RPO: Fine-Tuning Visual Generative Models via Rich Vision-Language Preferences

Hanyang Zhao[1,*] [1]Columbia University
Haoxian Chen[2,*,†] [2]Amazon
Yucheng Guo[3,*] [3]Princeton University
Genta Indra Winata[4,‡] [4]Capital One
Tingting Ou[1] [1]Columbia University
Ziyu Huang[1] [1]Columbia University
David D. Yao[1] [1]Columbia University
Wenpin Tang[1] [1]Columbia University

arXiv:2503.11720v5 [cs.LG] 10 Jan 2026

Abstract

Traditional preference tuning methods for LLMs/Visual Generative Models often rely solely on reward model labeling, which can be opaque, offer limited insights into the rationale behind preferences, and are prone to issues such as reward hacking or overfitting. We introduce Rich Preference Optimization (RPO), a novel pipeline that leverages rich feedback signals from Vision Language Models (VLMs) to improve the curation of preference pairs for fine-tuning visual generative models like text-to-image diffusion models. Our approach begins with prompting VLMs to generate detailed critiques of synthesized images, from which we further prompt VLMs to extract reliable and actionable image editing instructions. By implementing these instructions, we create refined images, resulting in synthetic, informative preference pairs that serve as enhanced tuning datasets. We demonstrate the effectiveness of our pipeline and the resulting datasets in fine-tuning state-of-the-art diffusion models.

1 Introduction

Learning from feedback and critiques is essential for enhancing the performance of a generative model by guiding the model to rectify unsatisfactory outputs, e.g. in RLHF (Ouyang et al., 2022). However, improvements can arise not only from distinguishing right from wrong, but also from receiving thoughtful and informative feedback that offers clear direction for enhancement, which we refer as *rich feedback*. For instance, in the natural language tasks, rich feedback has proved to be useful in LLM self-refinement (Saunders et al., 2022), reward model enhancement (Ye et al., 2024; Ankner et al., 2024), code debugging (Chen et al., 2023; McAleese et al., 2024), games (Paglieri et al., 2024; Hudi et al., 2025), and agents (Shinn et al., 2023; Wang et al., 2023; Renze and Guven, 2024). Rich feedback is comparatively less explored in vision tasks, such as human feedback generation model (Liang et al., 2024) and visual commonsense reasoning (Cheng et al., 2024; Chen et al., 2024; Li et al., 2024b).

To effectively leverage rich feedback in model training, it is crucial to ensure that the rich feedback is detailed, informative, and nuanced. Simply relying on numerical scores, as in classical reward models e.g. PairRM (Jiang et al., 2023; Kirstain et al., 2023a), falls short in identifying specific areas for model enhancement. Comprehensive feedback like heat maps of unaligned parts in (Liang et al., 2024) provides insights that extend beyond numerical evaluations, allowing for more targeted and substantial model improvements. In this context, exploring the use of critic models can be highly beneficial, as these models are growing rapidly and could offer deeper insights into the model's intricacies, thus contribute to a more robust understanding of improvement opportunities.

To leverage rich feedback in preference learning, we draw inspiration from the way students learn from their teachers. We introduce **R**ich **P**reference **O**ptimization **(RPO)**, a novel approach designed to enhance preference learning of images by letting the model learn from **rich preferences** that we curate with the aid of vision-language models (VLMs). These multi-modal models provide detailed critiques for downstream tasks need, offering rich feedback that mirrors the comprehensive guidance students receive in modern educational systems. Rather than merely receiving a final score on assignments or exams, students are given specific feedback that identifies logical errors, misunderstandings, or calculation mistakes. This feedback enables students to iteratively refine their initial responses, fostering deeper learning through a process of continuous improvement. Similarly, in RPO, we extract actionable editing instructions from VLMs and employ instruction-driven image-editing models for refinement. This

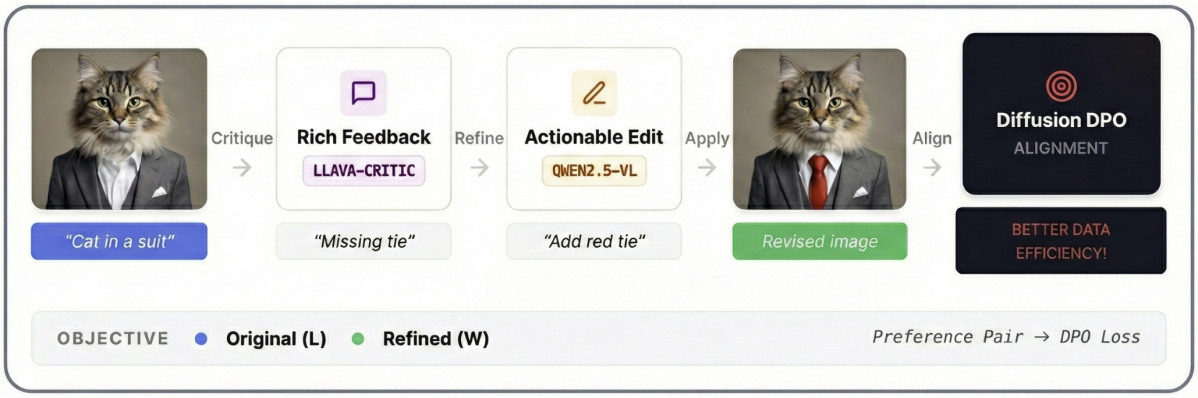


Figure 1: Our RPO pipeline for curating informative preference pairs from images generated from the base diffusion models: **(1)** Rich Feedback/Critic generation by a VLM (for which we choose LLaVA-Critic-7B), **(2)** Actionable editing instruction generation based on the critiques by another VLM (for which we chose Qwen2.5-VL-8B-Instruct), **(3)** Instruction-following image editing from the generated editing instructions (for which we choose ControlNet), and **(4)** Diffusion DPO training using reward model filtered preference pairs.

scalable method generates informative preference pairs that are crucial for effective preference learning. The process of receiving detailed feedback and making refinements is akin to learning from true preference pairs, reflecting the natural and effective way in which humans learn. The contribution of our paper is summarized as follows:

(1) We introduce RPO, a novel approach for generating preference datasets for images by leveraging VLMs to provide detailed critiques as rich feedback about misalignment between prompt and generated image. We further extract actionable editing instructions from VLMs, and employ instruction-driven image-editing models for refinement. This scalable data curation approach yields informative preference pairs, which we illustrate in Figure 1.

(2) We show that the intermediate critiques can improve the quality of editing instructions than directly generating them, which is reminiscent to the Chain-of-Thought (Wei et al., 2022) concept in mathematical reasoning for LLMs. We also propose to use ControlNet for image-editing by conditioning on the original image and combine prompts with editing instructions, which ensures fine-grained control over the image to be revised while keeping most of the image unchanged.

(3) We also empirically demonstrate that our curated preferences are scalable and useful offline synthetic preference dataset, by further training Diffusion DPO (trained on the original offline dataset) checkpoints on rich preferences we generate on the preferred images. We observe drastic performance gain, which validates the effectiveness of our pipeline.

Related Work. In contemporary RLHF pipelines for LLMs, preference pairs are generated by sampling various responses and subsequently ranking them using either human evaluators or pre-trained reward models, which serve as AI-based labels. This approach is widely utilized in both online reinforcement learning (RL)-based techniques (Ouyang et al., 2022; Bai et al., 2022) and offline methods for LLM preference learning, such as DPO (Rafailov et al., 2023), SimPO (Meng et al., 2025), and RainbowPO (Zhao et al., 2025a), which motivates Diffusion-DPO (Wallace et al., 2024) for diffusion models. See Winata et al. (2024) for a comprehensive review.

Concurrent work to ours, Wang et al. (2025) also investigate on utilizing dedicated feedback and then passing to the editing model in the contents of an inference time scaling method. As a comparison, we focus on alignment and our pipeline has an extra procedure of generating concrete editing instructions before passing to instruction-following editing models, which directly connects two well developed literature of VLMs and Image-editing models. Our pipeline is also easier for plug-in thanks to the recent rapid progress in instruction-following image-editing capability like in GPT-4o (Hurst et al., 2024). Our work is also reminiscent to Aligner (Ji et al., 2024) for pure language setup.

