Supplement Materials for Catch Missing Details: Image Reconstruction with Frequency Augmented Variational Autoencoder

Xinmiao Lin
Rochester Institute of Technology
xl3439@rit.edu

Yikang Li
OPPO US Research
yikang.li1@oppo.com

Jenhao Hsiao
OPPO US Research
mark@oppo.com

Chiuman Ho
OPPO US Research
chiuman@oppo.com

Yu Kong
Michigan State University
yukong@msu.edu

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 [1] 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 (256×256) 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 (256×256), when downsampled 16 times, the latent feature map will be of resolution (16×16), which also means that one codebook embedding in a (16×16) feature map would encode an image patch of (16×16). However, if the downsampling fac-

[1] https://github.com/openai/CLIP

Dataset	f	Channel Multiplier	dropout	attn resolution	FFL weight α	DSL Weight β	Disc weight	$\lvert\mathfrak{C}\rvert$	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 1. Hyperparameters and FA-VAE'settings for codebook training.

Dataset	n_{layer}	n_e	n_{heads}	dim head	image_encoded_dim	txt_cond_embed	dropout
CelebA-HQ [6]	24	1536	16	64	16	768	0.1

Table 2. Model parameters for text-to-image generation on CelebA-HQ-MM [11] of CAT model.

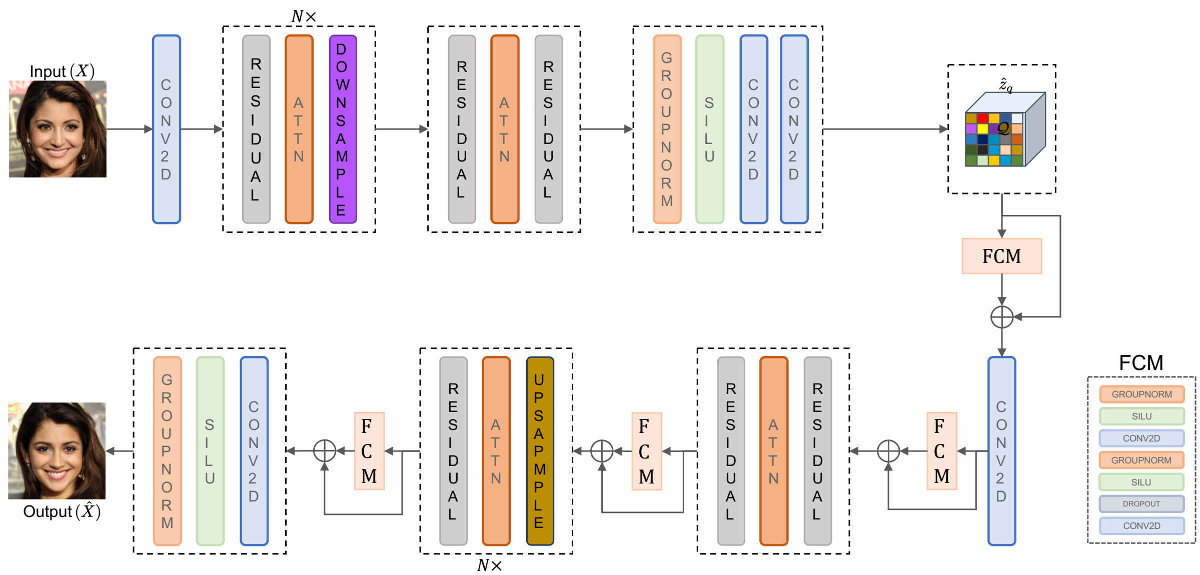

Figure 1. Entire FA-VAE model.

tor is 8, then one codebook embedding would encode an image patch of (8×8) because there are more codebook embeddings in a feature map of (32×32). Thus, a higher downsampling factor leads to more condensed information encoded within a codebook embedding.

Similarly, if we increase the input resolution from (256×256) to (512×512), then the encoded feature map would go from (16×16) to (32×32). While the image semantics remain the same regardless of the input image resolution, a higher resolution leads to a larger encoded feature map, which means that more local details are preserved and thus the reconstruction quality improved with higher input image resolution. In our explanation, we are simplifying the details for the sake of abstractness, such as during the decoding phase, the convolution kernels will merge features from all codebook embeddings.

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, top-p 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].

