Online RLHF for Diffusion Model

Vincent Tianyu Zhan
tianyu.zhan@nyu.edu

Abstract

Current visual generative models, while achieving remarkable progress through diffusion and flow matching frameworks, face persistent challenges in artifact reduction and prompt-image alignment. Existing solutions predominantly rely on offline preference optimization methods that suffer from distributional shift and policy degradation. This work proposes a paradigm shift toward online learning strategies for diffusion models and rectified flow architectures. Our approach aims to bridge the gap between offline contrastive training method and exploratory online method in text-to-image generation.

Advancements and Challenges

The field has evolved through three key technical trajectories. Supervised fine-tuning methods using parameter-efficient adaptation (e.g., LoRA) demonstrate effectiveness in low-data regimes but fail to scale effectively with increasing task complexity. Reinforcement learning approaches like DDPO[1] and DPOK[5] formulate generation as multi-step Markov Decision Processes, yet the performance of these methods degrades as the number of train/test prompts increases.˙

Preference-based optimization represents a breakthrough, with Diffusion-DPO[13] establishing direct alignment tuning without explicit reward modeling. Subsequent innovations like MaPO's margin-based objectives[6] and SPO's step-wise optimization[8] enhance temporal consistency and human preference matching. However, these offline methods fundamentally lack mechanisms for continuous policy improvement, relying instead on static offline datasets that perpetuate suboptimal trajectories[11] and performs poorly in OOD test sets[14].

Recent findings in language model alignment demonstrate the necessity of on-policy sampling for high-quality policy updates[12], suggesting similar principles may govern visual generation. Concurrently, the emergence of rectified flow frameworks[9, 10] presents new opportunities through their straight probability trajectories and single-step sampling capabilities. While implementations like SD3 and etc. [4, 7, 3] showcase competitive performance, the preference optimization community remains disproportionately focused on diffusion architectures. As a result, we aim to investigate whether DPO-based alignment methods can be effectively extended to rectified flow and other flow matching paradigms and how can online contrastive method imporove the post-training of Diffusion Models.

Proposed Method

Building upon the denoising diffusion models, we establish the discrete-time reverse process as:

$$p_\theta(x_t|x_{t-1}) = \mu_\theta(x_t) + \sigma^2_{t|t-1}\frac{\sigma^2_{t-1}}{\sigma^2_t}\epsilon_\theta \tag{1}$$

where ϵ_θ is a standard gaussian distribution. Unlike conventional RLHF approaches that use π for policy notation, we adopt p to emphasize the probabilistic nature of diffusion models. Accordingly, q is the forward process.

Extending Diffusion-DPO's core insight [13], we reformulate the alignment objective through an online learning lens. The baseline DPO loss becomes:

$$\mathcal{L}_{\text{Diffusion-DPO}}(\theta) = -\mathbb{E}_{\substack{(x_0^w,x_0^t)\sim\mathcal{D},t\sim U(0,T)\\ x_t^w\sim q(x_t^w|x_0^w),x_t^t\sim q(x_t^t|x_0^t)}} \log\sigma\left(-\beta T\left(s(c,x_t^w)-s(c,x_t^l)\right)\right) \tag{2}$$

Algorithm 1 EXPO Training Pipeline

Require: Reward Model r^*, Offline Dataset $\mathcal{D}_{\text{off}}$, Pre-trained Model p_θ, Prompt Set $\mathcal{D}_{\mathcal{C}}$, Negative Prompt Set $\mathcal{D}_{\mathcal{N}}$, Images per Prompt B, Total Training Iterations T, Number of Prompts per Iteration N

for $t = 1$ to T **do**
  $\mathcal{D}_{\text{hybrid}} \leftarrow \mathcal{D}_{\text{off}} \cup \mathcal{D}_{1:t-1}$
  Policy Update: ▷ Exploit historical data
  Update policy parameters by optimizing loss function $\mathcal{L}_{\text{EXPO}}$: $p_{\text{ex}} \leftarrow p_t$
  $p_{\text{main}} \leftarrow p_{t-1}$
  if $t < T$ **then** ▷ Explore new regions
    for each prompt $\mathbf{c}^n \sim \mathcal{D}_{\mathcal{C}}$ **do**
      Sample negative prompt $\mathbf{c}_{\text{neg}} \sim \rho(t, \mathcal{D}_{\mathcal{N}})$
      (Optional)Augmented prompt: $\mathbf{c} \leftarrow \mathbf{c}^n + \mathbf{c}_{\text{neg}}$
      $\mathcal{D}_{\text{main}} \leftarrow \{(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_B) \sim p_{\text{main}}(\cdot|\mathbf{c})\}$
      $\mathcal{D}_{\text{ex}} \leftarrow \{(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_B) \sim p_{\text{ex}}(\cdot|\mathbf{c})\}$
      Collect data $\mathcal{D}_t \leftarrow \mathcal{D}_{\text{main}} \cup \mathcal{D}_{\text{ex}}$
    end for
  end if
end for

where the prediction error differential $s(c, x_t) = \|\epsilon - \epsilon_{\text{ex}}(x_t)\|_2^2 - \|\epsilon - \epsilon_{\text{main}}(x_t)\|_2^2$ captures the implicit reward learned by the model.

