Aligning Text-to-Image Models using Human Feedback

Anonymous Author(s)
Affiliation
Address
email

Abstract

Deep generative models have shown impressive results in text-to-image synthesis. However, current text-to-image models often generate images that are poorly aligned with text prompts. We propose a fine-tuning method for aligning such models using human feedback, comprising three stages. First, we collect human feedback that assesses model output alignment over a set of diverse text prompts. We then use the human-labeled image-text dataset to train a reward function that predicts human feedback. Lastly, the text-to-image model is fine-tuned by maximizing *reward-weighted* likelihood to improve image-text alignment. Our method generates objects with specified colors, counts, and backgrounds more accurately than the pre-trained model. We also analyze several design choices and show that careful investigation of such design choices is crucial when balancing the alignment-fidelity tradeoff. Our results demonstrate the potential for learning from human feedback to significantly improve text-to-image models.

1 Introduction

Deep generative models have recently shown remarkable success in generating high-quality images from text prompts [33, 34, 37, 51, 35, 6]. This success has been driven in part by the scaling of deep generative models to large-scale datasets from the web such as LAION [39, 40]. However, major challenges remain in domains where large-scale text-to-image models fail to generate images that are well-aligned with text prompts [9, 25, 26]. For instance, current text-to-image models often fail to produce reliable visual text [26] and struggle with *compositional* image generation [9].

In language modeling, *learning from human feedback* has emerged as a powerful technique for aligning model behavior with human intent [53, 43, 46, 29, 30, 3]. Such methods first learn a *reward function* intended to reflect what humans care about in the task—they do so by exploiting *human feedback* on model outputs. The language model is then optimized using the learned reward function by a *reinforcement learning (RL)* algorithm, such as proximal policy optimization (PPO [41]). This *RL with human feedback (RLHF)* framework has successfully aligned large-scale language models (e.g., GPT-3 [5]) with complex human quality assessments.

Motivated by the success of RLHF in language domains, we propose a fine-tuning method for *aligning text-to-image models using human feedback*. Our method consists of the three main stages, illustrated in Figure 1: (1) We first generate diverse images from a set of text prompts designed to test output alignment of a text-to-image model. Specifically, we examine prompts where pre-trained models are more prone to error—generating objects with specific colors, counts, and backgrounds. We then collect binary human feedback assessing model outputs. (2) Using this human-labeled dataset, we train a reward function to predict human feedback given the image and text prompt. We propose an auxiliary task—identifying the original text prompt from a set of *perturbed* text prompts—to more effectively exploit human feedback for reward learning. This technique improves the generalization

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

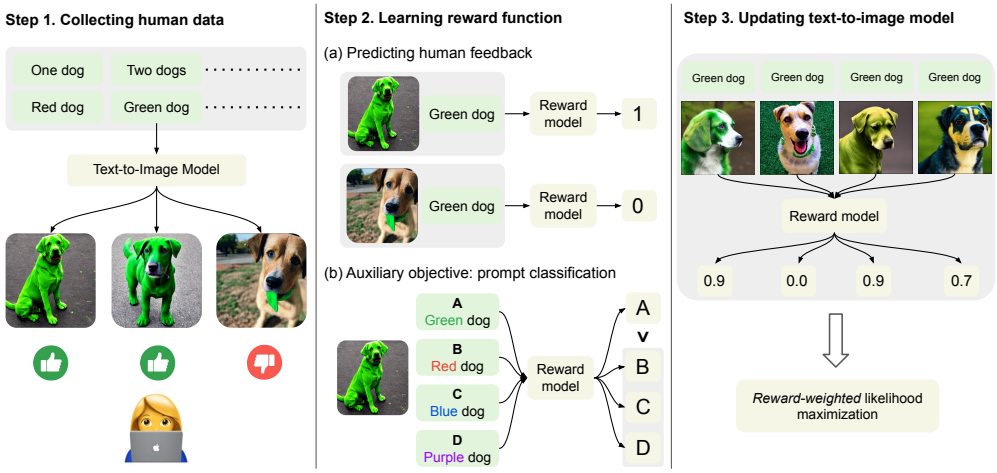


Figure 1: The three stages of our text-to-image fine-tuning method. (1) Multiple images are sampled from the model using the same text prompt, followed by collection of (binary) human feedback. (2) A reward function is learned from human these assessments to predict image-text alignment. We develop and exploit *prompt classification*, an auxiliary objective which identifies the original text prompt from among a set of *perturbed* text prompts. (3) We update the text-to-image model via reward-weighted likelihood maximization.

of the reward function to unseen images and text prompts. (3) We update the text-to-image model via *reward-weighted likelihood maximization* to better align it with human feedback. Unlike prior work that uses RL for optimization [43, 30], we update the model using semi-supervised learning to measure model-output quality w.r.t. the learned reward function.