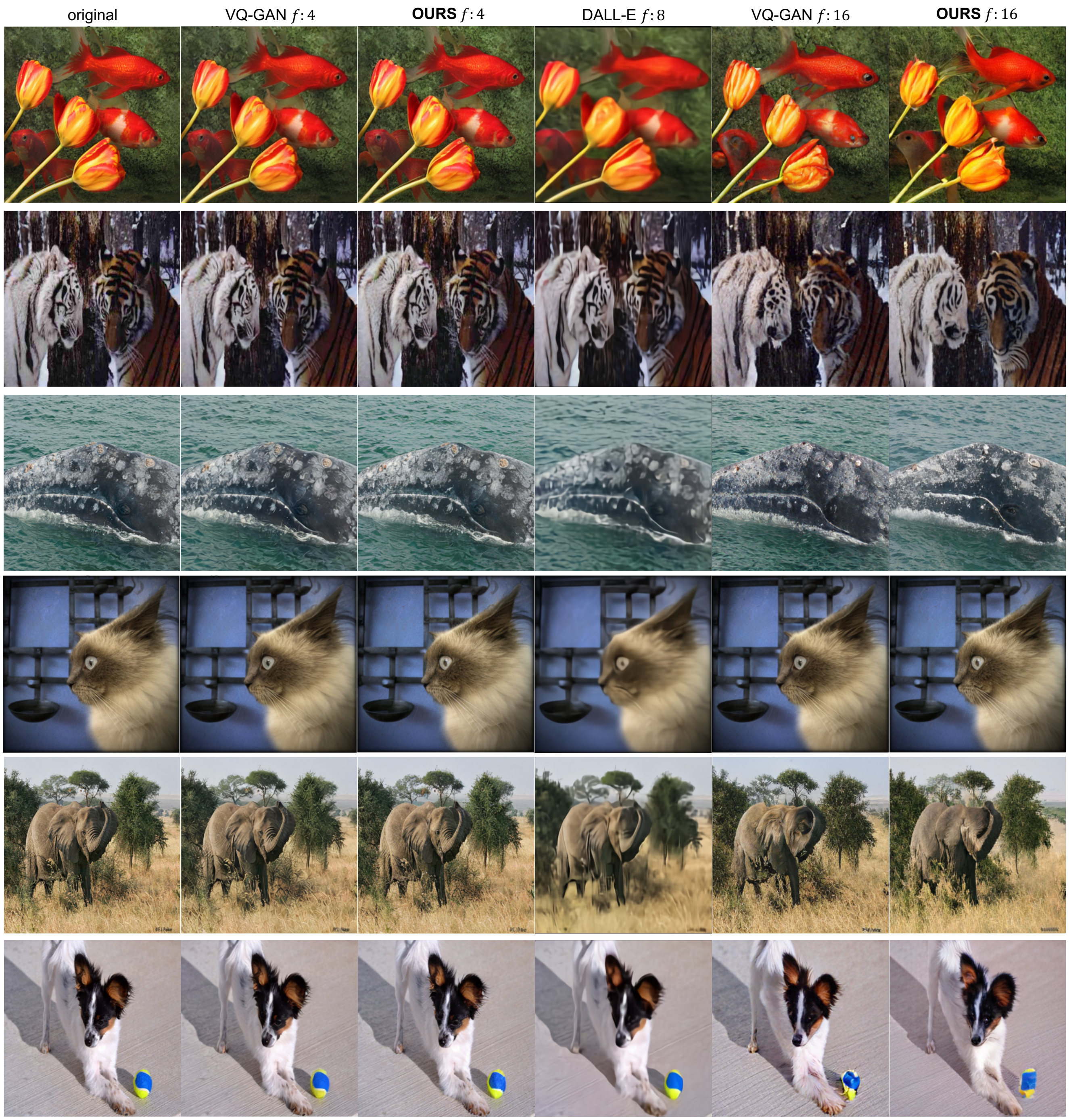

Figure 3. ImageNet reconstruction. VQ-GAN with downsampling factor $f = 4$ is from [2], VQ-GAN with $f = 16$ is from [10]. DALL-E is from [9]. The labels for each row of images are: goldfish, tiger, gray whale, Egyptian cat, African elephant, papillon.

