Supplementary Material for "DreamRenderer: Taming Multi-Instance Attribute Control in Large-Scale Text-to-Image Models"

Dewei Zhou, Mingwei Li, Zongxin Yang, Yi Yang
Zhejiang University, Zhongguancun Academy
{zdw1999,mingweili,yangzongxin,yangyics}@zju.edu.cn

July 30, 2025

Appendix

A. More Qualitative Results

Figs. A to C presents additional qualitative results of our method. Fig. A showcases our method's effectiveness and flexibility in graphic design applications. The example demonstrates how DreamRenderer can generate distinct design variations with only little input modifications. In the top design, we generate an orange bold "Dream Big!" text in the center, and with just a small modification to the text prompt from "Dream Big!" to two different color text prompts (orange "Dream" and blue "Renderer"), our method generates an entirely new variant (bottom).

As shown in Fig. B, we tackle the challenging task of simultaneously generating multiple specified person. Generating seven distinct person while maintaining consistent identity and appearance is particularly difficult, as it requires the model to understand and preserve individual characteristics across different poses and viewpoints. Our method successfully generates natural-looking results where not only is the identity consistently maintained across all instances, but the generated images also precisely align with the provided depth conditions. This demonstrates our model's robust capability in handling complex multi-instance person generation tasks.

B. More Results on COCO-POS Benchmark

Fig. D presents the qualitative results of our method compared with FLUX [3] and 3DIS [35] on both depth-guided and canny-guided generation. As shown in the figure, our DreamRenderer consistently outperforms both FLUX and 3DIS, particularly when generating multiple instances. In the **depth-guided** scenarios (top rows), our method accurately preserves the spatial relationships indicated in the depth maps while ensuring that each instance's attributes are correctly rendered according to the textual descriptions. The baseline methods struggle with attribute entanglement, often generating instances with incorrect colors, patterns, or other visual characteristics.

For the more challenging **canny-guided** generation (bottom rows), the performance gap is even more pronounced. While FLUX and 3DIS frequently produce instances with misaligned attributes or distorted appearances, our method maintains attribute fidelity even with minimal structural guidance from canny edges. This demonstrates the effectiveness of our **Hard Text Attribute Binding** mechanism, which ensures each instance's text embedding correctly binds with its corresponding visual features during the generation process.

Notably, our method achieves these improvements without compromising image quality. The generated images exhibit clear details, natural textures, and coherent global compositions, demonstrating that our **Image Attribute Binding** approach successfully preserves the model's inherent rendering capabilities while enhancing attribute control.

C. More Results on COCO-MIG Benchmark

Fig. E, Fig. F, Fig. G, and Fig. H present comparative results obtained by applying DreamRenderer to re-render outputs from various layout-to-image methods. The red boxes indicate that all methods exhibit challenges in attribute binding to varying degrees.

GLIGEN [9], one of the earliest methods with layout control, shows the most severe attribute confusion, often generating objects with incorrect colors or patterns and struggling with spatial consistency. **InstanceDiffusion** [21] improves instance separation but still struggles with attribute binding across multiple instances, and notably suffers from lower visual quality with blurry textures and less detailed renderings. **MIGC** [33] produces high-resolution results but frequently fails to adhere to depth conditions properly, and often generates images with overly saturated colors and unrealistic brightness. Even the most advanced

method **3DIS** [35] exhibits significant attribute binding errors when handling multiple instances with similar categories but different properties.

Our DreamRenderer consistently enhances performance across all methods by ensuring accurate attribute binding while preserving image quality. The improvements become more pronounced when controlling multiple similar instances with different visual properties, confirming our method's effectiveness in addressing attribute entanglement regardless of the underlying architecture.

D. More Results on Hard Text Attribute Binding

Fig. I demonstrates the effectiveness of our Hard Text Attribute Binding mechanism. **The naive approach** generally preserves basic attributes but suffers from severely degraded image quality with noticeable artifacts. The model **without** Hard Text Attribute Binding produces visually appealing images but frequently fails to correctly bind text attributes to the generated content, resulting in misaligned visual elements. In contrast, our full model with **Hard Text Attribute Binding** achieves both high image quality and accurate attribute preservation. Comparing the three approaches side by side, we observe that our method successfully addresses the limitations of both alternative approaches, delivering consistent text-image alignment without compromising visual fidelity.

