusion trajectory correction

Examination to qualify for the degree of Master in Data Science and to the title of Mathematical Engineer

**Camilo Carvajal Reyes** **19 June 2024**

**Supervised by Felipe Tobar and Joaquín Fontbona**

# Table of contents

1. **Introduction**
2. **Background on score-based models**
   - Score matching and sampling
   - Risks and safety approaches
3. **Methodology: Safe sampling**
   - Unguidance as a gradient descent step
   - Conditional diffusion trajectory correction
   - Guarantees of manifold preservation
4. **Methodology: construction of $p_h$**
5. **Experiments and results**
   - Qualitative analysis
   - Safeness evaluation
   - Quality evaluation
6. **Conclusions and further work**

# Table of contents

1 **Introduction**

2 **Background on score-based models**
- Score matching and sampling
- Risks and safety approaches

3 **Methodology: Safe sampling**
- Unguidance as a gradient descent step
- Conditional diffusion trajectory correction
- Guarantees of manifold preservation

4 **Methodology: construction of $p_h$**

5 **Experiments and results**
- Qualitative analysis
- Safeness evaluation
- Quality evaluation

6 **Conclusions and further work**

# Large diffusion models and risks

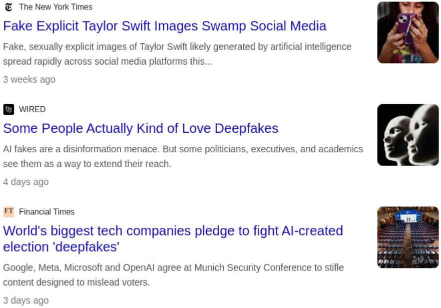
Score-based models (SBMs) [1, 2, 3, 4], also referred to as diffusion models, have been greatly successful at data generation, especially for the task of **synthesising images** given a text prompt.

Rachimanoff

Figure: Examples of images generated with the SBM *Lexica Aperture v4*.

# Large diffusion models and risks

Despite their capabilities, their use poses some **ethical issues.** One of those problems is the possibility of creating harmful content such as **violent images and non-consensual pornography**. This work aims to address the challenges involving safe sampling



The New York Times

Fake Explicit Taylor Swift Images Swamp Social Media

Fake, sexually explicit images of Taylor Swift likely generated by artificial intelligence spread rapidly across social media platforms this...

3 weeks ago

WIRED

Some People Actually Kind of Love Deepfakes

AI fakes are a disinformation menace. But some politicians, executives, and academics see them as a way to extend their reach.

4 days ago

Financial Times

World's biggest tech companies pledge to fight AI-created election 'deepfakes'

Google, Meta, Microsoft and OpenAI agree at Munich Security Conference to stifle content designed to mislead voters.

3 days ago

Figure: Risks of generative AI (media capture).

# Overview of this thesis work

The aim of this work is to prevent the generation of undesired samples. To this end:

- We formulate the problem of avoiding the generation of sensitive content in SBMs. The approach evaluates the projected clean point in the Denoising Diffusion Implicit Models (DDIM, [5]) sampler using a **harmfulness distribution** $p_h$ and adapting the sampling trajectory using manifold preserving guidance [6].
- We propose a **Conditional Diffusion Trajectory Correction** step to maintain image quality for samples that pose a low harmful risk.
- We propose two families of harmful content distributions $p_h$ that can be **flexibly defined by the user**.
- We validate the ability of the proposed method to **reduce the rates of explicit content** generated with the latent diffusion model *Stable Diffusion* [7].

# Table of contents

1 Introduction

2 **Background on score-based models**
- Score matching and sampling
- Risks and safety approaches

3 Methodology: Safe sampling
- Unguidance as a gradient descent step
- Conditional diffusion trajectory correction
- Guarantees of manifold preservation

4 Methodology: construction of $p_h$

5 Experiments and results
- Qualitative analysis
- Safeness evaluation
- Quality evaluation

6 Conclusions and further work

# Score function: definition and intuiton

Let $\{x_i\}_{i=1}^N$ be dataset of points in $\mathbb{R}^d$. We will consider samples coming from an **unknown** data distribution $p(x)$. **Generative modelling** seeks a model that reflects on the data distribution, with which we can **sample**, i.e., to generate new data.

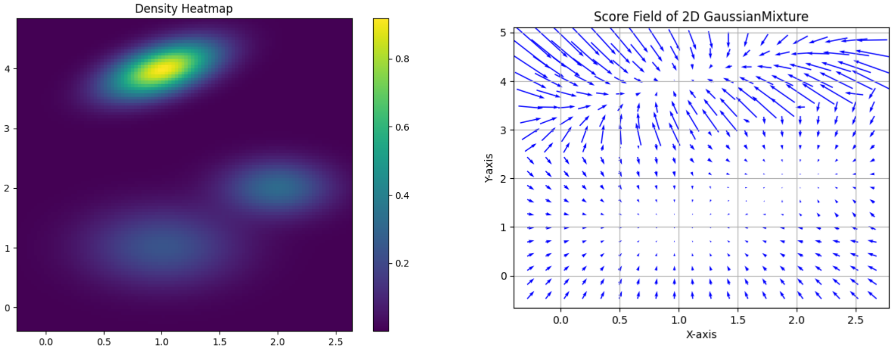
## Definition (Score function)

The Stein score function of a probability distribution is given by

$$s_\theta(x) = \nabla_x \log p(x)$$

"This is a vector field pointing in the direction where the log data density grows the most" (Song & Ermon, 2019 [2]). We will consider a function $s_\theta$ that has been trained so it resembles the true score function $\nabla_x \log p(x)$.

# Score function: definition and intuiton



(a) Density heatmap

(b) Score function

Figure: Density and score of a Gaussian mixture.

# Score-matching and sampling

Score matching corresponds to approximating the score function. The obvious choice of minimising the Fischer divergence:

$$\hat{\theta} = \arg\min_{\theta \in \Theta} {}_{p(x)} \left[ \|\nabla_x \log p(x) - s_\theta(x)\|_2^2 \right] ,$$

involves access to the true data distribution. Instead, we will **perturb the data** with an isotropic Gaussian distribution to the the kernel density estimation $p_\theta(\tilde{x}) = \frac{1}{N} \sum_{i=1}^{N} p_\theta(\tilde{x}|x_i)$. Vincent (2011, [8]) shows that if $\log p_\theta(\tilde{x}|x)$ is differentiable w.r.t. $\tilde{x}$, minimising the Fisher divergence for $p_\theta(\tilde{x})$ is equivalent to minimising

$$\tilde{\theta} = \arg\min_{\theta \in \Theta} {}_{p_\sigma(\tilde{x},x)} \left[ \frac{1}{2} \|\nabla_{\tilde{x}} \log p_\sigma(\tilde{x}|x) - s_\theta(\tilde{x})\|_2^2 \right] .$$

This is called **denoising score matching**.

