Direct Preference Optimization with Diffusion Models

2024.5.31 DMQA Open Seminar

DMQA 장건희

Introduction

❖ **장건희 (Geonhui Jang)**

- 고려대학교 산업경영공학과 석사 과정 (2023.03 ~ Present)
- Data Mining & Quality Analytics Lab (김성범 교수님)

❖ **Research Interests**

- Generative Models
- Computer Vision, Diffusion Models
- LLM, Multimodal Models

❖ **Contact**

- csleivear1@korea.ac.kr

Preliminary

❖ **RLHF (Reinforcement Learning from Human Feedback)**

- **인간 선호도 데이터셋**을 바탕으로 모델을 튜닝하는 방식

❖ **DPO (Direct Preference Optimization)**

- RLHF 방식에서 **reward model 학습 없이 인간 선호도 데이터셋만을 이용해** LLM (Large Language Models)을 바로 튜닝하는 방식

※ LLM을 인간 선호도 (human preference) 데이터셋으로 학습하는 이유?

- 사용자의 입력에 대해 안전하고 유용한 (Helpful, Honest, Harmless) 답변을 주는 LLM을 만들기 위해
- 그 외에 특정 도메인의 전문 지식으로 LLM을 fine-tuning 하고자 하는 경우 등

L = (폭력성/선정성 없는 텍스트를 생성하도록 하는 손실함수)

1. 위 손실함수 L을 이용해 학습

VS.

(a) 🥰😘☺️😚🤗

(b) 🤬😡😈☠️💩

2. (a)를 생성하고 (b)를 생성하지 않도록 학습

→ 인간 선호 자체를 모델링할 수 있는 손실함수는 만들기 매우 어려움

따라서 인간이 선호하는 방향으로 output이 생성될 수 있도록 인간 선호도 데이터셋을 통해 모델을 학습

Preliminary

❖ **RLHF (Reinforcement Learning from Human Feedback)**

- **인간 선호도 데이터셋**을 바탕으로 모델을 튜닝하는 방식

❖ **DPO (Direct Preference Optimization)**

- RLHF 방식에서 **reward model 학습 없이 인간 선호도 데이터셋만을 이용해** LLM (Large Language Models)을 바로 튜닝하는 방식

❖ **Diffusion-DPO**

- 디퓨전 모델 fine-tuning 시 L_{simple} 대신 DPO loss를 이용한 새로운 loss로 튜닝

❖ **DCO (Direct Consistency Optimization)**

- Personalization 시 L_{simple} 대신 DPO loss를 이용한 새로운 loss로 튜닝

→ LLM 분야에서 인간이 선호하는 텍스트를 생성하도록 학습한 것과 같이
디퓨전 분야에서도 **인간이 선호하는 이미지를 생성**하도록 학습하고자 함

요약

2017.6

강화학습 기반으로 인간 선호를 학습하는 RLHF

↓

2023.5

RLHF에서 강화학습 부분을 제거한 DPO

↓

2023.11 & 2024.2

DPO를 디퓨전 모델에 적용한 Diffusion-DPO와 DCO

Preliminary

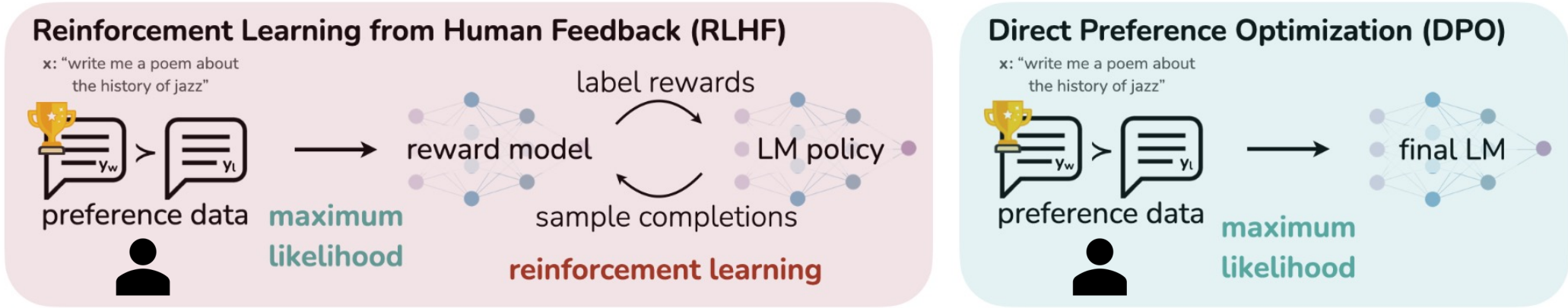
RLHF와 DPO

❖ RLHF (Reinforcement Learning from Human Feedback)

- **인간 선호도 데이터셋**을 바탕으로 모델을 튜닝하는 방식
- Supervised fine-tuning – Reward modeling – Fine-tuning with RM 세 단계로 구성
- ChatGPT, InstructGPT, LLaMA2 등 여러 LLM fine-tuning에 사용

❖ DPO (Direct Preference Optimization)

- RLHF 방식에서 **reward model 학습 없이 인간 선호도 데이터셋만을 이용해** LLM (Large Language Models)을 바로 튜닝하는 방식
- RLHF 목적식의 최적해를 LLM 목적 함수에 통합

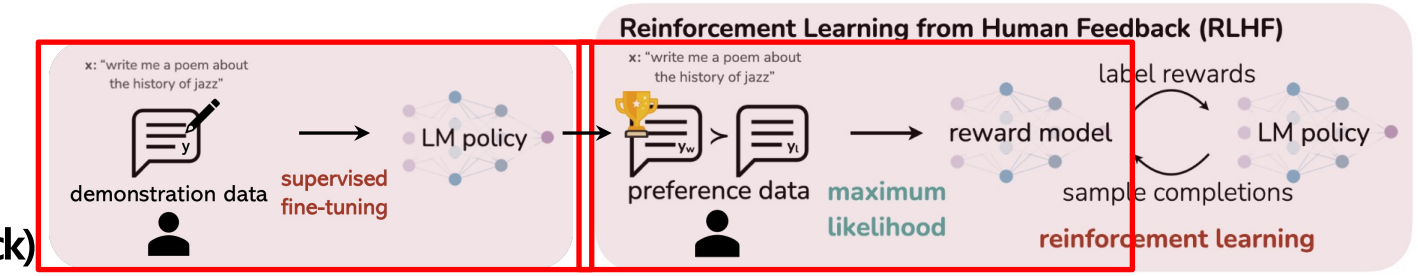


<RLFH와 DPO 비교>

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2024).
Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36.

Preliminary

RLHF와 DPO



❖ RLHF (Reinforcement Learning from Human Feedback)

- Supervised fine-tuning – Reward modeling – Fine-tuning with RM 세 단계로 구성

[Step 1] Supervised fine-tuning

- Human labeler가 Input x에 대해 desired output y를 작성 → 이 demonstration data로 **LLM fine-tuning**

[Step 2] Reward modeling

- Fine-tuned LLM이 Input x에 대해 두 개의 model output 생성 → human labeler가 이 두 output 중 더 선호하는 답변 선택
- Human preference dataset $\{(x, y_w, y_l)^i\}_{i=1}^N$으로 **reward model r_ϕ 학습** → 학습 후 reward model은 desirable output일수록 높은 점수 반환

$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log \sigma(r_\phi(x, y_w) - r_\phi(x, y_l))\right]$$

<Bradley-Terry Model (1952)>

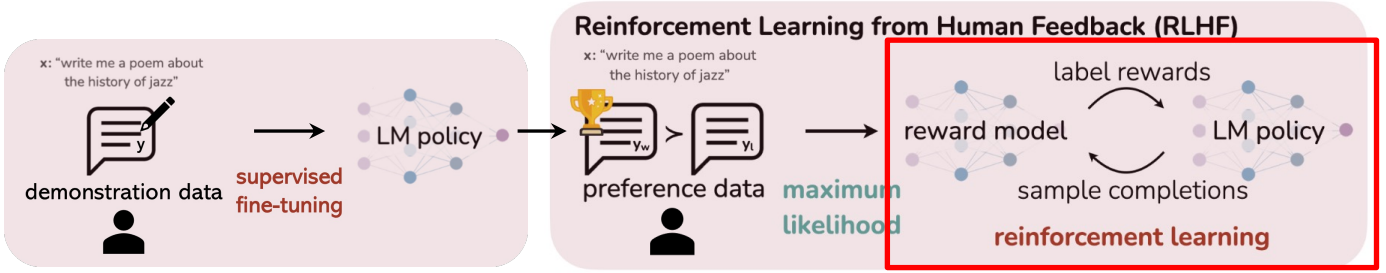
$$p^*(y_1 \succ y_2 \mid x) = \frac{\exp\left(r^*(x, y_1)\right)}{\exp\left(r^*(x, y_1)\right) + \exp\left(r^*(x, y_2)\right)}$$

→ Pair 데이터 사이 선호도에 대한 확률 분포를 정의 (선호 확률이 optimal reward와 비례)

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2024).
Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36.

