Rich Human Feedback for Text-to-Image Generation

Youwei Liang[*†1], Junfeng He[*‡2], Gang Li[*‡2], Peizhao Li[†5], Arseniy Klimovskiy[2], Nicholas Carolan[2], Jiao Sun[†§3], Jordi Pont-Tuset[2], Sarah Young[2], Feng Yang[2], Junjie Ke[2], Krishnamurthy Dj Dvijotham[2], Katherine M. Collins[†4], Yiwen Luo[2], Yang Li[2], Kai J Kohlhoff[2], Deepak Ramachandran[2], and Vidhya Navalpakkam[2]

[1]University of California, San Diego
[2]Google Research
[3]University of Southern California
[4]University of Cambridge
[5]Brandeis University

Abstract

Recent Text-to-Image (T2I) generation models such as Stable Diffusion and Imagen have made significant progress in generating high-resolution images based on text descriptions. However, many generated images still suffer from issues such as artifacts/implausibility, misalignment with text descriptions, and low aesthetic quality. Inspired by the success of Reinforcement Learning with Human Feedback (RLHF) for large language models, prior works collected human-provided scores as feedback on generated images and trained a reward model to improve the T2I generation. In this paper, we enrich the feedback signal by (i) marking image regions that are implausible or misaligned with the text, and (ii) annotating which words in the text prompt are misrepresented or missing on the image. We collect such rich human feedback on 18K generated images (RichHF-18K) and train a multimodal transformer to predict the rich feedback automatically. We show that the predicted rich human feedback can be leveraged to improve image generation, for example, by selecting high-quality training data to finetune and improve the generative models, or by creating masks with predicted heatmaps to inpaint the problematic regions. Notably, the improvements generalize to models (Muse) beyond those used to generate the images on which human feedback data were collected (Stable Diffusion variants). The RichHF-18K data set will be released in our GitHub repository: https://github.com/google-research/google-research/tree/master/richhf_18k.

*Co-first authors, equal technical contribution
†The work was done during an internship at Google.
Email: youwei@ucsd.edu
‡Corresponding authors, equal leading contribution.
Email: {junfenghe,leebird}@google.com
§Currently affiliated with Google Gemini Team

1. Introduction

Text-to-image (T2I) generation models [12, 17, 41, 42, 56, 58, 59] are rapidly becoming a key to content creation in various domains, including entertainment, art, design, and advertising, and are also being generalized to image editing [4, 27, 44, 50], video generation [23, 35, 53], among many other applications. Despite significant recent advances, the outputs still usually suffer from issues such as artifacts/implausibility, misalignment with text descriptions, and low aesthetic quality [30, 52, 54]. For example, in the Pick-a-Pic dataset [30], which mainly consists of images generated by Stable Diffusion model variants , many images (*e.g.* Fig. 1) contain distorted human/animal bodies (*e.g.* human hands with more than five fingers), distorted objects and implausibility issues such as a floating lamp. Our human evaluation experiments find that only ∼10% of the generated images in the dataset are free of artifacts and implausibility. Similarly, text-image misalignment issues are common too, *e.g.*, the prompt is "*a man jumping into a river*" but the generated image shows the man standing.

Existing automatic evaluation metrics for generated images, however, including the well-known IS [43] and FID [20], are computed over distributions of images and may not reflect nuances in individual images. Recent research has collected human preferences/ratings to evaluate the quality of generated images and trained evaluation models to predict those ratings [30, 52, 54], notably ImageReward [54] or Pick-a-Pic [30]. While more focused, these metrics still summarize the quality of one image into a single numeric score. In terms of prompt-image alignment, there are also seminal single-score metrics such as CLIPScore [19] and more recent question generation and answering pipelines [8, 10, 24, 57]. While more calibrated and explainable, these are expensive and complex models that still do not localize the regions of misalignment in the image.

In this paper, we propose a dataset and a model of fine-grained multi-faceted evaluations that are interpretable and attributable (*e.g.*, to regions with artifacts/implausibility or image-text misalignments), which provide a much richer understanding of the image quality than single scalar scores. As a first contribution, we collect a dataset of Rich Human Feedback on 18K images (RichHF-18K), which contains (i) point annotations on the image that highlight regions of implausibility/artifacts, and text-image misalignment; (ii) labeled words on the prompts specifying the missing or misrepresented concepts in the generated image; and (iii) four types of fine-grained scores for image plausibility, text-image alignment, aesthetics, and overall rating. Equipped with RichHF-18K, we design a multimodal transformer model, which we coin as Rich Automatic Human Feedback (RAHF) to learn to predict these rich human annotations on generated images and their associated text prompt. Our model can therefore predict implausibility and misalignment regions, misaligned keywords, as well as fine-grained scores. This not only provides reliable ratings, but also more detailed and explainable insights about the quality of the generated images. To the best of our knowledge, this is the first rich feedback dataset and model for state-of-the-art text-to-image generation models, providing an automatic and explainable pipeline to evaluate T2I generation.

