DetDiffusion: Synergizing Generative and Perceptive Models for Enhanced Data Generation and Perception

Supplementary Material

A. Details of Experiments

A.1. Fidelity

The experiments on Fidelity are conducted using the COCO-Thing-Stuff dataset [1], and its perception-aware attribute is derived from a pre-trained YOLOv4. As the detector can only identify the 80 categories in COCO2017 [8], objects that cannot be detected are assigned the perception aware attribute of [background]. For example, the obtained text prompt is ”`An image with (person,<23><44>,[easy]),(person,<45><80>,[hard]),(playingfield,<0><400>,[background])`”, where the location token follows the approach [5], using two location bins to represent the upper-left and lower-right coordinates of the object.

A.2. Trainability

We conducted training of our *DetDiffusion* model on the COCO-Thing-Stuff dataset at a resolution of 800x456. The perception-aware attribute of COCO-Thing-Stuff is derived from a pre-trained Faster R-CNN [10].

We then employ the trained generative model to generate a subset of the COCO2017 training set, comprising 47,429 images, using three distinct strategies: *DetDiffusion* $_{easy}$,*DetDiffusion* $_{hard}$, and *DetDiffusion* $_{origin}$. In this context, *DetDiffusion* $_{origin}$ is also obtained through detection utilizing Faster R-CNN on the 47,429 images, with the corresponding detection outcomes presented in Table 6. Finally, a Faster R-CNN with an R-50-FPN[mmdetection] backbone was trained using the combined 47,429 images and the coco2017 training set, followed by an evaluation of its performance on the coco2017 validation set.

B. More Results of Experiments

B.1. Ablation on perception-aware loss

This section investigates two important components of the perception aware loss, namely $\sqrt{\bar{\alpha}_t}$ [7] and dice loss. Table 1 shows that $\sqrt{\bar{\alpha}_t}$ is crucial for perception aware loss as it can reduce the impact of noise to some extent. And dice loss is also essential as a complement to mask loss.

C. More Discussion

Limitation. Currently, images generated by *DetDiffusion* can only be utilized to train object detectors. More flexible usage of generated images including the incorporation with the generative pre-training [3, 13] and the contrastive learning [2, 9] is an interesting future research direction. How to generate high-quality images aligned with human values without harmful and toxic content [4, 6, 11] is also important for the practical usage of *DetDiffusion*.

	mAP↑	AP$_{50}$↑	AP$_{75}$↑
DetDiffusion	**31.2**	**40.2**	**35.6**
w/o $\sqrt{\bar{\alpha}_t}$	29.6	38.6	34.3
w/o dice loss	29.6	39.4	34.5

Table 1. **Ablation on perception-aware loss.** We display the ablation experiments about $\sqrt{\bar{\alpha}_t}$ and dice loss.

D. More Qualitative Results

Figure 1 presents a comparison between our generated images and some state-of-the-art models [5, 12]. Our results demonstrate accurate hierarchical relationships, as evidenced by the realistic depiction of objects such as the car and dog, and the high realism of dynamic human figures. Additionally, our generated images exhibit high quality, as illustrated in (g) and (h).

Figure 2 and Figure 3 illustrate examples of images generated by providing easy and hard attributes. When $[hard]$ attributes are provided, the confidence score of the object is decreased to varying degrees, or some objects are missed in the detection. The changes in the objects in Figure 2 are particularly noticeable. For instance, in (a), the confidence score is significantly reduced due to the reflection on the monitor screen. Additionally, changes in color, blurring of text, occlusion, and deformation lead to decreased confidence scores and missed detection in other examples.

The changes in the objects in Figure 3 are minimal, yet they have a significant impact on the detector. This highlights the challenging examples that cannot be observed by the human eye but greatly affect the detector, which is what we aim to learn through attributes. For instance, in (a) and (c), only changes in color lead to a significant decrease in confidence score, while in (b), a minor change results in a substantial decrease in confidence score. Moreover, in (d) and (e), there are only slight deformations. Due to the sensitivity of detectors to subtle features, the use of prior-constructed image variations may not be effective for such cases. Our approach, which directly utilizes detection information to generate images, can reflect these differences, thereby further enhancing training.

DetDiffusion is capable of generating diverse scenes in Figure 4, thus demonstrating our fidelity and diversity.