E. Limitations

Despite the significant advancements achieved by DreamRenderer in multi-instance generation control, several limitations persist. For canny-guided generation, our method's performance is less robust compared to depth-guided generation, primarily constrained by the capabilities of the underlying FLUX-Canny model, as evidenced by the results in body part's Tab. 1. Furthermore, we observe a substantial decrease in the success ratio as the number of controlled instances increases, a phenomenon particularly pronounced in canny-guided generation, where the success rate drops from 23.28% with two instances to considerably lower values with additional instances. Although the Hard Text Attribute Binding mechanism significantly improves attribute binding accuracy, the attribute entanglement issue remains not fully resolved when handling complex scenes with multiple overlapping instances, indicating room for further improvement in this domain.

We would like to acknowledge several prior works that inspired this work [1–8, 8, 10–20, 22–36].

References

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2

[2] Shariq Farooq Bhat, Niloy Mitra, and Peter Wonka. Loosecontrol: Lifting controlnet for generalized depth conditioning. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

[3] BlackForest. Black forest labs; frontier ai lab, 2024. 1

[4] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*, 2023.

[5] Aditya Ramesh et al. Hierarchical text-conditional image generation with clip latents, 2022.

[6] Daiheng Gao, Shilin Lu, Shaw Walters, Wenbo Zhou, Jiaming Chu, Jie Zhang, Bang Zhang, Mengxi Jia, Jian Zhao, Zhaoxin Fan, et al. Eraseanything: Enabling concept erasure in rectified flow transformers. *arXiv preprint arXiv:2412.20413*, 2024.

[7] Leyang Li, Shilin Lu, Yan Ren, and Adams Wai-Kin Kong. Set you straight: Auto-steering denoising trajectories to sidestep unwanted concepts. *arXiv preprint arXiv:2504.12782*, 2025.

[8] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback. *arXiv preprint arXiv:2404.07987*, 2024. 2

[9] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. *CVPR*, 2023. 1, 7

[10] Han Lin, Jaemin Cho, Abhay Zala, and Mohit Bansal. Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. *arXiv preprint arXiv:2404.09967*, 2024. 2

[11] Rui Liu, Xiaohan Wang, Wenguan Wang, and Yi Yang. Bird's-eye-view scene graph for vision-language navigation. In *ICCV*, 2023.

[12] Rui Liu, Wenguan Wang, and Yi Yang. Vision-language navigation with energy-based policy. In *NeurIPS*, 2024.

[13] Rui Liu, Wenguan Wang, and Yi Yang. Volumetric environment representation for vision-language navigation. In *CVPR*, 2024.

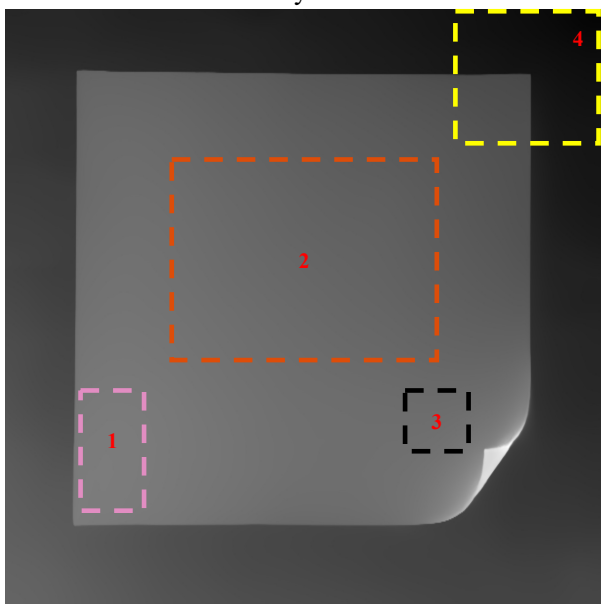
[14] Rui Liu, Sheng Fan, Wenguan Wang, and Yi Yang. Underwater visual slam with depth uncertainty and medium modeling. In *ICCV*, 2025.

[15] Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. Tf-icon: Diffusion-based training-free cross-domain image composition. In *ICCV*, 2023.

[16] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. *CVPR*, 2024.

