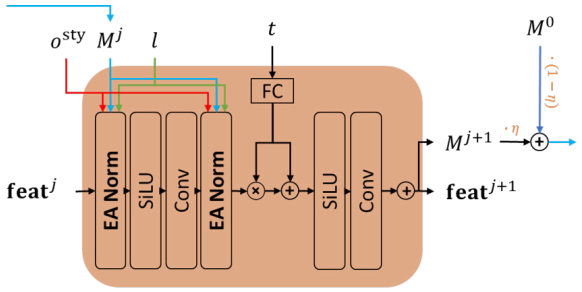
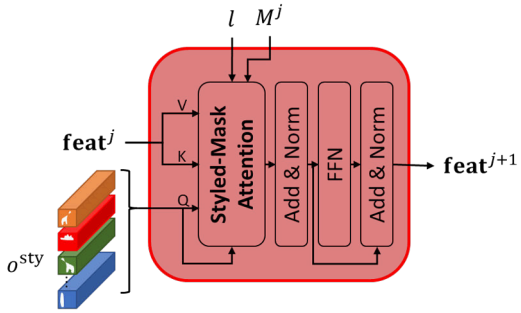
A. Implementation and Evaluation Details

A.1. Design of STAY Diffusion's Network Blocks

We illustrate the design of the Layout Diffusion Resblock and the Diffusion StyledMaskAttnBlock in Fig. 8. In the Layout Diffusion Resblock (see Fig. 8(a)), we replace the Group Normalization [48] layers with novel Edge-Aware Normalization (EA Norm) layers, and the updated object masks M^{j+1} are predicted at the end of the block. Before passing to the next block, we follow the design in [41] to further refine M_i^{j+1} with $M_i^{j+1} = (1-\eta) \cdot M_i^0 + \eta \cdot M_i^{j+1}$, where η is a learnable weight and i, j are the object and ResBlock indices, respectively. Note that the M^j used for the current Resblock is the weighted sum of the initial object masks M^0 and the predicted masks M^j from the previous Resblock. As for the Diffusion Styled-MaskAttnBlock, we replace the original self-attention by the proposed Styled-Mask Attention (SM Attention) layer as shown in Fig. 8(b).



(a)



(b)

Figure 8. The design of (a) the Layout Diffusion Resblock and (b) the Diffusion StyledMaskAttnBlock, respectively.

A.2. Evaluation Images of Compared Methods

For layout-to-image (L2I) synthesis methods, we acquired images of LostGAN v2 [41], PLGAN [44], LAMA [21], TwFA [50], LayoutDiffuse [6], and LayoutDiffusion (LayoutDM) [56] by sampling from the checkpoints released by the authors. Since there is no publicly available COCO-stuff checkpoint for GLIGEN [20], we followed the setting described by the authors to train a model on COCO-stuff based on LDM [34]. We then sampled from this model, following the instructions of the authors such as using scheduled sampling. Our evaluation results on the sampled images (cf. Tab. 1 and Tab. 6) closely matched the reported numbers in GLIGEN (cf. Tab. 2[1] in their main paper). We also provide a quick comparison in Tab. 4. As for Visual Genome (VG), we sampled from the checkpoint provided by the authors of GLIGEN.

Table 4. The reported FID and YoloScore (AP) in Tab. 2 of GLIGEN [20] and our reproduced results (marked with *).

Methods	FID ↓	Yolo ↑
GLIGEN [20]	21.04	22.4
GLIGEN [20]*	21.30	23.0

For semantic image synthesis benchmark methods, we selected ControlNet-v1.1 [52], FreestyleNet [49], PITI [45] and SDM [46] for evaluation. We used the checkpoints released by the authors and gave the models different sets of semantic maps (i.e., the ground truth map (**GT**), the self-supervised map with overlapping objects from Tab. 2(a) (**Base**) and the self-supervised map from our full model (**Ours**)) to generate images for evaluation.

A.3. Training and Sampling Details for STAY Diffusion

We reported the used hyperparameters of STAY Diffusion for training and sampling in Tab. 5. For models trained at resolution 256 × 256, we used four Tesla A100 for training. For models trained at resolution 128 × 128, we used four Tesla V100 for training. Finally, a Tesla V100 was used to sample images from both resolutions.