2 Preliminaries

We discussion some preliminaries about visual generative models like diffusion models, VLMs and relevant image-editing models.

Diffusion Models. Diffusion models are a class of generative models $p_\theta(x_0)$, whose goal is to turn noisy/non-informative initial distribution $p_{\text{noise}}(x_T)$ to a desired target distribution $p_{\text{tar}}(x_0)$ through a well-designed denoising process (Ho et al., 2020; Song et al., 2021a,b). Given noise scheduling functions α_t and σ_t (as defined in (Rombach et al., 2022)), the forward process is specified by $q(x_t \mid x_s) = \mathcal{N}\left(\frac{\alpha_t}{\alpha_s} x_s,\ \left(\sigma_t^2 - \frac{\alpha_t^2}{\alpha_s^2}\sigma_s^2\right) I\right)$ for $s < t$. A diffusion model is trained by minimizing the evidence lower bound (ELBO):

$$\min_\theta \mathcal{L}(\theta) := \mathbb{E}_{t,\epsilon,x_0,x_t}\left[\omega\left(\lambda_t\right)||\epsilon - \epsilon_\theta(t, x_t)||_2^2\right], \tag{1}$$

where $t \sim \mathcal{U}(0, T)$, $\epsilon \sim \mathcal{N}(0, I)$, $x_0 \sim p_{\text{tar}}(x_0)$, $x_t \sim q(x_t \mid x_0) = \mathcal{N}(\alpha_t x_0,\ \sigma_t^2 I)$, $\lambda_t := \frac{\alpha_t^2}{\sigma_t^2}$ is the signal-to-noise ratio, and $\omega : \mathbb{R}_+ \to \mathbb{R}_+$ is some weight function.

Rich Feedback. A good critic allows the recipient to learn and improve from the feedback. In T2I generation, RichHF (Liang et al., 2024) identifies misalignment in a multimodal instruction (i.e., an image-prompt pair), and hence, enriches the feedback. A by-product is the Rich Human Feedback dataset (RichHF-18k), consisting of fine-grained scores, and misalignment image regions and text descriptions on 18K Pick-a-Pic images (Kirstain et al., 2023b). However, the Rich Feedback model and code are both not released.

As an alternative to Rich Feedback, LLaVA-Critic (Xiong et al., 2024) is an open-source VLM that is primarily developed to give evaluation of multimodal tasks. e.g., VLM-as-a-judge and preference learning. It shows a high correlation and comparable performance to proprietary GPT models (GPT-4V/4o). In our approach, we use LLaVA-Critic as an VLM-as-a-judge: the input is a text-image pair, and the output is a critic to image-prompt misalignment. Following the Chain-of-Thought concept (Wei et al., 2022), such obtained critic will be subsequently passed to an open-source LLM to provide an editing instruction. Our experiment shows that the proposed open-source critic + editing pipeline yields more reliable improvements than directly querying a VLM, e.g., GPT4o, for editing instruction generation: see Section 3.2 for examples.

Instruction-following Image Editing Models. The image-editing part in our pipeline is related to the literature of instruction-following image-editing models. We focus on diffusion-based models in this paper due to both their advantage over autoregressive models, and that our base models for fine-tuning are also diffusion-based. The pioneer models include InstructPix2Pix (IP2P) (Brooks et al., 2023), which first enables editing from instructions that inform the model which action to perform. Follow-up works include Magic Brush (Zhang et al., 2023a), Emu Edit (Sheynin et al., 2024), UltraEdit (Zhao et al., 2025b) and HQEdit (Hui et al., 2024) by introducing additional datasets for further fine-tuning based on IP2P and enhancing the performance. Existing work like HIVE (Zhang et al., 2024) has also considered to align IP2P with human feedback to enhance generation capability.

As previously mentioned, we edit images based on ControlNet (Zhang et al., 2023b) to ensure a better coherence to the original image. ControlNet has been widely used for controlling image diffusion models by conditioning the model with an additional input image. Further applications include implementations for the state-of-the-art proprietary models like Stable Diffusion 3, Stable Diffusion XL (Esser et al., 2024) and FLUX (Labs, 2024), multi-image support extensions like ControlNet++ (Li et al., 2024a). We will stick to the original ControlNet in this paper (because the offline dataset is generated by the similar model scaled by Stable Diffusion v1.5), and leave the usage of more advanced ControlNet variants in future work.

Diffusion-DPO. Direct Preference Optimization (DPO) (Rafailov et al., 2023) is an effective approach for learning from human preference for language models. (Wallace et al., 2024) proposed Diffusion-DPO, a method to align diffusion models to human preferences by directly optimizing on human comparison data (x_0^w, x_0^l) given the conditional input c (the prompt). Let $\beta > 0$ be the regularization parameter, and $p_{\text{ref}}(x_{0:T}|c)$ be a (pre-trained) reference model. The DPO loss objective for diffusion models can be written as:

$$L_{\text{DPO}}(\theta) = -\mathbb{E}_{c,x_0^w,x_0^l} \log \sigma(\beta \mathbb{E}_{\substack{x_{1:T}^w \sim p_\theta(\cdot|x_0^w,c) \\ x_{1:T}^l \sim p_\theta(\cdot|x_0^l,c)}} \left[\log \frac{p_\theta(x_{0:T}^w|c)}{p_{\text{ref}}(x_{0:T}^w|c)} - \log \frac{p_\theta(x_{0:T}^l|c)}{p_{\text{ref}}(x_{0:T}^l|c)}\right]) \tag{2}$$

3 Curating Preference Pairs with Rich Feedback Signals

In this section, we present the concrete components in our pipeline for creating preference pairs. We utilize 1.6k rows of prompts and images from the test set of RichHF dataset provided by (Liang et al., 2024) as the validation set for ablation study. We first discuss instruction-following editing (despite it being the last part before preference tuning in our pipeline), and then use the best instruction-following image-editing model we find for further ablation study on the first two components: utilizing multimodal models/VLMs for generating feedback information, and providing concise and actionable editing instructions.

3.1 Instruction-Following Image Editing

Despite the numerous proposed image-editing methods, either based on diffusion models or not in the literature, we find that they struggle to perform fine-grained control to follow specific editing instructions, which is crucial to generating images that are direct improvements over the existing ones. Existing methods usually change the image to another one which may yield higher score but looks fundamentally different.

To tackle the above issue, we propose a ControlNet (Zhang et al., 2023b) based image-editing approach. Concretely, we exploit the image2image (pipeline) of the ControlNet by setting the conditional image to be the same as the original one, which guarantees that the changes adhere to the original image. In addition, we concatenate the prompt that generates the image with our generated editing instructions, which we find will greatly enhance the quality of the edited image generation as illustrated in Appendix B.

Additionally, we quantitatively assess the instruction-following capabilities of various diffusion-based image-editing models. We also examine the impact of incorporating prompts before editing instructions, using GPT-4o for pairwise comparisons on both a standard instruction-following dataset and a test set. Our findings indicate that ControlNet-based editing is particularly effective. However, we highlight that the choice of image-editing model is flexible and can be updated to include more advanced options, such as ControlNet SDXL and other variants. We plan to explore these options in future work.

3.2 Generating Rich Feedback Signals

Unlike RichFB (Liang et al., 2024) which requires to train a model to detect the heatmaps and misaligned words thus providing more accurate rewards, we leverage the textual feedback on the quality of generation, like a movie critic, as in Figure 2.

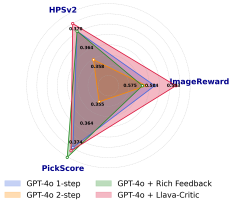


Figure 2: Comparison of RichFB generated informative feedback and our adopted textual criticism generate by carefully prompting a capable VLM.

