Rich Human Feedback for Text-to-Image Generation

Supplementary Material

7. Ethical conduct

Our data collection has been approved by an Institutional Review Board.

8. Data collection details

8.1. Image artifacts/implausibility definitions

1. Distorted human/animal bodies/faces
   (a) Distorted/combined faces and/or body parts (unless specified in the text caption)
   (b) Missing body parts (unless specified in the text caption)
   (c) Additional body parts (unless specified in the text caption)
2. Distorted objects (non human/animal)
   (a) Distorted objects (e.g., furniture, vehicles, buildings) (unless specified in the text caption)
3. Distorted/Nonsensical text
   (a) Text that is distorted, nonsensical, or misspelled (unless specified in the text caption)
4. Nonsensical Representations
   (a) Representations that are unrealistic/nonsensical (unless specified in the text caption), or difficult to understand
5. Excessive blurriness/lack of sharpness
   (a) The image contains excessive blurriness or quality that detracts from the overall image (focus on one part of the image is OK)
   (b) The image contains a lack of definition/sharpness that detracts from the overall image
6. Any other artifacts or implausibility not covered above

8.2. Text-image misalignment definitions and what-to-do

Since we require the annotators to mark the misaligned words in the text prompt, we differentiate this part from Sec. 8.1 by including a what-to-do under each definition.

1. **Something is missing**: a human/animal/object specified in the text caption is missing in the image
   - Click on that word of the human/animal/object in the text
2. **Incorrect attributes**: an attribute (e.g., color) of an object specified in the text is incorrect in the image
   - Click on that word of the attribute in the text and click on the region of the object on the image
3. **Incorrect actions**: an action specified in the text caption is not represented in the image
   - Click on that word of the action in the text and click on the region of the wrong actions on the image
4. **Incorrect numbers**: counts of humans/animals/objects in the image do not match those specified in the text
   - Click on the number in the text
5. **Incorrect position**: the spatial position of two entities in the image does not match that specified in the text
   - Click on the word of the position in the text
6. **Other**: any other inconsistency between text and image
   - Click on the word of the inconsistency in the text

8.3. Additional details

Annotation guideline To ensure the annotators understand the above definitions, we provide 4-10 examples for each definition of the annotation terms in the guideline. All of our annotators can read English and thus understand the text prompts. In some of the prompts, there are concepts or person names in the text prompts that are uncommon and may cause confusion to the annotators. Therefore, we instruct the annotators to do a quick search on the internet regarding any unfamiliar concepts in the text prompts and skip samples with confusing prompts full of strange concepts.

Annotation interface We designed a web UI to facilitate data collection with the following principles: 1) convenience for annotators to perform annotations, ideally within a short time for an image-text pair and, 2) allowing annotators to perform all annotations on the same UI, so that the fine-grained scores are based on the annotated regions and keywords. To this end, we created the interface as illustrated in Fig. 1. The main UI consists of an image displayed on the left and a panel on the right, where the text prompt is shown at the top of the panel. Annotators are asked to first click on the image to annotate the artifact/implausible regions and misalignment regions, and then select the misaligned keywords and the fine-grained scores on the right of the panel.

More details We created detailed annotation guidelines to instruct the annotators regarding the annotation steps, interactions with the web UI, examples of different types of implausibility, artifacts, and misalignment. All the annotators (27 in total) are trained with the annotation guidelines and calibrated, before they perform the annotation in order to reduce annotation discrepancy and improve quality. Our annotation took around 3,000 rater-hours in total. To improve the effectiveness of the collected dataset and control

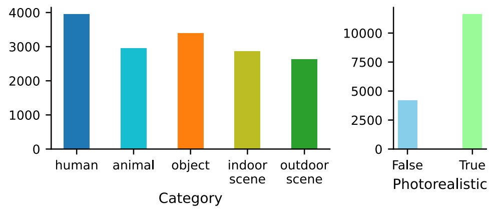


Figure 10. Histograms of the PaLI attributes of the images in the training set.

the time spent on annotation, we filter out any image-text pairs that have a text prompt with less than 3 words or more than 20 words. We also filter out non-English prompts or any prompts containing emoji.