B. Interactivity of STAY Diffusion

We demonstrate the interactivity of STAY Diffusion in Fig. 9. Although imperfect, the self-supervised maps generated by STAY Diffusion can help reduce human effort for image labeling or provide more comprehensive information for downstream tasks such as image blending. As shown in Fig. 9(a), the mask extracted from STAY Diffusion is more accurately aligned with the object shape than the one drawn from a raw bounding box. Furthermore, in Fig. 9(b), we gradually added an object to the layout to demonstrate that STAY Diffusion can be easily adjusted to reconfigurations.

[1]The authors of GLIGEN refer to COCO-stuff as COCO2017D.

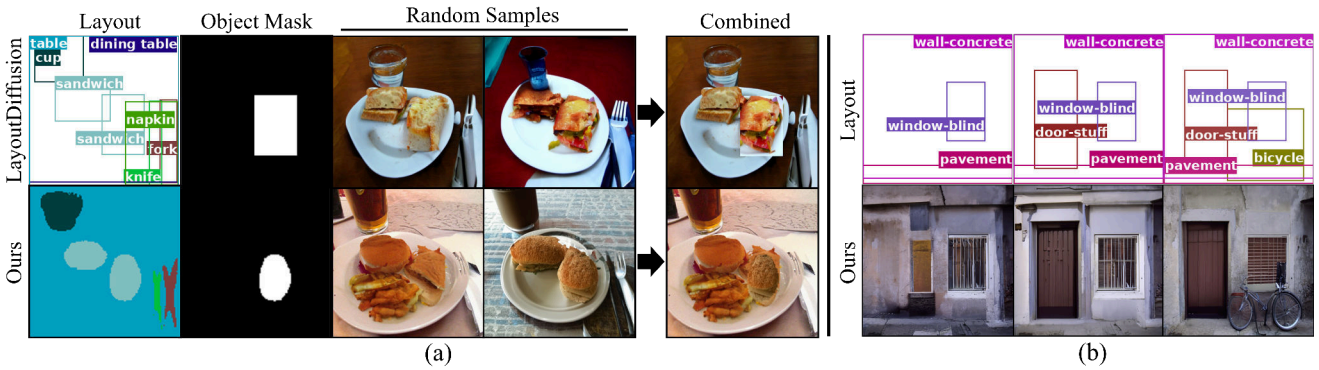


Figure 9. The interactivity of STAY Diffusion. (a) With the self-supervised semantic maps, STAY Diffusion provides more accurate object location for tasks like image blending. (b) STAY Diffusion can adapt to reconfigured layouts.

C. Additional Results

C.1. Full Quantitative Results

We reported the full quantitative results on COCO-stuff in Tab. 6 and Visual Genome (VG) in Tab. 7. Note that the YOLOScore is only applicable for COCO-stuff as defined in [21]. As shown in both tables, our STAY Diffusion presents superior performance in image diversity, generation accuracy, and controllability. As for image quality, our method shows comparable results to the previous state-of-the-art (SOTA) in FID and IS.

C.2. More Visualization Results

We show more visual comparison to previous methods on COCO-stuff in Fig. 10 and Fig. 11, and VG in Fig. 12 and Fig. 13. Additionally, we provided more demonstration of style variations on object appearance in Fig. 14, the learned self-supervised semantic maps and generated images based on the same layout in Fig. 15, and more visual comparisons between our ablated models in Fig. 16.

C.3. Additional Results for Mask Clarity

We present the full table of the quantitative results on mask clarity in Tab. 8. As for the visual comparison, we present the generated images from different semantic image synthesis and our STAY Diffusion in Fig. 17. It can be clearly seen that the self-supervised maps from our full model (**Ours**) produce images with recognizable objects and higher quality compared to the self-supervised maps from the **Base** (i.e., Tab. 2(a)). This again highlights the importance of the mask clarity. Moreover, it is evident that the semantic image synthesis methods highly rely on precisely labeled maps. When those are not available, their generation quality drops severely. This observation is also aligned with the quantitative results in Tab. 8. On the other hand, our STAY Diffusion still produces images with superior quality in this case due to the proposed EA Norm and SM Attention in Sec. 3.