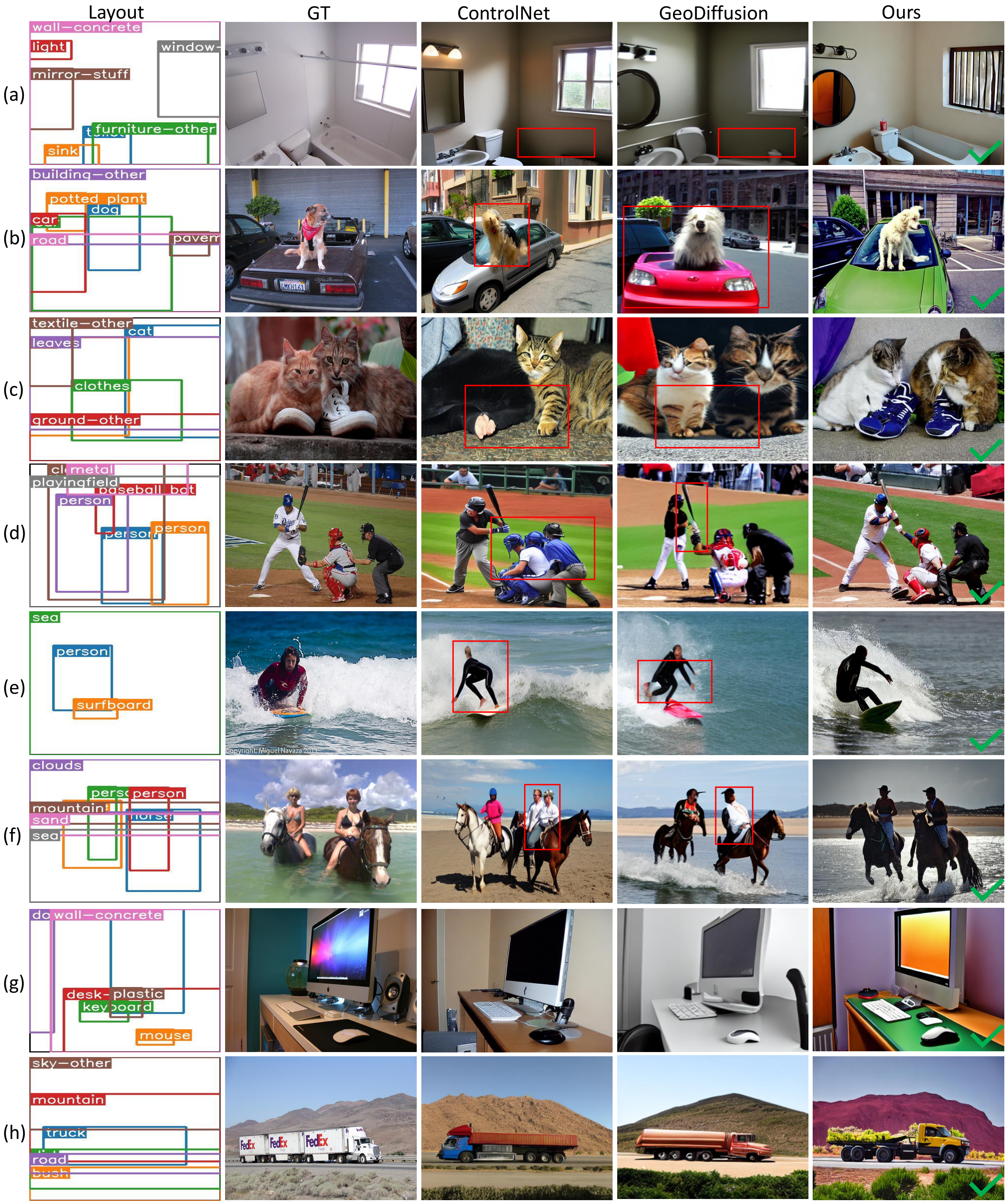

Figure 1. **More qualitative comparison on the COCO dataset.** We highlight some region with red boxes to facilitate comparison.