The main contributions of this paper are summarized below:

1. The first Rich Human Feedback dataset (RichHF-18K) on generated images (consisting of fine-grained scores, implausibility(artifact)/misalignment image regions, and misalignment keywords), on 18K Pick-a-Pic images.
2. A multimodal Transformer model (RAHF) to predict rich feedback on generated images, which we show to be highly correlated with the human annotations on a test set.
3. We further demonstrate the usefulness of the predicted rich human feedback by RAHF to improve image generation: (i) by using the predicted heatmaps as masks to inpaint problematic image regions and (ii) by using the predicted scores to help finetune image generation models (like Muse [6]), *e.g.*, via selecting/filtering finetuning data, or as reward guidance. We show that in both cases we obtain better images than with the original model.
4. The improvement on the Muse model, which differs from the models that generated the images in our training set, shows the good generalization capacity of our RAHF model.

2. Related works

Text-to-image generation Text-to-image (T2I) generation models have evolved and iterated through several popular model architectures in the deep learning era. An early work is the Generative Adversarial Network (GAN) [3, 16, 26], which trains a generator for image generation and a discriminator to distinguish between real and generated images, in parallel (also see [32, 38, 47, 55, 60, 62] among others). Another category of generation models develops from variational auto-encoders (VAEs) [21, 29, 48], which optimize evidence lower bound (ELBO) for the likelihood of the image data.

Most recently, Diffusion Models (DMs) [22, 36, 41, 46] have emerged as the state-of-the-art (SOTA) for Image Generation [13]. DMs are trained to generate images progressively from random noise, with the ability to capture more diversity than GANs and achieve good sample quality [13]. Latent Diffusion Models [41] are a further refinement that performs the diffusion process in a compact latent space for more efficiency.

Text-to-image evaluation and reward models There has been much recent work on evaluation of text-to-image models along many dimensions [9, 25, 30, 31, 37, 51, 52, 54]. Xu et al. [54] collect a human preference dataset by requesting users to rank multiple images and rate them according to their quality. They trained a reward model ImageReward for human preference learning, and proposed Reward Feedback Learning (ReFL) for tuning diffusion models with the ImageReward model. Kirstain et al. [30] built a web application to collect human preferences by asking users to choose the better image from a pair of generated images, resulting in a dataset called Pick-a-Pic with more than 500K examples generated by T2I models such as Stable Diffusion 2.1, Dreamlike Photoreal 2.05, and Stable Diffusion XL variants. They leveraged the human preference dataset to train a CLIP-based [39] scoring function, called PickScore, to predict human preferences. Huang et al. [25] proposed a benchmark called T2I-CompBench for evaluating text-to-image models, which consists of 6,000 text prompts describing attribute binding, object relationships, and complex compositions. They utilized multiple pretrained vision-language models such as CLIP [39] and BLIP [34] to calculate multiple evaluation metrics. Wu et al. [51, 52] collected a large scale dataset of human choices on generated images and utilized the dataset to train a classifier that outputs a Human Preference Score (HPS). They showed improvement in image generation by tuning Stable Diffusion with the HPS. Recently, Lee [31] proposed a holistic evaluation for T2I models with multiple fine-grained metrics.

Despite these valuable contributions, most existing works only use binary human ratings or preference rank-

Figure 1. **An illustration of our annotation UI**. Annotators mark points on the image to indicate artifact/implausibility regions (red points) or misaligned regions (blue points) w.r.t the text prompt. Then, they click on the words to mark the misaligned keywords (underlined and shaded) and choose the scores for plausibility, text-image alignment, aesthetics, and overall quality (underlined).

ing for construction of feedback/rewards, and lack the ability to provide detailed actionable feedback such as implausible regions of the image, misaligned regions, or misaligned keywords on the generated images. One recent paper related to our work is Zhang et al. [61], which collected a dataset of artifact regions for image synthesis tasks, trained a segmentation-based model to predict artifact regions, and proposed a region inpainting method for those regions. However, the focus of their work is artifact region only, while in this paper, we collected rich feedback for T2I generation containing not only artifact regions, but also misalignment regions, misaligned keywords, and four fine-grained scores from multiple aspects. To the best of our knowledge, this is the first work on heterogeneous rich human feedback for text-to-image models.

3. Collecting rich human feedback

3.1. Data collection process

In this section, we discuss our procedure to collect the RichHF-18K dataset, which includes two heatmaps (artifact/implausibility and misalignment), four fine-grained scores (plausibility, alignment, aesthetics, and overall score), and one text sequence (misaligned keywords).

For each generated image, the annotators are first asked to examine the image and read the text prompt used to generate it. Then, they mark points on the image to indicate the location of any implausibility/artifact or misalignment w.r.t the text prompt. The annotators are told that each marked point has an "effective radius" (1/20 of the image height), which forms an imaginary disk centering at the marked point. In this way, we can use a relatively small amount of points to cover the image regions with flaws. Lastly, annotators label the misaligned keywords and the four types of scores for plausibility, image-text alignment, aesthetic, and overall quality, respectively, on a 5-point Likert scale. Detailed definitions of image implausibility/artifact and misalignment can be found in the supplementary materials. We designed a web UI, as shown in Fig. 1, to facilitate data collection. More details about data collection process can be found in the supplementary materials.

