Appendix

Datasets

CelebA-HQ is a high-quality version of the CelebA dataset, consisting of 30000 images generated by PG-GAN. We followed [14] instructions to obtain the dataset.

LSUN [42] includes ten scenes and twenty object categories, totally about one million images with label. We mainly use the *Church*, which contains about 126,000 images. The image pre-processing method follows StyleGAN [15].

Discrete VAEs

Our architecture for discrete image representation follows that in [8]. For completeness, a brief description is as follows:

Encoder	Decoder
Conv2D 4×{ResDown} Middle Block GN, Swish, Conv2D	Conv2D Middle Block 4×{ResDown} GN, Swish, Conv2D

[1] ResDown is the combination of a Residual Block and Downsample Block, if the feature map size matches the preset value, there will be an addition non-local self-attention block.

[2] Middle Block is the cascade of one Residual Block, one Self-attention Block and one more Residual Block.

[3] GN means the group normalization [40]

Table 3. Brief Architecture of the VQ-GAN encoder and decoder

For *CelebA-HQ* and *ImageNet*, we obtain the pre-trained checkpoints from the official release, for *LSUN-Church*, we trained a model from scratch under the same configurations for ImageNet in [8]. Specifically, the embedding dimension is 256 and the number of embedded tokens is 1024. The channel numbers of the encoder-decoders is 128, the self-attention block is introduced when the feature map size meets 16 × 16. We set the learning rate is 4.5e-6 for each instance and the learning rate is fixed.

Discrete Diffusion Models

The network structures and hyperparameter settings of discrete diffusion models follow [10]. In detail, the model architecture is based on the backbone of PixelCNN++ [29], which is a U-Net [27] with group normalization. Instead of only adding a self-attention block at 16×16 feature map resolution level, we increase two more self-attention blocks on 8×8 and 4×4 separately. We have 117M parameters for the diffusion models.

For the logits of $\tilde{p}_\theta(\tilde{z}_0|z_t) = \mathrm{Cat}(\tilde{z}_0|p_\theta)$, we predict a noise using the neural network and add it to z_t instead of predicting the $\tilde{z}_0$ directly. As shown in Eq. 23, the desired logits is obtained by superimposing the predicted noise $\mathrm{nn}_\theta(z_t)$ on a calculated z_t

The noise schedule α_t is the same as [22]. The difference is that their parameter $\sqrt{\hat{\alpha}_t}$ is assigned to the mean of the Gaussian distribution, while our factor $\hat{\alpha}_t$ is the parameter of the categorical distribution. The definition is given in Eq. 19 and $s = 0.008$. We also sample t with $q(t) \propto \sqrt{\mathbb{E}[L_t^2]}$ instead of uniform sampling [22].

The batch size is 180 per GPU and the learning rate is 0.0001 with Adam optimizer with standard settings. The learning rate scheduler is the cosine annealing scheduler with 1 million steps. We have not employed any dropout in the model.

Additional Results

In Figures 9 & 10, we show additional generation results based on CelebA and on LSUN-Church. We also provide additional results for image inpainting in Fig. 11.

Risk of overfitting

As described in [8], FID scores cannot detect an overfitting, while early-stopping based on validation NLL can prevent overfitting. In Fig 12, we show top-10 nearest neighbors based on *LPIPS* distance [44] for the training image. We can find that the nearest neighboring generated image is not the reproduced original image and we can infer that there is no overfitting in such model.

Societal Impact

Our work is an extension of the diffusion model, which also belongs to the family of generative models. It can be used to generate fake images or videos to disseminate disinformation, however, as our adopted datasets are collected from the Internet, which will contain the biases, the generated images from our model are also difficult to escape from the bias caused by training data.