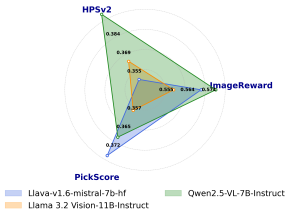
To compare and ablate about what type of feedback will be more useful for yielding better editing instructions for improving the images, we prompt GPT-4o to generate them based on the given feedback, and then compute the rewards of new images (evaluated by different reward models) after using ControlNet to edit the image with the instructions as in Figure 3a. Concretely, we let GPT-4o generate editing instructions based on various types of feedback, including those from RichFB, LLaVA-Critic, and even GPT-4o itself. From the RichFB dataset, we use the image, prompt, misalignment information within the prompt, and the misalignment heatmap of the image as inputs. From LLaVA-Critic, we incorporate its textual feedback. For GPT-4o, we first prompt it to also generate textual feedback based on the image-prompt pair, which is then used to derive editing instructions. For studing the necessity of rich feedback, we also let GPT 4o generate editing instructions directly from the input image and prompt, bypassing the intermediate step of critique.

Somewhat surprisingly, as in Figure 3a, we found that GPT-4o produced better editing instructions when generating them directly, rather than conditioning on its own critiques about the misalignment between the prompt and the image. However, GPT-4o generates much better editing instructions (measured by average reward) conditioning on the critiques provided by LLaVA-Critic. Based on this result, we adopt LLaVA-Critic as the feed-

back generator in our RPO pipeline.



(a) Comparison of RichFB generated informative feedback and our adopted textual criticism generated by carefully prompting a capable VLM.



(b) Comparison of open-source VLMs combined with LLaVA-Critic in image editing quality, showing that Qwen2.5-VL-7B-Instruct is the most capable VLM model.

Figure 3: Comparisons of feedback approaches and VLM performance for enhanced image editing, evaluated by ImageReward, HPSv2 and PickScore.

Notably, here we implicitly make a hypothesis that refined images of higher reward yield better preference pairs, which also makes the corresponding VLM more preferrable. We verify this hypothesis in the experiment section 4.4 by directly comparing the trained models upon these different preference pairs, which showcases that this intuition is indeed correct.

3.3 Generating Editing Instructions

We also compare the performance of Llama 3.2 Vision 11B Instruct (Grattafiori et al., 2024), Llava-v1.6 Mistral 7B (Liu et al., 2024), and Qwen2.5 VL 7B Instruct (Wang et al., 2024) generated editing instructions, in combined with LLaVA-Critic and ControlNet based editing as we argued earlier. As shown in Figure 3b, Qwen2.5-VL-7B-Instruct yield the best results in leading to highest rewards in both HPSv2 and ImageReward after adopting the image-editing instructions. We thus choose Qwen2.5-VL-7B-Instruct as the VLM for editing instruction generation in our RPO pipeline.

3.4 Reward Model Relabeling

After obtaining the pair of original and edited images, we rearrange them into preferences by further querying a reward model or an LLM-as-a-judge. On the test set of size 16K, we find that roughly 60% of our images yield a higher score than the original image under the ImageReward (Xu et al., 2023) metric. As a remark, it is possible to use RL fine-tuning methods to encourage the VLM to generate better editing instructions, which may get a higher score of edited images; we leave this for future work. Nevertheless, the edited images that fail to have higher scores than the original ones can still serve as the non-preferred images, and hence, yield the flipped preference pair.

To conclude, our RPO pipeline provides a generic and training-model-free way to generate preference pairs, as we do not need extra generations from the base model. We further use our curated dataset for fine-tuning large-scale SOTA diffusion models which validates the effectiveness of our pipeline.

4 Experiments

In this section, we evaluate our RPO pipeline by first fine-tuning different baseline models on our synthetic preferences data using the Diffusion-DPO algorithm, and then utilizing a couple of reward models to score the images generated by the fine-tuned models.

4.1 Settings

Baseline Models. We use SD1.5 (Rombach et al., 2022) and SDXL-1.0 (Podell et al., 2024) as our starting point. We create the following checkpoint models by Diffusion-DPO fine-tuning the two base models on the Pick-a-Pic v2 dataset: (a) DPO-SD1.5-100k, which is fine-tuned from SD1.5 using the first 100k rows; (b) DPO-SD1.5-200k, fine-tuned from SD1.5 using the first 200k rows; (c) DPO-SD1.5-100k (ImageReward-Aligned), fine-tuned from SD1.5 using the first 100k rows, with the preference modified by the relative order of ImageReward scores; (d) DPO-SDXL-100k, fine-tuned from SDXL-1.0 using the first 100k rows.

Produced Models. To evaluate our curated dataset, we produce the following models by Diffusion-DPO fine-tuning the baseline models using our 100k synthetic preferences data: (a) DPO-SD1.5-100k+RPO100k, which is fine-tuned from model DPO-SD1.5-100k; (b) DPO-SD1.5-100k (ImageReward-Aligned)+RPO100k, fine-tuned from model DPO-SD1.5-100k (ImageReward-Aligned); (c) DPO-SDXL-100k + RPO100k, fine-tuned from model DPO-SD1.5-100k (ImageReward-Aligned).

Diffusion-DPO Training. For Diffusion-DPO training, we follow the same setting and use the same hyperparameters in (Wallace et al., 2024) when fine-tuning SD1.5-based models. More specifically, we use AdamW (Loshchilov and Hutter, 2019) as the optimizer; the effective batch size

is set to 2048; we train at fixed square resolutions, and use a learning rate of $\frac{2000}{\beta} 2.048 \cdot 10^{-8}$ with 25% linear warmup; for the divergence penalty parameter, we keep $\beta = 5000$. When fine-tuning SDXL-based models, we use the LoRA implementation provided by the Github project (Suzukimain and Liu, 2024) to improve training efficiency.

Evaluation. We generate images using the prompts from Pick-a-Pic test set (Kirstain et al., 2023a) (which contains 500 unique prompts) and evaluate the generation with reward models including PickScore (Kirstain et al., 2023a), ImageReward (Xu et al., 2023), HPSv2 (Wu et al., 2023) and LAION-Aesthetic Predictor (a ViT-L/14 CLIP model trained with SAC dataset (Pressman et al., 2022)). The PickScore, ImageReward, HPSv2 reward models are used to evaluate human-preference alignment, and Aesthetic Score is expected to evaluate visual aesthetic appeal. For each model, we report average scores over all prompts.

4.2 Improving Diffusion-DPO via RPO

We present the performance of all models evaluated by different reward models in Table 1. Above all, models obtained by fine-tuning on our 100k synthetic preferences data outperforms the baseline models by all metrics. This confirms our hypothesis that the pipeline we designed for synthesizing preference pairs yields improved results when combined with preference-based training algorithms such as Diffusion-DPO.

Concretely, when utilizing the *same* amount of vanilla preference data (synthetic data for RPO is curated from this same vanilla preference data), our RPO pipeline stably improves the DPO baseline over all metrics, see DPO-SD1.5-100k and DPO-SD1.5-100k+RPO100k. Even comparing with the model DPO-SD1.5-200k which utilized twice the additional vanilla preference data, our RPO fine-tuned model still performs better across all metrics. This demonstrates that our pipeline is significantly more data efficient for Diffusion-DPO style of training than just on vanilla preference data, achieving more than 100% gain over the standard practice. We have thus verified the effectiveness and potential of the synthetic preferences data generated by rich feedback to augment the given (high-quality) dataset, which is of crucial importance when human annotated data is limited (which we believe is usually always the case in practice).

We also present a qualitative comparison of the three SDXL-based models in Appendix Sec C.1 (see Figure 8). Beyond generating images with better visual qualities, we remark that the RPO fine-tuned model is better at connecting different elements in the prompts in a deep and profound way (see the first and the third columns). Understanding how rich feedback and guided revision help fine-tuning diffusion models will be left for future work.

4.3 Data Scaling and Generalization Analysis

We evaluate the robustness of our method by comparing performance across training scales ranging from 5K to 100K samples. As illustrated in Figure 5, RPO-enhanced models consistently outperform the Diffusion-DPO baseline across all preference set sizes, with the performance margin widening as the training volume increases. Most notably, the model fine-tuned on only 100K RPO-enriched pairs outperforms the standard Diffusion-DPO model trained on nearly 1M pairs (denoted by stars in Figure 5) across all major metrics.