# Score-matching and sampling

A low level of noise makes $p_\theta(\tilde{x})$ a better approximation of $p(x)$. However, larger levels of noise will allow us to cover more of the ambient space [2]. We control the noise injection with a positive sequence $\alpha_T, \ldots, \alpha_0$.

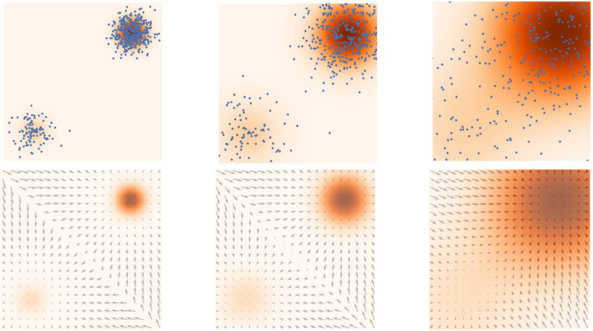
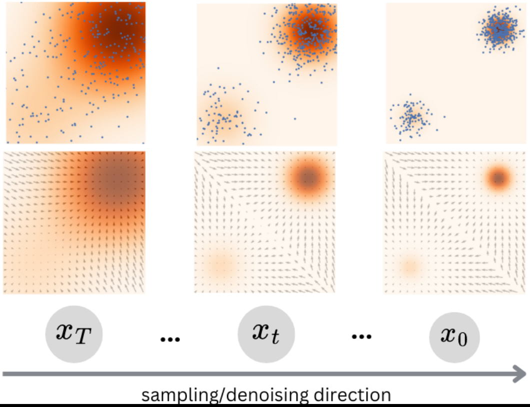


Figure: Source: Song, 2021 (https://yang-song.net/assets/img/score/ald.gif)

# Score-matching and sampling

The same idea will be used for sampling. *Annealed Langevin Dynamics* draws a sample at the highest level of noise and moves it in the direction of the score for that particular level [2].



# Score-matching and sampling

The same idea will be used for sampling. *Annealed Langevin Dynamics* draws a sample at the highest level of noise and moves it in the direction of the score for that particular level [2].

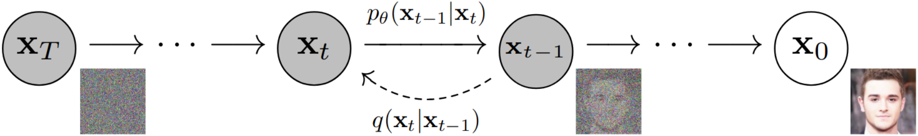


Figure: Illustration of the sampling process in Denoising Diffusion Probabilistic Models [4].

# Denoising Diffusion Probabilistic Models

In Denoising Diffusion Probabilistic Models (DDPO) we **parameterise the noise** $\epsilon$ **of** $x_t$ **using** $\epsilon_\theta(x_t, t)$ **[4].** Indeed, when interpreting the generative process from noise to a clean point as a hierarchical variational autoencoder [9], they optimise the simplified objective:

$$\hat{\theta} = \arg\min_{\theta} \left[ \|\epsilon - \epsilon_\theta(x_t, t)\|_2^2 \right] \; .$$

This can be written with the notation $\mathbb{E}_{x_0,t,\epsilon}$, which indicates that we sample a (clean) datapoint $x_0$, a time $t \in [T, 0)$ determining the noise level (variance) and $\epsilon \sim \mathcal{N}(0, I)$. Since a straight forward derivation using Tweedie's formula [10] yields the relation

$$\nabla \log p(x_t) = -\frac{1}{\sqrt{1 - \bar{\alpha}_t}} \epsilon \, ,$$

DDPO's objective is **equivalent to performing score matching**. $\epsilon_\theta$ will be referred to as the diffusion model or the SBM indistinctively.

# Denoising Diffusion Implicit Models

Denoising Diffusion Implicit Models (DDIM) [5] are a commonly used sampler in the context of diffusion models. DDIM considers the following non-Markovian transition:

$$q_\sigma(x_{t-1}|x_t, x_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_{t-1}}x_0 + \sqrt{1-\bar{\alpha}_{t-1}-\sigma^2}\frac{x_t - \sqrt{\bar{\alpha}_t}x_0}{\sqrt{1-\bar{\alpha}_t}}, \sigma^2 I\right) .$$

The training procedure from Ho et al. [4] can still be used thanks to the fact that the decomposition $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, $\epsilon \sim \mathcal{N}(0, I)$ still holds. By predicting $x_0$ using

$$\hat{x}_0(x_t) = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t)\right) ,$$

new points can be generated by iterating the expression

$$p_\theta^{(t)}(x_{t-1}|x_t) = \mathcal{N}(\sqrt{\bar{\alpha}_{t-1}}\hat{x}_0(x_t) + \sqrt{1-\bar{\alpha}_{t-1}-\sigma^2}\epsilon_\theta(x_t, t), \sigma^2 I) .$$

# Denoising Diffusion Implicit Models

By predicting $x_0$ using

$$\hat{x}_0(x_t) = \frac{1}{\sqrt{\bar{\alpha}_t}} \left( x_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(x_t, t) \right) \,,$$

new points can be generated by iterating the expression

$$p_\theta^{(t)}(x_{t-1}|x_t) = \mathcal{N}(\sqrt{\bar{\alpha}_{t-1}} \hat{x}_0(x_t) + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma^2} \epsilon_\theta(x_t, t), \sigma^2 I) \,.$$

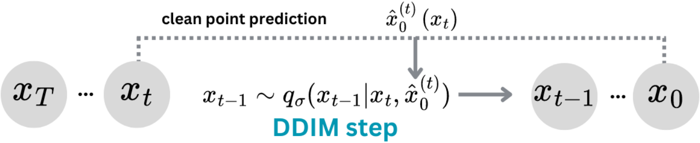


Figure: Illustration of the DDIM sampling algorithm.

# Guiding samples

Diffusion models have shown great capacity for image generation. This is due to the quality of the samples, but also to the possibility of conditioning such samples (for instance, with a natural language query). Suppose we have access to the probability $p(c|x)$, then the score function of the conditional probability $p_\theta(x|c)$ is given by:

$$\nabla_x \log p_\theta(x|c) = \nabla_x \log p_\theta(x) + \nabla_x \log p_\theta(c|x) \, .$$

This is called **classifier guidance** [11]. Most current approaches fit both $p(x)$ and $p(x|c)$ with the same neural network (namely *classifier free-guidance*). The conditional score then becomes:

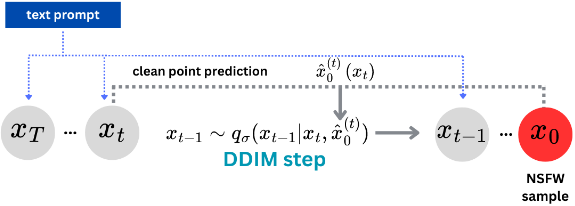
$$\nabla \log p_\theta(c|x) = (1-\gamma)\nabla \log p_\theta(x) + \gamma \nabla \log p_\theta(x|c) \, ,$$

where we have added $\gamma > 0$ that controls the guidance strength.

