Text-to-image Editing by Image Information Removal Supplementary

Zhongping Zhang[1] Jian Zheng[2] Jacob Zhiyuan Fang[2] Bryan A. Plummer[1]
[1]Boston University [2]Amazon Alexa AI
[1]{zpzhang,bplum}@bu.edu [2]{nzhengji, zyfang}@amazon.com

A. Additional Experiment Results

We present additional qualitative results in Figure 1 and Figure 2 to supplement the main paper. The results demonstrate that IIR-Net can modify image content base on user prompts while preserving the text-irrelevant content of the original image. In Figure 1, we observe that IIR-Net successfully preserves shape-related information of the target object in texture editing examples (*e.g.*, "A **wood** airplane" and "A woman skiing on **grassland**"), as well as texture-related information in color editing examples (*e.g.*, "A **red** horse." and "A **green** orange"). In contrast, Imagic [2] may modify the shape information, while ControlNet [3] may modify the texture information. Besides, we observe that our network produces visually more natural images compared to Text2LIVE [1]. *E.g.*, in the example of "A **red** horse," Text2LIVE applies some red effects to the horse, whereas our method directly produces "a red horse" with better consistency to the background. These observations are consistent with our conclusions in the main paper.

B. User study Interface

In our user study experiments, annotators were presented with an input image, a target text, and four edited images generated by different methods. They were asked to evaluate the accuracy of manipulated images according to two aspects: (1) the alignment of the image with the target text, and (2) the preservation of text-irrelevant content from original images. We provide a sample screenshoot in Figure 3.

References

[1] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 707–723. Springer, 2022. 1

[2] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[3] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1