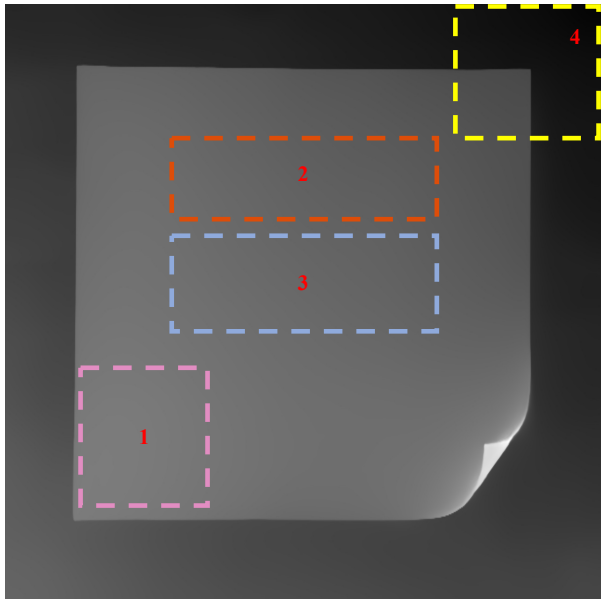
[17] Shilin Lu, Zihan Zhou, Jiayou Lu, Yuanzhi Zhu, and Adams Wai-Kin Kong. Robust watermarking using generative priors against image editing: From benchmarking to advances. *arXiv preprint arXiv:2410.18775*, 2024.

Layout



Designed Results

1) Decorative element: a **feather** 2) A paper with text **"Dream Big!"** in a bold, hand-painted orange font with a fiery, energetic style, letters slightly uneven for a dynamic feel; 3) A black apple logo 4) Decorative element: a **moon**

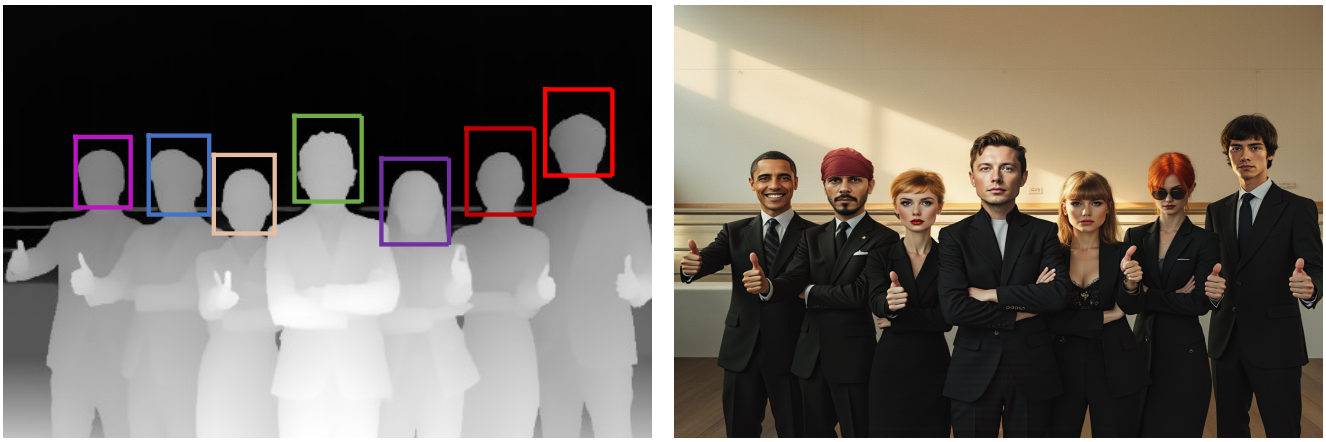


1) Decorative element: **feathers** 2) A paper with text **"Dream"** in a bold, hand-painted orange font with a fiery, energetic style, letters slightly uneven for a dynamic feel; 3) A paper with text **"Renderer"** in a bold, hand-painted blue font with a fiery, energetic style, letters slightly uneven for a dynamic feel; 4) Decorative element: shiny **stars**

Figure A. **Qualitative results of our method on multi-instance design generation. (§ A)** Our model enables efficient design iterations with precise control, enabling designers to explore multiple design layout by only little input modifications.

[18] 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. *ICML*, 2022.