Preliminary

RLHF와 DPO



❖ RLHF (Reinforcement Learning from Human Feedback)

- Supervised fine-tuning – Reward modeling – Fine-tuning with RM 세 단계로 구성

[Step 1] Supervised fine-tuning

- Human labeler가 Input x에 대해 desired output y를 작성 → 이 demonstration data로 **LLM fine-tuning**

[Step 2] Reward modeling

- Fine-tuned LLM이 Input x에 대해 두 개의 model output 생성 → human labeler가 이 두 output 중 더 선호하는 답변 선택
- Human preference dataset $\{(x, y_w, y_l)^i\}_{i=1}^N$으로 **reward model r_ϕ 학습** → 학습 후 reward model은 desirable output일수록 높은 점수 반환

$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log \sigma(r_\phi(x, y_w) - r_\phi(x, y_l))\right]$$

[Step 3] Fine-tuning with RM

- Reward가 최대화되도록 **LLM π_θ를 fine-tuning**함. 여기에 기존 모델 분포 π_{ref}와 너무 달라지지 않도록 KL divergence 제약 추가
- 일반적으로 강화학습 방식 중 하나인 PPO (Proximal Policy Optimization)을 통해 최적화

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(y|x)}\left[r_\phi(x, y)\right] - \beta\mathbb{D}_{\text{KL}}\left[\pi_\theta(y \mid x) \,||\, \pi_{\text{ref}}(y \mid x)\right]$$

<Bradley-Terry Model (1952)>

$$p^*(y_1 \succ y_2 \mid x) = \frac{\exp\left(r^*(x, y_1)\right)}{\exp\left(r^*(x, y_1)\right) + \exp\left(r^*(x, y_2)\right)}$$

→ Pair 데이터 사이 선호도에 대한 확률 분포를 정의 (선호 확률이 optimal reward와 비례)

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2024).
Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36.

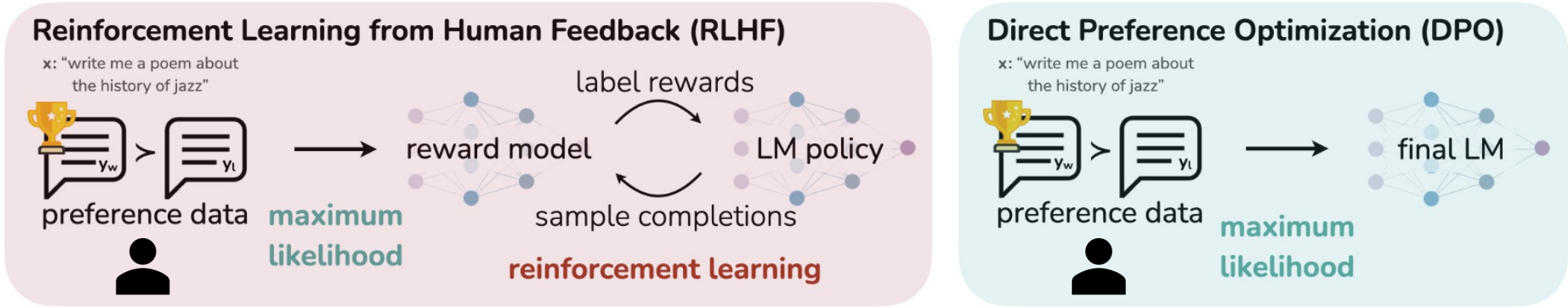
Preliminary

RLHF와 DPO

❖ DPO (Direct Preference Optimization)

- RLHF 방식에서 **reward model 학습 없이 인간 선호도 데이터셋만을 이용해** LLM을 바로 튜닝하는 방식
- **Why?** (1) Reward model, critic model 등 강화학습을 위해 추가 모델 학습 필요
  (2) LLM이 reward model 및 critic model과 상호작용할 때 학습 불안정성 + reward는 높지만 부자연스러운 문장을 생성하는 mode collapse
- **How?** → RLHF 목적식의 최적해를 LLM 목적 함수에 통합
- Reward model은 없어졌지만 reward 메커니즘을 LLM loss에 담음

  → "DPO를 통해 언어모델은 implicit하게 human preference를 이해하는 reward model로서 기능할 수 있음"



<RLFH와 DPO 비교>

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2024).
Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36.

Preliminary

RLHF와 DPO

- $\pi(y|x)$: Human preference data $\mathcal{D}$에 대해 fine-tuning되는 LLM
- $\pi_{\text{ref}}(y|x)$: RLHF Step 1에서 supervised fine-tuned된 LLM
- $r(x, y)$: Human preference data로 학습된 reward model

❖ DPO (Direct Preference Optimization)

- RLHF 목적식의 최적해를 구함

$$\max_{\pi} \mathbb{E}_{x\sim\mathcal{D},y\sim\pi}\left[r(x,y)\right] - \beta\mathbb{D}_{\text{KL}}\left[\pi(y|x) \,||\, \pi_{\text{ref}}(y|x)\right]$$

$$= \max_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi(y|x)}\left[r(x,y) - \beta\log\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}\right]$$

$$= \min_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi(y|x)}\left[\log\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} - \frac{1}{\beta}r(x,y)\right]$$

$$= \min_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi(y|x)}\left[\log\frac{\pi(y|x)}{\frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right)} - \log Z(x)\right]$$

$$= \min_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{D}_{\text{KL}}(\pi(y|x) \,||\, \pi^*(y|x)) - \log Z(x)\right]$$

$$\pi^*(y|x) = \frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right)$$

$$\therefore\ \pi(y|x) = \pi^*(y|x) = \frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right) \quad \textit{where} \quad Z(x) = \sum_y \pi_{\text{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right)$$

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2024).
Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36.

Preliminary

RLHF와 DPO

- $\pi(y|x)$: Human preference data $\mathcal{D}$에 대해 fine-tuning되는 LLM
- $\pi_{\text{ref}}(y|x)$: RLHF Step 1에서 supervised fine-tuned된 LLM
- $r(x, y)$: Human preference data로 학습된 reward model

❖ DPO (Direct Preference Optimization)

- Reward model formulation을 LLM output에 대한 수식으로 바꾸고 이를 이용해 목적식 재정의

$$\pi^*(y|x) = \frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right) \longrightarrow r(x,y) = \beta\log\frac{\pi^*(y|x)}{\pi_{\text{ref}}(y \mid x)} + \beta\log Z(x)$$

- **Reward model을 사용하지 않으면서** Bradely-Terry Model (선호도 확률분포)에 부합하는 LLM optimization 목적식을 만듦
- 최종적으로 human preference dataset $\mathcal{D}$와 reference LLM model π_{ref}만으로 모델을 학습할 수 있음

$$p^*(y_1 \succ y_2 \mid x)$$

Reward model output 수식을

$$= \frac{\exp\left(r^*(x, y_1)\right)}{\exp\left(r^*(x, y_1)\right) + \exp\left(r^*(x, y_2)\right)}$$

⇩ *LLM output 수식으로 대체*

$$= \frac{1}{1 + \exp\left(\beta\log\frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)} - \beta\log\frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)}\right)}$$

<선호도 확률분포 재정의>

Reward model output 수식을

$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma(r_\phi(x, y_w) - r_\phi(x, y_l))\right]$$

⇩ *LLM output 수식으로 대체*

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right]$$

<LLM 손실함수 재정의>

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2024).
Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36.

Preliminary

RLHF와 DPO

- $\pi_\theta(y|x)$: fine-tuning되는 LLM
- $\pi_{\text{ref}}(y|x)$: RLHF Step 1에서 supervised fine-tuned된 LLM
- $r(x, y)$: Human preference data로 학습된 reward model

❖ DPO (Direct Preference Optimization)

- Gradient analysis of DPO loss

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right]$$

$$\nabla_\theta\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) =$$

$$-\beta\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\underbrace{\sigma(\hat{r}_\theta(x, y_l) - \hat{r}_\theta(x, y_w))}_{\text{higher weight when reward estimate is wrong}}\Big[\underbrace{\nabla_\theta\log\pi(y_w \mid x)}_{\text{increase likelihood of } y_w} - \underbrace{\nabla_\theta\log\pi(y_l \mid x)}_{\text{decrease likelihood of } y_l}\Big]\Big]$$

where $\hat{r}_\theta(x, y) = \beta\log\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$

(=LLM으로 정의된 implicit reward estimate)

y_w(선호 응답)의 likelihood가 증가하도록 학습

y_l(선호하지 않는 응답)의 likelihood가 감소하도록 학습

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2024).
Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36.

Preliminary

RLHF와 DPO

- $\pi_\theta(y|x)$: fine-tuning되는 LLM
- $\pi_{\text{ref}}(y|x)$: RLHF Step 1에서 supervised fine-tuned된 LLM
- $r(x, y)$: Human preference data로 학습된 reward model

❖ DPO (Direct Preference Optimization)

- Gradient analysis of DPO loss

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right]$$

$$\nabla_\theta\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) =$$

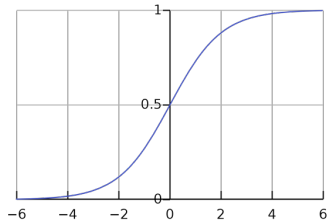
$$-\beta\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\underbrace{\sigma(\hat{r}_\theta(x, y_l) - \hat{r}_\theta(x, y_w))}_{\text{higher weight when reward estimate is wrong}}\Big[\underbrace{\nabla_\theta\log\pi(y_w \mid x)}_{\text{increase likelihood of } y_w} - \underbrace{\nabla_\theta\log\pi(y_l \mid x)}_{\text{decrease likelihood of } y_l}\Big]\Big]$$

where $\hat{r}_\theta(x, y) = \beta\log\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$

(=LLM으로 정의된 implicit reward estimate)

y_w(선호 응답)의 likelihood가 증가하도록 학습

y_l(선호하지 않는 응답)의 likelihood가 감소하도록 학습



<Sigmoid 함수>

y_l(선호하지 않는 응답)의 reward estimate가 클수록,

y_w(선호 응답)의 reward estimate가 작을수록

(=모델이 잘못된 reward estimate을 반환할수록) 큰 가중치 부여

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2024).
Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36.

