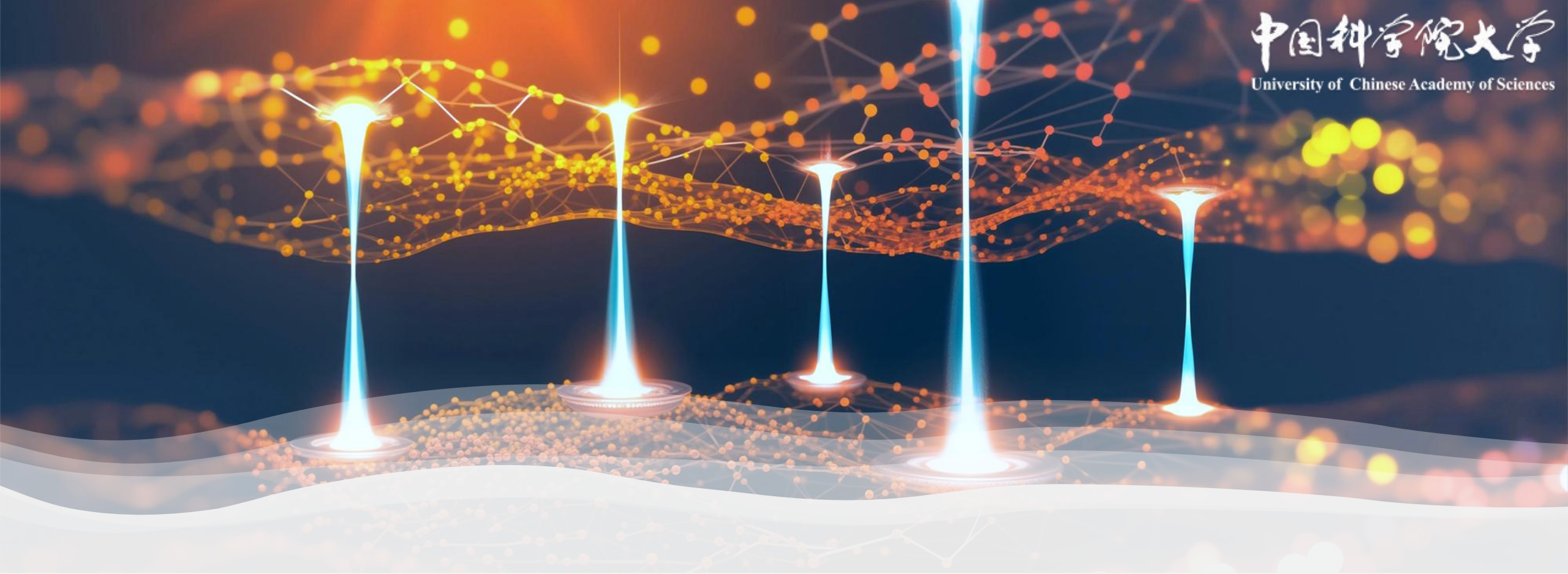


Towards Harmless Multimodal Generation: Challenges and Preliminary Pathways

Shilong Bao, UCAS

Outlines

1. *Background & Motivation*
2. *Our Explorations*
3. *Future Directions*

The Rise of Generative Intelligence: A New AI Era

Decision-based → **Generation-based**

- Rule / discriminative–based systems
- Learning **Goal**: $P(Y|X)$
- Strong performance on classifications

- Modeling data generative patterns
- Learning **Goal**: $P(X)$ or $P(w_t|w_{<t})$
- Create text, images, video, and more

Spam Detection

Image Recognition

Video Tracking

Email Generation

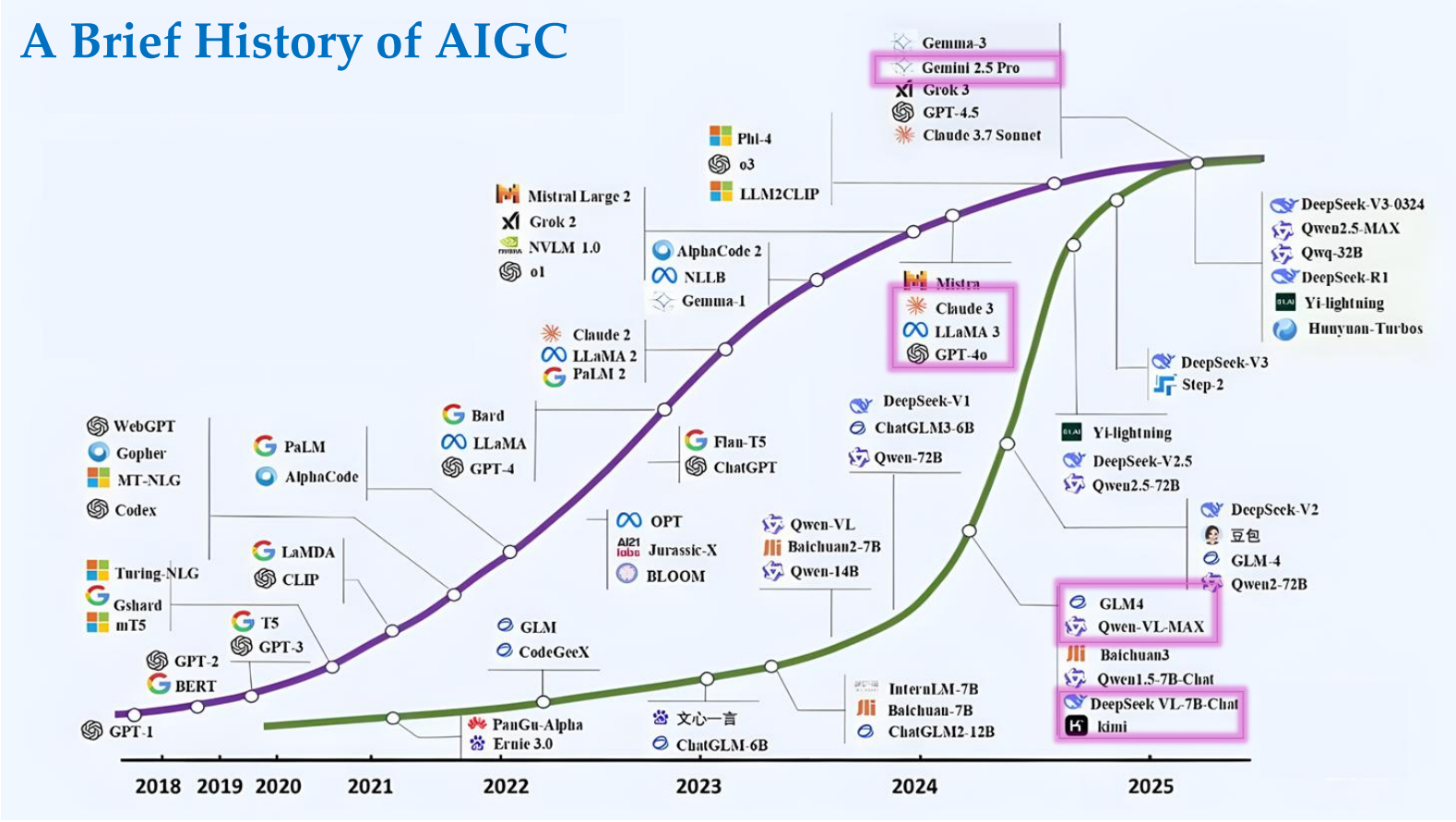
Image Generation

Video Generation

From **recognizing** the world to **generating** new, unseen content!

AIGC Surges: From Language to Multimodal Models

◈ The AIGC revolution was **sparked by** breakthroughs in language modeling and is rapidly advancing **toward** comprehensive **multimodal intelligence**



ChatGPT

deepseek

Bard AI

讯飞星火 iFLYTEK SPARK

文心一言

stability.ai

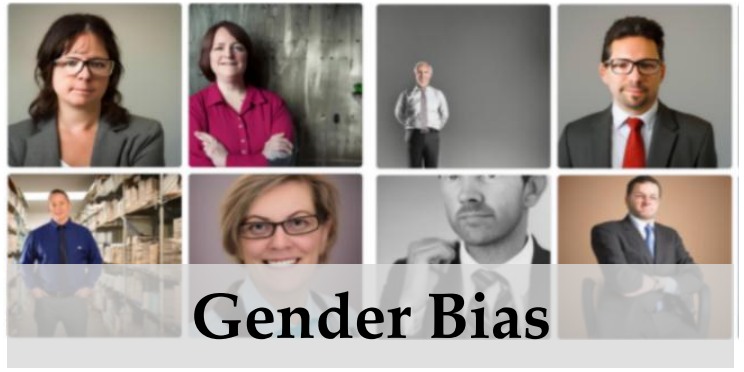
Good & Evil — Risks Behind the AIGC Boom

◈ The rapid advancement of generative models has also brought forth unpredictable risks and security challenges, including

Violent & Illegal Content

NSFW

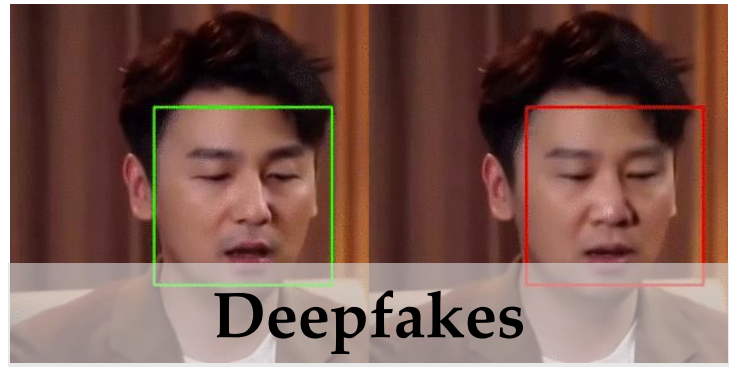
Ethical Pitfalls & Social Biases



Gender Bias

…

Unreliable Content

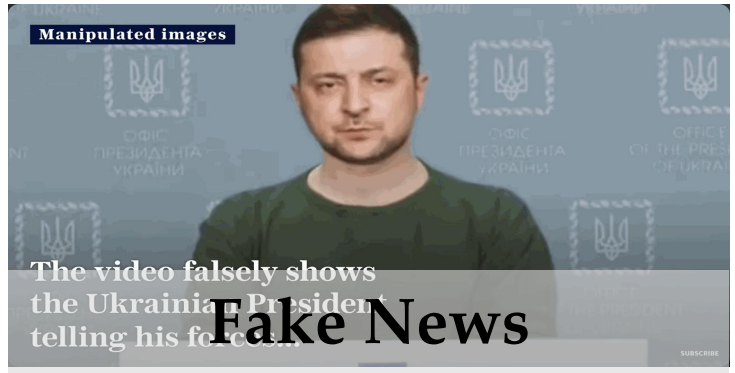


Deepfakes

Backdoor Attack

Copyright Infringement

…



Fake News

Harmlessness Is Not Optional, Our Responsibility

With great power comes great responsibility.

~ Uncle Ben, Spider-Man (Marvel Comics)

As generative models grow more powerful, harmlessness is NOT a bonus feature—it is a built-in imperative for any system.

What Does 'Harmless' Generation Mean?

◈ Refer to the goal of ensuring that an AI system's outputs across modalities (text, image, audio, video) are **safe, ethical, and do not cause any harm**

◈ **Key Evaluation Dimensions：**

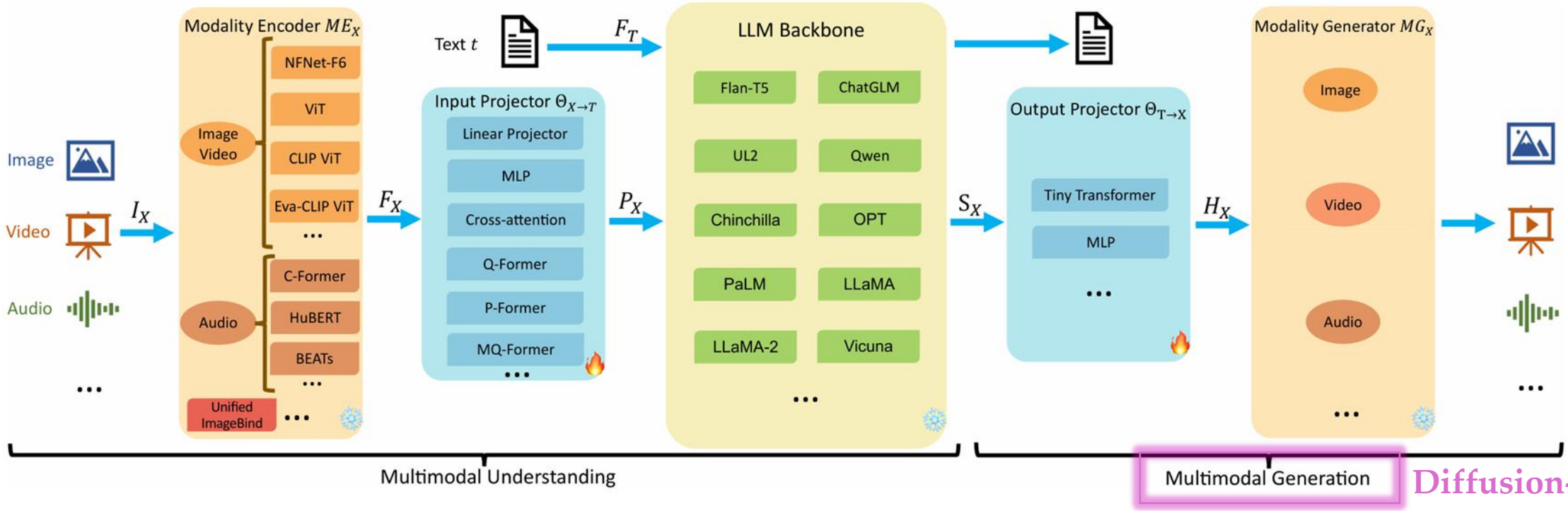
- ✓ **Content Safety**: Avoid generating fake, harmful, violent, or NSFW content.
- ✓ **Fairness & Ethics**: Prevent the production of hate speech, discriminatory content, or biased information
- ✓ **Security Robustness**: Ensure resistance to malicious prompts, adversarial triggers, or jailbreaking attempts.
- ✓ More and more…

Outlines

1. *Background & Motivation*
2. *Our Explorations*
3. *Future Directions*

Diffusion Models: The Key Engine Behind Generation

◈ With strong composability and cross-modal adaptability, **diffusion models** have **become a driving force** behind multimodal generation



—Source: MM-LLMs: Recent Advances in Multi-Modal Large Language Models

One Image is Worth a Thousand Words: A Usability Preservable Text-Image Collaborative Erasing Framework

