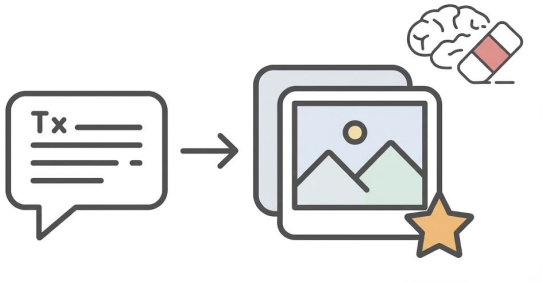
강화학습 기반의 Image Reward를 활용한 T2I 모델 및 Unlearning 방법론 분석

120250375 권범윤
20221633 홍지민



[github link]

목차

프로젝트 주제 및 목표

프로젝트 주제

- 인간의 선호도를 학습한 강화학습 모델 “Image Reward”[1]를 활용하여 concept erasure method를 비교한다.

프로젝트 목표

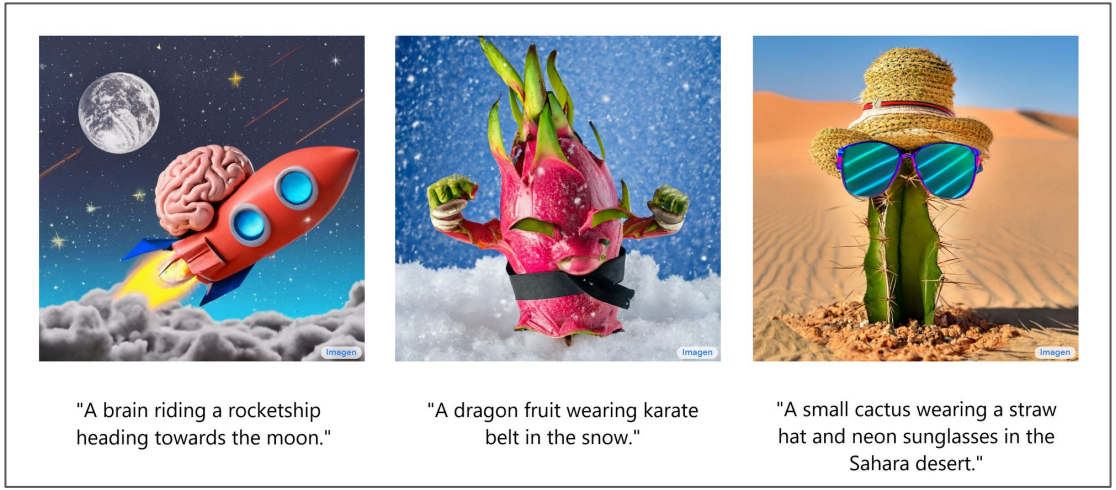
- Image Reward를 활용하여 주요 **T2I 모델**이 생성한 이미지를 human preference 관점에서 평가한다.
- 다양한 **concept erasure 기법**을 T2I 생성 모델에 적용한 뒤, 생성된 이미지를 Image Reward Metric으로 비교·분석한다.

[1] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong, “Imagereward: Learning and evaluating human preferences for text-to-image generation,” arXiv preprint arXiv:2304.05977, 2023.

Background

Text-to-image(T2I) model이란?

Text 형태의 설명(i.e., prompt)을 입력받아, 그 내용과 의미적으로 일치하는 고품질, 고해상도의 이미지를 생성하는 모델

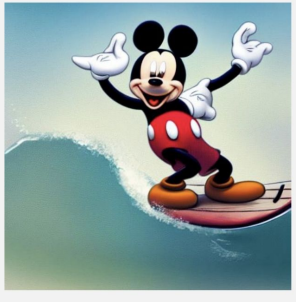


text-to-image 모델 예시

Background

Text-to-image model의 문제점

- 유해하거나 윤리적으로 문제가 있는 이미지를 생성할 수 있다.
  - 예를 들어, 적절한 제어 장치가 없는 경우 선정적인 신체 노출 (nudity) 이미지가 생성될 위험이 있다.
- 저작권 및 상표권이 포함된 이미지를 학습하거나 이를 생성할 경우, 법적·윤리적 문제가 발생할 수 있다.
  - 상표권을 침해하는 이미지를 생성한 예시

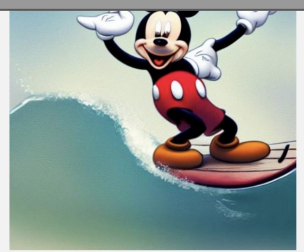



Background

Text-to-image model의 문제점

- 유해하거나 윤리적으로 문제가 있는 이미지를 생성할 수 있다.