This order-of-magnitude improvement in data efficiency highlights RPO's ability to achieve superior performance without additional human labeling. In contrast, we observe that the 1M-sample baseline exhibits inconsistent performance across reward metrics, such as lower LAION-Aesthetic scores on PartiPrompts despite the increased data volume, suggesting that it overfits to narrow preference signals rather than learning generalized visual alignment. Furthermore, RPO demonstrates strong generalization capabilities on out-of-distribution benchmarks, specifically PartiPrompts and HPSv2 (see additional results in Appendix E, Figure 13). These gains are consistent across diverse visual styles, confirming that the benefits of learning from rich feedback generalize effectively across novel prompt and stylistic distributions.

4.4 Direct Abalations on VLM Critics

To assess the impact of different VLM critics within our RPO pipeline, we conducted a controlled ablation study in which we varied only the critic component while keeping the rest of the training pipeline fixed. Specifically, we used 1.4k prompts to generate preference pairs based on different VLM critics, including GPT-4o (1-step and 2-step), RichFeedback, and Llava-critic. We then fine-tuned models using these preference datasets and evaluated their performance across four metrics: PickScore, ImageReward, Aesthetic, and HPSv2. As summarized in Table 2, Llava-critic consistently achieved the

Model	PickScore	ImageReward	Aesthetic	HPSv2
SD1.5	20.33	0.1733	5.949	0.2622
DPO-SD1.5-100k	20.66	0.2784	6.044	0.2650
DPO-SD1.5-200k	20.74	0.3638	6.088	0.2657
DPO-SD1.5-100k + RPO100k	**20.75**	**0.4395**	**6.113**	**0.2663**
DPO-SD1.5-100k (ImageReward-Aligned)	20.66	0.600	6.083	0.268
DPO-SD1.5-200k (ImageReward-Aligned)	20.63	0.606	6.109	0.268
DPO-SD1.5-100k (ImageReward-Aligned) + RPO100k	**20.72**	**0.686**	**6.213**	**0.269**
SDXL1.0	21.74	0.8473	6.551	0.2692
DPO-SDXL-100k	**21.92**	0.9183	6.566	0.2706
DPO-SDXL-100k + RPO100k	21.89	**0.9353**	**6.585**	**0.2707**

Table 1: Model performance on the Pick-a-Pic Test Set by four different metrics.

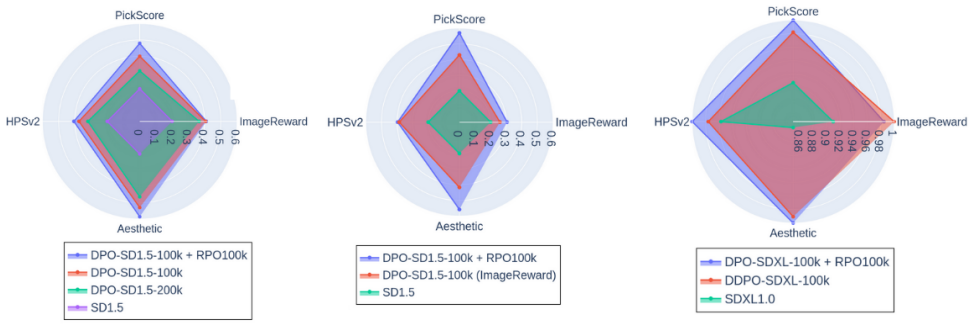


Figure 4: Model performance evaluated by PickScore, ImageReward, Aesthetic, and HPSv2.

best overall performance, with the highest scores in PickScore, ImageReward, and HPSv2. This trend is further confirmed by the radar plot in Figure 6, where Llava-critic demonstrates the most balanced and elevated performance profile. These results validate our design choice of using Llava-critic as the default VLM for generating rich feedback, given its ability to produce high-quality preference data that leads to more effective RPO fine-tuning.

Model	PickScore	ImageReward	Aesthetic	HPSv2
GPT-4o-1-step	20.37	0.0684	6.072	0.2629
GPT-4o-2-step	20.41	0.1675	**6.133**	0.2635
RichFeedback	20.46	0.1634	6.076	0.2639
Llava-critic	**20.47**	**0.2209**	6.039	**0.2640**

Table 2: RPO VLM critic ablation via training on 1.4k preference pairs.

4.5 Effectiveness of Critic-Instruction-Editing Data Curation and RPO Fine-Tuning

To evaluate the effectiveness of our curated preference data and the resulting RPO fine-tuned model, we compared images generated by the model against the ControlNet-edited images used during training, using multiple reward metrics (Table 3). The ControlNet-edited images consistently achieve higher scores, demonstrating the high quality of our training targets. Despite this, the RPO fine-tuned model yields competitive performance across all metrics, confirming the value of preference-based fine-tuning via Diffusion-DPO. More importantly, the RPO model eliminates the need for a separate critic and editing module at inference time, resulting in substantial savings in computational and memory resources. These results underscore both the quality of our critic-instruction-editing pipeline and the practical efficiency of our end-to-end generation framework.

Model	PickScore	ImageReward	Aesthetic	HPSv2
ControlNet	21.61	0.4689	6.161	0.2813
RPO	21.02	0.4395	6.113	0.2663

Table 3: Comparison of RPO fine-tuned data and RPO fine-tuned model generation.

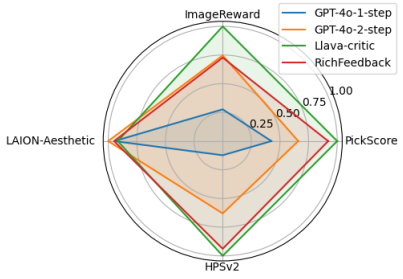


Figure 6: Performance comparison of RPO across different VLM critics.

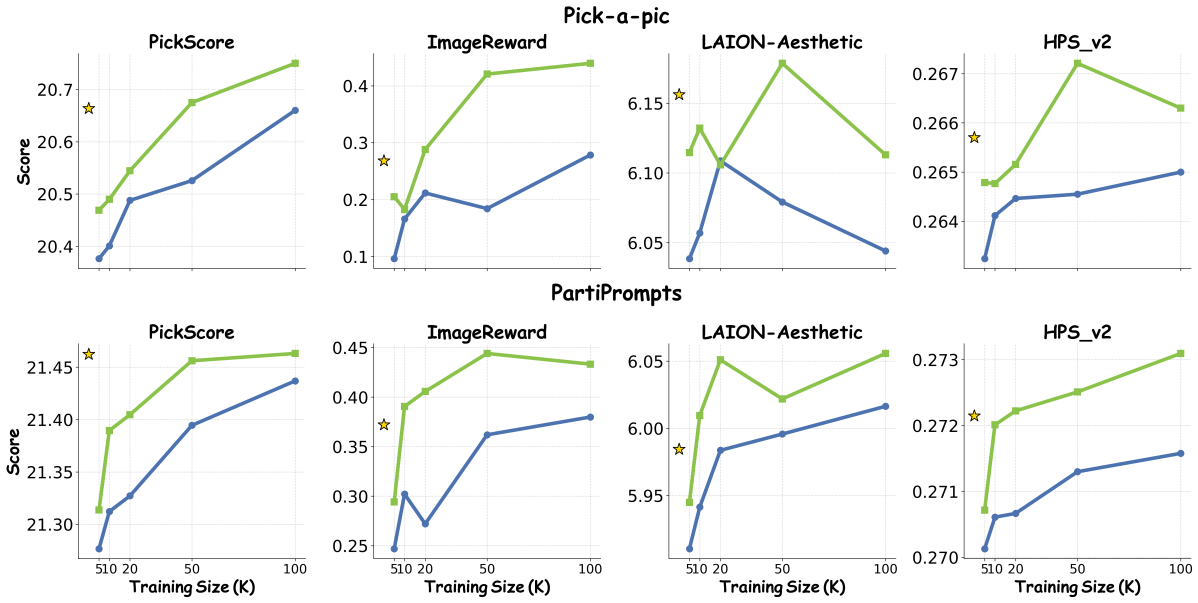


Figure 5: RPO improves DPO performance across data scales, prompt sets, and evaluation metrics. The x-axis indicates the number of Pick-a-Pic training samples used for fine-tuning. The first row presents in-distribution results on Pick-a-Pic, while the second row show out-of-distribution performance on the PartiPrompts prompt set. In all plots, stars indicate the performance of the full Diffusion-DPO model, based on publicly released checkpoints from Wallace et al. (2024), which were trained on nearly 1M offline preference pairs.

