Extending Reinforcement Learning Techniques for Diffusion Models

Cristóbal Alcázar

Universidad de Chile
cristobal.alcazar@ug.uchile.cl

Advisor: Felipe Tobar
Tesis para optar al grado de Magíster de Ciencia de Datos
October 22, 2024

FACULTAD DE
CIENCIAS FÍSICAS
Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

Table of Contents

Research Motivation: Why Does This Matter?

1. Generative models trained on large-scale data unlock powerful capabilities.
2. Embedding these models into real-world products requires more than generation—*control, reliability, and usability are critical*.
3. How can we orchestrate generative capabilities to drive downstream tasks effectively? Focus on building systems, not just models.

Research problem: bringing control to generation reinforcement learning (RL) on diffusion models (DM)

1. Can RL control DM? Yes, it can! (Black, 2023 [2])
2. Why RL? Exploration is key in RL, and reward models integrate human feedback into the process
3. RL serves as a building block for DM interfaces, enabling post-training as human-computer interaction (see Human-Centric Reward Design, Du 2023 [3]).

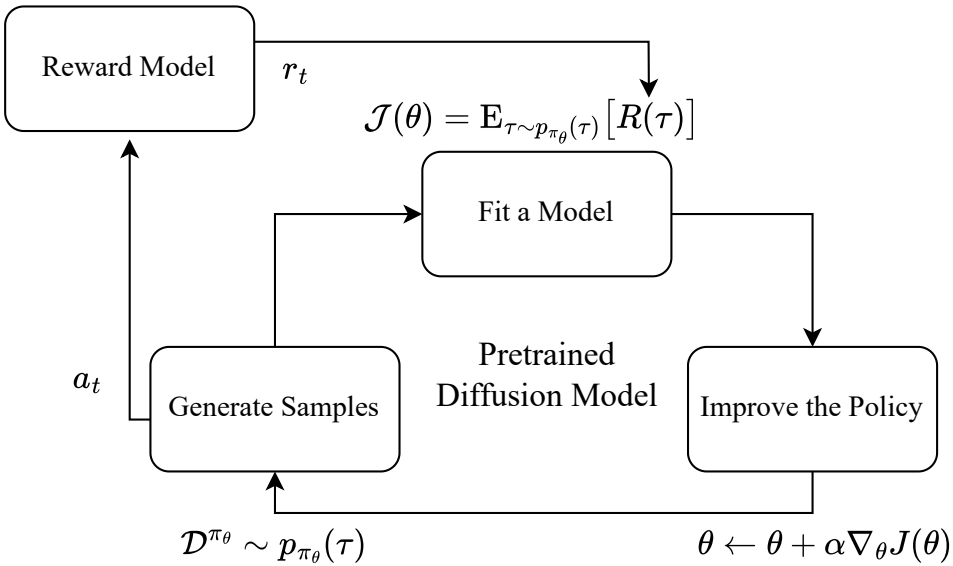


Figure: Interaction between RL and DM during post-training

Hypothesis and Objectives

Hypothesis: Reinforcement Learning (RL) can effectively finetune pretrained diffusion models (DM) for new tasks by optimizing reward functions that align with the desired behavior.

1. **General objective:** explore and validate the use of RL for finetuning diffusion models.
2. **Specific objectives:**
    1. Establish a foundational understanding on DM and RL (background).
    2. Understand the intersection between DM and RL, framing DM as a task to be optimized by an RL agent using policy gradient methods.
    3. Perform an empirical analysis of reward signals throughout the diffusion process.
    4. Implement and adapt DDPO for reward finetuning on smaller models following the DDPM framework.
    5. Assess the adaptability of RL techniques across various diffusion models and tasks.

Background

Denoising Diffusion Probabilistic Models, or DDPM (Ho, 2020 [4])

1. Represents a distribution $p(x_0 \mid c)$ over data samples $x_0 \in \mathcal{X}$ conditioned on contexts $c \in \mathcal{C}$.
2. Key idea is to successively corrupting data with noise and concurrently train a model to denoising it.
3. This distribution is defined by reversing the forward markovian process $q(x_t \mid x_{t-1})$ where the chain $\{x_t\}_{t=0}^{T}$ is a sequence of samples with increasing levels of noise
4. such that x_0 represents a clean data sample and x_T a completely noise

Training DDPM from the Variational Lower Bound

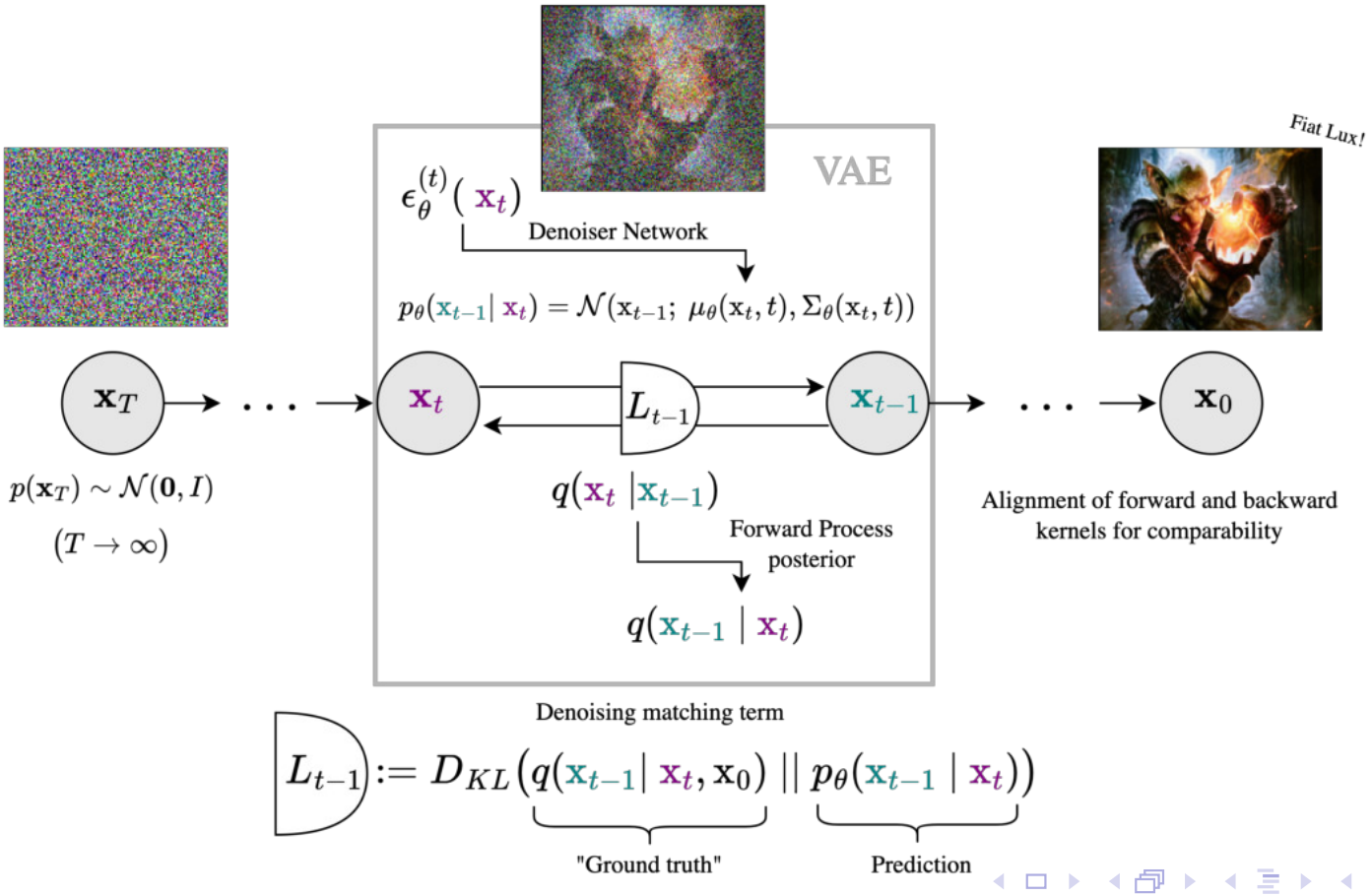
DDPMs are trained by optimizing the Variational Lower Bound (aka ELBO) [4, 5] on the negative likelihood of the target distribution using the backward transition kernel p_θ,

$$
\begin{aligned}
-\log p_\theta(\mathrm{x}_0) &= -\log \int p_\theta(\mathrm{x}_{0:T}) d\mathrm{x}_{1:T} \\
&= -\log \int \frac{p_\theta(\mathrm{x}_{0:T}) q(\mathrm{x}_{1:T} \mid \mathrm{x}_0)}{q(\mathrm{x}_{1:T} \mid \mathrm{x}_0)} d\mathrm{x}_{1:T} \\
&= -\log \mathbb{E}_{q(\mathrm{x}_{1:T}|\mathrm{x}_0)} \left[\frac{p_\theta(\mathrm{x}_{0:T})}{q(\mathrm{x}_{1:T} \mid \mathrm{x}_0)} \right] \\
&\leq \mathbb{E}_{q(\mathrm{x}_{1:T}|\mathrm{x}_0)} \left[-\log \frac{p_\theta(\mathrm{x}_{0:T})}{q(\mathrm{x}_{1:T} \mid \mathrm{x}_0)} \right] \qquad (1) \\
&\leq \mathbb{E}_{q(\mathrm{x}_{1:T}|\mathrm{x}_0)} \left[-\log p(\mathrm{x}_T) - \sum_{t \geq 1}^{T} \log \frac{p_\theta(\mathrm{x}_{t-1} \mid \mathrm{x}_t)}{q(\mathrm{x}_t \mid \mathrm{x}_{t-1})} \right]
\end{aligned}
$$

$$
L := \mathbb{E}_{q(\mathrm{x}_{1:T}|\mathrm{x}_0)} \left[-\log p(\mathrm{x}_T) - \sum_{t>1}^{T} \log \frac{p_\theta(\mathrm{x}_{t-1} \mid \mathrm{x}_t)}{q(\mathrm{x}_t \mid \mathrm{x}_{t-1})} - \log \frac{p_\theta(\mathrm{x}_0 \mid \mathrm{x}_1)}{q(\mathrm{x}_1 \mid \mathrm{x}_0)} \right].
$$