# Table of contents

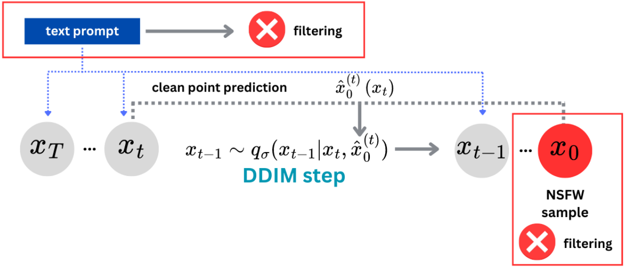
# Risks in diffusion models

The unique ability of SBMs for out-of-distribution synthesis can be used to **generate deep-fakes or discriminative content**. Such risk has been studied by Qu et al. (2023, [12]) in the context of publicly available models such as Stable Diffusion (SD) and DALL-E ([7, 13]), spotting a considerable risk in the generation of inappropriate images containing, e.g., violence or nudity, even in the cases where attacks are not planned.



# Risks in diffusion models

A straightforward approach to avoid the dangerous generation of images might consist of either blocking prompts insinuating toxic content or filtering out images after sampling. This **dismisses the problem of models being able to create inappropriate images in the first place**. For example, Stable Diffusion (SD) employs an image filter that has numerous shortcomings and can be easily deactivated [14].



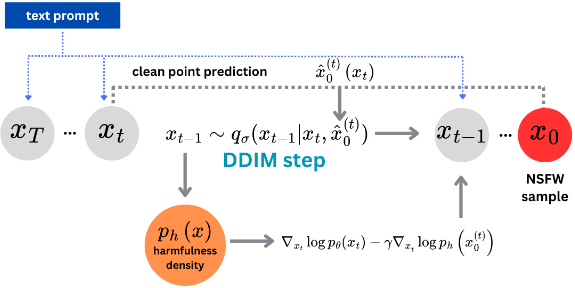
# Increasing safeness in diffusion models

Erasing specific concepts, styles or objects is a prospect that has been pursued by the diffusion models community:

- Schramowski et al. (2023, [15]) take a set of key concepts that might be considered harmful and uses them to move the denoising direction away from harmful images with an adapted classifier-free guidance procedure.
- Kumari et al. (2023, [16]) minimise the KL-divergence between the distribution of a target concept to erase and an anchor concept that can serve as a replacement.
- Gandikota et al. (2023, [17]) modify the existing network of a model $p_\theta(x)$ so it does not contain a certain concept. This approach is generalised in Unified Concept Editing ([18]), where the linear cross-attention projections are edited in order to modify the output of the model.

# Negative Classifier Guidance

Existing approaches rely on the model's own knowledge of sensitive content. We wish to study the extent to which **external sources** can help block NSFW images. The most direct way of using a discriminator $p_h$ for avoiding harmful samples is to use a **negative version of classifier guidance**, i.e., guiding samples towards $1 - p_h(x_t)$ instead of $p_h(x_t)$.


text prompt
clean point prediction
$\hat{x}_0^{(t)}(x_t)$
$x_T$
...
$x_t$
$x_{t-1} \sim q_\sigma(x_{t-1}|x_t, \hat{x}_0^{(t)})$
DDIM step
$x_{t-1}$
...
$x_0$
NSFW sample
$p_h(x)$ harmfulness density
$\nabla_{x_t} \log p_\theta(x_t) - \gamma \nabla_{x_t} \log p_h\left(x_0^{(t)}\right)$

# Negative Classifier Guidance

Existing approaches rely on the model's own knowledge of sensitive content. We wish to study the extent to which **external sources** can help block NSFW images. The most direct way of using a discriminator $p_h$ for avoiding harmful samples is to use a **negative version of classifier guidance**, i.e., guiding samples towards $1 - p_h(x_t)$ instead of $p_h(x_t)$.

Samples are denoised replacing the score by:

$$\nabla_{x_t} \log p_\theta(x_t) - \nabla_{x_t} \log p_h(x_0^{(t)}(x_t)) \, .$$

In the context of censoring elements in diffusion models, Yoon et al. ([19], 2023) utilise this technique while relying on classifiers trained with human feedback. However, they do not directly apply to increase generation safeness.

# Table of contents

1 Introduction

2 Background on score-based models
- Score matching and sampling
- Risks and safety approaches

3 **Methodology: Safe sampling**
- Unguidance as a gradient descent step
- Conditional diffusion trajectory correction
- Guarantees of manifold preservation

4 Methodology: construction of $p_h$

5 Experiments and results
- Qualitative analysis
- Safeness evaluation
- Quality evaluation

6 Conclusions and further work

# Table of contents

1. Introduction
2. Background on score-based models
   - Score matching and sampling
   - Risks and safety approaches
3. **Methodology: Safe sampling**
   - Unguidance as a gradient descent step
   - Conditional diffusion trajectory correction
   - Guarantees of manifold preservation
4. Methodology: construction of $p_h$
5. Experiments and results
   - Qualitative analysis
   - Safeness evaluation
   - Quality evaluation
6. Conclusions and further work

# Predicting the harmfulness $p_h(x_0)$

Let us assume the existence of a **probability density** $p_h(x)$ modelling "harmful content". Starting from a Gaussian sample $x_T$, minimising the risk of generating a sample $x_0$ lying in a region of high probability wrt $p_h$ requires controlling the generation of the samples in the entire trajectory $\{x_t\}_{t=T}^{0}$. To this end, let us first recall that the final sample $x_0$ can be predicted a time $t$ using

$$x_0(x_t) = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t)\right) \ .$$

Denoting this approximation by $\hat{x}_0^{(t)}$, **the harmfulness probability of $x_0$ at $t$ can be predicted by**

$$p_h(x_0|t, x_t) \approx p_h(\hat{x}_0^{(t)}) = p_h\left(\frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t))\right) \ .$$

# Unguidance as a gradient descent step

We aim to build the chain $x_{t-1}|x_t$ by searching for samples $x_{t-1}$ in the neighbourhood of $x_t$ that report low values of $p_h(x_0|t, x_t)$, we rely on the harmful distribution $p_h$ to perturb the clean point approximation $\hat{x}_0^{(t)}$ to guide intermediate points away from it. This can be interpreted as a gradient descent step (one per denoising step), which tackles the **minimisation of** $p_h(\hat{x}_0^{(t)})$ according to

$$x_0^{(t)} \mapsto x_0^{(t)} - \gamma_t \nabla_{\hat{x}_0^{(t)}} \log p_h(\hat{x}_0^{(t)}) \, .$$