Table 5. The used hyperparameters for the proposed STAY Diffusion in Sec. 4 experiments.

Dataset	COCO-stuff 256×256	COCO-stuff 128×128	VG 256×256
Model	STAY Diffusion	STAY Diffusion	STAY Diffusion
Layout-Conditional Diffusion Model			
In Channels	3	3	3
Hidden Channels	256	128	256
Channel Multiply	1,1,2,2,4,4	1,1,2,3,4	1,1,2,2,4,4
Number of Residual Blocks	2	2	2
Dropout	0	0	0
Diffusion Steps	1000	1000	1000
Noise Schedule	linear	linear	linear
λ	0.001	0.001	0.001
Object Representation			
Class Embedding Dimension	180	180	180
Style Embedding Dimension	128	128	128
Maximum Number of Objects	8	8	8
Maximum Number of Class Id	184	184	179
Edge-Aware Normalization Module			
α	0.5	0.5	0.5
Styled-Mask Attention Module			
Attention Method	Styled-Mask	Styled-Mask	Styled-Mask
Number of Head Channels	64	64	64
Training Hyperparameters			
Total Batch Size	32	32	32
Number of GPUs	4	4	4
Learning Rate	1e-4	1e-4	1e-4
Mixed Precision Training	No	No	No
Weight Decay	0	0	0
EMA Rate	0.9999	0.9999	0.9999
Iterations	1.25M	600K	1.45M
Sampling Hyperparameters			
Total Batch Size	4	8	4
Number of GPUs	1	1	1
Classifier-free Guidance s	1.5	1.5	1.0
Use DPM-Solver	True	True	True
DPM-Solver Algorithm	dpmsolver++	dpmsolver++	dpmsolver++
DPM-Solver Type	dpmsolver	dpmsolver	dpmsolver
DPM-Solver Skip Type	time_uniform	time_uniform	time_uniform
DPM-Solver Step Method	singlestep	singlestep	singlestep
DPM-Solver ODE Order	3	3	2
DPM-Solver Timesteps	50	50	50

Table 6. Quantitative results on COCO-stuff at resolution 256 × 256. The proposed STAY Diffusion outperforms LayoutDiffusion (LayoutDM) in diversity, accuracy and controllability metrics while maintaining close quality performance.

Methods	Coco-Stuff				
	FID ↓	IS ↑	DS ↑	CAS ↑	Yolo ↑
LostGAN v2 [41]	33.17	18.08±0.46	0.55±0.10	33.17	15.0
PLGAN [44]	30.67	18.92±0.65	0.52±0.10	29.15	13.6
LAMA [21]	33.00	19.77±0.66	0.48±0.12	9.97	20.4
TwFA [50]	23.78	23.02±0.94	0.43±0.13	20.09	23.9
LayoutDiffuse [6]	22.41	27.09±0.07	0.58±0.11	31.80	23.7
GLIGEN [20]	21.30	**27.71±0.79**	0.57±0.09	34.41	23.0
LayoutDM [56]	**15.74**	26.01±0.84	0.58±0.09	35.69	27.2
Ours	17.43	26.08±0.76	**0.59±0.09**	**37.18**	**29.5**

Table 7. Quantitative results on Visual Genome (VG) at resolution 256 × 256. The proposed STAY Diffusion outperforms LayoutDiffusion (LayoutDM) in diversity, accuracy and controllability metrics while maintaining close quality performance.

Methods	VG			
	FID ↓	IS ↑	DS ↑	CAS ↑
LostGAN v2 [41]	34.92	14.01±0.81	0.53±0.01	24.40
PLGAN [44]	-	-	-	-
LAMA [21]	38.51	13.70±0.76	0.54±0.10	24.16
TwFA [50]	18.57	17.75±0.68	0.50±0.10	18.30
LayoutDiffuse [6]	22.45	**22.89±1.69**	0.56±0.10	25.05
GLIGEN [20]	23.42	21.84±1.38	0.60±0.09	25.49
LayoutDM [56]	**15.26**	21.94±1.28	0.61±0.10	26.84
Ours	18.02	18.56±0.91	**0.65±0.08**	**27.23**

Table 8. Quantitative results of mask clarity on COCO-stuff. When GT maps are absent, the images generated from the self-supervised maps of our full model (**Ours**) outperform other baseline methods in quality and controllability metrics, highlighting the importance of clear masks.