**이러한 문제를 해결하기 위해 특정 개념(Target Concept)
선택적으로 제거하는 Concept Erasure 기법이 필수적임**

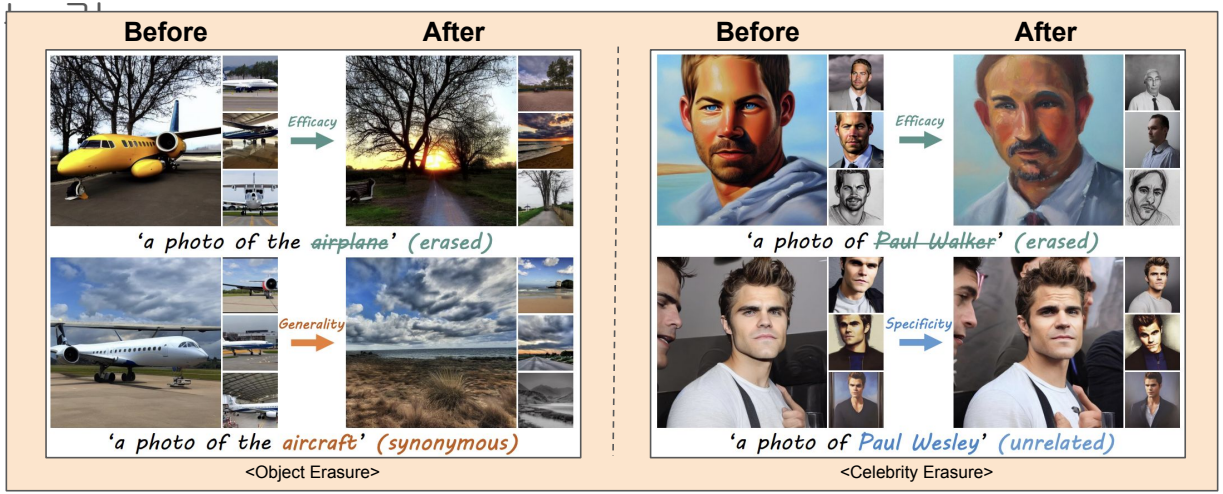
Background

Concept Erasure

: Pretrained된 T2I 모델에서 특정 개념(예: nudity, 인물 등)을 선택적으로 제거하는 과정

Concept Erasure의 목표

- Efficacy: 특정 개념을 효과적으로 제거하는 것
- Generality Preservation: 특정 개념만 제거하고 모델의 본래 생성 능력을 유지하는 것



<Example of Concept Erasure>

Background

Concept Erasure

: Pretrained된 T2I 모델에서 특정 개념(예: nudity, 인물 등)을 선택적으로 제거하는 기법

- 다음과 같은 concept erasure methods이 있음
  - **SLD**: 부정적인 프롬포트(negative prompts)를 사용하여 타겟 컨셉 생성 방지 [ref 1]
  - **AC**: 타겟 컨셉을 대체 컨셉으로 매핑하도록 fine-tuning하는 방식 [ref 2]
  - **ESD**: 타겟 컨셉의 생성을 억제하는 방향으로 모델을 역유도하도록 fine-tuning하는 방식[ref 3]
  - **UCE**: Closed-form solution을 사용하여 cross-attention 레이어를 직접 업데이트하는 방식 [ref 4]

[ref] 1 - 9 는 마지막 페이지에 기록해놓았습니다.

Background

Concept Erasure

: Pretrained된 T2I 모델에서 특정 개념(예: nudity, 인물 등)을 선택적으로 제거하는 기법

- 다음과 같은 concept erasure methods이 있음
  - **SA:** Continual learning 원리를 기반으로 설계된 unlearning 기법 [ref 5]
  - **RECELER**: 별도의 어댑터와 마스킹 기법을 활용하여 효율적으로 컨셉을 제거 [ref 6]
  - **MACE**: 입력 이미지에서 타겟 컨셉 영역을 식별하는 마스크를 활용하여 정밀한 unlearning 과정을 유도 [ref 7]

[ref] 1 - 9 는 마지막 페이지에 기록해놓았습니다.