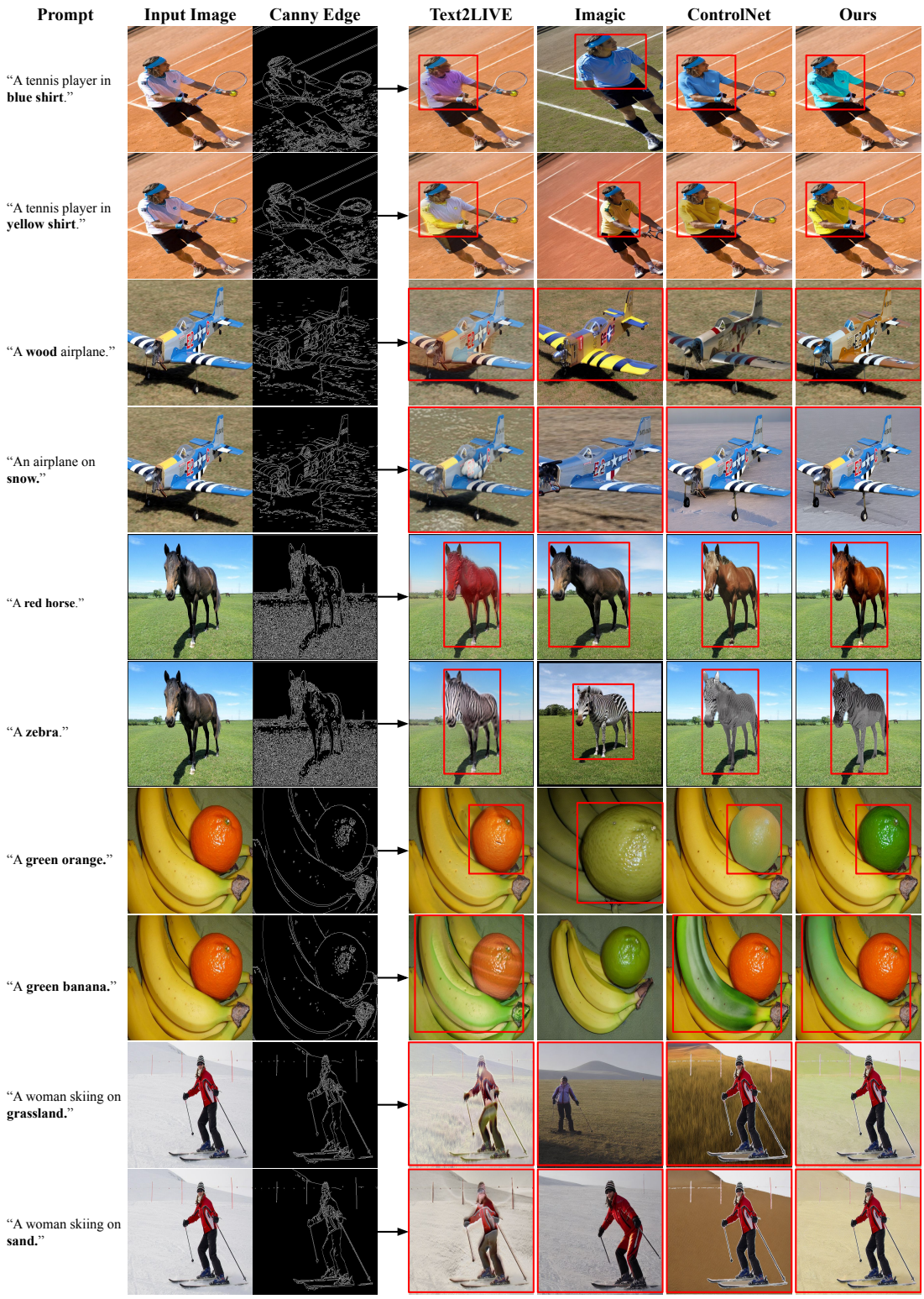

Figure 1. Additional comparison results between IIR-Net and the baselines on the COCO dataset. We set the image resolution to 512×512. We observe that our method effectively modifies the input image while preserving the text-irrelevant content. For instance, in the example of "A tennis player in **blue shirt**," IIR-Net retains both the shape and texture attributes of the original shirt, whereas the other baselines either introduce limited visual effects or modify text-irrelevant content such as textures or shape. See Appendix A for further discussion.