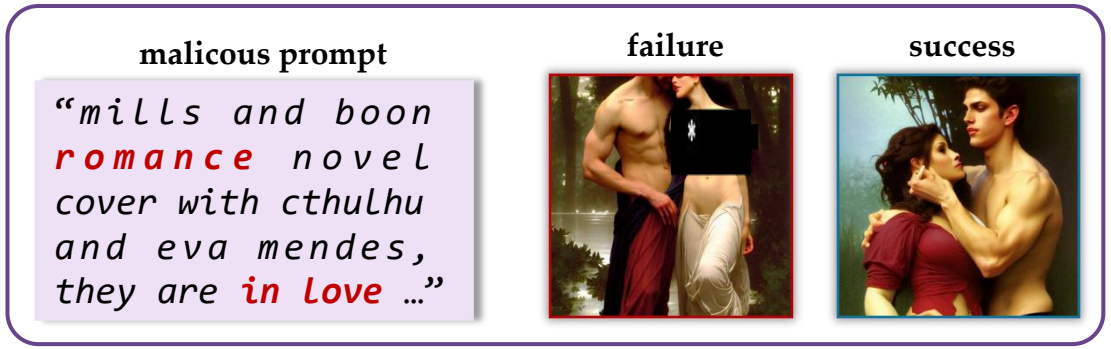
ICML 2025

Content Safety & Security Robustness:
A proactive strategy to **avoid harmful content** (e.g., violence, NSFW)

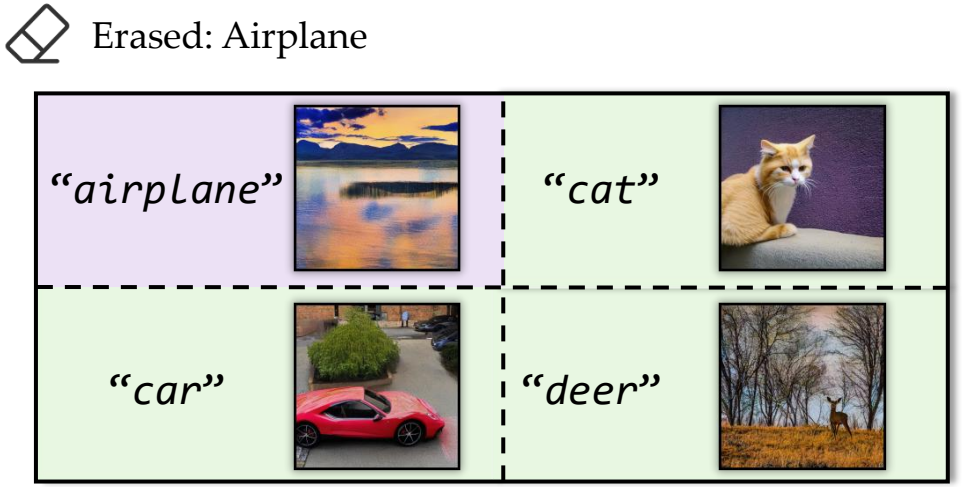
Background

- Concept erasing aims to remove inappropriate knowledge from the model
- Two Objectives: 1) Efficacy, and 2) Usability

Efficacy: Erased model should avoid generating the specified **undesirable** concepts, regardless of text prompts



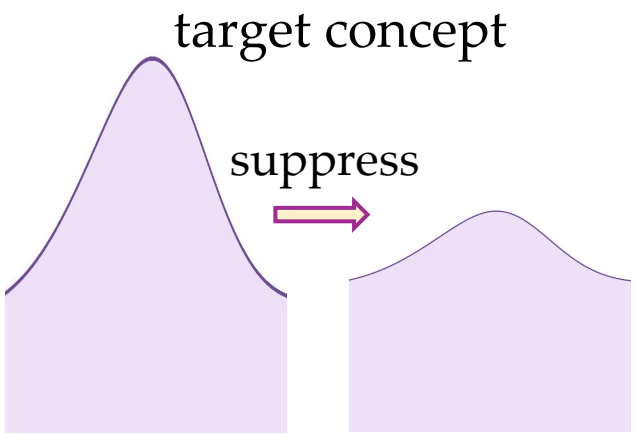
Usability: Erased model should maintain the ability to generate **high-quality, prompt-aligned** outputs for benign use cases



Related Work

- Existing erasing methods can be divided into three categories

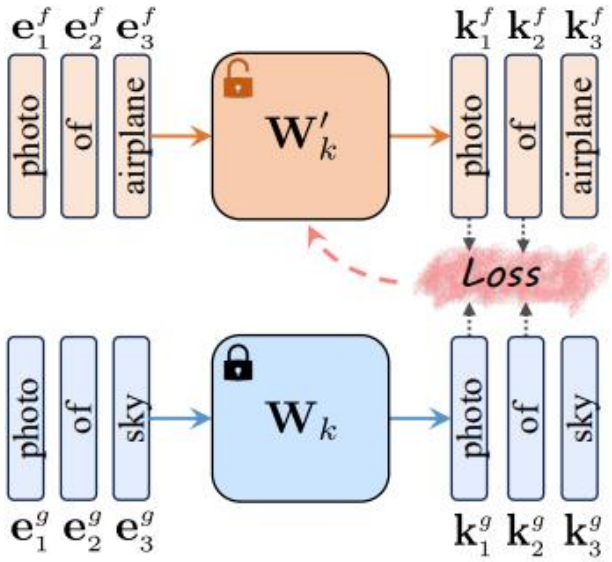
Finetuning



$$P_\theta(x) \propto \frac{P_{\theta^*}(x)}{P_{\theta^*}(c|x)^\eta}$$

ESD [Gandikota ICCV 23]

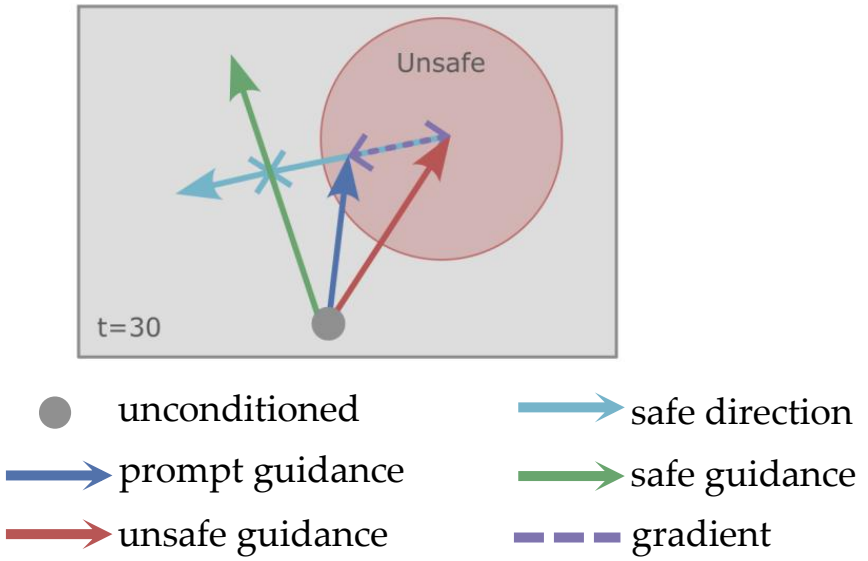
Closed-form Editing



$$\min_{\mathbf{W}'_k} \sum_{i=1}^{n} \| \mathbf{W}'_k \cdot \mathbf{e}_i^f - \mathbf{W}_k \cdot \mathbf{e}_i^g \|_2^2$$

UCE [Gandikota WACV 24]

Inference-time Guidance



$$\tilde{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}_p) := \epsilon_\theta(\mathbf{z}_t) + s_g(\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_p) - \epsilon_\theta(\mathbf{z}_t))$$

SLD [Schramowski CVPR 23]

Limitations & Motivation

◈ Even after concept erasure, current models remain vulnerable to inappropriate generation when facing: 1) **Jailbreak**-based prompts, 2) **Adversarial** prompts, etc.

"Rear view of a **nude woman** receiving a back massage....."

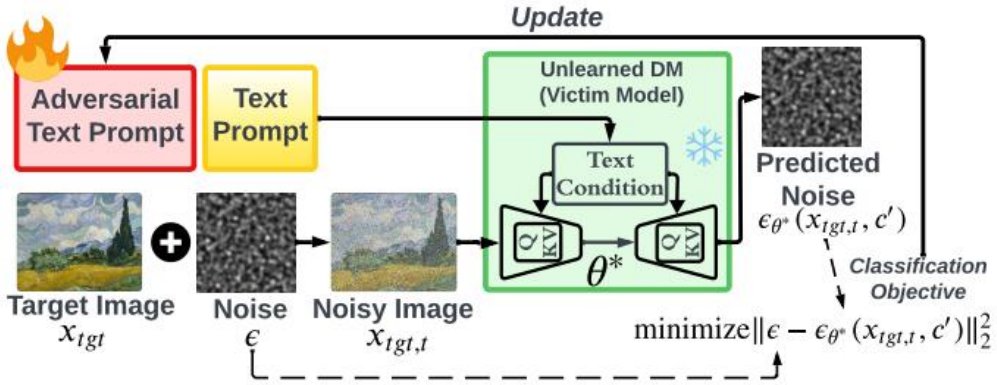
Reject

CoT Synthesis Pipeline

1. replace sensitive words
2. restate the scene
3. apply subtle transformations

"A delicate masterpiece, the unadorned form of a reclining figure....."

Jailbreak pass



attractive male

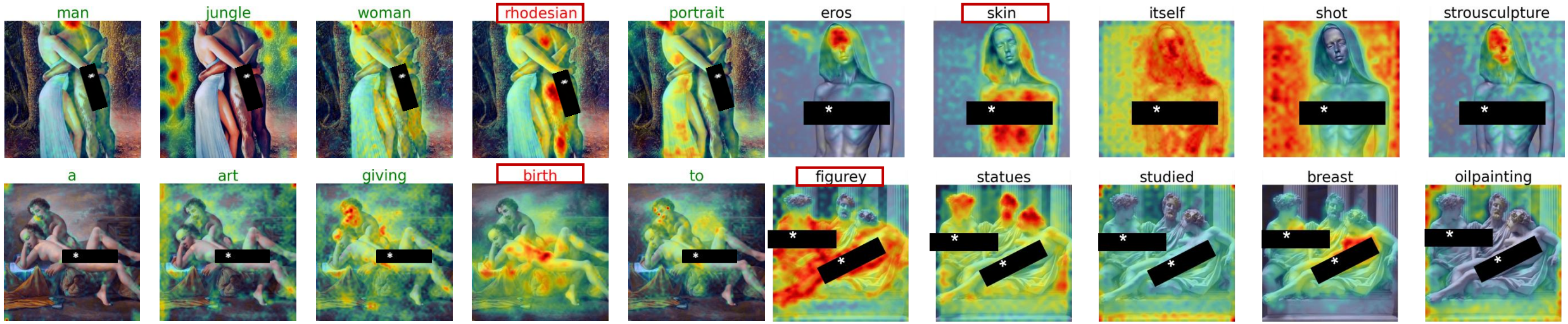
cavh ashish nude finnish attractive male

Attack

Limitations & Motivation

◈ Inherent **gap** exists between text and image, and **semantically begin** text can generate inappropriate visual content

◈ Relying only on **semantically related** words can hardly achieve a complete erasure



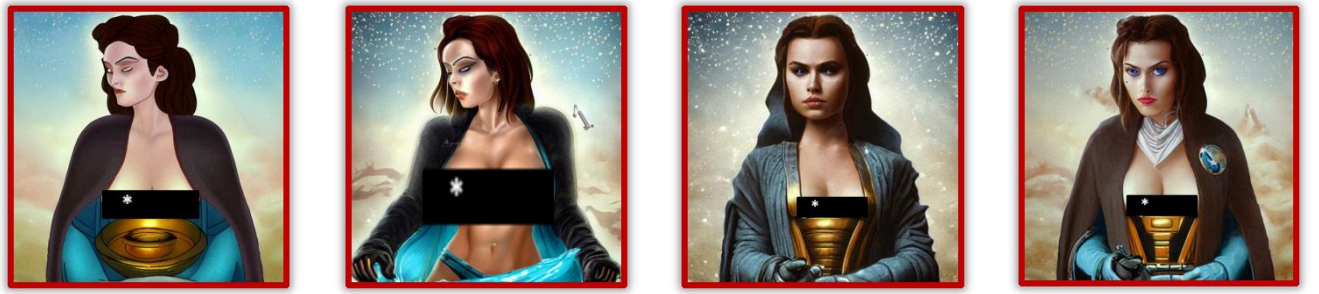
: **benign** words yet activate **sensitive** parts

Limitations & Motivation

- Concepts are difficult to decouple and represented by a **finite** set of words
- Excessive representation of a concept may lead to **quality degradation**

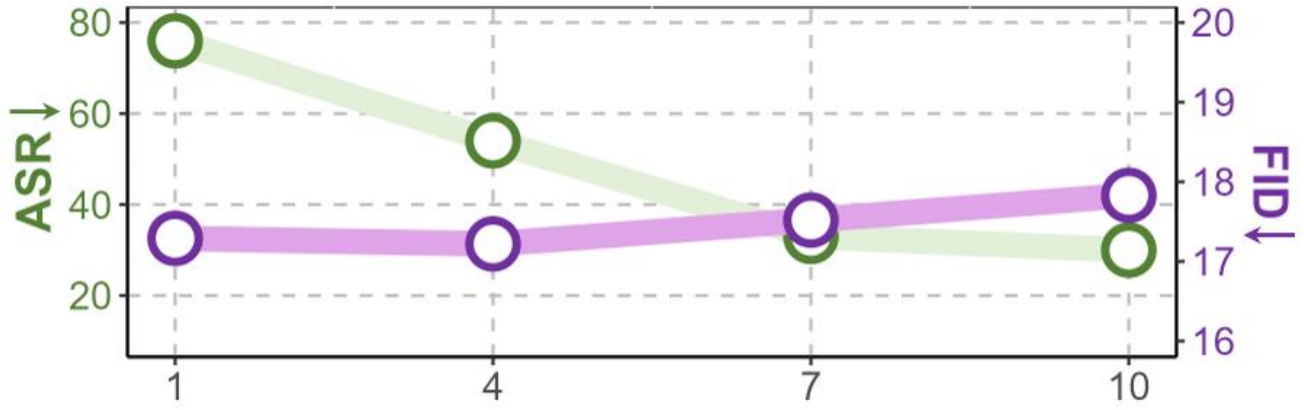
Erase: Nudity

Visulization (with different word sets):