3.2. Human feedback consolidation

To improve the reliability of the collected human feedback on generated images, each image-text pair is annotated by three annotators. We therefore need to consolidate the multiple annotations for each sample. For the scores, we simply average the scores from the multiple annotators for an image to obtain the final score. For the misaligned keyword annotations, we perform majority voting to get the final sequence of indicators of aligned/misaligned, using the most frequent label for the keywords. For the point annotations, we first convert them to heatmaps for each annotation, where each point is converted to a disk region (as discussed in the last subsection) on the heatmap, and then we compute the average heatmap across annotators. The regions with clear implausibility are likely to be annotated by all annotators and have a high value on the final average heatmap.

3.3. RichHF-18K: a dataset of rich human feedback

We select a subset of image-text pairs from the Pick-a-Pic dataset for data annotation. Although our method is general and applicable to any generated images, we choose the majority of our dataset to be photo-realistic images, due to its importance and wider applications. Moreover, we also want to have balanced categories across the images. To ensure balance, we utilized the PaLI visual question answering (VQA) model [7] to extract some basic features from the Pick-a-Pic data samples. Specifically, we asked the following questions for each image-text pair in Pick-a-Pic. 1) Is the image photorealistic? 2) Which category best describes the image? Choose one in 'human', 'animal', 'object', 'indoor scene', 'outdoor scene'. PaLI's answers to these two questions are generally reliable under our manual inspection. We used the answers to sample a diverse subset from Pick-a-Pic, resulting in 17K image-text pairs. We randomly split the 17K samples into two subsets, a training set with 16K samples and a validation set with 1K samples. The distribution of the attributes of the 16K training samples is shown in the supplementary materials. Additionally, we collect rich human feedback on the unique prompts and their corresponding images from the Pick-a-Pic test set as our test set. In total, we collected rich human feedback on the 18K image-text pairs from Pick-a-Pic. Our RichHF-18K dataset consists of 16K training, 1K validation, and 1K test samples.

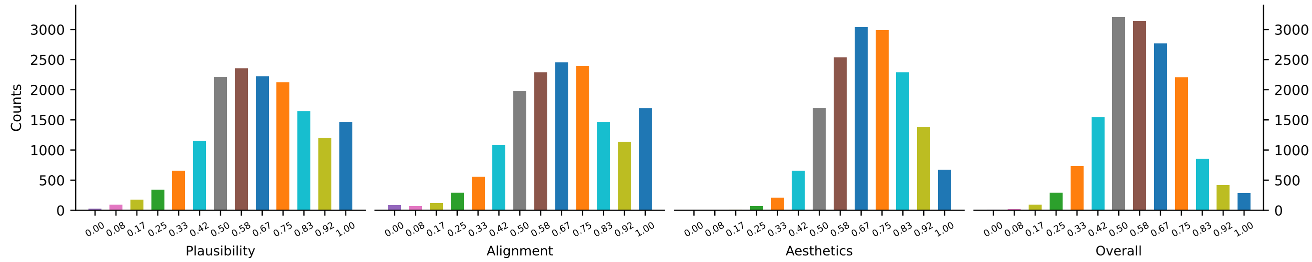


Figure 2. Histograms of the **average scores** of image-text pairs in the training set.

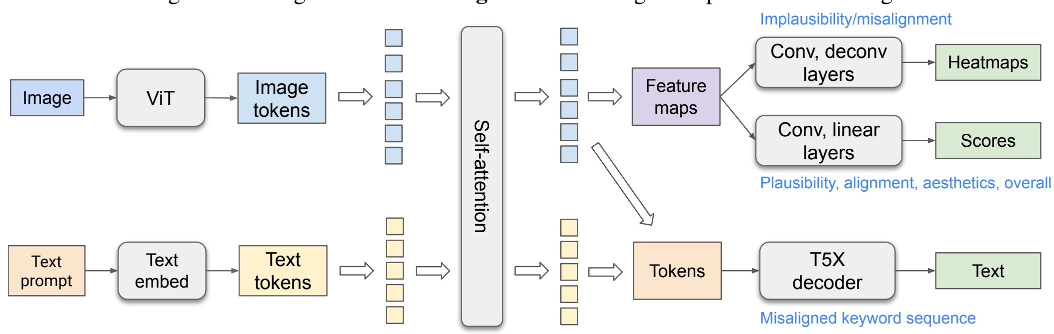


Figure 3. Architecture of our rich feedback model. Our model consists of two streams of computation: one vision and one text stream. We perform self-attention on the ViT-outputted image tokens and the Text-embed module-outputted text tokens to fuse the image and text information. The vision tokens are reshaped into feature maps and mapped to heatmaps and scores. The vision and text tokens are sent to a Transformer decoder to generate a text sequence.