Methods	Mask	FID ↓	IS ↑	DS ↑	CAS ↑	Yolo ↑
ControlNet-v1.1 [52]	GT	32.24	24.80±1.44	0.54±0.08	22.43	25.6
FreestyleNet [49]	GT	14.80	30.06±0.92	0.50±0.09	36.67	42.9
PITI [45]	GT	15.22	28.08±1.11	0.45±0.12	37.20	34.8
SDM [46]	GT	20.79	23.82±0.53	0.65±0.18	36.42	26.9
ControlNet-v1.1 [52]	Base	64.29	17.47±0.49	0.59±0.09	9.36	4.1
FreestyleNet [49]	Base	37.21	20.40±0.73	0.50±0.09	21.54	12.9
PITI [45]	Base	61.01	12.90±0.39	0.46±0.12	14.02	4.7
SDM [46]	Base	40.08	15.41±0.26	0.69±0.18	19.29	10.5
ControlNet-v1.1 [52]	Ours	50.25	20.65±0.93	0.60±0.09	15.38	6.5
FreestyleNet [49]	Ours	32.95	22.59±0.66	0.56±0.10	26.52	13.8
PITI [45]	Ours	37.41	17.24±0.91	0.53±0.13	18.63	12.0
SDM [46]	Ours	35.53	17.70±0.72	**0.69±0.19**	22.69	13.5
Ours	Ours	**17.43**	**26.08±0.76**	0.59±0.09	**37.18**	**29.5**

Figure 10. More comparison with previous methods on COCO-stuff 256 $\times$ 256. Zoom in for better view.