With this offline training objective, we begin by following the iterative RLHF design from [14] to develop an active online exploration algorithm using the West-of-n method combined with hybrid training. As shown in algorithm 1, we define the reference policy as the initial policy and train the model for one iteration to assimilate information from the offline dataset. Subsequently, we update the main agent and the exploration agent and use them to generate new online data, iterating this process to progressively improve the model.

Recent work in LLMs [2, 11] has demonstrated promising advances with active exploration mechanisms—many of which augment the DPO loss with a reward-based optimistic bonus to foster exploration. Inspired by these developments, we propose our preference-incentive optimization objective:

$$L_{\text{EXPO}}(\theta) = -\mathbb{E}_{\substack{(x_0^w, x_0^l)\sim\mathcal{D},\, t\sim U(0,T) \\ x_t^w \sim q(x_t^w|x_0^w),\, x_t^l \sim q(x_t^l|x_0^l)}} \log\sigma\Big(-\beta T\Big(s(c, x_t^w) - s(c, x_t^l)\Big)\Big) + \alpha\, \mathbb{E}_{x_0\sim\mathcal{D}_{\text{ex}},\, x_0'\sim\mathcal{D}_{\text{main}}}\Big[\sigma\Big(\beta\big(s(c, x_0) - s(c, x_0')\big)\Big)\Big]. \quad (3)$$

Here, α is the exploration temperature, and the exploration term is designed to encourage the exploration policy to discover more favorable outcomes compared to the reference policy.

Experimental Design and Evaluation

In this work, we aim to answer the following four research questions:

1. How can online exploration benefit offline training procedures, such as DPO training for diffusion models or flow matching?
2. Which dataset type is more effective: off-policy or on-policy?
3. Does the inclusion of an active exploration term genuinely enhance exploration?
4. Is it possible to train the model without relying on an external reward model?

To address these questions, we have designed a comprehensive experimental framework. We plan to use the Pick-A-Pic and HPDv2 datasets as our offline data sources, while iteratively generating new on-policy data during training. Our evaluation employs several reward models and benchmarks, including HPSv2, PickScore, CLIP, and Aesthetic, as well as external benchmarks such as text-to-image comparisons.

The experiments are structured as follows:

- **Experiment 1:** Train the model for 5–7 iterations and evaluate performance improvements using the metrics mentioned above. (Done, it improves the performance by a great margin)
- **Experiment 2:** Compare model performance when trained on off-policy data versus on-policy data.
- **Experiment 3:** Assess the impact of the active exploration term by comparing models trained with and without it.
- **Experiment 4:** Evaluate the feasibility of training the model using SD1.5 as the reward model.

Conclusion

This proposal introduces an online exploration framework for visual generative models that leverages active learning and rectified flow architectures. Our approach aims to overcome offline training limitations, improving prompt-image alignment and reducing artifacts. Planned experiments will validate the benefits of on-policy data and active exploration, paving the way for more adaptive generative systems.

References

[1] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

[2] Mingyu Chen, Yiding Chen, Wen Sun, and Xuezhou Zhang. Avoiding $\exp(r_{\max})$ scaling in rlhf through preference-based exploration. *arXiv preprint arXiv:2502.00666*, 2025.

[3] Shoufa Chen, Chongjian Ge, Yuqi Zhang, Yida Zhang, Fengda Zhu, Hao Yang, Hongxiang Hao, Hui Wu, Zhichao Lai, Yifei Hu, et al. Goku: Flow based video generative foundation models. *arXiv preprint arXiv:2502.04896*, 2025.

[4] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

[5] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

[6] Jiwoo Hong, Sayak Paul, Noah Lee, Kashif Rasul, James Thorne, and Jongheon Jeong. Margin-aware preference optimization for aligning diffusion models without reference. *arXiv preprint arXiv:2406.06424*, 2024.

[7] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024.

[8] Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Ji Li, and Liang Zheng. Step-aware preference optimization: Aligning preference with denoising performance at each step. *arXiv preprint arXiv:2406.04314*, 2024.

[9] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[10] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[11] Yuda Song, Gokul Swamy, Aarti Singh, Drew Bagnell, and Wen Sun. The importance of online data: Understanding preference fine-tuning via coverage. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[12] Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, Rémi Munos, Bernardo Ávila Pires, Michal Valko, Yong Cheng, et al. Understanding the performance gap between online and offline alignment algorithms, 2024. *URL https://arxiv. org/abs/2405.08448*, 2024.

[13] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024.

[14] Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. In *Forty-first International Conference on Machine Learning*, 2024.