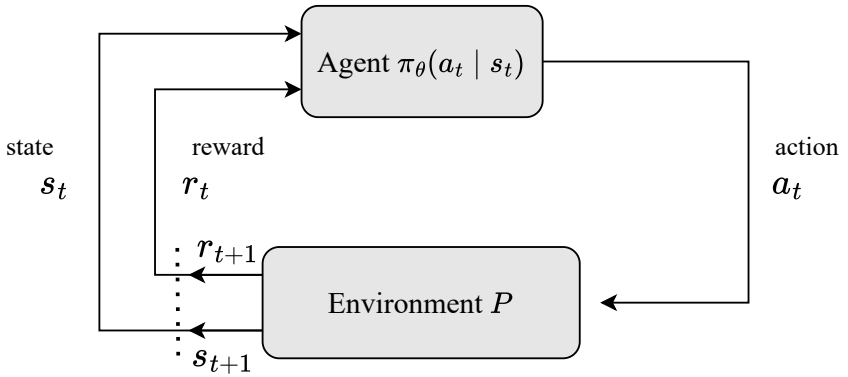
The most significant term in the cost function derived from the VLB is known as the *denoising matching term*

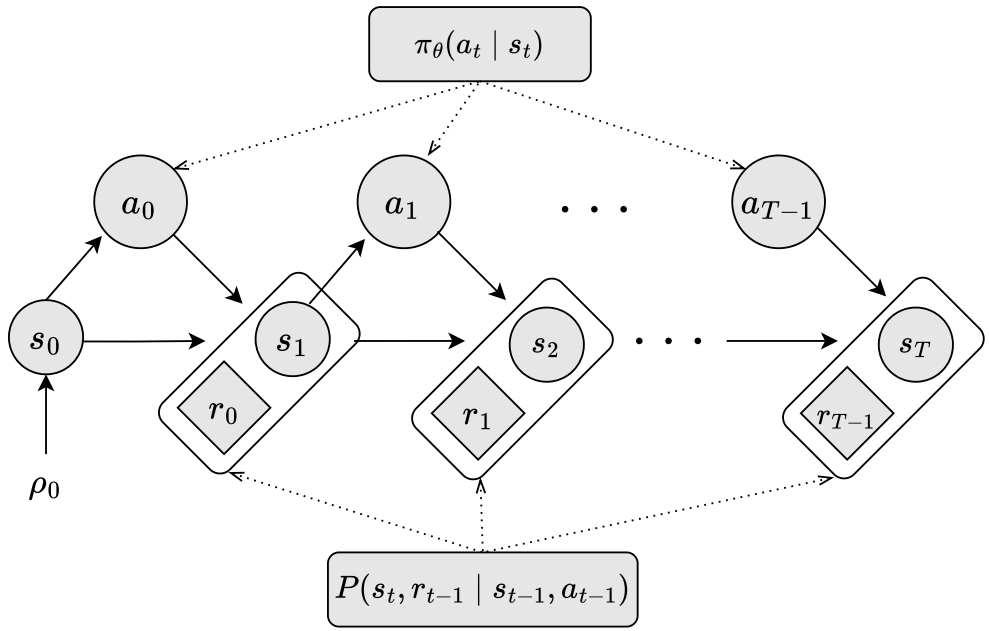
The forward process transforms the problem setting from unsupervised to a self-supervised one, providing a *"ground truth"* from the data itself to guide the learning process,



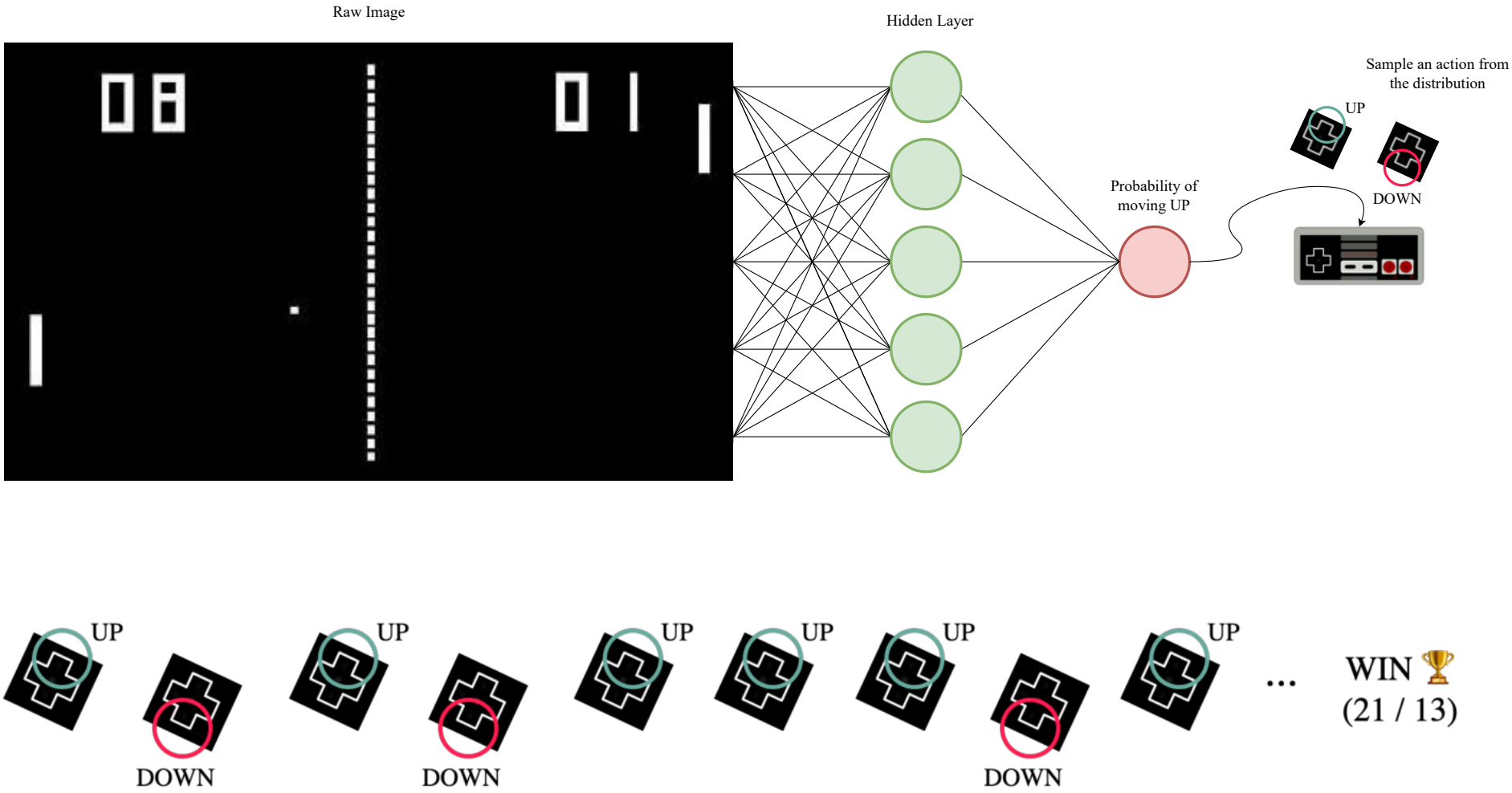
Reinforcement Learning: the framework for learning to act

A Markov Decision Process (MDP) is a mathematical object that describes the interaction between an agent and an environment. The agent interacts with the environment by taking actions, receiving rewards, and observing the state of the environment.





REINFORCE in action, mastering to play ATARI Pong



Deep Reinforcement Learning: Pong from Pixels (Karpathy, 2016)

Gradient Estimation via Score Function

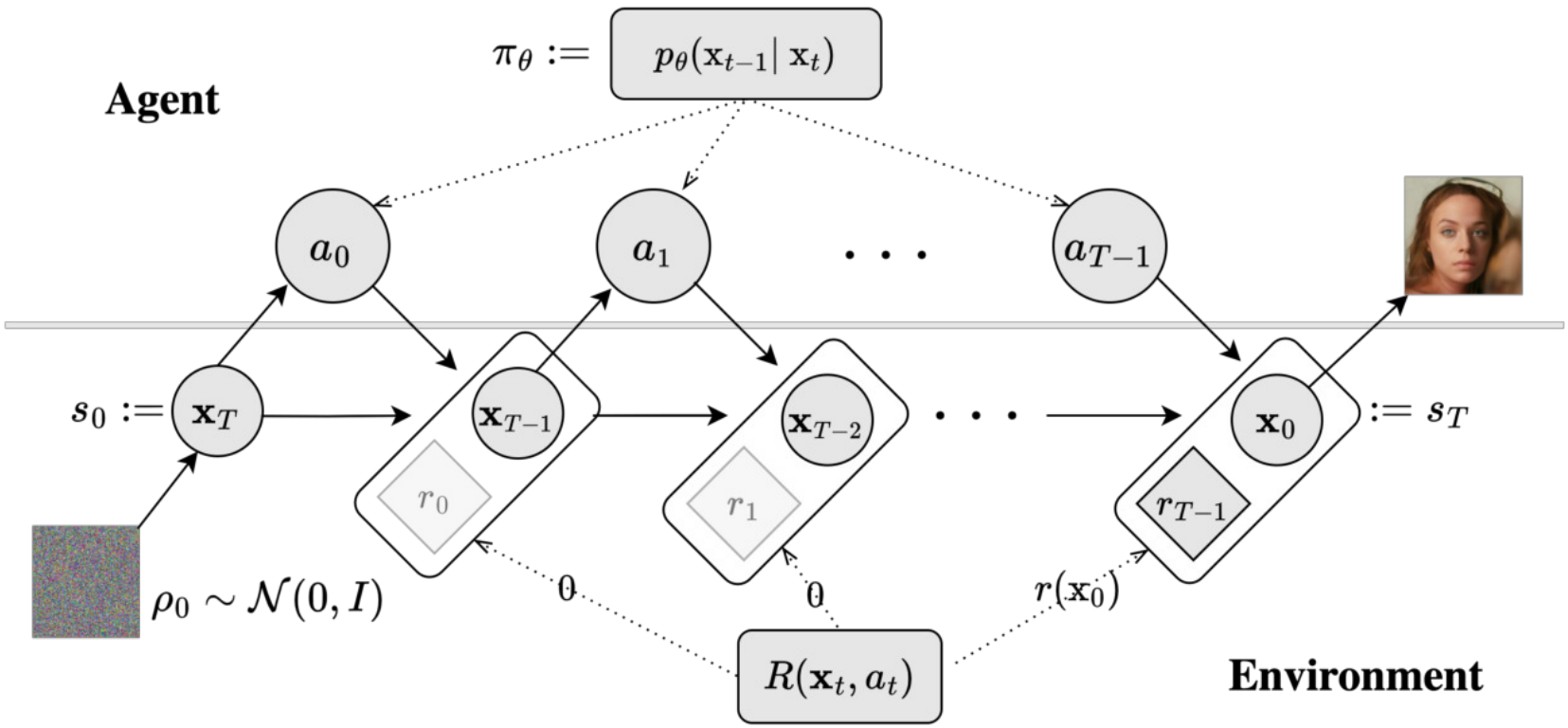
Intuitivelly, we want to collect the trajectories and make the good trajectories and actions more probable, and push the actions towards betters actions. This involves obtaining gradient information from sample trajectories, with performance assessed by a scalar-value function (i.e. reward).

$$
\begin{aligned}
g = \nabla_\theta \mathbb{E}_{p(\mathrm{x};\theta)}[f(\mathrm{x})] &= \nabla_\theta \int_{\mathcal{X}} p(\mathrm{x};\theta) f(\mathrm{x}) d\mathrm{x} \\
&= \int_{\mathcal{X}} \nabla_\theta \; p(\mathrm{x};\theta) f(\mathrm{x}) d\mathrm{x} \\
&= \int_{\mathcal{X}} p(\mathrm{x};\theta) \nabla_\theta \log p(\mathrm{x};\theta) f(\mathrm{x}) d\mathrm{x} \\
&= \mathbb{E}_{p(\mathrm{x};\theta)} \left[f(\mathrm{x}) \nabla_\theta \log p(\mathrm{x};\theta) \right].
\end{aligned} \tag{2}
$$

The connection with REINFORCE (Williams, 1992 [7]), is given by $\mathcal{J}_{\theta_{\mathrm{RL}}} = \mathbb{E}_{\tau \sim p_\theta(\tau)} [R(\tau)]$.