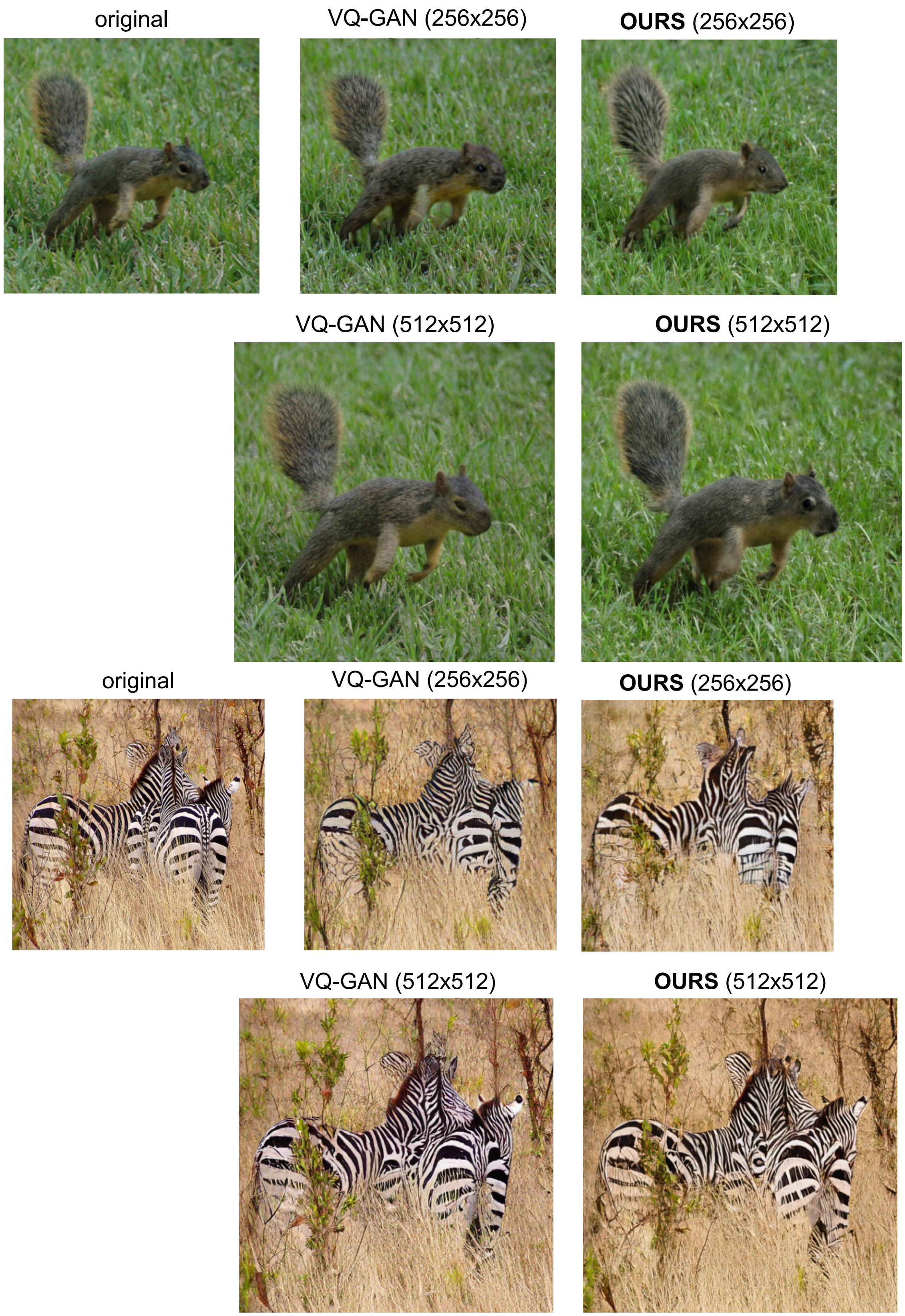

Figure 4. Reconstruction using inputs of different resolutions. The default resolution used for training is (256×256). When augmenting the input resolution to (512×512), reconstruction quality improves. The models used are with downsampling factor of 16. The images are fox squirrel and zebra from ImageNet dataset.

