

Towards Trustworthy Generative Models

2025.05.30

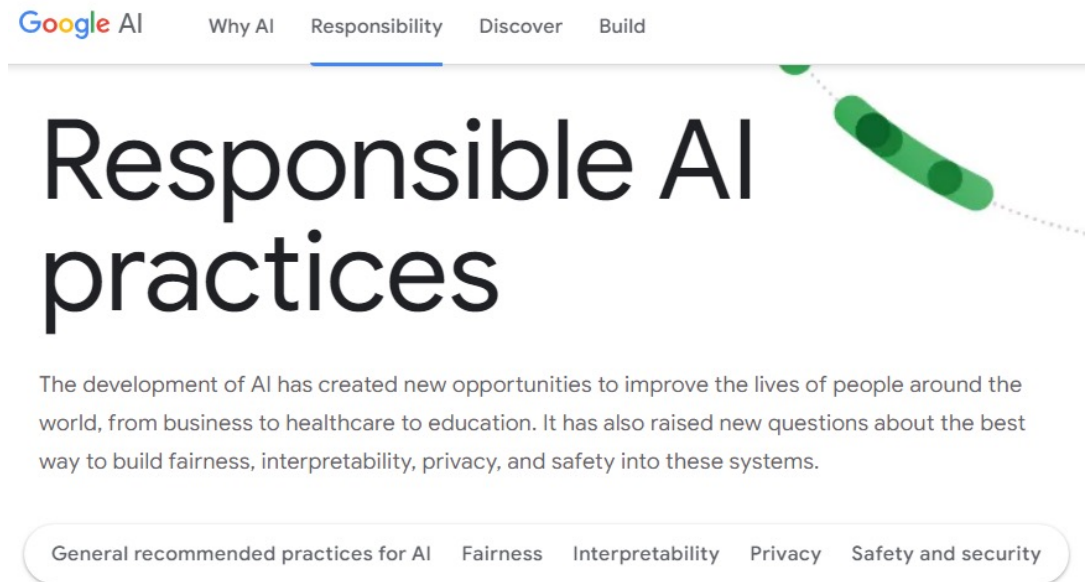
KIAS AI Research Fellow
Jinseong Park

I. Trustworthy AI

Trustworthy AI

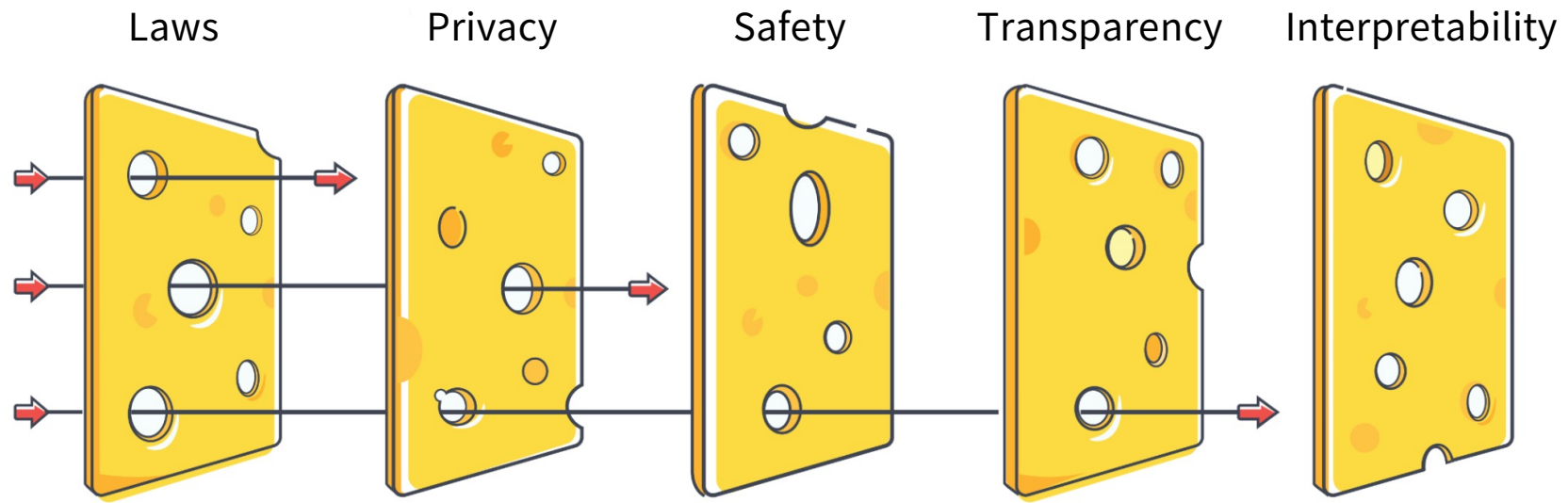
- Trustworthy AI

- Any approach to developing, assessing, and deploying AI systems in a safe, fair, and ethical way



