A. Appendix

A.1. Theoretical Results

Theorem 1. *For a given threshold $c > 0$, the generated text feature by* LAFITE$_G$ *satisfies Sim$(f_{img}(\mathbf{x}_i), \mathbf{h}'_i) \geq c$ with probability at least*

$$\begin{aligned}&Prob(Sim(f_{img}(\mathbf{x}_i), \mathbf{h}'_i) \geq c)\\ =&1 - \int_{-1}^{(c-1)/\xi+c} \frac{\Gamma(d/2+1/2)}{\sqrt{\pi}\Gamma(d/2)}(1-x^2)^{d/2-1}dx\end{aligned}$$

where d is the dimension number of features, $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$ is the Gamma function.

Proof. Without loss of generality, we omit the subscript for clearness.

$$\begin{aligned}&\text{Sim}(f_{\text{img}}(\mathbf{x}), \mathbf{h}')\\ =&\frac{\langle f_{\text{img}}(\mathbf{x}), \mathbf{h}'\rangle}{\|f_{\text{img}}(\mathbf{x})\|_2\|\mathbf{h}'\|_2}\\ =&\frac{\langle f_{\text{img}}(\mathbf{x}), f_{\text{img}}(\mathbf{x}) + \xi\epsilon\|f_{\text{img}}(\mathbf{x})\|_2/\|\epsilon\|_2\rangle}{\|f_{\text{img}}(\mathbf{x})\|_2\left\|f_{\text{img}}(\mathbf{x}) + \xi\epsilon\|f_{\text{img}}(\mathbf{x})\|_2/\|\epsilon\|_2\right\|_2}\\ =&\frac{\|f_{\text{img}}(\mathbf{x})\|^2 + \xi\epsilon^\intercal f_{\text{img}}(\mathbf{x})\|f_{\text{img}}(\mathbf{x})\|_2/\|\epsilon\|_2}{\|f_{\text{img}}(\mathbf{x})\|_2\left\|f_{\text{img}}(\mathbf{x}) + \xi\epsilon\|f_{\text{img}}(\mathbf{x})\|_2/\|\epsilon\|_2\right\|_2}\end{aligned}$$

Denote $a = f_{\text{img}}(\mathbf{x})/\|f_{\text{img}}(\mathbf{x})\|_2, b = \epsilon/\|\epsilon\|_2$, then we have

$$\begin{aligned}&\text{Sim}(f_{\text{img}}(\mathbf{x}), \mathbf{h}')\\ =&\frac{\|f_{\text{img}}(\mathbf{x})\|^2 + \xi\epsilon^\intercal f_{\text{img}}(\mathbf{x})\|f_{\text{img}}(\mathbf{x})\|_2/\|\epsilon\|_2}{\|f_{\text{img}}(\mathbf{x})\|_2\left\|f_{\text{img}}(\mathbf{x}) + \xi\epsilon\|f_{\text{img}}(\mathbf{x})\|_2/\|\epsilon\|_2\right\|_2}\\ =&\frac{1+\xi a^\intercal b}{\|a+\xi b\|_2}\\ \geq&\frac{1+\xi a^\intercal b}{\|a\|_2+\xi\|b\|_2}\\ =&\frac{1+\xi a^\intercal b}{1+\xi}\end{aligned}$$

Consequently,

$$\begin{aligned}&\text{Prob}(\text{Sim}(f_{\text{img}}(\mathbf{x}_i), \mathbf{h}'_i) \geq c)\\ \geq&\text{Prob}(\frac{1+\xi a^\intercal b}{1+\xi} \geq c)\\ =&\text{Prob}(1+\xi a^\intercal b \geq c + c\xi)\\ =&\text{Prob}(a^\intercal b \geq (c-1+c\xi)/\xi)\end{aligned}$$

By the cumulative distribution function (CDF) of inner product of random vectors on sphere [3], we know that

$$\text{Prob}(a^\intercal b \leq z) = \int_{-1}^{z} \frac{\Gamma(d/2+1/2)}{\sqrt{\pi}\Gamma(d/2)}(1-x^2)^{d/2-1}\text{d}x$$

Dataset	#train	#validation	caption/image
MS-COCO	82k	40k	5
CUB	9k	3k	10
LN-COCO	134k	8k	1
MM CelebA-HQ	24k	6k	10