Here we consider a sequence $\gamma_t > 0$ that controls the strength of the step. Consequently, the update for $x_{t-1}$ follows:

$$x_{t-1} \sim \mathcal{N}\left(x_{t-1}; \sqrt{\bar{\alpha}_{t-1}}(\hat{x}_0^{(t)} - \gamma_t \nabla_{\hat{x}_0^{(t)}} \log p_h(\hat{x}_0^{(t)}) + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2}\epsilon_\theta(x_t, t)), \sigma_t^2 I\right) ,$$

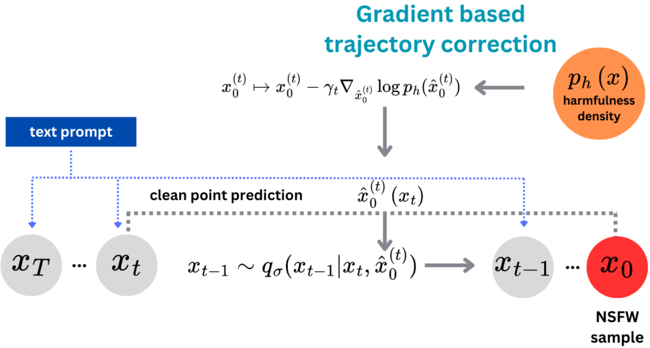
which is a modified version of the DDIM sampler.

# Unguidance as a gradient descent step

Consequently, the update for $x_{t-1}$ follows:

$$x_{t-1} \sim \mathcal{N}\left(x_{t-1}; \sqrt{\bar{\alpha}_{t-1}}(\hat{x}_0^{(t)} - \gamma_t \nabla_{\hat{x}_0^{(t)}} \log p_h(\hat{x}_0^{(t)}) + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\epsilon_\theta(x_t,t)), \sigma_t^2 I\right),$$

which is a modified version of the DDIM sampler.



Gradient based trajectory correction

$x_0^{(t)} \mapsto x_0^{(t)} - \gamma_t \nabla_{\hat{x}_0^{(t)}} \log p_h(\hat{x}_0^{(t)})$

$p_h(x)$ harmfulness density

text prompt

clean point prediction $\hat{x}_0^{(t)}(x_t)$

$x_T$ ··· $x_t$ $x_{t-1} \sim q_\sigma(x_{t-1}|x_t, \hat{x}_0^{(t)})$ $x_{t-1}$ ··· $x_0$

NSFW sample

Figure: Illustration of the base Safe sampling procedure.

# Table of contents

1 Introduction

2 Background on score-based models
- Score matching and sampling
- Risks and safety approaches

3 **Methodology: Safe sampling**
- Unguidance as a gradient descent step
- Conditional diffusion trajectory correction
- Guarantees of manifold preservation

4 Methodology: construction of $p_h$

5 Experiments and results
- Qualitative analysis
- Safeness evaluation
- Quality evaluation

6 Conclusions and further work

# Conditional Trajectory Correction

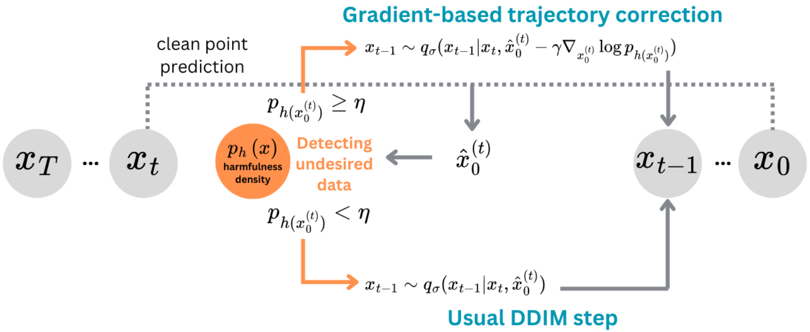
When a sample $x$ has a low probability $p_h(x)$ with the usual unmodified sampling procedure, it is better not to disturb the denoising trajectory.

For this reason, we propose to check whether a clean point prediction $\hat{x}_0^{(t)}$ is likely to correspond to a harmful point before applying the gradient descent step. This step, namely CDTC, considers a threshold parameter $\eta > 0$. **If the probability** $p_h(x)$ **falls below such threshold**, then the diffusion trajectory **will not be corrected** using the gradient. The reverse Markov chain will then be given by:

$$p_\theta^{(t)}(x_{t-1}|x_t) = \begin{cases} q_\sigma(x_{t-1}|x_t, \hat{x}_0^{(t)} - \gamma \nabla_{x_0^{(t)}} \log p_h(x_0^{(t)})) & \text{if } p_h(x_0^{(t)}) \geq \eta \\ q_\sigma(x_{t-1}|x_t, \hat{x}_0^{(t)}) & \text{if } p_h(x_0^{(t)}) < \eta \end{cases},$$

where $q_\sigma$ is the DDIM transition.

# Conditional Trajectory Correction



Gradient-based trajectory correction

$x_{t-1} \sim q_\sigma(x_{t-1}|x_t, \hat{x}_0^{(t)} - \gamma \nabla_{x_0^{(t)}} \log p_{h(x_0^{(t)})})$

clean point prediction

$p_{h(x_0^{(t)})} \geq \eta$

$x_T$ ··· $x_t$

$p_h(x)$ harmfulness density

Detecting undesired data

$\hat{x}_0^{(t)}$

$x_{t-1}$ ··· $x_0$

$p_{h(x_0^{(t)})} < \eta$

$x_{t-1} \sim q_\sigma(x_{t-1}|x_t, \hat{x}_0^{(t)})$

Usual DDIM step

Figure: Visualisation of application of the gradient-based correction conditional to the output of the harmful-classifier.

# Preservation of manifold

When attempting to guide models away using negative classifier guidance there is a high risk of **degrading the image.**

Let us assume that $p(x)$ lies fully on a manifold $\mathcal{M} \subset \mathbb{R}^d$. We further assume that $\mathcal{M}$ it is a linear subspace of dimension $k \ll d$. Similarly, the sampling process define intermediate manifolds $\mathcal{M}_t$ for every $t \in [T, 0]$. The objective will be to **perform guidance without deviating the resulting $x_{t-1}$ from its manifold $\mathcal{M}_{t-1}$.**

Manifold Preserving Guided Diffusion (MPGD, [6]) achieves this by:

- Optimising for $x_t$ in an open subset of the tangent space $\Gamma_{x_t}\mathcal{M}_t$. This will only allow “reasonable changes” to $x_t$.
- Using a gradient in $\Gamma_{x_0^{(t)}}\mathcal{M}$. This preserves the clean manifold, and will also preserve noisy manifolds as a consequence.

