

1 **A User interface for human study on categorizing text-to-image generation** 2 **errors**

3 Our user interface is shown in Figure 1. Annotators were asked to tick boxes of errors that they found
4 in the given synthesized images.

5 The error categories include:

- 6 • **People:** age, gender, type of clothing, color of clothing, weird face, weird body
- 7 • **Main object:** wrong, similar, inexistent, extra, weird
- 8 • **Other objects:** wrong, similar, inexistent, extra, weird
- 9 • **General:** stance, activity, position, number, inconsistent references, scene/event/location,
10 text, color, generally unrelated

Figure 1: Annotation interface for categorizing SD errors.

11 **B Examples of high loss training samples**

12 In Figure 2, we visualize the high loss training samples in the COCO dataset after the first epoch of
13 finetuning. These samples are target of our curation techniques. Compared to the average caption
14 length of 11 words, the top samples all have very long captions of around 30 words, making it difficult
15 for the model to learn. In the following finetuning epochs, we curate on these samples by either
16 removing the text-image pairs completely (REMOVE), replacing the caption (REPLACECAP) , or
17 replacing the image with a synthesized unseen image (REPLACEIMG).






Image	Caption	Length	Loss
	a picture of a clearly disrespectful person littered, abused alcohol, didn't flush their bad choices, and worst of all, let old glory touch a bathroom floor	26	213.24
	a picture of a rain-wet street view with lots of bike riders, rimmed with buildings that seem to bunch up and fight for space might look gray and unprepossessing, but doesn't, in part	33	200.14
	a picture of the scene shows outdoors, furthest to closest, shrubbery than a playing field with at least two uniformed and young players, and closest, a blue fence, and a long bench with	33	200.02
	a picture of it is outdoors, the exterior of a low roofed domicile, where a tiny grove of slender tropical trees makes a lean-to for super-modern blue and white motorcycle	30	199.90
	a picture of while a purple/blue sky with what looks like a kite or a loose para-sail floating in it covers most of a distance shot, the bottommost part shows grassy side banks	33	197.36

Figure 2: High loss training samples in COCO after the first epoch, ranked by loss in descending order. The top samples all have very long captions around 30 words, compared to the mean of 11 words.