Supplementary Material
Learn, Imagine and Create: Text-to-Image Generation from Prior Knowledge

Training details: LeicaGAN was trained for 650 epochs using the ADAM optimizer with $lr = 0.0002$ on an NVIDIA Tesla V100 machine with 16GB memory. The training took around 2.5 days. The parameter information of each model in the paper are reported in Table 2. Following [21,37,43], for calculating the Inception score and R-precision, each model first generated 30,000 images conditioned on each testing set. Then we calculated the inception score and R-precision results.

Table 1: The number of parameters of each module in LeicaGAN.

Module	Number of parameters
Image Encoder E_V	22.50M
Text-Image Encoder E_T^I	2.075M
Text-Mask Encoder E_T^M	2.075M
Generator G	7.618M
Discriminator D	97.596M

Generators: Our CAG adopts a cascaded architecture, in which images are generated in multi-steps. Figure 1 shows the detailed generator structures for initial step ($i = 0$) and the following steps ($i = \{1, 2, ..m\}$). The **Attention Module** denotes the attention generation process detailed in Eq.(13) and (14) in the paper. Details about each layer are explained in Table 2.

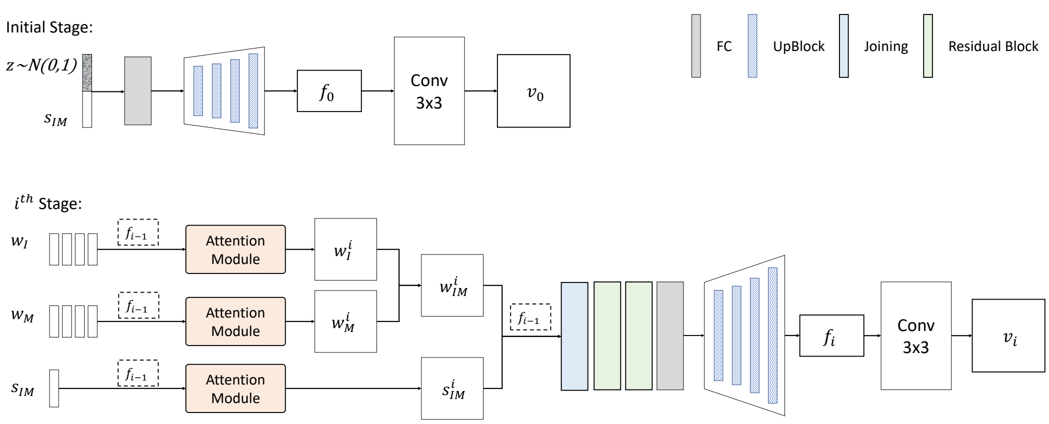

Figure 1: Detailed structures of generators at different steps.

Ablation studies and discussion: More ablation study results are reported in the Table 3. According to our experiment results, it is notice-worthy that a higher Inception Score does not necessarily correspond to a higher R-precision score. We suspect that this may be because these two evaluation metrics evaluate the generated images according to different aspects, $i.e.$ the Inception Score evaluates

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Table 2: Layer Information

Module	**Layer Information**
FC	Fully Connected Layer
UpBlock	Upsampling×2, Conv-(K3×3, S1, P1), BN, GLU
Residual Block	Conv-(K3×3, S1, P1), BN, GLU, Conv-(K3×3, S1, P1), BN
Conv3x3	Conv-(K3×3, S1, P1), Tanh
BN	Batch Normalization
GLU	Gated Linear Unit
K, S, P	kernel, stride, padding

diversity whereas the R-precision score evaluates semantic consistency. Take the Oxford-102 dataset as an example, this dataset contains many similar images with similar text descriptions. Therefore, a higher R-precision score for generated images means that these images semantically match the input text better, however, due to the high similarity among the text descriptions, it might lead to worse diversity and thus a lower Inception Score. We believe it would be valuable to explore more efficient modules to strengthen the impact of imagination and increase the diversity of the generated images, and a more objective evaluation method is also worthy to be explored in the future work.

Table 3: Other results of LeicaGAN trained with E_T^I and E_T^M of different loss weights on the CUB* and Oxford-102* datasets.

Evaluation Metric	**CUB* Dataset**	
	Inception Score	**R-precision [%]**
(1) LeicaGAN, $\gamma_1 = 1$, $\gamma_2 = 0$, $\gamma_3 = 4$, $\gamma_4 = 1$, $\gamma_5 = 1$, $\gamma_6 = 0.5$	5.65±0.06	85.38
(2) LeicaGAN, $\gamma_1 = 1$, $\gamma_2 = 1$, $\gamma_3 = 4$, $\gamma_4 = 1$, $\gamma_5 = 0$, $\gamma_6 = 0.5$	5.62±0.06	84.90
(3) LeicaGAN, $\gamma_1 = 1$, $\gamma_2 = 1$, $\gamma_3 = 4$, $\gamma_4 = 1$, $\gamma_5 = 5$, $\gamma_6 = 5$	5.39±0.05	86.75
(4) LeicaGAN, $\gamma_1 = 1$, $\gamma_2 = 1$, $\gamma_3 = 0$, $\gamma_4 = 1$, $\gamma_5 = 5$, $\gamma_6 = 5$	5.29±0.05	87.35
Evaluation Metric	**Oxford-102* Dataset**	
	Inception Score	**R-precision [%]**
(1) LeicaGAN, $\gamma_1 = 1$, $\gamma_2 = 0$, $\gamma_3 = 4$, $\gamma_4 = 1$, $\gamma_5 = 1$, $\gamma_6 = 0.5$	3.65±0.03	78.92
(2) LeicaGAN, $\gamma_1 = 1$, $\gamma_2 = 0$, $\gamma_3 = 4$, $\gamma_4 = 1$, $\gamma_5 = 0$, $\gamma_6 = 0.5$	3.72±0.03	77.60
(3) LeicaGAN, $\gamma_1 = 1$, $\gamma_2 = 1$, $\gamma_3 = 4$, $\gamma_4 = 1$, $\gamma_5 = 0$, $\gamma_6 = 0.5$	3.60±0.03	79.60
(4) LeicaGAN, $\gamma_1 = 1$, $\gamma_2 = 5$, $\gamma_3 = 0$, $\gamma_4 = 1$, $\gamma_5 = 0$, $\gamma_6 = 0.5$	3.64±0.02	86.37