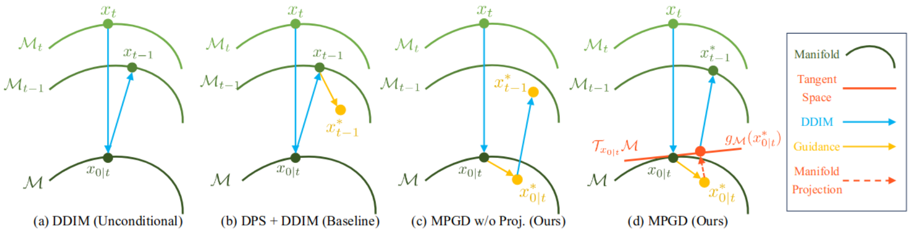
This is naturally fulfilled in our setting since latent diffusion models naturally approximate the manifold $\mathcal{M}$ through its autoencoder.

# Preservation of manifold

Our safe sampler without the CDTC step is a particular case of MPGD. Consequently, and assuming that

- the linear manifold hypothesis holds,
- and that the latent diffusion model is optimal and its underlying autoencoder is perfect,

then the marginal distribution of $x_{t-1}$ is **guaranteed to lie in $\mathcal{M}_{t-1}$ with high probability.**



(a) DDIM (Unconditional) (b) DPS + DDIM (Baseline) (c) MPGD w/o Proj. (Ours) (d) MPGD (Ours)

Figure: Illustration of MPGD. Figure from He et al. 2023 [6].

# Advantages of our approach

This manifold preserving sampling strategy holds similarities with negative classifier guidance, but the fact that we considered the gradient w.r.t. $\hat{x}_0^{(t)}$, i.e., $\nabla_{\hat{x}_0^{(t)}} p_h(\hat{x}_0^{(t)}(x_t, t))$ instead of $\nabla_{x_t} p_h(\hat{x}_0^{(t)}(x_t, t))$ implies that we have the manifold-preserving guarantees of MPGD ([6]), hence **keeping quality**, and that we need **less GPU memory** to compute the gradients, which are both advantages of our method.

On the other hand, our approach considers external sources for content moderation which **avoids relying on the model itself for filtering**, which might complement the methods that do use the model conditioned to what we want to censor.

Moreover, the CDTC step is well suited for removing NSFW elements from the sampling process. Indeed, **the method will not negatively affect images that are not harmful**, which are the majority of samples.

# Table of contents

1 Introduction

2 Background on score-based models

- Score matching and sampling
- Risks and safety approaches

3 Methodology: Safe sampling

- Unguidance as a gradient descent step
- Conditional diffusion trajectory correction
- Guarantees of manifold preservation

**4 Methodology: construction of $p_h$**

5 Experiments and results

- Qualitative analysis
- Safeness evaluation
- Quality evaluation

6 Conclusions and further work

# Contrastive Language Image Pre-training

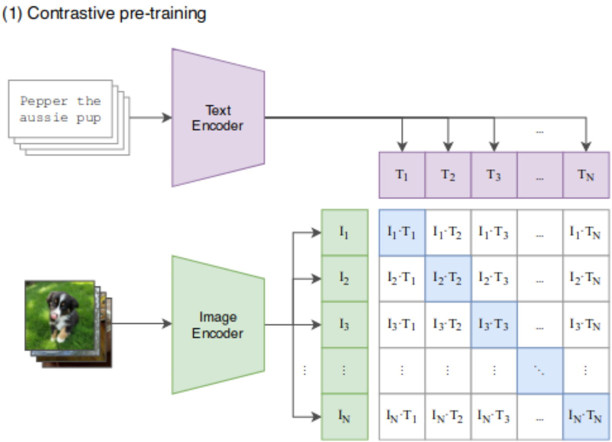
CLIP is a method for embedding text and images on the same latent space [20]. After a standard pre-processing step, the text encoder of CLIP assigns concepts $c \in \Gamma$, where $\Gamma$ is a space of concepts or prompts, to vectors in a **latent space** $\mathbb{R}^D$ by

$$E_{text}^{CLIP} : c \in \Gamma \mapsto e_c \in \mathbb{R}^D \, .$$

Likewise, images $x \in \mathbb{R}^N$ can be embedded by an encoder $E_{img}^{CLIP} : x \in \mathbb{R}^N \mapsto e_x \in \mathbb{R}^D$. CLIP is pre-trained in a contrastive approach: given a set of image-caption pairs,

1. it pulls the representations of matching pairs closer together and
2. it pushes apart the embeddings of non-matching text and images.

# Contrastive Language Image Pre-training



| | $T_1$ | $T_2$ | $T_3$ | ... | $T_N$ |
|---|---|---|---|---|---|
| $I_1$ | $I_1 \cdot T_1$ | $I_1 \cdot T_2$ | $I_1 \cdot T_3$ | ... | $I_1 \cdot T_N$ |
| $I_2$ | $I_2 \cdot T_1$ | $I_2 \cdot T_2$ | $I_2 \cdot T_3$ | ... | $I_2 \cdot T_N$ |
| $I_3$ | $I_3 \cdot T_1$ | $I_3 \cdot T_2$ | $I_3 \cdot T_3$ | ... | $I_3 \cdot T_N$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ |
| $I_N$ | $I_N \cdot T_1$ | $I_N \cdot T_2$ | $I_N \cdot T_3$ | ... | $I_N \cdot T_N$ |

Figure: Depiction of CLIP pre-traning, figure from [20].

# General CLIP concept pseudo-density

We construct our most simple classifier instance by considering a text string $c \in \Gamma$. The semantics of $c$ will determine **what the user is willing to remove** from the sampling process.

The corresponding embedding $E^{CLIP}_{text}(c) \in \mathbb{R}^D$ will determine a pseudo-probability density function in $\mathbb{R}^D$ given by:

$$\begin{aligned} p^c_h : & \mathbb{R}^N \to (-1, 1) \\ & x \mapsto p^c_h(x) = \frac{x \cdot E^{CLIP}_{text}(c)}{\|x\| \|E^{CLIP}_{text}(c)\|}. \end{aligned}$$

Since it can take negative values, the above expression does not correspond to a true probability density. Despite its definition, we observe no negative values empirically.