5 Discussions

Learning from AI-labeled preferences is the dominant methods in current alignment methods nowadays, but they often lack transparency since reward scores are typically black-box, and they are also not very informative, as they provide minimal insight into why a particular choice (an answer or a generated image) is preferred. Consequently, while the current preference curation pipeline is scalable, it is less efficient for aligning the model or agent based on these preferences. This inefficiency can even lead to issues like reward hacking, as noted in existing literature (Amodei et al., 2016; Denison et al., 2024).

Our RPO pipeline makes a step towards addressing the challenge of creating high-quality synthetic preferences for generative models post-training. Our procedures are straightforward compliment to existing offline RLHF methods like DPO. In addition, although not explored in this paper, our pipeline can also be catered to online algorithms, such as iterative DPO or RL based methods. We consider this as a promising avenue for future research. Our pipeline also combines the existing progress and improvements in two rapidly growing research directions: VLM and image-editing models, which makes our pipeline pretty robust for future improvement by adopting more capable models as plug-ins.

6 Conclusion

In this work, we present **Rich Preference Optimization (RPO)**, a method that utilizes rich feedback about the prompt image alignment from VLMs to improve the curation of synthetic preference pairs for fine-tuning text-to-image diffusion models. After extracting actionable editing instructions from VLMs, we employ ControlNet to modify the images, thereby producing a diverse range of refined samples. The edited images are then combined with the original versions and undergo a relabeling process using a reward model to create a curated set of preference data. By further fine-tuning checkpoint models on this synthetic dataset, we significantly enhance the performance of Diffusion-DPO training and achieve greater data efficiency. Moreover, we believe that our pipeline represents a promising direction and avenue for the data curation of synthetic vision language preference data, one that lies between two fast growing literature-VLMs and Image-editing, showcasing significant potential for future research advancements.

7 Limitations

Given the popularity of our research direction and huge amount of available models, we are only able to test on a set of representative models, which we believe that already provides valuable insights and potentials of our methods even that our tried models are not exclusive. There is also possible risk of image generation models, but our main contribution of this paper is instead novel training recipes/methodologies. In addition, we use AI to assist writing, help with figures design and more efficient coding.

References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

Zachary Ankner, Mansheej Paul, Brandon Cui, Jonathan D Chang, and Prithviraj Ammanabrolu. 2024. Critique-out-loud reward models. arXiv preprint arXiv:2408.11791.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In CVPR, pages 18392–18402.

Jiali Chen, Xusen Hei, Yuqi Xue, Yuancheng Wei, Jiayuan Xie, Yi Cai, and Qing Li. 2024. Learning to correction: Explainable feedback generation for visual commonsense reasoning distractor. In MM, pages 8209–8218.

Xinyun Chen, Maxwell Lin, Nathanael Schaerli, and Denny Zhou. 2023. Teaching large language models to self-debug. In ACL.

Kanzhi Cheng, Yantao Li, Fangzhi Xu, Jianbing Zhang, Hao Zhou, and Yang Liu. 2024. Vision-language models can self-improve reasoning via reflection. arXiv preprint arXiv:2411.00855.

Carson Denison, Monte MacDiarmid, Fazl Barez, David Duvenaud, Shauna Kravec, Samuel Marks, Nicholas Schiefer, Ryan Soklaski, Alex Tamkin, and Jared Kaplan. 2024. Sycophancy to subterfuge: Investigating reward-tampering in large language models. arXiv preprint arXiv:2406.10162.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, and Frederic Boesel. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In ICML.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In Neurips, volume 33, pages 6840–6851.

Frederikus Hudi, Genta Indra Winata, Ruochen Zhang, and Alham Fikri Aji. 2025. Textgames: Learning to self-play text-based puzzle games via language model reasoning. arXiv preprint arXiv:2502.18431.

Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. 2024. Hq-edit: A high-quality dataset for instruction-based image editing. arXiv preprint arXiv:2404.09990.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. arXiv preprint arXiv:2410.21276.

Jiaming Ji, Boyuan Chen, Hantao Lou, Donghai Hong, Borong Zhang, Xuehai Pan, Tianyi Alex Qiu, Juntao Dai, and Yaodong Yang. 2024. Aligner: Efficient alignment by learning to correct. Advances in Neural Information Processing Systems, 37:90853–90890.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. arXiv preprint arXiv:2306.02561.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023a. Pick-a-Pic: An open dataset of user preferences for text-to-image generation. In Neurips, pages 36652–36663.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023b. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In Neurips, volume 36, pages 36652–36663.

Black Forest Labs. 2024. Flux.1 [dev]: A 12 billion parameter rectified flow transformer for text-to-image generation. Accessed: 2025-03-05.

Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. 2024a. Controlnet++: Improving conditional controls with efficient consistency feedback: Project page: liming-ai. github. io/controlnet_plus_plus. In ECCV, pages 129–147.

Zhiyuan Li, Dongnan Liu, Chaoyi Zhang, Heng Wang, Tengfei Xue, and Weidong Cai. 2024b. Enhancing advanced visual reasoning ability of large language models. arXiv preprint arXiv:2409.13980.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, and Feng Yang. 2024. Rich human feedback for text-to-image generation. In CVPR, pages 19401–19411.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In CVPR, pages 26296–26306.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In ICLR.

Nat McAleese, Rai Michael Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. 2024. Llm critics help catch llm bugs. arXiv preprint arXiv:2407.00215.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2025. SimPO: Simple preference optimization with a reference-free reward. In Neurips, volume 37, pages 124198–124235.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In Neurips, volume 35, pages 27730–27744.

Davide Paglieri, Bartłomiej Cupiał, Samuel Coward, Ulyana Piterbarg, Maciej Wolczyk, Akbir Khan, Eduardo Pignatelli, Łukasz Kuciński, Lerrel Pinto, Rob Fergus, and 1 others. 2024. Balrog: Benchmarking agentic llm and vlm reasoning on games. arXiv preprint arXiv:2411.13543.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving latent diffusion models for high-resolution image synthesis. In ICLR.

John David Pressman, Katherine Crowson, and Simulacra Captions Contributors. 2022. Simulacra aesthetic captions. Technical Report Version 1.0, Stability AI. url https://github.com/JD-P/simulacra-aesthetic-captions .

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In Neurips, pages 53728–53741.

Matthew Renze and Erhan Guven. 2024. Self-reflection in llm agents: Effects on problem-solving performance. arXiv preprint arXiv:2405.06682.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In CVPR, pages 10684–10695.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. arXiv preprint arXiv:2206.05802.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. 2024. Emu edit: Precise image editing via recognition and generation tasks. In CVPR, pages 8871–8879.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In Neurips, volume 36, pages 8634–8652.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021a. Denoising diffusion implicit models. In ICLR.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021b. Score-based generative modeling through stochastic differential equations. In ICLR.

Suzukimain and Steven Liu. 2024. Github repository: Diffusion model alignment using direct preference optimization. Github.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. 2024. Diffusion model alignment using direct preference optimization. In CVPR, pages 8228–8238.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An open-ended embodied agent with large language models. arXiv preprint arXiv:2305.16291.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191.

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Daniel Egert, Ellie Evans, Hoo-Chang Shin, Felipe Soares, Yi Dong, and Oleksii Kuchaiev. 2025. Dedicated feedback and edit models empower inference-time scaling for open-ended general-domain tasks. arXiv preprint arXiv:2503.04378.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In Neurips, volume 35, pages 24824–24837.