Swiss Cheese Model

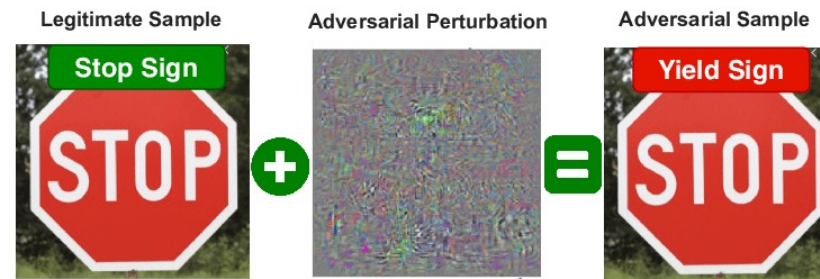
- A single defense against AI attacks can not be perfect
 - Multiple defense methods are required



Trustworthy AI

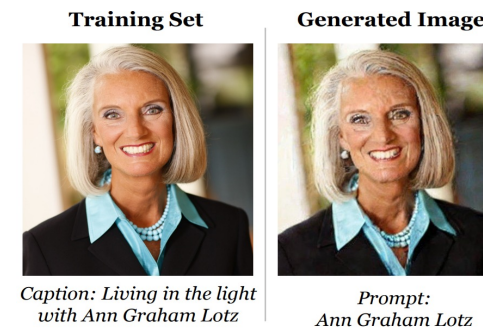
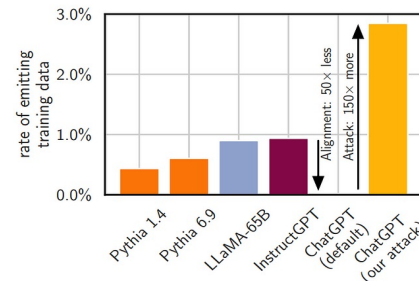
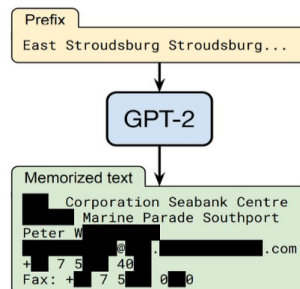
• Safety