Diffusion-DPO

DPO for Diffusion Models

❖ **디퓨전 모델에도 인간 선호도를 반영할 수 있는 DPO loss function을 적용하자**

- **LLM:** {입력 프롬프트 x, 선호 답변 y_w, 선호하지 않는 답변 y_l}
- **디퓨전 모델:** {입력 샘플링 프롬프트 x, 선호 이미지 y_w, 선호하지 않는 이미지 y_l}
- 단순히 이미지에 추가된 노이즈를 예측하는 것만 학습하지 않고, 인간이 선호하는 이미지를 더 잘 생성하도록 가이드를 주자

<Diffusion-DPO 모델로 생성한 이미지 (SDXL 1.0 fine-tuning)>

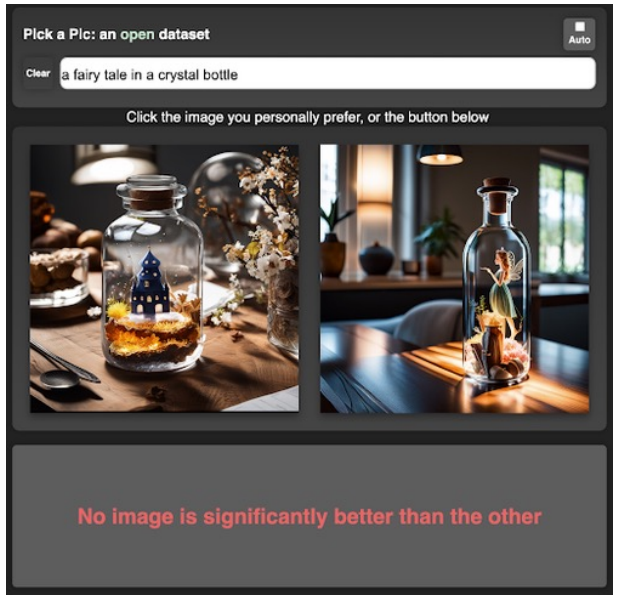
Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., ... & Naik, N. (2023). Diffusion model alignment using direct preference optimization. arXiv preprint arXiv:2311.12908.

Diffusion-DPO

DPO for Diffusion Models

❖ Dataset

- **Pick-a-Pic** [2023.5.2] 851k crowdsourced 데이터셋 사용. Pick-a-Pic은 웹 UI에서 사람들에게 이미지 쌍에 대해 선호 평가를 매기도록 하여 수집됨
- Synthesized image pair에 대한 (미적 수준, 프롬프트 반영 정도가 고려된) **인간 선호도** 정보 포함 (이미지는 SDXL-Beta, Dreamlike (fine-tuned SD 1.5)로부터 생성)
- Pick-a-Pic 논문에서는 데이터셋 제작과 더불어 이 데이터셋을 이용해 RLHF 방식으로 **선호도를 예측하는 reward model 'PickScore'**를 학습하기도 함



<웹 UI 예시. 유저는 이미지 생성모델이 만든 두 이미지 중 더 선호하는 이미지를 선택>

<데이터 구성 예시. (1) 텍스트 프롬프트 (2) 선호되지 않는 이미지 (3) 선호 이미지로 구성>

Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., & Levy, O. (2024). Pick-a-pic: An open dataset of user preferences for text-to-image generation. Advances in Neural Information Processing Systems, 36.

Diffusion-DPO

DPO for Diffusion Models

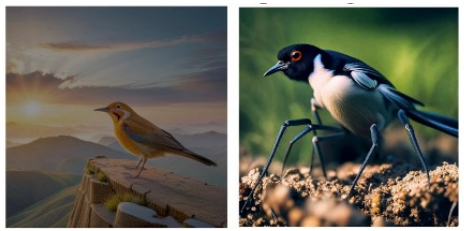
❖ Dataset

- **Pick-a-Pic** [2023.5.2] 851k crowdsourced 데이터셋 사용. Pick-a-Pic은 웹 UI에서 사람들에게 이미지 쌍에 대해 선호 평가를 매기도록 하여 수집됨
- Synthesized image pair에 대한 (미적 수준, 프롬프트 반영 정도가 고려된) **인간 선호도** 정보 포함 (이미지는 SDXL-Beta, Dreamlike (fine-tuned SD 1.5)로부터 생성)
- Pick-a-Pic 논문에서는 데이터셋 제작과 더불어 이 데이터셋을 이용해 RLHF 방식으로 **선호도를 예측하는 reward model 'PickScore'**를 학습하기도 함

"A rabbit"

기존: 이미지-캡션 pair에 대해 학습

"A bird with 8 spider legs"



Diffusion-DPO: 주어진 캡션과 이미지1, 이미지2에 대해 인간이 선호하는 이미지를 생성하도록 학습

Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., & Levy, O. (2024). Pick-a-pic: An open dataset of user preferences for text-to-image generation. Advances in Neural Information Processing Systems, 36.

Diffusion-DPO

DPO for Diffusion Models

❖ 손실함수 정의

기존 DPO loss: $$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right]$$

Diffusion-DPO loss: $$L_{\text{DPO-Diffusion}}(\theta) = -\mathbb{E}_{(\boldsymbol{x}_0^w,\boldsymbol{x}_0^l)\sim\mathcal{D}}\log\sigma\left(\beta\mathbb{E}_{\substack{\boldsymbol{x}_{1:T}^w\sim p_\theta(\boldsymbol{x}_{1:T}^w|\boldsymbol{x}_0^w)\\ \boldsymbol{x}_{1:T}^l\sim p_\theta(\boldsymbol{x}_{1:T}^l|\boldsymbol{x}_0^l)}}\left[\log\frac{p_\theta(\boldsymbol{x}_{0:T}^w)}{p_{\text{ref}}(\boldsymbol{x}_{0:T}^w)} - \log\frac{p_\theta(\boldsymbol{x}_{0:T}^l)}{p_{\text{ref}}(\boldsymbol{x}_{0:T}^l)}\right]\right)$$ (text condition c 생략)

∵ Since sampling from $p_\theta(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)$ is intractable, we utilize $q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)$

$$= -\log\sigma\left(\beta\mathbb{E}_{\boldsymbol{x}_{1:T}^w\sim q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0^w),\boldsymbol{x}_{1:T}^l\sim q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0^l)}T\mathbb{E}_t\left[\log\frac{p_\theta(\boldsymbol{x}_{t-1}^w|\boldsymbol{x}_t^w)}{p_{\text{ref}}(\boldsymbol{x}_{t-1}^w|\boldsymbol{x}_t^w)} - \log\frac{p_\theta(\boldsymbol{x}_{t-1}^l|\boldsymbol{x}_t^l)}{p_{\text{ref}}(\boldsymbol{x}_{t-1}^l|\boldsymbol{x}_t)}\right]\right)$$

∵ Jensen's inequality

$$\leq -\mathbb{E}_{t,\boldsymbol{x}_t^w\sim q(\boldsymbol{x}_t|\boldsymbol{x}_0^w),\boldsymbol{x}_t^l\sim q(\boldsymbol{x}_t|\boldsymbol{x}_0^l)}\log\sigma\left(\beta T\mathbb{E}_{\boldsymbol{x}_{t-1}^w\sim q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t^w,\boldsymbol{x}_0^w),\boldsymbol{x}_{t-1}^l\sim q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t^l,\boldsymbol{x}_0^l)}\left[\log\frac{p_\theta(\boldsymbol{x}_{t-1}^w|\boldsymbol{x}_t^w)}{p_{\text{ref}}(\boldsymbol{x}_{t-1}^w|\boldsymbol{x}_t^w)} - \log\frac{p_\theta(\boldsymbol{x}_{t-1}^l|\boldsymbol{x}_t^l)}{p_{\text{ref}}(\boldsymbol{x}_{t-1}^l|\boldsymbol{x}_t)}\right]\right)$$

$$= -\mathbb{E}_{t,\boldsymbol{x}_t^w\sim q(\boldsymbol{x}_t|\boldsymbol{x}_0^w),\boldsymbol{x}_t^l\sim q(\boldsymbol{x}_t|\boldsymbol{x}_0^l)}\log\sigma\Big(-\beta T\big(\mathbb{D}_{\text{KL}}(q(\boldsymbol{x}_{t-1}^w|\boldsymbol{x}_{0,t}^w)\|p_\theta(\boldsymbol{x}_{t-1}^w|\boldsymbol{x}_t^w)) - \mathbb{D}_{\text{KL}}(q(\boldsymbol{x}_{t-1}^w|\boldsymbol{x}_{0,t}^w)\|p_{\text{ref}}(\boldsymbol{x}_{t-1}^w|\boldsymbol{x}_t^w))$$
$$- \big(\mathbb{D}_{\text{KL}}(q(\boldsymbol{x}_{t-1}^l|\boldsymbol{x}_{0,t}^l)\|p_\theta(\boldsymbol{x}_{t-1}^l|\boldsymbol{x}_t^l)) + \mathbb{D}_{\text{KL}}(q(\boldsymbol{x}_{t-1}^l|\boldsymbol{x}_{0,t}^l)\|p_{\text{ref}}(\boldsymbol{x}_{t-1}^l|\boldsymbol{x}_t^l))\big)\Big)$$

Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., ... & Naik, N. (2023). Diffusion model alignment using direct preference optimization. arXiv preprint arXiv:2311.12908.

