A. Appendix

A.1. Experiment Details and Result Analysis

In Table 1 and Table 2, we compare our model with other models/methods. The comparison between GLIGEN and GLIGEN with our methods shows that our methods help improve the performance of layout-based generation models. The performance of DALL.E and SDXL is lower because these models are not conditioned on layouts. Their low performance shows that, the faithfulness of text-to-image generation is still a problem and the layout could improve the faithfulness. [35], [30] and [29] are three methods that improve the faithfulness of layout-based image generation models. [35] and [29] first compute the gradients with respect to a pre-defined loss function during the generation process, and then move the latents towards the gradients. [30] improves the transformer encoder to better encode the layout information. [27] focuses on image editing. Nevertheless, its methods can be applied to improve the faithfulness of generation - through detecting the generation errors and manipulating the initial noise to correct those errors. In our experiments, [35], [29] and [27] are applied to GLIGEN, same as our method. [30] is also a modified GLIGEN.

On VPEVAL, our method acquires the most gain on Scale and Spatial. The main reason is that, both object errors and relationship errors are counted in scale/spatial metrics. For example, for the prompt "a sheep to the right of a baseball glove", missing the baseball glove will also be counted as an error under spatial metric in the VPEVAL evaluation pipeline. Our methods reduce both object errors and relationship errors. Hence, the improvement is larger. For VPEVAL object, the accuracy of all models is high because the prompts are easy, which only ask the model to generate one object. On HRS, the improvement on count is high while that on spatial and scale is lower. We think it is because of the difficulty level of its prompts.

A.2. Inception Scores

We computed the Inception Scores for the original GLIGEN and the GLIGEN with our strategies, using the generated images on VPEval. The results are in Table 6. After applying our methods to GLIGEN, the score only changes slightly, indicating that the quality trade-off is minimal. We did not compute FID because we do not have reference real images for VPEval and HRS.

	IS
GLIGEN	26.21±3.31
GLIGEN with our methods	26.35± 4.03

Table 6. IS comparison

A.3. Additional Images

Figure 8 and 9 show generations with standard GLIGEN model, GLIGEN combined with our intervention-based inference method but without retrieval-based feedback, and GLIGEN with our full approach. Comparing the images in second and third column, one can find that, sometimes the model fails to generate the target object only with the text features. The use of retrieved image features helps the model to generate the missing objects.

A.4. Error Correction through Image Editing

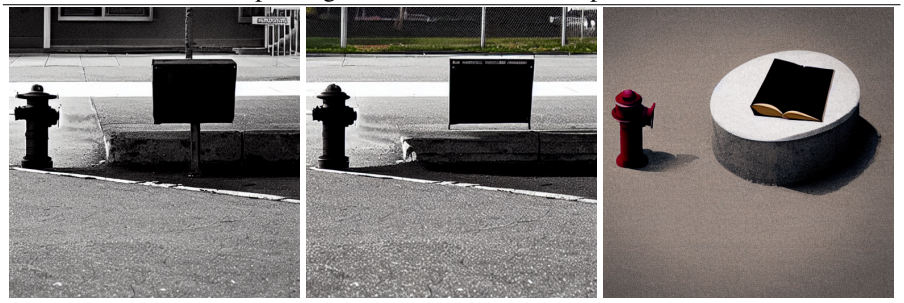
We also run image-editing model InstructPix2Pix repeatedly to correct the generation errors. The image guidance scale is set as 1.5. The model runs for 4 times and the average running time of each is around 5s (20s in total). At each turn, the model tries to correct the image generated in the previous turn. The editing prompt is designed by human, tailor-made to correct the errors of the input image. In Figure 10, we show the editing results for selected cases in Figure 4 and 5. From the figures one can see that the model fails to correct most errors in 4 turns. These examples show several disadvantages of "correcting by image editing", compared with our methods.

- The image editing model usually needs to run several times to correct the errors, which largely increases the running time.
- The same guidance scale cannot be applied to all images. For some cases, our guidance scale (1.5) leads to trivial changes.
- The model often fails to understand instructions about locations or scales.

Note that these problems are for InstructPix2Pix. In the future there may be an improved image editing model that addresses these problems.

GLIGEN No Retrieval Retrieval

"a parking meter is to the left of a person."



GLIGEN No Retrieval Retrieval

"a book is to the right of a fire hydrant."

GLIGEN No Retrieval Retrieval

"a clock and an apple. the clock is smaller than the apple."

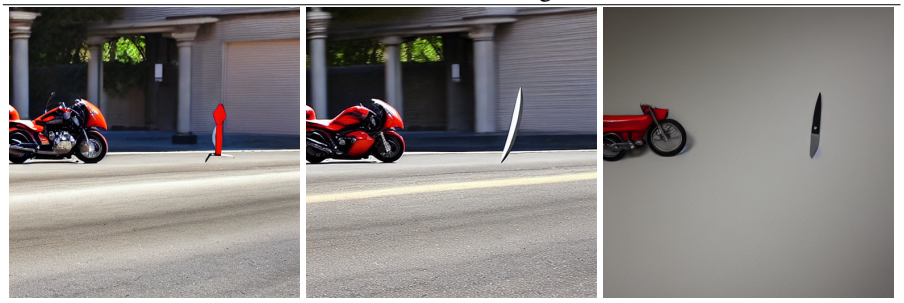
GLIGEN No Retrieval Retrieval

"a bicycle and a toothbrush. the bicycle is bigger than the toothbrush."

Figure 8. Comparing image generations with GLIGEN, GLIGEN with our proposed approach but without retrieval-based feedback, and GLIGEN with our full approach.