Background

Concept Erasure

: Pretrained된 T2I 모델에서 특정 개념(예: nudity, 인물 등)을 선택적으로 제거하는 기법

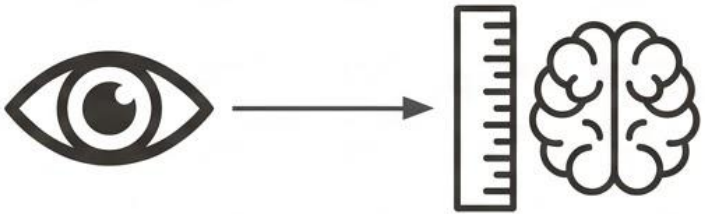
- 다음과 같은 concept erasure methods이 있음
  - **SAFREE**: 모델의 가중치를 수정하지 않고(training–free), 텍스트 임베딩과 시각적 latent space에서 유해한 개념을 실시간으로 필터링하여 안전한 생성을 유도하는 기법 [ref 8]
  - **ANT**: 디노이징 과정 중후반부의 궤적을 자동으로 조향하여, 이미지의 구조는 유지한 채, 원치 않는 개념만 정밀하게 회피하도록 모델을 fine-tuning하는 기법 [ref 9]

[ref] 1 - 9 는 마지막 페이지에 기록해놓았습니다.

Background

핵심 질문

모델이 생성한 이미지를 “인간의 관점"에서
어떻게 정량적으로 측정할 것인가?



Background

Image Reward

: 주어진 텍스트 프롬포트에 대해 생성된 이미지가 인간의 선호를 얼마나 잘 반영하는지를 정량적으로 평가하는 모델

- **출력**: 스칼라 값(scalar value)의 선호도 점수, 높을수록 선호도가 높음

Background

Image Reward

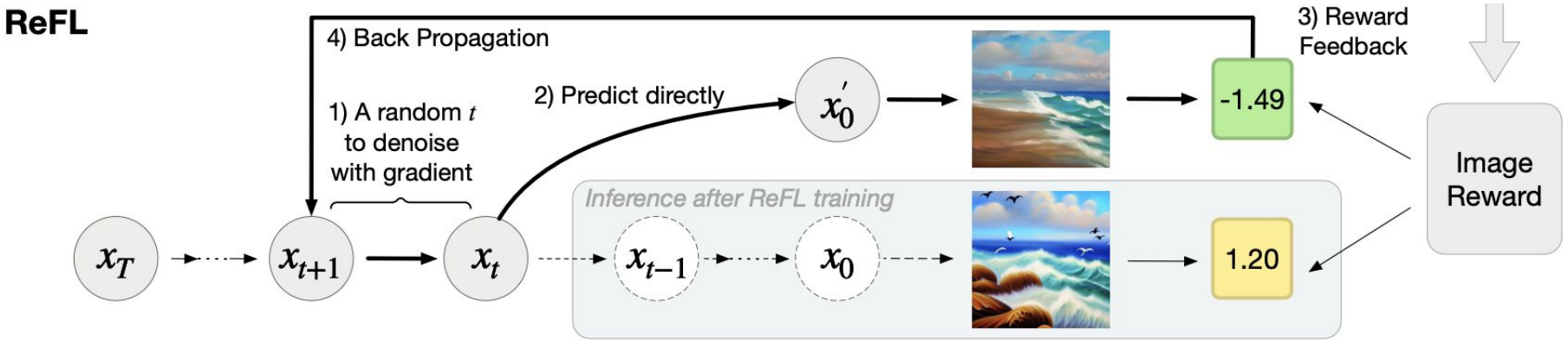
: 주어진 텍스트 프롬포트에 대해 생성된 이미지가 인간의 선호를 얼마나 잘 반영하는지를 정량적으로 평가하는 모델

- **목적**: T2I 모델을 통해 생성된 이미지의 품질을 인간의 시선으로 평가
  - 기존의 이미지 품질 평가 지표는 노이즈, 분포 차이 등을 기반으로 기계적인 품질만 평가할 뿐 사람이 실제로 어떤 이미지를 더 좋아하는지는 전혀 반영하지 못했다.
  - 이를 해결하기 위해, 대규모 Human Preference 데이터셋을 구축하고 강화학습 기반 Reward Model을 학습하여, 이미지의 Human Preference를 직접 반영하는 평가 지표를 제시하였다.