Diffusion-DPO

DPO for Diffusion Models

기존 디퓨전 모델 loss: $\mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}\left\|\boldsymbol{\epsilon}-\underbrace{\boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},t)}_{=\epsilon_\theta(x_t,t)}\right\|^2\right]$

❖ 손실함수 정의

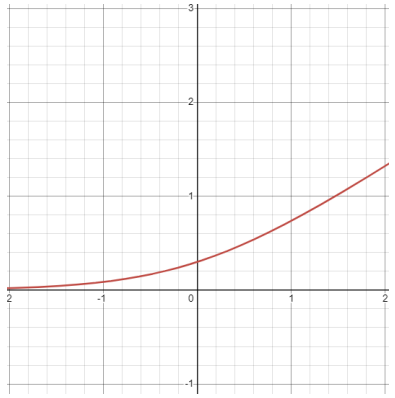
$$L_{\text{DPO-Diffusion}}(\theta) \leq -\mathbb{E}_{t,\boldsymbol{x}_t^w\sim q(\boldsymbol{x}_t|\boldsymbol{x}_0^w),\boldsymbol{x}_t^l\sim q(\boldsymbol{x}_t|\boldsymbol{x}_0^l)}\log\sigma\Big(-\beta T\big(\mathbb{D}_{\text{KL}}(q(\boldsymbol{x}_{t-1}^w|\boldsymbol{x}_{0,t}^w)\|p_\theta(\boldsymbol{x}_{t-1}^w|\boldsymbol{x}_t^w))-\mathbb{D}_{\text{KL}}(q(\boldsymbol{x}_{t-1}^w|\boldsymbol{x}_{0,t}^w)\|p_{\text{ref}}(\boldsymbol{x}_{t-1}^w|\boldsymbol{x}_t^w))$$

$$-\big(\mathbb{D}_{\text{KL}}(q(\boldsymbol{x}_{t-1}^l|\boldsymbol{x}_{0,t}^l)\|p_\theta(\boldsymbol{x}_{t-1}^l|\boldsymbol{x}_t^l))+\mathbb{D}_{\text{KL}}(q(\boldsymbol{x}_{t-1}^l|\boldsymbol{x}_{0,t}^l)\|p_{\text{ref}}(\boldsymbol{x}_{t-1}^l|\boldsymbol{x}_t^l))\big)\Big)$$

∵ 두 정규분포 사이의 KL divergence 정의

$$=-\mathbb{E}_{t,\boldsymbol{\epsilon}^w,\boldsymbol{\epsilon}^l}\log\sigma\left(-\beta T\omega(\lambda_t)\right(\underbrace{\|\boldsymbol{\epsilon}^w-\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^w,t)\|^2}_{(1)}-\underbrace{\|\boldsymbol{\epsilon}^w-\boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^w,t)\|^2}_{(2)}-\underbrace{\left(\|\boldsymbol{\epsilon}^l-\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^l,t)\|^2-\|\boldsymbol{\epsilon}^l-\boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^l,t)\|^2\right)}_{(3)\quad(4)}\left.\right)\right)$$

- 이 loss function은 어떤 의미를 가질까?



$y=-\log\sigma(-x)$

: 작을수록 좋음 →

선호 이미지에 대해

(1) fine-tuned model이 예측한 noise의 정확도가

(2) pre-trained model이 예측한 noise의 정확도보다 높을수록 좋음

: 클수록 좋음 →

선호하지 않는 이미지에 대해

(3) fine-tuned model이 예측한 noise의 정확도가

(4) pre-trained model이 예측한 noise의 정확도보다 낮을수록 좋음

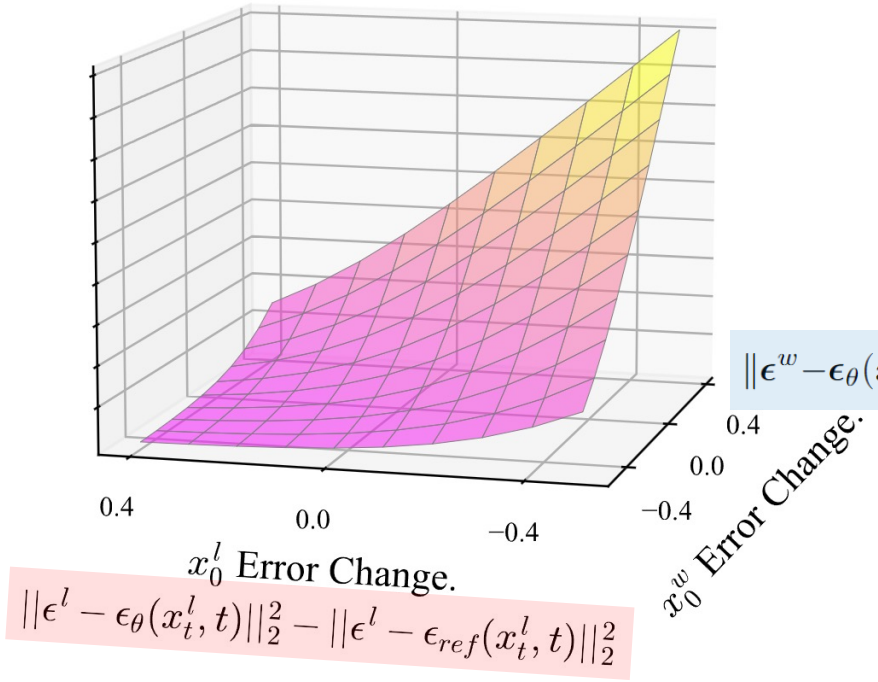
Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., ... & Naik, N. (2023). Diffusion model alignment using direct preference optimization. arXiv preprint arXiv:2311.12908.

Diffusion-DPO

DPO for Diffusion Models

❖ 손실함수 정의

$$L_{\text{DPO-Diffusion}}(\theta) \leq -\mathbb{E}_{t,\boldsymbol{\epsilon}^w,\boldsymbol{\epsilon}^l} \log \sigma\left(-\beta T \omega(\lambda_t)\left(\|\boldsymbol{\epsilon}^w - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^w, t)\|^2 - \|\boldsymbol{\epsilon}^w - \boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^w, t)\|^2 - \left(\|\boldsymbol{\epsilon}^l - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^l, t)\|^2 - \|\boldsymbol{\epsilon}^l - \boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^l, t)\|^2\right)\right)\right)$$



<Loss surface visualization>

- x_0^w를 잘 denoising할수록 loss 감소
- x_0^l를 denoising 못할수록 loss 감소

Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., ... & Naik, N. (2023). Diffusion model alignment using direct preference optimization. arXiv preprint arXiv:2311.12908.

Diffusion-DPO

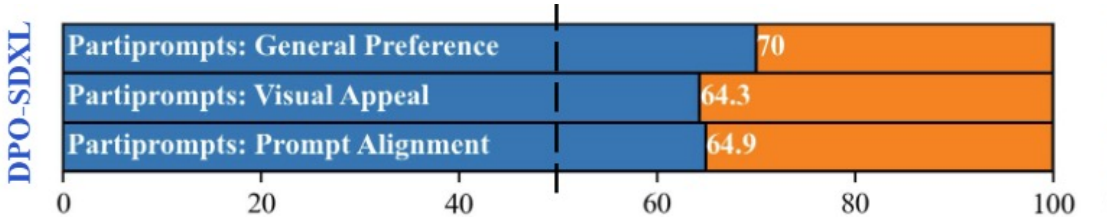
Experiments

<Pick-a-Pic 데이터셋>

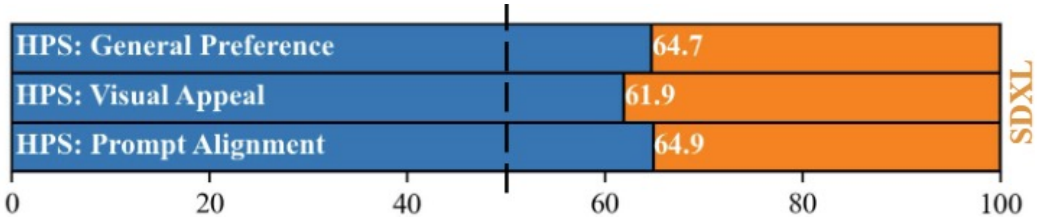
❖ 실험 및 평가

- Pick-a-Pic 851k 데이터셋을 이용해 **Stable Diffusion(SD) 1.5**와 **Stable Diffusion XL (SDXL) 1.0** 모델 fine-tuning
- 테스트 시 벤치마크 데이터셋 Partiprompts, HPSv2의 캡션 1632개/3200개로 평가 → Diffusion-DPO-tuned 모델과 baseline 모델 비교
- 모델이 만든 이미지에 대해 human evaluator가 선호도 평가. 이때 다음 기준 고려

