LAW-Diffusion: Complex Scene Generation by Diffusion with Layouts

Binbin Yang[1] Yi Luo[1] Ziliang Chen[2] Guangrun Wang[3] Xiaodan Liang[1] Liang lin[1*]
[1]Sun Yat-Sen University [2]Jinan University [3]University of Oxford
yangbb3@mail2.sysu.edu.cn, luoy97@mail3.sysu.edu.cn, c.ziliang@yahoo.com,
{wanggrun, xdliang328}@gmail.com, linliang@ieee.org

1. Further Implementation Details

Our experiments are conducted on 8 GPUs and implemented by PyTorch [11]. Following [6, 2], we set $T = 1000$ and the diffusion process with linearly decreasing $\{\alpha_t\}$ from $\alpha_1 = 1 - 10^{-4}$ to $\alpha_T = 0.98$. Different from previous text-to-image practices [10, 12, 15] which use pre-trained linguistic models to obtain text embeddings, LAW-Diffusion is **trained from scratch** by jointly optimizing the spatial dependency parser that generates the layout embedding $\mathcal{L}$, and the noise estimator $\tilde{\epsilon}_\theta(x_t, t|\mathcal{L})$ using the VLB loss defined in Eq. (6) of our paper. We use the same diffusion training strategies and U-Net architectures as ADM [2]. As for the generation of the layout embedding $\mathcal{L}$, we set the dimension of class embedding as $d_c = 32$ and the patch size of region fragments as $P = 8$. Then a two-layer MHSA with 8 attention heads and a learnable aggregation token $v_{[\text{Agg}]} \in \mathbb{R}^{P\times P\times d_c}$, is implemented as the fragment aggregation function in Eq. (8-11) of our paper.

Following [7, 15], we implement the conditional model $\epsilon_\theta(x_t, t|\mathcal{L})$ and unconditional model $\epsilon_\theta(x_t, t|\emptyset)$ in Eq. (12) as a single conditional model with 10% probability of replacing the conditional input $\mathcal{L}$ by a learnable null embedding $\emptyset$. Since the computational overhead is quadratic to the size of input image, directly training the 256×256 diffusion model in the pixel space would be costly expensive. Following [3, 14], we employ VQ-VAE [13] to downsample the 256×256 images to 64×64 low-dimensional latent representations. Therefore, our LAW-Diffusion is performed in the original pixel space for the sizes of 64×64 and 128×128, while it is trained in the compressed latent space for the size of 256×256. The ultimate 256×256 synthesized images are decoded from the denoised 64×64 latent codes using the decoder of VQVAE. In this way, we also demonstrate that LAW-Diffusion is a general and flexible model, which is effective for the generation in both pixel space and compressed latent space. For the hyper-parameters of our adaptive guidance in Eq. (15), we choose $\omega_{\max} = 3$ and $\omega_{\min} = 1$ and use the cosine-form annealing function.

*Corresponding author.

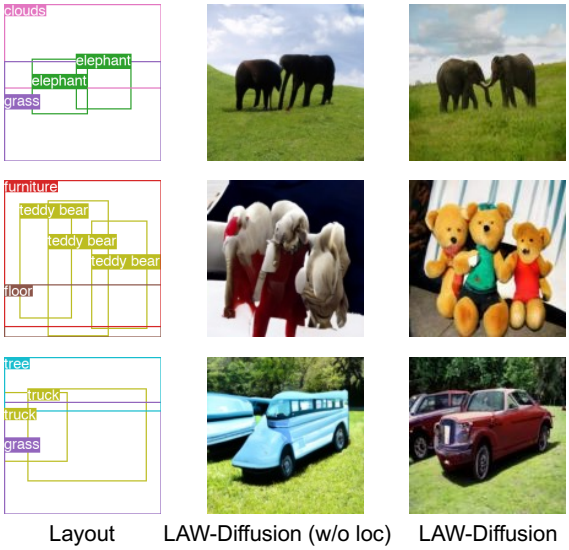

Figure 1. Comparison of the generated images between LAW-Diffusion (w/o loc) (baseline) and LAW-Diffusion. The images produced by LAW-Diffusion (w/o loc) exhibit noticeable spurious artifacts and distortions when multiple objects are overlapped and occluded. By contrast, LAW-Diffusion generates objects with photorealistic textures and coherent contextual relationships. Best viewed in color.