Dataset size Since the Pick-a-Pic v1 dataset contains some images and/or prompts that are inappropriate (*e.g*., containing nudity), we ask the annotators to mark these images with a special flag and skip the annotation. We filter out these inappropriate images and/or prompts during data post-processing. For this reason, the total number of images in our final training set is around 300 short of 16K.

Additional details of data collection The distribution of the attributes of the 16K training samples is shown in Fig. 10. We can see a relatively balanced distribution of the types of content in the generated images in our dataset.

8.4. Discussions and limitations

We choose **points over bounding boxes** in our region annotation because we find that points are much faster to mark and can have a reasonable coverage over image regions with various shapes when we specify an effective radius for each point as discussed in the main paper.

As a limitation in our region/heatmap annotations, we notice there is an **over-annotation** issue in the artifacts/implausibility region annotation. Specifically, our annotators tend to annotate more human faces and hands on the images than necessary. One reason is that human faces and hands in the Pick-a-Pic dataset indeed have more artifacts/implausibility than other parts. Moreover, the annotators, as humans, may naturally pay more attention to human faces and hands, resulting in over-annotation of these parts. Nevertheless, the over-annotation issue is minor in our final dataset, as we strive to provide feedback to the annotators frequently to make them less nitpicking about human faces and hands.

Another limitation is the **diversity** of the subjects in the prompts/images. The Pick-a-Pic dataset (and many others) is predominantly full of human, dog, and cat subjects. For this reason, it is challenging to find a very diverse dataset for annotation. We strive to make the subjects more diverse by using balanced categories as indicated by the PaLI attributes (as in Fig. 10). We didn't choose more fine-grained categories for PaLI to test as there would be an endless list of subjects we could consider. Therefore, we leave the goal of annotating more diverse images/prompts in future works.

9. Experimental details

Hyperparameters The main model components consist of a ViT B16 encoder for image encoding, a T5 base encoder for mixing image and text tokens, and three predictors for score, heatmap, and text misalignment, respectively. The ViT B16 encoder uses a 16x16 patch size, 12 layers with 12 heads with a hidden dimension of 768, wherein the MLP has a hidden dimension of 3072. The T5 base encoder uses 12 layers with 12 heads with a hidden dimension of 768, wherein the MLP has a hidden dimension of 2048. The score predictor consists of four convolutional layers with layer norm and ReLU activation, and the filter size, kernel size, and strides are $[768, 384, 128, 64]$, $[2, 2, 2, 2]$, $[1, 1, 1, 1]$, respectively. Three dense layers of output sizes 2048, 1024, and 1, respectively, are used to generate a scalar with ReLU activation for the first two layers and sigmoid for the last. The heatmap predictor consists of two convolution layers with filter size, kernel size, and stride as $[768, 384]$, $[3, 3]$, $[1, 1]$, respectively. It then uses four de-convolution layers to up-sample to the required output size, with the filter size, kernel size, and stride as $[768, 384, 384, 192]$, $[3, 3, 3, 3]$, $[2, 2, 2, 2]$, respectively. Each de-convolution layer is with two read-out convolution layers of kernel size 3 and stride 1. Layer norm and ReLU are used for each layer. In the end, two read-out convolution layers and a final sigmoid activation are used to generate the heatmap prediction. The text predictor is implemented using a T5 base decoder with 12 layers of 12 heads, MLP dimension 2048, and hidden dimension 768. The output token length is 64.

We train the model on the datasets with a batch size of 256 for 20K iterations. We utilize the AdamW optimizer with a base learning rate of 0.015. We linearly increase the learning rate from 0 to the base learning rate in the first 2000 iterations, and then decrease the learning rate with a reciprocal square root scheduler w.r.t the number of iterations. We trained the model using 64 Google Cloud TPU v3 chips.

Image augmentations For each image, we randomly crop it by sampling a bounding box with 80%-100% width and 80%-100% height. The cropping is applied by 50% chance and otherwise the original image is used. Note that we also crop the corresponding part of the implausibility heatmap and misalignment heatmap to match the cropped image. We then create an augmented version of the image by applying several random augmentations including random brightness (max delta 0.05), random contrast (random contrast factor between 0.8 and 1), random hue (max delta 0.025), random saturation (random saturation factor between 0.8 and 1) and random jpeg noise (jpeg quality between 70 and 100). By 10% chance the augmented version is used instead of the original image. We convert the image to grayscale by 10% probability as the final image.