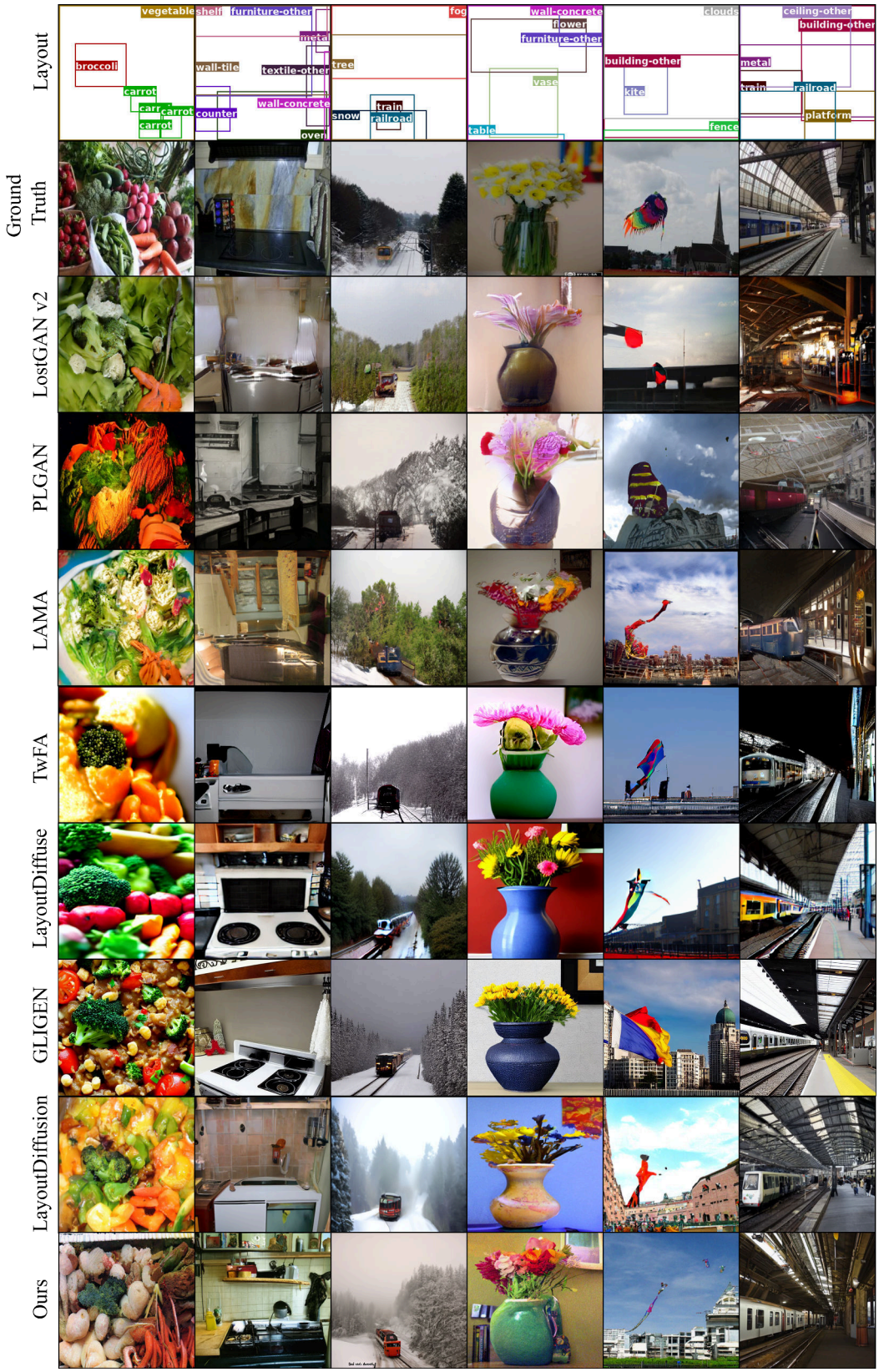


Figure 11. More comparison with previous methods on COCO-stuff 256 $\times$ 256. Zoom in for better view.

Figure 12. Visual comparison with previous methods on VG 256 × 256.

Figure 13. Visual comparison with previous methods on VG 256 × 256.