The Intersection of DM and RL

DDPM backward process, i.e. inference, as a sequential decision-making process



$$\mathcal{J}_{\text{DDRL}}(\theta) = \mathbb{E}_{c\sim p(c), x_0 \sim p_(\theta|c)}\left[r(x_0, c)\right]$$

Denoising Diffusion Policy Optimizatioon (DDPO)

Denoising as a multi-step MDP

$$s_t \triangleq (c, t, \mathbf{x}_t) \quad \pi(a_t \mid s_t) \triangleq p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, c) \quad P(s_{t+1} \mid s_t, a_t) \triangleq (\delta_c, \delta_{t-1}, \delta_{\mathbf{x}_{t-1}})$$

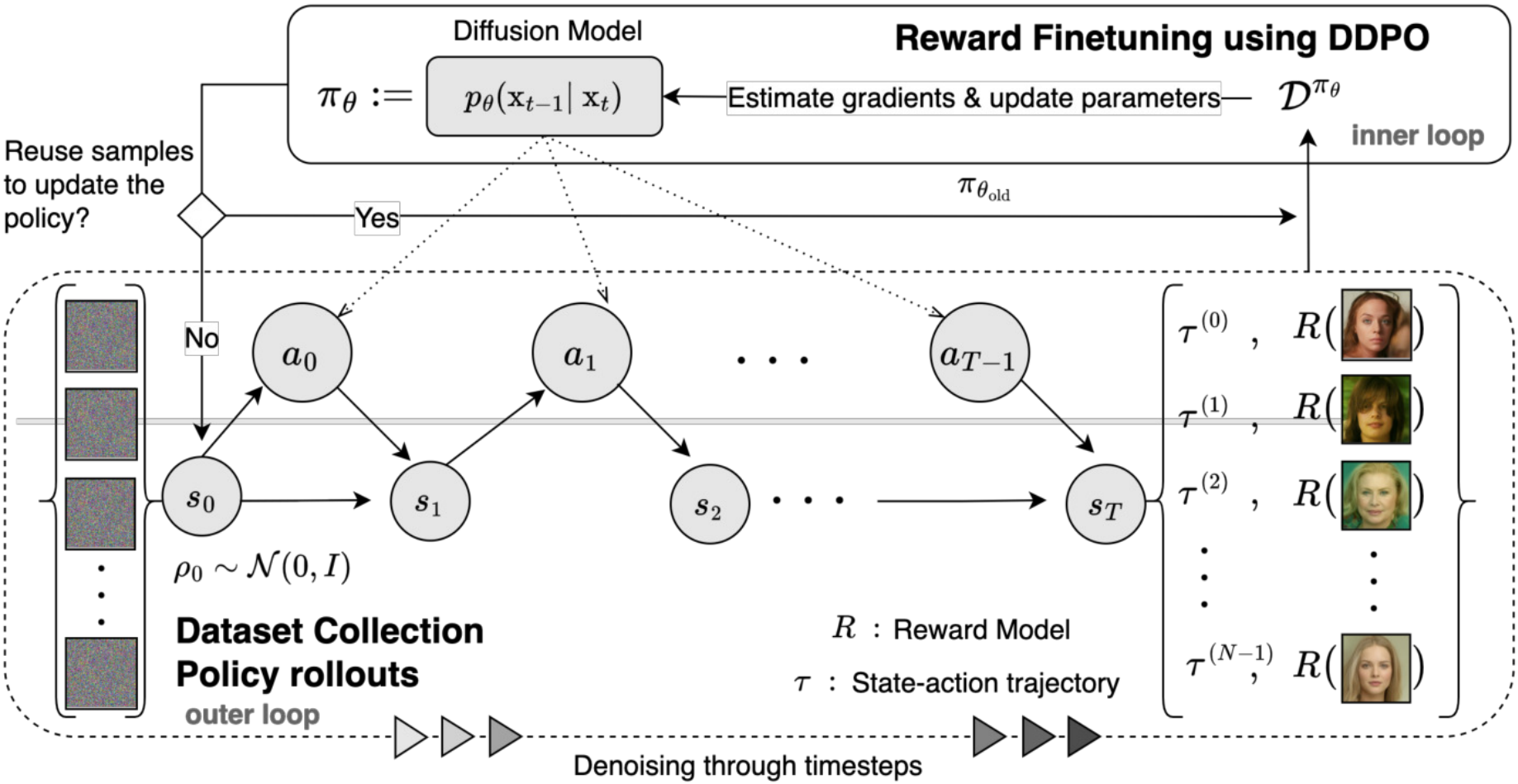
$$a_t \triangleq \mathbf{x}_{t-1} \quad \rho_0(s_0) \triangleq (p(c), \delta_T, \mathcal{N}(0, I)) \quad R(s_t, a_t) \triangleq \begin{cases} r(\mathbf{x}_0, c) & \text{if } t = 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\nabla_\theta \mathcal{J}_{\text{DDRL}} = \mathbb{E}\left[\sum_{t=0}^{T} \nabla_\theta \log p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, c)\, r(\mathbf{x}_0, c)\right] \qquad (\text{DDPO}_{\text{SF}})$$

$$\nabla_\theta \mathcal{J}_{\text{DDRL}} = \mathbb{E}\left[\sum_{t=0}^{T} \frac{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, c)}{p_{\theta_{\text{old}}}(\mathbf{x}_{t-1} \mid \mathbf{x}_t, c)} \nabla_\theta \log p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, c)\, r(\mathbf{x}_0, c)\right] \qquad (\text{DDPO}_{\text{IS}})$$

Training Diffusion Models with Reinforcement Learning (Black et al., 2023 [2])

Reward-Driven Finetuning with DDPO: Adapting diffusion models to downstream tasks



Empirical Analysis of Reward Dynamics in Denoising Trajectories

Empirical Analysis of Reward Dynamics in Denoising Trajectories

- Analyzing trajectories reward distribution on aesthetic quality and image filesize after JPEG compression.
- Human feedback *"in a bottle"*, LAION-Aesthetic predictor V2 (Schuhmann, 2022 [6]).
- A higher score indicates that the image has higher *aesthetic quality*.