Finetuning generative models with predicted scores To generate the training prompt set, we provide five handcrafted seed prompts as examples and then ask PaLM 2 [1] to generate similar textual prompts. We include additional instructions that specify the prompt length and the object category. We then explain why we do not use existing benchmark datasets for training. Theoretically, we can get an infinite number of prompts using the prompt synthesis technique we proposed above. Existing datasets are 1) relatively small (e.g., TIFA [24] has 4k prompts, Davidsonian Scene Graph (DSG) [10] has only 1k prompts), or 2) containing prompts that are simple and not diverse enough, for example, only measuring single objects in Parti benchmark [59]. This motivates us to synthesize a larger set of diverse prompts for training purposes. For the 100 prompts for our human evaluation, they are sampled from the existing benchmark: TIFA [24]. We only did our evaluation on 100 prompts due to the high cost of the human annotation.

10. Additional qualitative examples

Fig. 11 provides more examples of artifacts/implausibility heatmaps. We can see that our RAHF model can more accurately locate the positions of artifacts/implausibility on various subjects such as human hands, animals, vehicles, and concept arts.

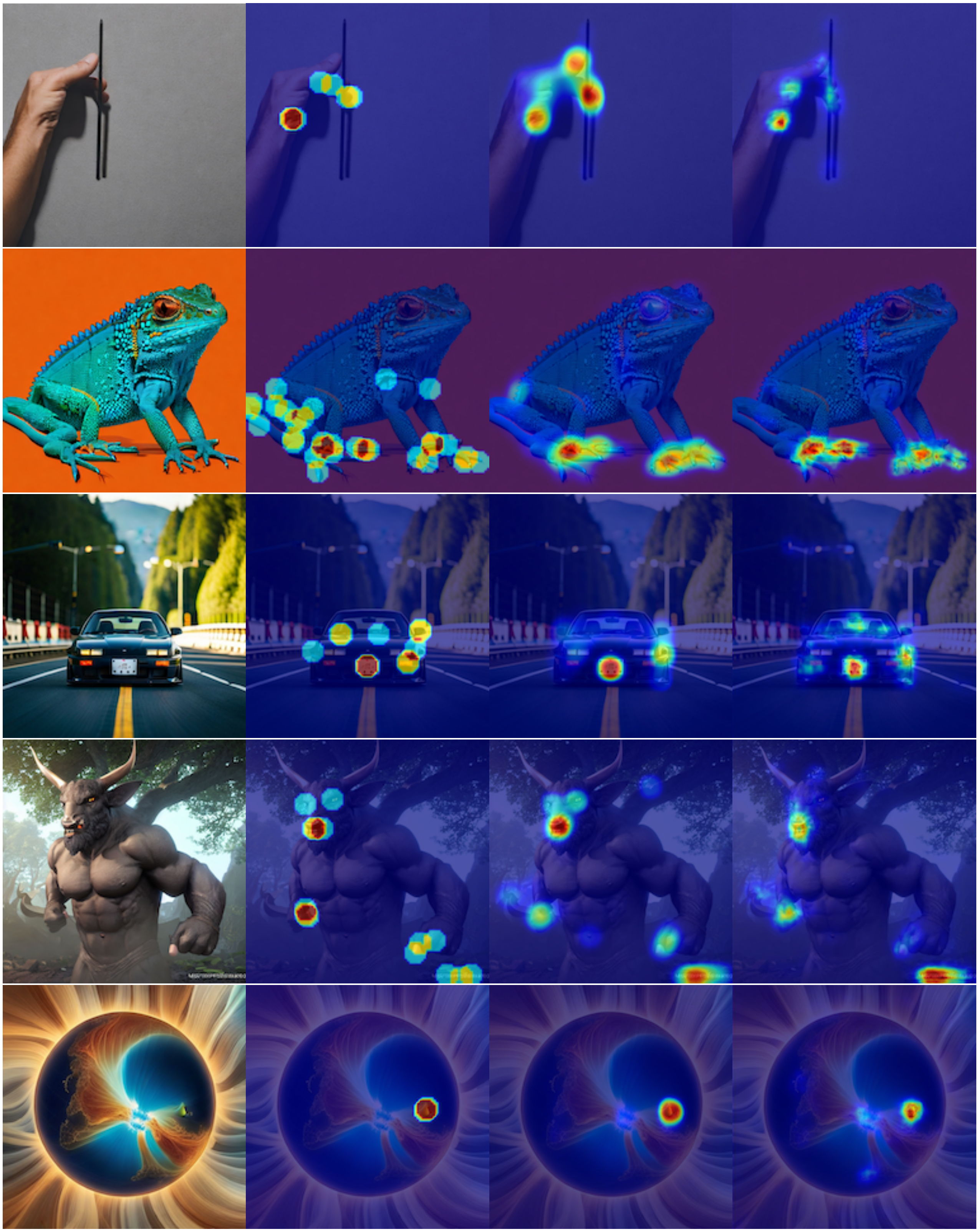
Fig. 12 provides more examples of misalignment heatmaps. We can see that our RAHF model can more accurately locate the positions of misalignment on various subjects such as animals, objects, and different outdoor scenes. For example, our model can identify the subtle difference between the real handlebar of a Segway and the one depicted in the image.