Table 7. Statistics of datasets. The last column indicates ratio of captions vs images.

where d is the dimension number of features, $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\text{d}x$ is the Gamma function. Thus we have

$$\begin{aligned}&\text{Prob}(\text{Sim}(f_{\text{img}}(\mathbf{x}_i), \mathbf{h}'_i) \geq c)\\ \geq&\text{Prob}(a^\intercal b \geq (c-1+c\xi)/\xi)\\ =&1 - \int_{-1}^{(c-1)/\xi+c} \frac{\Gamma(d/2+1/2)}{\sqrt{\pi}\Gamma(d/2)}(1-x^2)^{d/2-1}\text{d}x,\end{aligned}$$

which completes the proof. □

A.2. Experiment Details

Datasets The statistics of datasets are summarized in Table 7.

Algorithm 2 Image feature extraction process

Input: An image dataset $\{\mathbf{x}_i\}_{i=1}^N$, image resolution $w \times w$, pre-trained f_{img}, hyper-parameters $a > 0, k \geq 1$
// Image feature generation
for $i = 1$ **to** n **do**
  if use data augmentation **then**
    Initialize $\mathbf{h}'_i \leftarrow \mathbf{0}$;
    for $j = 1$ **to** k **do**
      $\mathbf{h}'_i \leftarrow \mathbf{h}'_i + f_{\text{img}}(\text{CROP}(\mathbf{x}_i))$, where CROP$(\cdot)$ denotes randomly cropping image to be $w' \times w'$, w' is an integer randomly sampled from the range $[a, w]$;
    end for
    $\mathbf{h}'_i \leftarrow \mathbf{h}'_i/k$;
  else
    Initialize $\mathbf{h}'_i \leftarrow f_{\text{img}}(\mathbf{x}_i)$;
  end if
end for

Image feature extraction In practice, we use random cropping as data augmentation when we extract the image features, which is presented in Algorithm 2. The pseudo text features will be generated by perturb the average feature of augmented samples. In our implementation, we set $k = 1, a = 256$ to extract image features used in generating $\mathbf{h}'$, while we set $k = 1, a = 128$ in contrastive loss (7).

RoF	SA	FID ↓	IS ↑	SOA-C ↑	SOA-I ↑
✓		24.85	23.74	30.54	48.72
	✓	25.42	21.14	23.14	38.32
✓	✓	18.04	27.20	36.84	54.16

Table 8. Ablation study on discriminator logits in language-free setting. RoF denotes "real or fake" term, SA denotes "semantic alignment" term.

Hyper-parameter The hyper-parameters are selected based on the performance on MS-COCO dataset. Specifically, τ is selected from $[0.1, 0.2, 0.5, 1.0, 2.0]$, λ, γ are selected from $[0, 1, 2, 5, 10, 20, 50]$.

Exponential sharpening In practice, we found that applying an extra exponential sharpening in contrastive loss makes it easier to reproduce the experiment results, i.e. we add an extra exponential operation right before the softmax function in (6) and (7). Our implementation can be found at https://github.com/drboog/Lafite.

A.3. More Results

We provide the implementation details of image generation with multi-modal conditions, an ablation study on discriminator, and more generated examples under language-free setting.

Generation with multi-modal conditions To generate an image conditioned on both a reference image and text description, we first extract the text feature $\mathbf{h}_1$ from the given text, and pseudo text feature $\mathbf{h}_2'$ from the image. Then $\mathbf{h}_1, \mathbf{h}_2'$ will be feed into the pre-trained generator, leading to two conditional style codes $\mathbf{u}_1$ and $\mathbf{u}_2$. We construct a new conditional style code, whose elements are randomly selected from the corresponding elements in either $\mathbf{u}_1$ or $\mathbf{u}_2$. The new conditional style code will be fed into the generator to generate the desired image.

Note that generation conditioned on image is not reconstruction. Thus when only a reference image is provided, the generated image may have differences with the given image. However, they will share some visible characteristics that are semantic meaningful as illustrated in our examples.

Ablation study on discriminator We test the impact of each term of 4 under language-free setting. The results are provided in Table 8, from which we can see that both terms are important, while the "real or fake" term seems to be more important.