Genta Indra Winata, Hanyang Zhao, Anirban Das, Wenpin Tang, David D Yao, Shi-Xiong Zhang, and Sambit Sahu. 2024. Preference tuning with human feedback on language, speech, and vision tasks: A survey. arXiv preprint arXiv:2409.11564.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2024. Llava-critic: Learning to evaluate multimodal models. arXiv preprint arXiv:2410.02712.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. ImageReward: Learning and evaluating human preferences for text-to-image generation. In Neurips, pages 15903–15935.

Zihuiwen Ye, Fraser Greenlee-Scott, Max Bartolo, Phil Blunsom, Jon Ander Campos, and Matthias Gallé. 2024. Improving reward models with synthetic critiques. arXiv preprint arXiv:2405.20850.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. 2023a. Magicbrush: A manually annotated dataset for instruction-guided image editing. In Neurips, volume 36, pages 31428–31449.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023b. Adding conditional control to text-to-image diffusion models. In ICCV, pages 3836–3847.

Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, and Stefano Ermon. 2024. Hive: Harnessing human feedback for instructional visual editing. In CVPR, pages 9026–9036.

Hanyang Zhao, Genta Indra Winata, Anirban Das, Shi-Xiong Zhang, David D Yao, Wenpin Tang, and Sambit Sahu. 2025a. Rainbowpo: A unified framework for combining improvements in preference optimization. In ICLR.

Haozhe Zhao, Xiaojian Shawn Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. 2025b. Ultraedit: Instruction-based fine-grained image editing at scale. In Neurips, volume 37, pages 3058–3093.

A Input Prompts

In this section, we present the prompts used for generating critiques and editing instructions.

A.1 Input Prompts to ChatGPT-4o for Instruction Generation

- **LLaVA-Critic feedback to editing instructions**:

Prompt

You are an AI assistant that provides 2-3 concise suggestions (separated by a semicolon) with each suggestion being no more than 8 words. Please make sure that each suggestion suggests concrete change, not just a high-level idea. Your goal is to improve images so they better align with the prompt: `prompt`.

I want you to give short, concise editing instructions based on the following inputs regarding misalignment information. Some instructions are (1) Keep it concise: “Change the red dog to yellow" is better than “Please make the dog that is red in the image a bright yellow color". (2) Be specific: Avoid ambiguous instructions like Make it more colorful. Instead, use Change the red dog to yellow and make the background green. (3) Avoid redundancy: Don’t repeat the same intent multiple times. The image is the generated image based on the prompt: `prompt`.

Here, we have the feedback given by the llava critic model: fb. Please give short editing instructions for the image to solve the misalignment as a text string, where instructions are separated by a semicolon.

- **Rich feedback to editing instructions:**

Prompt

You are an AI assistant that provides 2-3 concise suggestions (separated by a semicolon) with each suggestion being no more than 8 words. Please make sure that each suggestion suggests concrete change, not just a high-level idea. Your goal is to improve images so they better align with the prompt: `prompt`. The first image is the generated image that we want to improve. The second image is a heatmap highlighting areas that misalign with the prompt.

I want you to give short, concise editing instructions based on the following inputs regarding misalignment information. Some instructions are (1) Keep it concise: "Change the red dog to yellow" is better than "Please make the dog that is red in the image a bright yellow color". (2) Be specific: Avoid ambiguous instructions like Make it more colorful. Instead, use Change the red dog to yellow and make the background green. (3) Avoid redundancy: Don't repeat the same intent multiple times. The image is the generated image based on the prompt: `prompt`.

Here, we have the list of pairs where the first element is a word in the prompt, and the second element is 1 if there's misalignment for this word, and 0 otherwise.The list of pairs are: {misalignment-pairs}. If the pairs are None, this means that this info is unavailable, please use the image and the prompt to give advice. We also have a heatmap, which is the second image attached, highlighting the misalignment area in the original (first) generated image.
Now please give short editing instructions for the image to solve the misalignment as a text string, where instructions are separated by a semicolon.

- **ChatGPT image-prompt to editing instructions**

Prompt

You are an AI assistant that provides 2-3 concise suggestions (separated by a semicolon) with each suggestion being no more than 8 words. Please make sure that each suggestion suggests concrete change, not just a high-level idea. Your goal is to improve images so they better align with the prompt: `prompt`.

I want you to give short, concise editing instructions based on the following inputs regarding misalignment information. Some instructions are (1) Keep it concise: "Change the red dog to yellow" is better than "Please make the dog that is red in the image a bright yellow color". (2) Be specific: Avoid ambiguous instructions like Make it more colorful. Instead, use Change the red dog to yellow and make the background green. (3) Avoid redundancy: Don't repeat the same intent multiple times. The image is the generated image based on the prompt: `prompt`.

Now please give short editing instructions for the image to solve the misalignment as a text string, where instructions are separated by a semicolon.

- **ChatGPT image-prompt to feedback and then editing instructions**

Prompt

You are an AI assistant that helps improve a text-to-image model. Your task is to first analyze and critique whether the image aligns with the given prompt (i.e., give some feedback) and then provide 2-3 concise suggestions (separated by a semicolon) with each suggestion being no more than 8 words. Please separate the feedback and the editing instructions with an asterisk. Please make sure that each suggestion suggests concrete change, not just a high-level idea. Your goal is to improve images so they better align with the prompt: `prompt`.

I want you to first generate feedback based on the given input image (and the prompt) and then give short, concise editing instructions based on the given image and the given prompt. For the feedback, it should be a couple of sentences. Some instructions for the editing instructions are ((1) Keep it concise: “Change the red dog to yellow" is better than “Please make the dog that is red in the image a bright yellow color". (2) Be specific: Avoid ambiguous instructions like Make it more colorful. Instead, use Change the red dog to yellow and make the background green. (3) Avoid redundancy: Don’t repeat the same intent multiple times. The image is the generated image based on the prompt: `prompt`.

Now please give feedback and short editing instructions for the image to solve the misalignment as a text string, where feedback and instructions are separated by an asterisk and the instructions are separated by a semicolon.

A.2 Input Prompt to LLaVA-Critic for Critique Generation

Prompt

You are a visual-language critic. An image was generated by a diffusion model using this prompt:
`[prompt]`
Analyze the image relative to the prompt. Please:
1. Identify any deviations from the prompt and explain how they differ.
2. Note which keywords or phrases are poorly represented or misinterpreted.
3. Highlight any additional issues (e.g., artifacts, distortions, low detail) and their impact on quality.

Provide a clear, structured critique.

A.3 Input Prompts to VLMs for Instruction Generation

- **Generating editing instruction from Rich Feedback**

Prompt

You are an AI assistant providing exactly 2 to 3 concise, specific image editing suggestions (separated by semicolons), each no more than 8 words. Suggestions must describe only how to modify the *image itself* to better align with the prompt. Do not instruct changes to the text prompt. Formatting rules:

1. Output a single-line string with edits, separated by semicolons.
2. No explanations, bullet points, or extra details."
3. Do not repeat exact misaligned words; describe the needed visual change.
4. Avoid vague edits. Instead of 'Make it colorful,' say 'Turn the red dog bright yellow.'
5. Always generate a response unless no relevant objects exist.
The image is generated from this prompt: `prompt`.

Below is a list of (concept, flag) pairs. A flag of 0 means the image is misaligned; a flag of 1 means it is correct. For each concept flagged 0, provide one specific visual correction. List of pairs: `mis_pairs`. Output only the editing instructions in a single line. image: `base64_combined_image`.

- **Generating editing instruction from LLaVA-Critic**

Prompt

You are an AI assistant providing exactly 2 to 3 concise, specific image editing suggestions (separated by semicolons), each no more than 8 words. Suggestions must describe only how to modify the *image itself* to better align with the prompt. Do not instruct changes to the text prompt.
Formatting rules:

1. Output a single-line string with edits, separated by semicolons.
2. No explanations, bullet points, or extra details."
3. Do not repeat exact misaligned words; describe the needed visual change.
4. Avoid vague edits. Instead of 'Make it colorful,' say 'Turn the red dog bright yellow.'
5. Always generate a response unless no relevant objects exist.
The image is generated from this prompt: `prompt`.

