Freestyle Layout-to-Image Synthesis
Supplementary Material

Han Xue[1,2] Zhiwu Huang[2,3] Qianru Sun[2] Li Song[1,4] Wenjun Zhang[1]
[1]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University
[2]Singapore Management University [3]University of Southampton
[4]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
{xue_han, song_li, zhangwenjun}@sjtu.edu.cn qianrusun@smu.edu.sg zhiwu.huang@soton.ac.uk

This supplementary material includes an extensive description of Cross-Attention (CA) (§A), the algorithm of Rectified Cross-Attention (RCA) (§B), additional implementation details (§C), more qualitative results on freestyle layout-to-image synthesis (FLIS) (§D), more comparisons with layout-to-image synthesis (LIS) baselines (§E), the diversity evaluation (§F), discussions about the optimal form of textual inputs (§G), more failure cases of our approach (§H), some results on rectangular datasets (§I), and discussions about the societal impact (§J).

A. Cross-Attention (CA) in Stable Diffusion

This is supplementary to Section 4.1 "**rectifying diffusion model**". In this section, we provide an elaboration of Cross-Attention (CA) for a clearer comparison with our proposed Rectified Cross-Attention (RCA). For a CA layer in Stable Diffusion, let φ_I and φ_T denote the input image feature and text embeddings, respectively. Image queries Q, text keys K, and text values V can be calculated by:

$$Q = W_Q \cdot \varphi_I,\ K = W_K \cdot \varphi_T,\ V = W_V \cdot \varphi_T, \tag{S1}$$

where W_Q, W_K, and W_V are learnable projection matrices. Then attention score maps $\mathcal{M}$ are computed as:

$$\mathcal{M} = \frac{QK^T}{\sqrt{d}} \in \mathbb{R}^{C\times H\times W}, \tag{S2}$$

where d is the scaling factor that is set as the dimension of the queries and keys, and C, H, W are the channel number, height, and weight of $\mathcal{M}$, respectively. After that, we can calculate the output image feature $\mathcal{O}$ of this CA layer by:

$$\mathcal{O} = \mathrm{softmax}(\mathcal{M})V. \tag{S3}$$

A visual illustration of CA is shown in Figure S1. In contrast, the proposed RCA rectifies $\mathcal{M}$ via Eq. 2 in the main paper before applying softmax.

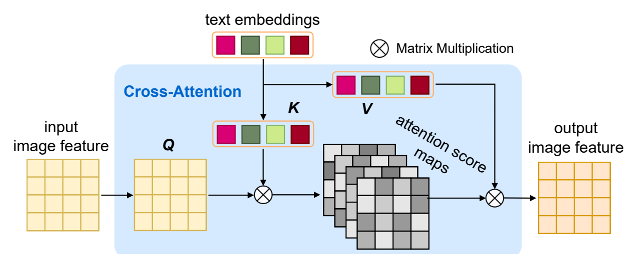

Figure S1. An illustration of Cross-Attention (CA).

B. Algorithm

This is supplementary to Section 4.1 "**rectifying diffusion model**". The computation pipeline of RCA is illustrated in Algorithm 1.

C. Additional implementation details

This is supplementary to Section 5.1 "**experimental settings**". Training on COCO-Stuff/ADE20K takes about 6/2 days on a single NVIDIA A100 GPU. All our experiments are conducted using Stable Diffusion v1.4.

D. More qualitative results on FLIS

This is supplementary to Section 5.2 "**qualitative evaluation on FLIS**". In Figures S2, S3, and S4, we present more FLIS results by using the proposed model. They demonstrate the capability of our method for FLIS and its high potential to spawn various applications.

E. More comparisons with LIS baselines

This is supplementary to Section 5.3 "**comparison with LIS baselines**". In this section, we provide more comparison results between SPADE [7], CC-FPSE [5], OASIS [9], SC-GAN [12], PITI [10], and our method. Figures S5 and S6 show the results on COCO-Stuff [2] and ADE20K [14],

Algorithm 1: RCA

Input: Input image feature φ_I, text embeddings φ_T, and layout l

Output: Output image feature $\mathcal{O}$

Get image queries Q, text keys K, and text values V by Eq. (S1)

Get attention score maps $\mathcal{M} \in \mathbb{R}^{C\times H\times W}$ by Eq. (1)

Initialize a mask $L \in \mathbb{R}^{C\times H\times W}$

Resize l to match the spatial size of $\mathcal{M}$

for k **in** $\{0, 1, ..., C-1\}$ **do**

  if *The k-th text embedding corresponds to a concept m* **then**

    Find the binary map $l^m \in \mathbb{R}^{H\times W}$ in l corresponding to this concept

    $L^k \leftarrow l^m$

  else

    $L^k \leftarrow 1$

  end

end

Get the rectified attention score maps $\widehat{\mathcal{M}}$ by Eq. (2)

Get the output image feature $\mathcal{O}$ by Eq. (3)

respectively. These results indicate the superiority of our method in generating high-fidelity images in the context of LIS.

For a fair comparison with PITI, we replace its pre-trained text-to-image diffusion model (GLIDE [6]) with Stable Diffusion [8]. Due to time limits, we carefully tune learning rates only when training its model (we call it PITI w/ SD). Some visual results are provided in Figure S7. The images synthesized by PITI w/ SD exhibit good visual quality but the spatial alignment with the input layout is poor (clearly poorer than ours). The quantitative comparison results are also provided in Table S1.