Generated examples Some text-to-image generation results on CUB, MS-COCO, MM CelebA-HQ, LN-COCO are provided in the following figures.

Figure 8. Generating examples on MS-COCO dataset.

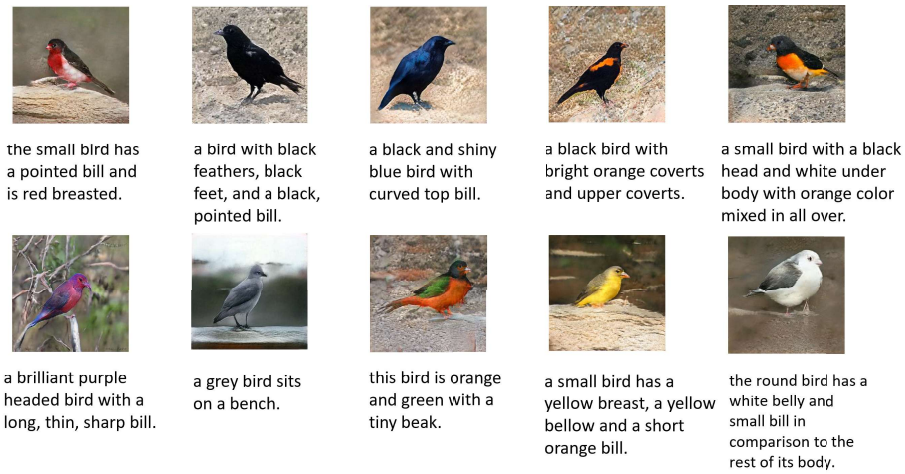

Figure 9. Generating examples on CUB dataset.

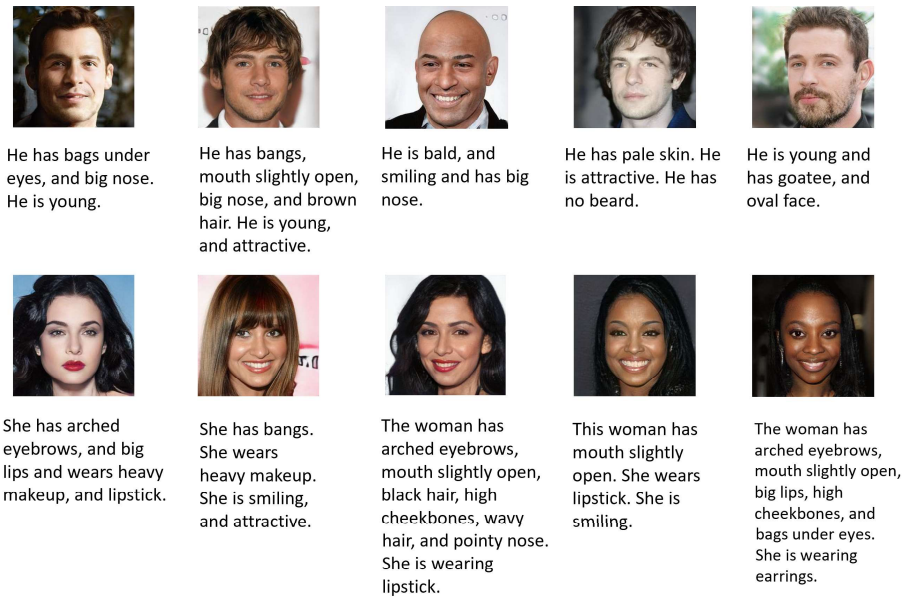

Figure 10. Generating examples on MM CelebA-HQ dataset.