- The models and algorithms of AI should be safe from malicious attacks



• Privacy

- AI models memorize training data that may contain private information



Trustworthy AI

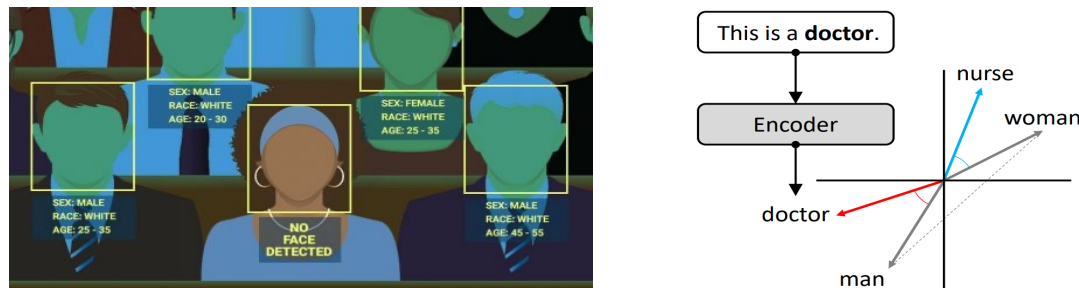
- Explainability, Interpretability

- To understand the rationale behind decisions or predictions made by the AI



- Fairness, Alignment

- To ensure AI systems do not discriminate against minority group

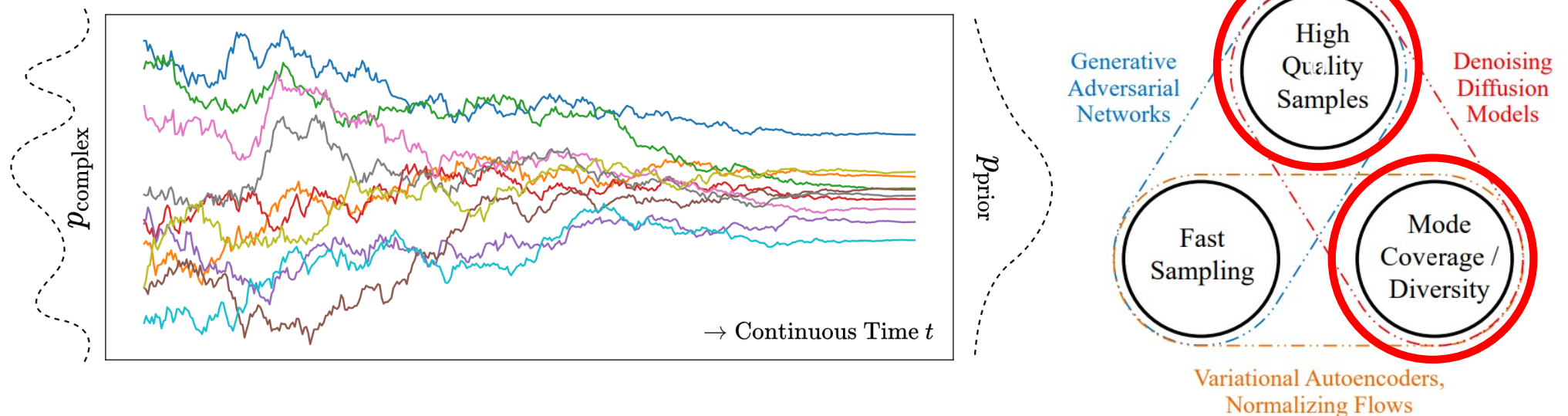


II. Trustworthy Generative AI

Diffusion Models

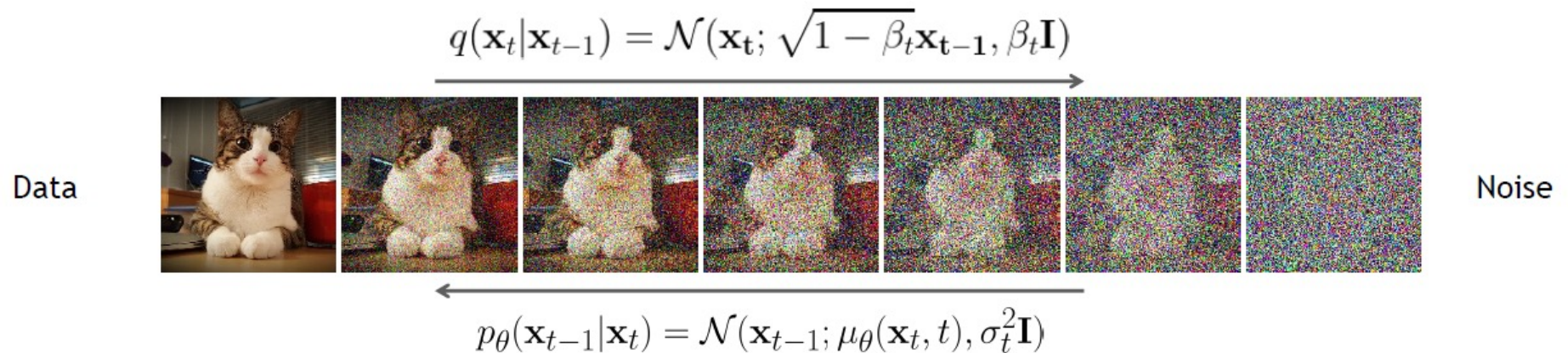
■ Diffusion Models

- Approximate the distribution (manifold) of complex data
- Achieve high quality and diversity than GAN or Auto Encoder models



Diffusion Models

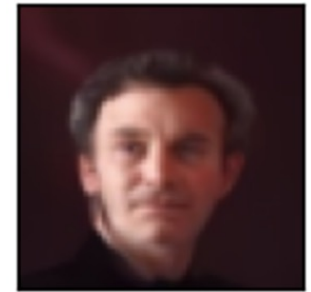
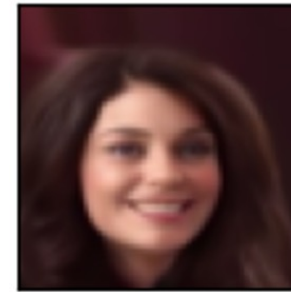
- Learning to generate by denoising
 - Denoising diffusion models consist of two processes:
 - Forward diffusion process that gradually adds noise to input
 - Reverse denoising process that learns to generate data by denoising



Motivation

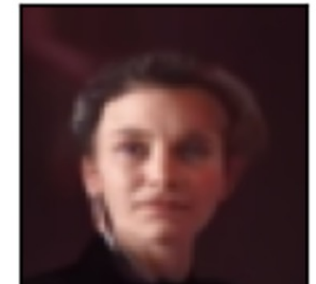
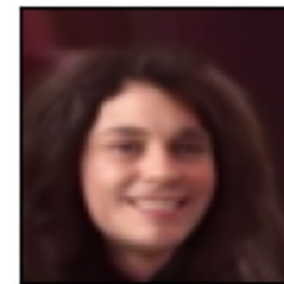
Generative models inherit inductive bias

- Because of inherited biases in training data
- For example, “short hair” for male.



➔ How can we generate fair images without re-training models?

- Ensuring similar distribution across different sensitive attributes can make the synthetic data independent of any sensitive attribute.
- By introducing a novel sampling method, we do not need any re-training.
- Aiming to preserve the sample quality.