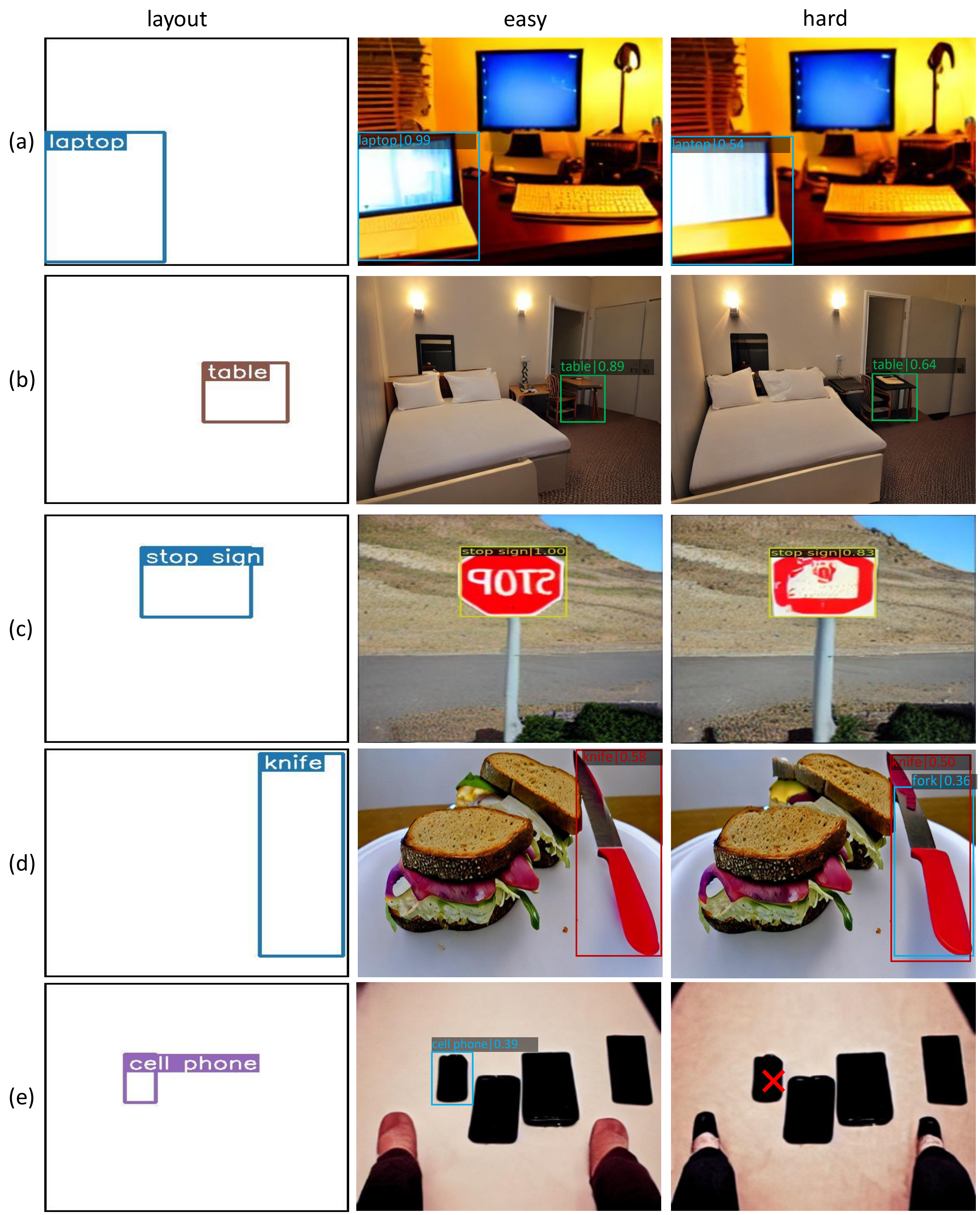

Figure 2. **More examples of easy and hard perception-aware attribute.** These examples are apparent to see the gap.