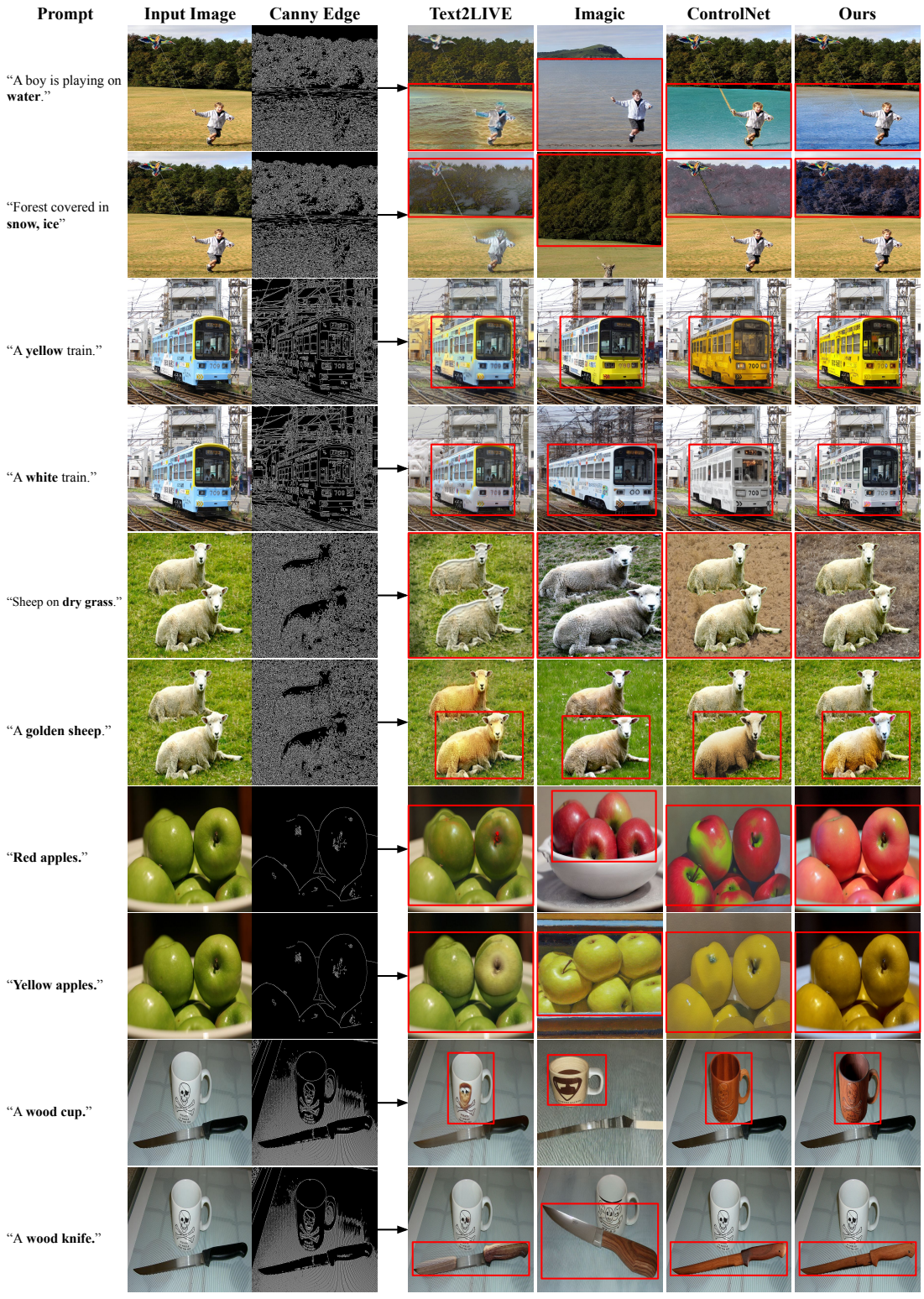

Figure 2. Additional comparison results between IIR-Net and baselines on the COCO dataset. We set the image resolution to 512×512. See Appendix A for discussion.

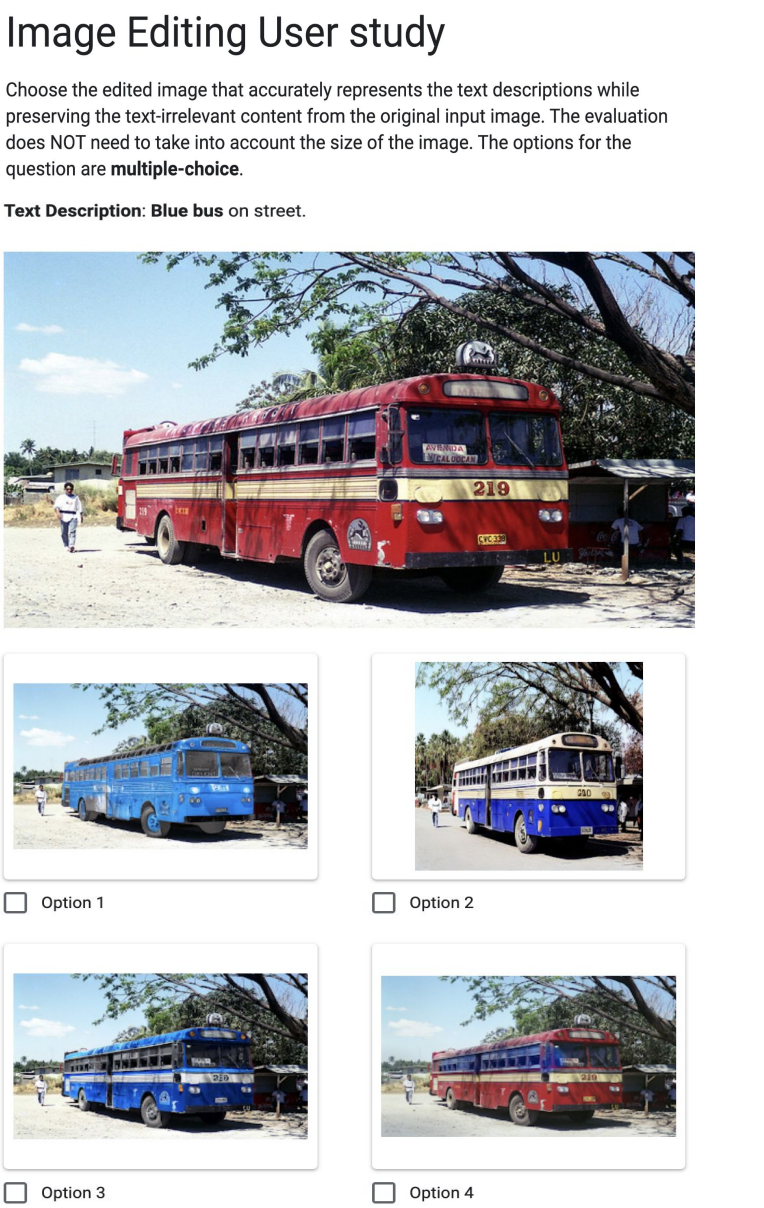

Figure 3. **User study screenshot.** A sample screenshot illustrating one of the questions presented to participants in our user study. See Appendix B for discussion.