As for our proposed new evaluation metric, *i.e.*, Scene Relation Score, we use VCTree-EBM-Predcls[16] pre-trained on Visual Genome [8] from https://github.com/mods333/energy-based-scene-graph as the scene graph generator to measure whether the correct relations are captured by our image generator. We report the mean Recall@K(mR@K) given by VCTree-EBM-Predcls as our Scene Relation Score (SRS).

2. Further Ablation Study

In Sec 4.4 of our paper, we have introduced our baseline diffusion model, namely LAW-Diffusion (w/o loc). It trivially extends ADM [2] for the task of L2I using a class-aware attention mechanism that has been widely employed in prior works [5, 18]. The Tab. 3 in our paper pro-

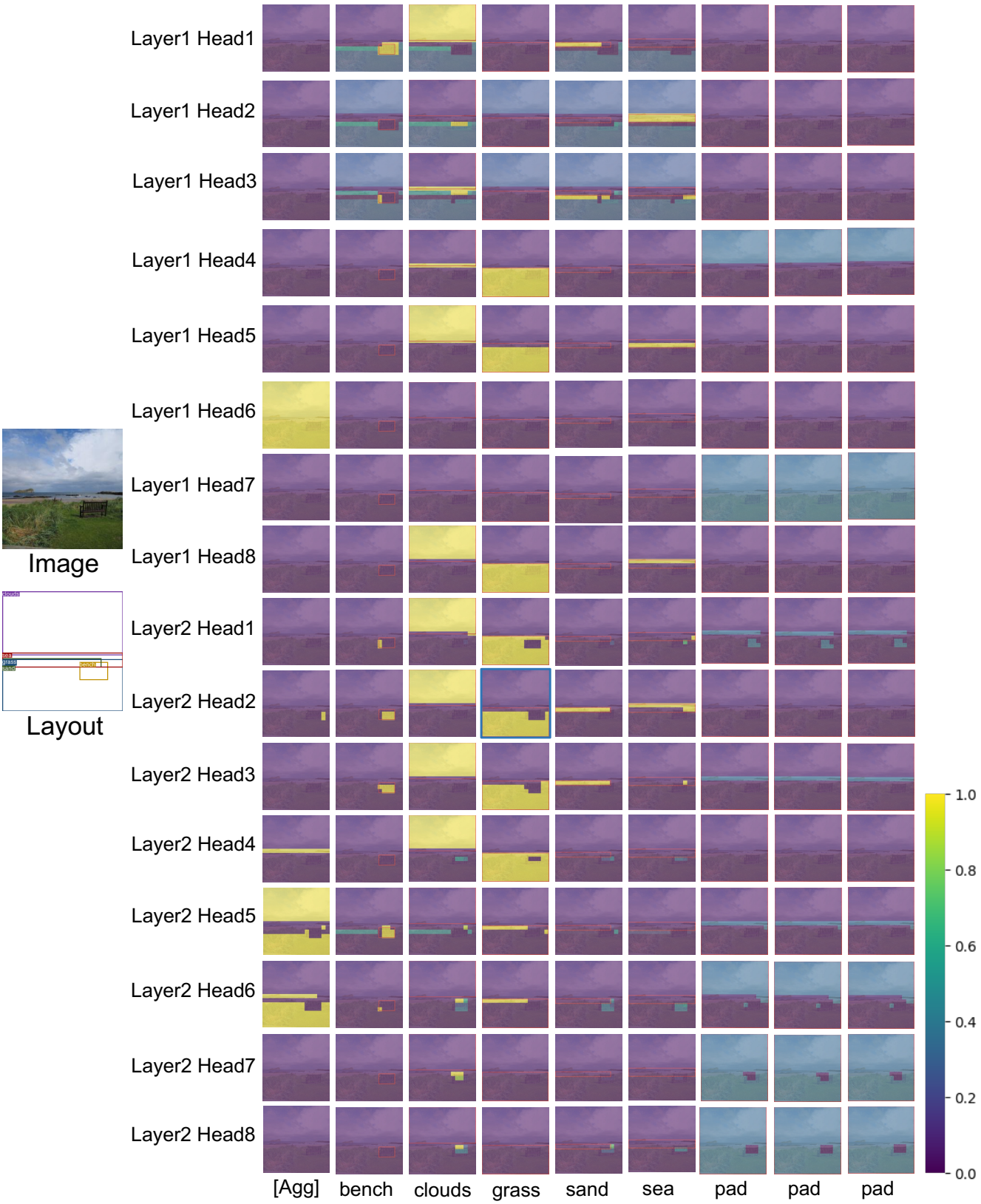

Figure 2. Visualization of the attention maps of our location-aware cross-object attention. Each column indicates the attention maps from the aggregation tokens to the fragments of different object region maps. Each row shows the attention maps of an attention head in specific layer. The red frames in attention maps represent the bounding boxes of objects. Best viewed in color.

vides a quantitative comparison between the baseline LAW-Diffusion (w/o loc) and our LAW-Diffusion, verifying the effectiveness of our location-aware cross-object attention. Here we provide a qualitative comparison between these two models (128×128), which both utilize the adaptive guidance strategy, *i.e.*, $\omega_t : 3 \searrow 1$. From Fig. 1, we observe that the images generated by LAW-Diffusion (w/o loc) exhibit noticeable blurry artifacts and distortions in the case where multiple objects are overlapped and occluded. By comparison, LAW-Diffusion is able to parse the spatial dependencies among instances that co-occur at the same position, which enables it to synthesize instances in the scene with clear shape and textures. As illustrated in the second row of Fig. 1, the teddy bears generated by the baseline LAW-Diffusion (w/o loc) are blended together with severe distortion. In contrast, LAW-Diffusion can generate identifiable teddy bears with photorealistic textures and coherent spatial relationships, thus demonstrating the effectiveness of our location-aware cross-object attention.

3. Visualization of Cross-object Attention

