Layered Rendering Diffusion Model for Controllable Zero-Shot Image Synthesis

Zipeng Qi*[1], Guoxi Huang*†[2], Chenyang Liu[1], and Fei Ye[3]

[1] Beihang University
[2] University of Bristol
[3] Mohamed bin Zayed University of Artificial Intelligence

1 The Derivation of Layered Rendering

In order to ensure that the prediction $\tilde{\mathbf{x}}_{t-1}$ for any step from $[T, \cdots, t_0]$ is within the data distribution $p_t(\mathbf{x})$ with the pre-trained score network $\mathbf{s}_{\boldsymbol{\theta}}$, the Layered Rendering method (*i.e.*, Eq.(7) from the main paper) is proposed by solving the following objective

$$\arg\min_{\tilde{\mathbf{x}}_t \in \mathcal{I}_t} \sum_{i=1}^{n} \left\| \boldsymbol{\mathcal{M}}_i \otimes \left(\tilde{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}^i \right) \right\|^2, \tag{1}$$

where $\{\mathbf{x}_{t-1}^i\}_{i=1}^n$ is referred to as the results of n independent denoising processes at timestep t; $\tilde{\mathbf{x}}_{t-1}$ is referred to as the desired result that integrates information from $\{\mathbf{x}_{t-1}^i\}_{i=1}^n$ based on the provided region contains from $\{\boldsymbol{\mathcal{M}}_i\}_{i=1}^n$.
The analytical solution to this objective is given by

$$\sum_{i=1}^{n} \boldsymbol{\mathcal{M}}_i \otimes \tilde{\mathbf{x}}_{t-1} = \sum_{i=1}^{n} \boldsymbol{\mathcal{M}}_i \otimes \mathbf{x}_{t-1}^i \tag{2}$$

$$\therefore \tilde{\mathbf{x}}_{t-1} = \frac{1}{\sum_{i=1}^{n} \boldsymbol{\mathcal{M}}_i} \sum_{i=1}^{n} \boldsymbol{\mathcal{M}}_i \otimes \mathbf{x}_{t-1}^i = \sum_{i=1}^{n} \frac{\boldsymbol{\mathcal{M}}^i}{\sum_{i=1}^{n} \boldsymbol{\mathcal{M}}^i} \otimes \mathbf{x}_{t-1}^i. \tag{3}$$

According to the reverse-time diffusion definition,

$$\mathbf{x}_{t-1}^i \approx \tilde{\alpha}_t \mathbf{x}_t + \underbrace{\tilde{\beta}_t \hat{\mathbf{s}}_t^i}_{\text{estimated direction pointing to } \mathbf{x}_t^i} + \underbrace{\sigma_t \epsilon_t^i}_{\text{random noise}}, \tag{4}$$

* Equal Contribution: Z. Qi and G. Huang
† Corresponding author: G. Huang (guoxi.huang@bristol.ac.uk)

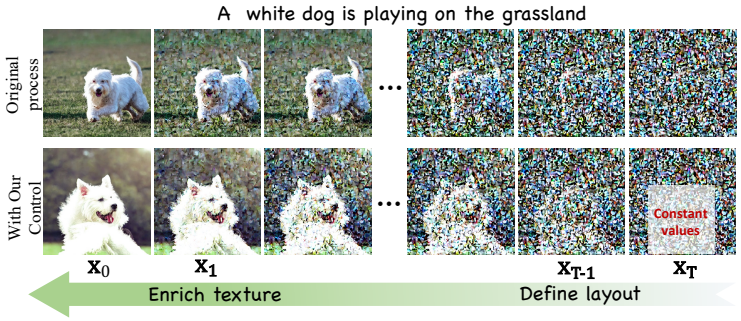

Fig. 1: Top: general reverse diffusion process; Bottom: a reverse diffusion when adding small constant values to a region in the initial noise. The altered region defines the location and size of the synthesised object.