Fig. 13 provides more examples of score predictions, where our RAHF model predicts scores that are quite close to the ground truth score from human evaluation.

Fig. 14 provides examples for the misaligned keywords prediction, which shows that our RAHF model can predict the majority of the misaligned keywords marked by human annotators.

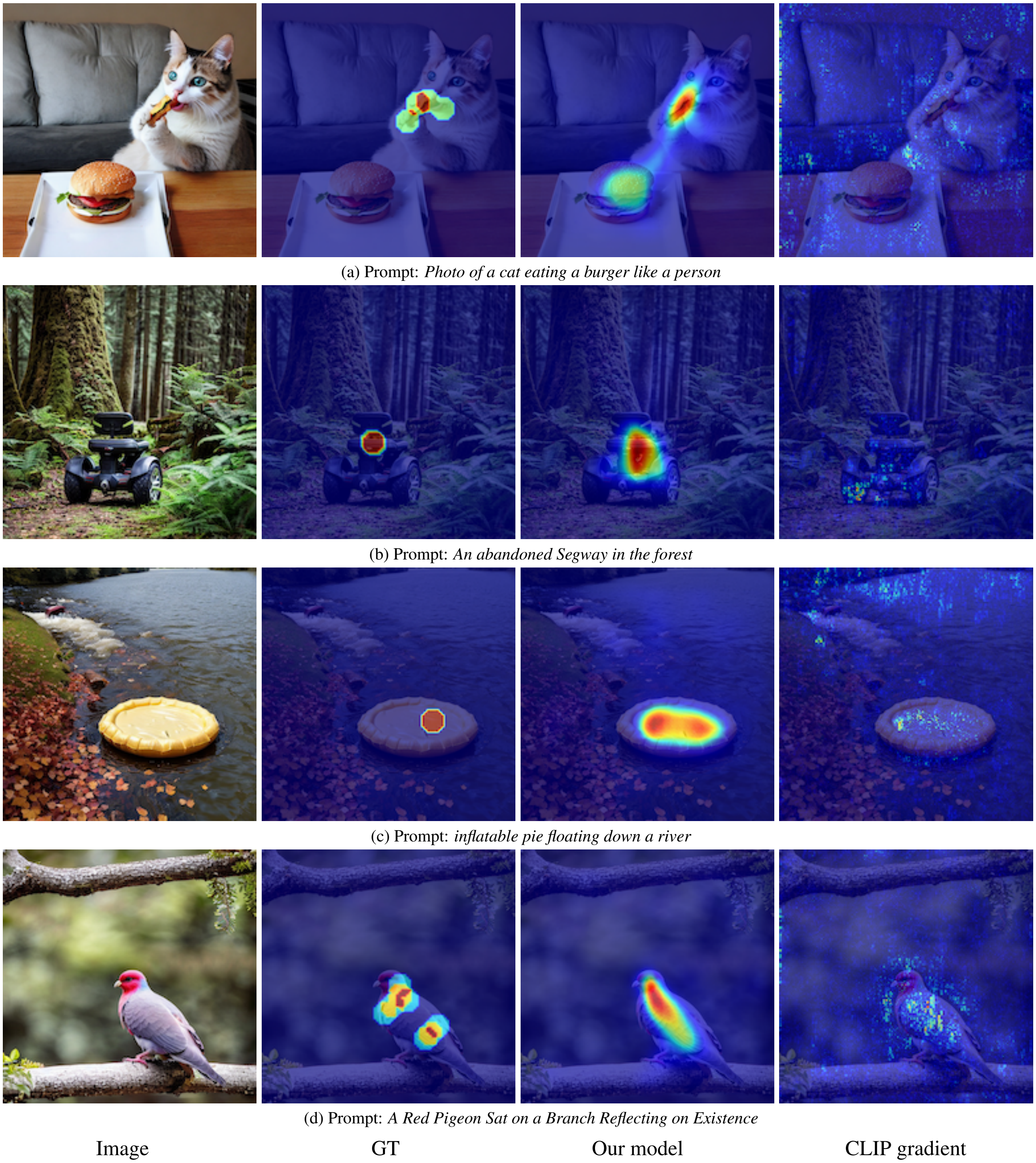
Fig. 15 provides more examples of the comparison before and after finetuning Muse with examples selected based on the predicted scores by our RAHF model and examples of using RAHF model predicted overall score as Classifier Guidance. We can see enhanced image quality of the generation from the finetuned Muse model and the Latent Diffusion model, which highlights the potential of improving T2I generation with our reward model.