To provide more insights into how LAW-Diffusion captures spatial properties using our cross-object attention mechanism, we visualize the attention maps of the location-aware cross-object attention in Fig. 2. As mentioned previously, our cross-object attention module is implemented using a segment-level two-layer eight-head self-attention mechanism, as described in Eq. (8-11) of our paper. For each attention head, we compute the attention scores between the aggregation token $v_{[\text{Agg}]}$ and the object region fragments $\{v_i^j\}_{i=1}^{N_{\max}}$ at the same location (the j^{th} fragment). Then we collect the attention scores of different positions and rearrange them to attention maps, which is shown in rows in Fig. 2. By comparing the bright areas with high attention activation within each object's attention map to their respective bounding boxes (shown as red frames here) in Fig. 2, we can conclude that our location-aware cross-object attention mechanism effectively integrates regional information from different objects, including their spatial occlusion relationships. For example, by observing the attention map (with blue frame) of "grass" of layer2 head 2 in Fig. 2, we find LAW-Diffusion definitely perceives that a bench is on the grass and the sand is partially occluded by the grass. Exploiting such spatial dependencies benefits coherently generating these co-existing objects.

4. Details about Layout-aware Latent Grafting

As the supplement to Sec 3.4 of our paper, Algorithm 1 summarizes the instance reconfiguration process using our layout-aware latent grafting strategy. In the cases of adding and removing objects, the reconfigured layout Γ^* is distinct from the source layout configuration Γ. When restyling an existing object in the source image x_0, the reconfigured layout Γ^* remains identical to Γ ($\mathcal{L}^*$ is also identical to $\mathcal{L}$ because the spatial dependencies are invariant) while the local restyling is determined by the re-initialization of x_T^* relative to the source noise x_T. The input source latents $\{x_t\}_{t=1}^T$ in Algorithm 1 can be obtained either by fetching them from the latent memory bank of the previous generation process guided by $\mathcal{L}$, or by approximating them using Eq. (3) of our paper, since the previously generated x_0 is known. It is worthy noting that, the reconfigured latent x_t^* is guided by a holistic semantics from L^* rather than the local manipulation within b^* which solely encodes the object-level information. Hence, our innovative design offers a simple yet effective way to preserve global coherence in instance reconfiguration. More examples of reconfiguration are exhibited in Fig. 3.

