Training-Free Layout-to-Image Generation with Marginal Attention Constraints
Supplementary Material

Anonymous CVPR submission

Paper ID 11

Appendix A: Experimental Details

In this section, we report the model architectures and hyperparameters used in the experiments across all layout-to-image schemes. We employed the DDIM scheduler [2] with 50 denoising steps in all experiments, performing latent variable optimization only within the first 10 steps, with a maximum of 5 iterations per step. The step size for updating the latent variable z_t was set to 70, and the loss threshold value for early stopping was set to 10^{-6}. In all experiments, the weight for classifier-free guidance was set to 7.5. In the experiments with MAC, λ and α for combining boundary-attention loss and regularization loss were both set to 0.25.

Appendix B: Evaluation Details

We employed the state-of-the-art GroundDINO [1] to detect objects in the synthetic images generated with the input prompts and layout instruction from DrawBench and HRS benchmarks. Specifically, GroundDINO generates multiple predicted bounding boxes corresponding to predicted categories on the synthetic images with threshold value for confidence 0.25. Those boxes are then used to compute various metrics for measuring the spatial controllability and counting accuracy.

Object Counting

To evaluate the layout-to-image schemes in object counting, we record the number of predicted bounding boxes $n_{\text{pred}}^{(i)}$ corresponding to the phrase $\mathbf{p}^{(i)}$, and compute the correct number of boxes and false number of boxes by

$$n_{\text{cor}}^{(i)} = \min(n_{\text{pred}}^{(i)}, n_{\text{gt}}^{(i)}), \tag{1}$$

$$n_{\text{fal}}^{(i)} = \max(n_{\text{pred}}^{(i)} - n_{\text{gt}}^{(i)}, 0), \tag{2}$$

$$n_{\text{neg}}^{(i)} = \max(n_{\text{gt}}^{(i)} - n_{\text{pred}}^{(i)}, 0), \tag{3}$$

where $n_{\text{gt}}^{(i)}$ is the ground-truth number. With $n_{\text{cor}}^{(i)}$ and $n_{\text{fal}}^{(i)}$, we can obtain

$$\textit{precision} = \frac{\sum_{i\in\mathcal{I}} n_{\text{cor}}^{(i)}}{\sum_{i\in\mathcal{I}} n_{\text{cor}}^{(i)} + \sum_{i\in\mathcal{I}} n_{\text{fal}}^{(i)}}, \tag{4}$$

$$\textit{recall} = \frac{\sum_{i\in\mathcal{I}} n_{\text{cor}}^{(i)}}{\sum_{i\in\mathcal{I}} n_{\text{cor}}^{(i)} + \sum_{i\in\mathcal{I}} n_{\text{neg}}^{(i)}}, \tag{5}$$

where $\mathcal{I}$ denotes indices of the phrases in the prompt, which is defined in the Section 3.1 in the main paper. With the metrics of precision and recall, we compute *F1* score by

$$\textit{F1} = \frac{2 * \textit{precision} * \textit{recall}}{\textit{precision} + \textit{recall}}. \tag{6}$$

Spatial Accuracy

In the experiments on spatial relationship, phrases in the input prompts are given grouth-truth relationship. For instance, in the prompt *a cat is on the left of a dog*, the two phrases *cat* and *dog* have the spatial relationship *on the left of*. By comparing the mean points of the predicted boxes for *cat* and *dog* as

$$(\mathbf{x}_1, \mathbf{y}_1) = (\frac{\mathbf{x}_1^{\text{left}} + \mathbf{x}_1^{\text{right}}}{2}, \frac{\mathbf{y}_1^{\text{left}} + \mathbf{y}_1^{\text{right}}}{2}), \tag{7}$$