Fig. 16 provides more examples of Muse inpainting with the predicted masks (converted from heatmaps) by our RAHF model, where the inpainted regions are significantly improved in plausibility.



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

Figure 11. More examples of implausibility heatmaps



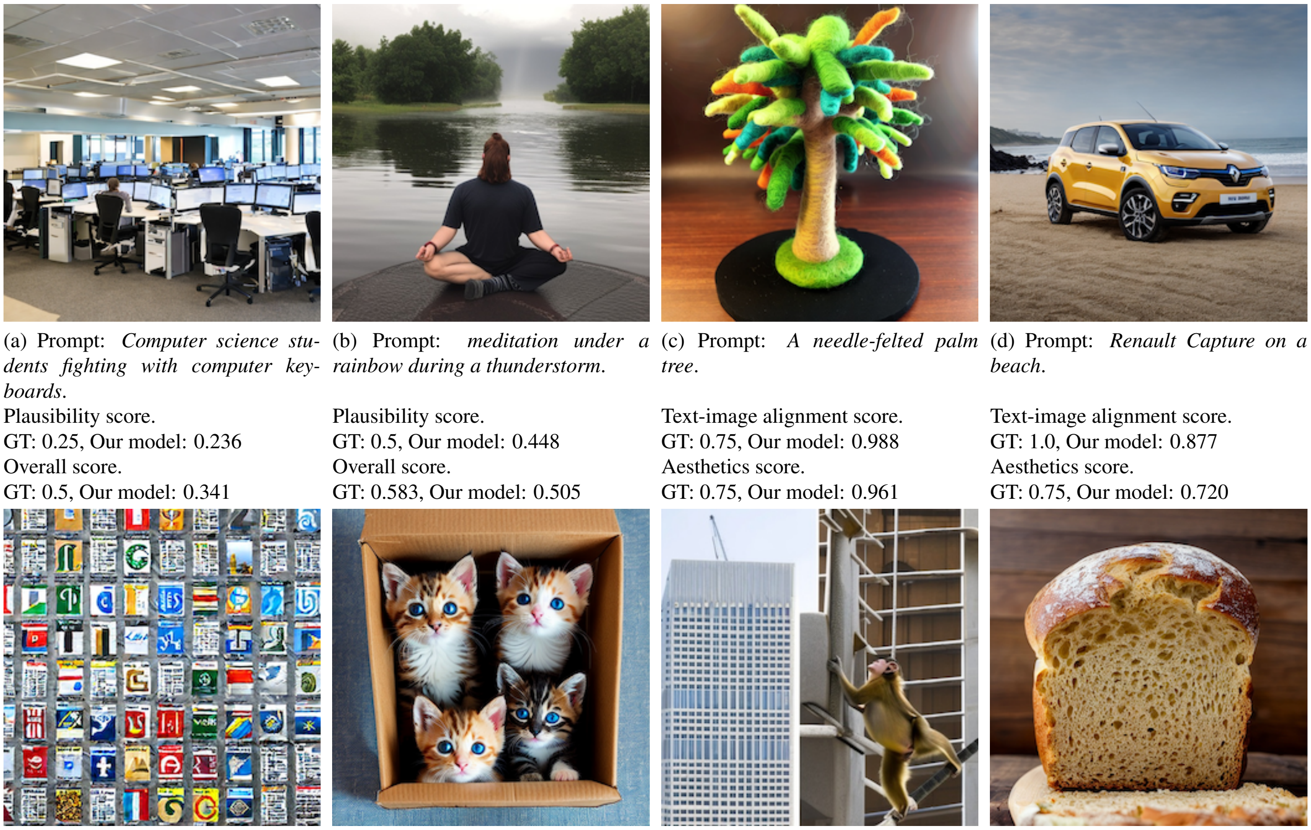
(a) Prompt: *Photo of a cat eating a burger like a person*