Algorithm 1 Layout-aware Latent Grafting Strategy

Input: a source image x_0 generated by LAW-Diffusion from a layout configuration Γ, with its layout embedding $\mathcal{L}$ and latents $\{x_t\}_{t=1}^T$; the learned layout-aware generation process of LAW-Diffusion $\{p_\theta(x_{t-1}|x_t, \mathcal{L})\}_{t=T}^1$; a reconfigured layout Γ^* where an object o^* within bounding box b^* is added/removed/restyled relative to Γ; a reconfiguration mask M indicating the rectangular region within b^* for object o^*; the new layout embedding $\mathcal{L}^*$ corresponding to Γ^*.
Output: a reconfigured image x_0^* where object o^* within b^* is added/removed/restyled relative to x_0 while the other contents in x_0 are preserved.
$x_T^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$;
for all t from T to 1 **do**
  $\hat{x}_t^* \leftarrow x_t^* \odot M \oplus x_t \odot (1 - M)$;
  $x_{t-1}^* \sim p_\theta(x_{t-1}^*|\hat{x}_t^*, \mathcal{L}^*)$;
end for
return x_0^*

5. Detailed Object-level Control

As a supplement, our LAW-Diffusion exhibits the ability of controlling the objects' attributes when sufficient fine-grained annotations are provided. For example, we select about 1K images from COCO-Stuff and manually annotate 11 colors for the objects in the images. Then we append a color-embedding block to the object class-embedding block of LAW-Diffusion, and tune it with our labels. As shown in Fig. 4, LAW-Diffusion is empowered to generate/restyle images with controlled colors. It shows that our L2I model's potential to control object's detailed attributes if auxiliary supervisions are provided.

6. Diverse Generation

As shown in Fig. 5, LAW-Diffusion can generate diverse images with completely different styles while ensuring faithful adherence to the input layouts in all synthesized