1. 일반적 선호 – 주어진 프롬프트에 대해 어떤 이미지를 더 선호하는지
2. 시각적 선호도 – 프롬프트는 고려하지 않고 어떤 이미지가 더 시각적으로 매력적인지
3. 프롬프트 정렬 – 어떤 이미지가 텍스트 프롬프트와 더 잘 맞는지



<Partiprompts 캡션 이용>



<HPSv2 캡션 이용>

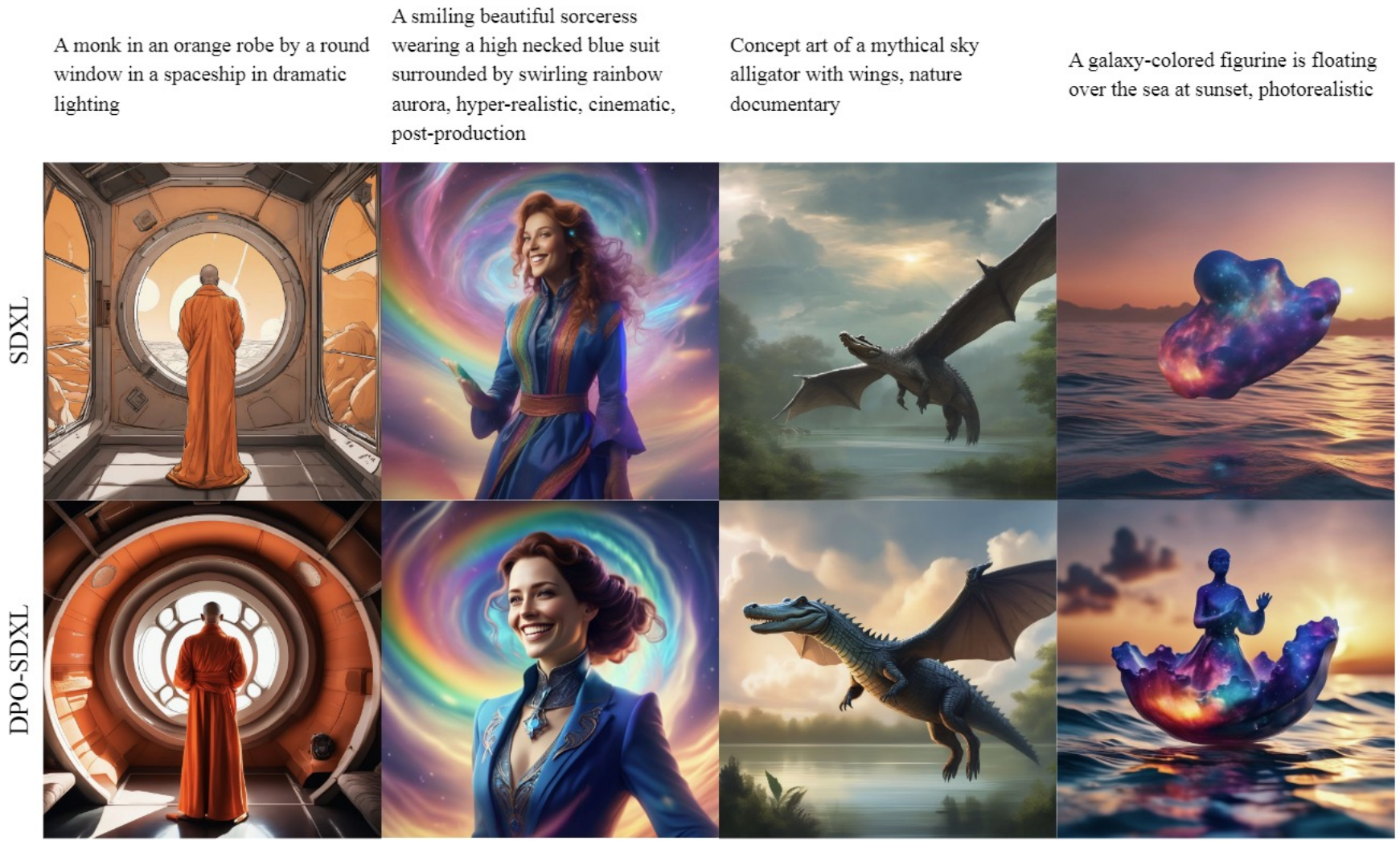
- 모든 선호도 평가지표에서 baseline 모델보다 Diffusion-DPO-tuned 모델이 더 높은 선호를 보임

Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., ... & Naik, N. (2023). Diffusion model alignment using direct preference optimization. arXiv preprint arXiv:2311.12908.

Diffusion-DPO

Experiments

❖ 실험 및 평가



- 샘플링 프롬프트를 더 잘 반영
- Human aesthetic preference에 더 align
- High contrast, vivid colors, fine detail, …

Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., ... & Naik, N. (2023). Diffusion model alignment using direct preference optimization. arXiv preprint arXiv:2311.12908.

Diffusion-DPO

Experiments

❖ 실험 및 평가

- DPO에서는 $\beta = 0.1$ 사용, Diffusion-DPO는 $\beta = 2 \sim 5$ 사용

  → β가 너무 작으면 디퓨전 모델이 pure reward scoring model로 악화, β가 너무 크면 KL divergence 제약이 강해져 학습 안됨
- RLHF 및 DPO의 첫 번째 step이었던 supervised fine-tuning은 크게 효과가 없었음 (SD는 기존 모델에 대해 win rate 5.5%p 증가, SDXL은 오히려 감소)

- 대략적인 스케치 정보를 주는 SDEdit 방식으로 이미지 생성
- DPO-SDXL이 더 visually appealing

Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., ... & Naik, N. (2023). Diffusion model alignment using direct preference optimization. arXiv preprint arXiv:2311.12908.

Direct Consistency Optimization (DCO)

DPO for Customization

❖ DPO loss를 Customization에 이용하자

- Diffusion-DPO는 디퓨전 모델이 인간의 선호도를 반영하는 이미지를 생성하도록 학습
- 이번엔 인간의 선호도가 아닌 consistency 유지를 목적으로 DPO loss를 이용하여 디퓨전 모델을 학습시켜보자
- 캡션 c에 대해, reference data $\mathcal{D}$와 모델이 생성한 이미지 x 사이의 consistency를 reward로 정의 → $r(\mathbf{x}, \mathbf{c}; \mathcal{D}_{\text{ref}}) := r(\mathbf{x}, \mathbf{c})$

<Personalization>

<Stylization>

Lee, K., Kwak, S., Sohn, K., & Shin, J. (2024). Direct Consistency Optimization for Compositional Text-to-Image Personalization. arXiv preprint arXiv:2402.12004.

Direct Consistency Optimization (DCO)

DPO for Customization

❖ 손실함수 정의

DPO에서 정의한 reward function: $$r(x, y) = \beta \log \frac{\pi_r(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x)$$

DCO에서 정의한 reward function: $$\hat{r}(\mathbf{x}_{0:T}, \mathbf{c}) = \beta \log \frac{p_\theta(\mathbf{x}_{0:T}|\mathbf{c})}{p_\phi(\mathbf{x}_{0:T}|\mathbf{c})} + \beta \log Z(\mathbf{c})$$ → 모든 marginal trajectories를 포함

marginal의 expectation

$$r(\mathbf{x}_0, \mathbf{c}) = \mathbb{E}_{\mathbf{x}_{1:T} \sim p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0, \mathbf{c})} \left[\beta \log \frac{p_\theta(\mathbf{x}_{0:T}|\mathbf{c})}{p_\phi(\mathbf{x}_{0:T}|\mathbf{c})} \right] + \beta \log Z(\mathbf{c})$$

Lee, K., Kwak, S., Sohn, K., & Shin, J. (2024). Direct Consistency Optimization for Compositional Text-to-Image Personalization. arXiv preprint arXiv:2402.12004.