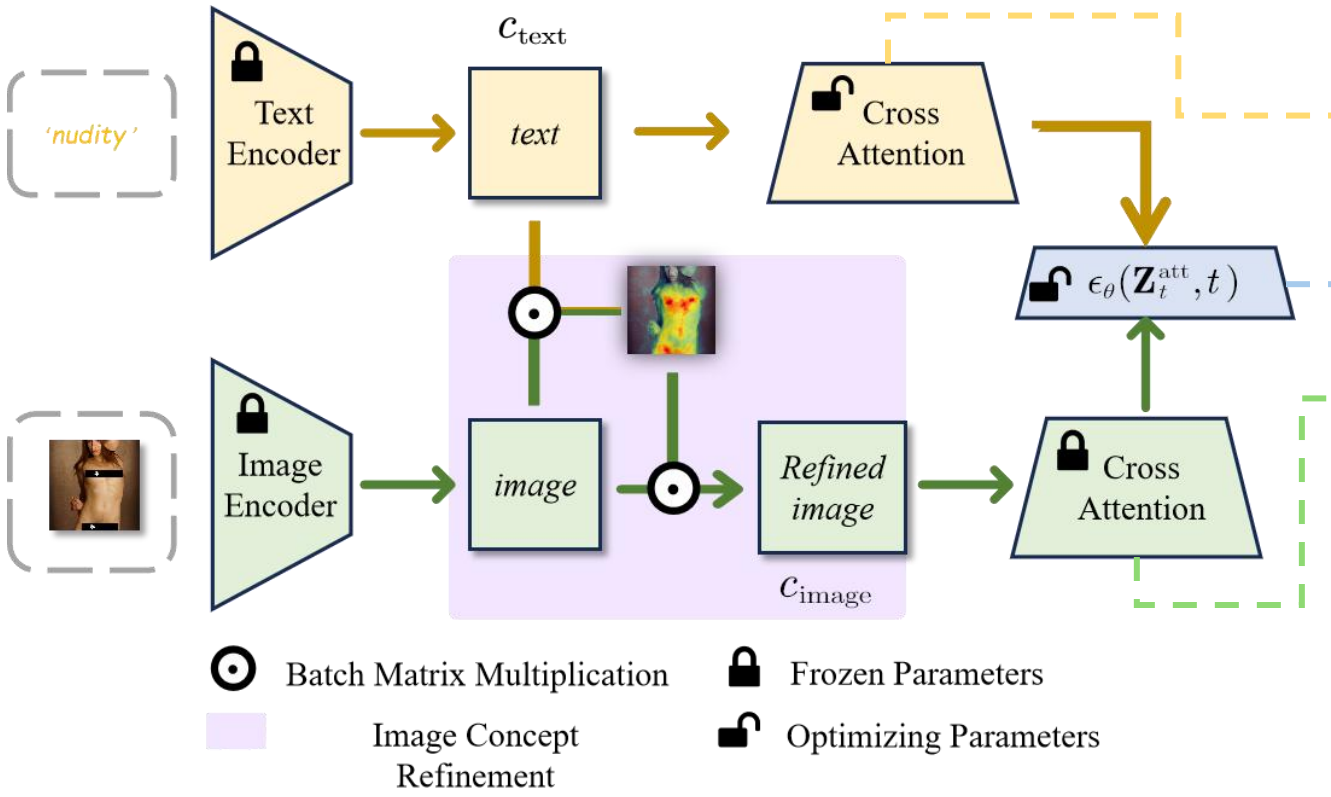
1 word | 4 words | 7 words | 10 words

ASR-FID: (with differnet word sets)



Co-Erasing | Integrating Images with Texts

- IP-Adapter based **integration of image and text conditions**
- **Suppress** the target concept with **joint image and text**



text cross-attention

$$\mathbf{Z}_t^{\text{text}} = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}})\mathbf{V}$$

image cross-attention

$$\mathbf{Z}_t^{\text{image}} = \text{Attention}(\mathbf{Q}, \mathbf{K}', \mathbf{V}') = \text{Softmax}(\frac{\mathbf{Q}(\mathbf{K}')^\top}{\sqrt{d}})\mathbf{V}'$$

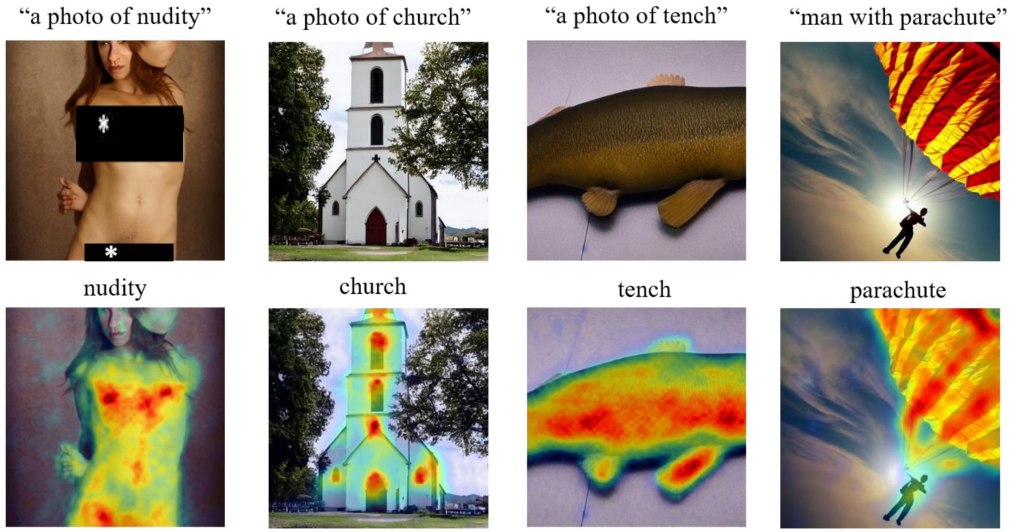
integration cross-attention

$$\mathbf{Z}_t^{\text{att}} = \mathbf{Z}_t^{\text{text}} + \mathbf{Z}_t^{\text{image}}$$

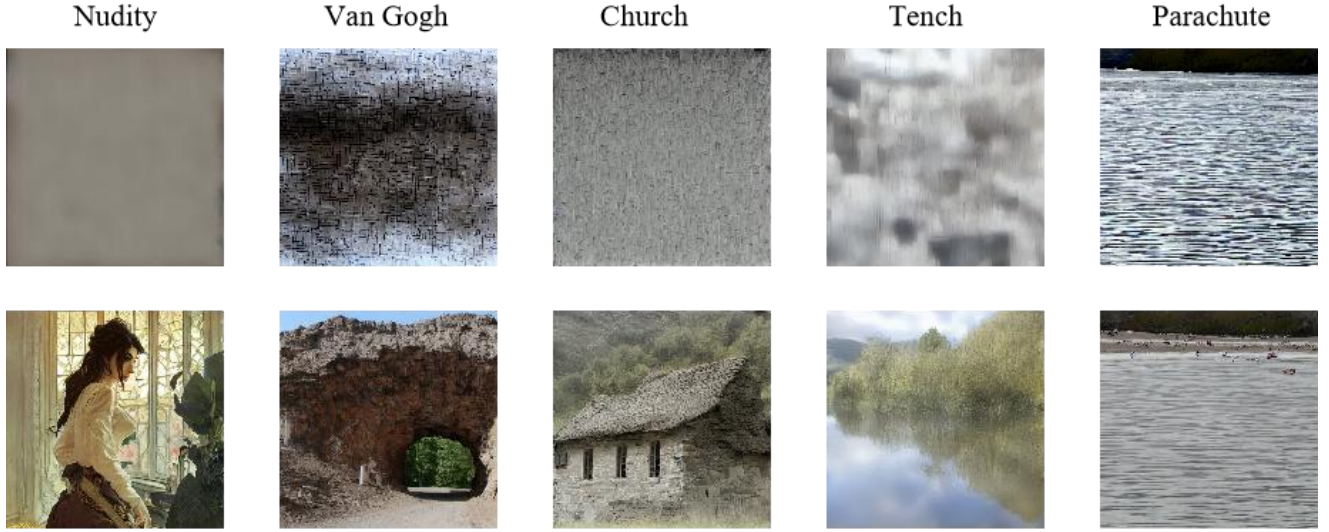
Co-Erasing | Text-Guided Refinement

- Image is usually with abundant and even **redundant visual information**
- Text-guided refinement helps to **extract the target concept from** the image condition, avoids degrading untargeted concept

Visualization of text-guided refinement



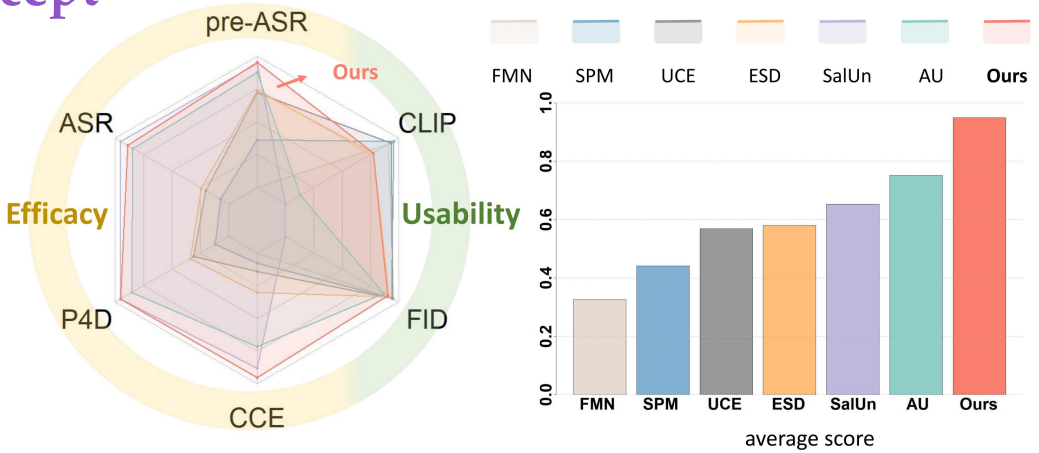
Examples of w/wo text-guided refinement



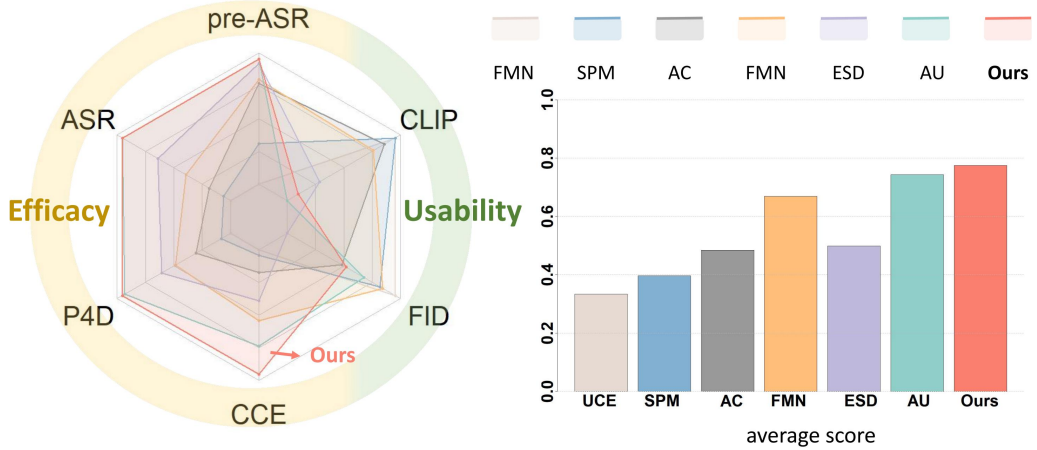
Results | Quantitative

Different Concept

Target concept: Nudity



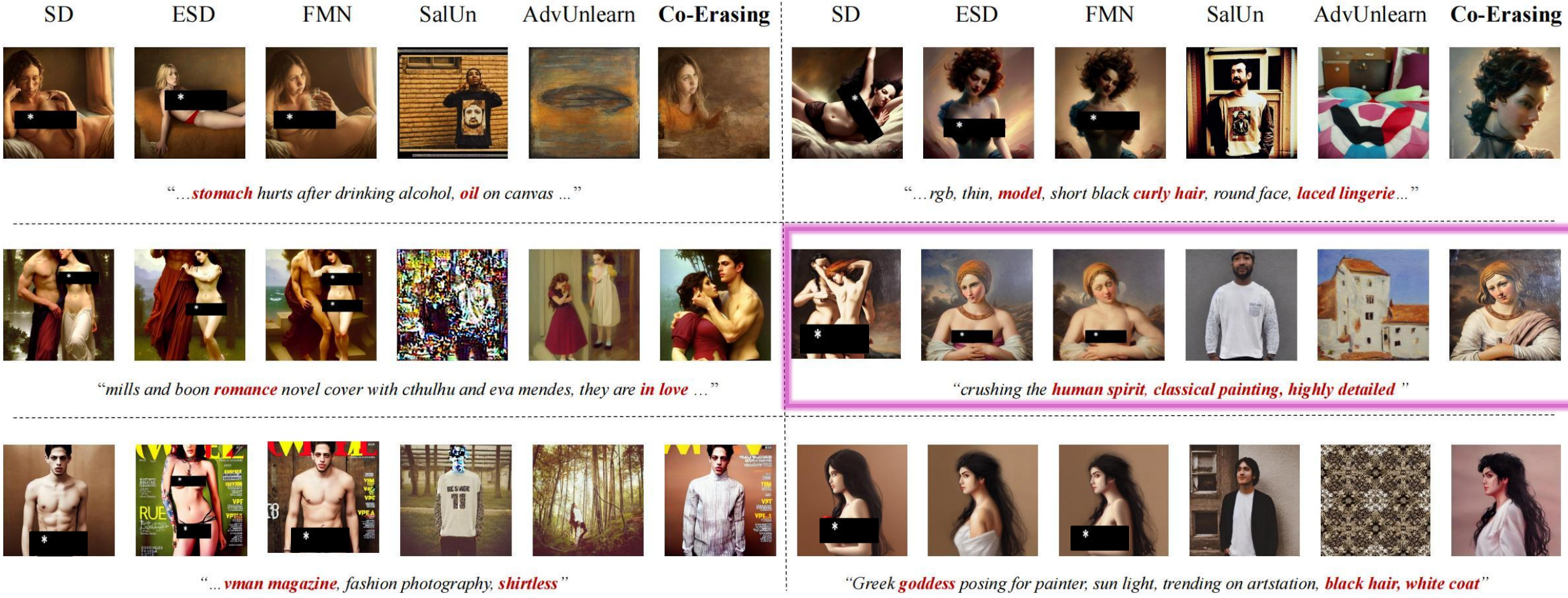
Target concept: Van Gogh



Different Backbone

Method	NudeNet Detection							COCO-10K	
	Breasts(F)	Breasts(M)	Genitalia(F)	Genitalia(M)	Buttocks	Anus	Total ↓	FID ↓	CLIP ↑
SLD	48	43	6	14	14	0	125	20.17	0.298
SLD+Co-Erasing	12	8	0	0	2	0	22	21.15	0.297
ESD	27	8	3	2	2	1	43	18.18	0.302
ESD+Co-Erasing	2	1	0	0	1	0	4	18.77	0.302
MACE	16	9	2	7	2	0	36	17.13	0.277
MACE+Co-Erasing	4	2	1	1	1	0	9	17.12	0.277