Here we compare our FreestyleNet with additional related works including Lab2Pix-V2 [15], sVQGAN-T [1], and PoE-GAN [4]. The comparison results under the in-distribution setting is reported in Table S2. As neither sVQGAN-T [1] nor PoE-GAN [4] provide code, their results are copied from their papers. These results showcase our superiority over the others.

F. Diversity evaluation

This is supplementary to Section 5.3 "**comparison with LIS baselines**". In this section, we conduct some experiments to evaluate the generation diversity of different methods. Note that our model naturally enables generation with high diversity from the same layout by using various texts (see Figures 1, 4, and 6 in the main paper). Here we perform the diversity evaluation in the conventional LIS setting. Following OASIS [9], we calculate LPIPS [13] between images generated from the same layout (and same text for our model) but with randomly sampled noise. The evaluation results are provided in Table S3. Our model achieves the highest LPIPS among all comparison methods. We also show some visual samples in Figure S8.

Method	PITI w/ SD	FreestyleNet (ours)
FID↓	15.5	**14.4**
mIoU↑	13.1	**40.7**

Table S1. Quantitative comparison results with PITI w/ SD on COCO-Stuff.

Method	COCO-Stuff		ADE20K	
	FID↓	mIoU↑	FID↓	mIoU↑
Lab2Pix-V2 [15]	18.1	40.5	31.3	41.0
sVQGAN-T [1]	28.8	-	38.4	-
PoE-GAN [4]	15.8	-	-	-
FreestyleNet (ours)	**14.4**	**40.7**	**25.0**	**41.9**

Table S2. Comparison results with additional related works.

Method	LPIPS↑	
	COCO-Stuff	ADE20K
CC-FPSE [5]	0.089	0.129
OASIS [9]	0.345	0.285
PITI [10]	0.523	0.480
FreestyleNet (ours)	**0.592**	**0.591**

Table S3. Diversity evaluation results. Pix2PixHD [11], SPADE [7], and SC-GAN [12] do not support diverse generation (*i.e.*, LPIPS is 0).

G. Optimal form of textual inputs

This is supplementary to Section 4.1 "**rectifying diffusion model**". As full-form image descriptions are expensive (or even intractable) to collect, we suggest using the stacked concepts which can be easily obtained from semantic labels. Moreover, stacked concepts fit naturally into the design of RCA, which builds the relationship between each individual semantic and its position on the image. We actually have explored several alternatives (which perform worse), including (1) keyword-to-sentence translation, (2) learnable prompts, and (3) manual construction of full-form prompts for inference. We believe that looking for the optimal form of textual inputs is important, and we will explore it for future work.

H. More failure cases

This is supplementary to Section 5.5 "**limitations**". In Figure S9, we show more failure cases of the proposed model. These results are in line with our conclusion that our method sometimes fails to synthesize counterfactual scenes. This limitation can possibly be alleviated in our future work, by 1) leveraging more powerful pre-trained text-to-image models, and 2) investigating better ways to retain the generative capability of the pre-trained model, perhaps by prompting techniques.

I. Results on rectangular datasets

The pre-trained Stable Diffusion that we leverage is designed to generate square (512×512) images. To verify the validity of the proposed method on rectangular datasets, we train our model on Cityscapes [3]. We resize all images of Cityscapes to 512×512 during training and resize the synthesized results back to the original size in testing phase. As shown in Figure S10, our method yields visually pleasing results.

J. Societal impact

Our method allows the users to generate diverse images using text and layout. This ability may be maliciously used for content, which incurs potential negative social impacts such as the spread of fake news and invasion of privacy. To mitigate them, powerful deepfake detection methods that automatically distinguish deepfake images from real ones are needed.

References

[1] Stephan Alaniz, Thomas Hummel, and Zeynep Akata. Semantic image synthesis with semantically coupled vq-model. *arXiv preprint arXiv:2209.02536*, 2022. 2

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR*, pages 1209–1218, 2018. 1

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 3

[4] Xun Huang, Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Multimodal conditional image synthesis with product-of-experts gans. In *ECCV*, pages 91–109, 2022. 2

[5] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, et al. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In *NeurIPS*, 2019. 1, 2

[6] 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. *arXiv preprint arXiv:2112.10741*, 2021. 2

[7] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, pages 2337–2346, 2019. 1, 2

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2

[9] Vadim Sushko, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. *arXiv preprint arXiv:2012.04781*, 2020. 1, 2

[10] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022. 1, 2

[11] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, pages 8798–8807, 2018. 2

[12] Yi Wang, Lu Qi, Ying-Cong Chen, Xiangyu Zhang, and Jiaya Jia. Image synthesis via semantic composition. In *ICCV*, pages 13749–13758, 2021. 1, 2

[13] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 2

[14] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017. 1

[15] Junchen Zhu, Lianli Gao, Jingkuan Song, Yuan-Fang Li, Feng Zheng, Xuelong Li, and Heng Tao Shen. Label-guided generative adversarial network for realistic image synthesis. *TPAMI*, 2022. 2