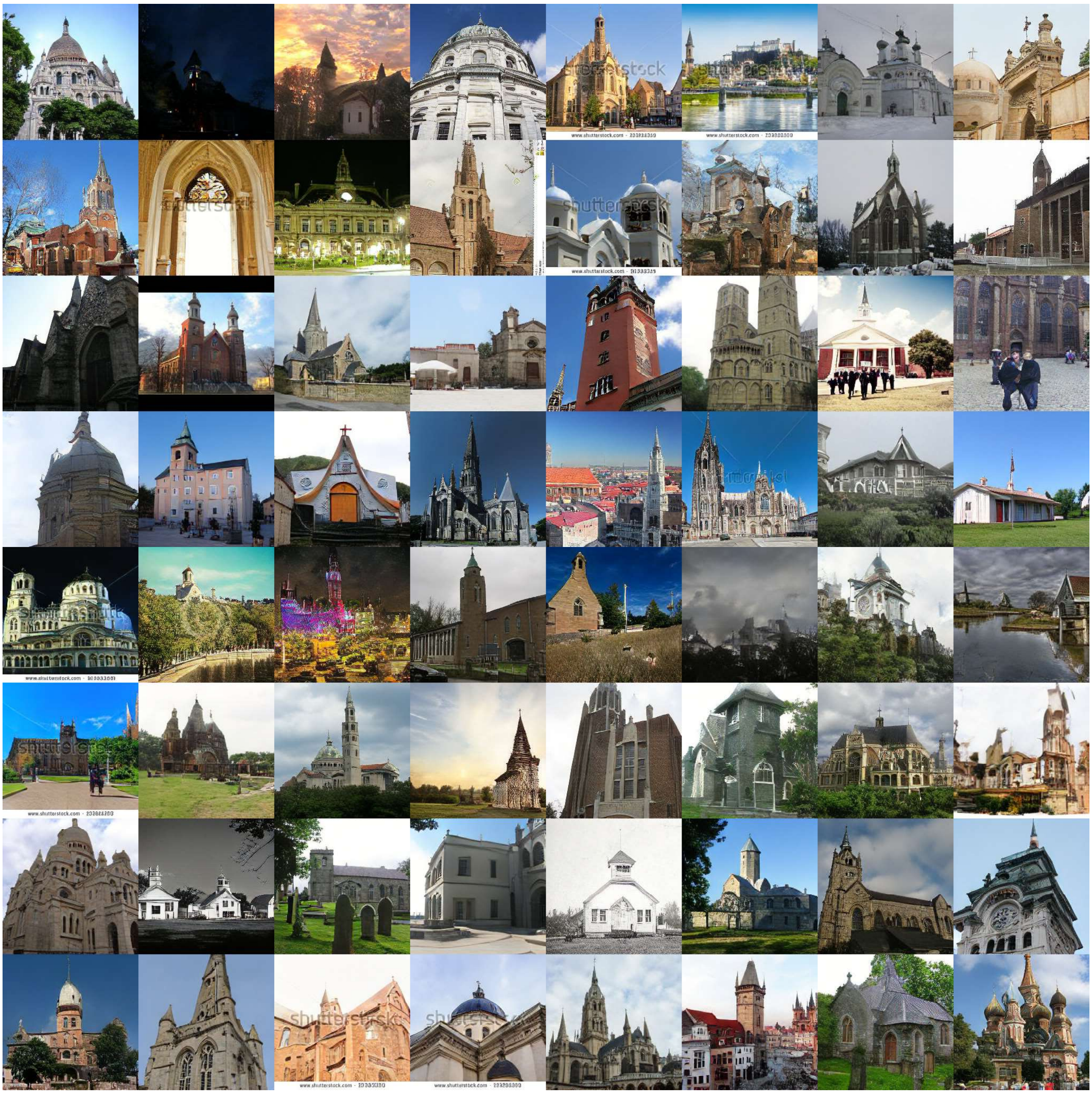

Figure 9. Additional samples on LSUN-Church.