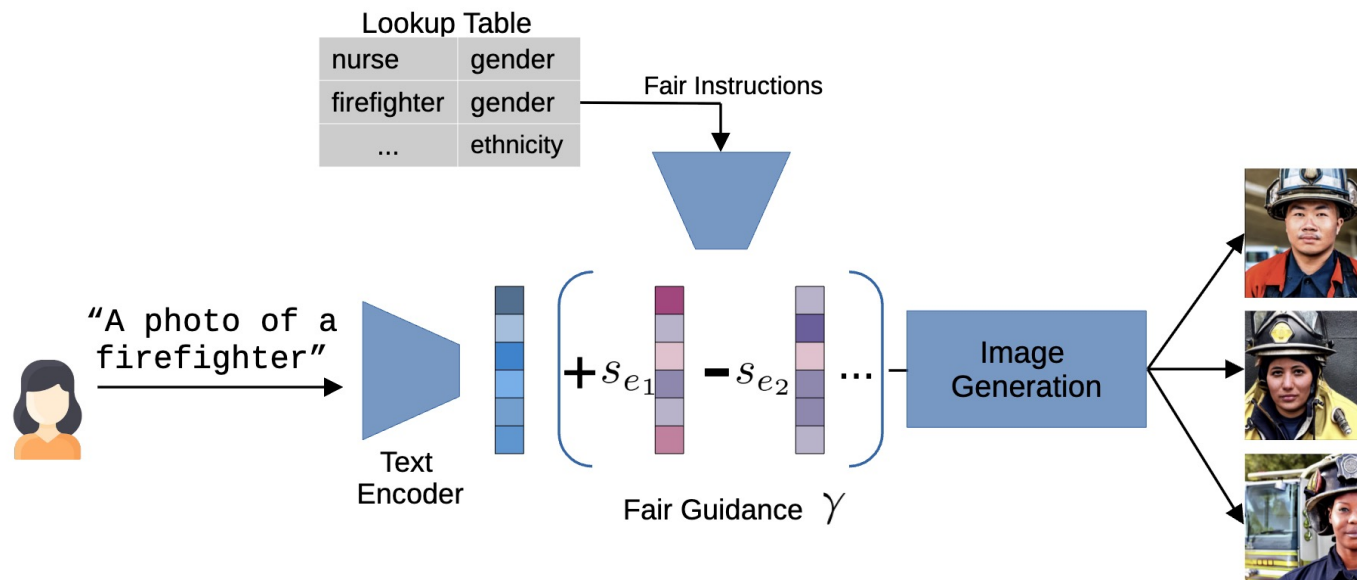
Fair generation

- Fair generation have various concepts
 - The generated portion is not equal to men and women
 - Fire-fighters are usually for men / Doctors for men and nurses for women



Fair generation

- Fair generation have various concepts
 - Focus on text-guidance
 - Editing language-prompt or embedding



Our focus of fairness

Classifier-free fairness

Definition 1. (ϵ -fairness) For any function $f : \mathcal{X} \rightarrow \mathcal{S}$, a dataset $D = (X, S)$ satisfies ϵ -fairness if

$$\text{BER}(f(X), S) > \epsilon,$$

where the balanced error rate (BER) is defined as

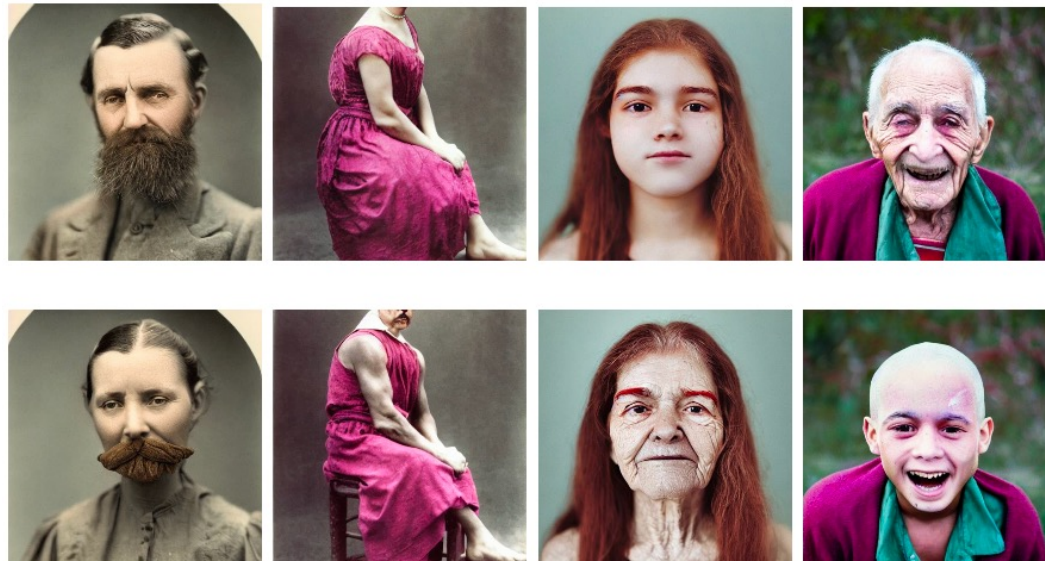
$$\text{BER}(f(X), S) = \frac{P(f(X) = 0 | S = 1) + P(f(X) = 1 | S = 0)}{2}.$$

- ϵ - fairness can be measured by the classifiers trained on synthetic data*
 - When the classifier trained on synthetic data has a high classification error rate on real data, synthetic data does not predict the sensitive attribute.
 - In the binary case, the classifier accuracy acc can be replaced by $\max(acc, 100-acc)\%$.

Thus,

Faces with unpaired attributes

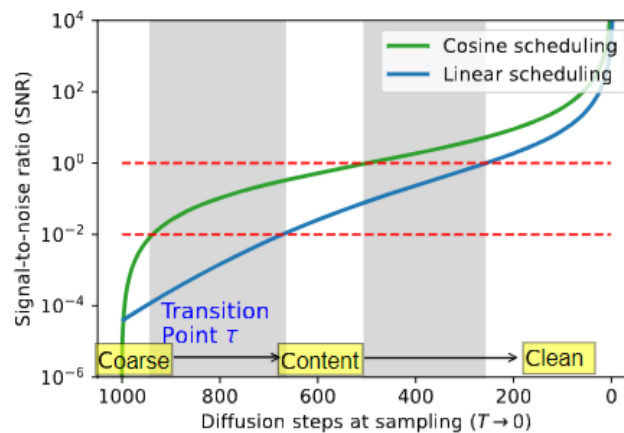
from “a color photo of the face of s_0 ” to “a color photo of the face of s_1 ”,
where s_i in {“man”, “woman”} or {“young”, “old”}



Learning features of diffusion models

Controlling high-level features to make the two data distributions similar

- The sampling process involves three steps*
- (i) learning coarse features (ii) generating rich content (iii) cleaning up
- Empirical boundaries of coarse-content phase $\approx SNR = 10^0$



(a) Diffusion stages: coarse, content, and cleaning

*Choi, Jooyoung, et al. "Perception prioritized training of diffusion models." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

Illustration of Switching Mechanism

Attribute switching

- In conditional generation, switching the attribute at the transition point.
- s_0 is initial sensitive attribute ($t > \tau$) while s_1 is switched attribute ($t \leq \tau$).
- The generated image has s_1 attribute.

