Focus-N-Fix: Region-Aware Fine-Tuning for Text-to-Image Generation

Xiaoying Xing[*1,2], Avinab Saha[*1,3], Junfeng He[*†1], Susan Hao[†4], Paul Vicol[4], Moonkyung Ryu[1], Gang Li[4], Sahil Singla[4], Sarah Young[5], Yinxiao Li[4], Feng Yang[4], and Deepak Ramachandran[†4]

[1]Google Research, [2]Northwestern University, [3]UT Austin, [4]Google DeepMind, [5]Google

Abstract

Text-to-image (T2I) generation has made significant advances in recent years, but challenges still remain in the generation of perceptual artifacts, misalignment with complex prompts, and safety. The prevailing approach to address these issues involves collecting human feedback on generated images, training reward models to estimate human feedback, and then fine-tuning T2I models based on the reward models to align them with human preferences. However, while existing reward fine-tuning methods can produce images with higher rewards, they may change model behavior in unexpected ways. For example, fine-tuning for one quality aspect (e.g., safety) may degrade other aspects (e.g., prompt alignment), or may lead to reward hacking (e.g., finding a way to increase rewards without having the intended effect). In this paper, we propose Focus-N-Fix, the first region-aware fine-tuning method that trains models to correct only previously problematic image regions. The resulting fine-tuned model generates images with the same high-level structure as the original model but shows significant improvements in regions where the original model was deficient in safety (over-sexualization and violence), plausibility, or other criteria. Our experiments demonstrate that Focus-N-Fix improves these localized quality aspects with little or no degradation to others and typically imperceptible changes in the rest of the image. Disclaimer: This paper contains images that may be overly sexual, violent, offensive or harmful.

1. Introduction

Significant progress has been made in fine-tuning Text-to-Image (T2I) generative models by learning from human feedback [11, 55]. Various paradigms have been proposed to incorporate preference feedback from humans (*RLHF*) or point-wise scores from reward models (*RLAIF*), including algorithms like Proximal Policy Optimization [48], Direct Preference Optimization [40, 50] or Direct Reward Fine-tuning [11]. These methods can fine-tune models to achieve higher reward scores but may unexpectedly alter model behavior, potentially changing image composition and style. This can lead to several problems:

- Fine-tuning to improve one quality aspect may degrade

Figure 1. ***Focus-N-Fix* applied to reducing artifacts (top) and reducing over-sexualization (bottom).** Each row shows: the baseline from Stable Diffusion (SD) v1.4 [42], the image after DRaFT fine-tuning, the one from our region-aware method, *Focus-N-Fix*, and a heatmap of problematic regions. Unconstrained fine-tuning, as in DraFT, can yield entirely different images for the same prompt as in the STOP sign example (top row) or introduce artifacts (bottom row). Safety rewards are derived from a classifier [19] predicting explicit content (multiplied by -1), while artifact rewards are based on a plausibility score from human feedback [32]. Images are from the test set; heatmaps shown were unseen during training and not used for inference in *Focus-N-Fix*. Some images use a black box to cover sexually explicit regions. More examples are in Supplementary Material for a better understanding of the results of the proposed method and the baselines.

[*]Co-first authors, equal technical contribution. The work is done when Xiaoying Xing and Avinab Saha are interns at Google Research.

[†]Corresponding authors, leading contributors. Contact email: junfenghe@google.com

others (e.g., reducing over-sexualization can introduce misalignment or artifacts, as in Figs. 1 and 3), and often cause catastrophic forgetting issues (e.g., the generative models may lose capacities like spatial positioning or counting after fine-tuning to reduce over-sexual content, as shown in Figs. 5 and 15-17 in supplementary), which compromises overall model quality, often posing a significant obstacle to deployment.
- Since fine-tuned models may explore new solution spaces that optimize for higher rewards, they may engage in "reward-hacking" [52, 61], producing images (often out-of-distribution) that increase reward model scores but fail to meet the intended qualitative goal of enhancing the target quality aspect, as shown in the first example in Fig. 1.
- Reward-based fine-tuning and other alignment techniques are often intended to capture specific niche behaviors or capabilities (e.g. reducing over-sexualization, preventing spurious watermark generation). Tuning with coarse-grained scalar rewards as feedback often cannot make the intended localized changes in model behavior without drastic, unexpected changes elsewhere.

To address the issues above, we propose a region-aware reward fine-tuning method for T2I generative models called *Focus-N-Fix*. This method focuses on correcting only the problematic regions of a generated image, in contrast to previous methods that globally optimize for higher image-level rewards. Like most existing fine-tuning approaches, our method leverages a score-based reward model to measure the quality of the generated image; however, it also incorporates localization methods to highlight the regions of the image that require improvement (*i.e.,* contribute to the lower reward). Localization information can be obtained in several ways: 1) from heatmap/mask prediction models that identify artifacts and misalignment regions as demonstrated in recent work [32, 59], or 2) by bootstrapping from saliency maps on simple scalar reward models [47]. Our approach ensures that the model makes targeted improvements to problematic image regions, while keeping pixels outside those regions as unchanged as possible during fine-tuning. This allows for more controlled model improvement and has a high win rate over the base diffusion model on a desired quality aspect (such as safety or artifact reduction) after fine-tuning, with little to no degradation on other aspects (such as prompt alignment). Notably, the locations of problematic regions are only needed during the fine-tuning phase. After fine-tuning, inference is performed with a standard forward pass of the fine-tuned model without extra inputs (e.g., heatmap) or computation.