(b) Prompt: *An abandoned Segway in the forest*

(c) Prompt: *inflatable pie floating down a river*

(d) Prompt: *A Red Pigeon Sat on a Branch Reflecting on Existence*

Figure 12. More examples of misalignment heatmaps.



(a) Prompt: *Computer science students fighting with computer keyboards*.
Plausibility score.
GT: 0.25, Our model: 0.236
Overall score.
GT: 0.5, Our model: 0.341

(b) Prompt: *meditation under a rainbow during a thunderstorm*.
Plausibility score.
GT: 0.5, Our model: 0.448
Overall score.
GT: 0.583, Our model: 0.505

(c) Prompt: *A needle-felted palm tree*.
Text-image alignment score.
GT: 0.75, Our model: 0.988
Aesthetics score.
GT: 0.75, Our model: 0.961

(d) Prompt: *Renault Capture on a beach*.
Text-image alignment score.
GT: 1.0, Our model: 0.877
Aesthetics score.
GT: 0.75, Our model: 0.720

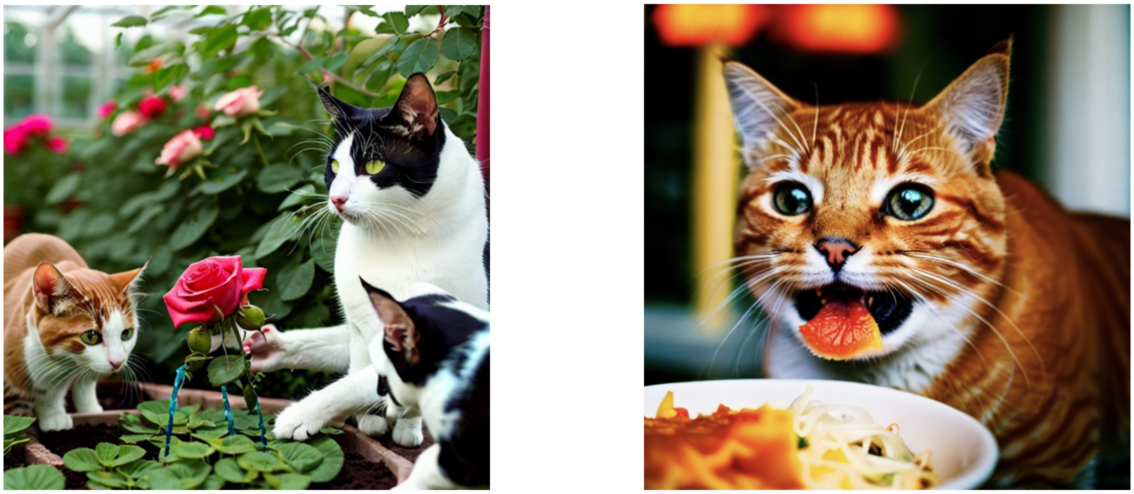
(e) Prompt: *all the letters of the greek alphabet*.
Plausibility score.
GT: 0.167, Our model: 0.331
Overall score.
GT: 0.250, Our model: 0.447