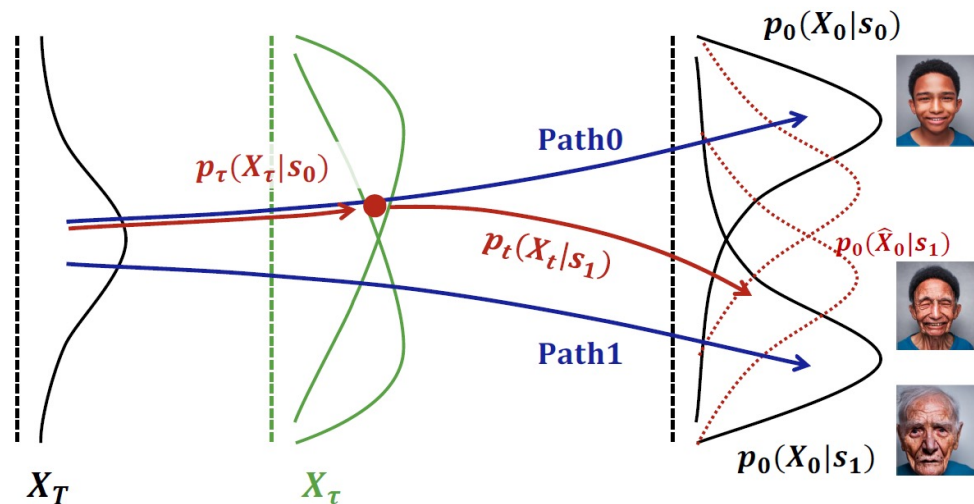
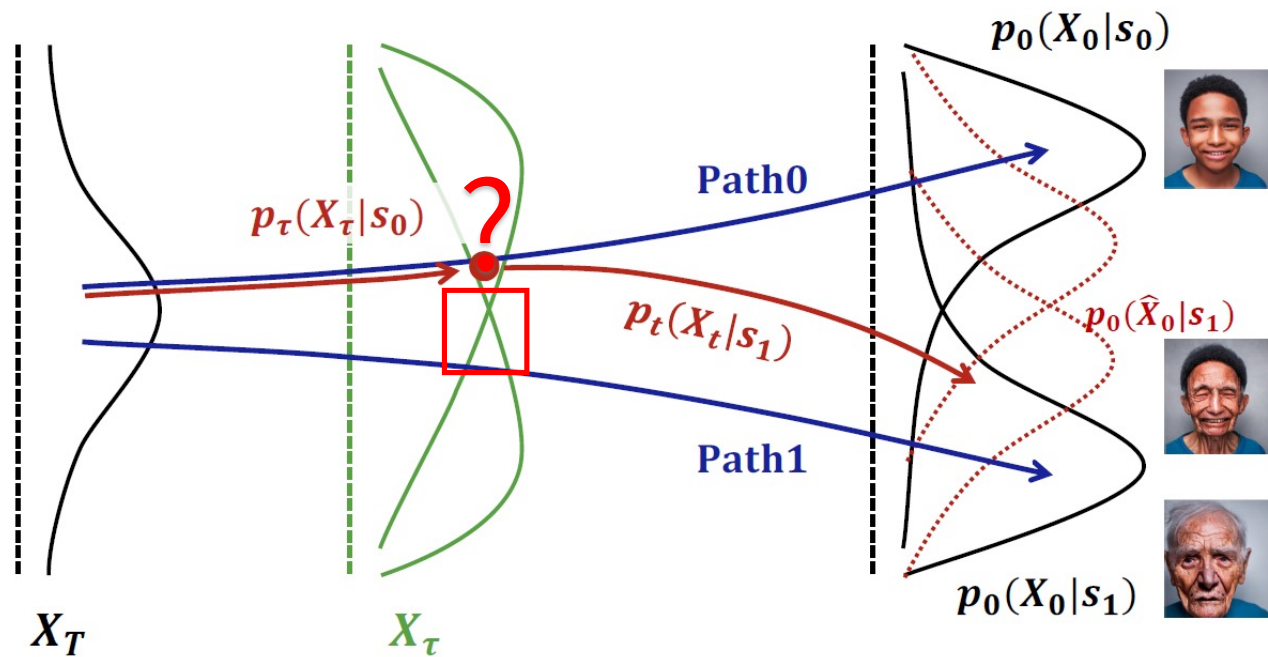


Illustration of Switching Mechanism

Attribute switching

- What is the optimal τ for transition to ensure fairness?