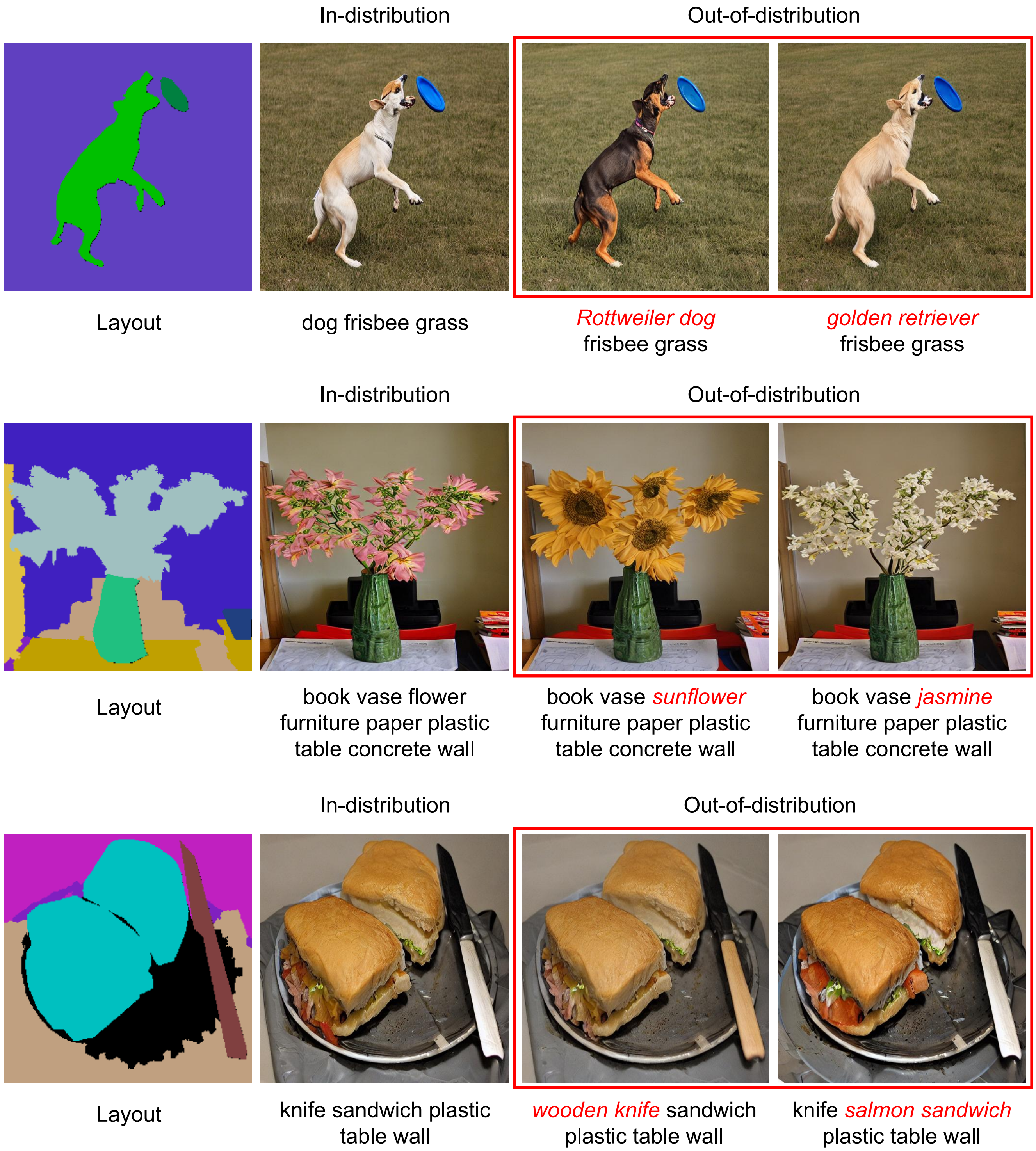

Figure S2. Supplementary to Figure 4. Our FreestyleNet is able to bind new attributes to the objects.

Figure S3. Supplementary to Figure 4. Our FreestyleNet is able to specify the styles for the synthesized images.