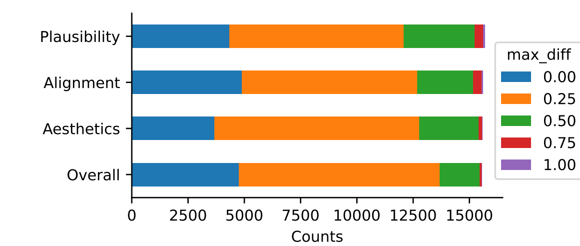


Figure 4. Counts of the samples with **maximum differences** of the scores in the training set.

3.4. Data statistics of RichHF-18K

In this section, we summarize the statistics of the scores and perform the annotator agreement analysis for the scores. We standardize the scores s with the formula $(s - s_{\min})/(s_{\max} - s_{\min})$ ($s_{\max} = 5$ and $s_{\min} = 1$) so that the scores lie in the range [0, 1].

The histogram plots of the scores are shown in Fig. 2. The distribution of the scores is similar to a Gaussian distribution, while the plausibility and text-image alignment scores have a slightly higher percentage of score 1.0. The distribution of the collected scores ensures that we have a reasonable number of negative and positive samples for training a good reward model.

To analyze the rating agreement among annotators for an image-text pair, we calculate the maximum difference among the scores: $\max_{\text{diff}} = \max(\text{scores}) - \min(\text{scores})$ where scores are the three score labels for an image-text pair. We plot the histogram of $\max_{\text{diff}}$ in Fig. 4. We can see that around 25% of the samples have perfect annotator agreement and around 85% of the samples have good annotator agreement ($\max_{\text{diff}}$ is less than or equal to 0.25 after the standardization or 1 in the 5-point Likert scale).

4. Predicting rich human feedback

4.1. Models

4.1.1 Architecture

The architecture of our model is shown in Fig. 3. We adopt a vision-language model based on ViT [14] and T5X [40] models, inspired by the Spotlight model architecture [33], but modifying both the model and pretraining datasets to better suit our tasks. We use a self-attention module [49] among the concatenated image tokens and text tokens, similar to PaLI [7], as our tasks require bidirectional information propagation. The text information is propagated to image tokens for text misalignment score and heatmap prediction, while the vision information is propagated to text tokens for better vision-aware text encoding to decode the text misalignment sequence. To pretrain the model on more diverse images, we add the natural image captioning task on the WebLI dataset [7] to the pretraining task mixture.

Specifically, the ViT takes the generated image as input and outputs image tokens as high-level representations. The text prompt tokens are embedded into dense vectors. The image tokens and embedded text tokens are concatenated and encoded by the Transformer self-attention encoder in T5X. On top of the encoded fused text and image tokens, we use three kinds of predictors to predict different outputs. For heatmap prediction, the image tokens are reshaped into a feature map and sent through convolution layers, deconvolution layers, and sigmoid activation, and outputs implausibility and misalignment heatmaps. For score prediction, the feature map is sent through convolution layers, linear layers, and sigmoid activation, resulting in scalars as fine-grained scores.

To predict the keyword misalignment sequence, the original prompt used to generate the image is used as text input to the model. A modified prompt is used as the prediction target for the T5X decoder. The modified prompt has a special suffix ('_0') for each misaligned token, *e.g.*, *a yellow_0 cat* if the generated image contains a black cat and the word *yellow* is misaligned with the image. During evaluation, we can extract the misaligned keywords using the special suffix.

4.1.2 Model variants

We explore two model variants for the prediction heads of the heatmaps and scores.

Multi-head A straightforward way to predict multiple heatmaps and scores is to use multiple prediction heads, with one head for each score and heatmap type. This will require seven prediction heads in total.

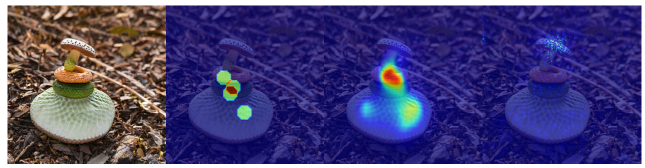
Augmented prompt Another approach is to use a single head for each prediction type, *i.e.*, three heads in total, for heatmap, score, and misalignment sequence, respectively. To inform the model of the fine-grained heatmap or score type, we augment the prompt with the output type. More specifically, we prepend a task string (*e.g.*, 'implausibility heatmap') to the prompt for each particular task of one example and use the corresponding label as the training target. During inference, by augmenting the prompt with the corresponding task string, the single heatmap (score) head can predict different heatmaps (scores). As we show in the experiments, this augmented prompt approach can create task-specific vision feature maps and text encodings, which performs significantly better in some of the tasks.