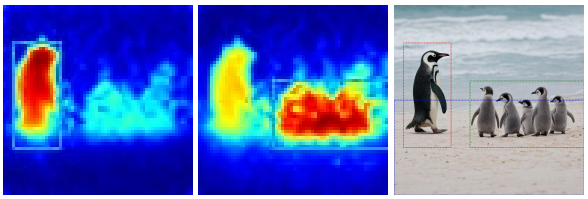
$$(\mathbf{x}_2, \mathbf{y}_2) = (\frac{\mathbf{x}_2^{\text{left}} + \mathbf{x}_2^{\text{right}}}{2}, \frac{\mathbf{y}_2^{\text{left}} + \mathbf{y}_2^{\text{right}}}{2}), \tag{8}$$

and record the number of correct prediction

$$n_{\text{cor}}^{(i)} = \begin{cases} 1, \text{ if } (\mathbf{x}_1, \mathbf{y}_1) \text{ is on the } \textit{ground-truth} \text{ of } (\mathbf{x}_2, \mathbf{y}_2), \\ 0, \text{ if } (\mathbf{x}_1, \mathbf{y}_1) \text{ is not on the } \textit{ground-truth} \text{ of } (\mathbf{x}_2, \mathbf{y}_2), \end{cases} \tag{9}$$

and compute the spatial accuracy

$$ACC_{\text{spatial}} = \frac{\sum_{i\in\mathcal{I}} n_{\text{cor}}^{(i)}}{|\mathcal{I}|}. \tag{10}$$



(a) one penguin (b) baby penguins (c) whole image

Figure 1. Example of overlapping objects in a box.

Size Accuracy

In the experiments on size relationship, phrases in the input prompts are given grouth-truth relationship. For instance, in the prompt *a cat is smaller than a dog*, the two phrases *cat* and *dog* have the size relationship *smaller*. By comparing the area of the predicted boxes $\mathcal{B}_1$ and $\mathcal{B}_2$ for *cat* and *dog* as

$$n_{\text{cor}}^{(i)} = \begin{cases} 1, & \text{if Area}(\mathcal{B}_1) \text{ is } \textit{ground-truth} \text{ than Area}(\mathcal{B}_2), \\ 0, & \text{if Area}(\mathcal{B}_1) \text{ is not } \textit{ground-truth} \text{ than Area}(\mathcal{B}_2), \end{cases} \tag{11}$$

and compute the size accuracy

$$ACC_{\text{size}} = \frac{\sum_{i \in \mathcal{I}} n_{\text{cor}}^{(i)}}{|\mathcal{I}|}. \tag{12}$$

Color Accuracy

In the experiments on color, phrases in the input prompts are given color instruction. For instance, in the prompt *a white cat and a black dog*, the two phrases *cat* and *dog* have the color *white* and *black*. Similarly, we can compute the color accuracy:

$$n_{\text{cor}}^{(i)} = \begin{cases} 1, & \text{if the predicted color is } \textit{ground-truth}, \\ 0, & \text{if the predicted color is not } \textit{ground-truth}, \end{cases} \tag{13}$$

and compute the color accuracy

$$ACC_{\text{color}} = \frac{\sum_{i \in \mathcal{I}} n_{\text{cor}}^{(i)}}{|\mathcal{I}|}. \tag{14}$$

Complex Layouts

As shown in Figure 1, our method excels at handling complex and crowded scenes compared to previous approaches, thanks to the boundary-attention constraint. For **overlapping objects**, the user can assign a unique bounding box to each object and use a distinct phrase to describe them. For example: "One penguin is standing on the left of the beach, and five baby penguins are on the right." The user can assign one box for the phrase One penguin while one box for the phrase baby penguins, as illustrated on the top right. However, the cross-attention maps for multiple objects within the same box inevitably overlap, leading to incorrect counting. We recognize that these training-free L2I schemes, including ours, may struggle with accurate counting in overlapping boxes, resulting in unsatisfactory generation in crowded scenarios. Addressing this limitation would be interesting for future work.

Hyper-parameter Sensitivity

The results in Table 4 highlight the roles of the three loss functions: (1) region-attention loss regulates object location, (2) boundary-attention loss improves object counting accuracy, and (3) regularization loss prevents the attention map from vanishing. Therefore, we set $\lambda = \alpha = 0.5 < 1$ as location regulation is the highest priority. To provide further insights, we will conduct additional experiments with varying λ and α in the revised version.

Text-Layout Mismatch

The entire discussion on improving object counting is based on the assumption that the user inputs a consistent layout and text prompts, which is a reasonable assumption, our method will not improve the counting otherwise. Under the same assumption, the existing schemes have shown poor performance in object counting accuracy, as illustrated in Table 1, because it is challenging to accurately determine the count relying solely on the text prompt.

References

[1] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1

[2] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1