Background

ReFL: Reward Feedback Learning Improves Text-to-Image Diffusion

- Image reward를 보상으로 활용한 강화학습 알고리즘
- 생성 결과에 대한 likelihood를 계산할 수 없어 기존 RLHF 알고리즘 적용이 어려운 문제를 해결했다.



환경 및 데이터셋

환경

- GPU: 1x NVIDIA RTX 4090

Compared Methods

- Baseline Models for ImageReward: Openjourney, Stable Diffusion v1.4, v2.1, DALL-E 2, Versatile Diffusion, CogView 2
- Concept Erasure Methods: SLD, AC, ESD, UCE, SA, RECELER, MACE, SAFREE, ANT

데이터셋

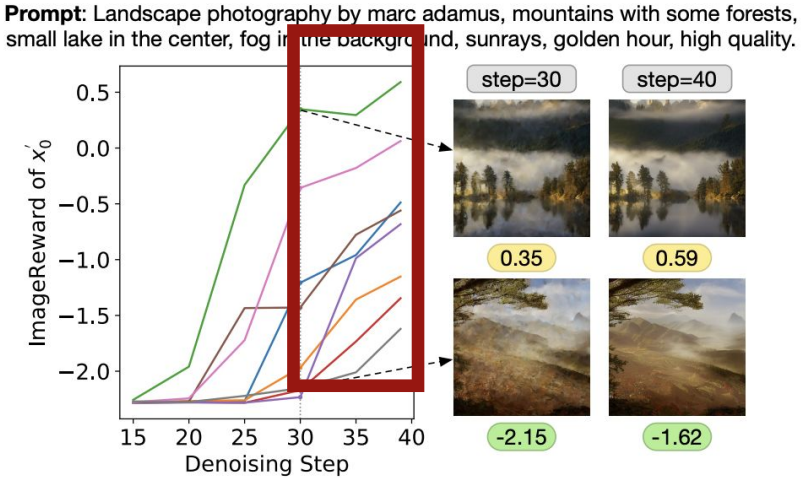
- MS-COCO 2014 Validation Set
- DiffusionDB

ReFL: State, Action, Reward 설계

- **State:** diffusion model의 생성 과정 중 특정 타임 스텝 t에서의 noisy latent vector x_t
- **Action:** T2I 모델(agent)이 주어진 프롬포트 T에 대해 이미지 x를 생성하는 행위
- **Reward:** (T, x) state에 대해 사전 학습된 Image Reward 모델이 부여하는 정량적 점수, 이 점수가 높을수록 해당 생성물이 인간의 선호도에 더 부합함을 의미한다.

ReFL: 강화학습 알고리즘

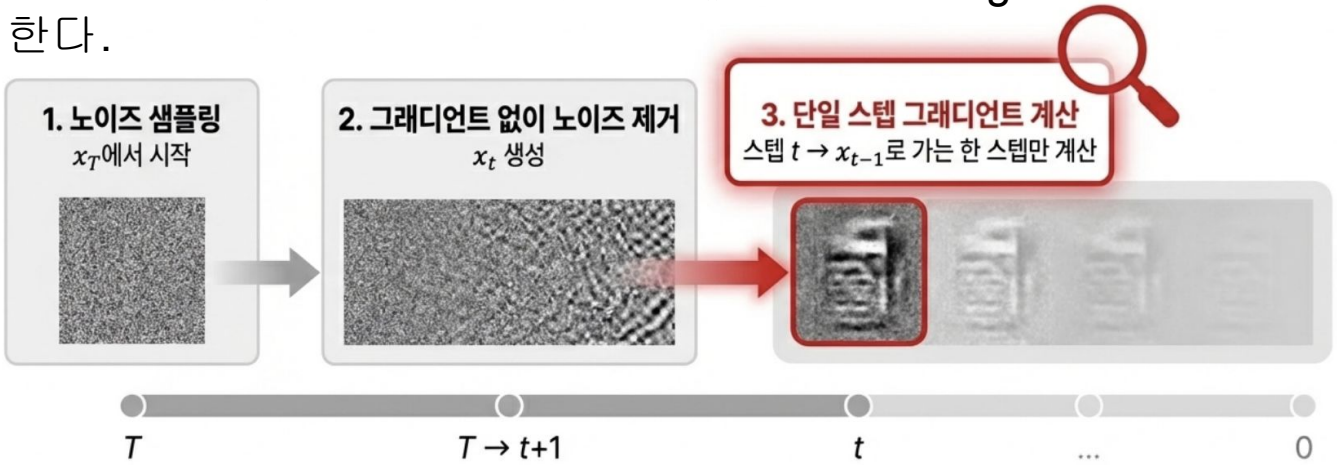
- Denoising 과정에서 다음과 같은 observation을 기반으로 피드백을 준다.
  - **Early-Step Stage**: 이미지가 아직 노이즈 상태가 품질을 판단하기 어렵다.
  - **Late-Step Stage:**이미지의 품질이 확실히 구분되며, 이 시점의 latent vector가 최종 이미지의 품질과 직결된다.