4.1.3 Model optimization

We train the model with a pixel-wise mean squared error (MSE) loss for the heatmap prediction, and MSE loss for the score prediction. For misalignment sequence prediction, the model is trained with teacher-forcing cross-entropy loss. The final loss function is the weighted combination of the heatmap MSE loss, score MSE loss, and the sequence teacher-forcing cross-entropy loss.

(a) Image (b) GT (c) Our model (d) ResNet-50

Figure 5. Examples of implausibility heatmaps. Prompt: *photo of a slim asian little girl ballerina with long hair wearing white tights running on a beach from behind nikon D5*



(a) Image (b) GT (c) Our model (d) CLIP gradient

Figure 6. Examples of misalignment heatmaps. Prompt: *A snake on a mushroom.*

4.2. Experiments

4.2.1 Experimental setup

Our model is trained on the 16K RichHF-18K training samples, and the hyperparameters were tuned using the model performance on the 1K RichHF-18K validation set. The hyperparameters setup can be found in supplementary material.

Evaluation metrics For score prediction tasks, we report the Pearson linear correlation coefficient (PLCC) and Spearman rank correlation coefficient (SRCC), which are typical evaluation metrics for score predictions [28]. For heatmap prediction tasks, a straightforward way to evaluate the results would be to borrow standard saliency heatmap evaluation metrics such as NSS/KLD [5]. However, these metrics cannot be applied directly in our case as all these metrics assume the ground truth heatmap is not empty; yet in our case, empty ground truth is possible (*e.g.*, for artifact/implausibility heatmap, it means the image does not have any artifact/implausibility). As such, we report MSE on all samples and on those with empty ground truth, respectively, and report saliency heatmap evaluation metrics like NSS/KLD/AUC-Judd/SIM/CC [5] for the samples with non-empty ground truth. For the misaligned keyword sequence prediction, we adopt the token-level precision, recall, and F1-score. Specifically, the precision/recall/F1 scores are computed for the misaligned keywords over all the samples.

Baselines For comparison, we finetune two ResNet-50 models [18], with multiple fully connected layers and deconvolution heads to predict the scores and heatmaps, respectively. We also use the off-the-shelf PickScore model [30] to compute the PickScores and calculate the metrics w.r.t each of our four ground truth scores. We use the off-the-shelf CLIP model [39] as a baseline to compute

	Plausibility		Aesthetics		Text-image Alignment		Overall	
	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑
ResNet-50	0.495	0.487	0.370	0.363	0.108	0.119	0.337	0.308
PickScore (off-the-shelf)	0.0098	0.0280	0.131	0.140	0.346	0.340	0.202	0.226
CLIP (off-the-shelf)	–	–	–	–	0.185	0.130	–	–
CLIP (fine-tuned)	0.390	0.378	0.357	0.360	0.398	0.390	0.353	0.352
Our Model (multi-head)	0.666	0.654	**0.605**	**0.591**	**0.487**	**0.500**	**0.582**	0.561
Our Model (augmented prompt)	**0.693**	**0.681**	0.600	0.589	0.474	0.496	0.580	**0.562**

Table 1. Score prediction results on the test set.

	All data	$GT = 0$	$GT > 0$				
	MSE ↓	MSE ↓	CC ↑	KLD ↓	SIM ↑	NSS ↑	AUC-Judd ↑
ResNet-50	0.00996	**0.00093**	0.506	1.669	0.338	2.924	0.909
Ours (multi-head)	0.01216	0.00141	0.425	1.971	0.302	2.330	0.877
Ours (augmented prompt)	**0.00920**	0.00095	**0.556**	**1.652**	**0.409**	**3.085**	**0.913**

Table 2. Implausibility heatmap prediction results on the test set. $GT = 0$ refers to empty implausibility heatmap, *i.e.*, no artifacts/implausibility (69 out of 995 test samples are empty), for ground truth. $GT > 0$ refers to heatmaps with artifacts/implausibility, for ground truth.

(a) Prompt: *gamer playing league of legends at night.*

Plausibility score.
GT: 0.333, Our model: 0.410
Overall score.
GT: 0.417, Our model: 0.457

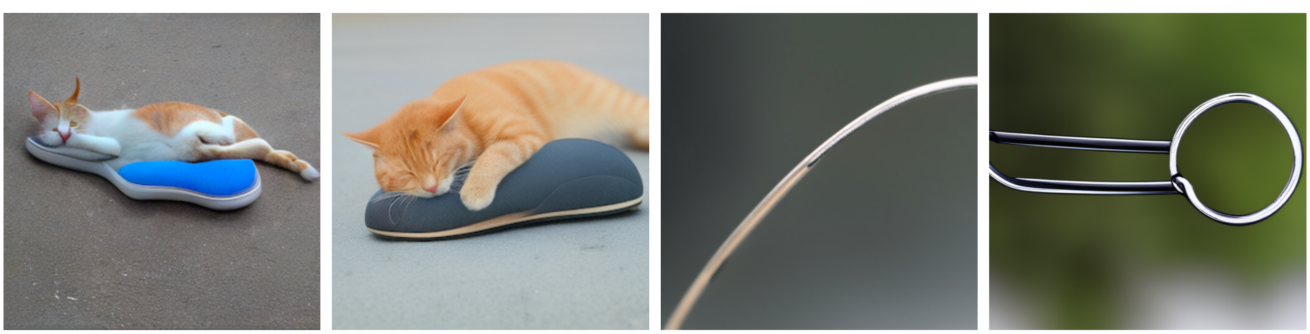
(b) Prompt: *An endless wavy ocean under a colorful night sky artistic painting pastel.*
Plausibility score.
GT: 1.0, Our model: 0.979
Overall score.
GT 1.0, Our model: 0.848