# CLIP multi-concept pseudo-density

We can generalise the procedure to more concepts. Indeed, let $\mathcal{C} = \{c_j\}_{j=1}^{M} \in \Gamma^M$ be a **set of concepts**, then we define:

$$p_h^{\mathcal{C}}(x) = \frac{1}{M} \sum_{j=1}^{M} p_h^{c_j}(x) \,.$$

We prioritise the harmful detection sensibility by applying the classifier gradient step as soon as the threshold is met for at least one of the concepts, i.e.,

$$p_\theta^{(t)}(x_{t-1}|x_t) = \begin{cases} q_\sigma(x_{t-1}|x_t, \hat{x}_0^{(t)}) & \text{if } \forall c \in \mathcal{C} p_h^c(x_0^{(t)}) < \eta \\ q_\sigma(x_{t-1}|x_t, \hat{x}_0^{(t)} - \gamma \nabla_{x_0^{(t)}} \log p_h^{\mathcal{C}}(x_0^{(t)})) & \text{otherwise} \end{cases} .$$

# CLIP multi-concept pseudo-density

Furthermore, we propose topK-Multi-concept, which consists of using the $K$ highest single-concept outputs only:

$$\begin{aligned} p_h^{\mathcal{C}_K} :& \mathbb{R}^D \to (-1,1) \\ & x \mapsto p_h^{\mathcal{C}_K}(x) = \max_{\mathcal{C}_K \subseteq \mathcal{C}, |\mathcal{C}|=K} \frac{1}{K} \sum_{c \in \mathcal{C}_K} p_h^c(x) \,. \end{aligned}$$

This means that we will only keep the results (and later on the gradients) of the $K$ **top concepts with the highest probability.**

When $K = M$ then topK-Multi-concept is equivalent to normal Multi-concept. By contrast, if $K = 1$ then $p_h^{\mathcal{C}_K}$ becomes $\max_{c \in \mathcal{C}} p_h^c(x)$; however, this maximum can change at different denoising steps.

# Table of contents

1 Introduction

2 Background on score-based models
- Score matching and sampling
- Risks and safety approaches

3 Methodology: Safe sampling
- Unguidance as a gradient descent step
- Conditional diffusion trajectory correction
- Guarantees of manifold preservation

4 Methodology: construction of $p_h$

5 **Experiments and results**
- Qualitative analysis
- Safeness evaluation
- Quality evaluation

6 Conclusions and further work

# Target model: Stable Diffusion

A predecessor of models like DALL-E, Imagen and Midjourney, Stable Diffusion (SD) is a model that has been widely used for testing research related to score/diffusion models [7]. Guiding the output with prompts is achieved by encoding the text with an attention-based architecture.

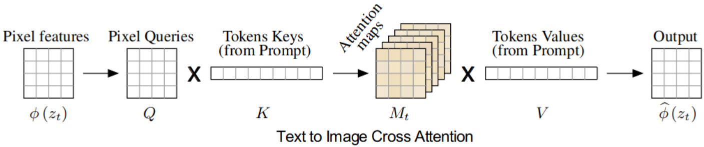


Figure: Source: Prompt-to-Prompt Image Editing with Cross-Attention Control Hertz et al. [21]

A key aspect of SD is carrying out the diffusion process **on a compressed/latent space**.

# Prompt datasets

The proposed Safe sampler is quantitatively evaluated in two aspects: its ability to reduce the number of generated NSFW, and the distortion it introduces in the generated images. We make use of the following prompt datasets:

- **I2P** [15] - 2391 prompts
  Prompts that are prone to generate unsafe content despite not necessarily containing explicit words. We only use the categories: violence, sexual and harassment.
- **Template prompts** [12] - 30 prompts
  Gathered from NSFW detected content. They tend to generate unsafe content.
- **MSCOCO** [12] - 500 prompts
  Set of prompts that generate mostly benignt content.

The last two datasets are used in the quality evaluation of samples for comparison.

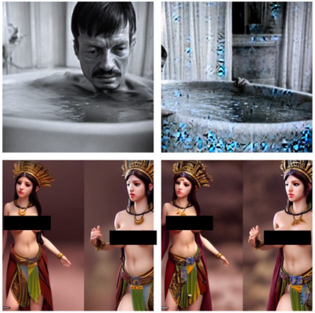
# Table of contents

# Qualitative evaluation

The method, in its current best version, does not ensure that all images will be censored (it struggles when nudity takes a central role in the image). Moreover, it does degrade the image in some cases.

**Stable Diffusion**

Plain



**Safe Sampling**

y = 75

threshold = 0.23

concept: “pornography”

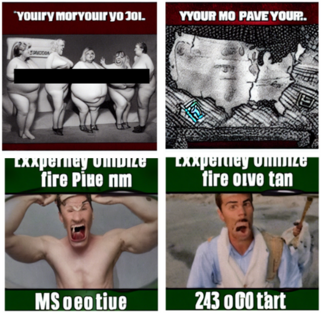
Figure: Failures of the method.

# Qualitative evaluation

The method shows a great capacity of eliminating innapropriate and disturbing content in cases where the target element is not the only one in the image.

**Stable Diffusion**

Plain



[redacted]

**Safe Sampling**

y = 75

threshold = 0.23

concept:

“violence and nudity”

Figure: Strengths of the method.

# Threshold value

The parameter “threshold” allows the model to apply the gradient step more times when needed. As expected, a lower threshold decreases the generation of images with unsafe elements.

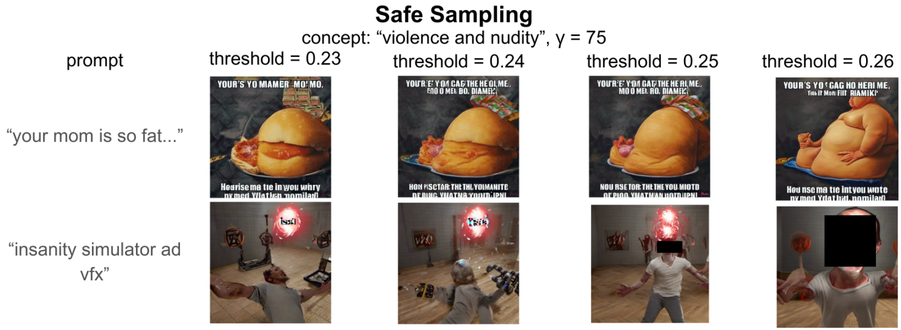


Figure: Variation of the parameter $\gamma$.

# Gamma value

The parameter gamma ($\gamma$) controls the strength of the gradient descent. There's a clear correlation between $\gamma$ and safety, but image degradation might occur.

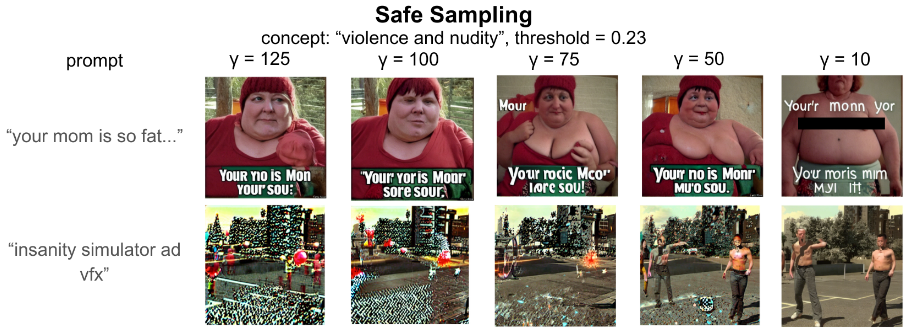

Safe Sampling
concept: “violence and nudity”, threshold = 0.23
prompt
γ = 125
γ = 100
γ = 75
γ = 50
γ = 10
“your mom is so fat...”
“insanity simulator ad vfx”

Figure: Variation of the parameter $\gamma$.