We illustrate our approach by fine-tuning the *stable diffusion* model [35] using 27K image-text pairs with human feedback. Our fine-tuned model shows improvement in generating objects with specified colors, counts, and backgrounds. Moreover, it improves compositional generation (i.e., better generates unseen objects[1] given unseen combinations of color, count, and background prompts). We also observe that the learned reward function is better aligned with human assessments of alignment than the CLIP score [31] on tested text prompts. We analyze several design choices, such as using an auxiliary loss for reward learning, and the effect of using "diverse" datasets for fine-tuning. Our main contributions are as follows:

- We propose a simple yet efficient fine-tuning method for aligning a text-to-image model using human feedback.
- We show that fine-tuning with human feedback significantly improves the image-text alignment of a text-to-image model. On human evaluation, our model achieves up to 47% improvement in image-text alignment at the expense of mildly degraded image fidelity.
- We show that our learned reward function predicts human assessments of the quality more accurately than the CLIP score [31]. Moreover, we demonstrate that rejection sampling based on our learned reward function can significantly improve the image-text alignment.
- Naive fine-tuning with human feedback can significantly reduce the image fidelity, despite better alignment. We find that the careful investigation of several key design choices is vital in balancing the alignment-fidelity tradeoff.

Though our results do not address all the failure modes of the existing text-to-image models, our work highlights the potential of learning from human feedback for improving and better aligning these models.

[1] "Unseen objects" are those included in the pre-training dataset but not in the human evaluation dataset.

2 Related Work

We briefly review related work on text-to-image models and human-feedback learning.

Text-to-image models. Various deep generative models, such as variational auto-encoders [16], generative adversarial networks [11], auto-regressive models [44], and diffusion models [42, 14] have been proposed for image distributions. Combined with the large-scale language encoders [31, 32], these models have shown impressive results in text-to-image generation [34, 37, 51, 35, 6]. However, text-to-image models frequently struggle to generate images that are well-aligned with text prompts [9, 25, 26]. Liu et al. [26] show that current models fail to produce reliable visual text and often perform poorly w.r.t. compositional generation [9, 25]. Several techniques, such as character-aware text encoders [49, 26] and structured representations of language inputs [9], have been investigated to address these issues. We study learning from human feedback, which aligns text-to-image models directly using human feedback on model outputs.

Fine-tuning with few images [36, 20, 10] for personalization of text-to-image diffusion models is also related to our work. The DreamBooth approach [36] shows that text-to-image models can generate diverse images in a personalized way by fine-tuning with just a few images, while Kumari et al. [20] propose a more memory and computationally efficient method. In this work, we demonstrate that one can fine-tune text-to-image models using simple binary (good/bad) human feedback.

Learning from human feedback. Human feedback has been used to improve various AI systems (e.g., translation [2, 18], web question-answering [29], story generation [52], training RL agents without hand-designed rewards [28, 7, 45, 15, 21]); and to induce more truthful, less harmful instruction-following and dialogue [30, 3, 53, 46, 43, 24, 38, 4]. Increasingly, learning from human feedback has been used to improve language models and game agents. Several lines of work concurrent with ours have collected human-feedback datasets, trained reward functions [48, 47, 17], and shown that learned rewards are better-aligned with human evaluation than existing score functions (e.g., CLIP [31], BLIP [22]). We demonstrate that certain aspects of text-to-image models can be improved by fine-tuning the model using a reward function trained on human feedback.

3 Main Method

To improve the alignment of generated images with their text prompts, we fine-tune a pre-trained text-to-image model [34, 37, 35] by repeating the following steps (see Figure 1). We first generate a set of diverse images from a collection of text prompts designed to test various capabilities of the text-to-image model. Human raters provide binary feedback on these images (Section 3.1). Next we train a *reward model* to predict human feedback given a text prompt and an image as inputs (Section 3.2). Finally, we fine-tune the text-to-image model using *reward-weighted* log likelihood to improve text-image alignment (Section 3.3).

3.1 Human Data Collection

Image-text dataset. To test specific capabilities of a given text-to-image model, we consider three categories of text prompts that generate objects with a specified count, color, or background.[2] For each category, we generate prompts by combining a word or phrase from that category with some object; e.g., combining `green` (or `in a city`) with `dog`. We also consider combinations of the three categories (e.g., `two green dogs in a city`). From each prompt, we generate up to 60 images using a pre-trained text-to-image model—in this work, we use Stable Diffusion v1.5 [35].

Human feedback. We collect simple binary feedback from multiple human labelers on the image-text dataset. Labelers are presented with three images generated from the same prompt and are asked to assess whether each image is well-aligned with the prompt ("good") or not ("bad").[3] We use binary feedback given the simplicity of our prompts—the evaluation criterion are fairly clear. More informative human feedback, such as ranking [43, 30], should prove useful when more complex or subjective text prompts are used (e.g., artistic or open-ended generation).

[2]For simplicity, we consider a limited class of text categories in this work, deferring the study of broader and more complex categories to future work.

[3]Labelers are instructed to skip a query if it is hard to answer. Skipped queries are not used in training.