[19] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and effi-



"many famous people in black suit are standing together. 1) Barack Obama 2) Pirates of the Caribbean 3) Audrey Hepburn 4) Elon Musk 5) Taylor Swift 6) Girl with brown hair & sunglasses 7) Bruce Lee

Figure B. **Qualitative results of our method on multi-instance person generation. (§ A)** Our model successfully generates 7 different persons simultaneously, which is a notably **challenging** task in image generation. The results demonstrate consistent identity preservation across all instances while maintaining natural appearance variations. Each generated image accurately follows the corresponding depth condition, showing our model's ability to handle complex spatial relationships and viewpoint variations.

Figure C. **Qualitative results of our method on different architecture style generation. (§ A)** Our model successfully generates images with the same layout input, demonstrating its versatility in capturing and reproducing diverse artistic styles.

cient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024.

[20] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NIPS*, 2022. 2

[21] Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit

Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation, 2024. 1, 7

[22] Yilin Wang, Haiyang Xu, Xiang Zhang, Zeyuan Chen, Zhizhou Sha, Zirui Wang, and Zhuowen Tu. Omnicontrolnet: Dual-stage integration for conditional image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7436–7448, 2024. 2

[23] Rui Xie, Ying Tai, Kai Zhang, Zhenyu Zhang, Jun Zhou, and Jian Yang. Addsr: Accelerating diffusion-based blind super-resolution with adversarial diffusion distillation. *ArXiv*, abs/2404.01717, 2024.

[24] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arxiv:2308.06721*, 2023.

[25] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation, 2022.

[26] Hui Zhang, Dexiang Hong, Tingwei Gao, Yitong Wang, Jie Shao, Xinglong Wu, Zuxuan Wu, and Yu-Gang Jiang. Creatilayout: Siamese multimodal diffusion transformer for creative layout-to-image generation. *arXiv preprint arXiv:2412.03859*, 2024.

[27] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023.

[28] Chen Zhao, Weiling Cai, Chenyu Dong, and Chengwei Hu. Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. *CVPR*, 2024.

[29] Chen Zhao, Weiling Cai, Chengwei Hu, and Zheng Yuan. Cycle contrastive adversarial learning with structural consistency for unsupervised high-quality image deraining transformer. *Neural Networks*, 2024.

[30] Chen Zhao, Chenyu Dong, and Weiling Cai. Learning a physical-aware diffusion model based on transformer for underwater image enhancement. *arXiv preprint arXiv:2403.01497*, 2024.

[31] Chen Zhao, Zhizhou Chen, Yunzhe Xu, Enxuan Gu, Jian Li, Zili Yi, Qian Wang, Jian Yang, and Ying Tai. From zero to detail: Deconstructing ultra-high-definition image restoration from progressive spectral perspective. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17935–17946, 2025.

[32] Dewei Zhou, Zongxin Yang, and Yi Yang. Pyramid diffusion models for low-light image enhancement. In *IJCAI*, 2023.

[33] Dewei Zhou, You Li, Fan Ma, Zongxin Yang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. *CVPR*, 2024. 1, 7

[34] Dewei Zhou, You Li, Fan Ma, Zongxin Yang, and Yi Yang. Migc++: Advanced multi-instance generation controller for image synthesis, 2024.

[35] Dewei Zhou, Ji Xie, Zongxin Yang, and Yi Yang. 3dis: Depth-driven decoupled instance synthesis for text-to-image generation. *arXiv preprint arXiv:2410.12669*, 2024. 1, 2, 8

[36] Dewei Zhou, Ji Xie, Zongxin Yang, and Yi Yang. 3dis-flux: simple and efficient multi-instance generation with dit rendering. *arXiv preprint arXiv:2501.05131*, 2025. 2

Figure D. **Qualitative comparison with FLUX and 3DIS on depth-guided (top) and canny-guided (bottom) generation. (§ A)** Our method produces images with more accurate attributes and better visual quality, while baseline methods often exhibit color and pattern inconsistencies with text prompts.