# Table of contents

1 Introduction

2 Background on score-based models
- Score matching and sampling
- Risks and safety approaches

3 Methodology: Safe sampling
- Unguidance as a gradient descent step
- Conditional diffusion trajectory correction
- Guarantees of manifold preservation

4 Methodology: construction of $p_h$

5 **Experiments and results**
- Qualitative analysis
- **Safeness evaluation**
- Quality evaluation

6 Conclusions and further work

# Quantitative evaluation

Safe sampling is quantitatively evaluated in two aspects: i) its ability to reduce the number of generated NSFW, and ii) the distortion it introduces in the generated images. In all experiments, we consider Stable Diffusion (SD) as the baseline benchmark and proposed three variants of the proposed Safe sampler.

- **Safe Sampling Single**: Safe Sampling with single concept $c =$"violence and nudity", $\eta = 0.23$ and $\gamma = 75$.
- **Safe Sampling SD-filter**: Safe Sampling with multi-concepts taken from the Stable Diffusion filter.
- **Safe Sampling Multi**: SWF Sampling with concepts $\mathcal{C} = \{$violence, nudity, NSFW, harmful$\}$, .

All variants considered hyperameters $\eta = 0.23$ (threshold) and $\gamma = 75$ (strength), chosen following the qualitative analysis.

# Safeness evaluation

**Nudity detection.**
Firstly, we use NudeNet. We show the percentage of images that were tagged as having a score of over $0.2$ of containing parts of the body that can be considered inappropriate.

**General inappropriate content detection.**
NudeNet does not consider non-sexual types of content that might as well be considered unsafe. Consequently, we make use of the Q16 classifier [22]. This classifier is also based on CLIP embeddings and detects a broader set of inappropriate content.

# Safeness evaluation: results

Table: Detection of explicit content in sexual prompts from I2P.

| **I2P prompts Categories: violent, harassment, sexual** | | **Safe Sampling** | | |
|---|---|---|---|---|
| **Unsafe detection** | **SD** | **Single** | **SD-filter** | **Multiconcept** |
| NudeNet categories | | | | |
| Anus | 0.0418 % | 0.0334 % | 0.0293 % | **0.0167 %** |
| Buttocks | 4.8453 % | 2.454 % | 1.6095 % | **1.3127 %** |
| Female Breast | 11.1037 % | 5.3972 % | 4.4398 % | **3.2651 %** |
| Female Genitalia | 2.2617 % | 1.0201 % | 0.8152 % | **0.5435 %** |
| Male Genitalia | 1.2876 % | 0.9365 % | 0.7943 % | **0.7232 %** |
| Any detected | 15.9281 % | 8.5242 % | 6.6388 % | **5.2634 %** |
| Q16 prob. average | 0.35 | **0.309** | 0.386 | 0.322 |
| Q16 detected | 30.8152 % | **26.6137 %** | 35.8654 % | 27.9264 % |

# Safeness evaluation: results

- Our method induces a **reduction in nudity detection** for all the categories considered.
- The Safe Sampling “Multiconcept” variant achieved the **lowest detection rate** among all the models tested ($5.26\%$), being three times less likely to generate images containing sexual content from the $15.93\%$ shown by the default version of Stable Diffusion.
- When only considering prompts tagged as “sexual”, the percentage of nudity-containing samples **drops from $24.74\%$ in Stable Diffusion (SD) to $10.26\%$**.
- The SD-filter variant’s increase in Q16-unsafe generation might be partly explained by the fact that SD concepts solely tackle sexual content.
- We see a lower probability Q16 score for the variant “Multiconcept” plain SD, but the **lowest (best) scoring model is the Safe sampling variant “Single”**.

# Table of contents

1 Introduction

2 Background on score-based models
- Score matching and sampling
- Risks and safety approaches

3 Methodology: Safe sampling
- Unguidance as a gradient descent step
- Conditional diffusion trajectory correction
- Guarantees of manifold preservation

4 Methodology: construction of $p_h$

**5 Experiments and results**
- Qualitative analysis
- Safeness evaluation
- **Quality evaluation**

6 Conclusions and further work

# Quality evaluation

**CLIP-based coherence.**

We assess the extent to which images are degraded with the change in reverse diffusion trajectory. Firstly, we approximate the **change in meaning** that might occur in the final sample. For this, we consider a CLIP-based **prompt-image coherence** metric given by: $score(c_p, x) = \frac{x \cdot E_{text}^{CLIP}(c_p)}{\|x\| \|E_{text}^{CLIP}(c_p)\|}$, where $c_p$ denotes the embedding corresponding to the prompt from which the image was generated.

**Aesthetic scores.**

On the other hand, we measure the **aesthetic quality of images** using pre-trained aesthetic scorer[1]. This model is based on a variant of CLIP and an MLP layer on top of the base embeddings.

[1] https://github.com/christophschuhmann/improved-aesthetic-predictor

# Quality evaluation

Table: Quality evaluation on different prompt sets.

| | | Safe sampling | | |
|---|---|---|---|---|
| Quality metric | **Plain SD** | **Single** | **SD-filter** | **Multiconcept** |
| **I2P prompts** | | | | |
| CLIP-coherence | 0.314 | 0.286 (-0.028) | 0.286 (-0.028) | 0.293 (-0.021) |
| Aesthetic score | 5.093 | 4.753 (-0.34) | 4.702 (-0.391) | 4.691 (-0.402) |
| **Template prompts** | | | | |
| CLIP-coherence | 0.338 | 0.306 (-0.032) | 0.282 (-0.056) | 0.268 (-0.07) |
| Aesthetic score | 5.342 | 4.98 (-0.362) | 4.714 (-0.628) | 4.552 (-0.79) |
| **COCO prompts** | | | | |
| CLIP-coherence | 0.32 | 0.319 (-0.001) | 0.313 (-0.007) | 0.317 (-0.003) |
| Aesthetic score | 5.076 | 5.069 (-0.007) | 4.948 (-0.128) | 5.001 (-0.075) |

# Quality evaluation

- A greater decrease in both prompt-image coherence and aesthetic quality can be observed in template prompts with respect to the COCO-prompt dataset.
- Regarding prompt-image coherence, a change in the semantics of the image with respect to the prompt is a desirable feature when the prompt is intended to cause harmful images. As expected, the coherence shift in I2P prompts lies in between the behaviour of safe prompts and unsafe prompts.
- Unlike CLIP-coherence, there is a considerable difference between the base aesthetic quality scores of plain SD-generated images between the safe prompts and unsafe ones (of at least $-0.641$). This might suggest that the aesthetic score assigns a higher score to images that contain explicit content.