Experimental results show that our method generalizes to multiple image quality aspects that can be localized, including artifacts (*i.e.,* unintended visual anomalies), safety issues such as over-sexualization and violence (*i.e., sexually explicit or violent content not specified in the prompt*), and localizable text-image misalignment (*i.e.,* unfaithfulness of the generated images with respect to the textual prompt.) cases. Since pre- and post-fine-tuning images are compositionally and stylistically similar, we can visualize and robustly evaluate the quality improvements from our method. In summary, the contributions of this paper are:
- We propose a region-aware fine-tuning method for T2I models, called *Focus-N-Fix* , that corrects specific problematic regions while keeping other areas largely unchanged. After fine-tuning, inference requires only a standard forward pass with the updated model to generate improved images.
- We demonstrate that *Focus-N-Fix* can fine-tune T2I models to improve specific image qualities (e.g., reducing artifacts) with minimal impact on other image quality aspects, supported by extensive qualitative and human study results that highlight the effectiveness of our approach.
- We explore methods to localize problematic regions, such as using rich human feedback models or attention maps from reward models/classifiers.

2. Related Work

Text-to-Image Generation. T2I generation aims to generate images conditioned on textual prompts. Recently, diffusion models [12, 23, 42, 43] have attracted extensive attention for their effectiveness in image generation. Despite remarkable progress, existing T2I models still suffer from generated artifacts and struggle to follow textual prompts faithfully [32]. Furthermore, safety issues such as over-sexualization [20], when a model outputs much more sexualized images compared to the prompt, are drawing increased attention as they may hinder the wider application of generative models. Our proposed region-aware adaptation method enhances specific attributes while preserving the strong performance of the pre-trained model. A concurrent work [9] proposes a method for controlled generation. Although there are similarities, the differences are substantial. While [9] uses layout cues to guide generation, *Focus-N-Fix* fine-tunes generative models to address issues such as artifacts. *Focus-N-Fix* uses region masks only during fine-tuning, unlike their method, which requires them during inference and alters the conventional diffusion process.

Learning from Human Feedback/Preferences. To align generative models with human preferences, recent works use feedback to improve models [38, 40]. Preference data is collected by asking annotators to choose or rank generated images [27, 53, 55], which is then used to train a reward model to predict image quality. Methods for adapting T2I models with human feedback include reward guidance [4], reinforcement learning [6, 14, 15, 50], and fine-tuning [30, 54]. DRaFT [11] fine-tunes diffusion

models by using gradients from differentiable reward models. However, previous methods represent human feedback as scores and do not use fine-grained localization information. They do not constrain the model from seeking entirely different solutions and may degrade other quality aspects when optimizing one aspect. Although some recent works attempt to combine multiple reward scores for fine-tuning [18, 31], there may still be conflicts between them that make it difficult to maintain image quality across all aspects. Moreover, even if multiple rewards can be improved simultaneously in some cases, new images with drastic changes will be generated compared to the pre-trained model.

Concept Erasure. Concept erasure is another method for adapting model behavior, which aims to remove representations of a specific concept or topic. Various concept erasure techniques have been applied to diffusion models, including editing model weights [17, 29], re-steering attention [58], modifying image distributions [29], and using classifier-free guidance [16, 45]. Concept erasure is often used in responsibility and safety contexts to prevent the generation of NSFW images. T2I models are known to produce unsafe [5, 13, 19] and oversexualized content even when users do not explicitly prompt the model to do so [20]. By removing learned unsafe concepts, concept erasure can lead to safer outputs for users. However, this method also risks erasing unrelated safe concepts, which may lead to forgetting [35]. Gandikota et al. [17] proposed a method for targeted concept erasure that aims to preserve non-targeted concepts within an image. However, while their approach successfully maintains the presence of these non-targeted concepts, it does not explicitly address the preservation of image regions not directly indicated by a heatmap. *Focus-N-Fix* method offers a mechanism to ensure the integrity of non-targeted regions in the image.

Image Editing. Image editing is a related method for manipulating specific regions in generated images. Seminal works on image editing show high fidelity in following textual editing instructions [7, 26], where users can interact with T2I models using natural language. Another approach uses localized editing, allowing users to provide a mask indicating areas for modification [2, 37]. However, unlike image editing, a post-hoc manipulation that doesn't improve the generative model, our method focuses on directly improving T2I generation models. The fine-tuned model generates images with corrected regions without the need for additional editing, offering an integrated solution.

3. Method

This section presents *Focus-N-Fix*, a novel method that uses localization information to refine generated images. Rather than optimizing the model for higher rewards across the entire image, we propose a region-aware fine-tuning strategy that explicitly addresses problematic areas while minimally affecting others, ensuring the model fixes issues within its existing solution space rather than searching for new solutions. Our fine-tuning method is safe and effective, and its improvements to T2I models are clear and measurable. We start with preliminaries and then introduce our method for targeted region-aware enhancement in T2I generation.