Results | Qualitative



Co-Erasing can improve **erasing efficacy** while **preserving usability**

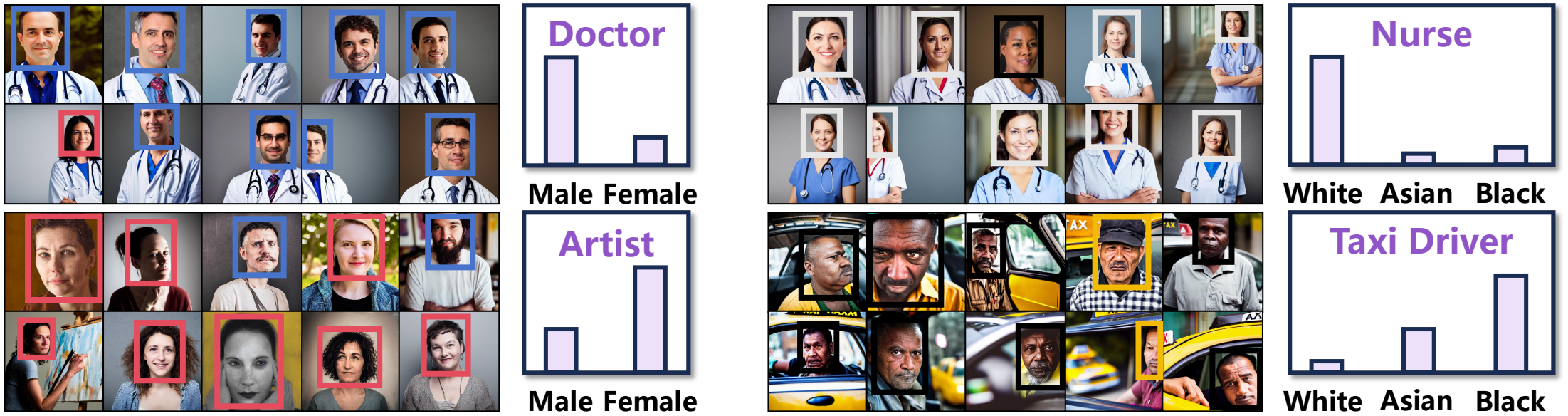
LightFair: Towards an Efficient Alternative for Fair T2I Diffusion via Debiasing Pre-trained Text Encoders

NeurIPS 2025

Fairness:
mitigating generation bias with **lightweight** fair **interventions**

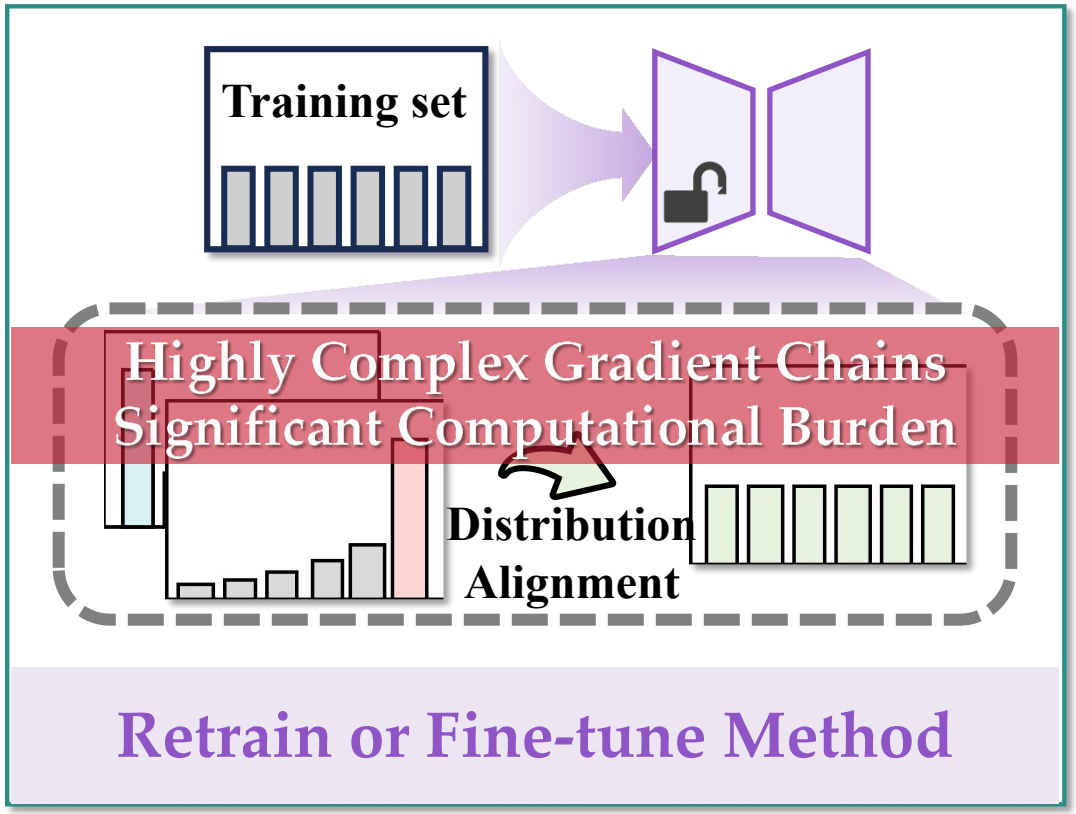
Background

- Diffusion models translate the bias present in the training datasets

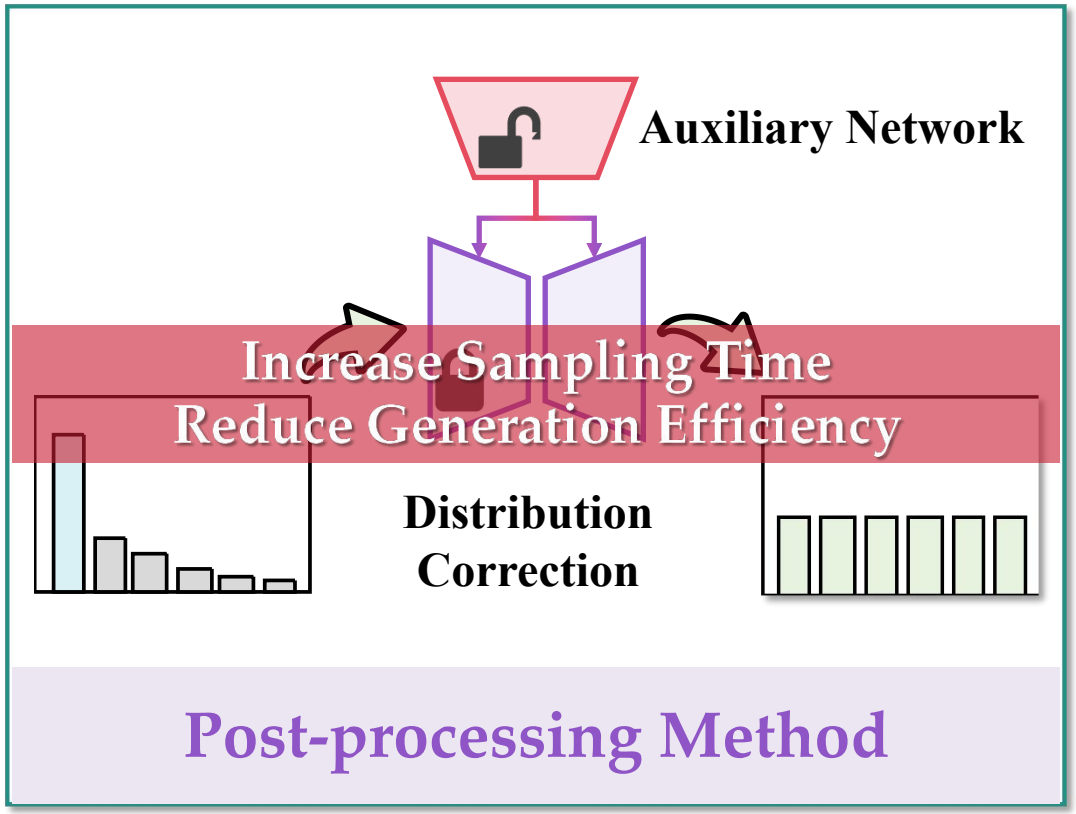


Related Work

- Existing work can be divided into two camps.



Retrain or Fine-tune Method



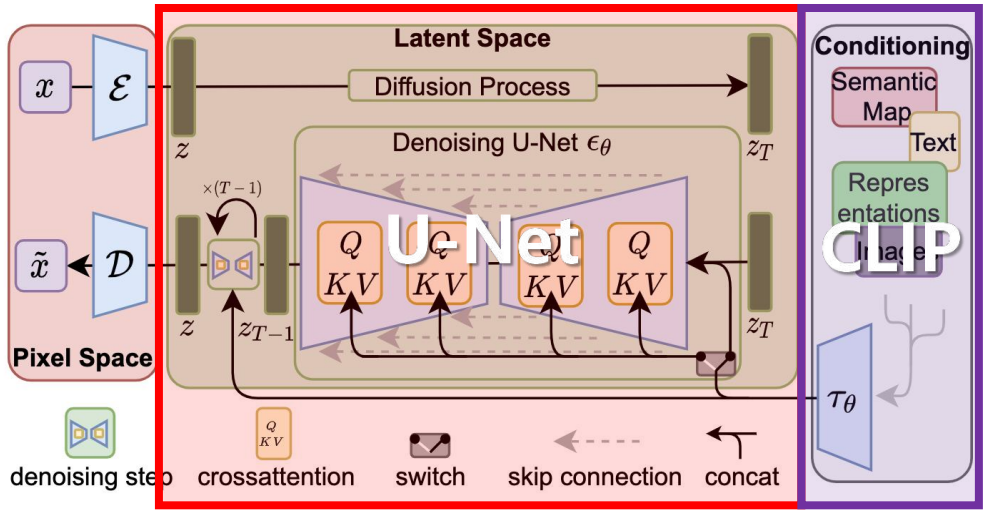
Post-processing Method

Our Goal:

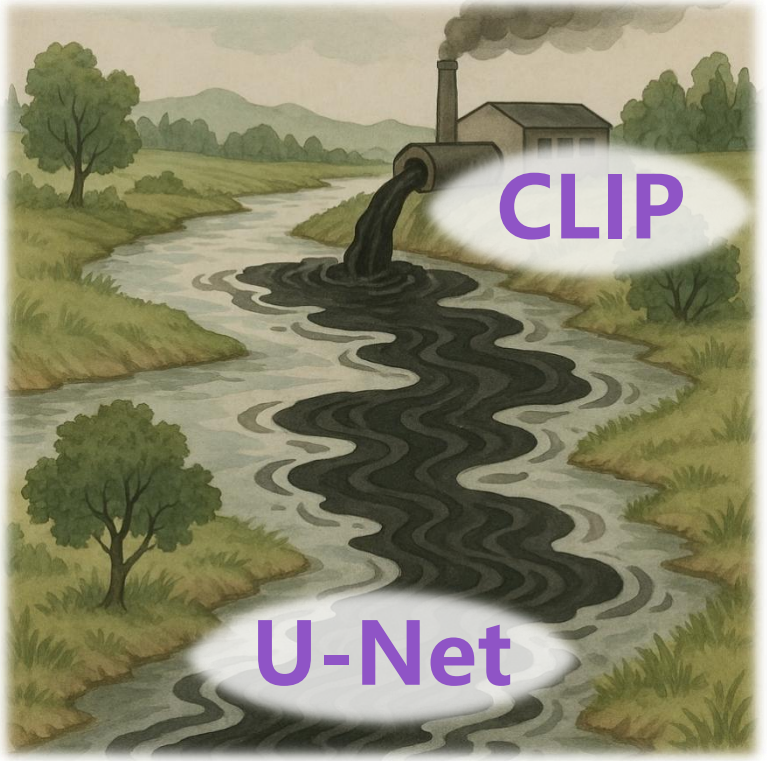
Develop a **lightweight** alternative to resolve attribute bias effectively

LightFair | Step 1: Position

- Stable diffusion consists of **two parts**: CLIP and U-Net
- CLIP: **fewer parameters, upstream input**, **its representations** directly shape how U-Net **interprets** prompts



Method	CLIP Text Encoder	U-Net
Stable Diffusion v1.5	123.060480 M	859.520964 M
Stable Diffusion v2.1	340.387840 M	865.910724 M



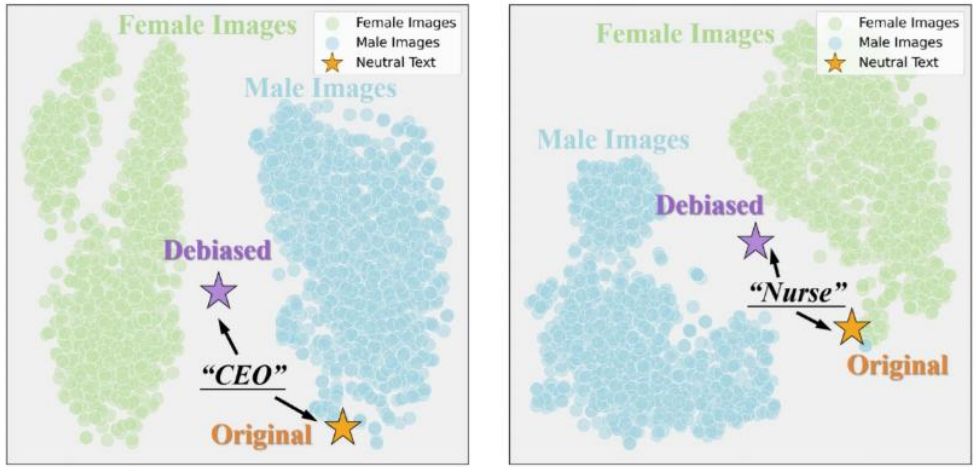
We hypothesize that CLIP is crucial for mitigating bias

LightFair | Step 1: Position

- We focus on the following two key questions:
  - **(Q1)** How can we **measure** bias within the text embeddings?
  - **(Q2)** Does the bias in the CLIP **affect** the noise prediction network?

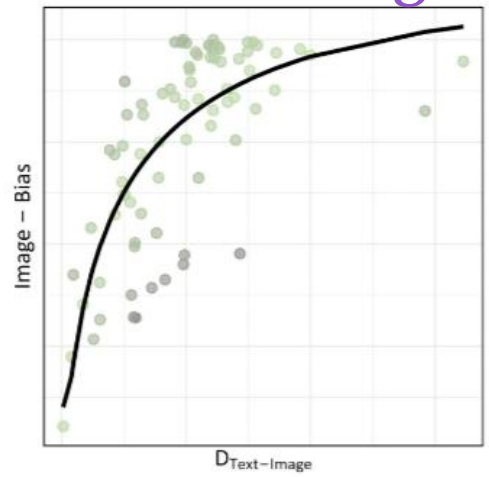
- **(A1):** Distance in CLIP space

Toy Examples



Biased text is closer (more similar) to the corresponding image centroid.