Below is an image critique highlighting deviations from the prompt. Identify the specific visual misalignments and suggest precise edits to correct them.

Critique: `critique`.

Output only the editing instructions in a single line. image: `base64_combined_image`.

B ControlNet Examples

C Qualitative Examples

C.1 Qualitative Comparison of Base Model and Fine-tuned Variants

In this section, we provide a side-by-side comparison of three distinct model configurations: the base SDXL 1.0 , the standard DPO-SDXL-100k baseline , and our RPO-augmented version, DPO-SDXL-100k + RPO100k. As discussed in the main body, while the DPO-SDXL-100k baseline improves general image fidelity, it often fails to coordinate multiple complex prompt elements simultaneously.

As shown in Figure 8, the RPO-enhanced model demonstrates a significantly higher capacity for semantic connectivity. For instance, in the "samurai pizza cat" and "tiger wearing casual outfit" examples, the RPO-enhanced model demonstrates a significantly higher capacity for semantic connectivity, accurately rendering specific clothing and objects that the baseline models omit or distort. These results confirm that rich preference signals provide the nuanced guidance necessary for the model to learn deep relationships between disparate prompt concepts.

C.2 Qualitative Examples from Feedback Mechanism–VLM Combinations

We present qualitative results from different combinations of feedback methods and VLMs, using a fixed image and prompt.

GPT-4o with Various Feedback Mechanisms. Table 10 shows that GPT-4o is able to generate accurate instructions to improve alignment with the prompt. However, in the 1-step generation, it enhances only two aspects in the prompt, which are *old* and *empty*. In contrast, 2-step generation, GPT-4o with Rich Feedback and GPT-4o with LLaVA-Critic successfully reinforce three elements of the prompt. In particular, GPT-4o with LLaVA-Critic's feedback produces more detailed and informative instructions.

Rich Feedback and LLaVA-Critic with open-source VLMs. Tables 11 and 12 show the editing instructions generated by LLaVA, Llama and Qwen using Rich Feedback and LLaVA-Critic. We observe that LLaVA-1.6, when paired with Rich Feedback and LLaVA-Critic, can occasionally generate unhelpful instructions like "Fill empty refrigerator" and "repair the broken parts". Llama-3.2 with both feedback types, as well as Qwen-2.5 with



Figure 7: **(L)** We utilize ControlNet by using the same input image as the conditional image, and concatenating the prompt with the editing instruction as the textual control.
(R) We compare ControlNet with InstructPix2Pix and also ablate the necessity of concatenating the prompt with the editing instructions, which are generated by ChatGPT-4o. ***(Top)***: The prompt is *"Italian coastline, buildings, ocean, architecture, surrealism by Michiel Schrijver."* The editing instruction is *"Incorporate iconic Italian elements like olive trees or Vespa scooters. Enhance the coastline with more distinct Mediterranean features. Add intricate architectural details typical of Italian structures. Intensify the ocean's depth with gradient blues, and ensure the surrealism reflects Michiel Schrijver's style by blending dreamlike elements."* In this case, InstructPix2Pix struggles to make any fine-grained modifications. ***(Bottom)***: The prompt is *"Mickey Mouse in a Superman outfit bodybuilding."* The editing instruction is *"Adjust the character to have Mickey Mouse's face, including distinct ears, in a Superman outfit. Include bodybuilding elements such as visible muscles or weights. Ensure the outfit is accurate with the Superman logo prominently displayed."* In this case, InstructPix2Pix distorts the image. In both cases, adding the image prompt to the editing instruction for the final instruction yields better results in terms of following the instructions while keeping most of the original image unchanged.



Figure 8: A comparison of generations made by SDXL1.0, DPO-SDXL-100k, and DPO-SDXL-100k+RPO100k. Prompts (from left to right): “tiger wearing casual outfit", “an adventurer walking along a riverbank in a forest during the golden hour in autumn", “samurai pizza cat", “anime portrait of a beautiful vamire witch, sci fi suit, intricate, elegant, highly detailed, digital painting, artstation, concept art, smooth, sharp focus, illustration, art by grep rutkowski and" (truncated due to the limit of number of tokens).



Table 9: Prompt: *An old, and broken refrigerator, open door and empty.*

Prompt	An old, and broken refrigerator, open door and empty.	
GPT-4o	1-step	Make exterior rusty and dented; Add dust on surfaces; Remove all items inside.
	2-step	Add dust and dirt to surfaces; Remove all items inside; Add visible damage to fridge exterior.
	Rich Feedback	Add more dirt and grime; Ensure the door is open wider; Remove all items inside the fridge.
	LLaVA-Critic	Add rust and scratches to the exterior; cover surfaces with visible dust; make shelves appear broken and empty.

Table 10: Qualitative examples of critiques generated by GPT-4o with different feedback generating mechanisms.

Rich Feedback, produce reasonably good instructions but still sometimes include unnecessary parts like "lighten door" and "Add a small, empty shelf" etc. In contrast, Qwen-2.5 with LLaVA-Critic produces concise and helpful instructions that enhance three key aspects of the prompt.

Prompt	An old, and broken refrigerator, open door and empty.	
Rich Feedback	Llava-v1.6-mistral-7b-hf	Clean dusty refrigerator; Fix broken door; Fill empty refrigerator.
	Llama-3.2-Vison-11B-Instruct	Adjust refrigerator brightness; darken dusty areas; lighten door.
	Qwen2.5-VL-7B-Instruct	Add a small, empty shelf inside the refrigerator; make the refrigerator appear older and worn.

Table 11: Qualitative examples of critiques generated by Rich Feedback with different VLMs.

Prompt	An old, and broken refrigerator, open door and empty.	
LLaVA-Critic	Llava-v1.6-mistral-7b-hf	Clean the dust; repair the broken parts; remove any trash
	Llama-3.2-Vison-11B-Instruct	Turn the fridge light bright white; turn the floor brown; darken the edges.
	Qwen2.5-VL-7B-Instruct	Add dust and scratches; replace contents with old food items.

Table 12: Qualitative examples of critiques generated by LLaVA-Critic with different VLMs.

D Failure Case Analysis

While our RPO pipeline produces informative preference pairs and improves downstream alignment, we also observed several systematic failure cases in data curation that limit the effectiveness of preference data pair curation. To diagnose these issues, we screened examples where the majority of reward models reported negative improvement after editing. Our analysis suggests that most failures originate in the instruction generation stage, with two recurring patterns:

- **Stray content.** Generated instructions occasionally introduce content not mentioned in the critique or even contradictory to it. For example, the instruction in Case 1 asked to replace ice cream with a giant chocolate bar, despite the original prompt containing no mention of chocolate.
- **Non-executable instructions.** In other cases, the editing instructions require targets absent in the image. For instance, in Case 2, an instruction requested "removing IKEA branding" when no such branding was present.

Beyond these instruction-level errors, we identified two additional sources of failure during the image editing stage:

- **Failure to follow instructions.** Even when the instructions are valid, the editing model sometimes diverges from them. For example, in Case 3, an instruction to correct a road sign from "TURN EFT" to "TURN LEFT" was not successfully completed by ControlNet.
- **Low editing quality in human-related images.** Edits involving human faces or bodies often introduce artifacts or unrealistic distortions, leading to degraded image fidelity. An example is shown in Case 4.

Failure Case 1: Stray content

Prompt:

Giant ice cream cone melting and creating a river through a city

Critique:

1. Deviations from the Prompt

The image depicts a giant ice cream cone melting and creating a river through a city, which aligns well with the prompt. However, there are a few deviations:

- **Scale**: The ice cream cone is indeed giant in scale, but it appears to be more of an artistic representation rather than a realistic one. This deviation might be intentional to create a surreal effect.
- **Melting Effect**: The melting effect is exaggerated, with the ice cream cone appearing to be melting into a river. This effect is visually striking and aligns with the prompt's description of "melting and creating a river."