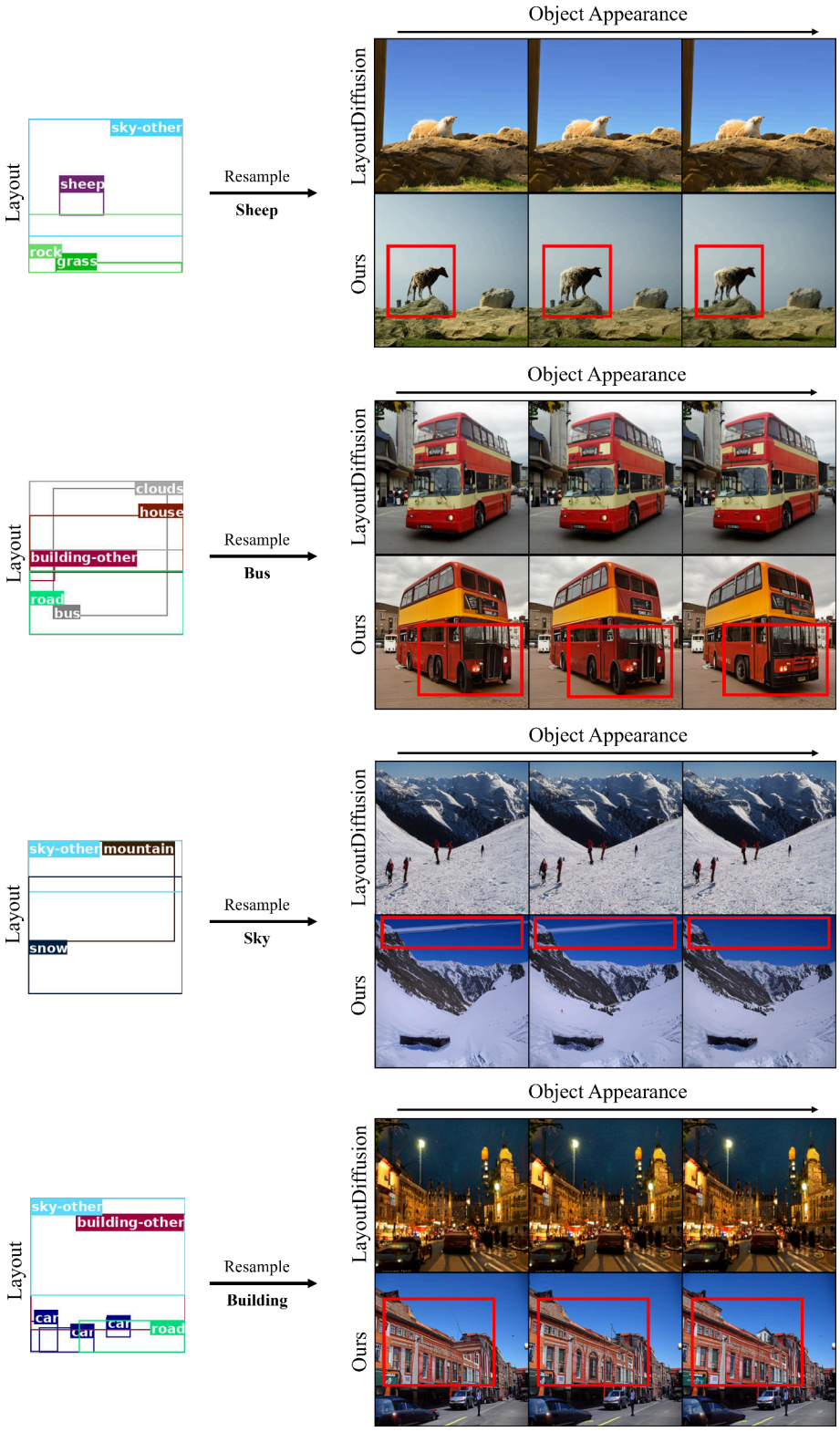


Figure 14. The demonstration of fine-grained style variations offered by our STAY Diffusion. Note that only one object is resampled in each image (Zoom in for better view).