ReFL은 전체 과정을 다 학습시키는 대신, 생성 후반부에 집중하여 피드백을 준다.

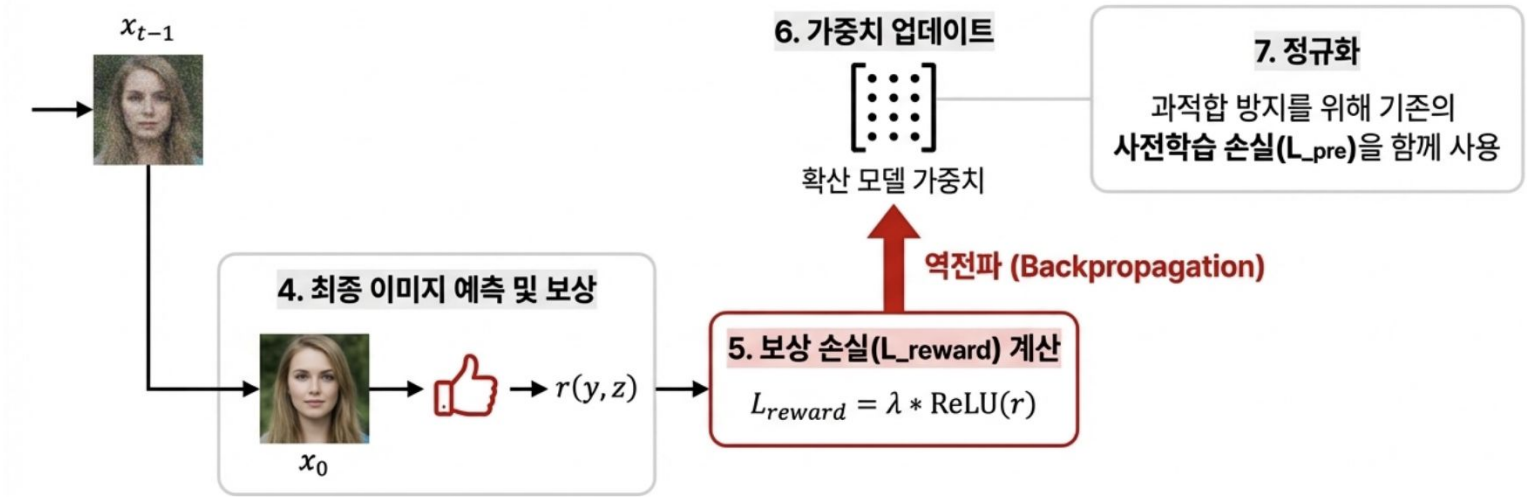
ReFL: 강화학습 알고리즘

- Diffusion 모델의 전체 denoising steps 중, 후반부의 임의 시점 t를 선택한다.
  - 본 논문에서는 30 단계 이상일 때 서로 다른 image reward 점수를 가진 생성물들이 일반적으로 구분 가능해지기 때문에 30단계에서 40단계 사이의 단계에서만 reward를 통해 학습한다.
- 시작점(T)부터 선택한 시점(t+1) 까지는 gradient 계산을 하지 않고 Text2Image 모델을 통과시킨다.
- 선택된 시점 **t에서 t-1 로 가는 딱 한단계**에 대해서만 gradient를 계산하며 노이즈를 제거한다.