3.2 Reward Learning

To measure image-text alignment, we learn a *reward function* $r_\phi(\mathbf{x}, \mathbf{z})$, parameterized by ϕ, that maps the CLIP embeddings [31] of an image $\mathbf{x}$ and a text prompt $\mathbf{z}$ to a scalar value.[4] It is trained to predict human feedback $y \in \{0, 1\}$ (1 = good, 0 = bad). Formally, given the human feedback dataset $\mathcal{D}^{\texttt{human}} = \{(\mathbf{x}, \mathbf{z}, y)\}$, r_ϕ is trained by minimizing the mean-squared-error (MSE):

$$\mathcal{L}^{\texttt{MSE}}(\phi) = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{z},y)\sim\mathcal{D}^{\texttt{human}}} \big[\, (y - r_\phi(\mathbf{x}, \mathbf{z}))^2 \,\big]. \tag{1}$$

Prompt classification. Data augmentation can significantly improve the data-efficiency and performance of learning [19, 8]. To effectively exploit the feedback dataset, we design a simple data augmentation scheme and auxiliary loss for reward learning. For each image-text pair that has been labeled *good*, we generate $N - 1$ text prompts with different semantics than the original text prompt. For example, we might generate $\{\texttt{Blue dog}, \ldots, \texttt{Green dog}\}$ given the original prompt `Red dog`.[5] This process generates a dataset $\mathcal{D}^{\texttt{txt}} = \{(\mathbf{x}, \{\mathbf{z}_j\}_{j=1}^N, i')\}$ with N text prompts $\{\mathbf{z}_j\}_{j=1}^N$, including the original, for each image $\mathbf{x}$. We denote by $i' \in [N]$ the index of the original text prompt.

We use the augmented prompts in an auxiliary task, namely, classifying the original prompt for reward learning. Our prompt classifier uses the reward function r_ϕ as follows:

$$P_\phi(i \mid \mathbf{x}, \{\mathbf{z}_j\}_{j=1}^N) = \frac{\exp(r_\phi(\mathbf{x}, \mathbf{z}_i)/T)}{\sum_j \exp(r_\phi(\mathbf{x}, \mathbf{z}_j)/T)}, \quad \forall i \in [N],$$

where $T > 0$ is a temperature parameter. Our auxiliary loss is then defined as

$$\mathcal{L}^{\texttt{pc}}(\phi) = \mathop{\mathbb{E}}_{(\mathbf{x},\{\mathbf{z}_j\}_{j=1}^N,i')\sim\mathcal{D}^{\texttt{txt}}} \big[\mathcal{L}^{\texttt{CE}}\big(P_\phi(i \mid \mathbf{x}, \{\mathbf{z}_j\}_{j=1}^N), i'\big)\big], \tag{2}$$

where $\mathcal{L}^{\texttt{CE}}$ is the standard cross-entropy loss. This encourages r_ϕ to produce low values for prompts with different semantics than the original. Our experiments show this auxiliary loss improves generalization to unseen images and text prompts. Finally, we define our reward learning loss by combining the losses in (1) and (2) as

$$\mathcal{L}^{\texttt{reward}}(\phi) = \mathcal{L}^{\texttt{MSE}}(\phi) + \lambda \mathcal{L}^{\texttt{pc}}(\phi), \tag{3}$$

where λ is a penalty parameter. The pseudo-code for reward learning is provided in Appendix E.

3.3 Updating the Text-to-Image Model

We use our learned r_ϕ to update the text-to-image model p with parameters θ by minimizing the loss

$$\mathcal{L}(\theta) = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{z})\sim\mathcal{D}^{\texttt{model}}} \big[- r_\phi(\mathbf{x}, \mathbf{z}) \log p_\theta(\mathbf{x}|\mathbf{z}) \big] + \beta \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{z})\sim\mathcal{D}^{\texttt{pre}}} \big[- \log p_\theta(\mathbf{x}|\mathbf{z}) \big]. \tag{4}$$

In (4), $\mathcal{D}^{\texttt{model}}$ is the model-generated dataset (i.e., images generated by the text-to-image model on the tested text prompts), $\mathcal{D}^{\texttt{pre}}$ is the *pre-training dataset*, and β is a penalty parameter. The first term in (4) minimizes the *reward-weighted* negative log-likelihood (NLL) on $\mathcal{D}^{\texttt{model}}$. To increase diversity, we collect an unlabeled dataset $\mathcal{D}^{\texttt{unlabel}}$ by generating more images from the text-to-image model, and use both the human-labeled dataset $\mathcal{D}^{\texttt{human}}$ and the unlabeled dataset $\mathcal{D}^{\texttt{unlabel}}$ for training, i.e., $\mathcal{D}^{\texttt{model}} = \mathcal{D}^{\texttt{human}} \cup \mathcal{D}^{\texttt{unlabel}}$. By evaluating the quality of the outputs using a reward function aligned with the text prompts, this term improves the image-text alignment of the model.

Minimizing the loss in (4) in autoregressive models like Parti [51] is relatively straightforward. However, this minimization is challenging for diffusion models, such as Imagen [37], Stable Diffusion [35], and Dalle-2 [34], as computing the exact log-likelihood is difficult in these models. In this work, we minimize the reward-weighted MSE loss [14] to fine-tune diffusion-based text-to-image models (Stable Diffusion in our experiments).

Typically, the diversity of the model-generated dataset $\mathcal{D}^{\texttt{model}}$ is limited, which can result in overfitting. To mitigate this, similar to Ouyang et al. [30], we also minimize the *pre-training loss*, the second

[4] To improve generalization ability, we use CLIP embeddings pre-trained on various image-text samples.

[5] We use a rule-based strategy to generate different text prompts (see Appendix E for more details).