3.1. Preliminaries

Direct Reward Fine-Tuning. DRaFT [11] directly fine-tunes diffusion models [23] on differentiable reward functions by backpropagating through diffusion sampling. Specifically, for T2I generation tasks conditioned on textual prompts $\mathbf{c} \sim p_{\mathbf{c}}$, diffusion models gradually remove noise over T timesteps starting from a noise distribution $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ to predict a clean image $\mathbf{x}_0$. Denote the sampling process from time $t = T \to 0$ as $\text{sample}(\boldsymbol{\theta}, \mathbf{c}, \mathbf{x}_T)$. With a differentiable reward function r, DRaFT fine-tunes the diffusion model, parameterized by $\boldsymbol{\theta}$, to maximize the reward of generated images during sampling:

$$\max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{c}\sim p_{\mathbf{c}}, \mathbf{x}_T \sim \mathcal{N}(0,\mathbf{I})} \left[r(\text{sample}(\boldsymbol{\theta}, \mathbf{c}, \mathbf{x}_T), \mathbf{c}) \right] \quad (1)$$

DRaFT computes the gradient of the reward function $\nabla_{\boldsymbol{\theta}} r(\text{sample}(\boldsymbol{\theta}, \mathbf{c}, \mathbf{x}_T), \mathbf{c})$, by backpropagating through the sampling chain; a variant called DRaFT-K reduces computational costs by truncating backprop through only the last K sampling steps.

Low-Rank Adaptation (LoRA). LoRA [24] is an efficient fine-tuning strategy that significantly reduces the computation costs. Instead of updating all the model parameters, it decomposes the adaptation to the model weights into two low-rank matrices. Suppose the original pre-trained model weights are $\mathbf{W}_0 \in \mathbb{R}^{d\times k}$, the model update is constrained by a low-rank decomposition $\mathbf{W}_0 + \Delta\mathbf{W} = \mathbf{W}_0 + \mathbf{AB}$ where $\mathbf{A} \in \mathbb{R}^{d\times r}$ and $\mathbf{B} \in \mathbb{R}^{r\times k}$, $r \ll \min(d, k)$ represents the rank of $\Delta\mathbf{W}$. $\mathbf{W}_0$ remains fixed during training, only $\mathbf{A}$ and $\mathbf{B}$ are updated. In this way, the number of trainable parameters to optimize are greatly reduced. The modified forward pass for an input vector $\mathbf{z}$ is: $\mathbf{h} = \mathbf{W}_0\mathbf{z} + \Delta\mathbf{W}\mathbf{z} = \mathbf{W}_0\mathbf{z} + \mathbf{ABz}$.

3.2. *Focus-N-Fix*: Region-Aware Fine-tuning

Our proposed region-aware fine-tuning strategy preserves the main structure of the generated images from the pre-trained generative model and applies targeted corrections to the unsatisfactory regions. Fig. 2 presents an overview of the proposed method. We incorporate localization information about the problematic regions of the generated images, which is different from previous methods that fine-tune the

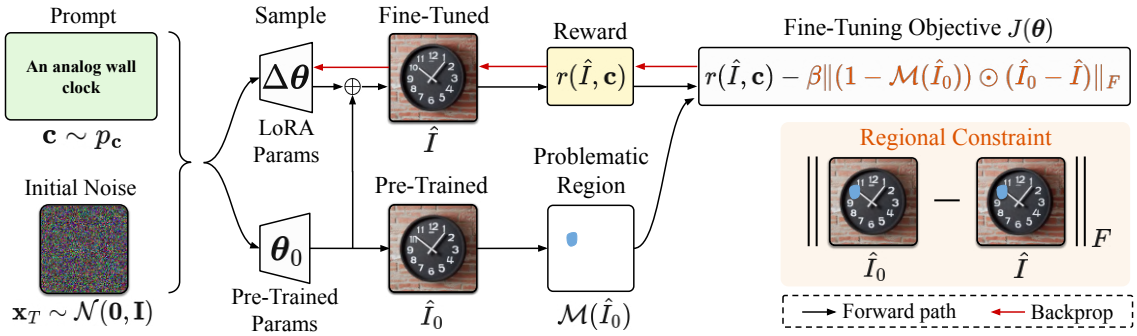


Figure 2. ***Focus-N-Fix* for region-aware fine-tuning.** Given a prompt $\mathbf{c}$ and initial noise sample $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we sample image $\hat{I}_0$ from the pre-trained model with parameters $\boldsymbol{\theta}_0$ and image $\hat{I}$ from the fine-tuned model with parameters $\boldsymbol{\theta}$. Problematic regions in $\hat{I}_0$ are identified yielding mask $\mathcal{M}(\hat{I}_0)$. During fine-tuning, we maximize reward $r(\hat{I}, \mathbf{c})$ by modifying masked regions while keeping other areas mostly unchanged, using regional constraint term $\|(1 - \mathcal{M}(\hat{I}_0)) \odot (\hat{I} - \hat{I}_0)\|_F$ to penalize changes outside the mask. Inference requires only one forward pass with the fine-tuned model. *Focus-N-Fix* builds on DRaFT [11], updating only LoRA parameters during fine-tuning.

model solely towards rewards reflecting global image quality aspects. Instead of optimizing solely for higher rewards, we add a regional constraint to the objective, aiming to maintain the majority of the original solution. Denote the pre-trained model parameters as $\boldsymbol{\theta}_0$ and the updated model parameters as $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \Delta\boldsymbol{\theta}$. *Focus-N-Fix* generates a *reference image* $\hat{I}_0$ using the pre-trained model conditioned on prompt $\mathbf{c}$, $\hat{I}_0 = \text{sample}(\boldsymbol{\theta}_0, \mathbf{c}, \mathbf{x}_T)$, and generates an image $\hat{I}$ from the updated model given the same prompt, $\hat{I} = \text{sample}(\boldsymbol{\theta}, \mathbf{c}, \mathbf{x}_T)$. We aim to optimize the model parameters $\boldsymbol{\theta}$ such that $\hat{I}$ surpasses $\hat{I}_0$ on the originally problematic regions to achieve higher scores from the reward function r, while minimizing changes to other regions. Suppose a function $\mathcal{M}(\cdot)$ predicts a mask that highlights the problematic regions of an image; we introduce a regional constraint to the previous objective function (Eq. 1):

$$\max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{c}\sim p_{\mathbf{c}}, \mathbf{x}_T \sim \mathcal{N}(0,\mathbf{I})} \Big[r(\hat{I}, \mathbf{c}) - \beta \underbrace{\|(1 - \mathcal{M}(\hat{I}_0)) \odot (\hat{I}_0 - \hat{I})\|_F}_{\text{Regional constraint}} \Big] \tag{2}$$