ReFL: 강화학습 알고리즘

- 현재 상태(x_{t-1})에서 최종적으로 생성될 깨끗한 이미지(x_0)를 한 번에 예측한다.
- 예측된 이미지(z_i)와 prompt(y_i)를 ImageReward 모델에 넣어 점수를 받고 이를 손실값($\mathcal{L}_{reward}$)으로 변환한다.

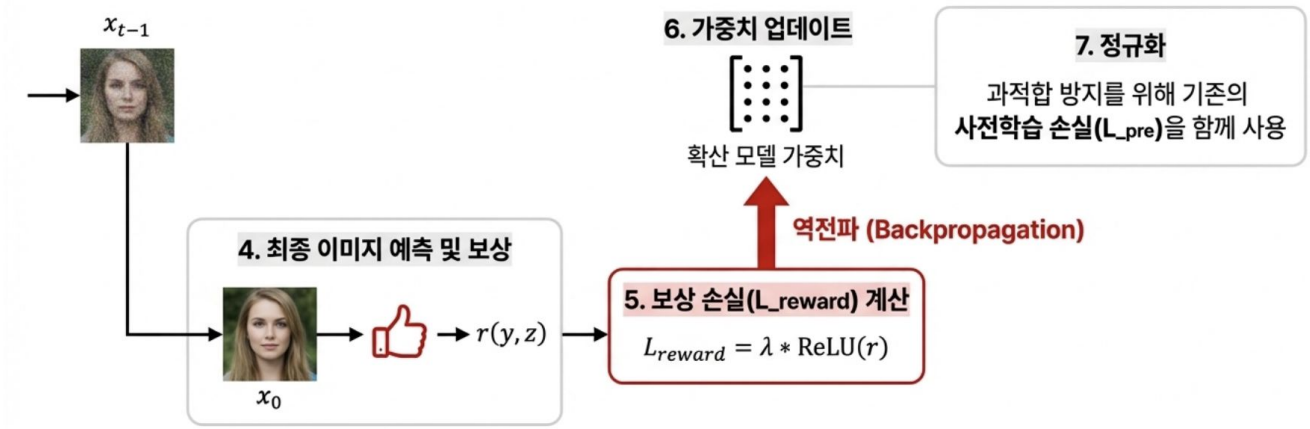


ReFL: 강화학습 알고리즘

- reward만 높이려고 하면 overfitting 문제가 발생할 수 있기 때문에, 이를 방지하기 위해 사전 학습 손실($\mathcal{L}_{pre}$)을 함께 사용한다.
- 각 loss function은 다음과 같다.

$$\mathcal{L}_{reward} = \lambda \mathbb{E}_{y_i \sim \mathcal{Y}}(\phi(r(y_i, g_\theta(y_i))))$$

$$\mathcal{L}_{pre} = \mathbb{E}_{(y_i,x_i)\sim\mathcal{D}}(\mathbb{E}_{\mathcal{E}(x_i),y_i,\epsilon\sim\mathcal{N}(0,1),t}[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y_i))\|_2^2])$$



ReFL: hyperparameter

- 주요 hyperparameter는 다음과 같다.

hyperparameter	설명	값
보상 가중치(λ)	보상 손실의 영향력을 조절	1e-3
전체 타임스텝(T)	노이즈 제거 스케쥴러의 전체 스텝 수	40
학습 타임스텝 범위($[T_1, T_2]$)	그래디언트를 계산할 스텝 t	종료 1-40 스텝 전
보상 매핑 함수	보상 점수를 손실 값으로 변환하는 함수	ReLU
guidance score	생성 이미지에 프롬포트 강도를 조절하는 값	1.5 ~ 5
생성 이미지의 해상도	생성되는 이미지의 resolution을 조절	256

실험 셋업

1. ImageReward 논문 실험 재현

- MS-COCO Validation Set과 DiffusionDB를 활용하여 6개의 T2I baseline 모델로 이미지를 생성한다.
- 생성된 이미지는 Pretrained ImageReward 모델을 사용해 **Human Preference 점수**를 평가한다.