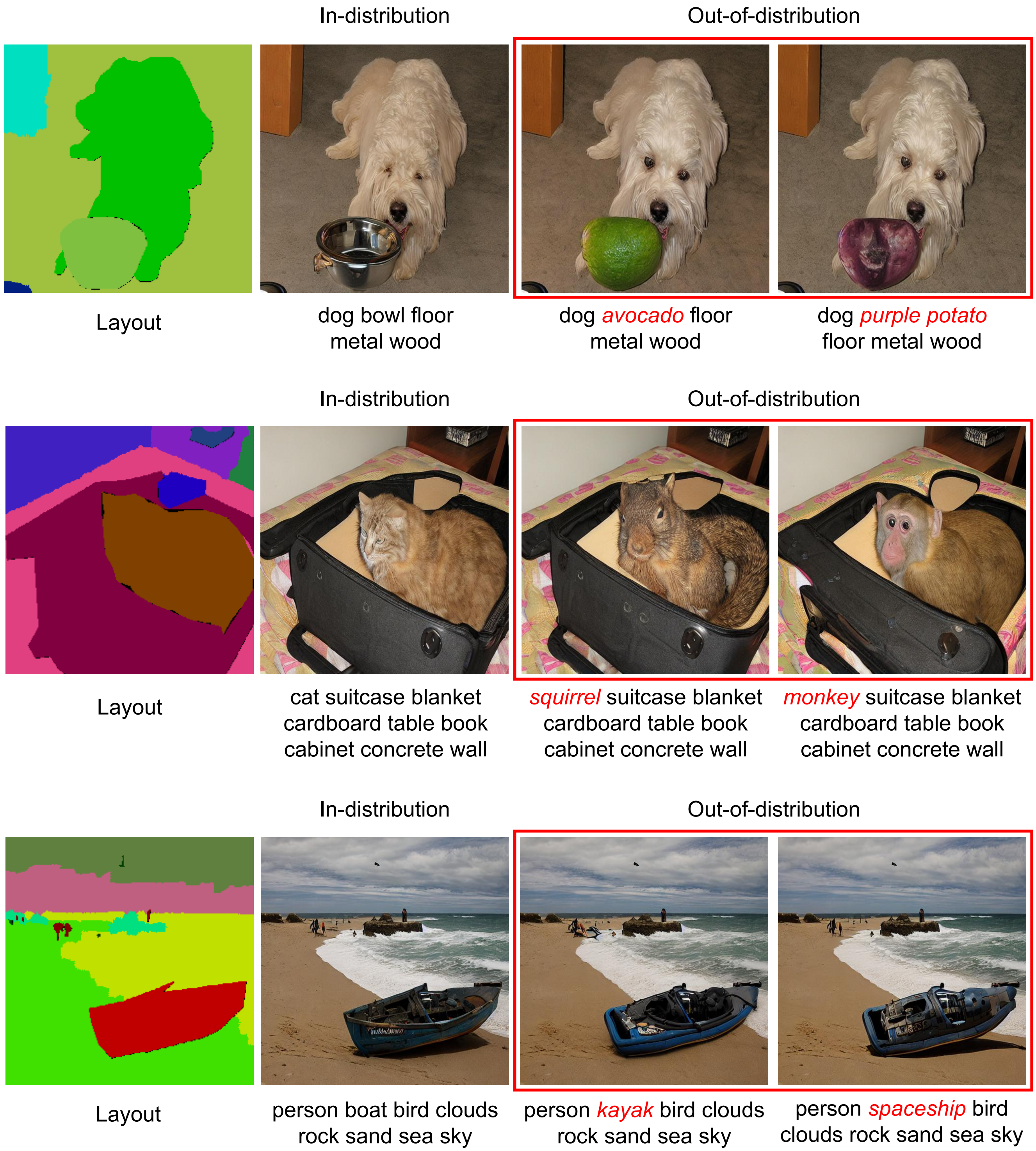

Figure S4. Supplementary to Figure 4. Our FreestyleNet is able to generate unseen objects.