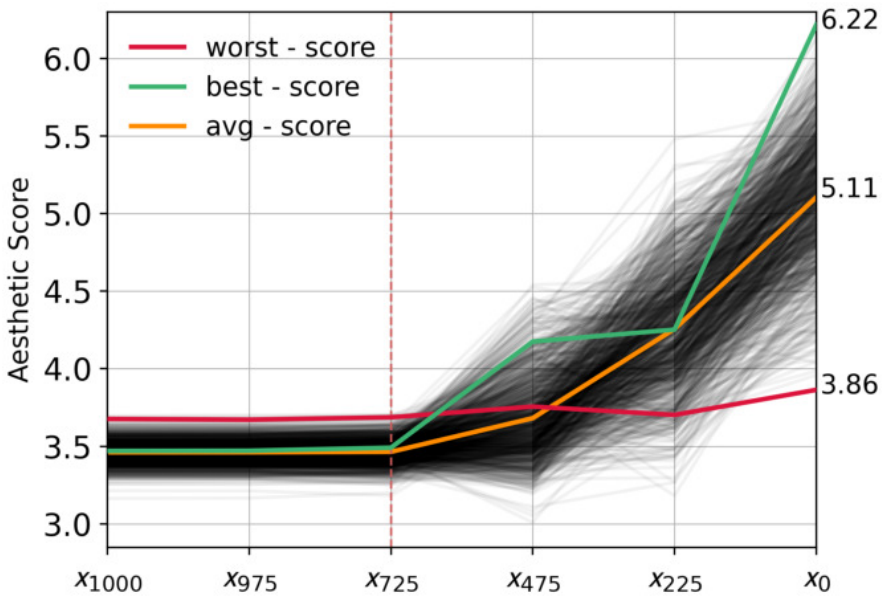


Figure: LAION Aesthetic Score Predictor

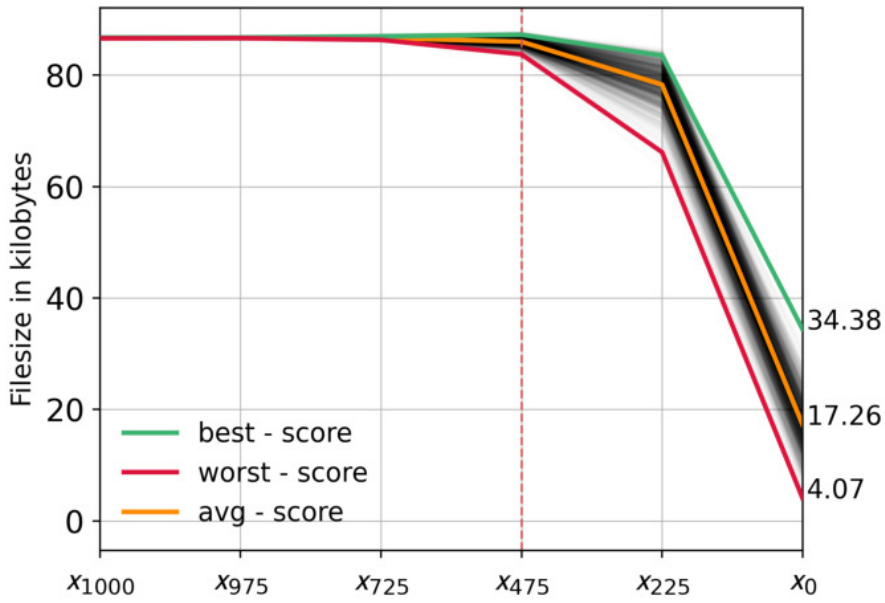
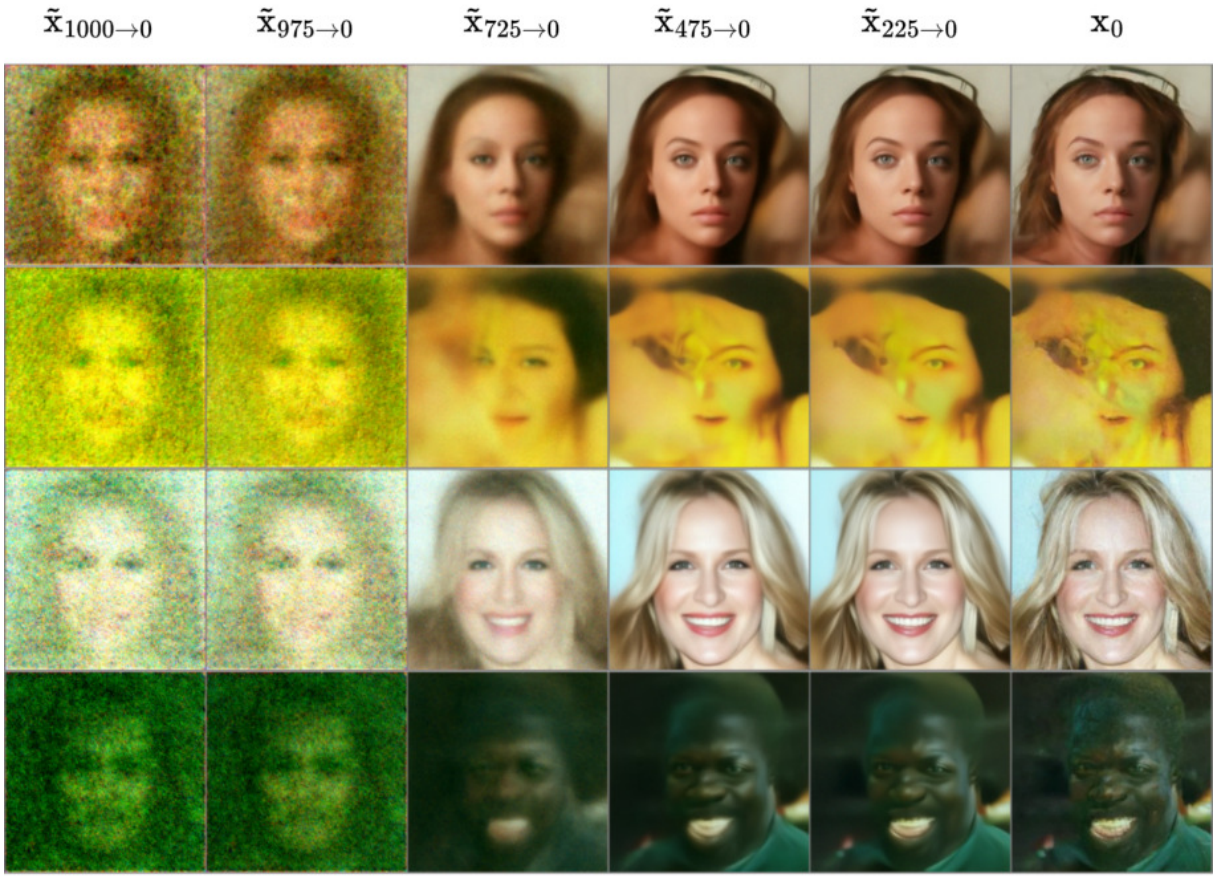


Figure: JPEG compression

Can we do better in extract signal from intermediate rewards?

We can obtain a denoised observations $\tilde{\mathrm{x}}_{t\to 0}$ from the intermediate states x_t using the amount of noise for t, and noise estimation by the trained model .



Can we exploit rewards on intermediate states?

Yes, using the denoised trajectoriees we can extract the reward signal from the intermediate states.

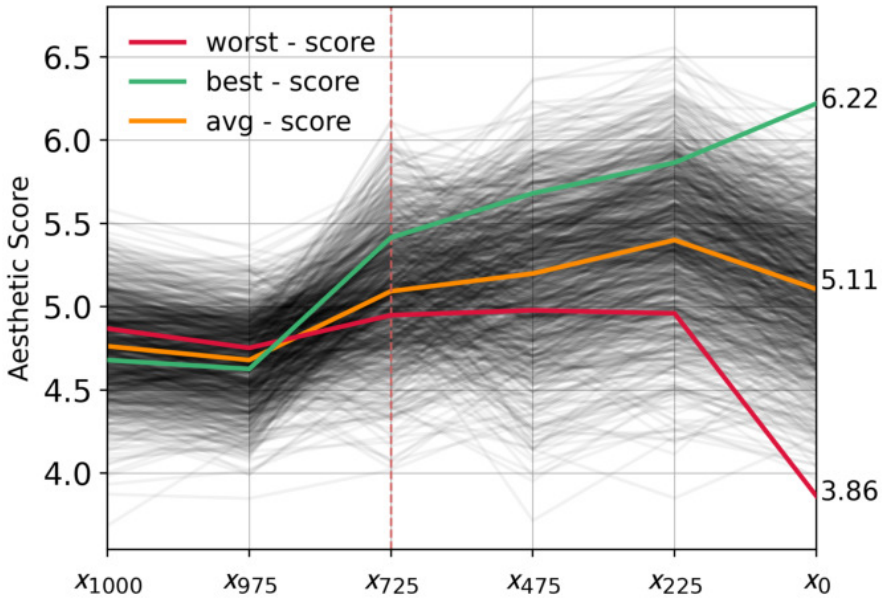


Figure: LAION Aesthetic Score Predictor

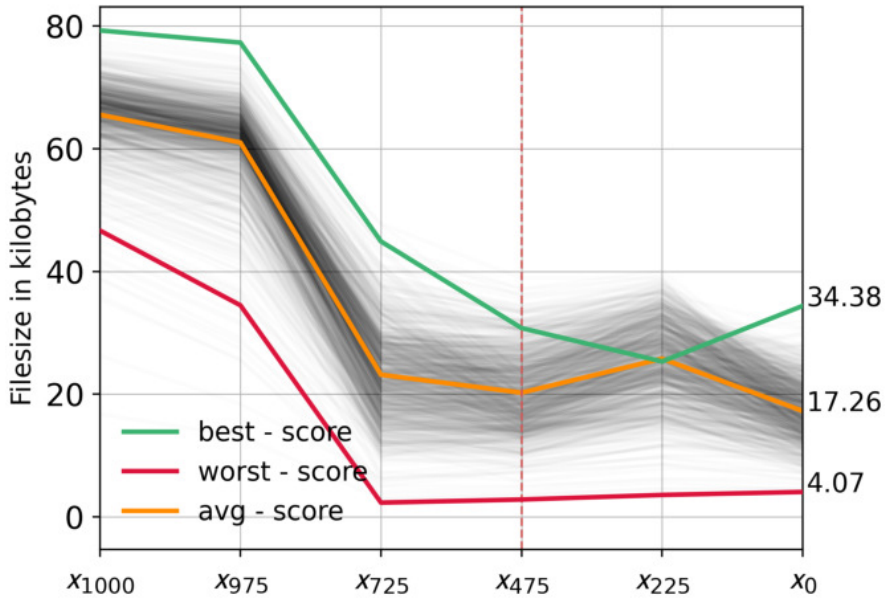


Figure: JPEG compression

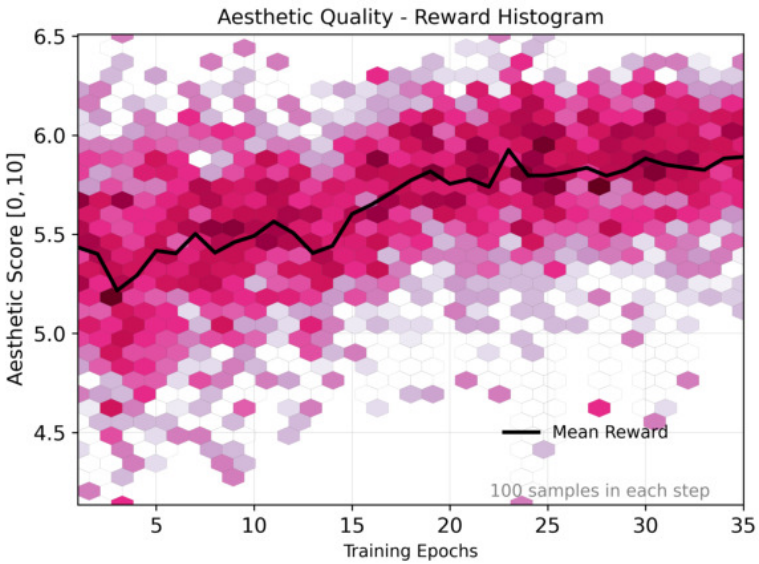
Reward Finetuning Results

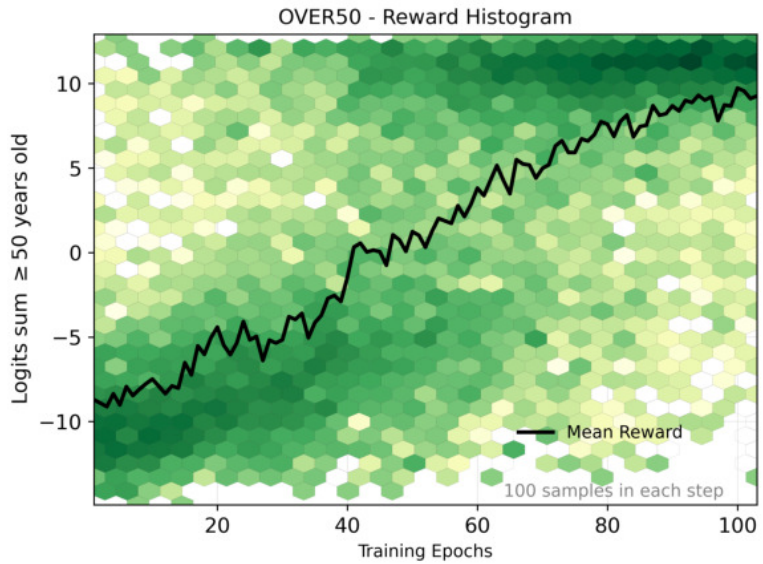
Summary of reward finetuning experiments on 2 models and 4 downstream tasks