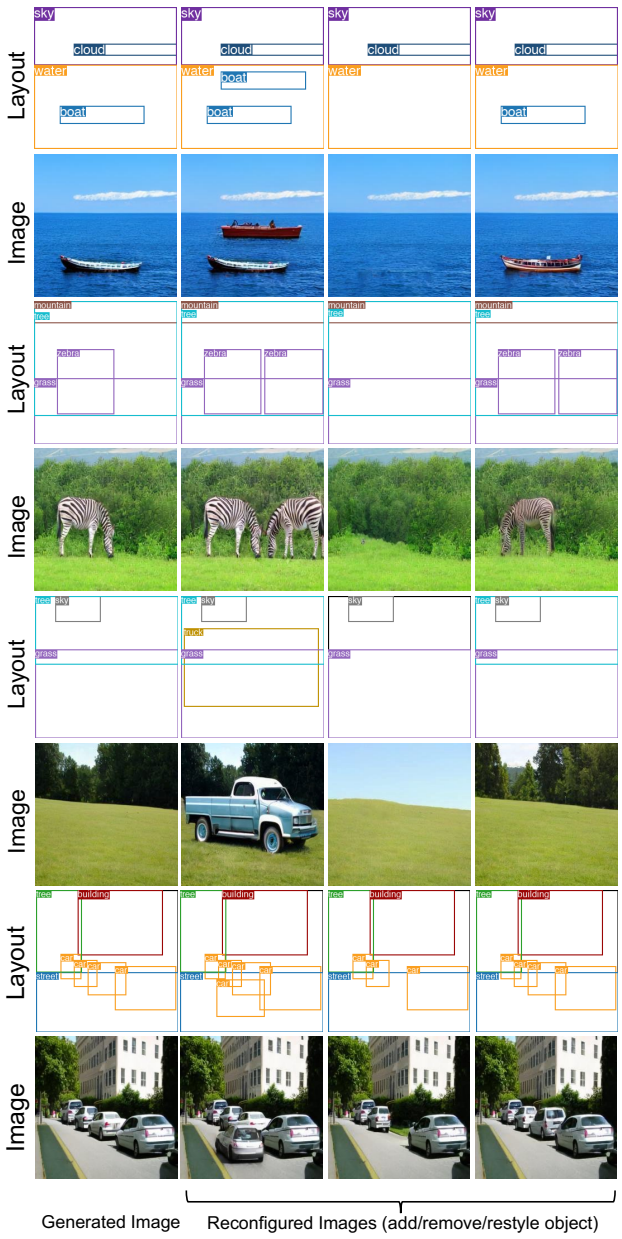

Figure 3. More examples of instance reconfiguration generated by LAW-Diffusion using our layout-aware latent grafting strategy.

scenes.

7. Human Evaluation

To perform a human evaluation and further assess the fidelity and coherence of the generated images, we conducted a user study. Specifically, each participant is randomly assigned 20 groups of 256×256 images synthesized by LostGAN-V2 [17], LAMA [9], Frido [4], TwFA [18], and our LAW-Diffusion. Each group of images was generated from the same layout by different methods. For example, four groups of sample images are presented in Fig. 6.

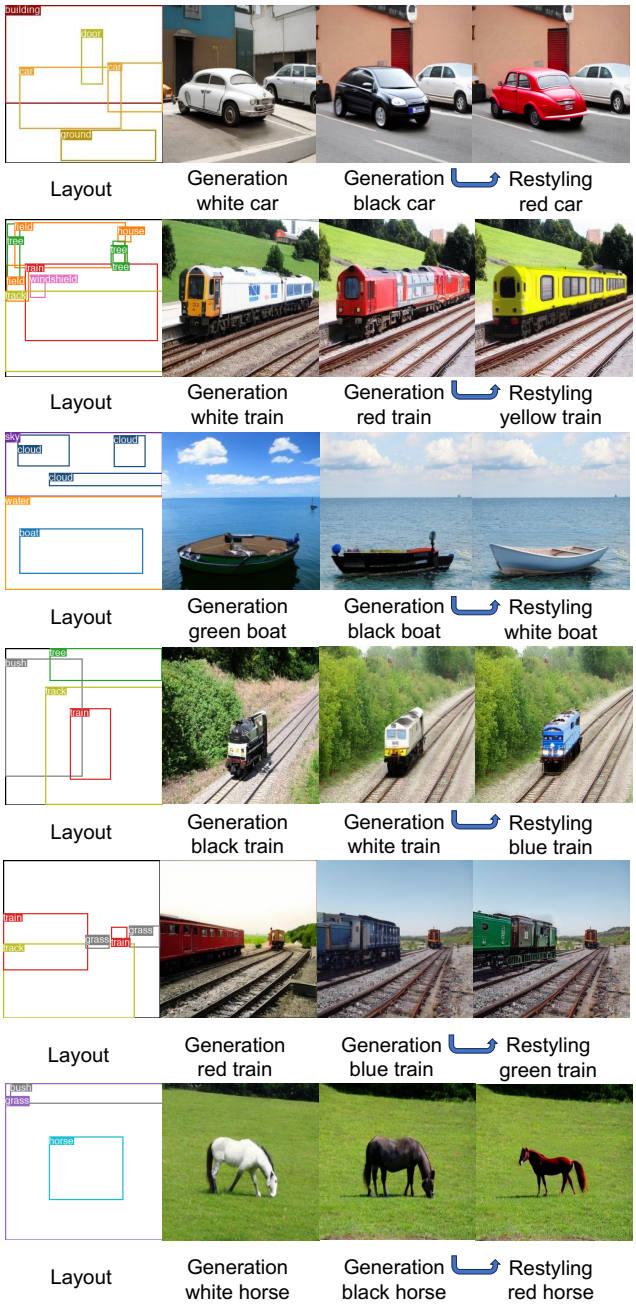

Figure 4. Examples of controlling objects' colors.