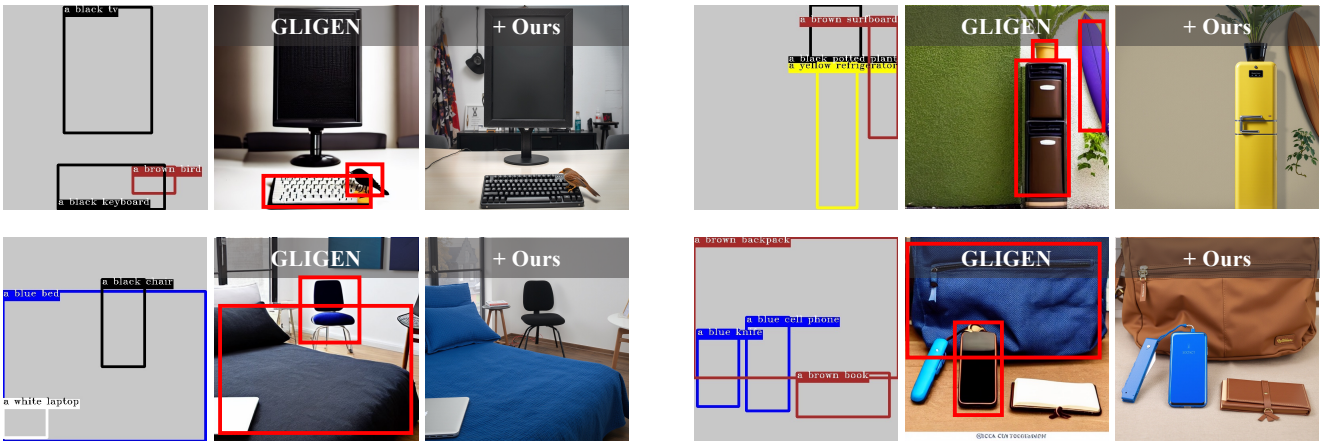


Figure E. **Additional qualitative comparison on the COCO-MIG benchmark. (§ A)** We show more results of re-rendering on GLIGEN [9]. We highlight with red boxes the areas where the compared method exhibits noticeable attribute generation errors.

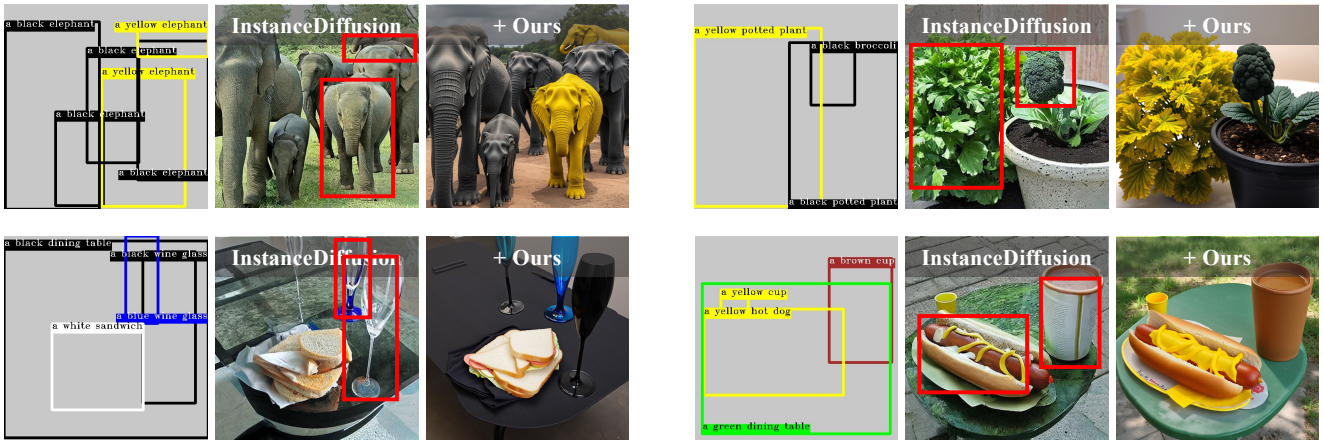


Figure F. **Additional qualitative comparison on the COCO-MIG benchmark. (§ A)** We show more results of re-rendering on InstanceDiffusion [21]. We highlight with red boxes the areas where the compared method exhibits noticeable attribute generation errors.