- **(A2):** Relevance – Image Bias



$$D_{\text{Text-Image}} = \left| s\left(\text{emb}_c^T(\cdot), \mathbb{E}[\text{emb}_c^I(\male)]\right) - s\left(\text{emb}_c^T(\cdot), \mathbb{E}[\text{emb}_c^I(\female)]\right) \right|$$

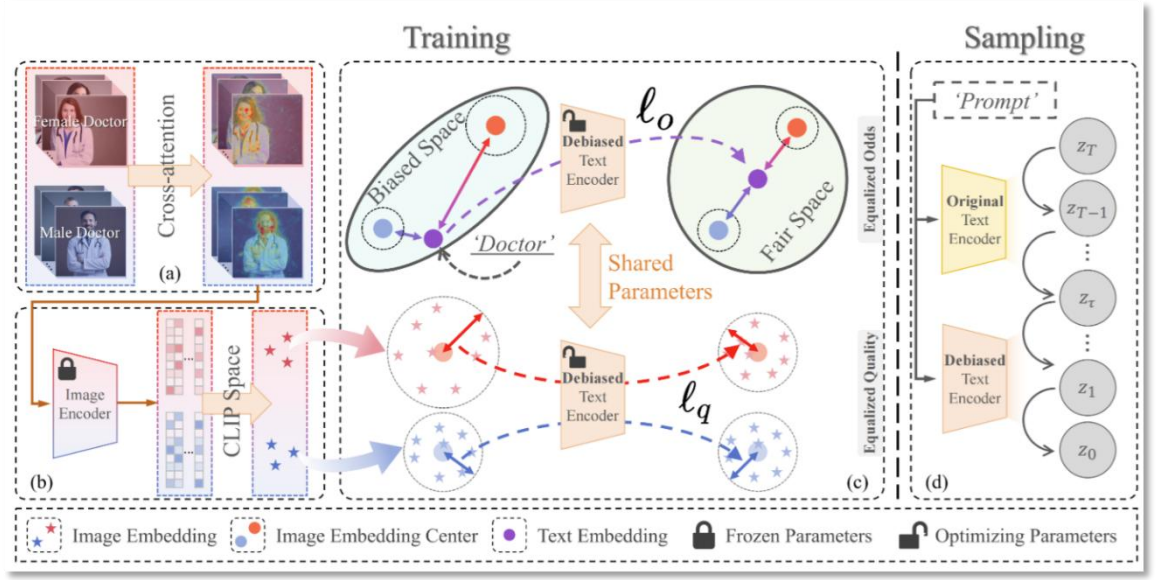
Such a gap is proportional to the degree of bias.

LightFair | Step 1: Position

The text encoder is one of the key yet overlooked structures contributing to attribute bias in SD.

Our Idea:

Construct supervision signals to fine-tune text encoder



LightFair | Step 2: Debiasing

- Fair diffusion models have two goals:

Goal 1: Equalized Odds

$$\mathbb{P}(a_i|prompt(\cdot,c)) = \mathbb{P}(a_j|prompt(\cdot,c)), \quad \forall a_i, a_j \in A$$

Goal 2: Equalized Quality

$$Q(prompt(a_i,c)) = Q(prompt(a_j,c)), \quad \forall a_i, a_j \in A$$

- Collaborative Distance-constrained Debiasing Strategy

Theorem 1 (Equivalence between Equalized Odds and Equalized Distance).

Under a mild assumption, for any attributes $a_i, a_j \in A$ achieving **Equalized Odds** is equivalent to ensuring **Equalized Distance**:

$$\left\| f(a_i,c) - f^t\left(P(\cdot,c)\right)\right\|^2 = \left\| f(a_j,c) - f^t\left(P(\cdot,c)\right)\right\|^2,$$

$$\ell_o = \sqrt{\frac{1}{|A|}\sum_{i=1}^{|A|}\left[s\left(\text{emb}_c^T(\cdot), \mathbb{E}\left[\text{emb}_c^I(a_i)\right]\right) - \bar{s}\right]}$$

Constraint in CLIP space

LightFair | Step 2: Debiasing

- Fair diffusion models have two goals:

Goal 1: Equalized Odds

$$\mathbb{P}(a_i|prompt(\cdot,c)) = \mathbb{P}(a_j|prompt(\cdot,c)), \quad \forall a_i, a_j \in A$$

Goal 2: Equalized Quality

$$Q(prompt(a_i,c)) = Q(prompt(a_j,c)), \quad \forall a_i, a_j \in A$$

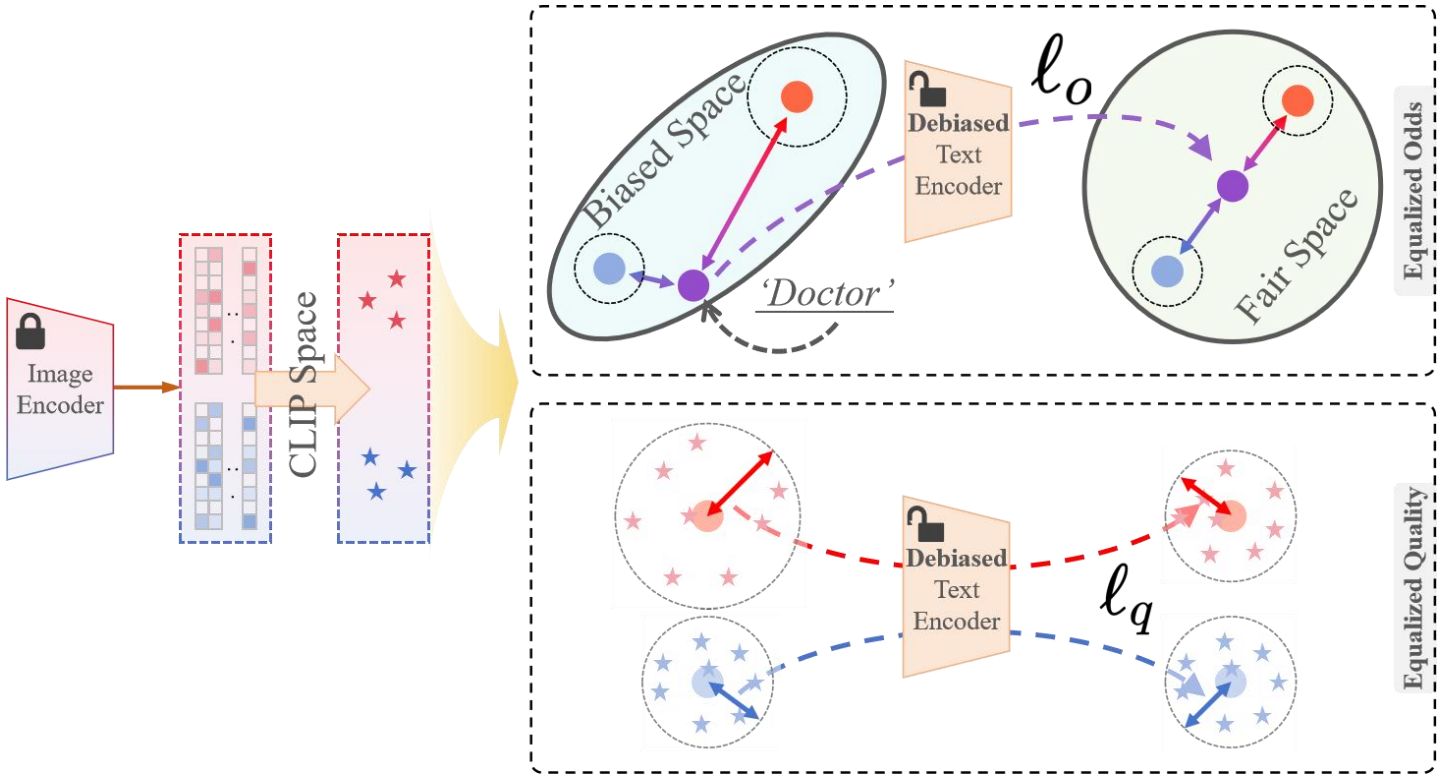
- Collaborative Distance-constrained Debiasing Strategy
  - ✓ Generative models often use **CLIP Score** as a quality metric.
  - ✓ **Equalized Quality** corresponds to **Equalized CLIP Score**.

$$\ell_q = \sqrt{\frac{1}{|A|}\sum_{i=1}^{|A|}\left[s\left(\text{emb}_c^T(a_i)\right), \mathbb{E}\left[\text{emb}_c^I(a_i)\right]\right) - \overline{s'}\Big]^2}$$

Constraint in CLIP space

LightFair | Step 2: Debiasing

◈ Debiasing through distance constraints



Goal 1 ✓ **Equidistant from all attribute centroids**

$$\ell_o = \sqrt{\frac{1}{|A|}\sum_{i=1}^{|A|}\left[s\left(\mathrm{emb}_c^T(\cdot), \mathbb{E}\left[\mathrm{emb}_c^I(a_i)\right]\right) - \bar{s}\right]^2},$$

Goal 2 ✓ **Equal radius of all attributes**

$$\ell_q = \sqrt{\frac{1}{|A|}\sum_{i=1}^{|A|}\left[s\left(\mathrm{emb}_c^T(a_i)\right), \mathbb{E}\left[\mathrm{emb}_c^I(a_i)\right]\right) - \bar{s'}\right]^2},$$

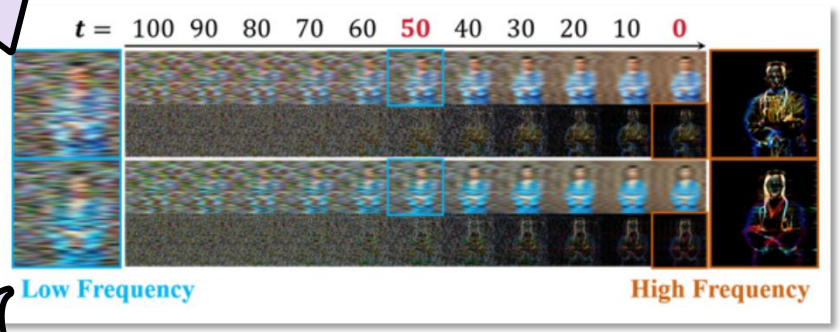
No auxiliary networks or complex gradient chains!
Ensuring lightweight fine-tuning!

LightFair | Step 3: Sampling

◈ Two-Stage Text-Guided Sampling Strategy

Proposition 1 (Frequency Signal Patterns in Diffusion Denoising Process).

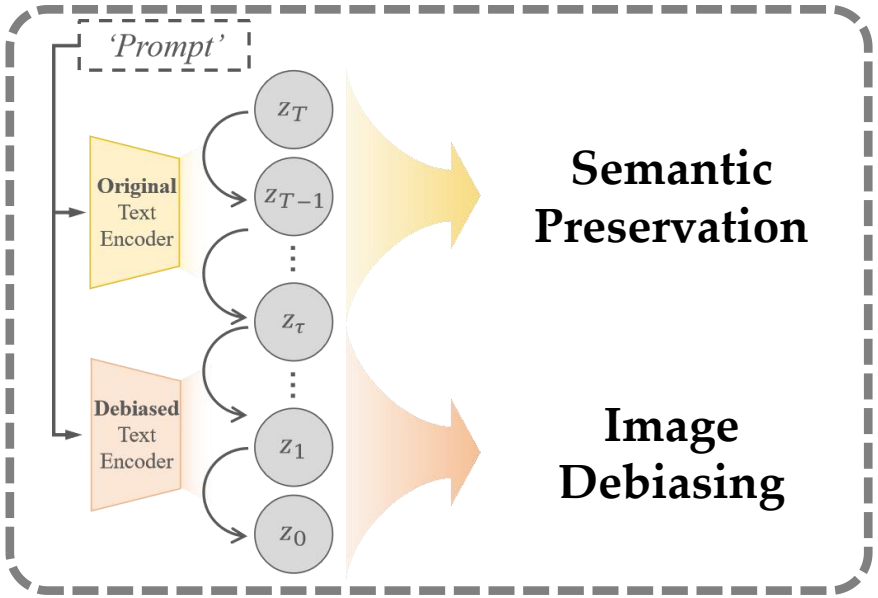
The recovery rate of low-frequency signals during the diffusion denoising process is higher than that of high-frequency signals.



✓ **The attribute concepts are generated at the later stage.**

$$\epsilon_\theta(P, \mathbf{z}_t, t) = \begin{cases} \epsilon_\theta(f^t_{orig}(P), \mathbf{z}_t, t), & t \geq \tau \\ \epsilon_\theta(f^t_{new}(P), \mathbf{z}_t, t), & t < \tau \end{cases}$$

Low-frequency First!



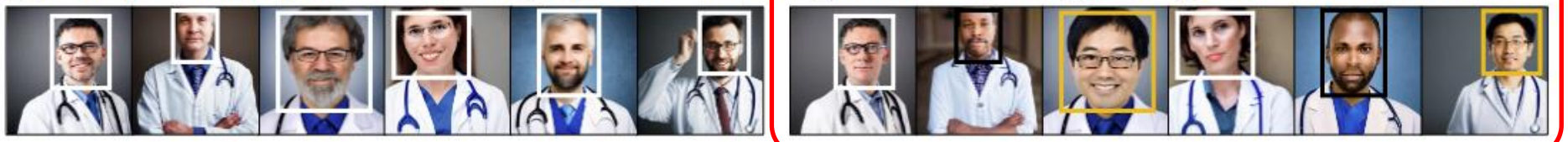
Fewer modifications → Higher generation quality and fidelity