we have

$$
\begin{aligned}
\tilde{\mathbf{x}}_{t-1} &= \sum_{i=1}^{n} \frac{\boldsymbol{\mathcal{M}}^i}{\sum_{i=1}^{n} \boldsymbol{\mathcal{M}}^i} \otimes \left(\tilde{\alpha}_t \mathbf{x}_t^i + \tilde{\beta}_t \hat{\mathbf{s}}_t^i + \sigma_t \epsilon_t^i\right) \\
&= \tilde{\alpha}_t \left(\sum_{i=1}^{n} \frac{\boldsymbol{\mathcal{M}}^i}{\sum_{i=1}^{n} \boldsymbol{\mathcal{M}}^i} \otimes \mathbf{x}_t^i\right) + \tilde{\beta}_t \left(\sum_{i=1}^{n} \frac{\boldsymbol{\mathcal{M}}^i}{\sum_{i=1}^{n} \boldsymbol{\mathcal{M}}^i} \otimes \hat{\mathbf{s}}_t^i\right) + \sigma_t \left(\sum_{i=1}^{n} \frac{\boldsymbol{\mathcal{M}}^i}{\sum_{i=1}^{n} \boldsymbol{\mathcal{M}}^i} \otimes \epsilon_t^i\right) \\
&= \tilde{\alpha}_t \tilde{\mathbf{x}}_t + \tilde{\beta}_t \left(\sum_{i=1}^{n} \frac{\boldsymbol{\mathcal{M}}^i}{\sum_{i=1}^{n} \boldsymbol{\mathcal{M}}^i} \otimes \hat{\mathbf{s}}_t^i\right) + \sigma_t \tilde{\epsilon}_t \qquad (\text{Apply } Eq.\ (3)) \\
&= \tilde{\alpha}_t \tilde{\mathbf{x}}_t + \tilde{\beta}_t \left\{\sum_{i=1}^{n} \frac{\boldsymbol{\mathcal{M}}^i}{\sum_{i=1}^{n} \boldsymbol{\mathcal{M}}^i} \otimes \left[\gamma\, \mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_t + \boldsymbol{\xi}^i, t, y^i) + (1-\gamma)\mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_t, t, \varnothing)\right]\right\} + \sigma_t \tilde{\epsilon}_t \qquad (\text{Apply CFG}) \\
&= \tilde{\alpha}_t \tilde{\mathbf{x}}_t + \tilde{\beta}_t \left\{\gamma \left[\sum_{i=1}^{n} \frac{\boldsymbol{\mathcal{M}}^i}{\sum_{i=1}^{n} \boldsymbol{\mathcal{M}}^i} \otimes \mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_t + \boldsymbol{\xi}^i, t, y^i)\right] + (1-\gamma)\mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_t, t, \varnothing)\right\} + \sigma_t \tilde{\epsilon}_t
\end{aligned} \tag{5}
$$

Let $\boldsymbol{\Phi}_t$ be the compositional estimation at timestep t:

$$
\boldsymbol{\Phi}_t = \sum_{i=1}^{n} \frac{\boldsymbol{\mathcal{M}}^i}{\sum_{i=1}^{n} \boldsymbol{\mathcal{M}}^i} \otimes \mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_t + \boldsymbol{\xi}^i, t, y^i). \tag{6}
$$

Finally, we have

$$
\tilde{\mathbf{x}}_{t-1} = \tilde{\alpha}_t \tilde{\mathbf{x}}_t + \tilde{\beta}_t \left[\gamma\, \boldsymbol{\Phi}_t + (1-\gamma)\mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_t, t, \varnothing)\right] + \sigma_t \tilde{\epsilon}_t. \tag{7}
$$

2 More analysis about Vision Guidance

We first illustrate the effectiveness of the vision guidance mechanism from a simple experiment. A common knowledge of diffusion models is that the spatial layout of a synthesised image is established during the early reverse process. This phenomenon

triggers our thoughts to conduct a preliminary exploration: for each channel, we manually add a small value (a constant version of vision guidance) to a local region within the initial noise, such as adding white colour feature values for a white dog. The lower part of Fig. 1 illustrates the denoising process for this altered initial noise. Intriguingly, we discovered that the region we altered in the initial noise corresponded with where an object, mentioned in the text caption, appeared in the synthesised image. This finding indicates that altering the mean value of a local area in the initial noise can effectively guide the direction of the denoising process.