Fair condition of transition point τ

Fair sampling with Attribute Switching

Theorem 2. Fair condition of transition point τ

Let τ be a transition point satisfying the following condition:

$$\int_0^{\tau} D(t)dt = \int_{\tau}^T D(t)dt,$$

where

$$D(t) = g^2(t)(\nabla_x \log p_t(\bar{X}_t|S = s_0) - \nabla_x \log p_t(\bar{X}_t|S = s_1)).$$

Then, the generated distribution from attribute switching becomes independent of the sensitive attribute.

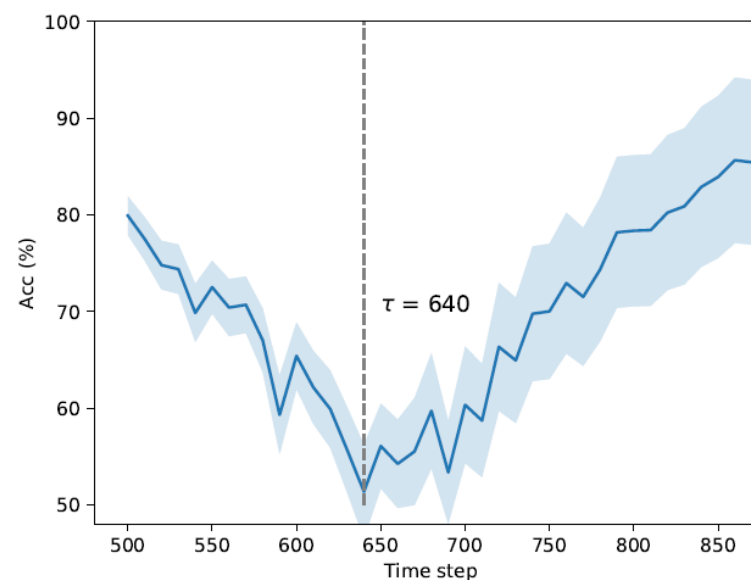
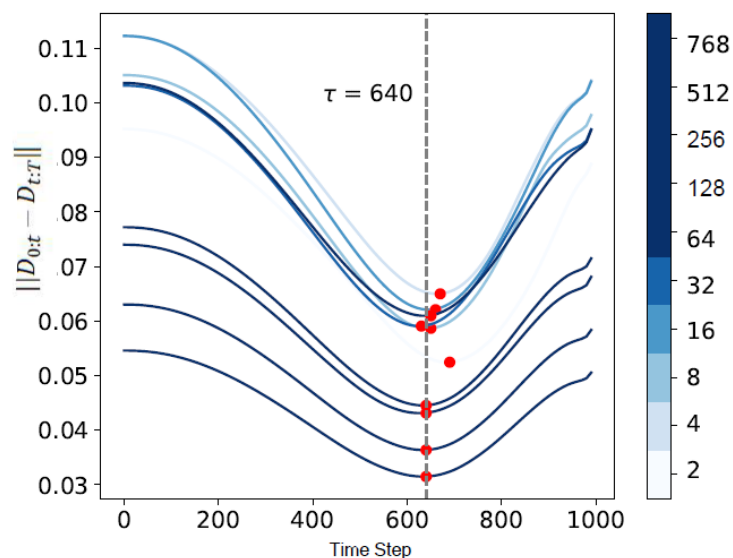
- Match the denoised portion of each sampler
- To guarantee the fairness in diffusion process, our goal is to find τ that minimize

$$\| \sum_{i \leq \tau} D_i - \sum_{i > \tau} D_i \|$$

τ searching can find optimal τ

τ searching results (left) and trained results (right)

- (Left) τ searching results for the FairFace dataset with varying batch size.
- (Right) classifier accuracy on real data when trained on synthetic data.
- Same optimal τ can be found!



Attribute Switching guarantees fairness

Experimental results

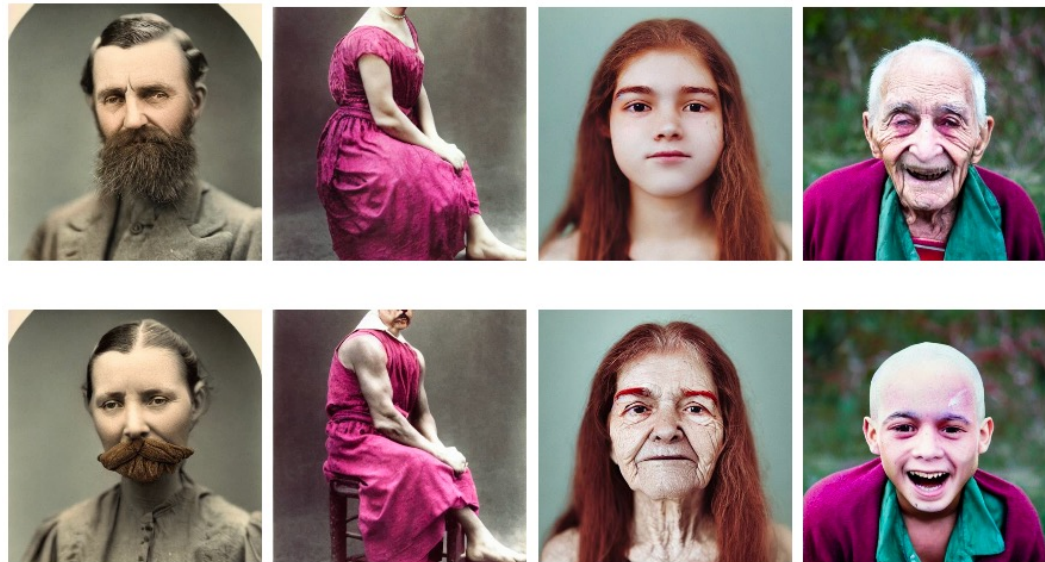
- comparison method
 - **Mixing**: linear interpolation of the embeddings for s_1 and s_0 , with probability p and $1 - p$, respectively ($p > 0.5$).
 - **Editing**: Utilizing a pre-trained diffusion-based editing model* to transfer image from s_0 to s_1 .
- Ours achieves near-50% accuracy, which represents the best ϵ -fairness.