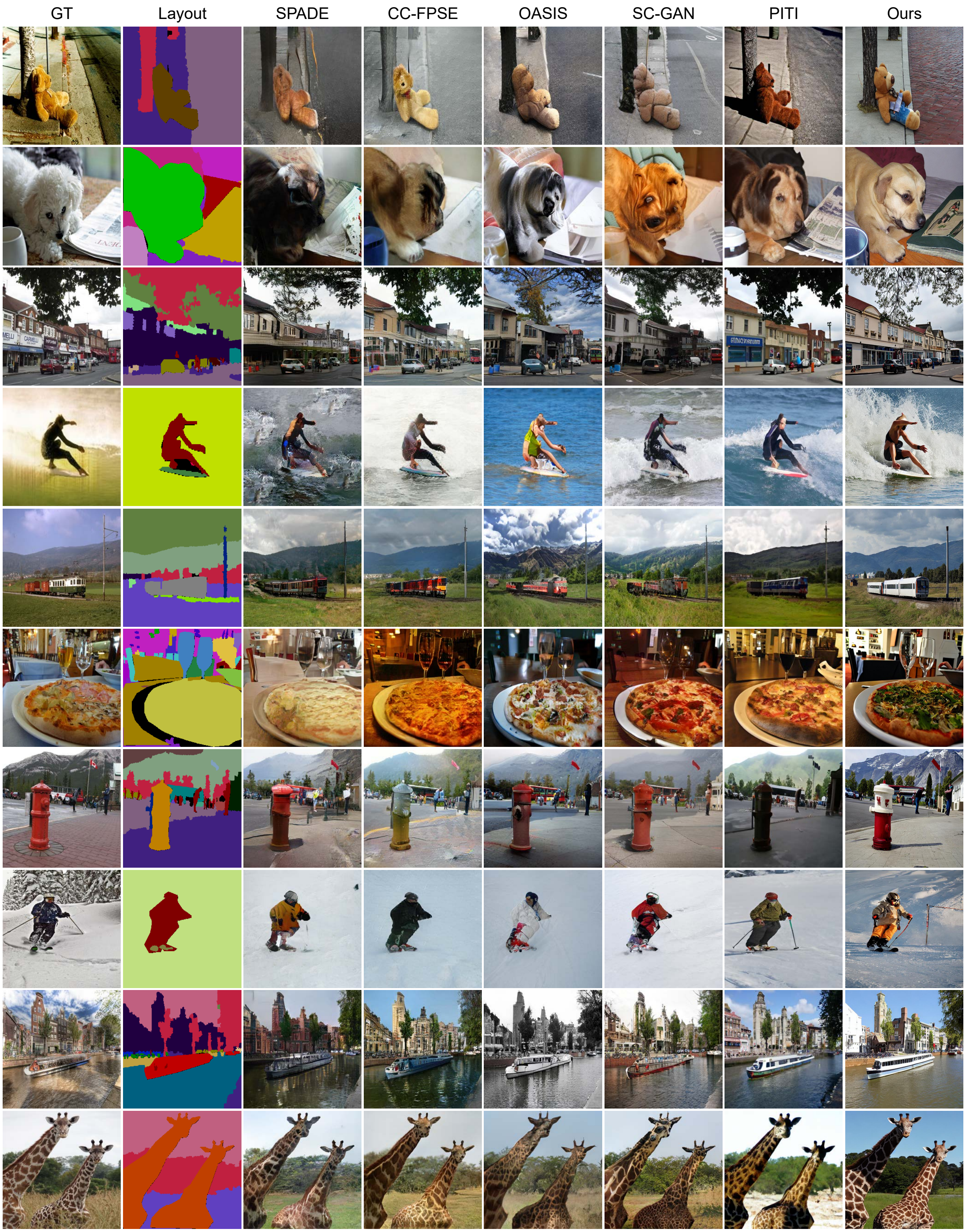

Figure S5. Supplementary to Figure 5. Visual comparison results with LIS baselines on COCO-Stuff.

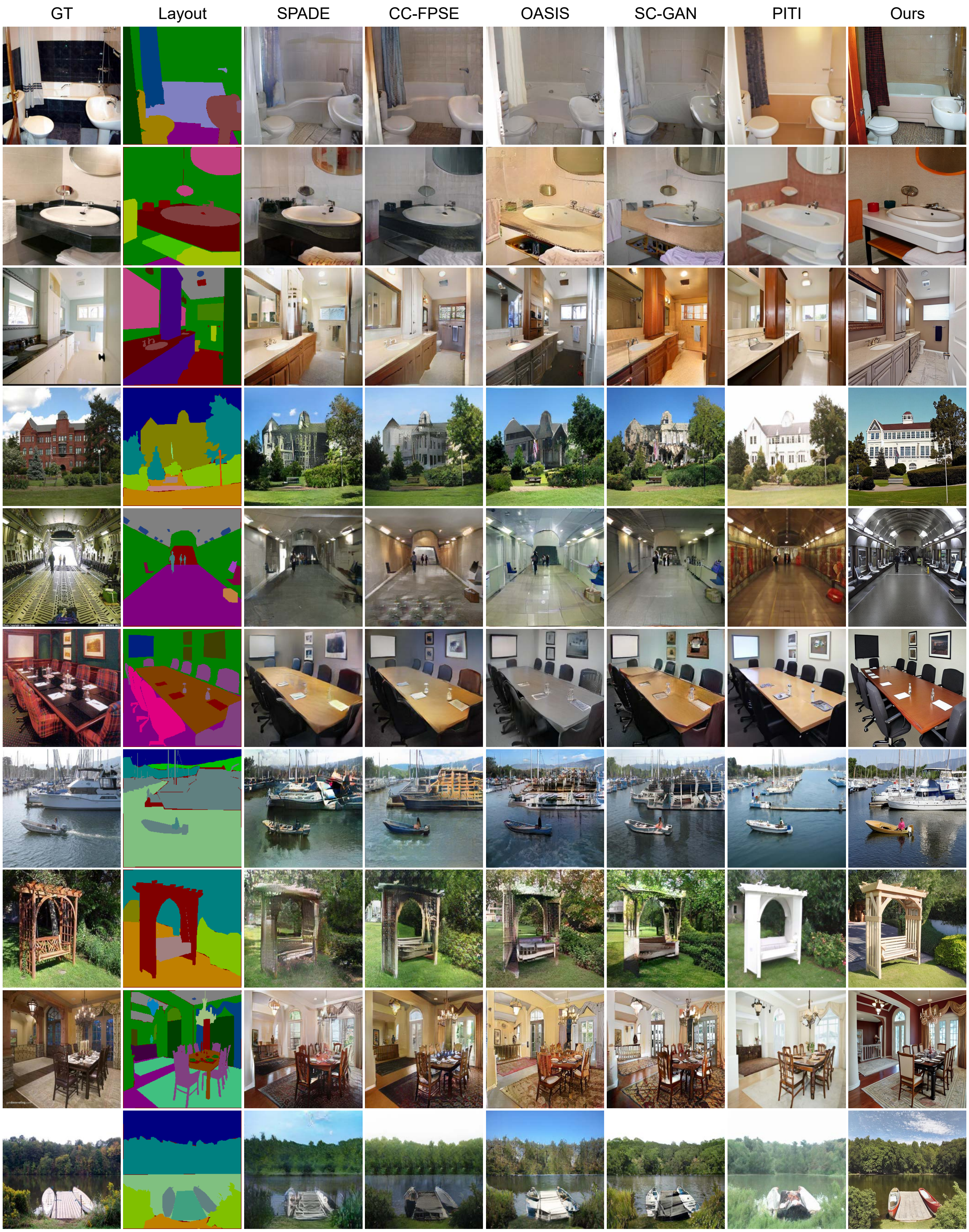

Figure S6. Supplementary to Figure 5. Visual comparison results with LIS baselines on ADE20K.