Category	Total # of images	Human feedback (%)			Examples
		Good	Bad	Skip	
Count	6480	34.4	61.0	4.6	`One dog; Two dogs; Three dogs; Four dogs; Five dogs;`
Color	3480	70.4	20.8	8.8	`A green colored dog; A red colored dog;`
Background	2400	66.9	33.1	0.0	`A dog in the forest; A dog on the moon;`
Combination	15168	35.8	59.9	4.3	`Two blue dogs in the forest; Five white dogs in the city;`
Total	27528	46.5	48.5	5.0	

Table 1: Details of image-text datasets and human feedback.

term in (4). This reduces NLL on the pre-training dataset $\mathcal{D}^{\texttt{pre}}$. In our experiments, we observed that regularization in the loss function $\mathcal{L}(\theta)$ in (4) enables the generation of more natural images.

Different objective functions and algorithms (e.g., PPO; [41]) could be considered for updating the text-to-image model, much like RLHF fine-tuning [30]. We believe RLHF fine-tuning may lead to better models because it uses online sample generation during updates and KL-regularization over the prior model. However, RL usually requires extensive hyperparameter tuning and engineering, thus, we defer proper investigation of RLHF fine-tuning in text-to-image models to future work.

4 Experiments

In this section, we describe a set of experiments designed to test the efficacy of our approach to fine-tune text-to-image models with human feedback and report our findings and results.

4.1 Experimental Setup

Models. For our baseline generative model, we use Stable Diffusion v1.5 [35], which has been pre-trained on large image-text datasets [39, 40].[6] For fine-tuning, we freeze the CLIP language encoder [31] and fine-tune only the diffusion module. For the reward model, we use ViT-L/14 CLIP model [31] to extract image and text embeddings and train an MLP using these embeddings as input. Experimental details (e.g., model architectures, hyperparameters) are given in Appendix B.

Datasets. From a set of 2700 English prompts (see Table 1 for some instances), we generate 27K images using the stable diffusion model (see Appendix C for further details). Table 1 shows the feedback distribution provided by multiple human labelers, which has been class-balanced. We note that the stable diffusion model struggles to generate the number of objects specified by the prompt, but reliably generates specified colors and backgrounds.

We use 23K samples for training and leave the remaining samples for validation. We also use 16K unlabeled samples for the reward-weighted loss and a 625K subset[7] of LAION-5B [40] filtered by an aesthetic score predictor[8] for the pre-training loss.

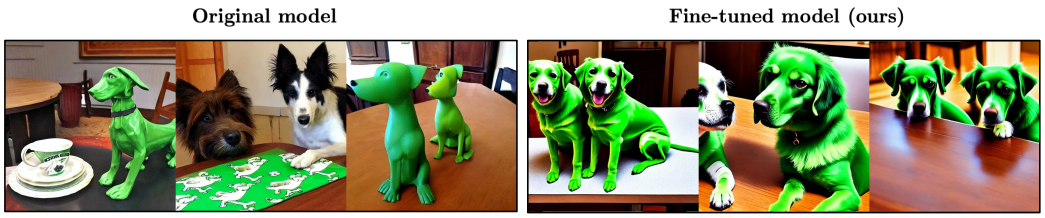
4.2 Text-Image Alignment Results

Human evaluation. We measure human ratings of image alignment with 120 text prompts (60 seen text prompts and 60 unseen text prompts), testing the ability of the models to render different colors, number of objects, and backgrounds (see Appendix C for the full set of prompts). Given two (anonymized) sets of images, one from our fine-tuned model and one from the stable diffusion model,

[6] Our fine-tuning method can be used readily with other text-to-image models, such as Imagen [37], Parti [51], and Dalle-2 [34].

[7] https://huggingface.co/datasets/ChristophSchuhmann/improved_aesthetics_6.5plus

[8] https://github.com/christophschuhmann/improved-aesthetic-predictor



(a) Seen text prompt: `Two green dogs on the table.`

(b) Unseen text prompt (unseen object): `Four tigers in the field.`

(c) Unseen text prompt (artistic generation): `Oil painting of sunflowers.`

Figure 2: Samples from the original Stable Diffusion model (left) and our fine-tuned model (right). (a) Our model generates a high-quality seen object (`dog`) with specified color, count, and background. (b) Our model generates an unseen object (`tiger`) with specified color, count, and background. (c) Our model generates reasonable images even given unseen text categories (artistic generation).

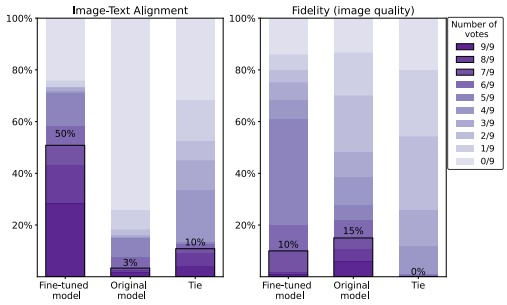
we ask human raters to assess which is better w.r.t. image-text alignment and fidelity (i.e., image quality). We ask raters to declare a *tie* if they have similar quality. Each query is evaluated by 9 independent human raters. We show the percentage of queries based on the number of positive votes.