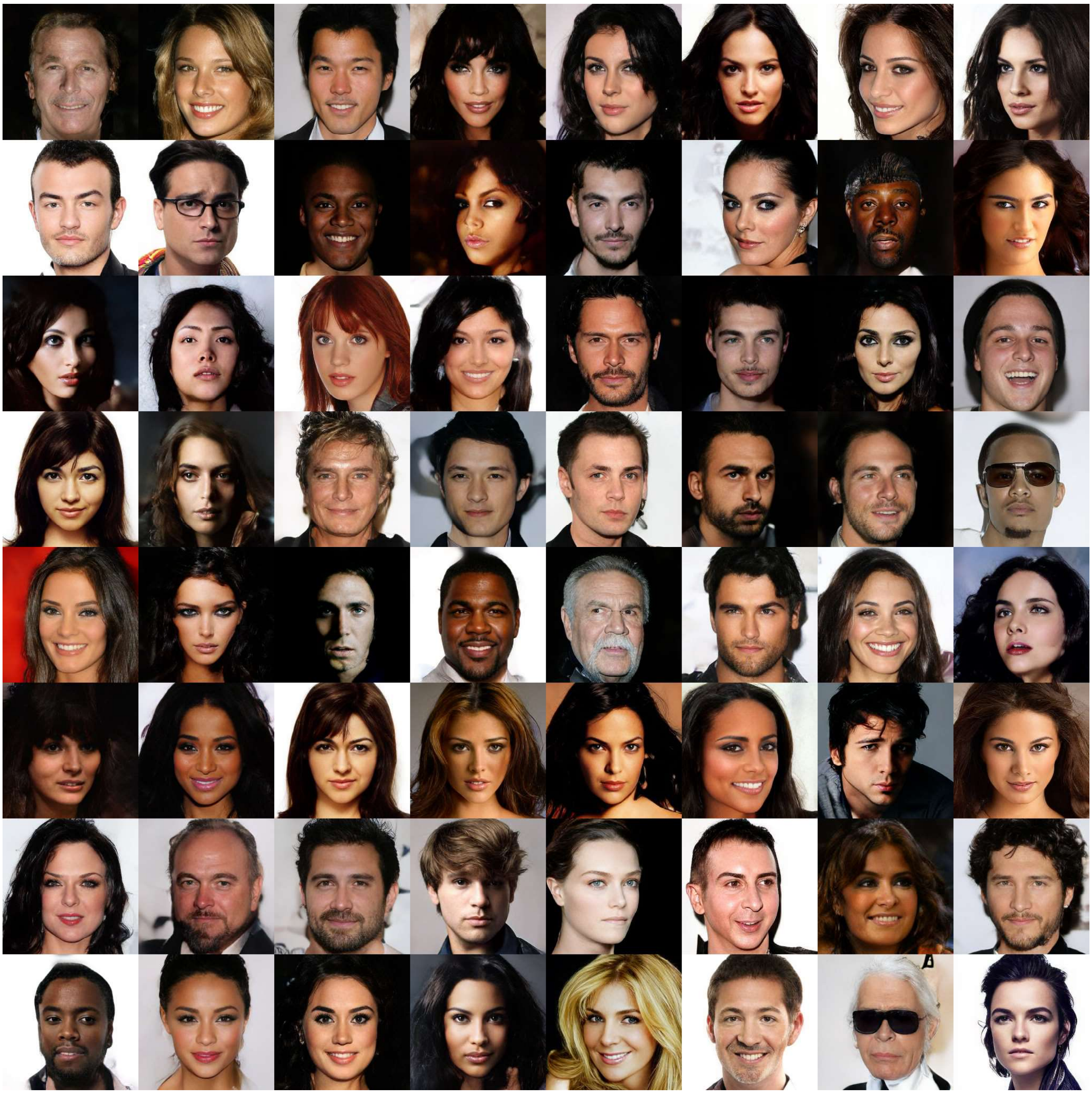

Figure 10. Additional samples on CelebA-HQ.