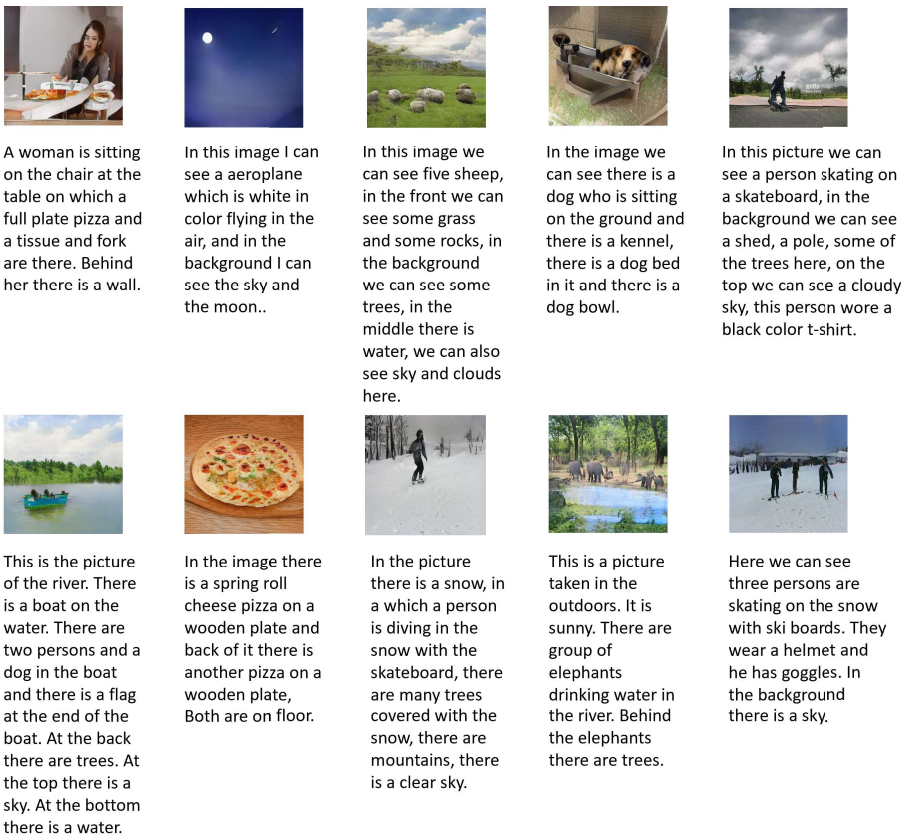

Figure 11. Generating examples on LN-COCO dataset.

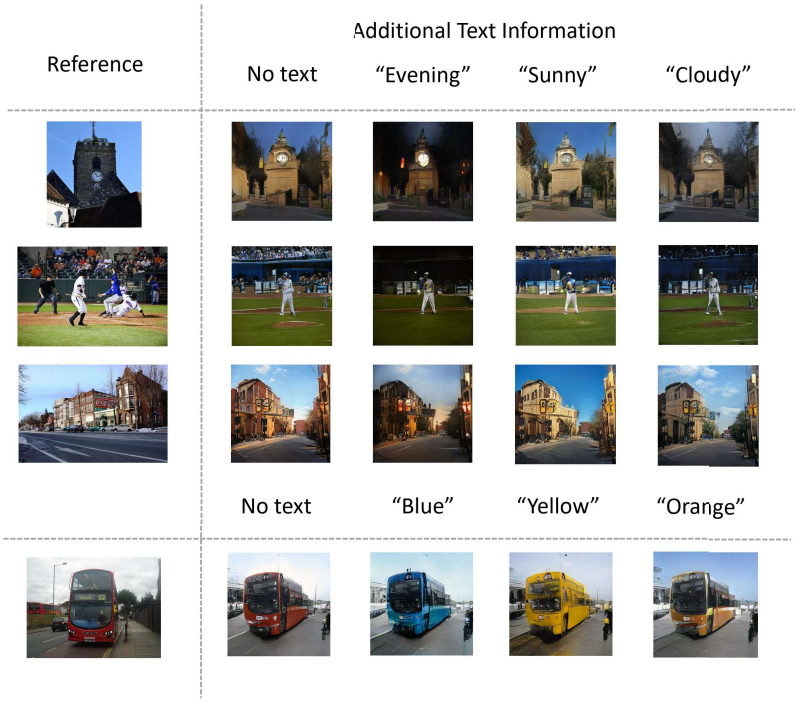

Figure 12. Generating images with multi-modal conditions (conditioned on both image and text) on MS-COCO dataset.