As shown in Figure 3(a), our method significantly improves image-text alignment relative to the original model. Specifically, 50% of samples from our model receive at least two-thirds of the votes (7 or more positive votes) for image-text alignment. However, fine-tuning somewhat degrades image fidelity (15% compared to 10%). We expect that this is because (i) we asked the labelers to provide feedback mainly on alignment, (ii) the diversity of our human data is limited, and (iii) we used a small subset of the pre-training dataset for fine-tuning.[9] This issue can presumably be mitigated with greater number of raters and a larger pre-training dataset.

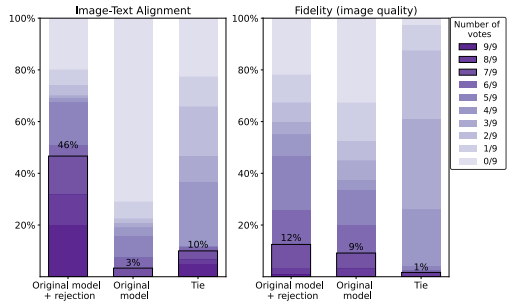
Qualitative comparison. Figure 2 shows image samples from the original model and our fine-tuned counterpart (see Appendix A for more image examples). While the original often generates images with missing details (e.g., color, background or count) (Figure 2(a)), our model generates objects that adhere to the prompt-specified colors, counts, and backgrounds. Of special note, our model generates high-quality images on unseen text prompts that specify unseen objects (Figure 2(b)). Our model also generates reasonable images given unseen text categories, such as artistic generation (Figure 2(c)).

Despite its strong performance, we do observe some difficulties with our fine-tuned model. For certain text prompts, our fine-tuned model generates oversaturated and non-photorealistic images. Our model occasionally duplicates entities within the generated images, or produces images for the same prompt with low diversity. We believe these issues can be addressed with larger (and more diverse) human datasets and better optimization (e.g., with RL).

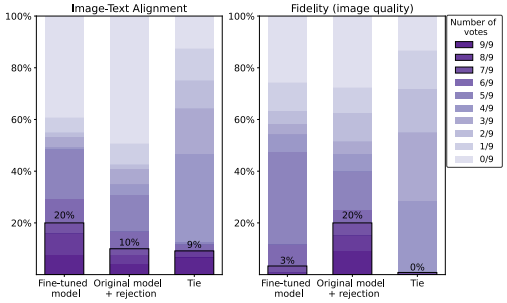
[9] Similar issue, which is akin to the *alignment tax*, has been observed in language domains [1, 30].



(a) Original model vs. Fine-tuned model.

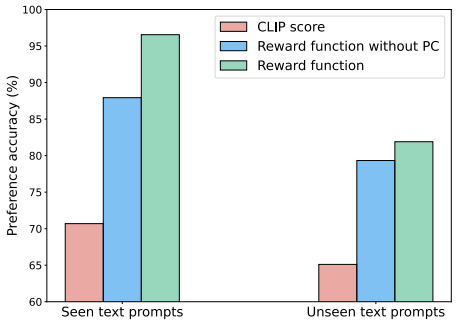


(b) Original model with/without Rejection sampling.

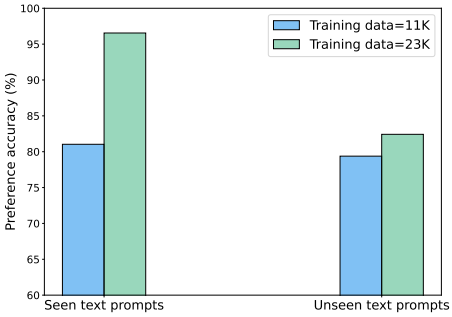


(c) Fine-tuned vs. Original + Rejection sampling.

Figure 3: Human evaluation on 120 tested text prompts (60 seen and 60 unseen). We generate two sets of images with the same text prompt. Human raters indicate which set is better (or if it is a tie, i.e., the two sets are similar), in terms of both image-text alignment and image fidelity. Each query is evaluated by nine raters and we report the percentage of queries that receive a given number of positive votes (0–9). We also highlight the percentage of queries receiving at least two-thirds (7 or more) of the positive votes in the black box. (a) Comparison of the fine-tuned and original models. (b) Comparison of the original model with and without the rejection sampling. (c) Comparison of the fine-tuned model and the original model with rejection sampling.



(a) Comparison with CLIP score [31]

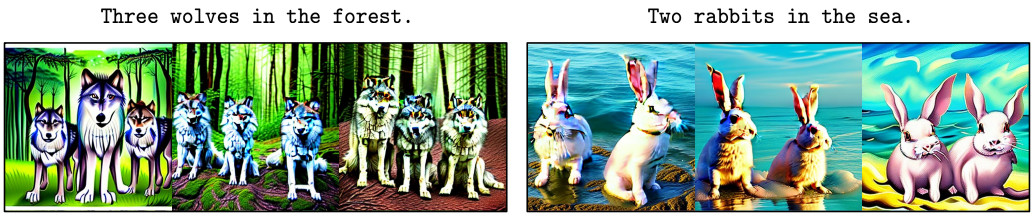


(b) Performance of reward function vs. data size

Figure 4: (a) Accuracy of CLIP score [31] and our reward functions when predicting the assessments of human labelers. We consider a variant of our reward function which is *not* trained with the prompt classification (PC) loss in (2). (b) Performance of reward functions when varying the size of the training dataset.