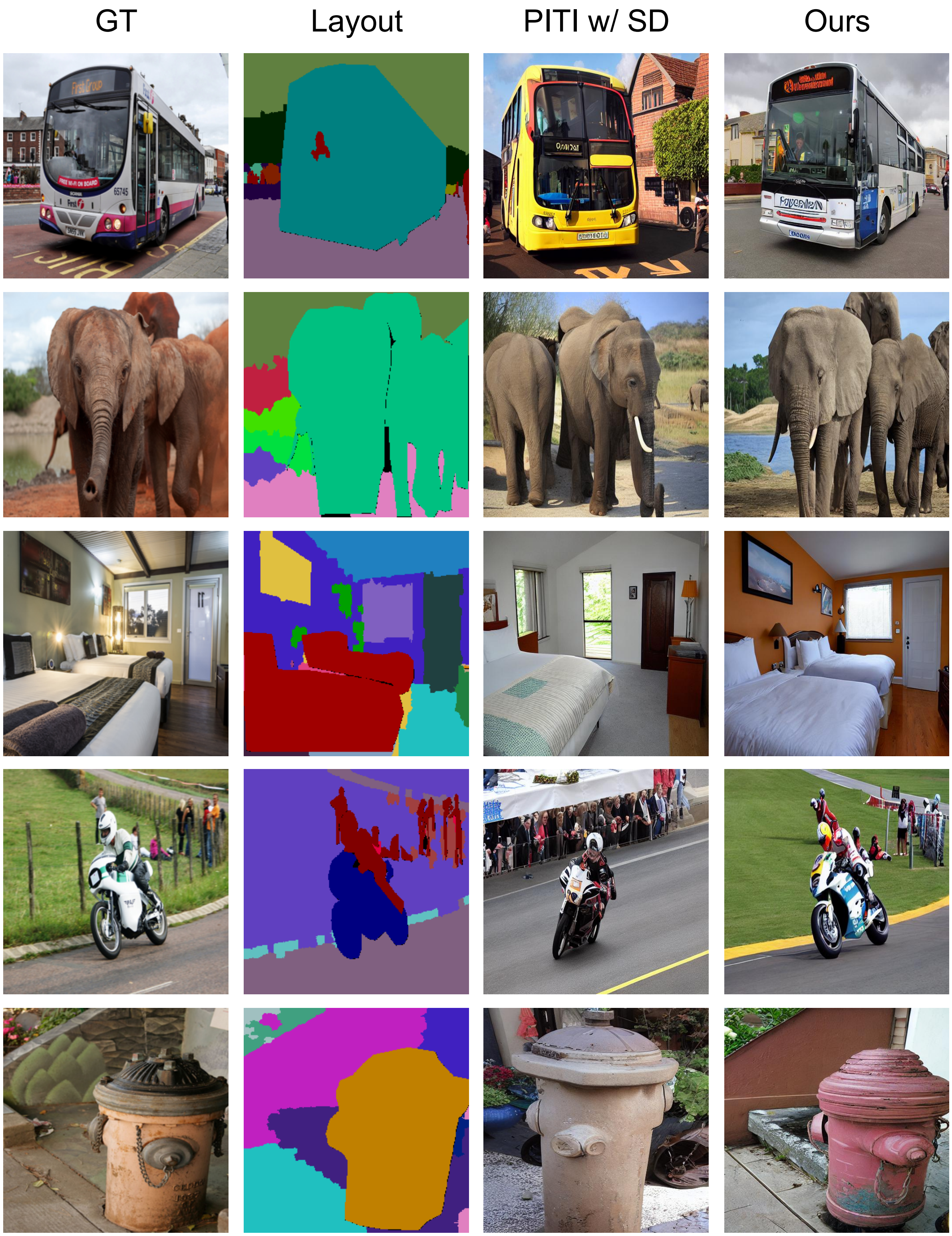

Figure S7. PITI w/ SD represents the PITI method whose diffusion model (GLIDE) is replaced by Stable Diffusion.