(c) Prompt: *Mechanical bee flying in nature electronics motors wires buttons lcd.*
Text-image alignment score.
GT: 0.583, Our model: 0.408
Aesthetics score.
GT: 0.75, Our model: 0.722

(d) Prompt: *anime fortnite character.*

Text-image alignment score.
GT: 1.0, Our model: 0.897
Aesthetics score.
GT: 0.75, Our model: 0.713

Figure 7. Examples of ratings. "GT" is the ground-truth score (average score from three annotators).

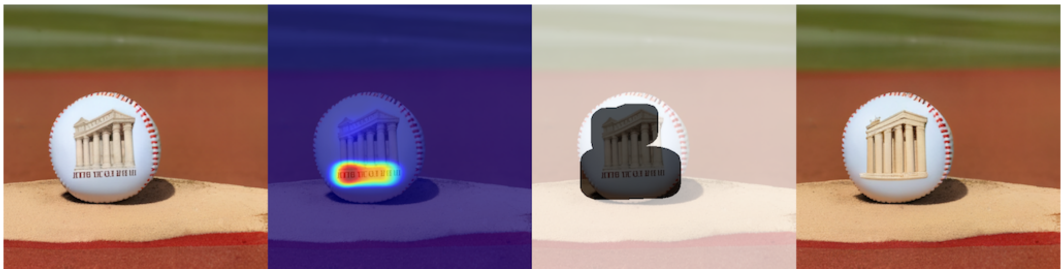


(a) Muse [6] before finetuning (b) Muse [6] after finetuning (c) LD [41] without guidance (d) LD [41] after aesthetic guidance

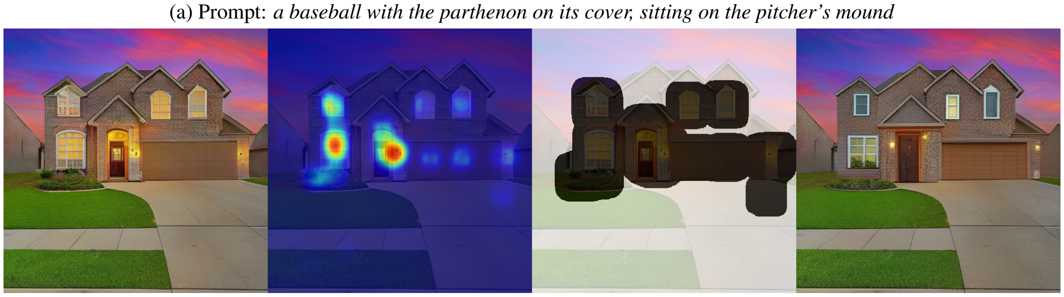
Figure 8. Examples illustrating the impact of RAHF on generative models. (a-b): Muse [6] generated images before and after finetuning with examples filtered by plausibility scores, prompt: *A cat sleeping on the ground using a shoe as a pillow*. (c-d): Results without and with aesthetic score used as Classifier Guidance [2] on Latent Diffusion (LD) [41], prompt: *a macro lens closeup of a paperclip.*

	All data	$GT = 0$	$GT > 0$				
	MSE ↓	MSE ↓	CC ↑	KLD ↓	SIM ↑	NSS ↑	AUC-Judd ↑
CLIP gradient	0.00817	0.00551	0.015	3.844	0.041	0.143	0.643
Our Model (multi-head)	**0.00303**	0.00015	0.206	**2.932**	0.093	1.335	0.838
Our Model (augmented prompt)	0.00304	**0.00006**	**0.212**	2.933	**0.106**	**1.411**	**0.841**

Table 3. Text misalignment heatmap prediction results on the test set. $GT = 0$ refers to empty misalignment heatmap, *i.e.*, no misalignment (144 out of 995 test samples are empty), for ground truth. $GT > 0$ refers to heatmaps with misalignment, for ground truth.



(a) Prompt: *a baseball with the parthenon on its cover, sitting on the pitcher's mound*



(b) Prompt: *A photograph of a beautiful, modern house that is located in a quiet neighborhood. The house is made of brick and has a large front porch. It has a manicured lawn and a large backyard.*

Figure 9. Region inpainting with Muse [6] generative model. From left to right, the 4 figures are: original images with artifacts from Muse, predicted implausibility heatmaps from our model, masks by processing (thresholding, dilating) the heatmaps, and new images from Muse region inpainting with the mask, respectively.

