1. Experimental Details

1.1. Implementation Details

The hyperparameters and training settings for the FA-VAE experiments on different datasets are in Table 1. The training parameters for the CAT model on the CelebA dataset are in Table 2.

1.2. Model Details

The full model of FA-VAE model is in Figure 1. The number of blocks $N$ here corresponds to the number of channel multipliers in Table 1. For instance, the channel multiplier of the FA-VAE model on the CelebA-HQ dataset is $[1, 1, 2, 2, 4]$, then $N = 5$. In all the FA-VAE models for all datasets, we use 4 FCMs as illustrated in Figure 1. FCMs all have the architecture illustrated in Figure 1. FCMs with residual connection in Figure 5 in the paper have the same architecture as the FCMs with convolutional connection as illustrated in Figure 1. FCMs with attention mechanism have the architecture illustrated in Figure 5 in the paper. Due to memory limitation, the last FCM block near the output layer is replaced with FCM with residual connection architecture.

The CLIP model used in the CAT model for training text-to-image generation on CelebA-HQ-MM [11] has text condition embedding dimension of 768.

2. Additional Results

2.1. Reconstruction

Ablation Studies In paper, we give the quantitative and qualitative results of ablation studies when varying the architecture of FCM and settings for the SL and DSL, as in Table 2 and Figure 7. Figure 2 gives additional visualization results of the ablation studies with the frequency spectrums provided as well. We see that the FCM with convolution architecture shows better alignment on the frequency space compared to the original image’s spectrum (Ours w/ DSL* conv) than the residual (Ours w/ DSL* Residual) or attention architecture (Ours w/ DSL* Attention). When comparing different kernel sizes, Ours w/ DSL* $\mu = 3$ to Ours w/ DSL* $\mu = 15$, we see that the frequency spectrum of $\mu = 3$ contains more features on the higher frequency spectrum while a larger kernel size tends to smoothe more the images and we see less high frequency features being captured.

Reconstruction on ImageNet Figure 3 gives additional reconstruction results on the ImageNet dataset [1]. All images are from the validation dataset. We see that FA-VAE shows better reconstruction in local details, such as the flower petals in Figure 3 row 1 column 6 than the baseline VQ-GAN [2] and DALL-E [9]. As discussed in the paper, DALL-E and VQ-GAN tend to produce images that are over-smoothed because the high-frequency spectrum is not accurately reconstructed.

Reconstruction on different input resolution. In Figure 4, we vary the input resolutions of the input image and reconstruct using FA-VAE and the baseline model VQ-GAN. Note that the models used are all trained with image resolution of $(256 \times 256)$ on the ImageNet dataset with a downsampling factor of 16. Figure 4 shows that when the input resolution increases, the reconstruction improves as well, our method FA-VAE shows also better reconstruction in the local details, such as the zebra patterns. As motivated in the introduction of the paper, higher downsampling factor leads to more compressed codebook embeddings. For instance, an image of resolution $(256 \times 256)$, when downsampled 16 times, the latent feature map will be of resolution $(16 \times 16)$, which also means that one codebook embedding in a $(16 \times 16)$ feature map would encode an image patch of $(16 \times 16)$. However, if the downsampling fac-

---

1. https://github.com/openai/CLIP
Table 1. Hyperparameters and FA-VAE's settings for codebook training.

| Dataset        | f   | Channel Multiplier | dropout | attn resolution | FFL weight α | DSL Weight β | Disc weight | | n_z |
|----------------|-----|--------------------|---------|-----------------|--------------|--------------|-------------|------|
| CelebA-HQ [6]  | 16  | [1,1,2,2,4]        | 0.0     | [16]            | 1.0          | 0.01         | 0.75        | 1024 | 256 |
| ImageNet [1]   | 4   | [1,2,4]            | 0.0     | []              | 1.0          | 0.01         | 0.75        | 8192 | 3   |
| ImageNet [1]   | 16  | [1,1,2,2,4]        | 0.0     | [16]            | 1.0          | 0.01         | 0.75        | 16384| 256 |
| FFHQ [7]       | 16  | [1,1,2,2,4]        | 0.0     | [16]            | 1.0          | 0.01         | 0.75        | 2048 | 256 |


<table>
<thead>
<tr>
<th>Dataset</th>
<th>n_layer</th>
<th>n_e</th>
<th>n_heads</th>
<th>dim head</th>
<th>image_encoded_dim</th>
<th>txt_condEmbed</th>
<th>dropout</th>
</tr>
</thead>
<tbody>
<tr>
<td>CelebA-HQ [5]</td>
<td>24</td>
<td>1536</td>
<td>16</td>
<td>64</td>
<td>16</td>
<td>768</td>
<td>0.1</td>
</tr>
</tbody>
</table>

2.2. Generation

Figure 5 gives additional text-to-image generation results on the CelebA-HQ dataset [11] of our method CAT compared with LAFITE [12]. The top-k used is 1024, topp is 0.95 and the temperature is 1.0 for all the generation results in the paper and the supplement. As mentioned in the background section of the paper, LAFITE uses StyleGAN [8] as the decoder which has better decoding capability than VQ-GAN in the face domain because StyleGAN has layers to encode fine-grained image semantics. This also comes with a limitation which is that StyleGAN is hard to scale to large datasets because the number of decoder layers cannot increase beyond a dozen of layers.

Figure 5 shows that the images generated by LAFITE sometimes exhibit unnatural looks. For instance, the second image of second row, the face generated looks rigid and the image texture is similar to the oil painting. While CAT is able to generate more naturalistic images than LAFITE. Note that existing evaluation metrics fall short at capturing the “naturalistic” criterion of the generated image [3,5]. For instance, FID measures the similarity between the latent features of the original and generated images.
Figure 2. Visualization on CelebA-HQ dataset for the ablation studies in Table 2 of the paper. Ours* is FA-VAE model with FCM and FFL, no DSL. Ours** is FA-VAE model with FCM with SL without gaussian filters. DSL* is DSL with non pair-wise sigmas. VQ-GAN is from [2]. VQ-GAN + FFL is with FFL from [4].
Figure 3. ImageNet reconstruction. VQ-GAN with downsampling factor $f = 4$ is from [2], VQ-GAN with $f = 16$ is from [10]. DALL-E is from [9]. The labels for each row of images are: goldfish, tiger, gray whale, Egyptian cat, African elephant, papillon.
Figure 4. Reconstruction using inputs of different resolutions. The default resolution used for training is \((256 \times 256)\). When augmenting the input resolution to \((512 \times 512)\), reconstruction quality improves. The models used are with downsampling factor of 16. The images are fox squirrel and zebra from ImageNet dataset.
"The woman has big lips and is wearing heavy makeup."

"She wears lipstick. She is smiling, has wavy hair, and brown hair.
She has brown hair, and straight hair and wears earrings. She is young."

"This man has big lips, oval face, arched eyebrows, receding hairline, and big nose."

Figure 5. Text-to-image generation on the CelebA-HQ-MM dataset [11]. The first row is our method CAT, the second row is the baseline LAFITE [12]. From row 3-5, the left 3 images are from CAT and the right 3 images are from LAFITE.
References