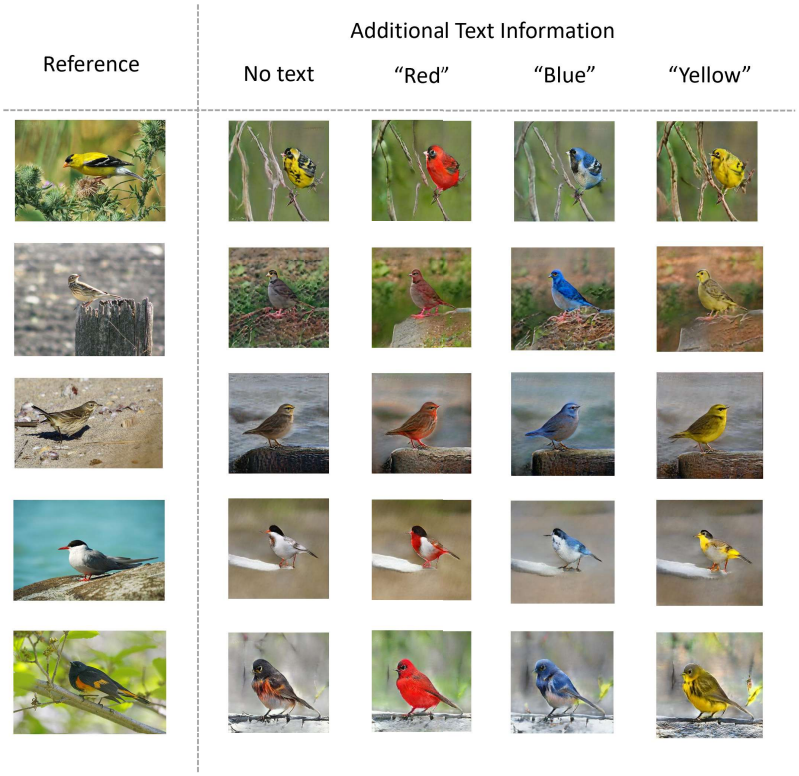

Figure 13. Generating images with multi-modal conditions (conditioned on both image and text) on CUB dataset.

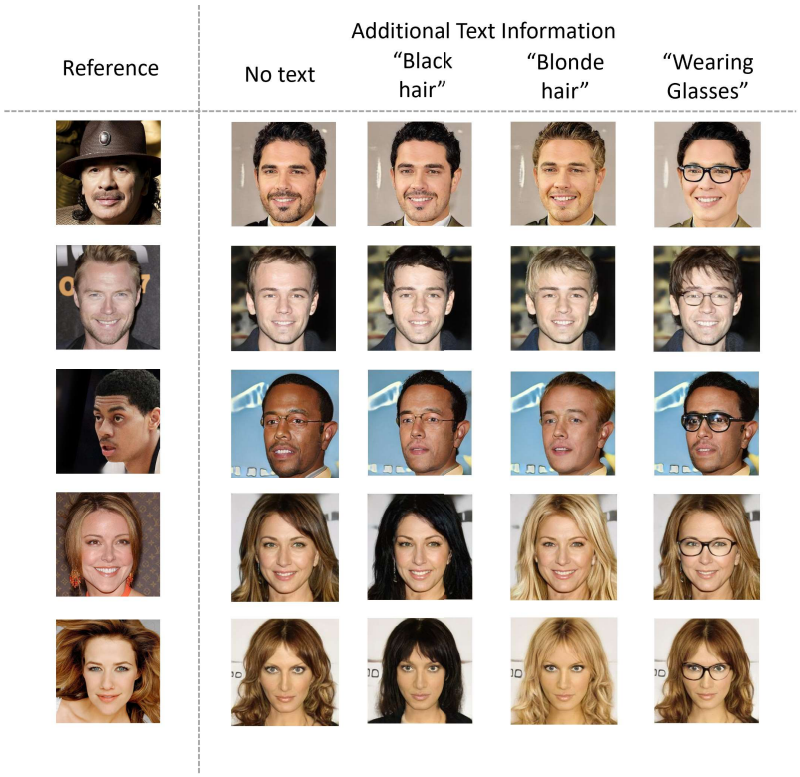

Figure 14. Generating images with multi-modal conditions (conditioned on both image and text) on MM CelebA-HQ dataset.

Figure 15. Generating images with multi-modal conditions (conditioned on both image and text) on LN-COCO dataset.