	Precision	Recall	F1 Score
Multi-head	**62.9**	33.0	43.3
Augmented prompt	61.3	**34.1**	**43.9**

Table 4. Text misalignment prediction results on the test set.

Preference	$\gg$	$>$	$\approx$	$<$	$\ll$
Percentage	21.5%	30.33%	31.33%	12.67%	4.17%

Table 5. **Human Evaluation Results: Finetuned Muse vs original Muse model preference**: Percentage of examples where finetuned Muse is significantly better ($\gg$), slightly better ($>$), about the same ($\approx$), slightly worse ($<$), significantly worse ($\ll$) than original Muse. Data was collected from 6 individuals in a randomized survey.

the cosine similarity of the image and text embeddings and use it to calculate the text-image alignment metric, as the CLIP cosine similarity is designed to reflect the alignment between images and prompts. Besides, we also fine-tune a CLIP model to predict the four types of scores using our training dataset. For misalignment heatmap prediction, we use CLIP gradient [45] map as a baseline.

4.2.2 Prediction result on RichHF-18K test set

Quantitative analysis The experimental results of our model prediction on the four fine-grained scores, the implausibility heatmap, misalignment heatmap, and misalignment keyword sequence on our RichHF-18K test set are presented in Tab. 1, Tab. 2, Tab. 3, and Tab. 4 respectively.

In both Tab. 1 and Tab. 3, the two variants of our proposed model both significantly outperform ResNet-50 (or CLIP for text-image alignment score). Yet, in Tab. 2, the multi-head version of our model performs worse than ResNet-50, but our augmented prompt version outperforms ResNet-50. The main reason might be that in the multi-head version, without augmenting the prediction task in the prompt, the same prompt is used for all the seven prediction tasks, and hence the feature maps and text tokens will be the same for all tasks. It might not be easy to find a good tradeoff among these tasks, and hence the performance of some tasks such as artifact/implausibility heatmap became worse. However, after augmenting the prediction task into a prompt, the feature map and text token can be adapted to each particular task with better results. Additionally, we note that misalignment heatmap prediction generally has worse results than artifact/implausibility heatmap prediction, possibly because misalignment regions are less well-defined, and the annotations may therefore be noisier.

Qualitative examples We show some example predictions from our model for implausibility heatmap (Fig. 5), where our model identifies the regions with artifact/implausibility, and for misalignment heatmap (Fig. 6), where our model identifies the objects that don't correspond to the prompt. Fig. 7 shows some example images and their ground-truth and predicted scores. More examples are in the supplementary material.

5. Learning from rich human feedback

In this section, we investigate whether the predicted rich human feedback (*e.g*., scores and heatmaps) can be used to improve image generation. To ensure that the gains from our RAHF model generalize across generative model families, we mainly use Muse [6] as our target model to improve, which is based on a masked transformer architecture and thus different from the Stable Diffusion model variants in our RichHF-18K dataset.

Finetuning generative models with predicted scores We first illustrate that finetuning with RAHF scores can improve Muse. First, we generate eight images for each of the 12,564 prompts (the prompt set is created via PaLM 2 [1, 11] with some seed prompts) using the pre-trained Muse model. We predict RAHF scores for each image, and if the highest score for the images from each prompt is above a fixed threshold, it will be selected as part of our finetuning dataset. The Muse model is then finetuned with this dataset. This approach could be viewed as a simplified version of Direct Preference Optimization [15].

In Fig. 8 (a)-(b), we show one example of finetuning Muse with our predicted plausibility score (threshold=0.8). To quantify the gain from Muse finetuning, we used 100 new prompts to generate images, and asked 6 annotators to perform side-by-side comparisons (for plausibility) between two images from the original Muse and the fine-tuned Muse respectively. The annotators choose from five possible responses (image A is significantly/slightly better than image B, about the same, image B is slightly/significantly better than image A), without knowledge of which model is used to generate the image A/B. The results in Tab. 5 demonstrate that the finetuned Muse with RAHF plausibility scores possesses significantly fewer artifacts/implausibility than the original Muse.

Moreover, in Fig. 8 (c)-(d), we show an example of using the RAHF aesthetic score as Classifier Guidance to the Latent Diffusion model [41], similar to the approach in Bansal et al. [2], demonstrating that each of the fine-grained scores can improve different aspects of the generative model/results.

Region inpainting with predicted heatmaps and scores We demonstrate that our model's predicted heatmaps and scores can be used to perform region inpainting to improve the quality of generated images. For each image, we first predict implausibility heatmaps, then create a mask by processing the heatmap (using thresholding and dilating). Muse inpainting [6] is applied within the masked region to generate new images that match the text prompt. Multiple images are generated, and the final image is chosen by the highest predicted plausibility score by our RAHF.