Figure 2: Examples of images generated by our LeicaGAN on the CUB testing set.

Figure 3: Examples of images generated by our LeicaGAN on the Oxford-102 testing set.

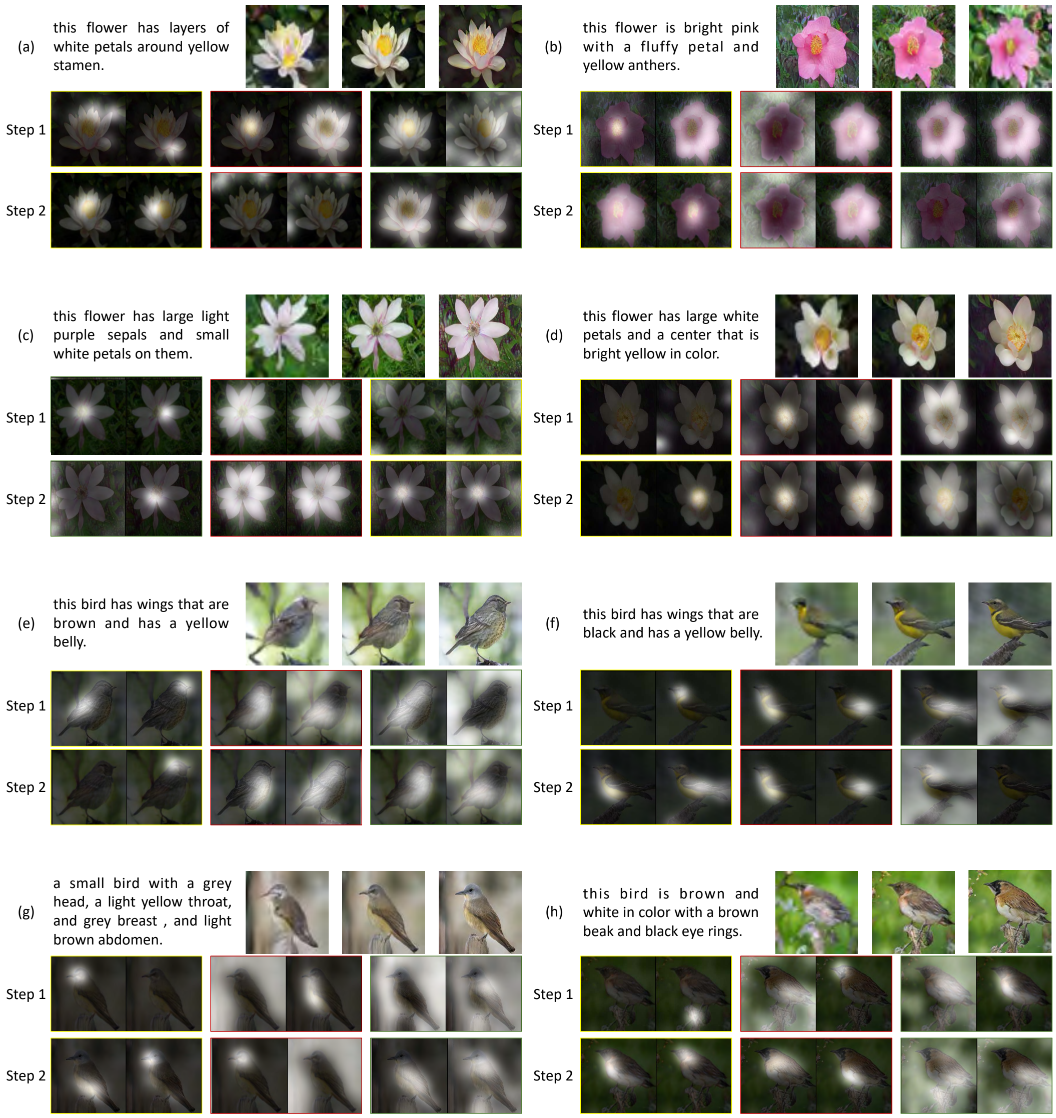

Figure 4: Visualization of attention maps and images generated by LeicaGAN. The first row shows the generated images in three steps. The attention maps shown in yellow, red and green frames correspond to intermediate attention for producing w_I^i, w_M^i and $\hat{s}_{IM}^i$ in Eq.(13) and (14).

Figure 5: **Human perceptual test:** To obtain higher-quality human perception study results, we recruited 60 undergraduate and graduate students. The first 5 trials of each test were considered as training. Then the remaining process consisted of *Real&Fake* and *Shape-Wise Test*, and *Pair-Wise Test*. Images generated by different methods were randomly shown to students. The interfaces of these tests are shown below.