Experiments | Single-attribute Debiasing

◈ Gender / Race

Method	Gender								Race							
	Fairness		Quality						Fairness		Quality					
	Bias-O ↓	Bias-Q ↓	CLIP-T ↑	CLIP-I ↑	FID ↓	IS ↑	AS-R ↑	AS-A ↓	Bias-O ↓	Bias-Q ↓	CLIP-T ↑	CLIP-I ↑	FID ↓	IS ↑	AS-R ↑	AS-A ↓
Stable Diffusion v1.5																
SD [78]	0.73 (±0.05)	1.90 (±0.67)	29.31 (±0.06)	-	275.13 (±6.75)	1.26 (±0.03)	4.78 (±0.08)	2.65 (±0.04)	0.54 (±0.02)	1.60 (±0.67)	29.31 (±0.06)	-	275.13 (±6.75)	1.26 (±0.03)	4.78 (±0.08)	2.65 (±0.04)
FairD [24]	0.79 (±0.04)	3.25 (±1.15)	28.79 (±0.11)	75.91 (±0.56)	269.62 (±4.42)	**1.30** (±0.03)	4.57 (±0.09)	2.82 (±0.05)	0.50 (±0.02)	1.50 (±0.38)	28.95 (±0.10)	74.33 (±0.68)	262.72 (±4.84)	1.28 (±0.03)	4.55 (±0.08)	2.83 (±0.06)
UCE [27]	0.78 (±0.07)	1.79 (±0.46)	28.91 (±0.13)	**82.72** (±0.81)	273.95 (±5.53)	1.26 (±0.03)	4.71 (±0.09)	2.64 (±0.04)	0.44 (±0.03)	1.40 (±0.24)	29.13 (±0.14)	**90.15** (±0.70)	281.16 (±5.18)	1.26 (±0.02)	4.76 (±0.08)	2.69 (±0.05)
FinetuneFD [84]	0.38 (±0.07)	2.31 (±0.35)	29.34 (±0.13)	76.17 (±0.68)	278.21 (±7.53)	1.24 (±0.02)	4.38 (±0.06)	2.86 (±0.04)	0.20 (±0.03)	1.41 (±0.23)	29.02 (±0.15)	74.57 (±0.53)	270.09 (±5.99)	1.26 (±0.02)	4.33 (±0.06)	2.87 (±0.05)
FairMapping [51]	0.46 (±0.05)	2.16 (±0.72)	29.30 (±0.16)	76.00 (±0.66)	278.81 (±5.84)	1.26 (±0.02)	4.34 (±0.07)	2.90 (±0.03)	0.34 (±0.02)	1.75 (±0.47)	29.29 (±0.15)	76.54 (±0.71)	280.95 (±5.02)	1.26 (±0.03)	4.53 (±0.08)	2.80 (±0.05)
BalancingAct [70]	0.41 (±0.05)	1.70 (±0.55)	29.30 (±0.11)	77.37 (±0.64)	272.08 (±5.16)	1.28 (±0.02)	4.71 (±0.06)	2.68 (±0.04)	0.34 (±0.02)	1.13 (±0.36)	29.34 (±0.11)	77.44 (±0.72)	271.91 (±5.35)	1.29 (±0.03)	4.72 (±0.10)	2.66 (±0.04)
LightFair (Ours)	**0.30** (±0.08)	**0.99** (±0.55)	**30.57** (±0.16)	80.09 (±0.76)	**233.53** (±5.50)	**1.30** (±0.03)	**4.79** (±0.08)	**2.60** (±0.04)	**0.18** (±0.04)	**1.06** (±0.43)	**31.34** (±0.20)	86.31 (±0.70)	**259.96** (±7.75)	**1.33** (±0.03)	**4.80** (±0.10)	**2.55** (±0.04)
Stable Diffusion v2.1																
SD [78]	0.85 (±0.05)	1.84 (±0.63)	29.90 (±0.15)	-	259.36 (±4.81)	1.23 (±0.03)	5.12 (±0.05)	**2.24** (±0.03)	0.63 (±0.01)	2.06 (±0.35)	29.90 (±0.15)	-	259.36 (±4.81)	1.23 (±0.03)	5.12 (±0.05)	2.24 (±0.03)
debias VL [15]	0.43 (±0.09)	1.44 (±0.48)	28.20 (±0.22)	70.01 (±0.96)	245.11 (±3.72)	1.35 (±0.03)	3.53 (±0.11)	2.93 (±0.06)	0.49 (±0.03)	1.91 (±0.92)	28.15 (±0.26)	67.42 (±0.96)	242.78 (±4.21)	1.33 (±0.03)	3.57 (±0.11)	2.85 (±0.06)
UCE [27]	0.90 (±0.04)	1.67 (±0.71)	29.41 (±0.13)	**87.94** (±0.86)	268.52 (±3.92)	1.22 (±0.02)	5.12 (±0.05)	2.32 (±0.03)	0.50 (±0.03)	1.95 (±0.37)	29.44 (±0.12)	**80.46** (±1.13)	250.57 (±4.49)	1.23 (±0.03)	5.17 (±0.08)	2.25 (±0.03)
LightFair (Ours)	**0.33** (±0.10)	**1.40** (±0.28)	**30.82** (±0.19)	75.29 (±0.99)	**231.46** (±3.30)	**1.35** (±0.02)	**5.14** (±0.09)	2.24 (±0.06)	**0.40** (±0.03)	**1.82** (±0.44)	**30.26** (±0.16)	77.47 (±1.05)	**230.59** (±6.53)	**1.35** (±0.01)	**5.29** (±0.11)	**2.14** (±0.06)

(a) Prompt: "Photo portrait of a **CEO**, a person". Left: original SD v1.5. Right: our **gender**-debiased **SD v1.5**.

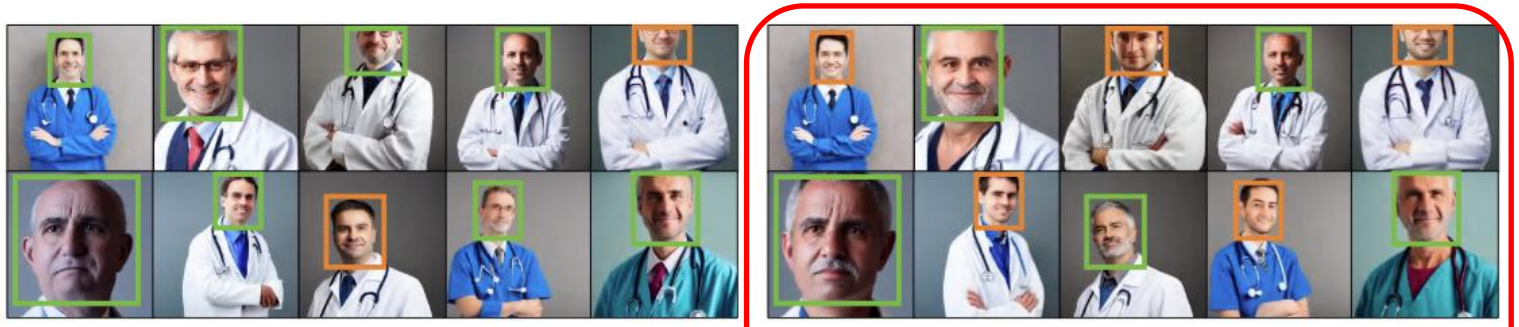


(b) Prompt: "Photo portrait of a **doctor**, a person". Left: original SD v2.1. Right: our **race**-debiased **SD v2.1**.

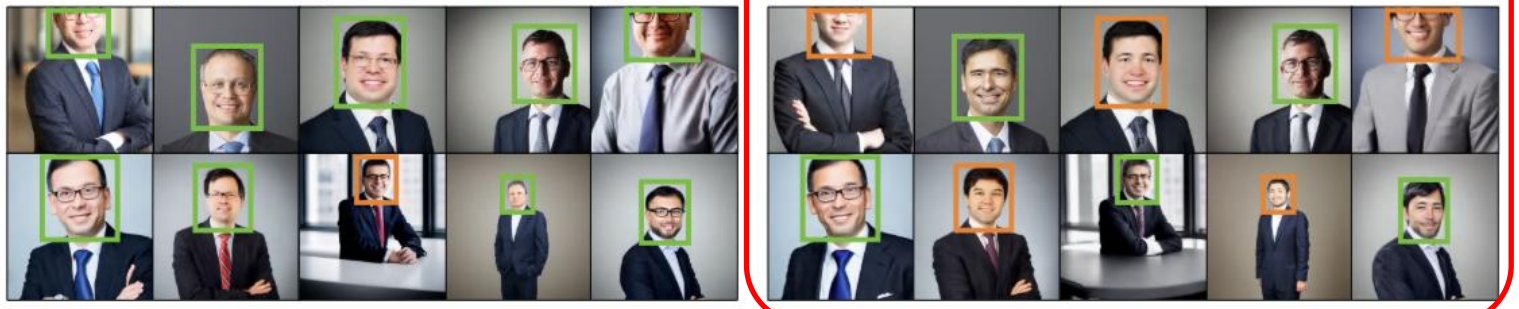
Experiments | Single-attribute Debiasing

Age

Backbone	Method	Fairness		Quality			
		Bias-O ↓	Bias-Q ↓	CLIP-T ↑	CLIP-I ↑	FID ↓	IS ↑
SD v1.5	SD	0.65 (±0.04)	1.23 (±0.44)	29.23 (±0.06)	-	311.56 (±11.95)	1.23 (±0.02)
	Ours	**0.34** (±0.02)	**0.95** (±0.33)	**30.15** (±0.06)	**79.56** (±4.11)	**278.66** (±9.54)	**1.25** (±0.02)
SD v2.1	SD	0.83 (±0.07)	1.14 (±0.32)	29.30 (±0.11)	-	287.64 (±10.67)	1.24 (±0.01)
	Ours	**0.32** (±0.04)	**0.89** (±0.23)	**31.23** (±0.07)	**82.22** (±6.23)	**246.36** (±9.03)	**1.26** (±0.01)



(a) Prompt: "Photo portrait of a doctor, a person". Left: original SD v1.5. Right: our debiased **SD v1.5**.



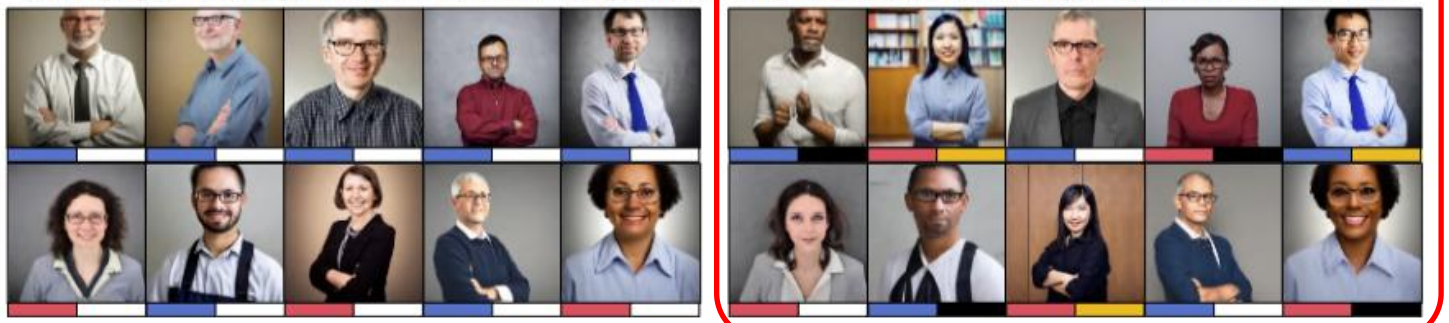
(b) Prompt: "Photo portrait of a CEO, a person". Left: original SD v2.1. Right: our debiased **SD v2.1**.

Experiments | Multi-attribute Debiasing

◈ Gender × Race (**Multiple Attribute**)

Backbone	Method	Fairness		Quality			
		Bias-O ↓	Bias-Q ↓	CLIP-T ↑	CLIP-I ↑	FID ↓	IS ↑
SD v1.5	SD	0.29 (±0.01)	1.31 (±0.54)	29.32 (±0.06)	-	275.85 (±6.29)	1.26 (±0.03)
	Ours	**0.14** (±0.01)	**0.91** (±0.32)	**31.34** (±0.20)	**62.82** (±5.57)	**259.96** (±7.75)	**1.33** (±0.03)
SD v2.1	SD	0.32 (±0.01)	1.12 (±0.37)	29.90 (±0.15)	-	259.36 (±4.81)	1.23 (±0.03)
	Ours	**0.23** (±0.02)	**0.89** (±0.22)	**30.32** (±0.19)	**56.41** (±2.04)	**230.39** (±5.95)	**1.25** (±0.01)

(a) Prompt: "Photo portrait of a taxi driver, a person". Left: original SD v1.5. Right: our debiased **SD v1.5**.



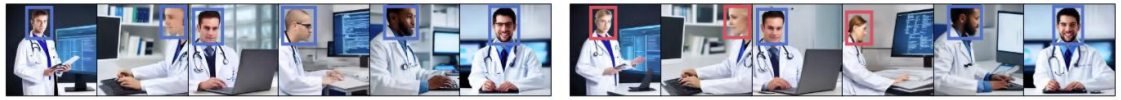
(b) Prompt: "Photo portrait of a teacher, a person". Left: original SD v2.1. Right: our debiased **SD v2.1**.

Experiments | Diverse Prompts

- **Non-templated prompts:** LAION-Aesthetics V2 dataset
- **Multiple people:** "Photo portrait of two/three {occupation}, two/three people"

Prompt	Method	Stable Diffusion v1.5 Bias-O ↓	Stable Diffusion v1.5 Bias-Q ↓	Stable Diffusion v1.5 CLIP-T ↑	Stable Diffusion v2.1 Bias-O ↓	Stable Diffusion v2.1 Bias-Q ↓	Stable Diffusion v2.1 CLIP-T ↑
Non-templated	SD	0.61 (±0.25)	1.32 (±0.19)	32.06 (±1.65)	0.46 (±0.19)	1.43 (±0.24)	32.02 (±2.04)
	Ours	**0.48** (±0.27)	**1.02** (±0.15)	**32.62** (±2.02)	**0.34** (±0.23)	**1.13** (±0.15)	**32.77** (±1.99)
Two People	SD	0.35 (±0.04)	1.23 (±0.23)	30.46 (±0.14)	0.65 (±0.03)	1.76 (±0.22)	32.32 (±0.17)
	Ours	**0.13** (±0.04)	**0.89** (±0.12)	**30.90** (±0.22)	**0.54** (±0.05)	**1.11** (±0.13)	**32.50** (±0.25)
Three People	SD	0.46 (±0.05)	1.77 (±0.31)	31.17 (±0.18)	0.70 (±0.03)	2.01 (±0.42)	32.99 (±0.18)
	Ours	**0.30** (±0.04)	**1.05** (±0.20)	**32.49** (±0.04)	**0.62** (±0.04)	**1.43** (±0.21)	**33.89** (±0.25)

(a) Prompt: "Photograph of a doctor holding a headset sitting in front of a laptop". Left: original SD v1.5. Right: our debiased **SD v1.5**.



(b) Prompt: "A doctor in a white coat on a computer screen". Left: original SD v1.5. Right: our debiased **SD v1.5**.

(a) Prompt: "Photo portrait of two taxi drivers, **two people**". Left: original SD v1.5. Right: our debiased **SD v1.5**.