2. Keywords or Phrases Misinterpreted

- **Giant Ice Cream Cone**: The image accurately represents this keyword, showing a massive ice cream cone.
- **Melting**: The melting effect is well-executed, though it might be slightly overdone.
- **River Through a City**: The image correctly depicts a river flowing through a city, although the city appears to be frozen or covered in snow, which adds an interesting twist.

3. Additional Issues

- **Artifacts and Distortions**: There are no significant artifacts or distortions visible in the image. The visual quality is high, with clear details and vibrant colors.
- **Low Detail**: The image has a high level of detail, especially in the cityscape surrounding the ice cream cone. The textures and structures of the buildings are well-defined.

Conclusion

Overall, the image is a creative and imaginative interpretation of the prompt. It captures the essence of the giant ice cream cone melting and creating a river through a city, with only minor deviations that enhance the surreal and artistic nature of the scene. The

image is visually striking and effectively conveys the intended concept.

Instruction:

Replace the ice cream with a giant chocolate bar; change the river to a stream of milk.



Original (left) and Edited (right).

Failure Case 2: Non-executable instructions

Prompt:

Detailed Ikea ad, messy Room, perfect Lighting and shadows

Critique:

1. Deviations from the Prompt

- **Messy Room**: The room appears tidy and well-organized, which contradicts the prompt's request for a "messy" room.
- **Perfect Lighting and Shadows**: The lighting in the room is indeed well-balanced and creates natural shadows, but it does not appear overly dramatic or exaggerated as one might expect in a perfect lighting scenario.

2. Poorly Represented or Misinterpreted Keywords

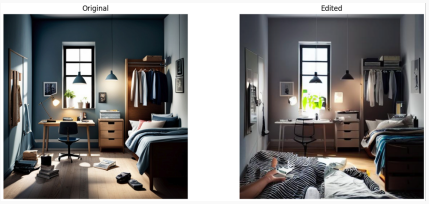
- **Ikea Ad**: The image does not clearly represent an Ikea product or advertisement. It shows a room with various items that could be associated with an Ikea style, but there are no explicit signs of being an Ikea ad.
- **Perfect Lighting and Shadows**: While the lighting is well-balanced, it does not have the dramatic effect that would make it "perfect" according to the prompt.

3. Additional Issues

- **Artifacts and Distortions**: There are no visible artifacts or distortions in the image.
- **Low Detail**: The image has a high level of detail, with clear textures and sharpness in the objects and furniture.

Conclusion
The image does not fully align with the prompt due to the room being tidy and the lighting being well-balanced rather than dramatic. The keywords "Ikea Ad" and "Perfect Lighting and Shadows" are not well-represented in the image. Overall, the image quality is high, but it does not meet the specific criteria set by the prompt.

Instruction:
Add cluttered furniture; enhance lighting contrast; remove IKEA branding.



Original (left) and Edited (right).

Failure Case 3: Failure to follow instructions

Prompt:

A road sign with text saying "turn left for food"

Critique:

1. Deviations from the Prompt
The image provided does not match the prompt accurately. The prompt specifies a road sign with text saying "turn left for food," while the image shows a road sign that says "turn eft for food." The word "eft" is incorrect and does not align with the prompt's instruction to say "left."

2. Keywords or Phrases Misinterpreted

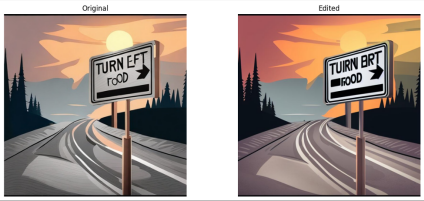
- **"Turn Left for Food"**: The prompt correctly identifies the intended message, but the image incorrectly uses "eft" instead of "left."
- **"Turn EFT for Food"**: The image uses "eft" instead of "left," which is a deviation from the prompt.

3. Additional Issues

- **Artifacts and Distortions**: There are no visible artifacts or distortions in the image.
- **Low Detail**: The image has a moderate level of detail, but the main issue is the incorrect wording on the sign, which affects the overall quality and accuracy.

Conclusion
The image deviates significantly from the prompt due to the incorrect wording on the road sign. This misinterpretation of the prompt's instructions results in a noticeable discrepancy between the expected and actual content of the image.

Instruction:
Correct "TURN EFT" to "TURN LEFT"; Change "FOOD" to "RESTAURANT".



Original (left) and Edited (right).

Failure Case 4: Low editing quality in human-related images

Prompt:

photo of asian little ballet dancers resting in dance studio, nikon D5

Critique:

1. Deviations from the Prompt

- **Photography Equipment**: The image does not show any specific brand or type of camera, such as a Nikon D5, which is mentioned in the prompt.
- **Setting**: The image appears to be taken in a dance studio, but the specific details about the studio (e.g., lighting, decor) are not clearly visible or described.
- **Dancers**: The image shows four young dancers, which aligns with the prompt's mention of "little ballet dancers."

2. Keywords and Phrases Misinterpreted

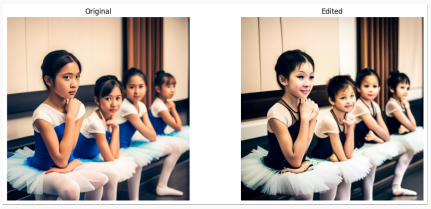
- **Asian**: The prompt specifies "asian," but the image does not provide any information about the ethnicity of the dancers.
- **Resting**: The dancers are sitting and appear to be resting, which aligns with the prompt.

3. Additional Issues

- **Artifacts and Distortions**: There are no apparent artifacts or distortions in the image.
- **Detail**: The image has a clear focus on the dancers, but the background details are not very distinct.

Conclusion
The image generally aligns with the prompt regarding the presence of young ballet dancers in a studio setting. However, it misses specific details about the photography equipment and the exact nature of the studio environment. The prompt's mention of "asian" is not verifiable from the image provided. Overall, the image quality is good, with clear focus on the subjects, but it could benefit from more detailed context.

Instruction:
Add a floral backdrop behind the dancers; adjust lighting for a softer glow.



Original (left) and Edited (right).

E Additional Generalization and Scalability Analysis

To complement the analysis in the main paper, we additionally conducted experiments on a second prompt set using the HPSv2 benchmark, together with an extended suite of reward models that place greater emphasis on stylistic attributes. The results in Figure 13 further confirm that the RPO pipeline consistently outperforms the Diffusion-DPO baseline across all training sizes and scales effectively. Moreover, RPO demonstrates stronger efficiency and generalization: with a training size of 100K, RPO outperforms both the Diffusion-DPO model and its checkpoint trained on nearly 1M preference pairs across all the metrics.

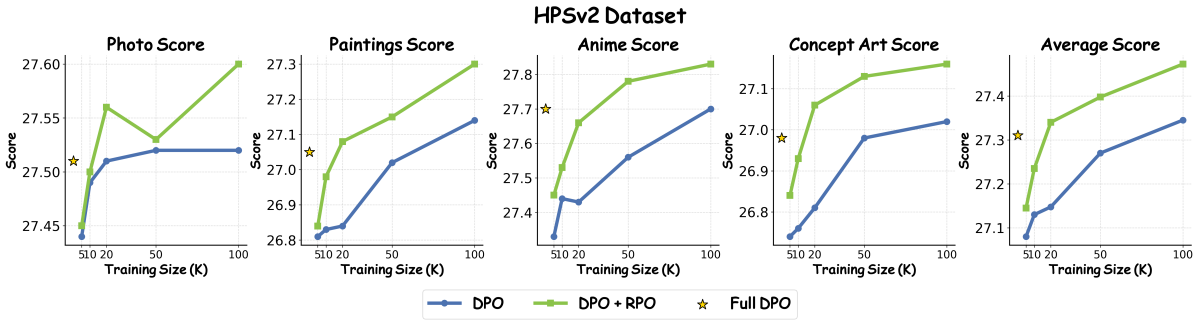


Figure 13: Style-specific scaling performance on the HPSv2 dataset. RPO-enhanced models (green) consistently outperform the Diffusion-DPO baseline (blue) and the 1M-sample "Full Diffusion-DPO" baseline (star) across all visual categories.