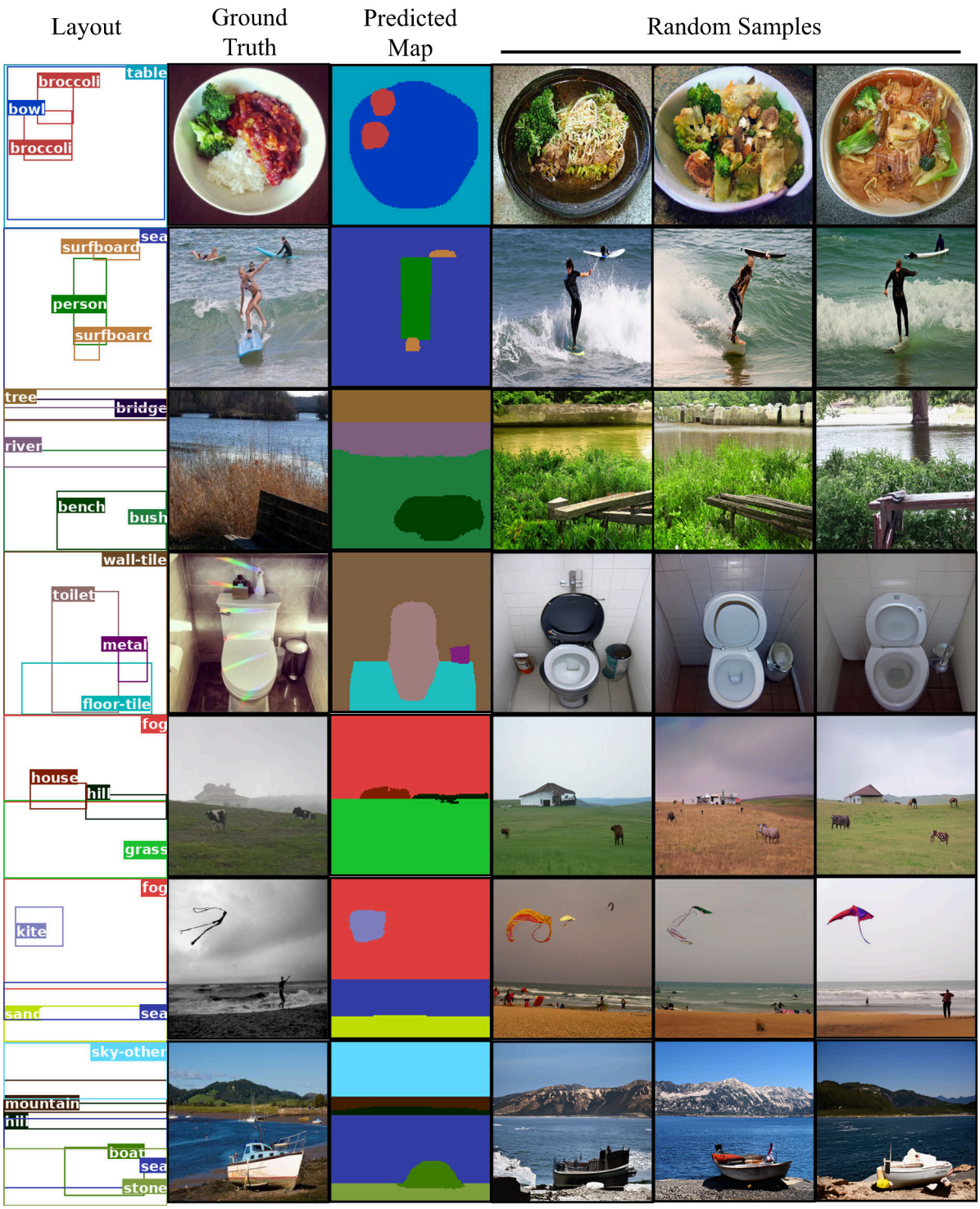


Figure 15. The demonstration of self-supervised semantic maps learned by our STAY Diffusion.

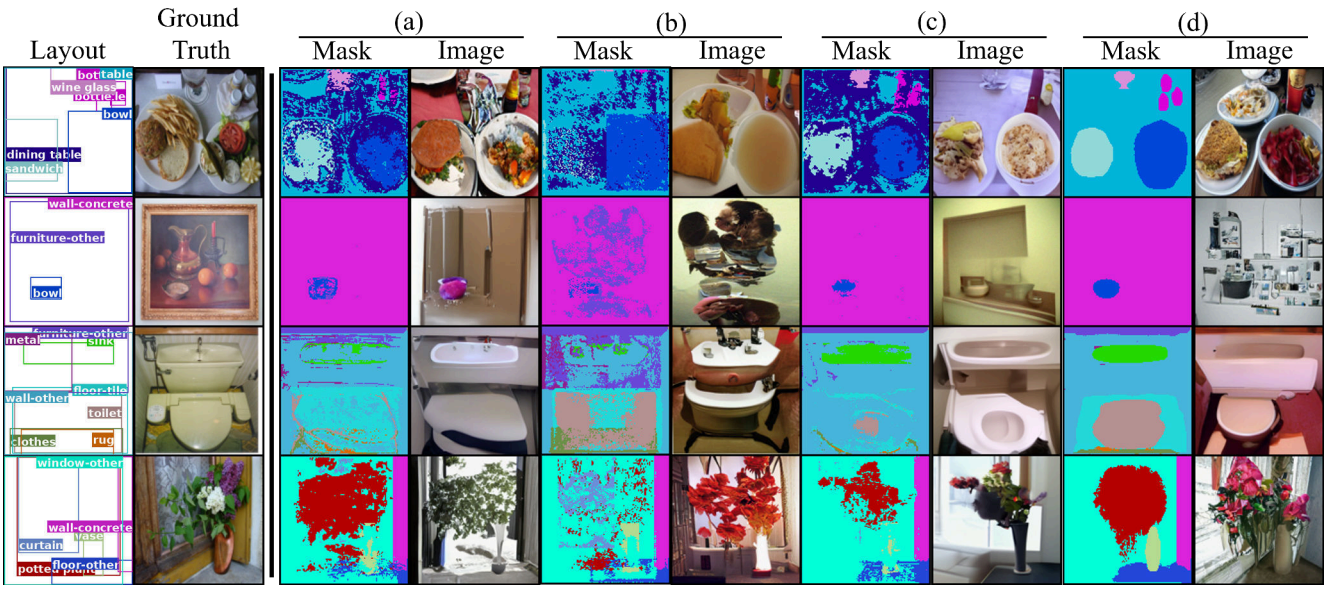


Figure 16. The generated images and their predicted self-supervised semantic maps from ablated models mentioned in Sec. 4.5. (a) ISLA Norm + Self Attention. (b) EA Norm + Self Attention. (c) ISLA Norm + SM Attention. (d) EA Norm + SM Attention.