Direct Consistency Optimization (DCO)

DPO for Customization

❖ 손실함수 정의

$$r(\mathbf{x}_0, \mathbf{c}) = \mathbb{E}_{\mathbf{x}_{1:T}\sim p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0,\mathbf{c})}\left[\beta \log \frac{p_\theta(\mathbf{x}_{0:T}|\mathbf{c})}{p_\phi(\mathbf{x}_{0:T}|\mathbf{c})}\right] + \beta \log Z(\mathbf{c})$$

$$r(\mathbf{x}_0, \mathbf{c})/\beta \approx \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\log \frac{p_\theta(\mathbf{x}_{0:T}|\mathbf{c})}{p_\phi(\mathbf{x}_{0:T}|\mathbf{c})}\right]$$

$$= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\sum_{t=1}^{T} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c})}{p_\phi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c})}\right]$$

$$= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\sum_{t=1}^{T} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c})}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} - \sum_{t=1}^{T} \log \frac{p_\phi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c})}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)}\right]$$

$$= \sum_{t=1}^{T} \mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)}\big[- D_{\mathrm{KL}}(q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)\|p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c})) + D_{\mathrm{KL}}(q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)\|p_\phi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c}))\big]$$

$$= \mathbb{E}_{\boldsymbol{\varepsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t\sim\mathcal{U}[1,T]}\big[- T\omega(t)\big(\|\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t; \mathbf{c}, t) - \boldsymbol{\varepsilon}\|_2^2 - \|\boldsymbol{\varepsilon}_\phi(\mathbf{x}_t; \mathbf{c}, t) - \boldsymbol{\varepsilon}\|_2^2\big)\big],$$

- Diffusion-DPO와 유사하게
  intractable p_θ 대신 q 이용 + Jensen's inequality 이용

$$\mathcal{L}_{\mathrm{DCO}}(\theta) := \mathbb{E}_{(\mathbf{x},\mathbf{c})\sim\mathcal{D}_{\mathrm{ref}},t\sim\mathcal{U}(0,T),\boldsymbol{\varepsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\big[- \log\sigma\big(- \beta T\omega(t)(\|\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t; \mathbf{c}, t) - \boldsymbol{\varepsilon}\|_2^2 - \|\boldsymbol{\varepsilon}_\phi(\mathbf{x}_t; \mathbf{c}, t) - \boldsymbol{\varepsilon}\|_2^2)\big)\big]$$

Lee, K., Kwak, S., Sohn, K., & Shin, J. (2024). Direct Consistency Optimization for Compositional Text-to-Image Personalization. arXiv preprint arXiv:2402.12004.

Direct Consistency Optimization (DCO)

DPO for Customization

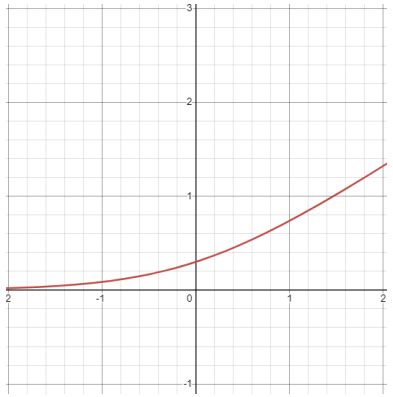
❖ 손실함수 정의

$$\mathcal{L}_{\text{DCO}}(\theta) := \mathbb{E}_{(\mathbf{x},\mathbf{c})\sim\mathcal{D}_{\text{ref}},t\sim\mathcal{U}(0,T),\boldsymbol{\varepsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\Big[-\log\sigma\Big(-\beta T\omega(t)\big(\|\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t;\mathbf{c},t)-\boldsymbol{\varepsilon}\|_2^2-\|\boldsymbol{\varepsilon}_\phi(\mathbf{x}_t;\mathbf{c},t)-\boldsymbol{\varepsilon}\|_2^2\big)\Big)\Big]$$

모든 reference data x, 캡션 c,
timestep t, Gaussian noise ε에 대해

(1) (2)

- 이 loss function은 어떤 의미를 가질까?



$$y = -\log\sigma(-x)$$

: 작을수록 좋음 →

선호 이미지에 대해

(1) fine-tuned model이 예측한 noise의 정확도가

(2) pre-trained model이 예측한 noise의 정확도보다 높을수록 좋음

Lee, K., Kwak, S., Sohn, K., & Shin, J. (2024). Direct Consistency Optimization for Compositional Text-to-Image Personalization. arXiv preprint arXiv:2402.12004.

Direct Consistency Optimization (DCO)

DPO for Customization

❖ 손실함수 정의

$$\mathcal{L}_{\text{DCO}}(\theta) := \mathbb{E}_{(\mathbf{x},\mathbf{c})\sim\mathcal{D}_{\text{ref}},t\sim\mathcal{U}(0,T),\boldsymbol{\varepsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\big[-\log\sigma\big(-\beta T\omega(t)(\|\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t;\mathbf{c},t)-\boldsymbol{\varepsilon}\|_2^2-\|\boldsymbol{\varepsilon}_\phi(\mathbf{x}_t;\mathbf{c},t)-\boldsymbol{\varepsilon}\|_2^2)\big)\big]$$

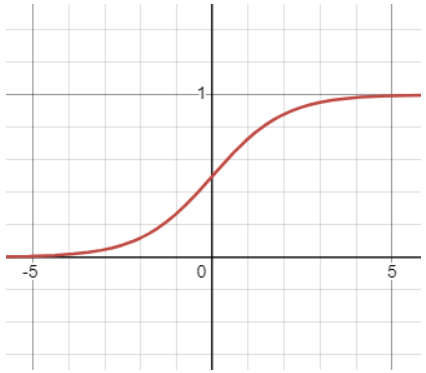
모든 reference data x, 캡션 c,
timestep t, Gaussian noise ε에 대해

- Gradient analysis of DCO loss

기존 디퓨전 loss와 동일

$$\nabla_\theta\mathcal{L}_{\text{DCO}}(\theta)\propto(1-\sigma(d_t))\,\nabla_\theta\|\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t;\mathbf{c},t)-\boldsymbol{\varepsilon}\|_2^2 \quad \text{where } d_t=-\beta T(\|\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t,\mathbf{c};t)-\boldsymbol{\varepsilon}\|_2^2-\|\boldsymbol{\varepsilon}_\phi(\mathbf{x}_t;\mathbf{c},t)-\boldsymbol{\varepsilon}\|_2^2)$$

fine-tuned model이 노이즈를 잘 예측하도록 학습



$y = 1-\sigma(-x)$

fine-tuned model이 예측한 노이즈의 오차가 클수록,
Pre-trained model이 예측한 노이즈의 오차가 작을수록
(=모델이 노이즈를 잘못 예측할수록) 큰 가중치 부여

Lee, K., Kwak, S., Sohn, K., & Shin, J. (2024). Direct Consistency Optimization for Compositional Text-to-Image Personalization. arXiv preprint arXiv:2402.12004.

Direct Consistency Optimization (DCO)

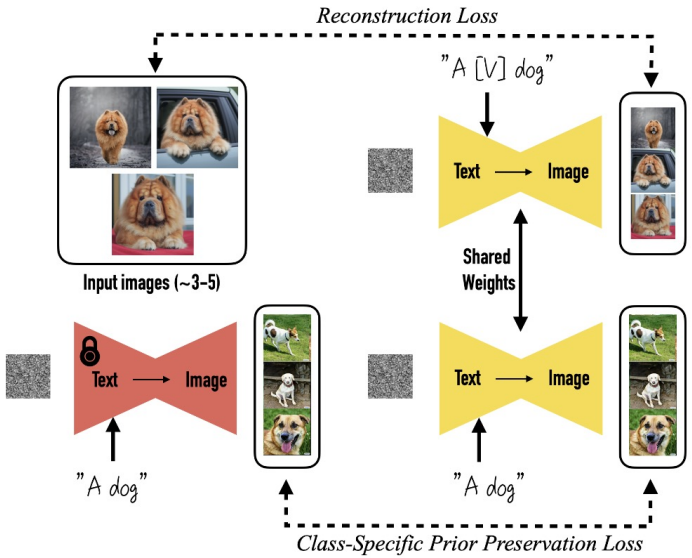
DPO for Customization

❖ 손실함수 정의

기존 디퓨전 loss와 동일

$$\nabla_\theta \mathcal{L}_{\text{DCO}}(\theta) \propto (1-\sigma(d_t))\,\nabla_\theta \|\varepsilon_\theta(\mathbf{x}_t;\mathbf{c},t)-\varepsilon\|_2^2 \quad \text{where } d_t = -\beta T\left(\|\varepsilon_\theta(\mathbf{x}_t,\mathbf{c};t)-\varepsilon\|_2^2 - \|\varepsilon_\phi(\mathbf{x}_t;\mathbf{c},t)-\varepsilon\|_2^2\right)$$

- DCO loss의 장점



<DreamBooth>

- 기존 personalization 방법론인 DreamBooth는 prior preservation을 위해 추가 데이터셋을 구축하고 학습해야 했음
- 그러나 DCO는 기존 pre-trained 모델과 너무 멀어지지 않도록 하는 DPO의 KL divergence 개념을 포함하고 있으므로 prior preservation이 불필요

Lee, K., Kwak, S., Sohn, K., & Shin, J. (2024). Direct Consistency Optimization for Compositional Text-to-Image Personalization. arXiv preprint arXiv:2402.12004.