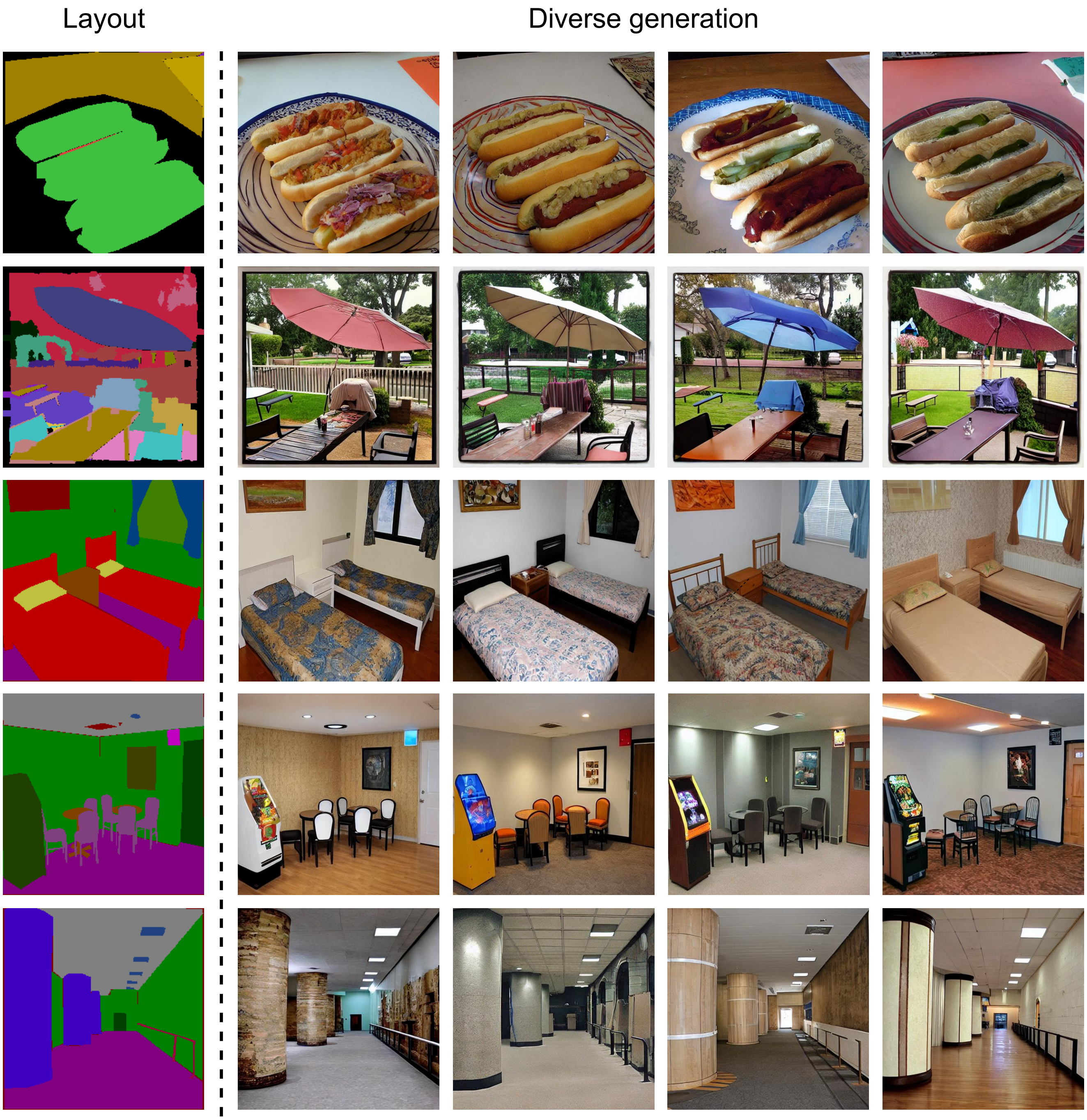

Figure S8. Diverse generation results of our FreestyleNet.