In Fig. 9, we show several inpainting results with our predicted implausibility heatmaps and plausibility scores. As shown, more plausible images with fewer artifacts are generated after inpainting. Again, this shows that our RAHF generalizes well to images from a generative model very different from the ones whose images are used to train RAHF. More details and examples can be found in the supplementary material.

6. Conclusions and limitations

In this work, we contributed RichHF-18K, the first rich human feedback dataset for image generation. We designed and trained a multimodal Transformer to predict the rich human feedback, and demonstrated some instances to improve image generation with our rich human feedback.

While some of our results are quite exciting and promising, there are several limitations to our work. First, the model performance on the misalignment heatmap is worse than that on the implausibility heatmaps, possibly due to the noise in the misalignment heatmap. It is somewhat ambiguous how to label some misalignment cases such as absent objects on the image. Improving the misalignment label quality is one of the future directions. Second, it would be helpful to collect more data on generative models beyond Pick-a-Pic (Stable Diffusion) and investigate their effect on the RAHF models. Moreover, while we present three promising ways to leverage our model to improve T2I generation, there is a myriad of other ways to utilize rich human feedback that can be explored, *e.g*., how to use the predicted heatmaps or scores as a reward signal to finetune generative models with reinforcement learning, and how to use the predicted heatmaps as a weighting map, or how to use the predicted misaligned sequences in learning from human feedback to help improve image generation, etc. We hope RichHF-18K and our initial models inspire quests to investigate these research directions in future work.

References

[1] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri,

Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023. 8

[2] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models, 2023. 6, 8

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[4] 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. 1

[5] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 5

[6] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2, 6, 7, 8

[7] 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, 2022. 3, 4

[8] Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian Scene Graph: Improving Reliability in Fine-Grained Evaluation for Text-to-Image Generation. In *arXiv:2310.18235*, 2023. 2

[9] Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for text-to-image generation and evaluation. In *NeurIPS*, 2023. 2

[10] Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for text-to-image generation and evaluation. In *NeurIPS*, 2023. 2

[11] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 8

[12] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[13] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[15] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, 2023. 8

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[17] Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Josh Susskind, and Navdeep Jaitly. Matryoshka diffusion models, 2023. 1

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[19] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2022. 2

[20] 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

[21] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2016. 2

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[23] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1

[24] 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 *ICCV*, 2023. 2

[25] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *arXiv preprint arXiv:2307.06350*, 2023. 2

[26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[27] 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 *CVPR*, 2023. 1

[28] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 5

[29] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[30] 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. *arXiv preprint arXiv:2305.01569*, 2023. 1, 2, 5

[31] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Li Fei-Fei, Jiajun Wu, Stefano Ermon, and Percy Liang. Holistic evaluation of text-to-image models. In *Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 2

[32] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[33] Gang Li and Yang Li. Spotlight: Mobile ui understanding using vision-language models with a focus. In *International Conference on Learning Representations*, 2023. 4

[34] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2

[35] Haomiao Ni, Changhao Shi, Kai Li, Sharon X. Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18444–18455, 2023. 1

[36] 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. 2

[37] Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14277–14286, 2023. 2

[38] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1505–1514, 2019. 2

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 5

[40] Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, Curtis Hawthorne, Aitor Lewkowycz, Alex Salcianu, Marc van Zee, Jacob Austin, Sebastian Goodman, Livio Baldini Soares, Haitang Hu, Sasha Tsvyashchenko, Aakanksha Chowdhery, Jasmijn Bastings, Jannis Bulian, Xavier Garcia, Jianmo Ni, Andrew Chen, Kathleen Kenealy, Jonathan H. Clark, Stephan Lee, Dan Garrette, James Lee-Thorp, Colin Raffel, Noam Shazeer, Marvin Ritter, Maarten Bosma, Alexandre Passos, Jeremy Maitin-Shepard, Noah Fiedel, Mark Omernick, Brennan Saeta, Ryan Sepassi, Alexander Spiridonov, Joshua Newlan, and Andrea Gesmundo. Scaling up models and data with `t5x` and `seqio`. *arXiv preprint arXiv:2203.17189*, 2022. 4

[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. 1, 2, 6, 8

[42] 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. 1

[43] 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

[44] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks, 2023. 1

[45] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014. 7

[46] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2

[47] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16515–16525, 2022. 2

[48] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[50] S. Wang, C. Saharia, C. Montgomery, J. Pont-Tuset, S. Noy, S. Pellegrini, Y. Onoe, S. Laszlo, D. J. Fleet, R. Soricut, J.

Baldridge, M. Norouzi, P. Anderson, and W. Chan. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *CVPR*, 2023. 1

[51] 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

[52] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023. 1, 2

[53] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *arXiv preprint arXiv:2310.10647*, 2023. 1

[54] 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 *Neural Information Processing Systems*, 2023. 1, 2

[55] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 2

[56] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 2022. 1

[57] 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, 2023. 2

[58] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 1

[59] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023. 1

[60] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 2

[61] 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. 3

[62] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5802–5810, 2019. 2