We further clarify the effectiveness of vision guidance from a distribution transfer perspective. The traditional diffusion denoising process can be seen as transferring the distribution, guided by text or other conditions, from a Gaussian distribution to a dataset-style distribution. In each denoising step, Unet estimates the noise from X_t. Subsequently, we caculate the mean of the distribution at the previous step $t-1$ and sample X_{t-1}. From the main text, vision guidance implies the layout by suppressing the tendency of object generation in the background area while enhancing it in the target area. Within ξ, we define δ as calculated features associated with the object text in the noise feature space, increasing attention and thus influencing the tendency. Adding a constant tensor δ to the i_{th} point in X_t^i, denoted as $X_t^i + \delta$, implies sampling a point from the distribution $\mathcal{N}(u_t + \delta, \sigma_t)$, indicating a preference for generating approximate object features. Conversely, subtracting a constant tensor δ from the i_{th} point in X_t^i, denoted as $X_t^i - \delta$, suuggests sampling a point from the distribution $\mathcal{N}(u_t - \delta, \sigma_t)$, reflecting a preference for not generating approximate object features. The layout ability arises from points in different areas being drawn from distributions with varying tendencies to generate objects. To maintain consistency across distributions, we introduce a small coefficient in δ. Based on the above, before each normal denoising, vision guidance slightly advances the distribution that conforms to the object prompt to one that also conforms to the given layout. After several times, there will be a high probability of generating objects in the target area, and the pre-trained model can generate a result consistent with the layout from the current noise gradually.

From the perspective of attention mechanisms, a more intuitive explanation is that by incorporating features more relevant to the object text in specific areas, we enhance the similarity coefficient between these area features and the object text features. This, in turn, improves the accuracy of generating objects in these areas, based on probabilities.

3 Dataset Construction

We construct a dataset comprising 1,134 global captions along with corresponding bounding boxes, object categories, and instance masks sourced from the MS COCO dataset. The dataset encompasses 55 categories in total (including 11 animal classes and 44 object classes) for spatially conditional text-to-image synthesis, which is summarised in In this dataset, for the layers from $[1, n-1]$, we configured the layered captions in the form of "a $< \text{category name} >$", whereas, for the final layer n, the scene descriptions or other object categories are extracted using LLM from the global captions. Thanks to the exceptional capability of LLM in few-shot learning, we only provide a few examples, *e.g.*, "A bear is running the grassland", the description of the scene is "The beautiful

grassland", and the LLM effectively extracted the context and produced a suitable response. Notably, LLM is simply used for batch processing the global captions in the dataset. In practical applications, our method supports user-defined scene descriptions or provides a default description if not specified. Fig. 2.

{Animals: [bear, bird, cat, cow, dog, elephant, giraffe, horse, person, sheep, zebra],
Objects: [apple, airplane, banana, bed, bench, bowl, boat, book, bottle, bus, car, cake, carrot, cell-phone, chair, clock, couch, donut, fire-hydrant, fork, frisbee, kite, knife, laptop, mouse, motorcycle, orange, oven, pizza, remote, scissors, sink, skis, stop-sign, suitcase, surfboard, snowboard, teddy-bear, toilet, toothbrush, traffic-light, train, truck, vase]

Fig. 2: The constructed dataset contains 55 categories in total.

4 More Visualisation Results

In Fig.4, we display the intermediate results for each layer. Meanwhile, Fig.3 shows additional visualisation results obtained with different random seeds.