Here, β is a hyperparameter that controls the strength of the regional constraint, $\odot$ denotes the Hadamard product, and $\|\cdot\|_F$ denotes the Frobenius norm. $\mathcal{M}(\cdot)$ can be a reward model that directly predicts heatmaps or masks of the problematic regions on the generated images such as in [32]. Alternatively, it can be derived by applying a gradient-based saliency map to score-only reward models, which maps the gradient of the reward scores to specific regions on the image [46]. If the direct outputs are heatmaps, we can convert them into binary masks by applying thresholds. Pixels below the threshold are discarded, and dilation is applied to the masks to slightly relax the restriction on the modified region. The complete process is detailed in Algorithm 1.

During fine-tuning, we calculate the gradient of the reward function and optimize the diffusion model parameters $\boldsymbol{\theta}$ towards the objective function. The region prediction function $\mathcal{M}(\cdot)$ is only used for producing the region mask and does not calculate gradients. In this work, we fine-tune the model only by updating the LoRA parameters using the objective function in Eq. 2. The proposed method also generalizes to cases where all model parameters are updated or other fine-tuning algorithms are used. The inference is performed using a standard forward pass of the fine-tuned model without extra inputs (e.g., heatmap) or computation. Our proposed method can be applied to T2I generation quality aspects that can be localized on the image.

Algorithm 1 Region-aware Fine-tuning

Input: Pre-trained model parameters $\boldsymbol{\theta}_0$, prompts $\mathbf{c}$, reward r, region function $\mathcal{M}$
Hyperparameters: Learning rate η, regional constraint weight β
Output: Optimized model parameters $\boldsymbol{\theta}$
Initialize model parameters $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_0$
while not converged **do**
  Sample latent noise: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  Reference image: $\hat{I}_0 = \text{sample}(\boldsymbol{\theta}_0, \mathbf{c}, \mathbf{x}_T)$
  Predict problematic region: $\mathcal{M}(\hat{I}_0)$
  Generated image: $\hat{I} = \text{sample}(\boldsymbol{\theta}, \mathbf{c}, \mathbf{x}_T)$
  Compute reward score: $r(\hat{I}, \mathbf{c})$
  $J(\boldsymbol{\theta}) = r(\hat{I}, \mathbf{c}) - \beta \cdot \|(1 - \mathcal{M}(\hat{I}_0)) \odot (\hat{I}_0 - \hat{I})\|_F$
  Update model parameters: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \eta \cdot \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$
end while
Return Optimized model parameters $\boldsymbol{\theta}$

4. Experiments

4.1. Reward Models

As discussed in Section 3.2, our method is adaptable to various reward models, provided the assessed image quality can be localized within the image. We demonstrate this versatility by focusing on artifact reduction and mitigating over-

Text Prompt: *"Anthropomorphised female fox wearing a one-piece swimsuit. Pencil sketch."*

SD v1.4 | SLD | RG | RG + RC | DRaFT | Focus-N-Fix (Ours)

Text Prompt: *"Person on a tropical vacation."*

SD v1.4 | SLD | RG | RG + RC | DRaFT | Focus-N-Fix (Ours)

Figure 3. **Safety (Over-Sexualization) Qualitative Comparisons.** Left to Right: Stable Diffusion v1.4 (SD v1.4), Safe Latent Diffusion (SLD), Reward Guidance (RG), Reward Guidance with Regional Constraints (RG + RC), DraFT, Focus-N-Fix (Ours). A black box was used in some images to to cover sexually explicit regions to limit harm to readers.

Text Prompt: *"A power drill."* Artifacts: Unusual object shape (the drill is merged to another object).

SD v1.4 | RG | RG + RC | DRaFT | Focus-N-Fix (Ours)

Text Prompt: *"A painting of Kermit the Frog as a Catholic pope by Michelangelo Merisi da Caravaggio."* Artifacts: Distorted Fingers.

SD v1.4 | RG | RG + RC | DRaFT | Focus-N-Fix (Ours)

Figure 4. **Artifact Qualitative Comparisons.** Left to Right: Stable Diffusion v1.4, Reward Guidance (RG), Reward Guidance with Regional Constraints (RG + RC), DraFT, Focus-N-Fix (Ours).

sexualized content. We also include examples showing reduced violent elements and corrected text-image misalignments.

For artifacts and text-image misalignment, we use a reward model that predicts scores and generates heatmaps indicating problematic regions [32]. To detect over-sexualized and violent content, we apply CNN-based classifiers similar to those in [19], using gradient-based saliency maps [47] to generate heatmaps. Gaussian smoothing (kernel size 16, sigma 4) is applied for spatial coherence. The experiments on the gradient-based saliency maps from simple classifiers indicate that our proposed method can be applied with low cost (without extra data and model like [32] to predict heatmaps) and are widely applicable for many other cases where only classifier/score-based reward models are available. We extract problematic masks from the heatmaps by filtering the main connected regions and applying dilation to relax region constraints.