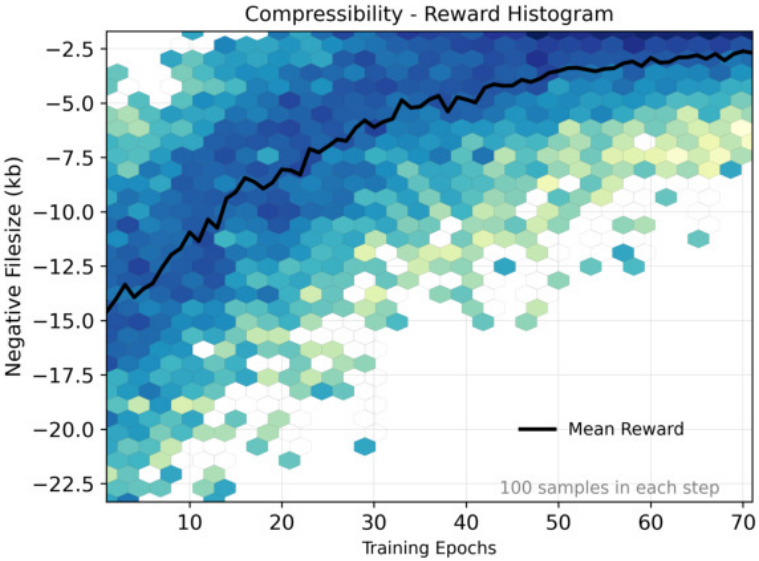
Downstream Task	Baseline	DDPO
`google/ddpm-celebahq-256`		
Aesthetic Score (↑ better)	5.11 ± 0.01	**5.58** ± 0.01
Compressibility (↓ better)	17.26 ± 0.15	**6.01** ± 0.13
Incompressibility (↑ better)	17.26 ± 0.15	**21.6** ± 0.12
Over 50 years old (↑ better)	-7.72 ± 0.17	**7.39** ± 0.16
`google/ddpm-church-256`		
Aesthetic Score (↑ better)	4.77 ± 0.01	**5.13** ± 0.01
Compressibility (↓ better)	29.57 ± 0.29	**10.62** ± 0.18
Incompressibility (↑ better)	29.57 ± 0.29	**50.21** ± 0.34

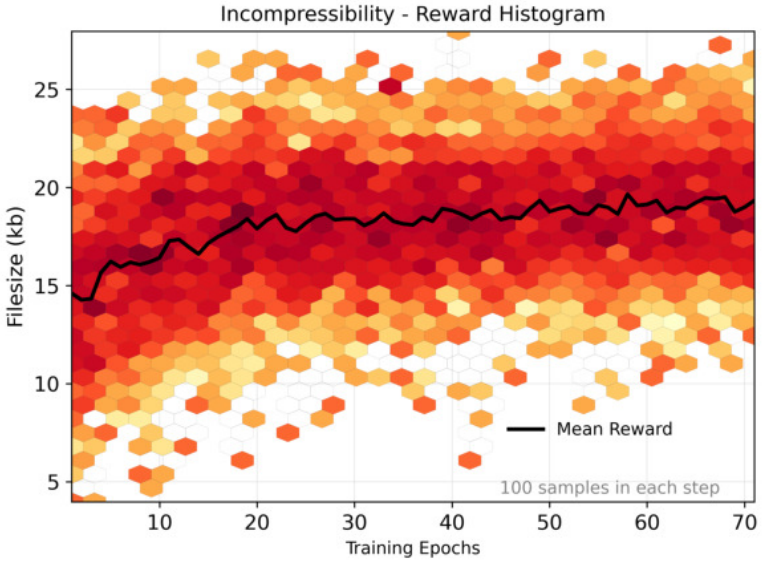
Table: **Mean and standard deviation for each downstream task across two pretrained models**. All samples were generated using the same initial noise to ensure a fair comparison.

Evolution of Mean Reward and Reward Distribution Across Downstream Tasks in Face Generation









Visual Comparison between a Pretrained DDPM Model and Reward Finetuned Models with DDPO

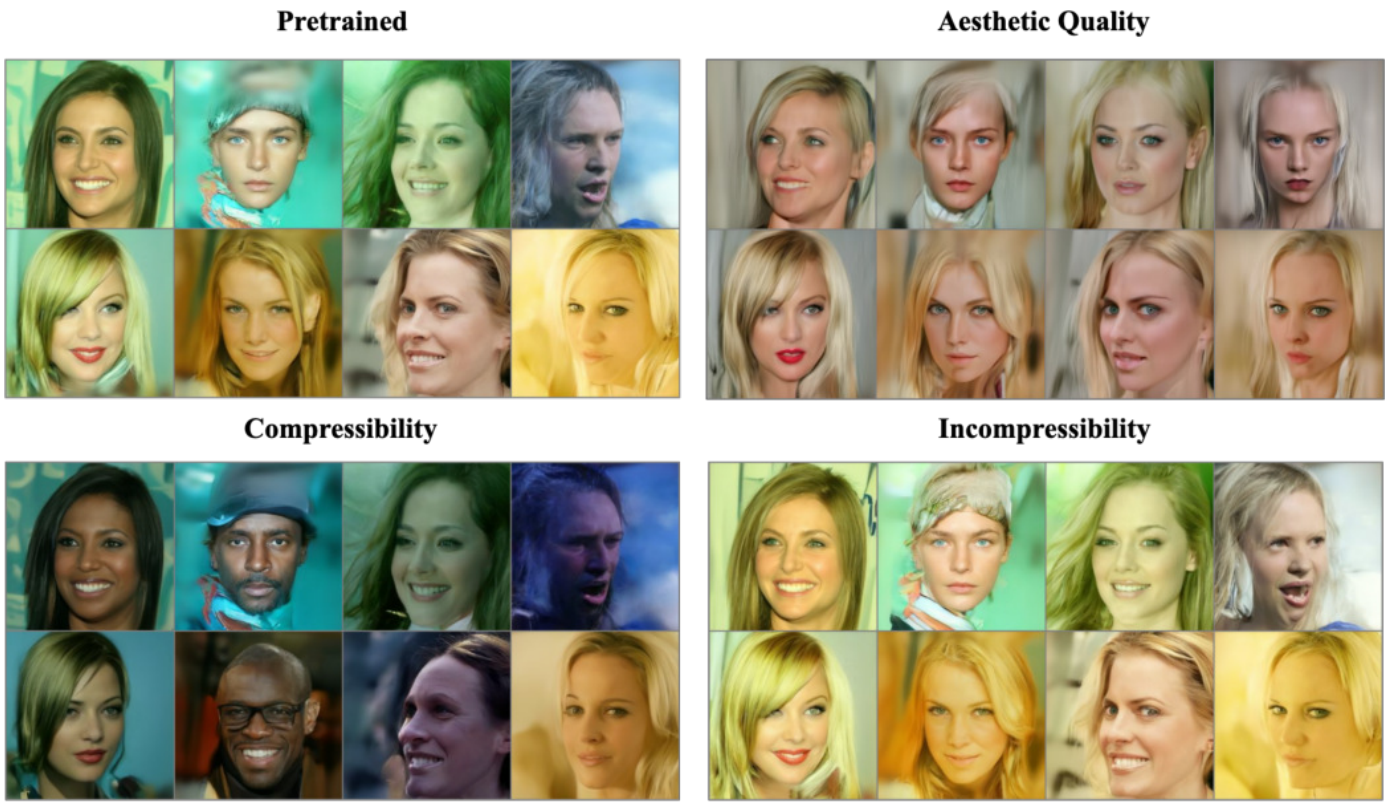


Figure: Top left panel shows samples from the DDPM pretrained model google/ddpm-celebahq-256. The other panels display samples from the same initial noise, using models finetuned with DDPO for different rewards.

Putting faces on the reward dynamics when optimized by compressibility

What are the emerging effects of optimizing for JPEG compressibility?

Gender Presentation / Melanotropism / Attenuation

Melanotropism / Blurring

Realistic-looking / Attenuation / Blurring

Gender Presentation / Attenuation / Realistic-looking

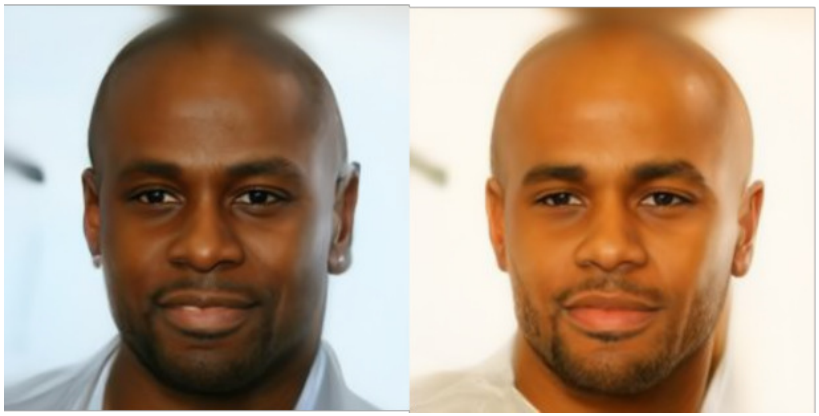
Incompressibility: Optimizing Images to Be Harder to Compress Using JPEG

What are the emerging effects of optimizing for JPEG incompressibility?