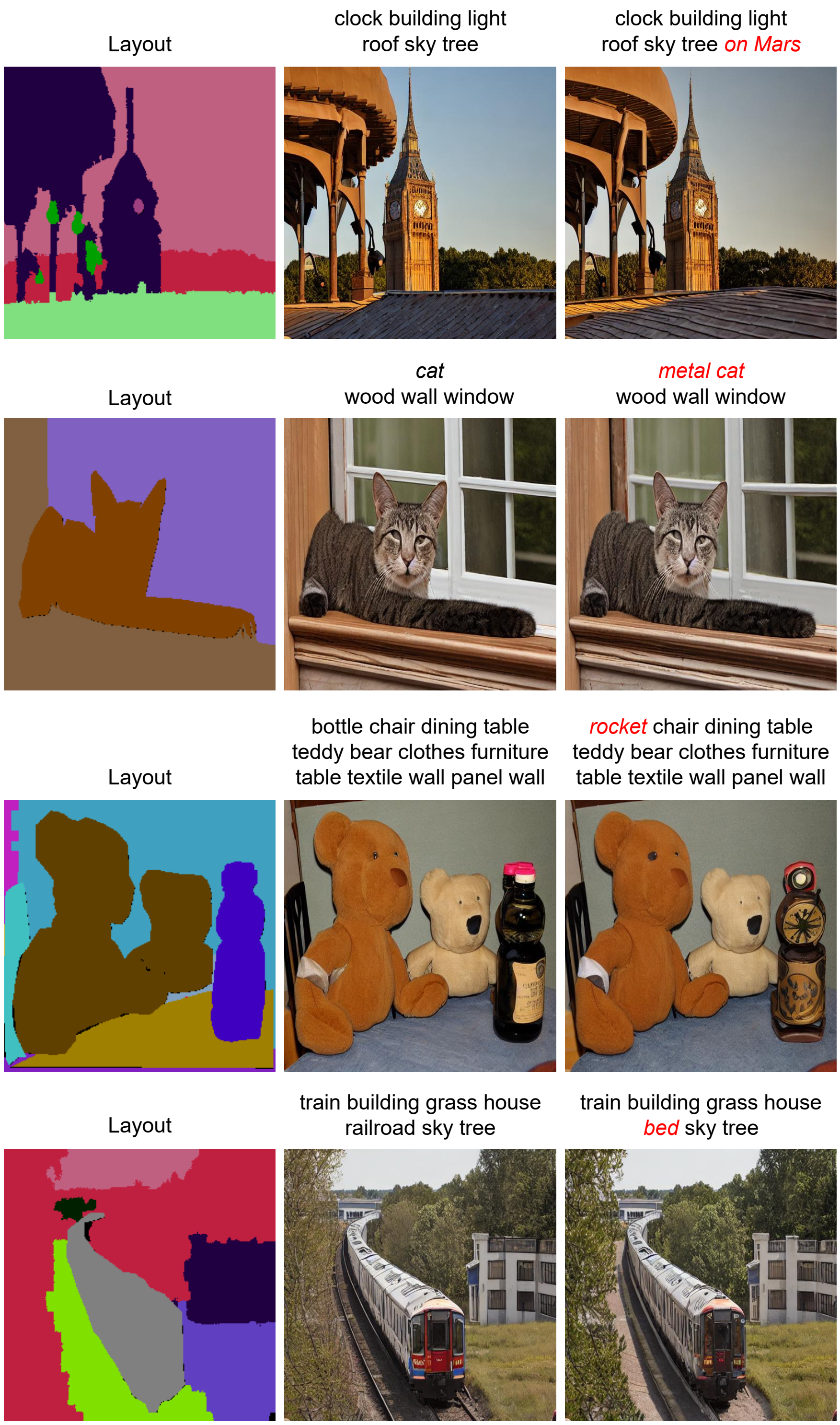

Figure S9. Supplementary to Figure 7. Failure cases. It is difficult for our FreestyleNet to generate some rare semantics or unreasonable scenes.

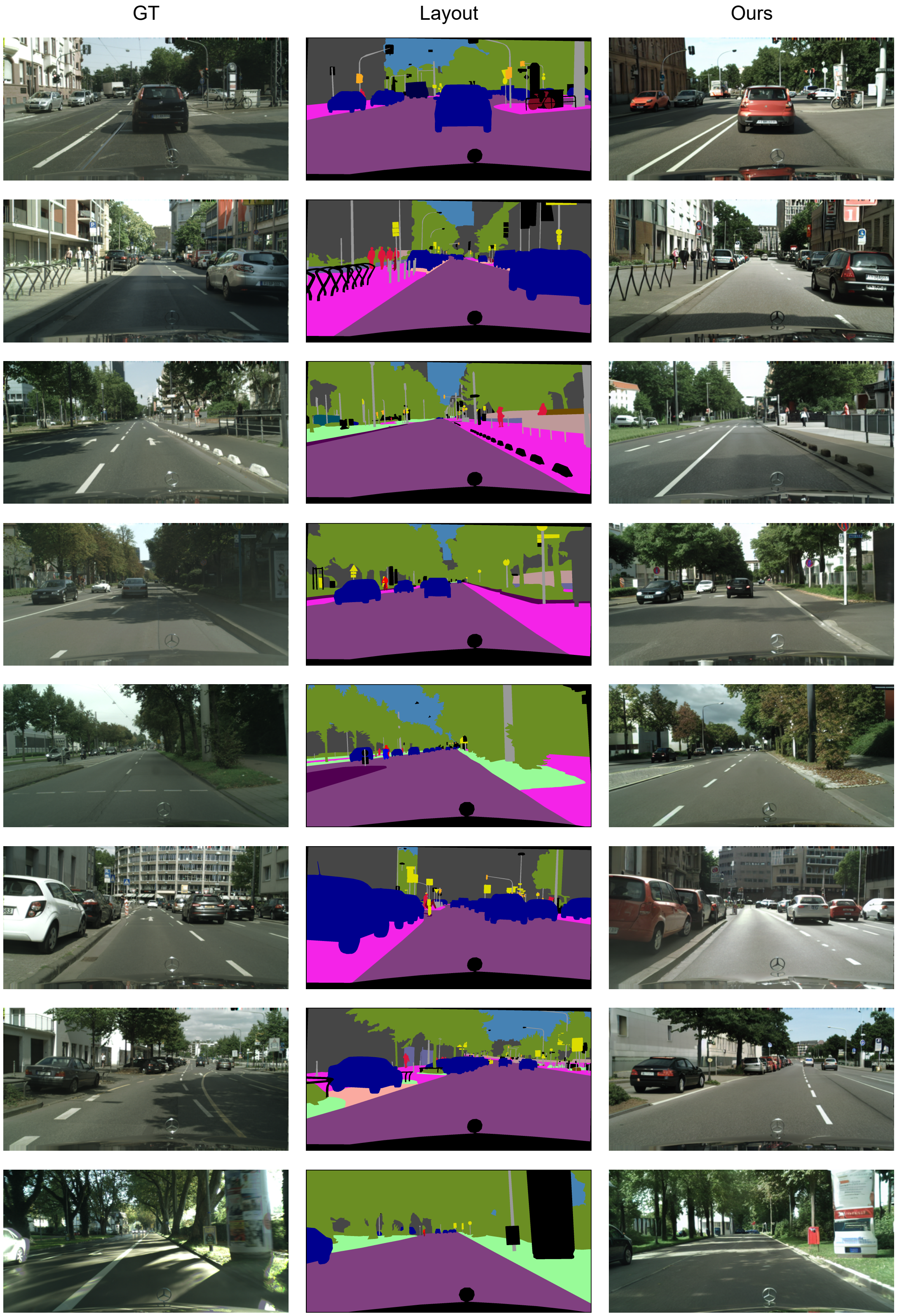

Figure S10. Generation results of our FreestyleNet on Cityscapes.