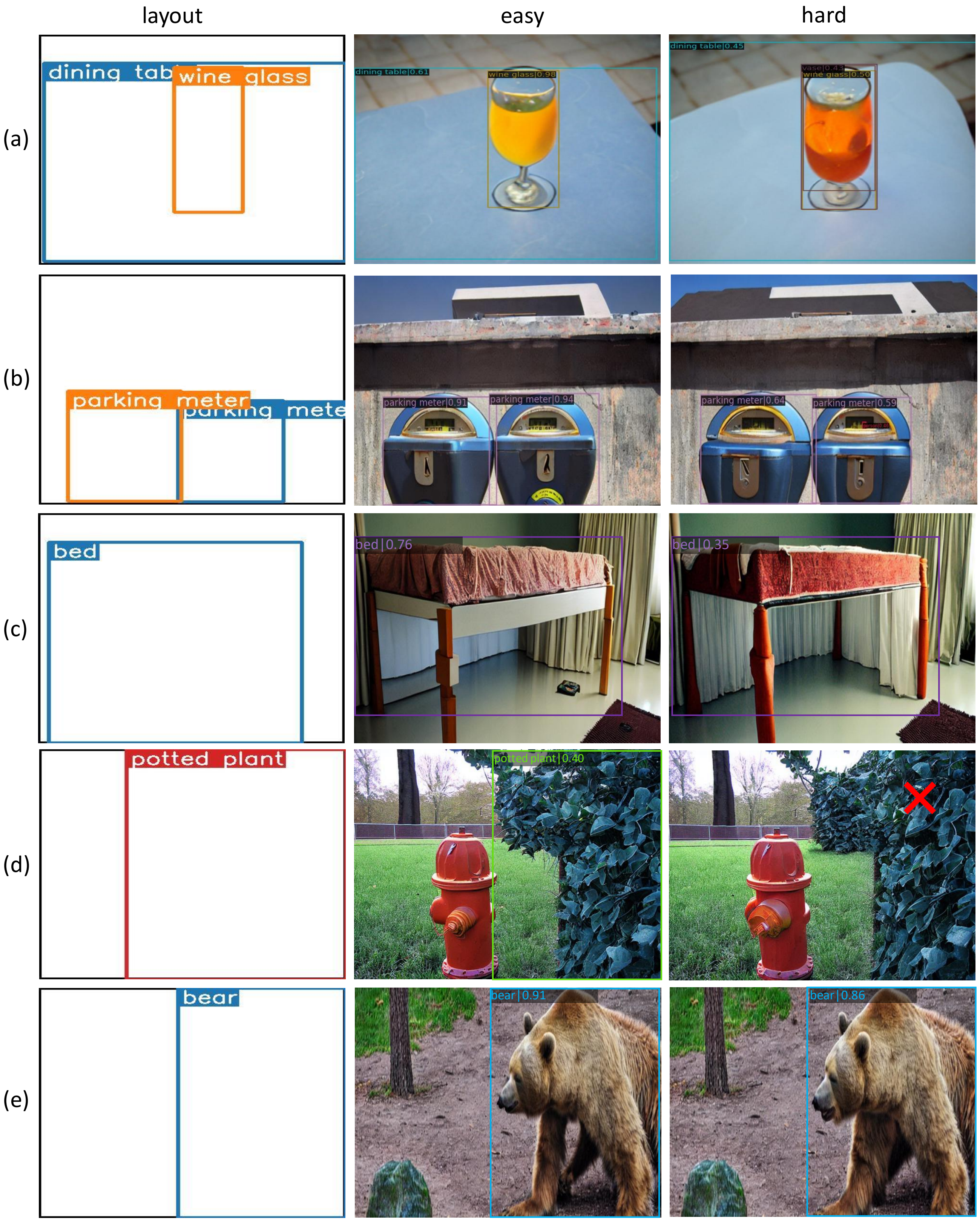

Figure 3. **More examples of easy and hard perception-aware attribute.** The gap between easy and hard examples is not obvious, while the gap between confidence scores is large.