Gender Presentation / Hair Voluminization / Illumination Increase

Hair Voluminization & Definition / Skin Tone Lightening / Illumination Increase

Skin Tone Lightening / Illumination Increase



Hair Voluminization & Definition / Gender Presentation / Makeup Application / Illumination Increase

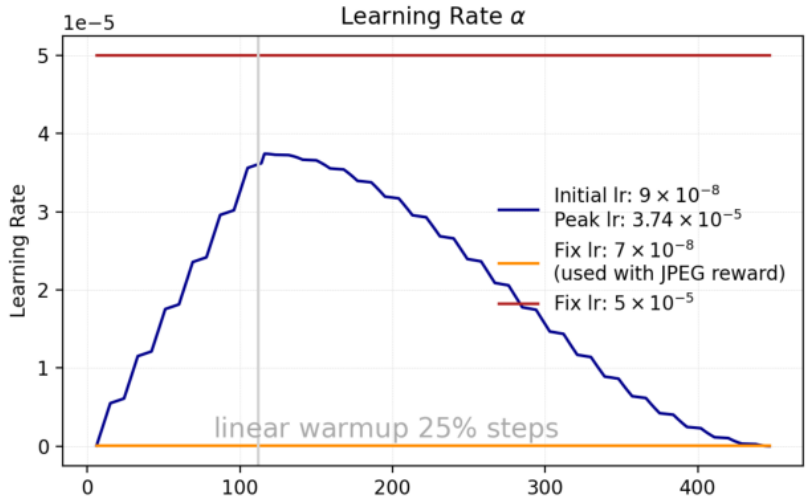
Compressibility and Incompressibility: both sides of the coin

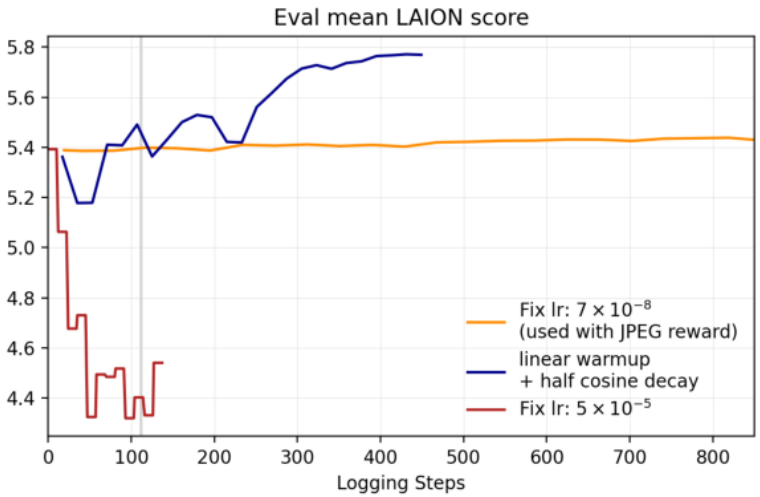
1. Not all sides are weighted equally.
2. Destruction of information leads to quality loss:
   - File size reduction can always be accomplished by diminishing the model's generative capacity
3. Creating information presents unique challenges:
   - This process is constrained by the generative capabilities of the model, limiting effective exploitation.
   - *Removing information is inherently simpler than adding or generating new information.*

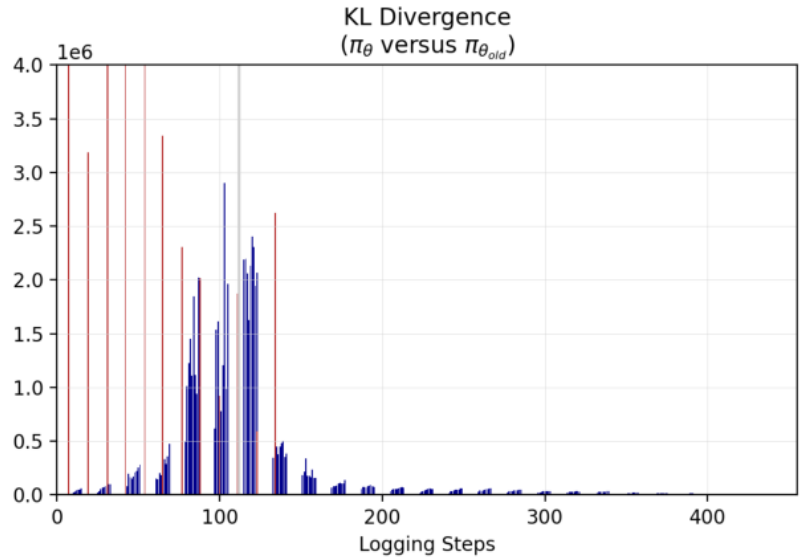
Optimizing towards stereotypical cover magazine images

Improving *aesthetic quality* with the perfect recipe: a learning rate with linear warm-up and half cosine decay

1. Previous tasks used a fixed learning rate (orange), with low rate and minimal experimentation
2. Problem: low learning rates show little improvement; high rates degrade the model (red)







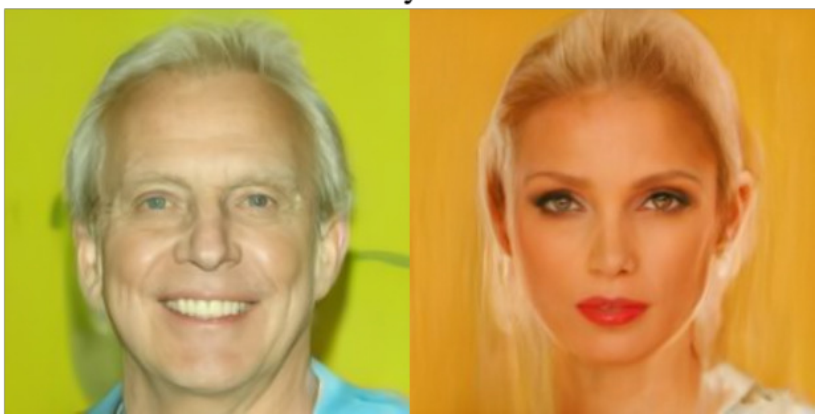
What are the emerging effects of optimizing for aesthetic quality?

Rejuvenation / Sketched / Profiled / Gaze Intensity

Warmer Tones / Gaze Intensity / Blonde Coloration

Rejuvenation / Profiled / Gaze Intensity / Warmer Tones

Gender Presentation / Rejuvenation / Warmer Tones / Gaze Intensity / Sketched



New downstream task: Increasing the Frequency of Generated Celebrity-Like Faces Over 50 Years Old

1. Roughly 6% of the samples generated by the pretrained model are $50 \geq$ years old.
2. Can we leverage RL to generate more samples of this kind?

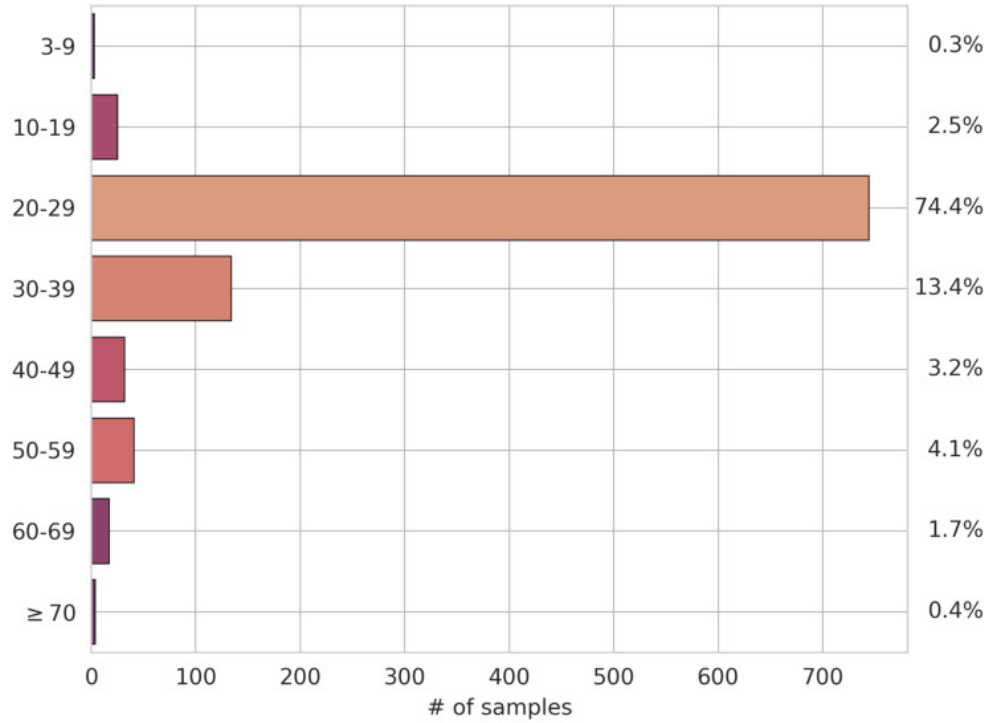
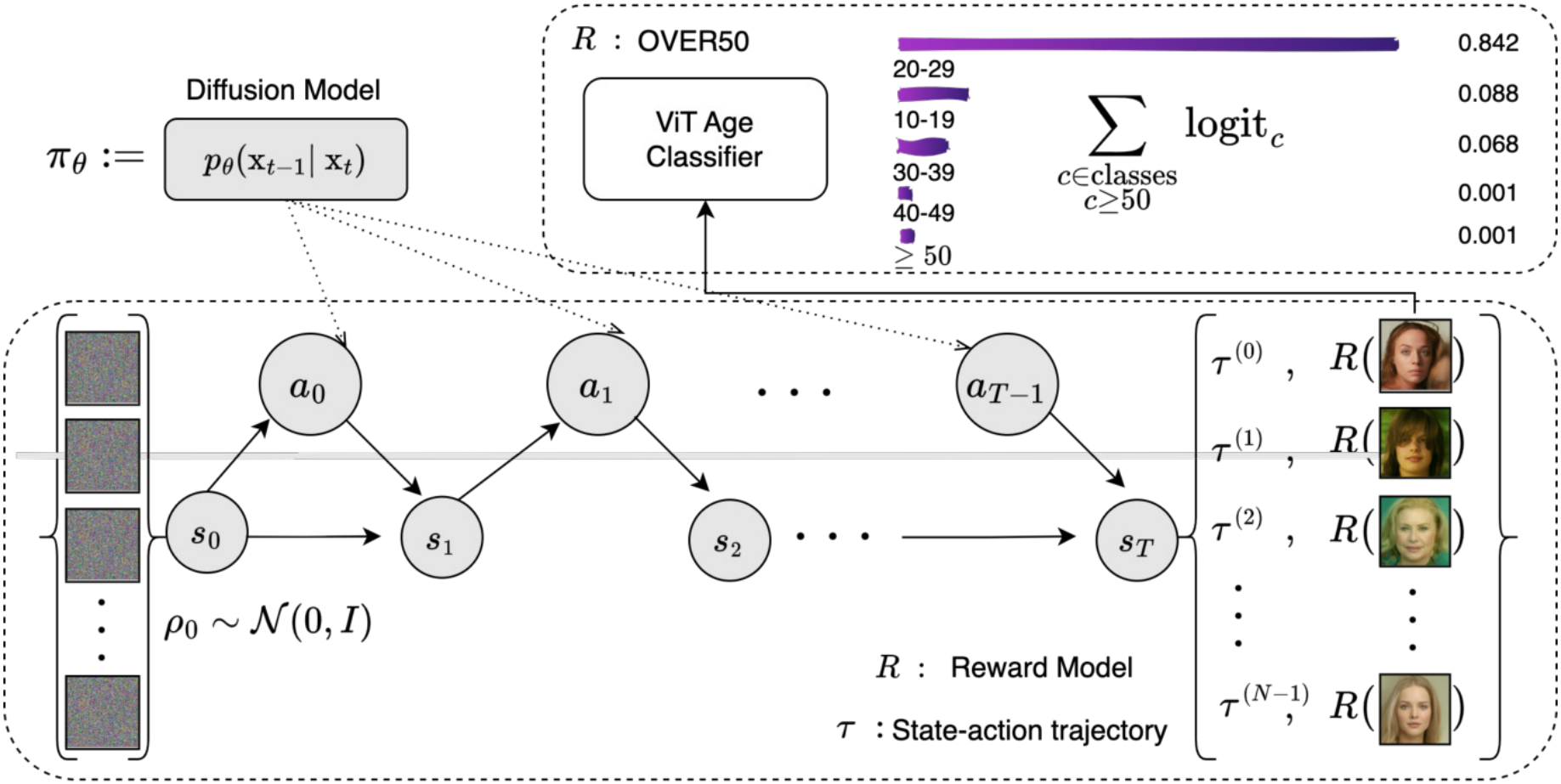


Figure: **Age samples distribution on images generated by the pretrained model**. Estimate using the ViT Age Classifier.