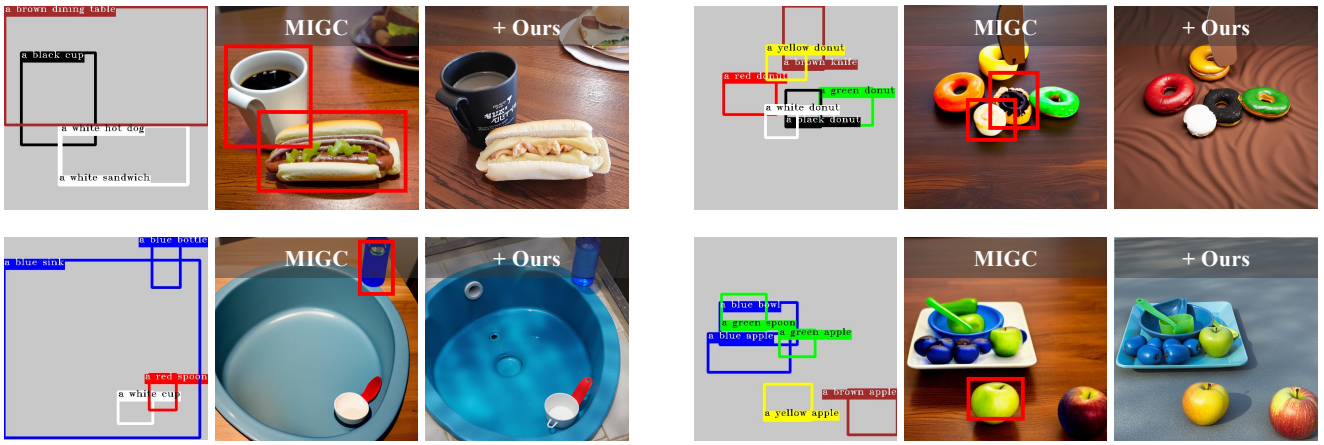


Figure G. **Additional qualitative comparison on the COCO-MIG benchmark. (§ A)** We show more results of re-rendering on MIGC [33]. We highlight with red boxes the areas where the compared method exhibits noticeable attribute generation errors.

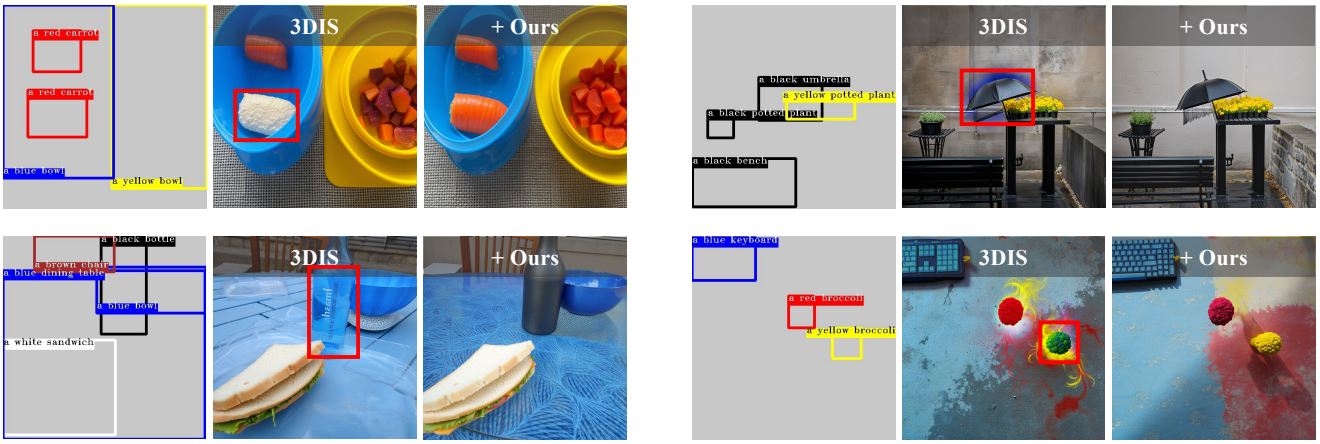


Figure H. **Additional qualitative comparison on the COCO-MIG benchmark. (§ A)** We show more results of re-rendering on 3DIS [35]. We highlight with red boxes the areas where the compared method exhibits noticeable attribute generation errors.