4.3 Results on Reward Learning

Predicting human preferences. We investigate the quality of our learned reward function by evaluating its prediction of the assessments of human labelers. Given two images $\mathbf{x}_1$ and $\mathbf{x}_2$ from the same text prompt $\mathbf{z}$, we check whether our reward r_ϕ generates a higher score for the human-preferred



(a) Fine-tuned model only with human-labeled dataset.

(b) Fine-tuned model with human-labeled and unlabeled datasets.

(c) Fine-tuned model with human-labeled, unlabeled and pre-training datasets.

Figure 5: Samples from fine-tuned models trained with different datasets on unseen text prompts. (a) Model fine-tuned with only human data generates low-quality images due to overfitting. (b) Unlabeled samples improve the quality of generated images. (c) Fine-tuned model can generate high-fidelity images by using the pre-training dataset.

image, i.e., $r_\phi(\mathbf{x}_1, \mathbf{z}) > r_\phi(\mathbf{x}_2, \mathbf{z})$ when rater prefers $\mathbf{x}_1$ to $\mathbf{x}_2$. As a baseline, we compare it with the CLIP score [12], which measures image-text similarity in the CLIP embedding space [31].

Figure 4(a) compares the accuracy of r_ϕ and the CLIP score on unseen images from both seen and unseen text prompts. Our reward (green) more accurately predicts human evaluation than the CLIP score (red), hence is better aligned with typical human assessments. To show the benefit of our auxiliary loss (prompt classification) in (2), we also assess a variant of our reward function which ignores this loss (blue). The results show that the auxiliary classification task improves reward performance on both seen and unseen text prompts. The gain from the auxiliary loss clearly shows the importance of text diversity and our auxiliary loss in improving data efficiency. Although our reward function is more accurate than the CLIP score, its performance on unseen text prompts ($\sim$80%) suggests that it may be necessary to use more diverse and larger human datasets.

Rejection sampling. Similar to Parti [51] and DALL-E [33], we evaluate a rejection sampling technique, which selects the best output w.r.t. the learned reward function.[10] Specifically, we generate 16 images per text prompt from the original stable diffusion model and select the four with the greatest reward scores. We compare these to four randomly sampled images in Figure 3(b). Rejection sampling significantly improves image-text alignment (46% with two-thirds preference vote by raters) without sacrificing image fidelity. This result illustrates the significance of learning a reward function in improving text-to-image models even *without any fine-tuning*.

We also compare our fine-tuned model to the original with rejection sampling in Figure 3(c).[11] Our fine-tuned model achieves a 10% gain in image-text alignment (20%-10% two-thirds vote) but sacrifices 17% in image fidelity (3%-20% two-thirds vote). However, as discussed in Section 4.2,

[10] Parti and DALL-E use similarity scores of image and text embeddings from CoCa [50] and CLIP [31], respectively.

[11] We note that rejection sampling can also be applied on top of our fine-tuned model.

	Original model	Fine-tuned model w.o unlabeled & pre-train	Fine-tuned model w.o pre-train	Fine-tuned model
FID on MS-CoCo (↓)	13.97	26.59	21.02	16.76
Average rewards on tested prompts (↑)	0.43	0.69	0.79	0.79

Table 2: Comparison with the original Stable Diffusion. To evaluate image fidelity, we measure FID scores on MS-CoCo. To evaluate image-text alignment, we measure reward scores and CLIP scores on 120 tested text prompts. (↑ (resp., ↓) indicates that higher (resp., lower) metrics are better.)

we expect degradation in fidelity to be mitigated with larger human datasets and better hyper-parameters. Note also that rejection sampling has several drawbacks, including increased inference-time computation and the inability to improve the model (since it is a post-processing technique).

4.4 Ablation Studies

Effects of the size of the human dataset. To investigate how the quality of human data affects reward learning, we conduct an ablation study, reducing the number of images per text prompt by half before training the reward function. Figure 4(b) shows that model accuracy decreases on both seen and unseen prompts as data size decreases, clearly demonstrating the importance of diversity and the size of the rater data.

Effects of using diverse datasets. To verify the importance of data diversity, we incrementally include unlabeled and pre-training datasets during fine-tuning. We measure the reward score (image-text alignment) on 120 tested text prompts and FID score [13]—the similarity between generated images and real images—on MS-CoCo validation data [23]. Table 2 shows that FID score is significantly reduced when the model is fine-tuned using only human data, despite better image-text alignment. However, by adding the unlabeled and pre-training datasets, FID score is improved without impacting the image-text alignment. We provide image samples from unseen text prompts in Figure 5. We observe that fine-tuned models generate more natural images when exploiting more diverse datasets.

5 Discussion

In this work, we showed that fine-tuning with human feedback can effectively improve the image-text alignment in three domains: generating objects with a specified count, color, or backgrounds. We analyzed several design choices (such as using an auxiliary loss and collecting diverse training data) and found that it is challenging to balance the alignment-fidelity tradeoff without careful investigation of such design choices. While our results do not address all the failure modes of existing text-to-image models, our work serves as a starting point for the deeper study of learning from human feedback for improving text-to-image models.

Limitations, future directions. Several limitations of our work suggest interesting future directions:

More nuanced human feedback. Some of the poor image generation we observe (e.g., highly saturated image colors) are likely due to similar images being highly ranked in our training set. Instructing raters to watch for a more diverse set of failure modes (e.g., oversaturated colors, unrealistic animal anatomy, physics violations) should improve performance along these dimensions.