Direct Consistency Optimization (DCO)

DPO for Customization

❖ Reward Guidance

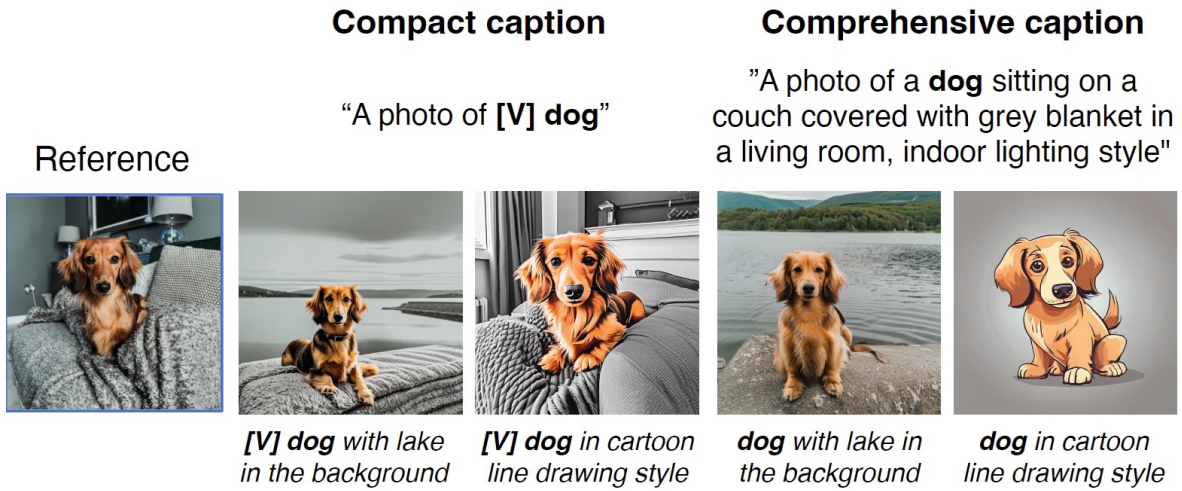
$$\hat{\varepsilon}(\mathbf{x}_t; \mathbf{c}, t) = \omega_{\text{rg}} \left(\varepsilon_\theta(\mathbf{x}_t; \mathbf{c}, t) - \varepsilon_\phi(\mathbf{x}_t; \mathbf{c}, t)\right) + \omega_{\text{text}} \left(\varepsilon_\phi(\mathbf{x}_t; \mathbf{c}, t) - \varepsilon_\phi(\mathbf{x}_t, t)\right) + \varepsilon_\phi(\mathbf{x}_t, t)$$

ε_θ: fine-tuned 모델

ε_ϕ: pre-trained 모델

- 기존에는 fine-tuned 모델로만 노이즈를 계산. Reward guidance를 통해 pre-trained 모델과 fine-tuned 모델의 output을 분리하여 조절할 수 있도록 함
- ω_{text}는 일반적인 CFG (classifier free guidance) 값, ω_{rg}는 fine-tuned 모델의 output 힘을 조절하는 scaler (논문에선 2.0, 3.0, 4.0, 5.0 사용)

❖ Comprehensive caption



- 단순한 캡션으로 학습하는 것보다 이미지를 최대한 잘 설명하는 캡션으로 학습하면 과적합을 방지하고 목표 대상을 다른 대상들과 분리하여 더 잘 표현할 수 있음

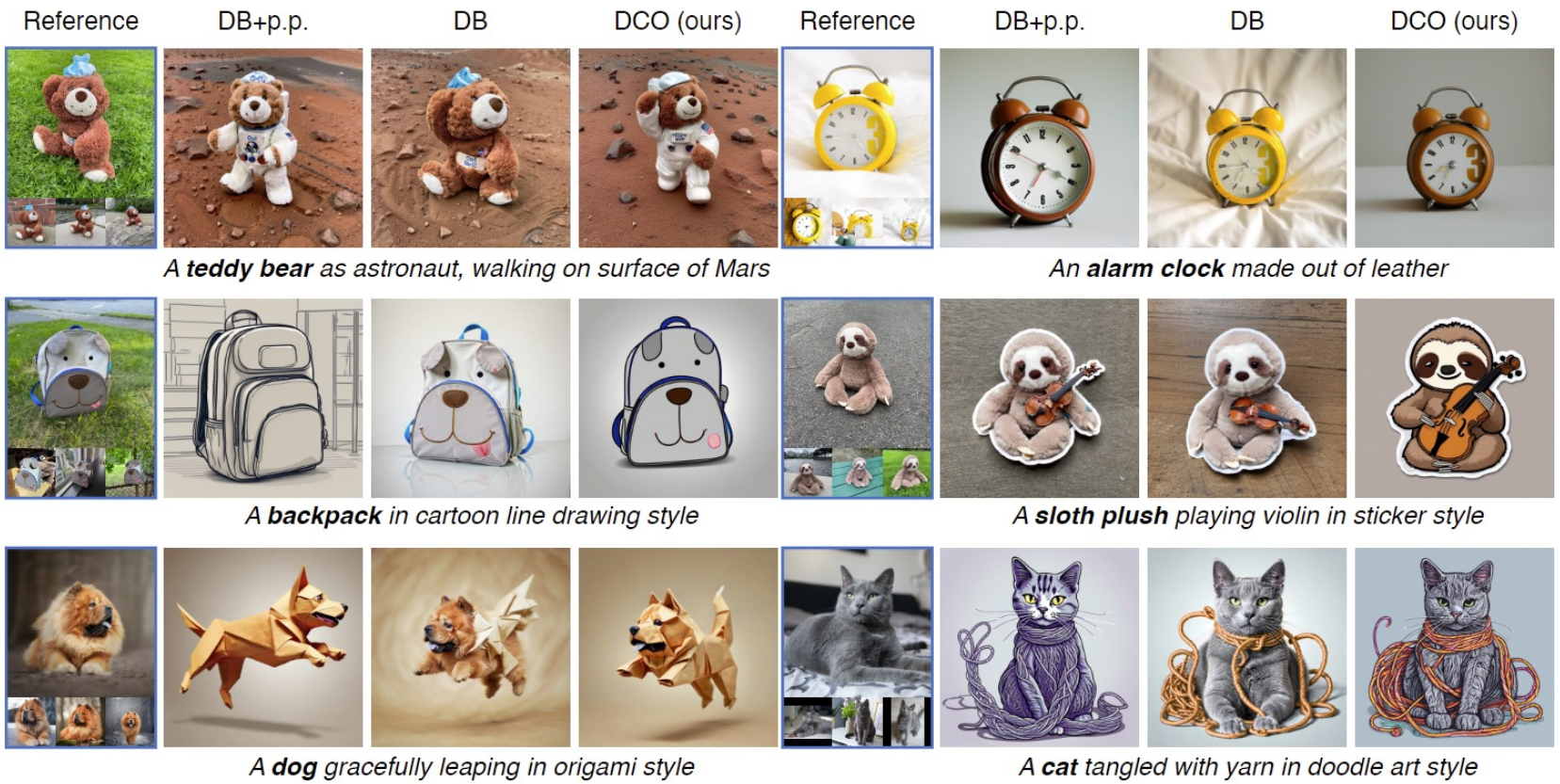
Lee, K., Kwak, S., Sohn, K., & Shin, J. (2024). Direct Consistency Optimization for Compositional Text-to-Image Personalization. arXiv preprint arXiv:2402.12004.

Direct Consistency Optimization (DCO)

DPO for Customization

❖ 실험 및 평가

- Stable Diffusion XL (SDXL) 1.0 + LoRA fine-tuning



<personalization 비교>

DB: DreamBooth

DB+p.p.: DreamBooth + prior preservation

- 기존 방법론보다 대상의 identity를 잘 유지함+샘플링 프롬프트에 더 충실

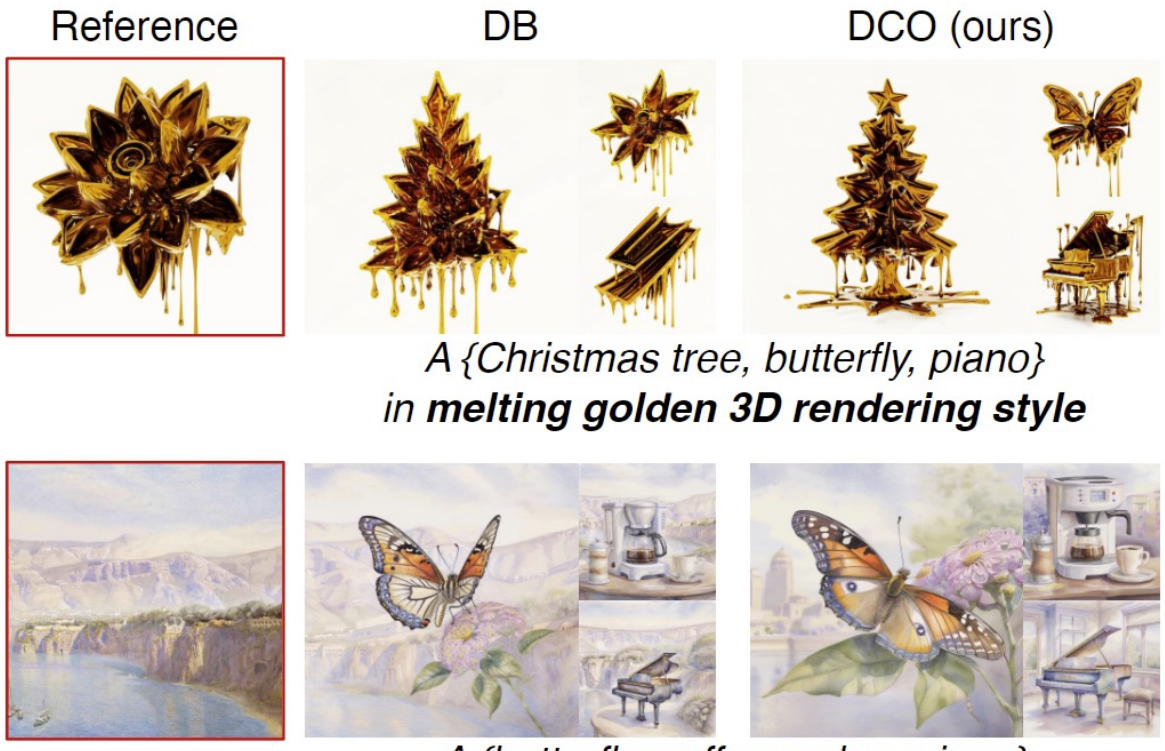
Lee, K., Kwak, S., Sohn, K., & Shin, J. (2024). Direct Consistency Optimization for Compositional Text-to-Image Personalization. arXiv preprint arXiv:2402.12004.