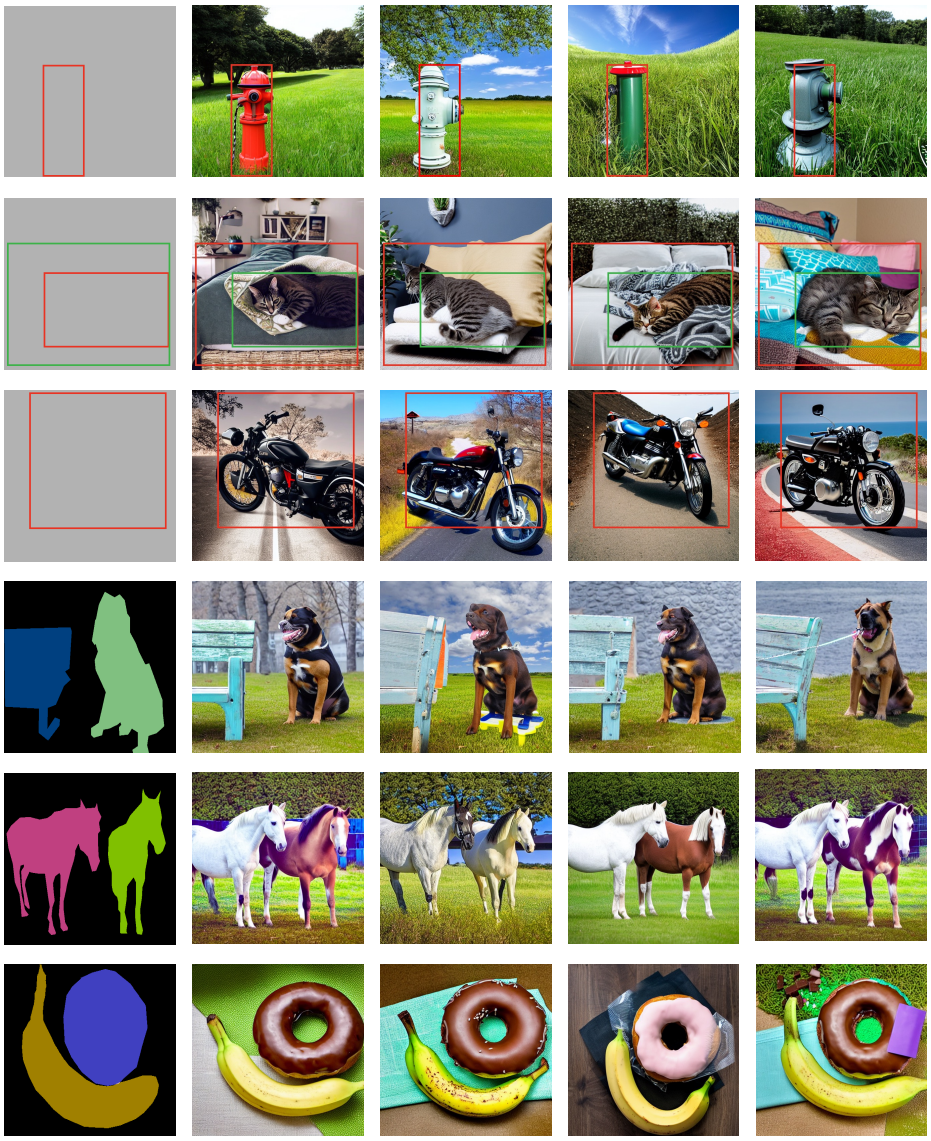

Fig. 3: More visualisation results are presented using our method. Columns 2-5 showcase the outcomes with various random seeds.

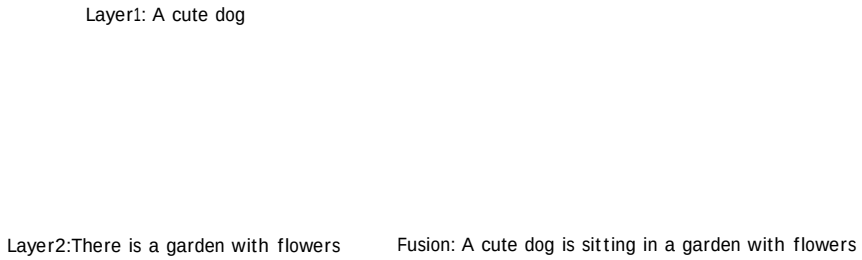

Fig. 4: The visualisation of intermediate results of each layer.