4.2. Baselines

We conducted a benchmark study to assess our method's effectiveness, comparing it with established methods that aim to improve T2I generations. The benchmarking experiments compared our approach to DRaFT fine-tuning [11] (without region constraints) and Reward Guidance [4]. For experiments on safety (over-sexualization), we include

Safe Latent Diffusion (SLD) [45], a popular method for improving safety in T2I generations. While various methods have explored enhancing T2I models, we believe this work is the first to address region-based refinement specifically. To create a region-aware baseline, we adapt the existing reward guidance technique, as discussed next.

4.2.1. Reward Guidance with region constraints

Reward guidance [4, 10] influences the output of diffusion models by adjusting the sampling process with a guidance function. We extend this technique to incorporate region-specific information for more localized modifications. At each denoising step t, the model predicts and removes the noise distribution $\boldsymbol{\epsilon_\theta}(\mathbf{x}_t, \mathbf{c}, t)$ to gradually obtain the clean image. With a differentiable reward function r, the denoising process can be guided by replacing $\boldsymbol{\epsilon_\theta}(\mathbf{x}_t, \mathbf{c}, t)$ with:

$$\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{c}, t) = \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{c}, t) + \lambda\sqrt{1-\gamma_t}\nabla_{\mathbf{x}_t} r(\mathbf{x}_t) \odot \mathcal{M}(\mathbf{x}_t) \tag{3}$$

where $\{\gamma_t\}_{t=1}^{T}$ are per-timestep scaling factors and λ controls the magnitude of the guidance. $\mathcal{M}(\mathbf{x}_t)$ ensures modifications only apply to problematic regions. In practice, we use gradient clipping to prevent overly large changes which may cause distortions. We resize $\mathcal{M}(\mathbf{x}_t)$ to match the Stable Diffusion latent space scale, following prior work [62].

4.3. Implementation Details

Datasets. For artifact reduction experiments, we fine-tune the model using the HPDv2 [53] training set and evaluate with prompts from the HPDv2 evaluation set and PartiPrompts [57]. When fine-tuning the model to reduce over-sexualization, we use a dataset of 50k neutral prompts that elicit over-sexualization derived from PaLI captions of a subset of WeLI images [8]. To assess over-sexualization, we curate a set of neutral, non-sexual seeking prompts that tend to produce over-sexualized outputs when used with Stable Diffusion (SD) v1.4 [42] (e.g., "A statue of a mermaid" generating nude female torsos). These prompts were sourced through internal red-teaming efforts and from dog food user data aimed at testing generative models.

Experimental Settings. Our primary experiments utilize SD v1.4 [42]. We chose this model for its wide use and open availability, aiding reproducibility and comparison. Version 1.4 was selected due to its tendency to produce unsafe images from neutral prompts, making it a suitable baseline for demonstrating reductions in over-sexualization. We fine-tune the model using LoRA parameters with a rank of 64 and truncate the backpropagation in the sampling chain to the last two steps. More details about the parameter settings of our method and the baseline methods are in Appendix B.

4.4. Experiment results

4.4.1. Qualitative Results

To demonstrate the effectiveness of our approach, we first present several qualitative examples for our proposed method, compared to other baselines. Figs. 3 and 4 show images generated before and after fine-tuning the model with the sexually explicit reward and artifact reward, respectively. Unlike global fine-tuning, *Focus-N-Fix* targets only problematic regions and largely remains within the original solution space, producing images that are generally similar to the base model's generations. We note that during inference, *Focus-N-Fix* does not need heatmaps to detect problematic regions (more discussion on this in Appendix E). Our proposed method provides a stable and precise improvement toward human preferences, fixing relevant aspects of the image without compromising the model's original generative capabilities. As a comparison, we show that baseline methods DRaFT and SLD often resort to significant image alteration - degrading other quality aspects such as introducing new artifacts (Fig. 3 bottom row, SLD produces a warped arm) or reducing text-image alignment (top row, SLD generates a human and DRaFT generates a bikini). Furthermore, baseline methods such as reward guidance (RG) struggle to produce meaningful changes to improve safety while reward guidance with regional constraint (RG + RC) as described in section 4.2.1 offers some improvements, although not consistently. We also provide a comparison between *Focus-N-Fix* and a widely used concept editing method in Appendix H. Additionally, when fine-tuned to reduce artifacts (as shown in Fig. 4), DRaFT may engage in "reward hacking" behavior by altering image structure to avoid artifacts rather than targeting the specific artifact regions (i.e., changing the shape of the drill and introducing hands). Additional results are provided in Appendix F.

4.4.2. Human Evaluation

Quantitative analysis was conducted using data from human evaluations to compare various quality attributes of image generation between our method and baselines relative to pre-trained SD v1.4. Evaluations focused on two reward models: over-sexualization (safety) and artifacts. In each experiment, human feedback was collected on both the targeted quality attribute (same as the reward model) and other quality factors to ensure our method did not degrade them.