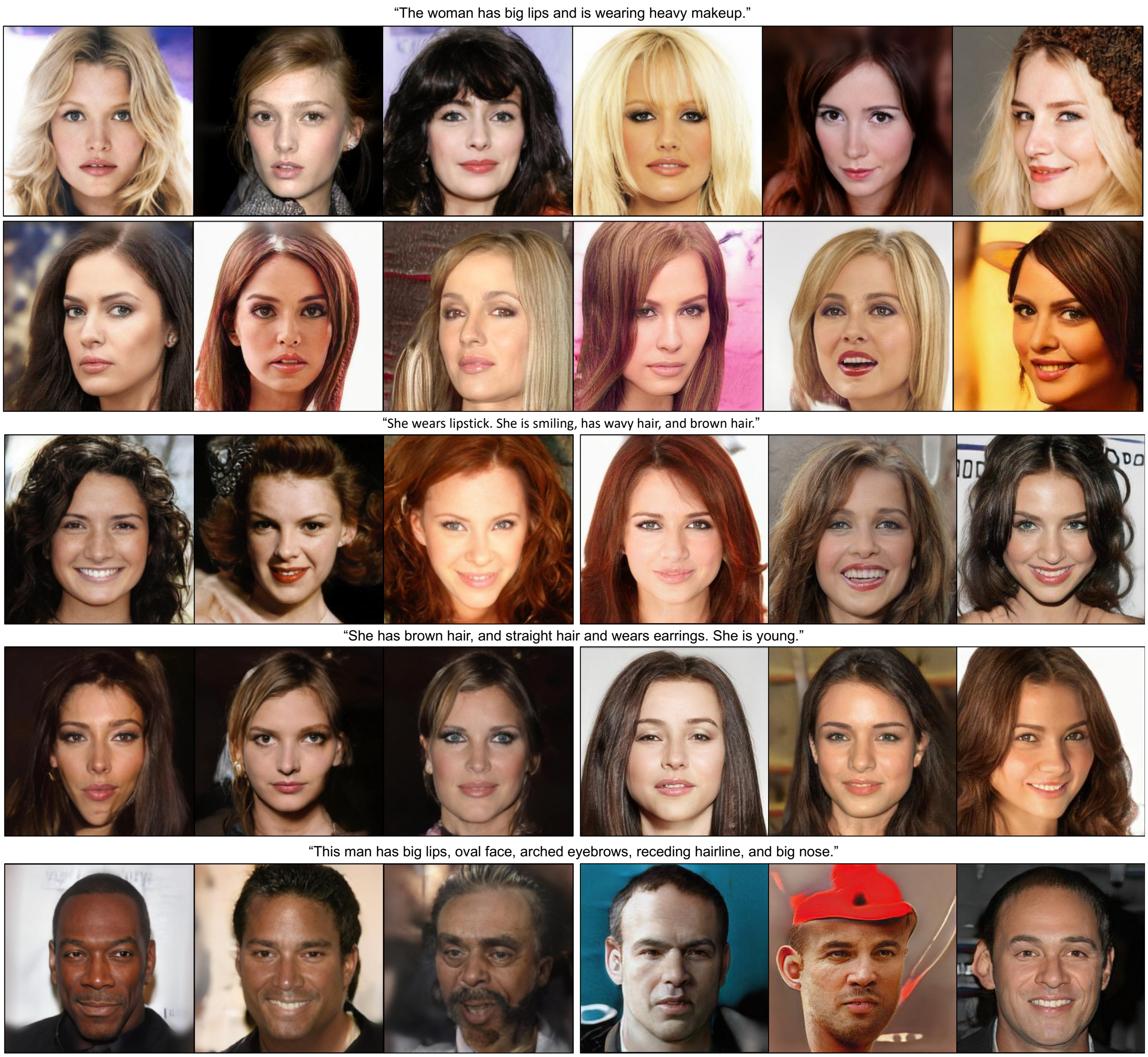

Figure 5. Text-to-image generation on the CelebA-HQ-MM dataset [11]. The first row is our method CAT, the second row is the baseline LAFITE [12]. From row 3-5, the left 3 images are from CAT and the right 3 images are from LAFITE.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 2

[2] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 1, 3, 4

[3] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 2

[4] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *ICCV*, 2021. 3

[5] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 2

[6] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017. 2

[7] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2

[8] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 2

[9] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 1, 4

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 4

[11] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *CVPR*, 2021. 1, 2, 6

[12] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Lafite: Towards language-free training for text-to-image generation. *CVPR*, 2022. 2, 6