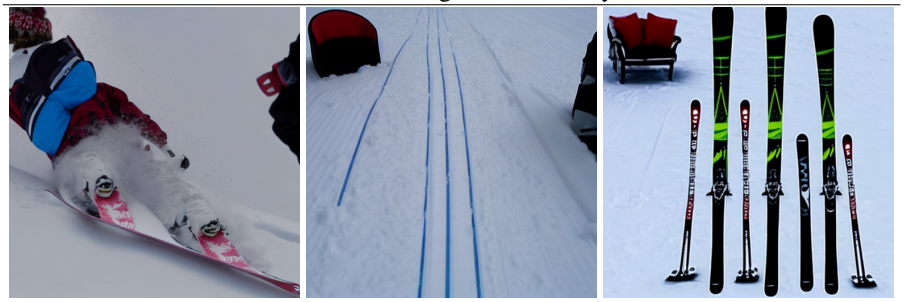
GLIGEN No Retrieval Retrieval

"a bus is below a wine glass."



GLIGEN No Retrieval Retrieval

"a knife is to the right of a motorcycle."



GLIGEN No Retrieval Retrieval

"a couch and a skis. the couch is smaller than the skis."

GLIGEN No Retrieval Retrieval

"a cup and a motorcycle. the cup is bigger than the motorcycle."

Figure 9. Comparing image generations with GLIGEN, GLIGEN with our proposed approach but without retrieval-based feedback, and GLIGEN with our full approach.

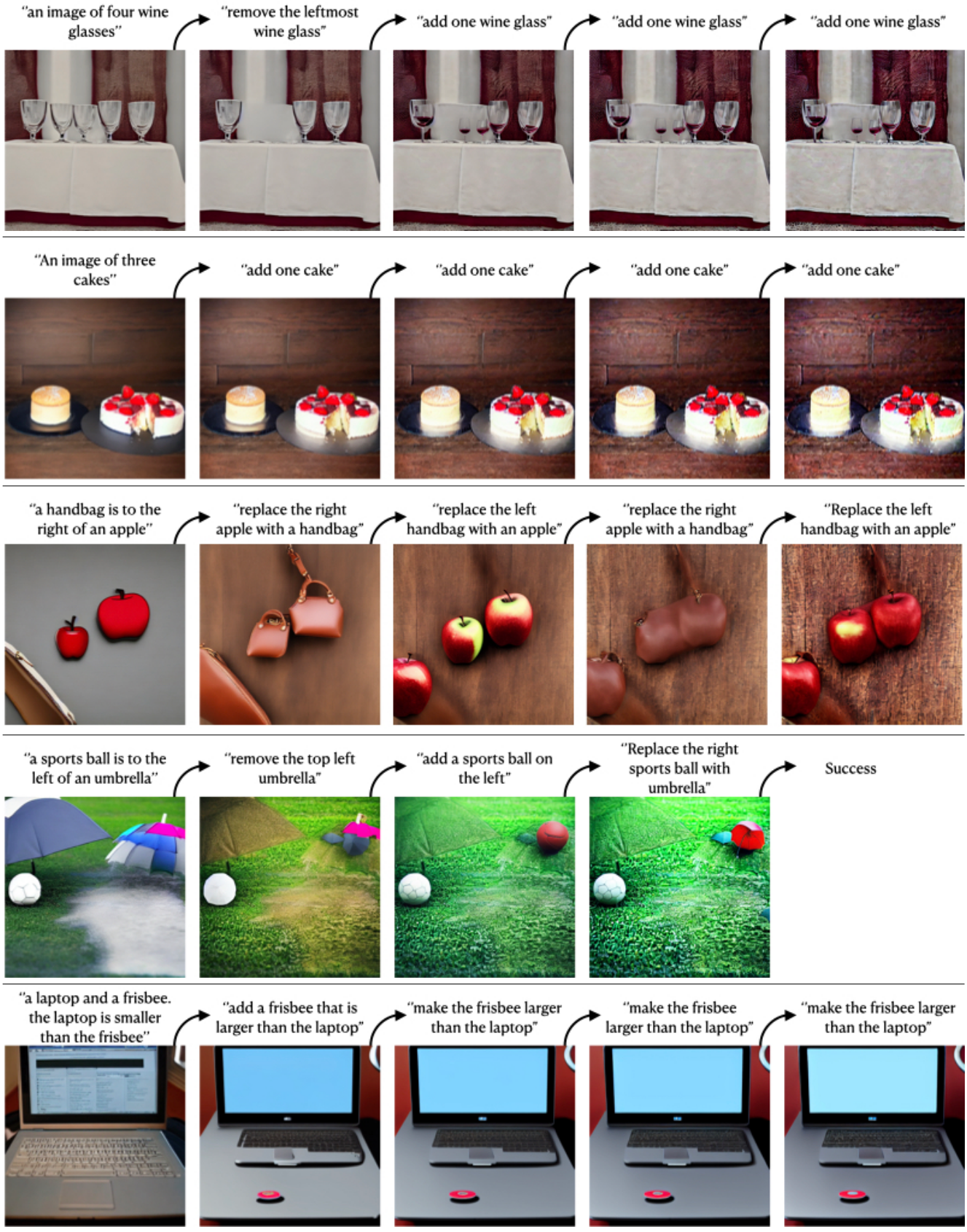


Figure 10. Correcting generation errors via image editing with InstructPix2Pix. In each example, we run the model for 4 times. For each run, we manually design the editing prompt based on the output of the previous run. The editing prompt is shown above the images and used to edit the image at previous stage. The corresponding output is shown below the prompt.