OVER50: Use an *off-the-shelf* classifier to design the reward function



ViT Age Classifier Trained on Fairface dataset (Nate Raw, 2021)

Using OVER50 the frequency generation of older faces increase from 6.1% to 78.7%

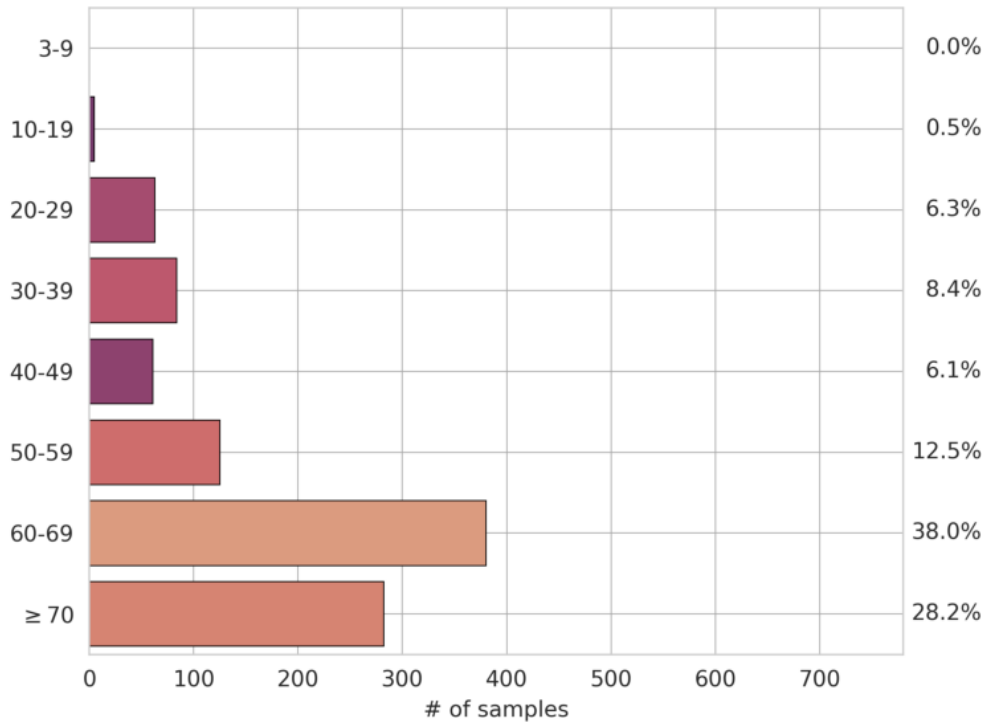
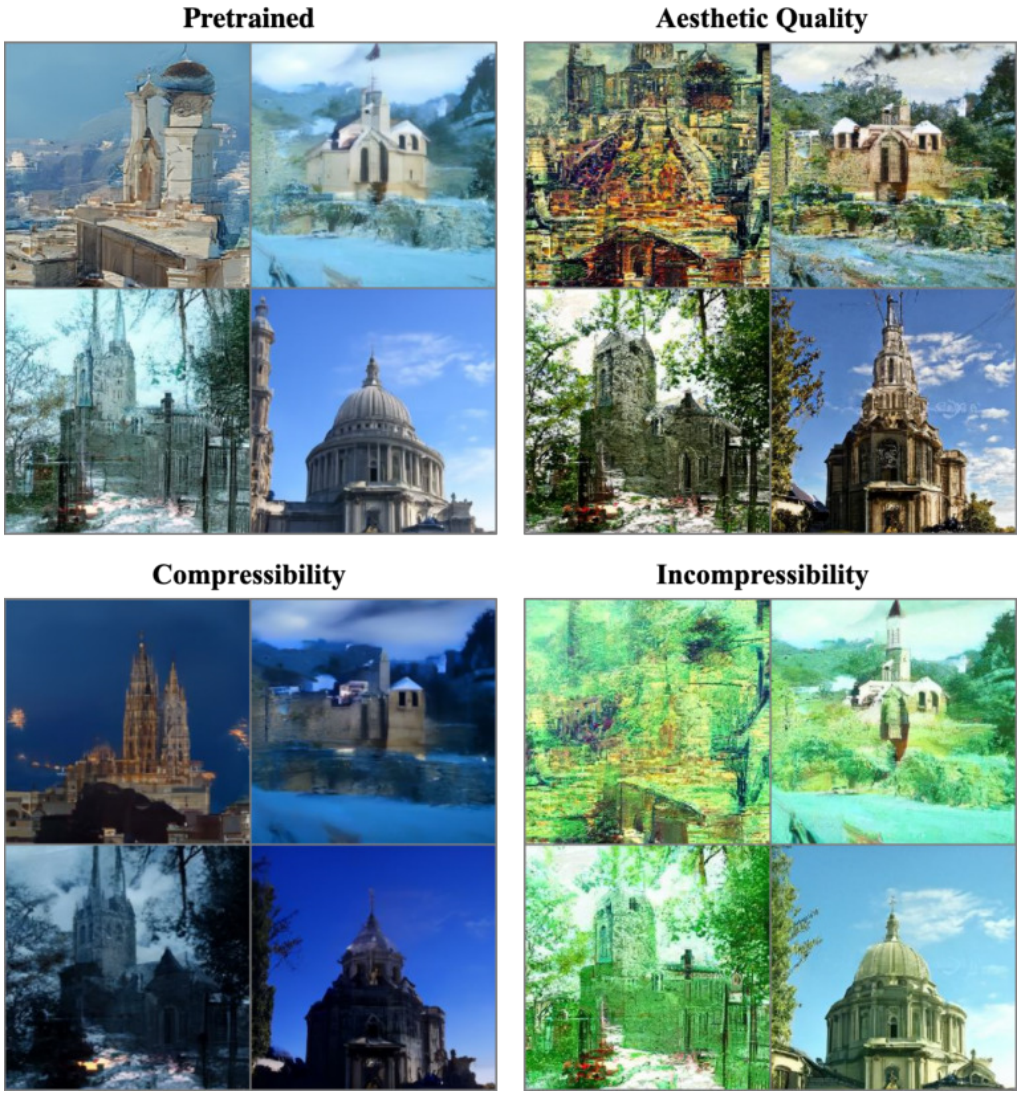


Figure: **Age samples distribution on images generated by the postrained model** `OVER50`. Estimate using the ViT Age Classifier.

1. Can we leverage RL to generate more samples of this kind?
2. **Yes!** and considering that we use a really simple exploration strategy (not all) indeed

OVER50: *the rhythm of time as the parameter update of a model*

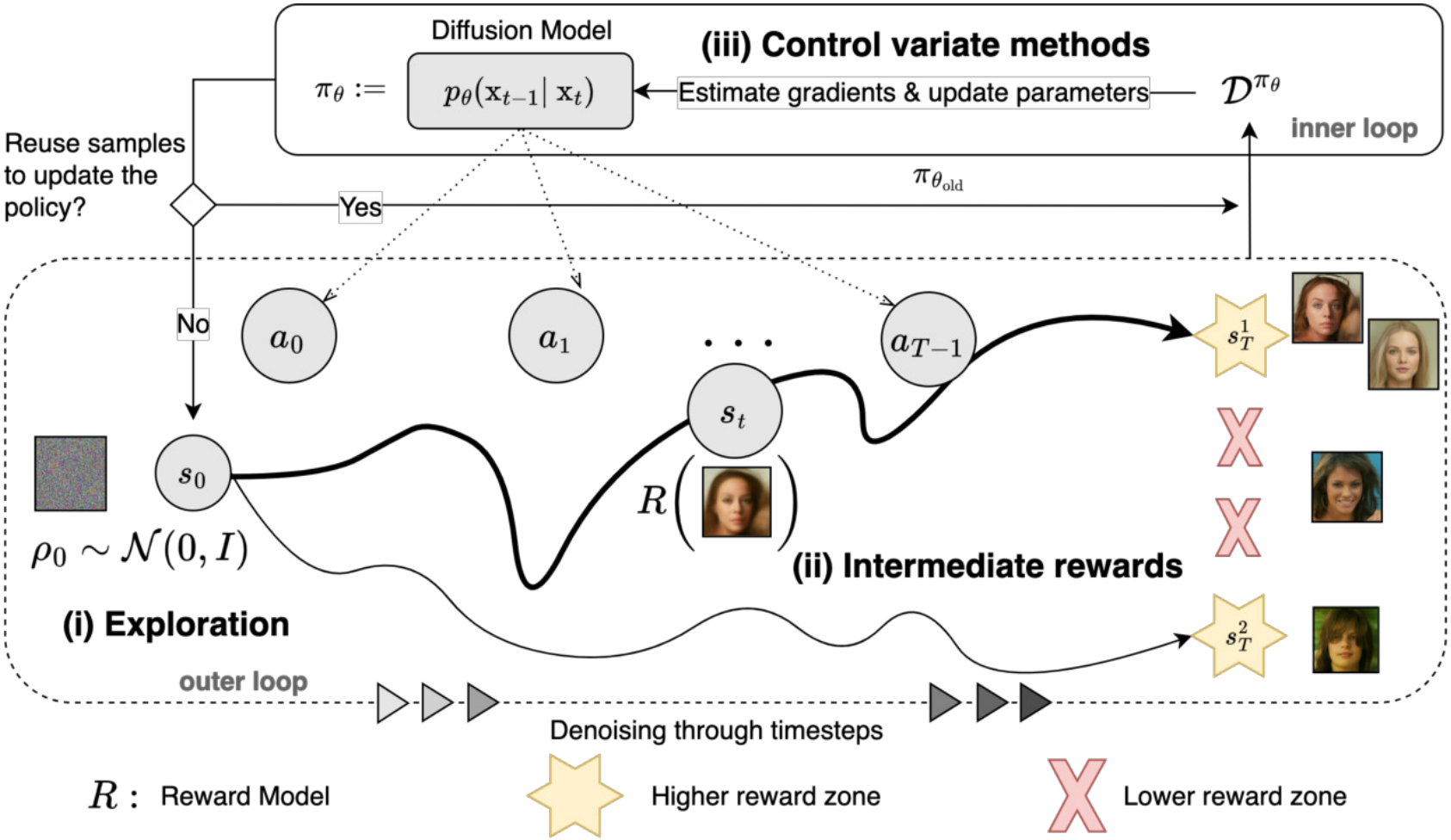
Beyond Face Generation: *does this setting generalize to other domains?*