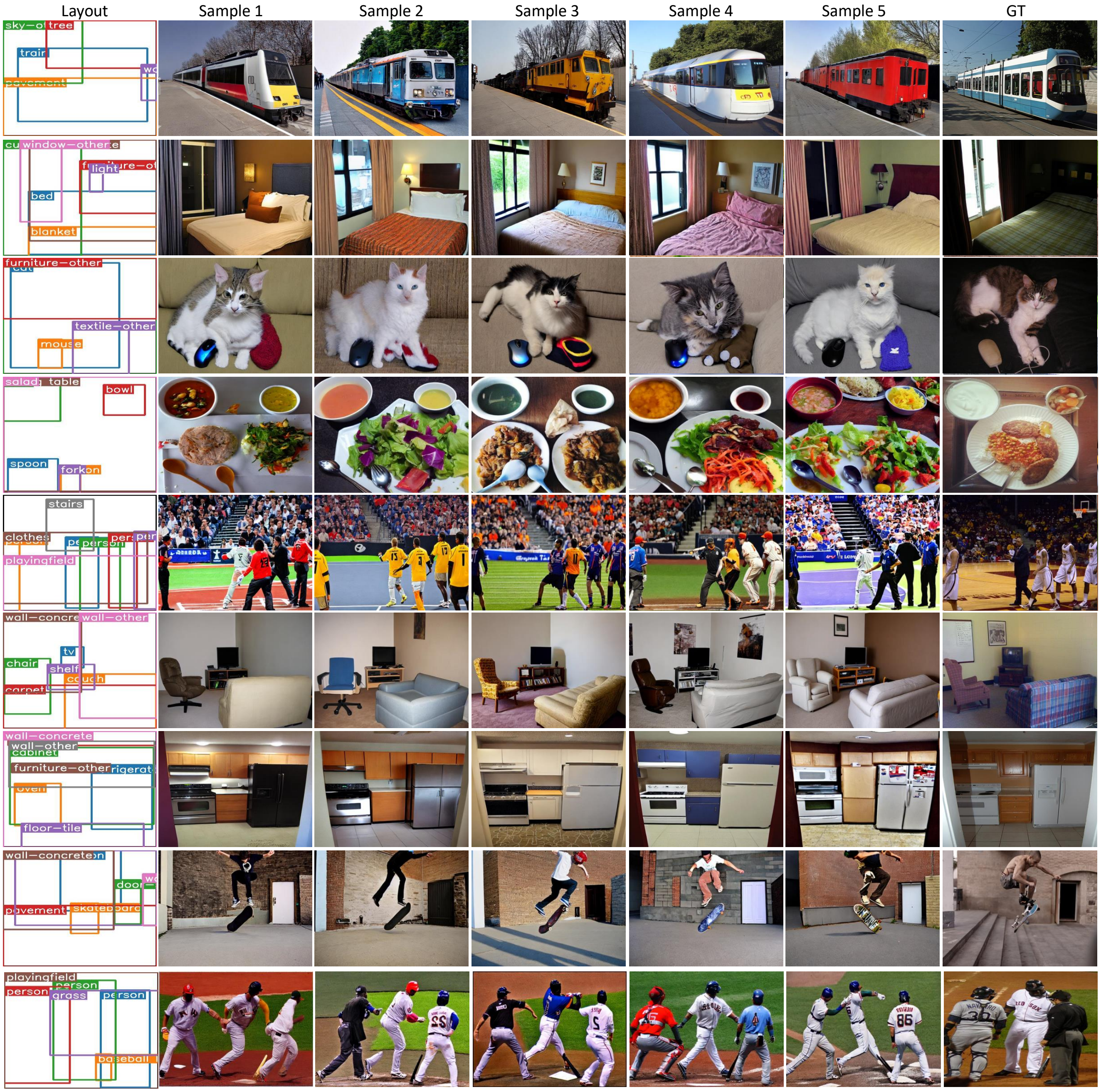

Figure 4. **More qualitative results of the same layout with random noises.**

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, 2018. 1

[2] Kai Chen, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Multisiam: Self-supervised multi-instance siamese representation learning for autonomous driving. In *ICCV*, 2021. 1

[3] Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Mixed autoencoder for self-supervised visual representation learning. In *CVPR*, 2023. 1

[4] Kai Chen, Chunwei Wang, Kuo Yang, Jianhua Han, Lanqing Hong, Fei Mi, Hang Xu, Zhengying Liu, Wenyong Huang, Zhenguo Li, et al. Gaining wisdom from setbacks: Aligning large language models via mistake analysis. *arXiv preprint arXiv:2310.10477*, 2023. 1

[5] Kai Chen, Enze Xie, Zhe Chen, Lanqing Hong, Zhenguo Li, and Dit-Yan Yeung. Integrating geometric control

into text-to-image diffusion models for high-quality detection data generation via text prompt. *arxiv preprint arXiv: 2306.04607*, 2023. 1

[6] Yunhao Gou, Zhili Liu, Kai Chen, Lanqing Hong, Hang Xu, Aoxue Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. Mixture of cluster-conditional lora experts for vision-language instruction tuning. *arXiv preprint arXiv:2312.12379*, 2023. 1

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[9] Zhili Liu, Jianhua Han, Kai Chen, Lanqing Hong, Hang Xu, Chunjing Xu, and Zhenguo Li. Task-customized self-supervised pre-training with scalable dynamic routing. In *AAAI*, 2022. 1

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1

[11] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *CVPR*, 2023. 1

[12] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 1

[13] LIU Zhili, Kai Chen, Jianhua Han, HONG Lanqing, Hang Xu, Zhenguo Li, and James Kwok. Task-customized masked autoencoder via mixture of cluster-conditional experts. In *ICLR*, 2023. 1