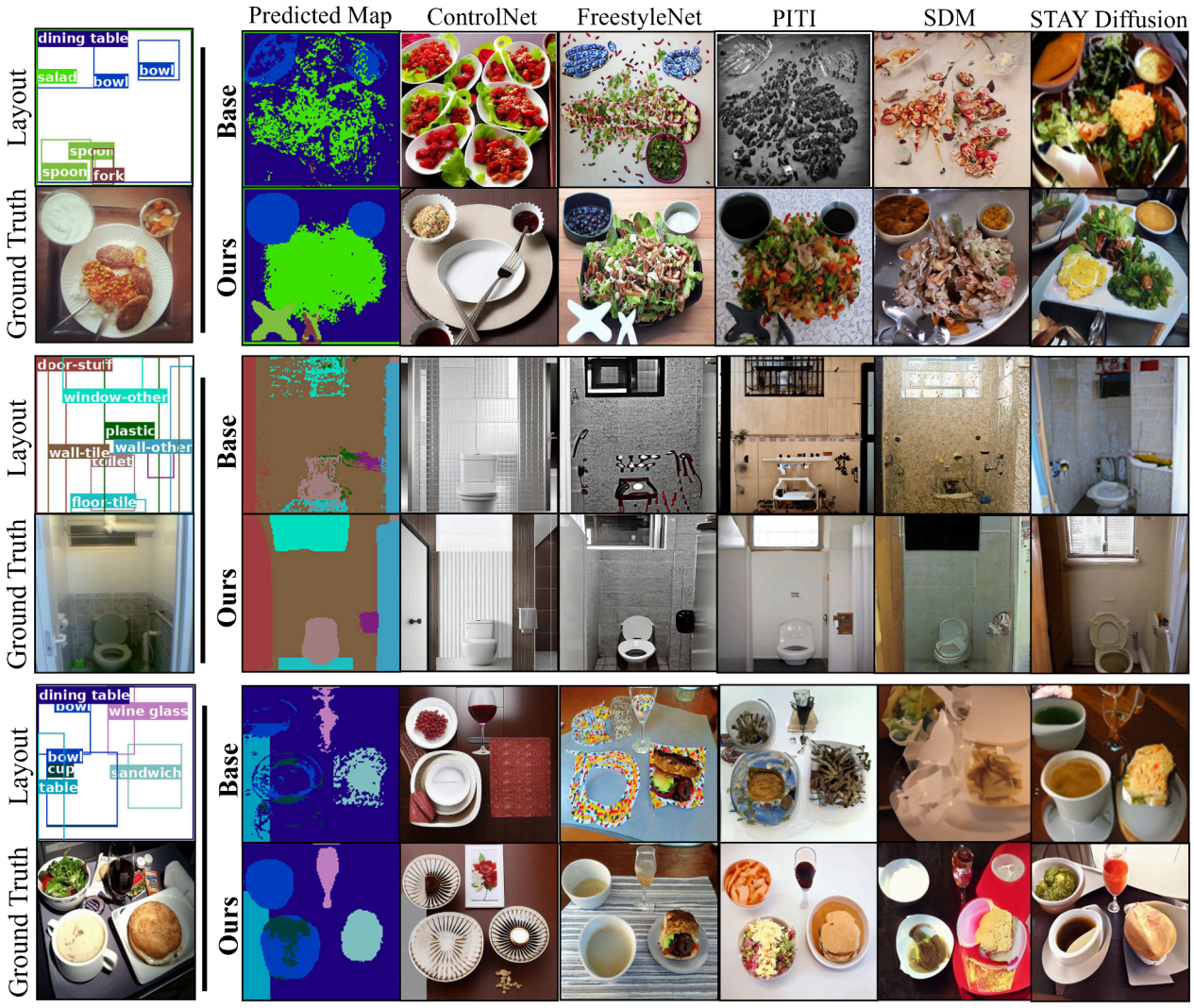


Figure 17. The visual comparison of images generated from different semantic image synthesis methods and our STAY Diffusion. Note that **Base** indicates the baseline self-supervised map from Tab. 2 setting (a) and **Ours** indicates the self-supervised map from the full model (i.e., Tab. 2 setting (d)).