Beyond Face Generation: *does this setting generalize to other domains?*

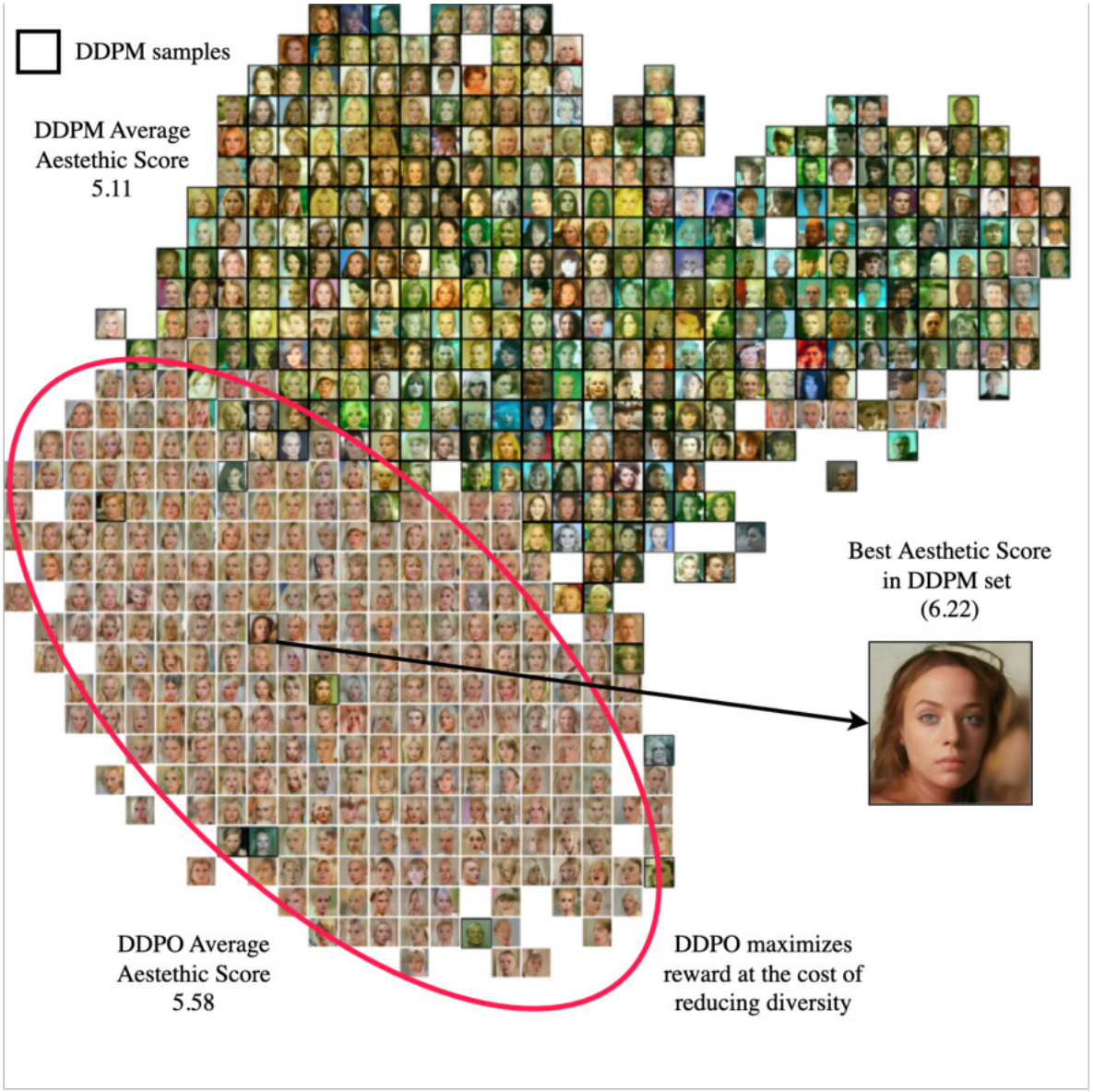
Figure: **Transitions from DDPM toward optimizting rewards functions with DDPO**. were generated using the same initial noise to ensure a fair comparison.

Future Directions: *what are the next steps?*

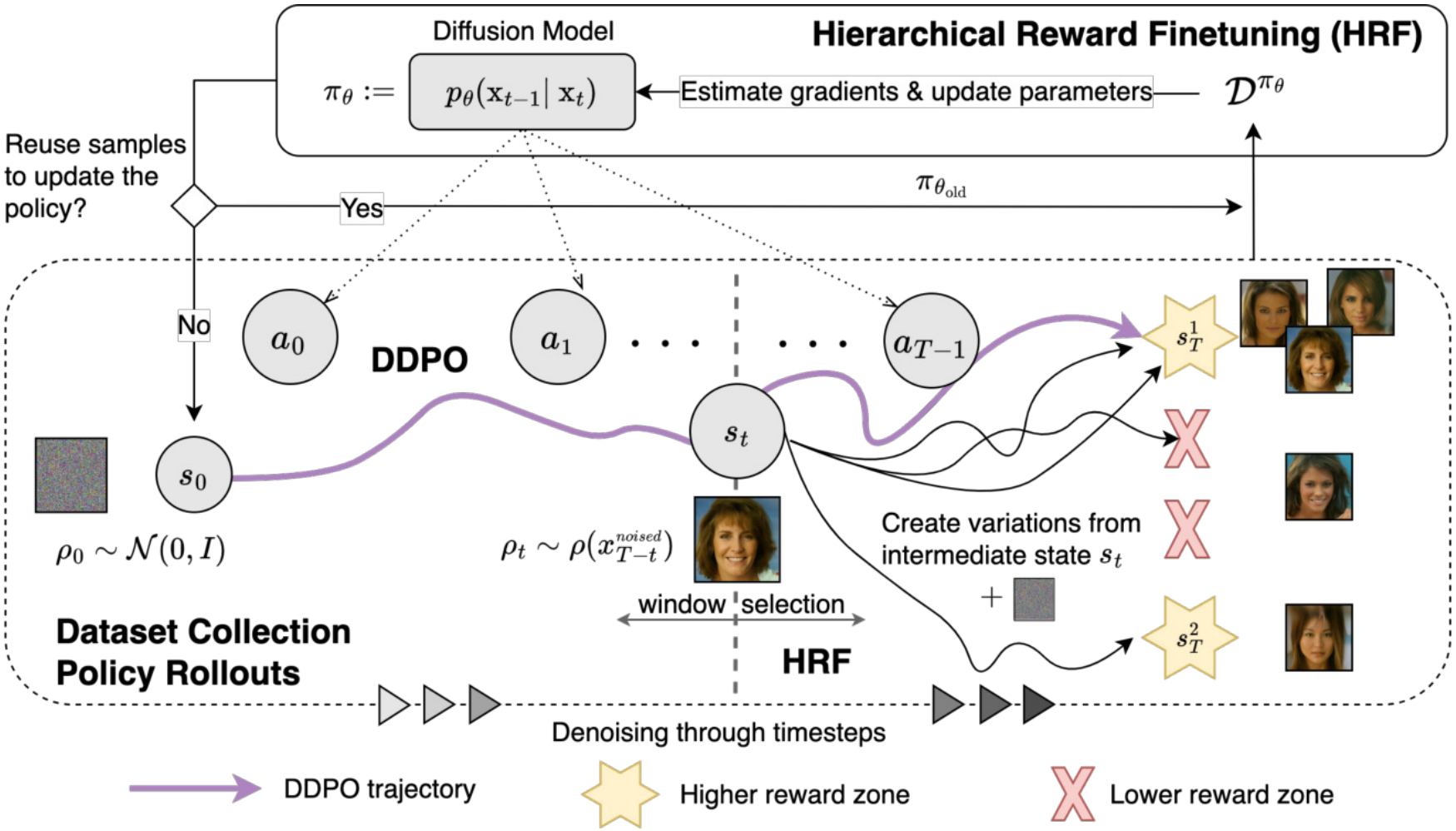


Overoptimization leads to *mode collapse*

1. Starting from the same initialization, both the pretrained and finetuned models on aesthetic quality.
2. 2D projection of CLIP embedding space shows both distributions.
3. Excessive reward alignment results in a loss of diversity (red ellipse).



Hierarchical Reward Finetuning (HRF)



Avoiding Mode Collapse in Diffusion Models Fine-tuned with Reinforcement Learning (Barceló, Alcázar, Tobar, 2024 [1])

Conclusions

1. The hypothesis of the effectiveness of using RL to finetune pretrained diffusion models on new tasks was validated, specifically using the DDPO methodology on two DDPM models: one for face generation and the other for churches.
2. Experiments were conducted on tasks from Black 2023's work [2]; and an additional task was proposed and implemented for generating faces of people OVER50, using a reward based on an age classifier.
3. A repository was created with the implementation, simplifying the model architecture used in DDPO, reducing cognitive overhead, and enabling the exploration of new ideas with lower VRAM usage.
4. Checkpoints from the trained tasks were made available in Hugging Face repositories.

References I

[1] Roberto Barceló, Cristóbal Alcázar, and Felipe Tobar. *Avoiding mode collapse in diffusion models fine-tuned with reinforcement learning*. 2024. arXiv: 2410.08315 [stat.ML]. URL: https://arxiv.org/abs/2410.08315.

[2] Kevin Black et al. "Training diffusion models with reinforcement learning". In: *arXiv preprint arXiv:2305.13301* (2023).

[3] Yu Qing Du. "Human-Centric Reward Design". PhD thesis. EECS Department, University of California, Berkeley, Nov. 2023. URL: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2023/EECS-2023-231.html.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. arXiv: 2006.11239 [cs.LG].

[5] Calvin Luo. *Understanding Diffusion Models: A Unified Perspective*. 2022. arXiv: 2208.11970 [cs.LG].

References II

[6] Christoph Schuhmann. "LAION-Aesthetics V2". In: (2022). URL: https://laion.ai/blog/laion-aesthetics/.

[7] Ronald J Williams. "Simple statistical gradient-following algorithms for connectionist reinforcement learning". In: *Machine learning* 8 (1992), pp. 229–256.