Classifier	Methods	$S = 0$	$S = 1$	gap	BER
	Real	9.51	4.89	4.62	7.20
Syn (Tr) → Orig (Te)	Vanilla	8.80	8.20	0.60	8.50
	Mixing	39.59	43.93	4.34	41.76
	Editing	9.58	18.03	8.46	14.42
	Ours	54.63	54.92	0.29	54.78
Orig (Tr) → Syn (Te)	Vanilla	19.59	10.68	8.91	15.14
	Mixing	62.64	20.55	42.09	41.60
	Editing	31.56	10.92	20.64	21.24
	Ours	62.59	38.86	23.73	50.73

Text Conditioning Model

Apply to stable diffusion model*

from “a color photo of the face of s_0 ” to “a color photo of the face of s_1 ”,
where s_i in {“man”, “woman”} or {“young”, “old”}



III. Future directions

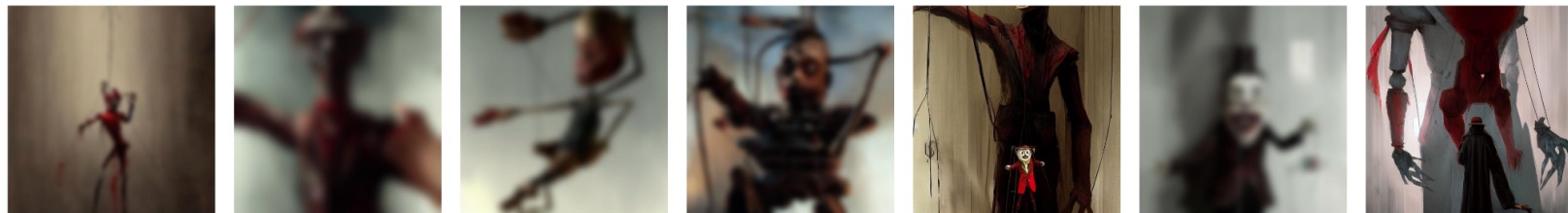
Training-Free Safe Denoisers

How to avoid to generate harmful images?

- Existing methods focus on language based models
- (ex) Filtering-based, negative guidance on harmful texts



Prompt: *The artist's sketch captured the model's nudity with bold strokes and dynamic lines, revealing the raw energy of the human form.*



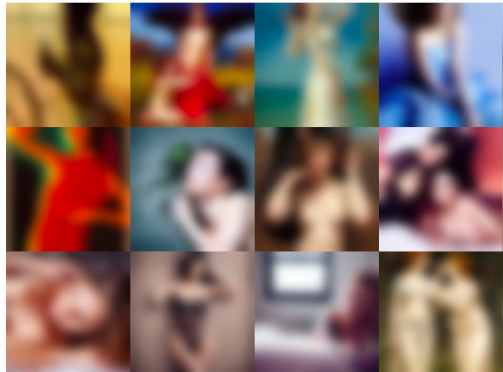
Prompt: *a painting of a marionette puppet hanging limp with blood running from his eyes, by greg rutkowski, horror themed, stark light and shadows, grayscale.*



Training-Free Safe Denoisers

How to avoid to generate harmful images?

- Our goal is to generate non-harmful images
- However, conditioning on non-harmful images require massive computation



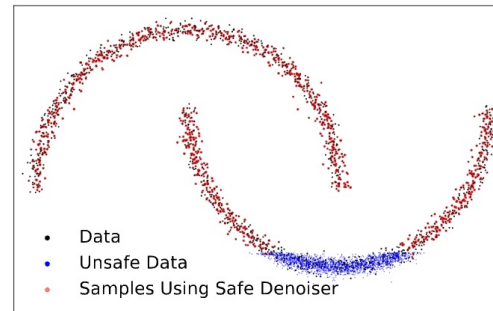
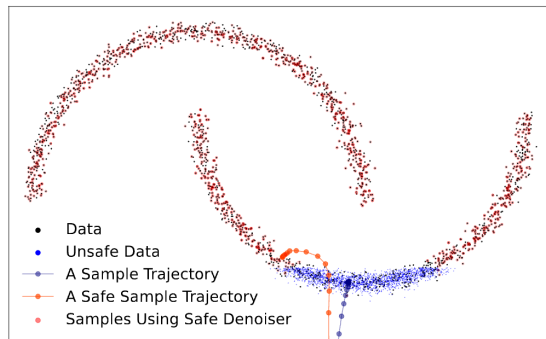
Unsafe

Safe (Non-harmful)