For each layout, a participant was asked to select his/her most preferred image based on the photorealism and its adherence to the layout. Additionally, the participant is also asked to give a rating of 1 to 5 for the coherence and harmony of relations among contextual objects in each generated image.(1→5:worst→best)

The results of human evaluation from 539 users are summarized in Fig. 7 and Tab. 1. As shown in Fig. 7, the comparison of preference percentages demonstrates that the scene images generated by LAW-Diffusion were preferred

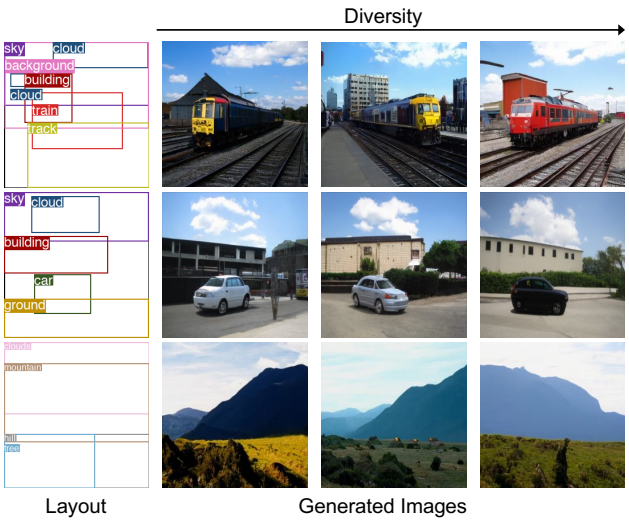

Figure 5. Illustration of diverse generation. Each row shows the diverse images synthesized from the same layout on the left.

Figure 6. Sample images provided to the participants of the human evaluation.

more by the human evaluators. The results of relation coherence rating presented in Tab. 1 provide additional evidence that LAW-Diffusion is able to produces scenes with more coherent object relationships.

Methods	SRS(mR@20)	Percentage(relation coherence rating) 1	2	3	4	5	Average Rating
LostGAN-V2 [17]	0.1241	26.7%	24.3 %	27.3%	15.2%	6.5%	2.50
LAMA [9]	0.1260	20.5%	26.6%	25.5%	17.5%	9.9%	2.70
Frido [4]	0.1375	8.3%	15.5%	34.3%	20.4%	21.5%	3.31
TwFA [18]	0.1407	4.7%	9.4%	28.9%	32.8%	24.2%	3.62
LAW-Diffusion	0.1485	2.1%	6.5%	16.9%	38.3%	36.2%	4.00

Table 1. Results of relation coherence rating for different methods.

8. The effectiveness of Scene Relation Score

To verify the effectiveness of our propose new metric, the Scene Relation Score (SRS), we examine its relationship with the existing metrics FID/IS and the human rating

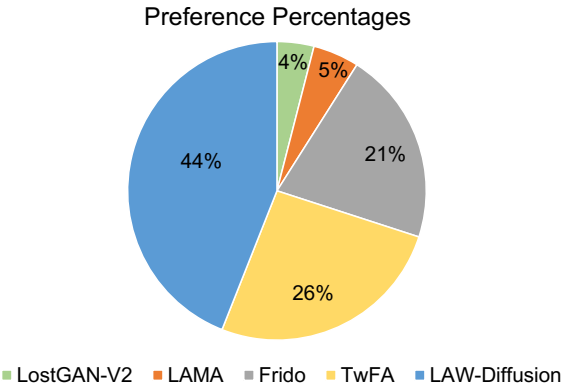

Figure 7. The preference percentages for different methods.

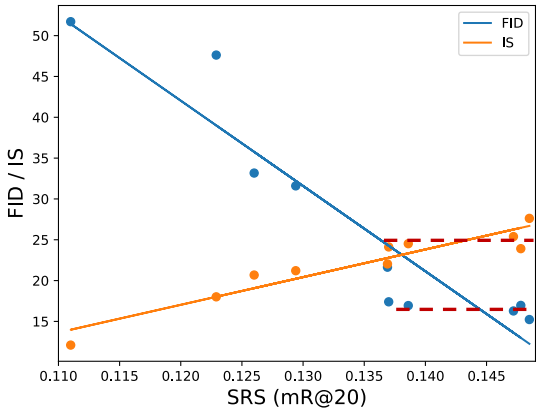

Figure 8. The relationship between SRS and FID/IS.