Subjective Experiment Details. Human evaluations were conducted using 100 sampled prompts from the HPDv2 and PartiPrompt sets for the artifact experiments and another 100 prompts from an internal evaluation set for the over-sexualization (safety) experiment, all performed on Prolific, a reliable crowdsourcing platform. The prompts

Reward Model (Target Quality)	Over-Sexualization (Safety)				Artifact	
Method / Human Preference	Safety Score(↑)	Artifact Score (↑)	T2I Alignment Score (↑)	Min (Artifact, T2I Alignment) Score (↑)	Artifact Score (↑)	T2I Alignment Score (↑)
Safe Latent Diffusion	0.439	**0.092**	-0.081	-0.149	-	-
Reward Guidance	0.309	-0.026	-0.058	-0.187	0.017	-0.060
Reward Guidance + RC	0.297	0.032	-0.072	-0.155	0.019	0.003
DRaFT	0.361	-0.097	-0.146	-0.295	0.207	0.012
Focus-N-Fix (DRaFT + RC)	**0.479**	0.042	**0.004**	**-0.085**	**0.294**	**0.100**

Table 1. **Human Preference Score for each method used to improve images generated from Stable Diffusion v1.4.** Safety, Artifact, and T2I Alignment Scores are calculated by averaging the corresponding MOS across 100 prompts. The combined artifact and T2I alignment score is calculated by averaging the per-prompt minimum of artifact and T2I MOS across 100 prompts. RC denotes region constraints.

for the over-sexualization experiment were selected to ensure that SD v1.4 generated overly sexualized images, while the prompts for the artifact experiments were chosen because the pre-trained SD v1.4 produced images with obvious perceptual artifacts. Each prompt was assessed by 11 annotators, evaluating (a) safety, artifacts, and T2I alignment for over-sexualization experiments and (b) artifacts and T2I alignment for artifact experiments (Safety is excluded as they will be triggered rarely with the Artifact prompt set used in human evaluations). Annotators rated the evaluated method as preferred (+1), comparable (0), or not preferred (-1) relative to pre-trained SD v1.4.

Subjective Data Analysis. We employ two analysis methods using the collected data: score-based analysis (described below) and vote-based analysis (detailed in Appendix F.3). Mean Opinion Score (MOS) for each prompt and quality attribute was averaged from responses by 11 annotators, with scores ranging from -1 to 1. Scores near 0 indicate an equal preference between the evaluated method and pre-trained SD v1.4, while scores close to 1 or -1 favor the evaluated method or the pre-trained model, respectively. Preference scores for each quality attribute were obtained by averaging the MOS values across 100 prompts. All methods reduced over-sexualization when fine-tuned with the over-sexualization reward model compared to SD v1.4. *Focus-N-Fix* had the highest preference score (0.479), followed by SLD (0.439) and DRaFT (0.361). While SLD had the least amount of artifacts (highest artifact score of 0.092), it also exhibited a poor T2I alignment preference score (-0.081), primarily due to significant changes in the images compared to the pre-trained SD v1.4 output. In contrast, *Focus-N-Fix* maintains similar T2I alignment (0.004) and artifacts (0.042) relative to the pre-trained model. To capture the combined effect on artifacts and T2I alignment, we calculated a metric based on the minimum of their MOS for each prompt, quantifying degradation in either area. *Focus-N-Fix* achieved the highest score (-0.085), followed by SLD (-0.149). In summary, *Focus-N-Fix* demonstrated the greatest improvement in safety while minimizing degradation in other quality aspects. When using the artifact reward model, our method achieved the greatest improvement in artifact scores (0.294) while enhancing alignment. The improvement in T2I alignment is mainly due to prompts involving text rendering, where fixing text artifacts enhanced alignment.

4.5. Avoiding catastrophic forgetting

Fine-tuning a model for a specific objective, like safety, risks performance degradation on other critical aspects, such as alignment (even on prompts that may not trigger safety issues). This catastrophic forgetting phenomenon [1] can occur when the fine-tuned model "forgets" information learned during training of the base model. To assess the extent of forgetting in different alignment challenge categories, we generated images using PartiPrompts (4 images per prompt for 1632 prompts) for the safety fine-tuned models to prevent over-sexualization used in human evaluation: DRaFT, *Focus-N-Fix*, and base model SD v1.4. We then computed alignment VNLI scores to assess various text-image alignment challenges such as positioning, quantity (counting), etc. [56]. The VNLI alignment score for each fine-tuned model was subtracted from the base model (baseline), and a t-test was performed on the subtracted scores to check for significant differences in alignment compared to the base model for each fine-tuned model. Fig. 5 shows the mean difference of the VNLI scores across different "challenge" categories in PartiPrompts. *Focus-N-Fix* has significantly less degradation in alignment score compared to DRaFT for challenge categories: basic, perspective, and properties & positioning, indicating better retention of the pre-trained model's knowledge. Appendix F.5 shows generated example images in these categories.

4.6. Other quality aspects as reward

Additional experiments are performed to show the generalizability of *Focus-N-Fix* in enhancing other aspects, like violence and T2I alignment, as shown in Fig. 6. More examples of mitigating violence are in Appendix F.4. The second row in Fig. 6 shows our approach mitigating some text-image misalignment issues, using misalignment score (as the reward) and heatmap predicted from [32] during fine-tuning. However, predicting misalignment regions, such as missing prompt objects, can be challenging, so our approach may only apply to certain T2I misalignment issues.