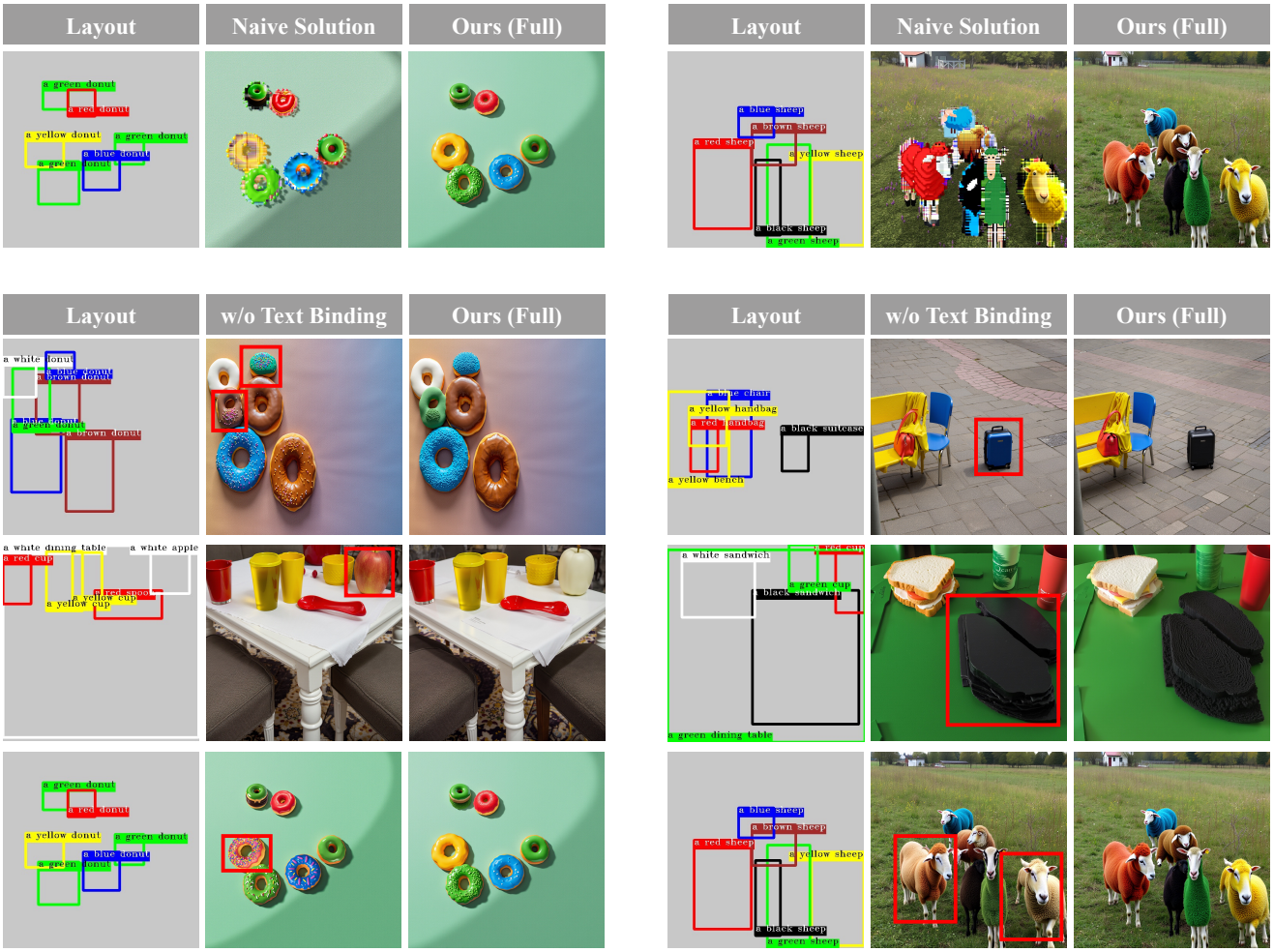


Figure I. **Ablation study on the Hard Text Attribute Binding mechanism. (§ D)** Top: Results from the naive approach, which maintains basic attribute correctness but produces poor image quality with significant artifacts. Bottom: Results without hard text binding, showing good visual quality but frequent attribute binding failures. Results from our full model, demonstrating both high-quality image generation and accurate text attribute binding. The comparison highlights how our method effectively balances visual quality and text-prompt adherence.