2. Concept Erasure 기법 비교 실험

- Stable Diffusion v1.4 모델을 기반으로, 다양한 Concept Erasure 기법을 적용하여 특정 개념(nudity)을 제거한다.
- 그 후 MS-COCO 데이터셋을 활용해 이미지를 생성하고, 생성된 이미지에 대해 **ImageReward 점수**를 산출하여 기법 간 성능을 비교한다.

실험 결과

1-1. ImageReward 논문 실험 재현 - MS COCO Dataset

- MS-COCO Validation Set을 통해 측정한 ImageReward Score이다.
- 생성되는 이미지의 해상도를 바꿔가며 비교한 결과, 더 큰 해상도에서 높은 Score를 달성했다.

Model & Size	256 x 256 (No Fixed)	512 x 512 (No Fixed)	256 x 256 (Paper)
StableDiffusion v 1.4	-1.5598	0.2531	-0.0857
StableDiffusion v 2.1	-1.1548	0.4946	0.1553
Openjourney	-1.4600	0.2624	-0.0455

실험 결과

1-1. ImageReward 논문 실험 재현 - MS COCO dataset

- MS-COCO Validation Set을 통해 측정한 ImageReward Score이다.
- Guidance Scale 값에 따른 점수를 비교한 결과, 더 큰 Guidance 값에서 높은 Score를 달성했다.

Model & Guidance Scale	256 x 256 (Scale =1.5)	256 x 256 (Scale = 2)	256 x 256 (Scale = 3)	256 x 256 (Scale = 4)	256 x 256 (Scale = 5)
StableDiffusion v 1.4	-2.1816	-2.0689	-1.8729	-1.7352	-1.6454
StableDiffusion v 2.1	-2.0159	-1.8253	-1.5429	-1.3625	-1.2592
Openjourney	-2.1356	-1.9855	-1.7448	-1.6041	-1.5277

실험 결과

1-2. ImageReward 논문 실험 재현 - DiffusionDB dataset

- DiffusionDB dataset을 통해 측정한 ImageReward Score이다.

	Reproduce		Paper	
Model & Metric	ImageReward	CLIP Score	ImageRward	CLIP Score
StableDiffusion v 1.4	0.1343	0.2763	0.1344	0.2726
StableDiffusion v 2.1	0.2467	0.2683	0.2458	0.2683
Openjourney	0.2615	0.2726	0.2614	0.2726
DALL-E 2	0.2115	0.2684	0.2114	0.2684
Versatile Diffusion	-0.2471	0.2606	-0.2470	0.2606
Cogview 2	-1.2376	0.2044	-1.2376	0.2044

실험 결과

1-3. ImageReward 논문 실험 재현 - ImageReward dataset

- ImageReward 모델을 평가하기 위해 사용한 테스트 세트의 정확도 결과이다.

	Reproduce	Paper
Metric	Preference Accuracy (%)	Preference Accuracy (%)
CLIP Score	54.81	54.82
Aestheic Score	57.32	57.35
BLIP Score	57.84	57.76
ImageReward	**65.15**	**65.14**

실험 결과

2. Concept Erasure 방법론 결과 - MS COCO dataset

- Stable Diffusion 모델에 Concept Erasure 기법을 적용한 후 생성한 이미지들에 대한 ImageReward 및 CLIP Score 결과이다.

Methods	ImageReward Score
Original	0.172
SLD	0.107
AC	0.116
ESD	-0.284
UCE	0.190
SA	-0.781
RECELER	-0.066
MACE	-1.403
SAFREE	0.253
ANT	-0.508

토의 및 결론

하이퍼파라미터 변화에 따른 Text2Image 모델의 Image Reward 분석

- *해상도(resolution)*
  - Text2Image 모델이 생성한 이미지의 해상도가 높을수록 Image Reward가 높게 측정됨을 확인하였다.
  - 이를 통해 해상도가 높을 수록 세부적인 디테일과 선명도가 인간의 선호에 더 잘 맞음을 알 수 있다.