Different objectives and algorithms. To update a text-to-image model, we used a reward-weighted likelihood maximization. However, similar to prior work in language domains [30], it would be an interesting direction to use RL [41]. We expect RLHF fine-tuning to lead to better models because (a) it uses online sample generation during updates and (b) KL-regularization over the prior model can mitigate overfitting to the reward function.

References

[1] Askell, Amanda, Bai, Yuntao, Chen, Anna, Drain, Dawn, Ganguli, Deep, Henighan, Tom, Jones, Andy, Joseph, Nicholas, Mann, Ben, DasSarma, Nova, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

[2] Bahdanau, Dzmitry, Brakel, Philemon, Xu, Kelvin, Goyal, Anirudh, Lowe, Ryan, Pineau, Joelle, Courville, Aaron, and Bengio, Yoshua. An actor-critic algorithm for sequence prediction. *arXiv preprint arXiv:1607.07086*, 2016.

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

[4] Bai, Yuntao, Kadavath, Saurav, Kundu, Sandipan, Askell, Amanda, Kernion, Jackson, Jones, Andy, Chen, Anna, Goldie, Anna, Mirhoseini, Azalia, McKinnon, Cameron, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[5] Brown, Tom B, Mann, Benjamin, Ryder, Nick, Subbiah, Melanie, Kaplan, Jared, Dhariwal, Prafulla, Neelakantan, Arvind, Shyam, Pranav, Sastry, Girish, Askell, Amanda, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[6] Chang, Huiwen, Zhang, Han, Barber, Jarred, Maschinot, AJ, Lezama, Jose, Jiang, Lu, Yang, Ming-Hsuan, Murphy, Kevin, Freeman, William T, Rubinstein, Michael, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

[7] Christiano, Paul F, Leike, Jan, Brown, Tom, Martic, Miljan, Legg, Shane, and Amodei, Dario. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 2017.

[8] Cubuk, Ekin D, Zoph, Barret, Mane, Dandelion, Vasudevan, Vijay, and Le, Quoc V. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[9] Feng, Weixi, He, Xuehai, Fu, Tsu-Jui, Jampani, Varun, Akula, Arjun, Narayana, Pradyumna, Basu, Sugato, Wang, Xin Eric, and Wang, William Yang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022.

[10] Gal, Rinon, Alaluf, Yuval, Atzmon, Yuval, Patashnik, Or, Bermano, Amit H, Chechik, Gal, and Cohen-Or, Daniel. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[11] Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

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

[13] Heusel, Martin, Ramsauer, Hubert, Unterthiner, Thomas, Nessler, Bernhard, and Hochreiter, Sepp. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, 2017.

[14] Ho, Jonathan, Jain, Ajay, and Abbeel, Pieter. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

[15] Ibarz, Borja, Leike, Jan, Pohlen, Tobias, Irving, Geoffrey, Legg, Shane, and Amodei, Dario. Reward learning from human preferences and demonstrations in atari. In *Advances in Neural Information Processing Systems*, 2018.

[16] Kingma, Diederik P and Welling, Max. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[17] Kirstain, Yuval, Polyak, Adam, Singer, Uriel, Matiana, Shahbuland, Penna, Joe, and Levy, Omer. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569*, 2023.

[18] Kreutzer, Julia, Khadivi, Shahram, Matusov, Evgeny, and Riezler, Stefan. Can neural machine translation be improved with user feedback? *arXiv preprint arXiv:1804.05958*, 2018.

[19] Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[20] Kumari, Nupur, Zhang, Bingliang, Zhang, Richard, Shechtman, Eli, and Zhu, Jun-Yan. Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2212.04488*, 2022.

[21] Lee, Kimin, Smith, Laura, and Abbeel, Pieter. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. In *International Conference on Machine Learning*, 2021.

[22] Li, Junnan, Li, Dongxu, Xiong, Caiming, and Hoi, Steven. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022.

[23] Lin, Tsung-Yi, Maire, Michael, Belongie, Serge, Hays, James, Perona, Pietro, Ramanan, Deva, Dollár, Piotr, and Zitnick, C Lawrence. Microsoft coco: Common objects in context. In *European conference on computer vision*, 2014.

[24] Liu, Hao, Sferrazza, Carmelo, and Abbeel, Pieter. Chain of hindsight aligns language models with feedback. *arXiv preprint arXiv: Arxiv-2302.02676*, 2023.

[25] Liu, Nan, Li, Shuang, Du, Yilun, Torralba, Antonio, and Tenenbaum, Joshua B. Compositional visual generation with composable diffusion models. *arXiv preprint arXiv:2206.01714*, 2022.

[26] Liu, Rosanne, Garrette, Dan, Saharia, Chitwan, Chan, William, Roberts, Adam, Narang, Sharan, Blok, Irina, Mical, RJ, Norouzi, Mohammad, and Constant, Noah. Character-aware models improve visual text rendering. *arXiv preprint arXiv:2212.10562*, 2022.

[27] Loshchilov, Ilya and Hutter, Frank. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[28] MacGlashan, James, Ho, Mark K, Loftin, Robert, Peng, Bei, Roberts, David, Taylor, Matthew E, and Littman, Michael L. Interactive learning from policy-dependent human feedback. In *International Conference on Machine Learning*, 2017.