(f) Prompt: *a kittens in box*.
Plausibility score.
GT: 0.75, Our model: 0.851
Overall score.
GT: 0.75, Our model: 0.855

(g) Prompt: *monkey climbing a skyscraper*.
Text-image alignment score.
GT: 0.833, Our model: 0.536
Aesthetics score.
GT: 0.583, Our model: 0.467

(h) Prompt: *bread*.
Text-image alignment score.
GT: 1.0, Our model: 0.975
Aesthetics score.
GT: 1.0, Our model: 0.984

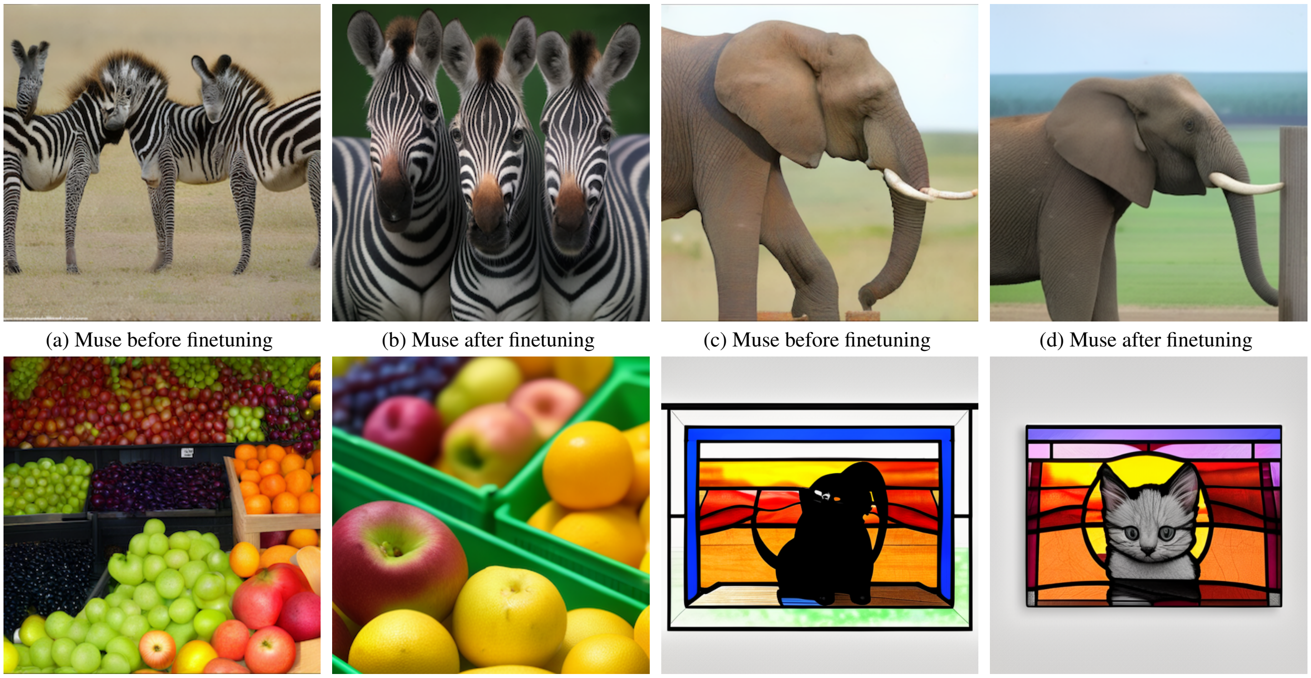
Figure 13. Examples of ratings. "GT" is the ground-truth score (average score from three annotators).



(a) The prompt is: *Two cats watering roses in a greenhouse*. The ground truth labels *two, watering, greenhouse* as misaligned keywords and our model predicts *two, greenhouse* as misaligned keywords.

(b) The prompt is: *A close up photograph of a fat orange cat with lasagna in its mouth, shot on Leica M6*. The ground truth labels *fat, lasagna, Leica, M6* as misaligned keywords and our model predicts *lasagna, Leica, M6* as misaligned keywords.