Direct Consistency Optimization (DCO)

DPO for Customization

❖ 실험 및 평가

- Stable Diffusion XL (SDXL) 1.0 + LoRA fine-tuning



<Stylization 비교>

- 기존 방법론보다 대상의 style만을 잘 포착하여 과적합 방지

Lee, K., Kwak, S., Sohn, K., & Shin, J. (2024). Direct Consistency Optimization for Compositional Text-to-Image Personalization. arXiv preprint arXiv:2402.12004.

Direct Consistency Optimization (DCO)

DPO for Customization

❖ 실험 및 평가

- **Merge:** personalization (content) + Stylization (style)

- 기존 방법론보다 대상의 identity 및 style을 더 잘 유지하면서 병합
- Pre-trained 모델과 너무 멀어지지 않고 잘 어울릴 수 있도록 하는 DCO loss 자체 제약 덕분에 LoRA fine-tuned 개별 모델들도 서로와 잘 어울리는 것

DB Merge: DreamBooth loss로 학습한
content LoRA와 style LoRA weight을 단순히 병합

DB ZipLoRA: DreamBooth loss로 학습한
content LoRA와 style LoRA weight을 ZipLoRA 방식으로 병합

DCO Merge: DCO loss로 학습한
content LoRA와 style LoRA weight을 단순히 병합

Lee, K., Kwak, S., Sohn, K., & Shin, J. (2024). Direct Consistency Optimization for Compositional Text-to-Image Personalization. arXiv preprint arXiv:2402.12004.

Direct Consistency Optimization (DCO)

DPO for Customization

❖ 실험 및 평가

- **Merge:** personalization (content) + Stylization (style) – 다양한 샘플링 프롬프트와 결합

- Content와 style을 동시에 잘 반영하면서 추가적인 프롬프트에 충실한 이미지 생성

Lee, K., Kwak, S., Sohn, K., & Shin, J. (2024). Direct Consistency Optimization for Compositional Text-to-Image Personalization. arXiv preprint arXiv:2402.12004.

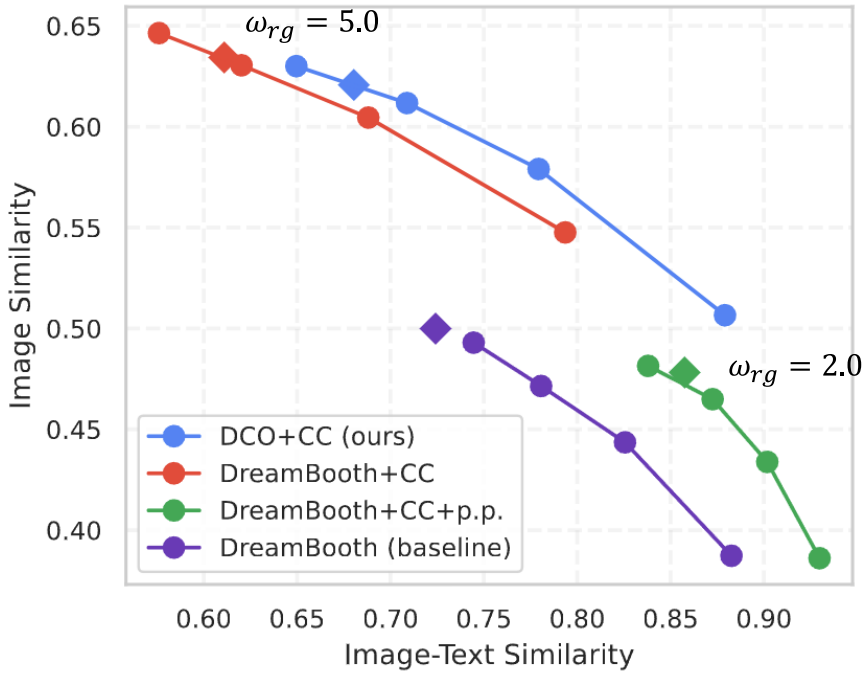
Direct Consistency Optimization (DCO)

DPO for Customization

※ **SigLIP:** CLIP과 동일하게 텍스트와 이미지를 같은 공간 상에서 학습한 모델
CLIP과 다른 점은 loss를 softmax에서 sigmoid로 바꾼 것

❖ 실험 및 평가

- 정량적 평가



CC: Comprehensive caption

p.p.: prior preservation

- **Image similarity:** DINOv2 score 이용
  Reference 이미지와 synthesized 이미지를 DINOv2에 넣고 임베딩 추출
  두 임베딩의 코사인 유사도를 image similarity로 정의
  → 값이 클수록 레퍼런스 이미지의 특징을 잘 잡도록 학습된 것

- **Image-text similarity:** SigLIP score 이용
  샘플링 프롬프트를 SigLIP 텍스트 인코더에 넣어 임베딩 추출
  + synthesized 이미지를 SigLIP 이미지 인코더에 넣어 임베딩 추출
  두 임베딩의 코사인 유사도를 image similarity로 정의
  → 값이 클수록 텍스트와 이미지 사이 align이 잘 맞는것

- **원형 point 4개:** 각각 $\omega_{rg} = 5.0, 4.0, 3.0, 2.0$으로 샘플링
- **사각형 point:** 일반적인 CFG scale만으로 샘플링

- 기존 방법론에 비해 DCO가 **Image similarity**와 **Image-text similarity**에서 모두 높은 성능 기록

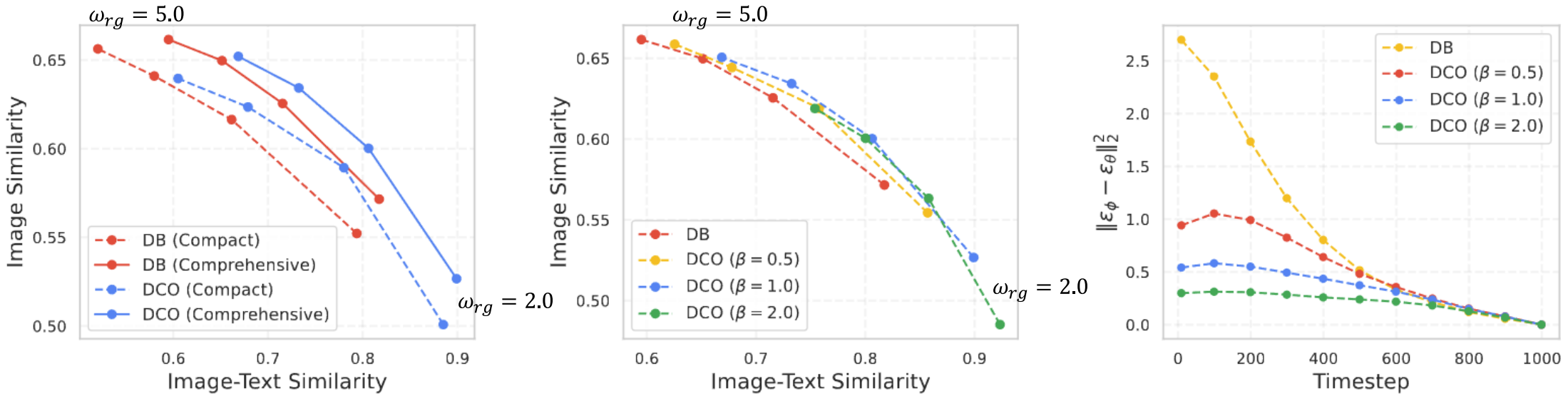
Lee, K., Kwak, S., Sohn, K., & Shin, J. (2024). Direct Consistency Optimization for Compositional Text-to-Image Personalization. arXiv preprint arXiv:2402.12004.

Direct Consistency Optimization (DCO)

DPO for Customization

❖ 실험 및 평가

- Ablation study



(a) Ablation on comprehensive caption

(b) Ablation on β

(c) Noise distance

DCO가 pre-trained model output과 fine-tuned model output 사이 오차가 더 작음

+ 큰 값의 β로 학습시킬수록 (=KL regularization을 강하게 걸수록) 오차가 더 작음

Lee, K., Kwak, S., Sohn, K., & Shin, J. (2024). Direct Consistency Optimization for Compositional Text-to-Image Personalization. arXiv preprint arXiv:2402.12004.

고맙습니다