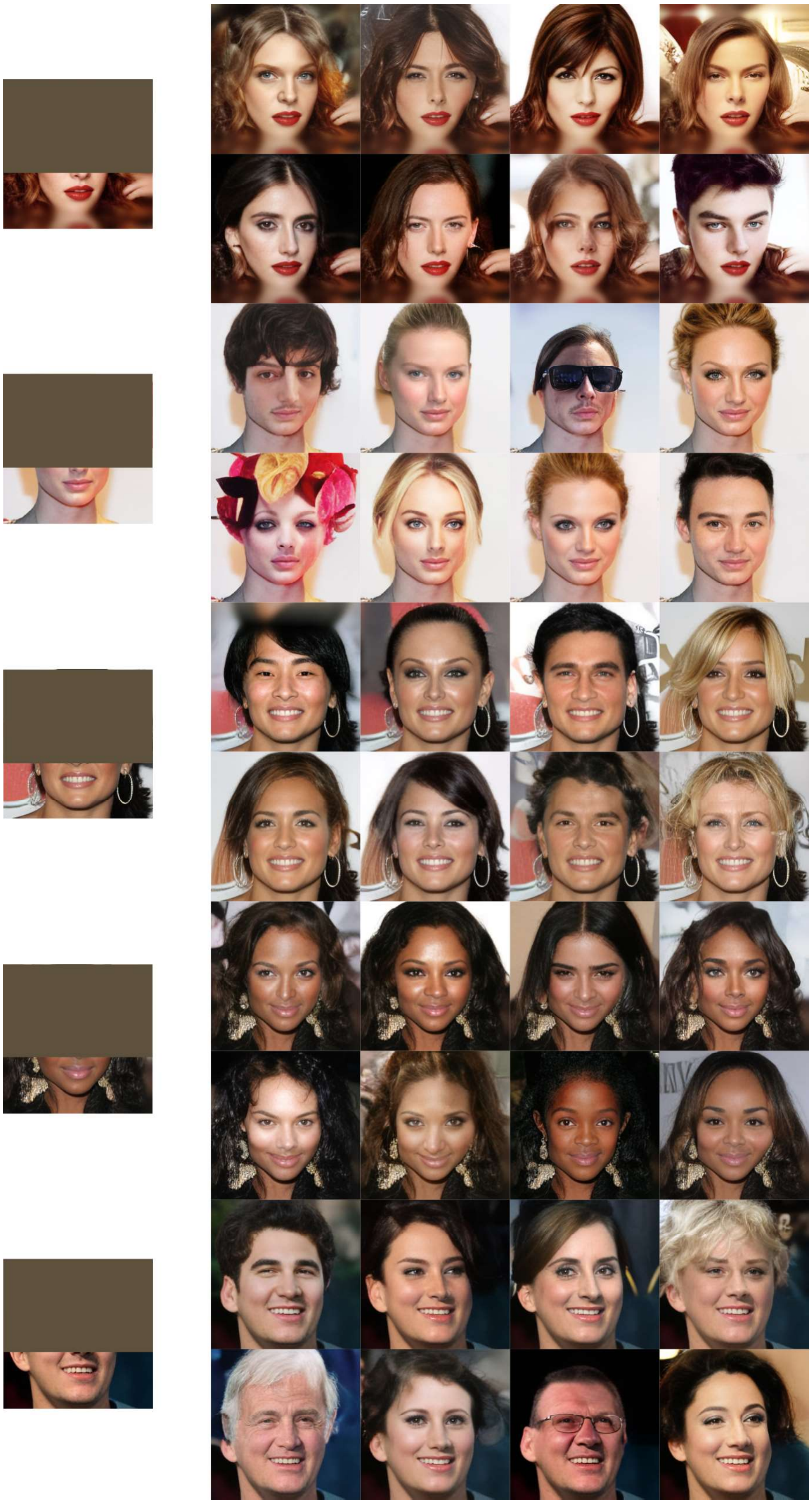

Figure 11. Additional samples on image inpainting for CelebA-HQ.