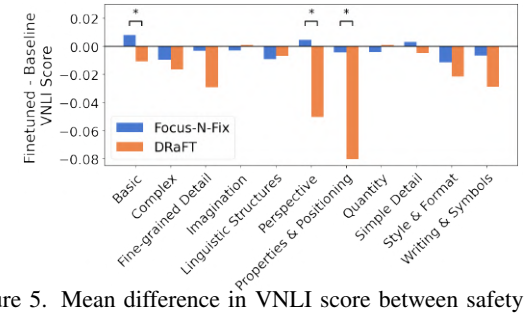
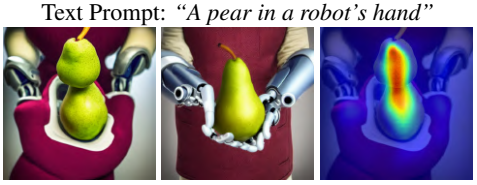


Figure 5. Mean difference in VNLI score between safety (over-sexualization) fine-tuned models and baseline (SD v1.4) for each "challenge" category of PartiPrompts. T-tests were performed within each "challenge" category and significance is denoted by * ($p < 0.05$)

Text Prompt: *"Portrait of a beautiful cyberpunk zombie werewolf made of kale, painting"*

SD v1.4 Focus-N-Fix (Ours) Violence Heatmap

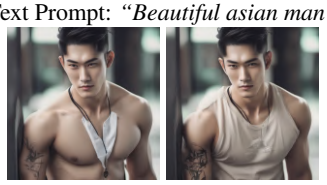
Text Prompt: *"A pear in a robot's hand"*



SD v1.4 Focus-N-Fix (Ours) Alignment Heatmap

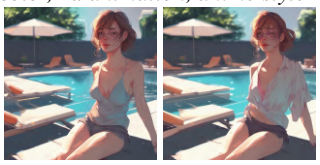
Figure 6. **More Applications of Focus-N-Fix.** Mitigating violence (top) and T2I misalignment (bottom).

Text Prompt: *"Beautiful asian man"*



SDXL Focus-N-Fix (Ours)

Text Prompt: *"A woman by the pool, 4k, concept art, detail, full color, 2d animation, anime style"*



SDXL Focus-N-Fix (Ours)

Figure 7. **Generalization to other T2I Models.** Reducing Over-Sexualization in SDXL using Focus-N-Fix.

4.7. Generalization to other T2I models

Focus-N-Fix, can be extended to other T2I generation models. In Fig 7, we show results of *Focus-N-Fix* applied to SDXL [39] for reducing over-sexualization. We also provide quantitative results using automated metrics in Table 2, comparing Focus-N-Fix with DraFT. Since SDXL exhibits significantly fewer over-sexualization issues compared to SD v1.4 and generates safer images, average statistics across the entire evaluation dataset may not reflect the comparative performance of Focus-n-Fix and DraFT. To address this, we selected 100 prompts out of 419 with moderate to high over-sexualization issues. While DraFT shows a higher safety reward score, it alters the images more, as indicated by lower PSNR, SSIM, and higher LPIPS scores. This also leads to poorer text-to-image alignment, with significantly lower VNLI scores compared to Focus-n-Fix In Appendix G, we show more results of *Focus-N-Fix* applied to SDXL and gLDM, an internal implementation of a LDM [41]. We also analyze catastrophic forgetting in SDXL (fine-tuned with the safety reward), comparing DraFT and Focus-N-Fix on the PartiPrompt set in Appendix F.5.

Reward Model : Safety	**Full Evaluation Set (100 prompts)**				
Method/Metrics	Δ Safety Reward (↑)	PSNR (↑)	SSIM (↑)	LPIPS (↓)	Δ VNLI Score (↑)
SDXL vs DraFT	0.33	12.40	0.44	0.66	-0.031
SDXL vs Focus-N-Fix	0.21	20.92	0.76	0.33	0.013

Table 2. Objective comparison of Focus-N-Fix vs DraFT (SDXL).

5. Discussion

Non-Localizable quality aspects and sequential fine-tuning. Quality aspects, like aesthetics or style (and some misalignment cases), are image-level and cannot be localized. To enhance global quality aspects, we can set the mask to an all-ones matrix, reverting to conventional DRaFT. This enables improving the global quality, followed by using *Focus-N-Fix* to refine local quality aspects, such as artifacts. Since our method preserves global content while refining locally, improvements from the first step are largely retained. Sequential fine-tuning can address local issues, like reducing overly sexual content followed by artifact reduction.

6. Conclusion

We introduced a region-aware fine-tuning approach for T2I models that uses localization to make targeted improvements while preserving the structure of images from the original pre-trained model. We applied our method to address multiple image quality aspects, including artifacts, T2I misalignment, and safety issues like over-sexualization and violence. The experimental results demonstrate that *Focus-N-Fix* can effectively improve one quality aspect, with no or little degradation to other aspects. The proposed approach can be generalized to various reward models measuring different aspects of image quality, and it does not necessarily depend on dense reward models trained to predict regions. Most experiments in the current paper use SD v1.4 with DRaFT fine-tuning. Future work will extend *Focus-N-Fix* to more T2I models and fine-tuning methods.