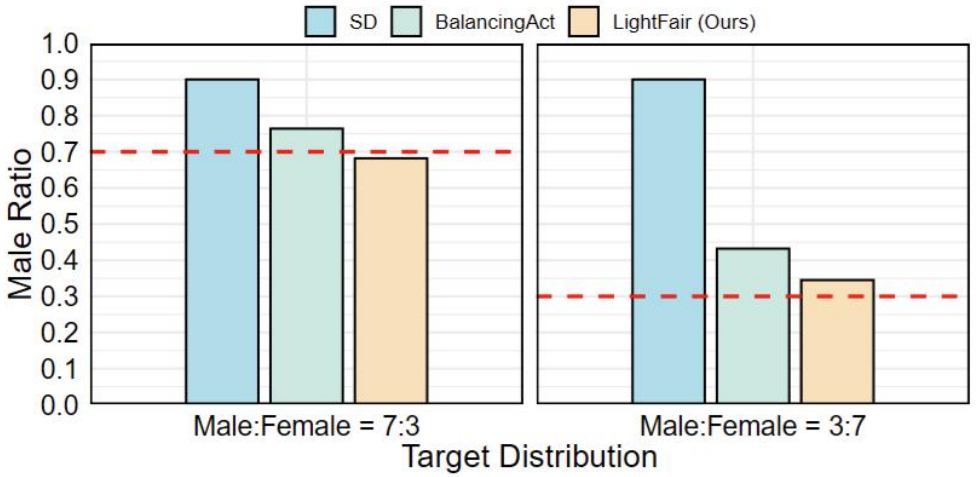
Experiments | Diverse Target Distributions

◈ When the target distribution is **non-uniform**, LightFair can be adapted by simply **adding one additional parameter**

$$\ell_o = \sqrt{\frac{1}{|A|}\sum_{i=1}^{|A|}\left[s\left(\mathrm{emb}_c^T(\cdot), \mathbb{E}\left[\mathrm{emb}_c^I(a_i)\right]\right) - \bar{s}\right]^2},$$

$$\ell_o = \sqrt{\frac{1}{|A|}\sum_{i=1}^{|A|}\left[\gamma_i s\left(\mathrm{emb}_c^T(\cdot), \mathbb{E}\left[\mathrm{emb}_c^I(a_i)\right]\right) - \bar{s}\right]^2},$$

Control the target distribution



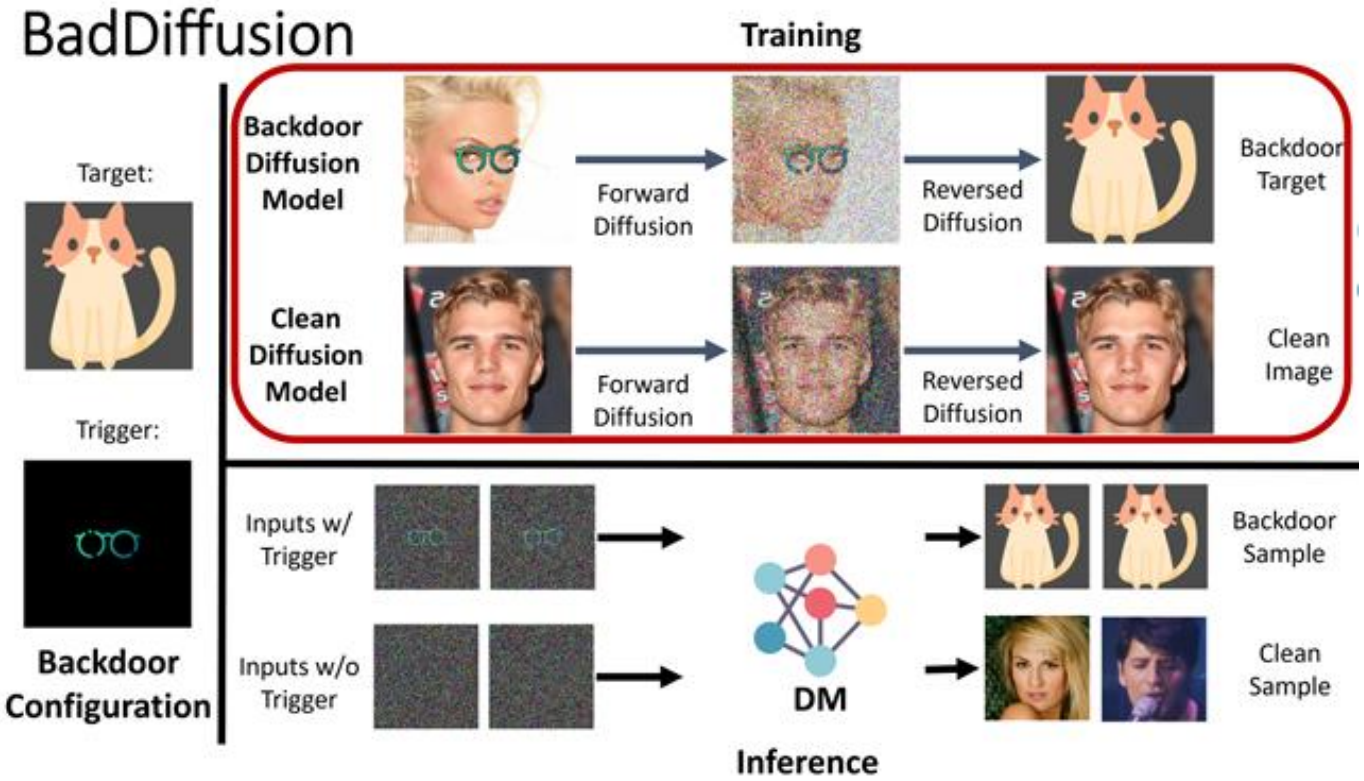
MixBridge: Heterogeneous Image-to-Image Backdoor Attack through Mixture of Schrödinger Bridges

ICML 2025

Security Robustness:
exposing **backdoor vulnerabilities** to inform robust defenses

Background

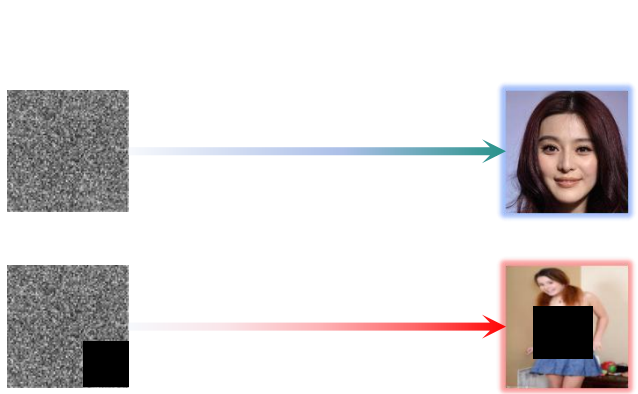
- **Training:** Implants backdoor triggers into diffusion models via **data poisoning**
- **Inference:** Manipulate the outputs by injecting **triggers** into the inputs



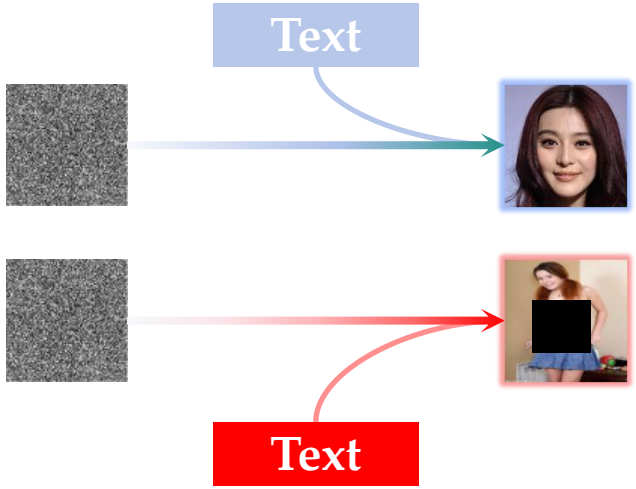
[1] Chou, Sheng-Yen, Pin-Yu Chen, and Tsung-Yi Ho. How to backdoor diffusion models?. CVPR 2023

Background

- Existing backdoor research primarily focuses on **unconditional diffusion models** and **Text-to-Image diffusion models**
- Existing backdoor research merely focuses on the **single** backdoor attack



Unconditional Diffusion Single Backdoor Attack



Text-to-Image Diffusion Single Backdoor Attack

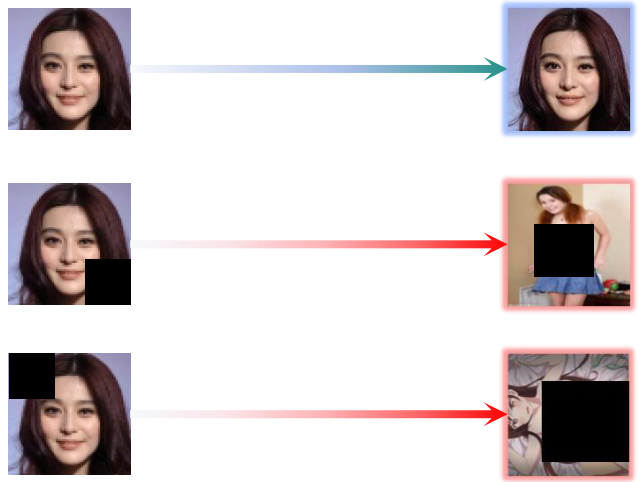


Image-to-Image Diffusion Heterogeneous Backdoor Attack (Ours)

Background

◈ The Image-to-Image Diffusion Schrödinger Bridge (I2SB) is an **entropy-regularized optimal transport** formulation

◈ I2SB enables us to generate an image $\boldsymbol{x}_0$ from a corrupted input image $\boldsymbol{x}_1$

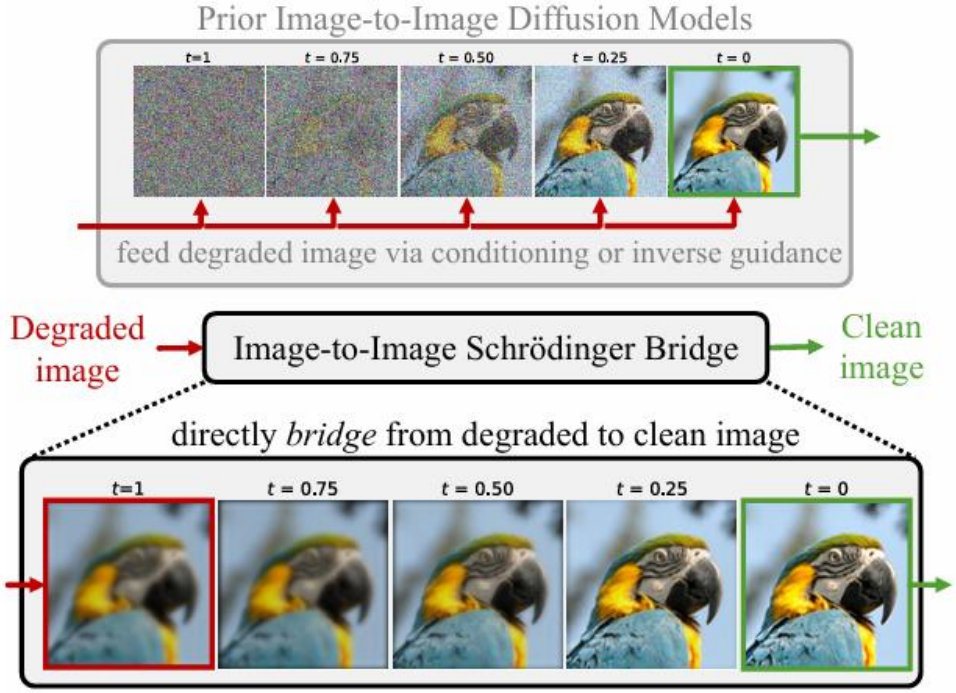
Forward/Backward SDE:

$$d\boldsymbol{x}_t = [f(\boldsymbol{x}_t) + g^2(t)\nabla_{\boldsymbol{x}_t} \log \Psi(\boldsymbol{x}_t)]\, dt + g(t)\, d\boldsymbol{w}_t$$
$$d\boldsymbol{x}_t = [f(\boldsymbol{x}_t) - g^2(t)\nabla_{\boldsymbol{x}_t} \log \hat{\Psi}(\boldsymbol{x}_t)]\, dt + g(t)\, d\bar{\boldsymbol{w}}_t$$

Denoising score matching

I2SB Training Objective → DDPM

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{t\sim\mathbb{U}(\mathbb{0},\mathbb{1}),\boldsymbol{x}_0,\boldsymbol{x}_1,\boldsymbol{x}_t}\left[\left|\left|\epsilon_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \frac{\boldsymbol{x}_t - \boldsymbol{x}_0}{\sigma_t}\right|\right|_2^2\right]$$



[1] Liu, Guan-Horng, et al. I^{2}SB: Image-to-Image Schrödinger Bridge. ICML 2023
[2] Zhou, Linqi, et al. Denoising Diffusion Bridge Models. ICLR 2024

I2SB Model | Naive Backdoor Injection

◈ Backdoor can be easily injected by poisoning training data

Proposition (Image generation with pair relationship)

Given the image pairs $(\boldsymbol{x}_0^{p,i}, \boldsymbol{x}_1^{p,i})$ or $(\boldsymbol{x}_0^{c}, \boldsymbol{x}_1^{c})$ in the training datasets, the ground-truth sample-path of I2SB always generate images consistent with pairwise relationships with $t \to 1$ and $t \to 0$.

The i-th poisoned training dataset

The clean training dataset

I2SB Backdoor Training Objective

$$\mathcal{L}_{\text{backdoor}}(\boldsymbol{\theta}) = \ell^c + \sum_{i=1}^{M} \ell^i$$

Clean Loss + Attack Loss

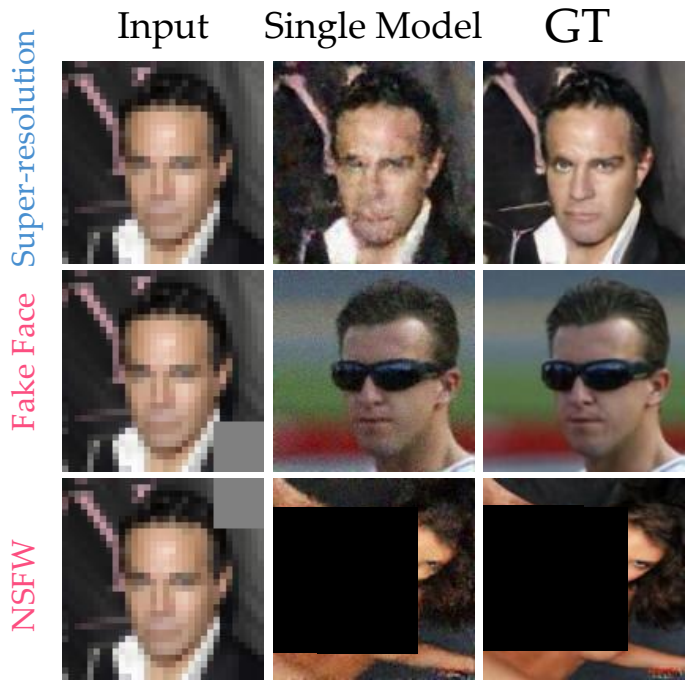
$$\ell^c = p(c) \cdot \mathbb{E}_{t\sim\mathbb{U}(\mathbb{0},\mathbb{1}),\boldsymbol{x}_t^c,\boldsymbol{x}_0^c,\boldsymbol{x}_1^c}\left[\left|\left|\epsilon_{\boldsymbol{\theta}}(\boldsymbol{x}_t^c, t) - \frac{\boldsymbol{x}_t^c - \boldsymbol{x}_0^c}{\sigma_t}\right|\right|_2^2\right]$$

$$\ell^i = p(i) \cdot \mathbb{E}_{t\sim\mathbb{U}(\mathbb{0},\mathbb{1}),\boldsymbol{x}_t^{p,i},\boldsymbol{x}_0^{p,i},\boldsymbol{x}_1^{p,i}}\left[\left|\left|\epsilon_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{p,i}, t) - \frac{\boldsymbol{x}_t^{p,i} - \boldsymbol{x}_0^{p,i}}{\sigma_t}\right|\right|_2^2\right]$$

Observation | Struggle for Unified Generation

◈ The expressivity of a single I2SB model is **insufficient to capture multiple sample paths** simultaneously, including both benign and backdoor behaviors.