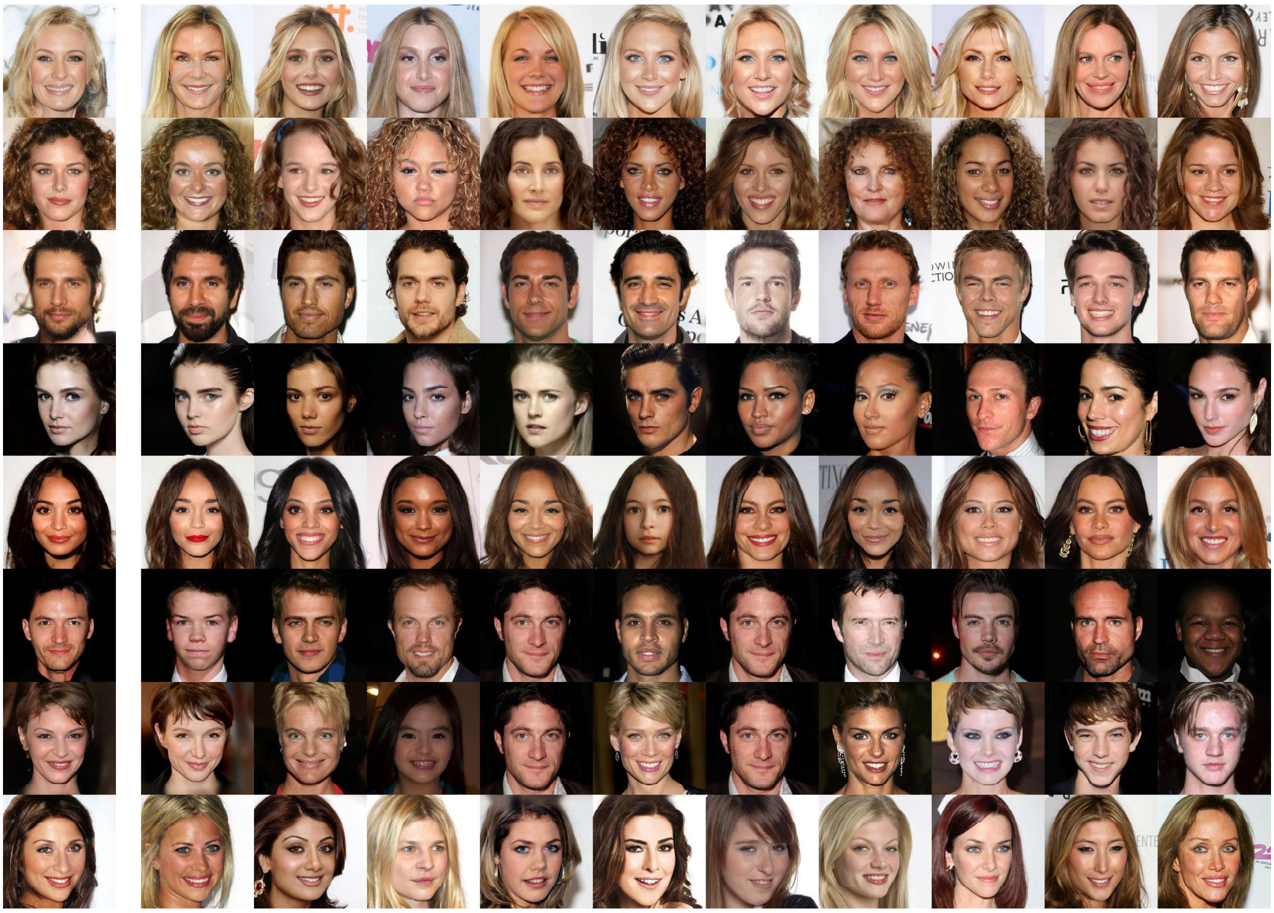

Figure 12. Nearest Neighbours for CelebA-HQ 256 × 256 model. The left column are the images generated by our model, and the remaining images are the nearest neighbors(with minimum LPIPS distance) from the training set.