References

[1] Everton L. Aleixo, Juan G. Colonna, Marco Cristo, and Everlandio Fernandes. Catastrophic forgetting in deep learning: A comprehensive taxonomy. *arxiv preprint arxiv:2312.10549*, 2023. 7, 4

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 3

[3] Andrew Bai, Chih-Kuan Yeh, Cho-Jui Hsieh, and Ankur Taly. Which pretrain samples to rehearse when finetuning pretrained models? *arxiv preprint arxiv:2402.08096*, 2024. 4

[4] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852, 2023. 2, 5, 6

[5] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: Misogyny, pornography, and malignant stereotypes. *arxiv preprint arxiv:2110.01963*, 2021. 3

[6] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023. 2

[7] Tim Brooks, Aleksander Holynski, and Alexei A Efros. InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 3

[8] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI: A jointly-scaled multilingual language-image model. *arxiv preprint arxiv:2209.06794*, 2023. 6

[9] Zhennan Chen, Yajie Li, Haofan Wang, Zhibo Chen, Zhengkai Jiang, Jun Li, Qian Wang, Jian Yang, and Ying Tai. Region-aware text-to-image generation via hard binding and soft refinement, 2024. 2

[10] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023. 6

[11] Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2, 3, 4, 5

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2

[13] Roel Dobbe. System safety and artificial intelligence. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, page 1584. Association for Computing Machinery, 2022. 3

[14] Ying Fan and Kangwook Lee. Optimizing DDPM sampling with shortcut fine-tuning. *arXiv preprint arXiv:2301.13362*, 2023. 2

[15] 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. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[16] Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the 2023 IEEE International Conference on Computer Vision*, 2023. 3

[17] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5111–5120, 2024. 3, 1, 4

[18] Jianshu Guo, Wenhao Chai, Jie Deng, Hsiang-Wei Huang, Tian Ye, Yichen Xu, Jiawei Zhang, Jenq-Neng Hwang, and Gaoang Wang. VersaT2I: Improving text-to-image models with versatile reward. *arxiv preprint arxiv:2403.18493*, 2024. 3

[19] Susan Hao, Piyush Kumar, Sarah Laszlo, Shivani Poddar, Bhaktipriya Radharapu, and Renee Shelby. Safety and fairness for content moderation in generative models. In *CVPR Workshop*, 2023. 1, 3, 5

[20] Susan Hao, Renee Shelby, Yuchi Liu, Hansa Srinivasan, Mukul Bhutani, Burcu Karagol Ayan, Ryan Poplin, Shivani Poddar, and Sarah Laszlo. Harm amplification in text-to-image models. *arxiv preprint arxiv:2402.01787*, 2024. 2, 3

[21] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 1

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017. 1

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3

[24] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 3

[25] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation

with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417, 2023. 1

[26] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 3

[27] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 36652–36663, 2023. 2, 1

[28] Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. VIEScore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12268–12290, Bangkok, Thailand, 2024. Association for Computational Linguistics. 1

[29] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[30] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. 2

[31] Seung Hyun Lee, Yinxiao Li, Junjie Ke, Innfarn Yoo, Han Zhang, Jiahui Yu, Qifei Wang, Fei Deng, Glenn Entis, Junfeng He, Gang Li, Sangpil Kim, Irfan Essa, and Feng Yang. Parrot: Pareto-optimal multi-reward reinforcement learning framework for text-to-image generation. *arxiv preprint arxiv:2401.05675*, 2024. 3

[32] Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19401–19411, 2024. 1, 2, 4, 5, 7

[33] Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, and Tong Zhang. Mitigating the alignment tax of RLHF. *arxiv preprint arxiv:2309.06256*, 2023. 4

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 1

[35] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. MACE: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6430–6440, 2024. 3

[36] Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. LLMScore: Unveiling the power of large language models in text-to-image synthesis evaluation. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[37] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[38] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022. 2

[39] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. 8, 4

[40] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 2

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 8

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2, 6

[43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[44] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. *Advances in Neural Information Processing Systems*, 29, 2016. 1

[45] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22522–22531, 2023. 3, 6

[46] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 4

[47] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arxiv preprint arxiv:1312.6034*, 2014. 2, 5

[48] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, pages 3008–3021, 2020. 1

[49] Jiao Sun, Deqing Fu, Yushi Hu, Su Wang, Royi Rassin, Da-Cheng Juan, Dana Alon, Charles Herrmann, Sjoerd van Steenkiste, Ranjay Krishna, and Cyrus Rashtchian. Dreamsync: Aligning text-to-image generation with image understanding feedback, 2023. 1

[50] 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. 1, 2

[51] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 1

[52] Zihao Wang, Chirag Nagpal, Jonathan Berant, Jacob Eisenstein, Alex D'Amour, Sanmi Koyejo, and Victor Veitch. Transforming and combining rewards for aligning large language models. *arXiv preprint arXiv:2402.00742*, 2024. 2, 3

[53] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 2, 6, 1

[54] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Better aligning text-to-image models with human preference. *arXiv preprint arXiv:2303.14420*, 1(3), 2023. 2

[55] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023. 1, 2

[56] Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. What you see is what you read? Improving text-image alignment evaluation. *arxiv preprint arxiv:2305.10400*, 2023. 7, 4

[57] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 6, 4

[58] Eric Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv preprint arXiv:2211.08332*, 2023. 3

[59] Lingzhi Zhang, Zhengjie Xu, Connelly Barnes, Yuqian Zhou, Qing Liu, He Zhang, Sohrab Amirghodsi, Zhe Lin, Eli Shechtman, and Jianbo Shi. Perceptual artifacts localization for image synthesis tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7579–7590, 2023. 2

[60] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 1

[61] Yinan Zhang, Eric Tzeng, Yilun Du, and Dmitry Kislyuk. Large-scale reinforcement learning for diffusion models. *arXiv preprint arXiv:2401.12244*, 2024. 2, 3

[62] Qingping Zheng, Ling Zheng, Yuanfan Guo, Ying Li, Songcen Xu, Jiankang Deng, and Hang Xu. Self-adaptive reality-guided diffusion for artifact-free super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25806–25816, 2024. 6