토의 및 결론

하이퍼파라미터 변화에 따른 Text2Image 모델의 Image Reward 분석

- *Guidance Scale*
  - Guidance scale 변화에 따른 추이를 분석한 결과, scale이 5일 때 가장 높은 image reward 점수를 기록하였다.
  - 일반적인 권장 설정이 scale 7~8임을 감안할 때, 다소 낮은 Scale에서 더 높은 인간 선호도가 나타난 점이 주목된다.

⇨ 기존 CLIP Score가 텍스트와의 정합성(Alignment)을 중시하여 높은 Guidance Scale을 선호하는 경향이 있는 반면, 인간의 선호도는 과도한 텍스트 강제성보다는 자연스러운 이미지 품질을 더 중요하게 평가할 가능성이 있을 것이다.

토의 및 결론

Image Reward를 통한 Test Set Accuracy 분석

- 기존의 Text–Image alignment metric(예: CLIP, Aesthetic, BLIP score)과 비교했을 때, ImageReward는 더 높은 accuracy를 보인다.

- 이는 ImageReward가 인간의 선호도를 보다 정밀하게 반영하는 평가 지표임을 의미한다.

토의 및 결론

Image Reward를 통한 Concept Erasure Methods 분석

- SAFREE와 UCE의 경우 Image Reward에서 높은 점수를 기록하였다.
- MACE, SA, ANT의 경우 negative scores를 보였다.

→ concept erasure가 과도하게 적용될 경우, 모델이 학습한 일반적인 이미지 품질이나 미적 요소까지 함께 망각하게 되는 trade-off 현상이 극명하게 드러난다는 것을 알 수 있다.

토의 및 결론

Limitations

- Image Reward의 학습 데이터셋인 DiffusionDB가 가진 내재적 결함 (데이터 깨짐 등)과 같은 한계로 인해, 해당 데이터셋을 활용한 Training은 진행하지 못하였다.
- 그러나 Image Reweard 모델을 평가 지표로 적극 활용하는 방향으로 진행하였다.

토의 및 결론

결론

- Image Reward를 핵심 metric으로 채택하여, Text2Image model의 하이퍼파리미터를 변경하며 성능을 '인간 선호도' 관점에서 재해석하였다.
- 단순히 개념이 지워졌는지만 확인하는 기존의 평가를 넘어서, 지워진 후의 이미지가 인간에게 어떻게 받아들여지는지를 Image Reward로 평가함으로써, 모델의 실용성을 보다 입체적으로 검증하였다.
- 향후 연구에서는 Image Reward를 고려하여 개념 삭제와 품질 유지를 동시에 학습할 수 있는 고도화된 method가 필요할 것이다.

Reference

[1]: Patrick Schramowski, Manuel Brack, Bj¨ orn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappro- priate degeneration in diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22522–22531, 2023. 2, 3, 6, 8, 24

[2]: Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In Proceedings of
the IEEE/CVF International Conference on Computer Vision, pages 22691–22702, 2023. 2, 6, 24

[3]: Rohit Gandikota, Joanna Materzynska, Jaden Fiotto Kaufman, and David Bau. Erasing concepts from diffusion models. In Proceedings of the IEEE/CVF International Con-
ference on Computer Vision, pages 2426–2436, 2023. 2, 3, 6, 24

[4]: Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzy´ nska, and David Bau. Unified concept editing in diffusion models. In Proceedings of the IEEE/CVF Win-
ter Conference on Applications of Computer Vision, pages 5111–5120, 2024. 2, 3, 4, 6, 25

[5]: Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. Advances in Neural Information Processing Systems, 36,
2024. 2, 3, 6, 24

[6]: Chi-Pin Huang, Kai-Po Chang, Chung-Ting Tsai, Yung Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Receler: Reliable concept erasing of text-to-image diffusion
models via lightweight erasers. European Conference on Computer Vision, 2024. 2, 3, 6, 25

[7]: Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6430– 6440, 2024. 2, 3, 5, 6, 25

[8]: Jaehong Yoon, Shoubin Yu, Vaidehi Patil, Huaxiu Yao, and Mohit Bansal. Safree: Training-free and adaptive guard for safe text-to-image and video generation. arXiv preprint arXiv:2410.12761, 2024

[9]: Li, L., Lu, S., Ren, Y., and Kong, A. W.-K. Set you straight: Auto-steering denoising trajectories to sidestep unwanted concepts. arXiv preprint arXiv:2504.12782, 2025.