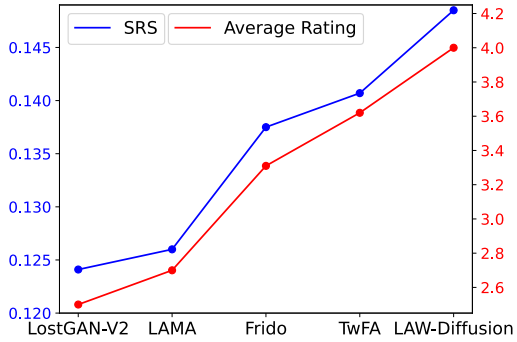

Figure 9. The relationship between SRS and human rating.

results given in Tab. 1.

We examine the relationship between SRS (in terms of mR@20) and FID/IS in Fig. 8. To perform this analysis, we randomly sampled 10 checkpoints of our model and evaluated their SRS, FID, and IS scores. First, the blue and orange solid lines generated by linear regression show that a higher SRS score often implies a better FID/IS score. Second, the red dotted lines demonstrate that two models may have different SRS scores even though they have similar FID/IS scores. Therefore, our proposed SRS metric is consistent with the widely used metrics such as FID and IS, while also providing a more comprehensive measure of a

model's generative performance. Moreover, the results in Fig. 9 reveal that SRS (in terms of mR@20) is consistent with the human ratings of scene relation coherence, providing further evidence of its effectiveness.

9. More Generated Samples

In this section, we show more generated samples of our LAW-Diffusion on COCO-Stuff [1] and VG [8]. Fig. 10 and Fig. 11 respectively show the results of 64×64 and 128×128. Additionally, Fig. 12 provides more 256×256 samples and demonstrates the effectiveness of accurate semantic alignment and high fidelity.

10. Discussions

Potential negative social impact While all benchmarks used in this paper are public and transparent, it is important to consider the negative impacts that may occur when our model is fine-tuned on socially biased datasets collected by other users. In this case, our generative model may produce undesired images that incorporate harmful social biases, which potentially leads to privacy concerns and issues of intellectual property infringement.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 6, 7, 8, 9

[2] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1

[3] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 1

[4] Wan-Cyuan Fan, Yen-Chun Chen, DongDong Chen, Yu Cheng, Lu Yuan, and Yu-Chiang Frank Wang. Frido: Feature pyramid diffusion for complex scene image synthesis. *arXiv preprint arXiv:2208.13753*, 2022. 4, 5

[5] Sen He, Wentong Liao, Michael Ying Yang, Yongxin Yang, Yi-Zhe Song, Bodo Rosenhahn, and Tao Xiang. Context-aware layout to image generation with enhanced object appearance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15049–15058, 2021. 1

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1

[7] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1

[8] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 1, 6, 7, 8, 9

[9] Zejian Li, Jingyu Wu, Immanuel Koh, Yongchuan Tang, and Lingyun Sun. Image synthesis from layout with locality-aware mask adaption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13819–13828, 2021. 4, 5

[10] 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. 1

[11] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NIPS Autodiff Workshop*, 2017. 1

[12] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[13] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 1

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1

[15] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1

[16] Mohammed Suhail, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gerard Medioni, and Leonid Sigal. Energy-based learning for scene graph generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13936–13945, 2021. 1

[17] Wei Sun and Tianfu Wu. Learning layout and style reconfigurable gans for controllable image synthesis. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5070–5087, 2021. 4, 5

[18] Zuopeng Yang, Daqing Liu, Chaoyue Wang, Jie Yang, and Dacheng Tao. Modeling image composition for complex scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7764–7773, 2022. 1, 4, 5

Figure 10. More generated 64×64 samples of LAW-Diffusion on COCO-Stuff [1] and VG [8]. Best viewed in color.

Figure 11. More generated 128×128 samples of LAW-Diffusion on COCO-Stuff [1] and VG [8]. Best viewed in color.

Figure 12. More generated 256 × 256 samples of LAW-Diffusion on COCO-Stuff [1] and VG [8] with the input layout on the left. Best viewed in color.