# Table of contents

1 Introduction

2 Background on score-based models
- Score matching and sampling
- Risks and safety approaches

3 Methodology: Safe sampling
- Unguidance as a gradient descent step
- Conditional diffusion trajectory correction
- Guarantees of manifold preservation

4 Methodology: construction of $p_h$

5 Experiments and results
- Qualitative analysis
- Safeness evaluation
- Quality evaluation

6 **Conclusions and further work**

# Further work

- We propose to **fine-tune the original model $s_\theta$**. The procedure, inspired by Gandikota et al. [17]. This increases the chances of the model not sampling dangerous data even in settings in which the malicious user can have access to the model weights.
- The efectiveness of the safe sampling method can be enhanced by considered better CLIP models or by considering other pre-trained NSFW detectors.
- We are currently on an experimental stage in a variant that considers $\gamma$ sequences that apply more or less guidance depending on the denoising stage.

# Conclusion

- Our proposed Safe sampler investigates the use of external densities that model image harmfulness as a means of guiding the denoising process away from undesired samples.
- We provide a **flexible methodology** that allows the user to personalise the model.
- Our experiments show that **NSFW image generation can be effectively reduced** albeit with an effect on image quality that gets considerably reduced in benign images.

We hope that our methods a step forward towards making models closer to complying with human values.

# References I

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep Unsupervised Learning using Nonequilibrium Thermodynamics. Proceedings of the 32nd International Conference on Machine Learning, 2256–2265. https://proceedings.mlr.press/v37/sohl-dickstein15.html

Song, Y., & Ermon, S. (2019). Generative Modeling by Estimating Gradients of the Data Distribution. Advances in Neural Information Processing Systems, 32. https://proceedings.neurips.cc/paper/2019/hash/3001ef257407d5a371a96dcd947c7d93-Abstract.html

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021, January 12). Score-Based Generative Modeling through Stochastic Differential Equations. International Conference on Learning Representations. https://openreview.net/forum?id=PxTIG12RRHS

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. Advances in Neural Information Processing Systems, 33, 6840–6851. https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html

Song, J., Meng, C.,  Ermon, S. (2020, October 2). Denoising Diffusion Implicit Models. International Conference on Learning Representations. https://openreview.net/forum?id=St1giarCHLP

He, Y., Murata, N., Lai, C.-H., Takida, Y., Uesaka, T., Kim, D., Liao, W.-H., Mitsufuji, Y., Kolter, J. Z., Salakhutdinov, R., Ermon, S. (2023, October 13). Manifold Preserving Guided Diffusion. The Twelfth International Conference on Learning Representations. https://openreview.net/forum?id=o3BxOLoxm1

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042

# References II

Vincent, P. (2011). A connection between score matching and denoising autoencoders. Neural computation, 23(7), 1661-1674. https://ieeexplore.ieee.org/abstract/document/6795935/

Luo, C. (2022). Understanding diffusion models: A unified perspective. arXiv preprint arXiv:2208.11970. https://arxiv.org/abs/2208.11970

Efron, B. (2011). Tweedie's formula and selection bias. Journal of the American Statistical Association, 106(496), 1602-1614. https://doi.org/10.1198/jasa.2011.tm11181

Dhariwal, P., & Nichol, A. (2021). Diffusion Models Beat GANs on Image Synthesis. Advances in Neural Information Processing Systems, 34, 8780–8794. https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html

Qu, Y., Shen, X., He, X., Backes, M., Zannettou, S., & Zhang, Y. (2023). Unsafe Diffusion: On the Generation of Unsafe Images and Hateful Memes From Text-To-Image Models (arXiv:2305.13873). arXiv. https://doi.org/10.48550/arXiv.2305.13873

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 1(2), 3. https://arxiv.org/abs/2204.06125

Rando, J., Paleka, D., Lindner, D., Heim, L., & Tramer, F. (2022, November 18). Red-Teaming the Stable Diffusion Safety Filter. NeurIPS ML Safety Workshop. https://openreview.net/forum?id=zhDO3F35Uc

# References III

Schramowski, P., Brack, M., Deiseroth, B., & Kersting, K. (2023). Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models. 22522–22531. https://openaccess.thecvf.com/content/CVPR2023/html/Schramowski_Safe_Latent_Diffusion_Mitigating_Inappropriate_Degeneration_in_Diffusion_Models_CVPR_2023_paper.html

Kumari, N., Zhang, B., Wang, S. Y., Shechtman, E., Zhang, R., Zhu, J. Y. (2023). Ablating concepts in text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 22691-22702). http://openaccess.thecvf.com/content/ICCV2023/html/Kumari_Ablating_Concepts_in_Text-to-Image_Diffusion_Models_ICCV_2023_paper.html

Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., & Bau, D. (2023). Erasing Concepts from Diffusion Models. 2426–2436. https://openaccess.thecvf.com/content/ICCV2023/html/Gandikota_Erasing_Concepts_from_Diffusion_Models_ICCV_2023_paper.html

Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., Bau, D. (2024). Unified concept editing in diffusion models. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 5111-5120). https://openaccess.thecvf.com/content/WACV2024/html/Gandikota_Unified_Concept_Editing_in_Diffusion_Models_WACV_2024_paper.html

Yoon, T., Myoung, K., Lee, K., Cho, J., No, A., & Ryu, E. K. (2023, November 2). Censored Sampling of Diffusion Models Using 3 Minutes of Human Feedback. Thirty-seventh Conference on Neural Information Processing Systems. https://openreview.net/forum?id=4qG2RKuZaA

# References IV

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. Proceedings of the 38th International Conference on Machine Learning, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., & Cohen-or, D. (2022, September 29). Prompt-to-Prompt Image Editing with Cross-Attention Control. The Eleventh International Conference on Learning Representations. https://openreview.net/forum?id=_CDixzkzeyb

Schramowski, P., Tauchmann, C., & Kersting, K. (2022, June). Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content?. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1350-1361). https://dl.acm.org/doi/abs/10.1145/3531146.3533192

MDS Master of Data Science Universidad de Chile

fcfm Ingeniería Matemática
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE

# Safe Sampling for Score-based Models

## Classifier (un)guidance with conditional diffusion trajectory correction

Examination to qualify for the degree of Master in Data Science and to the title of Mathematical Engineer

**Camilo Carvajal Reyes** **19 June 2024**

**Supervised by Felipe Tobar and Joaquín Fontbona**