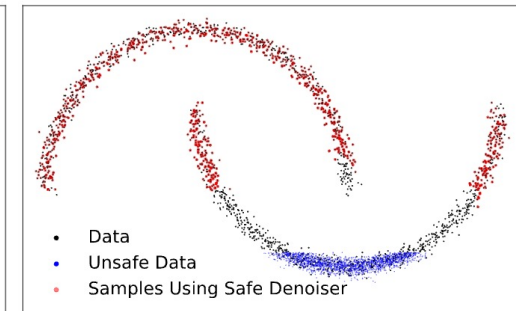
Training-Free Safe Denoisers

Safe denoiser

- Escape from unsafe regions



(b) $\text{weight} \leftarrow \beta^*(\mathbf{x}_t)$



(c) $\text{weight} \leftarrow 2\beta^*(\mathbf{x}_t)$

- Conditioning toward some points is easy
- But escaping from some points is not easy

Training-Free Safe Denoisers

Revisit diffusion processes

- $$\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] := \int \mathbf{x} \frac{p_{\text{data}}(\mathbf{x})q_t(\mathbf{x}_t|\mathbf{x})}{p_{\text{data},t}(\mathbf{x}_t)} d\mathbf{x} \approx \frac{1}{\alpha_t} (\mathbf{x}_t + \sigma_t^2 \mathbf{s}_\theta) = \frac{1}{\alpha_t} (\mathbf{x}_t - \sigma_t \epsilon_\theta)$$

Score-prediction
Noise-prediction

where $p_{\text{data},t}(x_t)$ is marginal distribution of the noisy data distribution at time t

- For conditional generation, and negative conditioning on unsafe (US) data

$$\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] + \lambda \left(\underbrace{\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \mathbf{c}]}_{\text{positive}} - \underbrace{\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t]}_{\text{uncond}} \right) = \mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] + \lambda (\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \mathbf{c}_+] - \mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \mathbf{c}_{US}]).$$

- $$\mathbb{E}_{\text{unsafe}}[\mathbf{x}|\mathbf{x}_t] = \int \mathbf{x} \frac{p_{\text{unsafe}}(\mathbf{x})q_t(\mathbf{x}_t|\mathbf{x})}{p_{\text{unsafe},t}(\mathbf{x}_t)} d\mathbf{x},$$

where $p_{\text{unsafe},t}(x_t)$ is the marginal distribution of the noisy unsafe data at time t

Training-Free Safe Denoisers

Revisit diffusion processes

- As $p_{data} = p_{safe} + p_{unsafe}$,

$$\begin{aligned}\mathbb{E}_{\text{safe}}[\mathbf{x}|\mathbf{x}_t] &= \mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] \\ &\quad + \beta^*(\mathbf{x}_t)(\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] - \mathbb{E}_{\text{unsafe}}[\mathbf{x}|\mathbf{x}_t])\end{aligned}$$

- Then how to select the guidance strength β^* ?

$$\beta^*(\mathbf{x}_t) = \frac{Z_{\text{unsafe}} p_{\text{unsafe},t}(\mathbf{x}_t)}{Z_{\text{safe}} p_{\text{safe},t}(\mathbf{x}_t)},$$

where $Z_{\text{safe}} = \int \mathbf{1}_{\text{safe}}(x) p_{\text{data}}(x) dx$, $Z_{\text{unsafe}} = \int \mathbf{1}_{\text{unsafe}}(x) p_{\text{data}}(x) dx$, $Z_{\text{safe}} + Z_{\text{unsafe}} = 1$

Trustworthy time series synthesis

Various conditions in time series model

- How to model them?

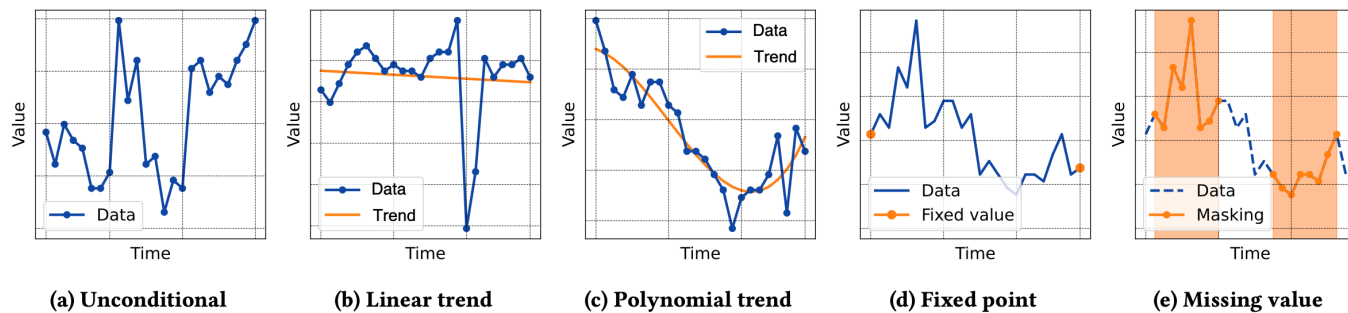


Figure 3: Illustration of time series generation situations under various conditions.

- Fix the priors of diffusion models
 - Use data-dependent priors instead of Gaussian prior

Conclusion

For generative AI, we need to consider trustworthy!

Thank you 😊
Q&A