Figure 14. Examples for text misalignment prediction.



(a) Muse before finetuning

(b) Muse after finetuning

(c) Muse before finetuning

(d) Muse after finetuning

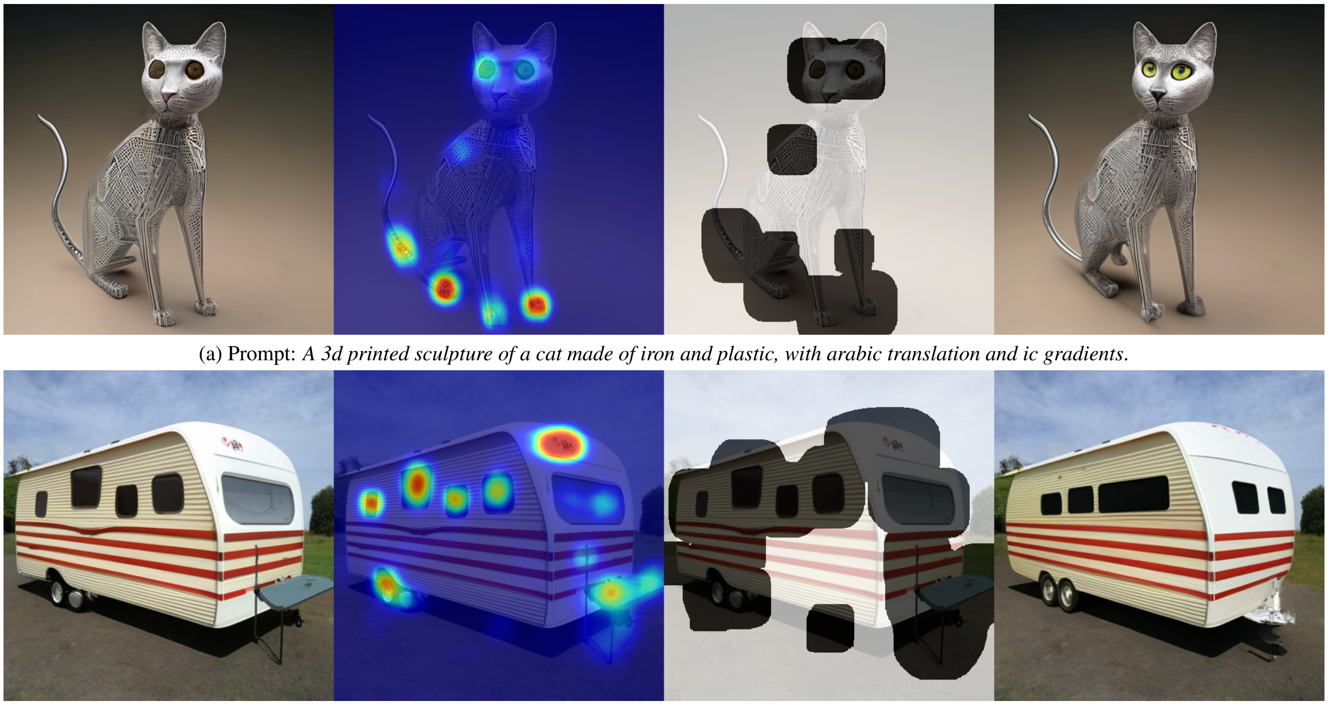
(e) Muse before finetuning

(f) Muse after finetuning

(g) LD without guidance

(h) LD with overall guidance

Figure 15. More examples illustrating the impact of RAHF on generative models. (a-f): Muse [6] generated images before and after finetuning with examples filtered by plausibility scores. Prompt: (a-b): *Three zebras are standing together in a line.* (c-d): *An elephant scratching it's neck on a post.* (e-f): *Apples, lemons, grapes, oranges and other fruits in crates.* (g-h): Results without and with overall score used as Classifier Guidance [2] on Latent Diffusion (LD) [42], prompt: *Kitten sushi stained glass window sunset fog.*



(a) Prompt: *A 3d printed sculpture of a cat made of iron and plastic, with arabic translation and ic gradients*.

(b) Prompt: *A 1960s slide out camper with a blonde, white and red color scheme*

Figure 16. 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.