[29] Nakano, Reiichiro, Hilton, Jacob, Balaji, Suchir, Wu, Jeff, Ouyang, Long, Kim, Christina, Hesse, Christopher, Jain, Shantanu, Kosaraju, Vineet, Saunders, William, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

[30] Ouyang, Long, Wu, Jeff, Jiang, Xu, Almeida, Diogo, Wainwright, Carroll L, Mishkin, Pamela, Zhang, Chong, Agarwal, Sandhini, Slama, Katarina, Ray, Alex, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

[31] Radford, Alec, Kim, Jong Wook, Hallacy, Chris, Ramesh, Aditya, Goh, Gabriel, Agarwal, Sandhini, Sastry, Girish, Askell, Amanda, Mishkin, Pamela, Clark, Jack, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[32] Raffel, Colin, Shazeer, Noam, Roberts, Adam, Lee, Katherine, Narang, Sharan, Matena, Michael, Zhou, Yanqi, Li, Wei, Liu, Peter J, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

[33] Ramesh, Aditya, Pavlov, Mikhail, Goh, Gabriel, Gray, Scott, Voss, Chelsea, Radford, Alec, Chen, Mark, and Sutskever, Ilya. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 2021.

[34] Ramesh, Aditya, Dhariwal, Prafulla, Nichol, Alex, Chu, Casey, and Chen, Mark. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

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

[36] Ruiz, Nataniel, Li, Yuanzhen, Jampani, Varun, Pritch, Yael, Rubinstein, Michael, and Aberman, Kfir. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022.

[37] Saharia, Chitwan, Chan, William, Saxena, Saurabh, Li, Lala, Whang, Jay, Denton, Emily, Ghasemipour, Seyed Kamyar Seyed, Ayan, Burcu Karagol, Mahdavi, S Sara, Lopes, Rapha Gontijo, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, 2022.

[38] Scheurer, Jérémy, Campos, Jon Ander, Chan, Jun Shern, Chen, Angelica, Cho, Kyunghyun, and Perez, Ethan. Training language models with language feedback. *arXiv preprint arXiv: Arxiv-2204.14146*, 2022.

[39] Schuhmann, Christoph, Vencu, Richard, Beaumont, Romain, Kaczmarczyk, Robert, Mullis, Clayton, Katta, Aarush, Coombes, Theo, Jitsev, Jenia, and Komatsuzaki, Aran. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[40] Schuhmann, Christoph, Beaumont, Romain, Vencu, Richard, Gordon, Cade, Wightman, Ross, Cherti, Mehdi, Coombes, Theo, Katta, Aarush, Mullis, Clayton, Wortsman, Mitchell, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[41] Schulman, John, Wolski, Filip, Dhariwal, Prafulla, Radford, Alec, and Klimov, Oleg. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[42] Sohl-Dickstein, Jascha, Weiss, Eric, Maheswaranathan, Niru, and Ganguli, Surya. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015.

[43] Stiennon, Nisan, Ouyang, Long, Wu, Jeff, Ziegler, Daniel M, Lowe, Ryan, Voss, Chelsea, Radford, Alec, Amodei, Dario, and Christiano, Paul. Learning to summarize from human feedback. *arXiv preprint arXiv:2009.01325*, 2020.

[44] Van Den Oord, Aäron, Kalchbrenner, Nal, and Kavukcuoglu, Koray. Pixel recurrent neural networks. In *International conference on machine learning*, pp. 1747–1756, 2016.

[45] Warnell, Garrett, Waytowich, Nicholas, Lawhern, Vernon, and Stone, Peter. Deep tamer: Interactive agent shaping in high-dimensional state spaces. In *Conference on Artificial Intelligence*, 2018.

[46] Wu, Jeff, Ouyang, Long, Ziegler, Daniel M, Stiennon, Nisan, Lowe, Ryan, Leike, Jan, and Christiano, Paul. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*, 2021.

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

[48] Xu, Jiazheng, Liu, Xiao, Wu, Yuchen, Tong, Yuxuan, Li, Qinkai, Ding, Ming, Tang, Jie, and Dong, Yuxiao. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023.

[49] Xue, Linting, Barua, Aditya, Constant, Noah, Al-Rfou, Rami, Narang, Sharan, Kale, Mihir, Roberts, Adam, and Raffel, Colin. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022.

[50] Yu, Jiahui, Wang, Zirui, Vasudevan, Vijay, Yeung, Legg, Seyedhosseini, Mojtaba, and Wu, Yonghui. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[51] Yu, Jiahui, Xu, Yuanzhong, Koh, Jing Yu, Luong, Thang, Baid, Gunjan, Wang, Zirui, Vasudevan, Vijay, Ku, Alexander, Yang, Yinfei, Ayan, Burcu Karagol, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

[52] Zhou, Wangchunshu and Xu, Ke. Learning to compare for better training and evaluation of open domain natural language generation models. In *Conference on Artificial Intelligence*, 2020.

[53] Ziegler, Daniel M, Stiennon, Nisan, Wu, Jeffrey, Brown, Tom B, Radford, Alec, Amodei, Dario, Christiano, Paul, and Irving, Geoffrey. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.