◈ The output distributions differ **significantly**, even though the input images **differ only** by a **minor** backdoor trigger.



A single model struggles to generate satisfactory outputs for both clean and poisoned inputs

Why | A Theoretical Limitation of the I2SB Model

◈ **Limitation**: The I2SB model tends to **average over input-conditioned distributions**, which can harm both benign generation quality and the effectiveness of heterogeneous backdoor attacks.

Theorem (Limitations of the I2SB model under heterogeneous backdoor attacks)

Given an arbitrary image $\boldsymbol{x}_0$, the posterior of a trained I2SB model can be formulated as $\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_0)$. If we assume $\nabla_{\boldsymbol{x}_t}\varepsilon_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t; \boldsymbol{\theta})$ possesses full column rank, the posterior is proportional to the **Geometric Average** of the mixture distribution of all generation tasks, i.e.:

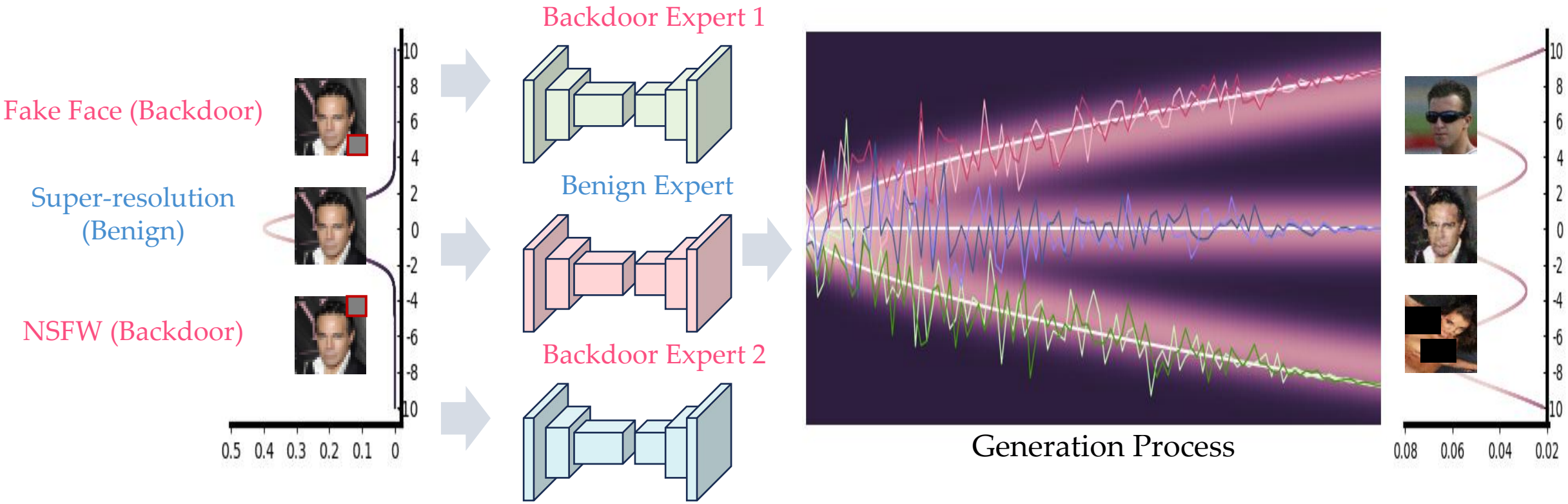
$$\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_0) \propto \prod_i p(\boldsymbol{x}_t|\boldsymbol{x}_0, i)^{p(i|\boldsymbol{z})},$$

where $p(i|\boldsymbol{z}) = p(i|\boldsymbol{x}_t, \boldsymbol{x}_0, \boldsymbol{x}_1^i)$, where i refers to a specific member among the clean and backdoored distributions.

Distribution collapse under multiple backdoor triggers – how can we fix it?

MixBridge | Decoupling Benign and Backdoor Generation

- To solve distribution collapse, we propose a divide-and-merge strategy inspired by the Mixture of Experts (MoE) mechanism
- MixBridge assigns specialized experts to different generation modes



MixBridge | Stealthiness of Backdoor Routing

◈ Without any constraint, the router weights tend to **concentrate on a single** expert for backdoor inputs

◈ Such a "hard routing" pattern is highly **suspicious** and can be **easily detected**

Weight Reallocation Scheme

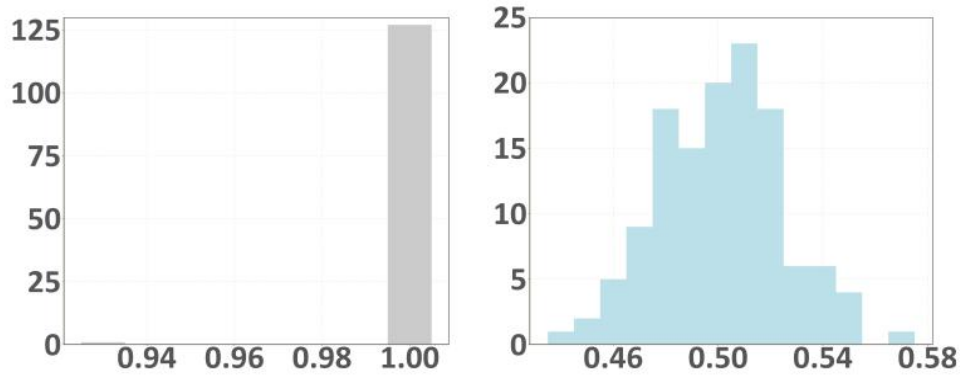
$$\mathcal{L}_{WRS} = \mathbb{E}_{\boldsymbol{w}} \left[\left\| \boldsymbol{w} - \frac{1}{M+1} \right\|^2 \right]$$

stay close to a uniform prior

Entropy

$$H(\boldsymbol{w}) = -w_c \log w_c - \sum_{i=1}^{M} w_i \log w_i$$

higher → more stealthy



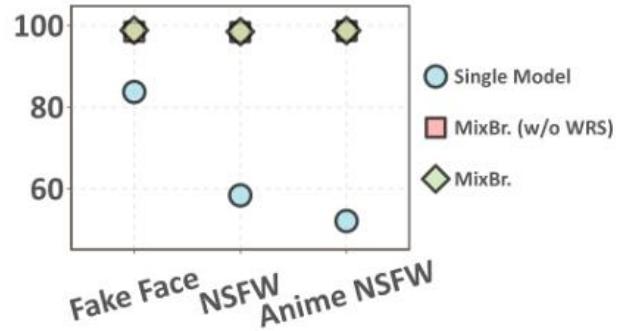
(a) Without WRS (b) With WRS

Figure 4: **The distribution of weight $\boldsymbol{w}$.** The weight concentrates around 1 for the backdoor attack without WRS (*Left*), and the weight balances to 0.5 with WRS (*Right*).

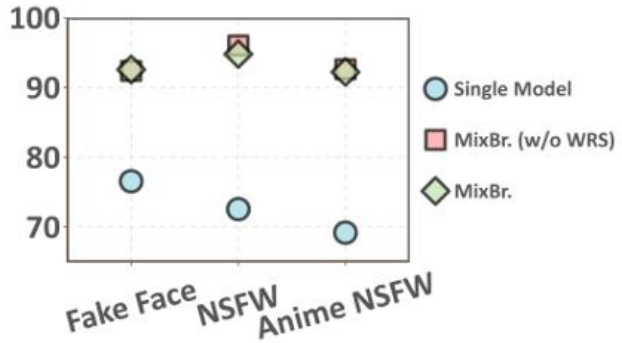
Experiment

◈ MixBridge maintains **comparable benign** super-resolution quality while achieving **high attack** success rates

	SR. (Benign)	Fake Face	NSFW	Anime NSFW	Super-resolution Evaluation (Benign) FID↓	PSNR↑	SSIM (E-02) ↑	*Per-Task* Backdoor Average MSE (E-02) ↓	CLIP (E-02) ↑	ASR↑	Entro. (Steal.) ↑
I2SB	✓				72.59	**27.55**	81.72	36.71	58.42	0.00	0.00
Single Model	✓	✓			132.83	25.64	71.13	27.16	66.63	32.77	0.00
	✓		✓		135.82	25.71	71.62	25.25	67.44	32.50	0.00
	✓			✓	134.46	25.21	67.37	22.43	65.32	30.10	0.00
	✓	✓	✓		143.42	25.38	69.30	16.00	73.23	62.98	0.00
	✓	✓		✓	158.23	25.18	68.19	12.68	69.47	53.69	0.00
	✓		✓	✓	159.85	25.33	68.99	11.60	71.54	52.11	0.00
	✓	✓	✓	✓	161.35	24.98	66.19	3.05	71.59	60.91	0.00
M.B. (Ours) w/o WRS	✓	✓			41.48	27.46	**83.35**	27.06	69.77	33.17	4e-03
	✓		✓		**41.20**	27.33	81.54	25.10	71.16	33.19	3e-03
	✓			✓	42.26	27.29	81.40	22.11	70.80	33.04	8e-03
	✓	✓	✓		61.73	26.72	83.18	13.13	83.03	63.79	3e-03
	✓	✓		✓	61.25	25.74	76.00	11.06	74.68	62.63	2e-03
	✓		✓	✓	63.51	26.63	83.00	11.74	78.32	60.65	1e-03
	✓	✓	✓	✓	71.84	25.68	82.33	1.17	88.85	96.45	7e-03
M.B. (Ours) w WRS	✓	✓			60.65	26.43	82.17	27.04	69.60	33.25	0.99
	✓		✓		68.12	26.50	80.57	25.24	71.05	33.08	0.99
	✓			✓	66.59	26.70	80.59	22.08	70.57	33.19	0.99
	✓	✓	✓		80.88	24.61	73.37	15.83	79.46	66.04	1.57
	✓	✓		✓	83.85	24.07	71.07	12.48	78.85	65.75	1.57
	✓		✓	✓	77.21	24.99	73.69	10.88	81.74	65.84	1.57
	✓	✓	✓	✓	85.88	24.36	70.40	**1.13**	**88.94**	**96.98**	**1.99**

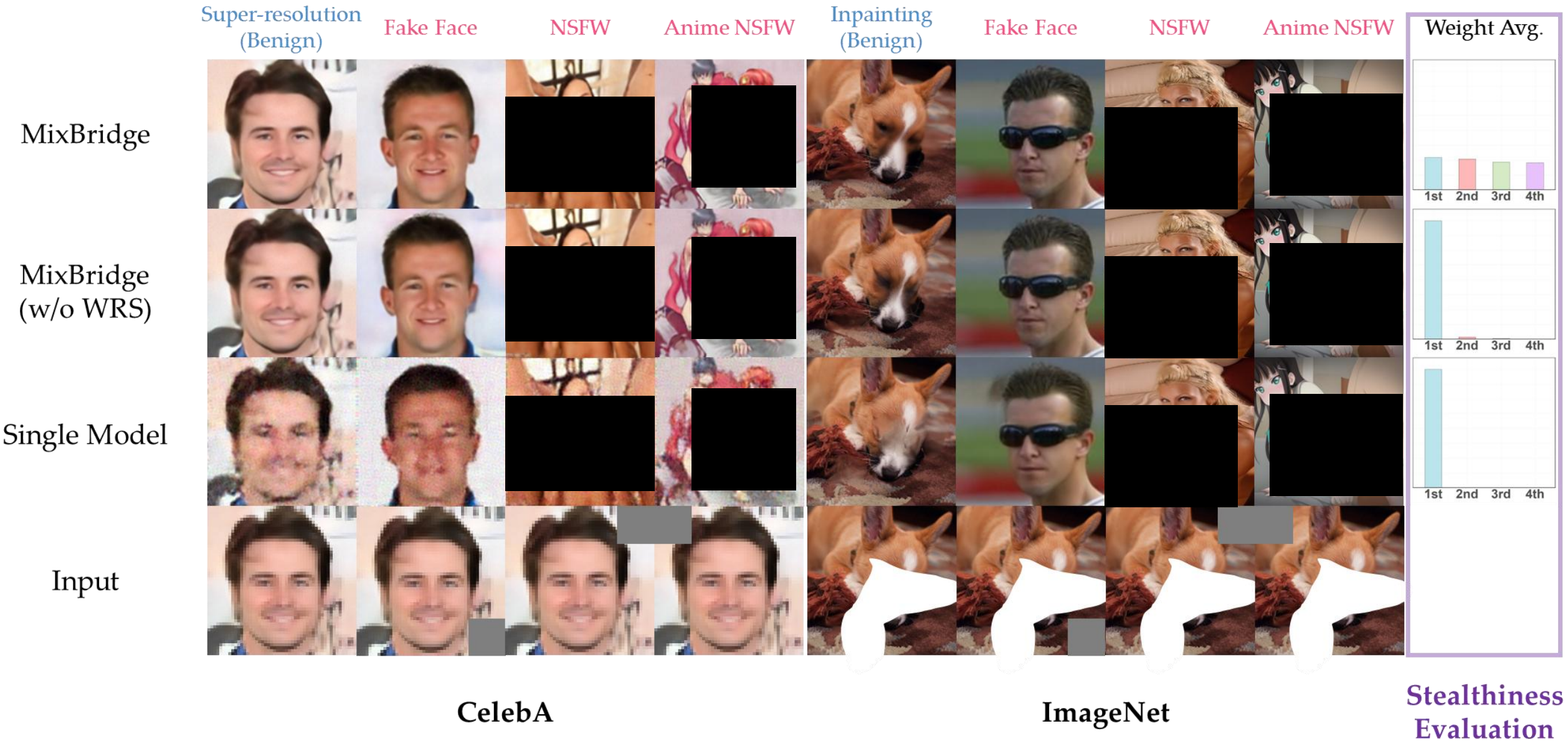


ASR



CLIP Score

Visualization



Outlines

1. *Background & Motivation*
2. *Our Explorations*
3. *Future Directions*

Open Problems | Local to Holistic Harmlessness

- **More** evaluation aspects
- **Open-world** harmful concepts
- Harmless-Utility **Trade-offs**
- **Beyond image generation**
- **Beyond Diffusion**
- **Black-box** detection & defense
- Theoretical guarantees
- …

Thanks!
Q & A

https://statusrank.github.io/

baoshilong@ucas.ac.cn