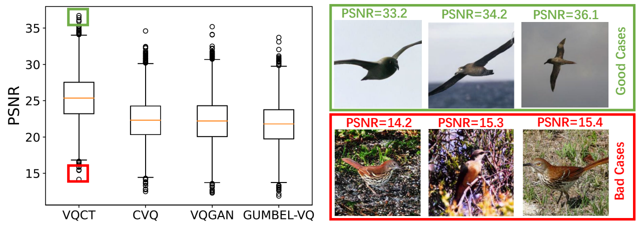
Supplementary Materials: Codebook Transfer with Part-of-Speech for Vector-Quantized Image Modeling

Baoquan Zhang[1], Huaibin Wang[1], Chuyao Luo[1], Xutao Li[1,3], Guotao Liang[1,3], Yunming Ye[*1,3], Xiaochen Qi[2], Yao He[2]
[1] Harbin Institute of Technology, Shenzhen; [2] ShenZhen SiFar Co., Ltd.; [3] Peng Cheng Laboratory
baoquanzhang@hit.edu.cn, 22S051022@stu.hit.edu.cn, luochuyao.dalian@gmail.com,
lianggt@pcl.ac.cn, {lixutao, yeyunming}@hit.edu.cn, {joeqxc1974, heyao18818}@gmail.com

model	codes	trainable parameters
VQ-VAE	1024	262k
VQ-GAN	1024	262k
Gumbel-VQ	1024	328k
CVQ	1024	262k
VQCT(ours)	adj 1949 noun 4258	197k

Table 1. details of 5 VQIM methods's codebook



(a) PSNR (↑)

(b) Extreme Cases

Figure 1. Statistical result analysis of image levels on CUB

1. Codebook Details.

The vocabulary filtering of codebook is not dataset-specific, which is shared for all datasets. We introduce WordNet as part-of-speech prior and then filter vocabulary by 1) using WordNet's "pos_tag()" to remove vision-unrelated words (e.g., pron.); and 2) using word frequency (WF) in corpus to remove infrequent words ($WF < 10$). Finally, only the adjective and noun with $WF \geq 10$ are retained as final vocabularies. In Tabel 1, we report the details (including the number of codes and the number of trainable parameters) of codebook of VQ-VAE, VQ-GAN, Gumbel-VQ, CVQ, and our VQCT. From results, we can see that although our VQCT has a bigger codebook,our VQCT has fewer trainable parameters. This is very reasonable because our codebook is generated from pretrained codebook rather than directly learned. The advantage of such design is that the semantic relationships between codes can be fully exploited for achieving cooperative optimization between codes.

2. Encoder Details.

For the encoder, we divide its feature vectors in a 1:1 manner into adj and noun parts. This is because our adjective and noun codebooks are all generated by our codebook transfer network, which have the same feature dimensions.

3. More Experiment Results

3.1. Statistical Qualitative Results on Image Levels

In Figure 1(a), we also report the statistical results of image levels on CUB. Here, we evaluate all test images by PSNR, and then show statistical results in a box-plots manner. From results, we find that our VQCT is also superior over existing baselines on overall distribution. Beside, we also visualize some cases with extreme performances in Figure 1(b), which suggests that our VQCT performs better on some images with clear objects, while worse on images with clutter background.

3.2. Image Completion

We also apply our VQCT method on image completion on CelebA-HQ, generating images based on masked images. The results are shown in Figure 2. From these experimental results, we can see that our VQCT indeed can achieve high quality image completion.

3.3. Image Synthesis

In Figure 3, we show more generation examples of our VQCT on image synthesis. From results, we find that our VQCT indeed can achieve high quality image synthesis.

[*] Corresponding author.

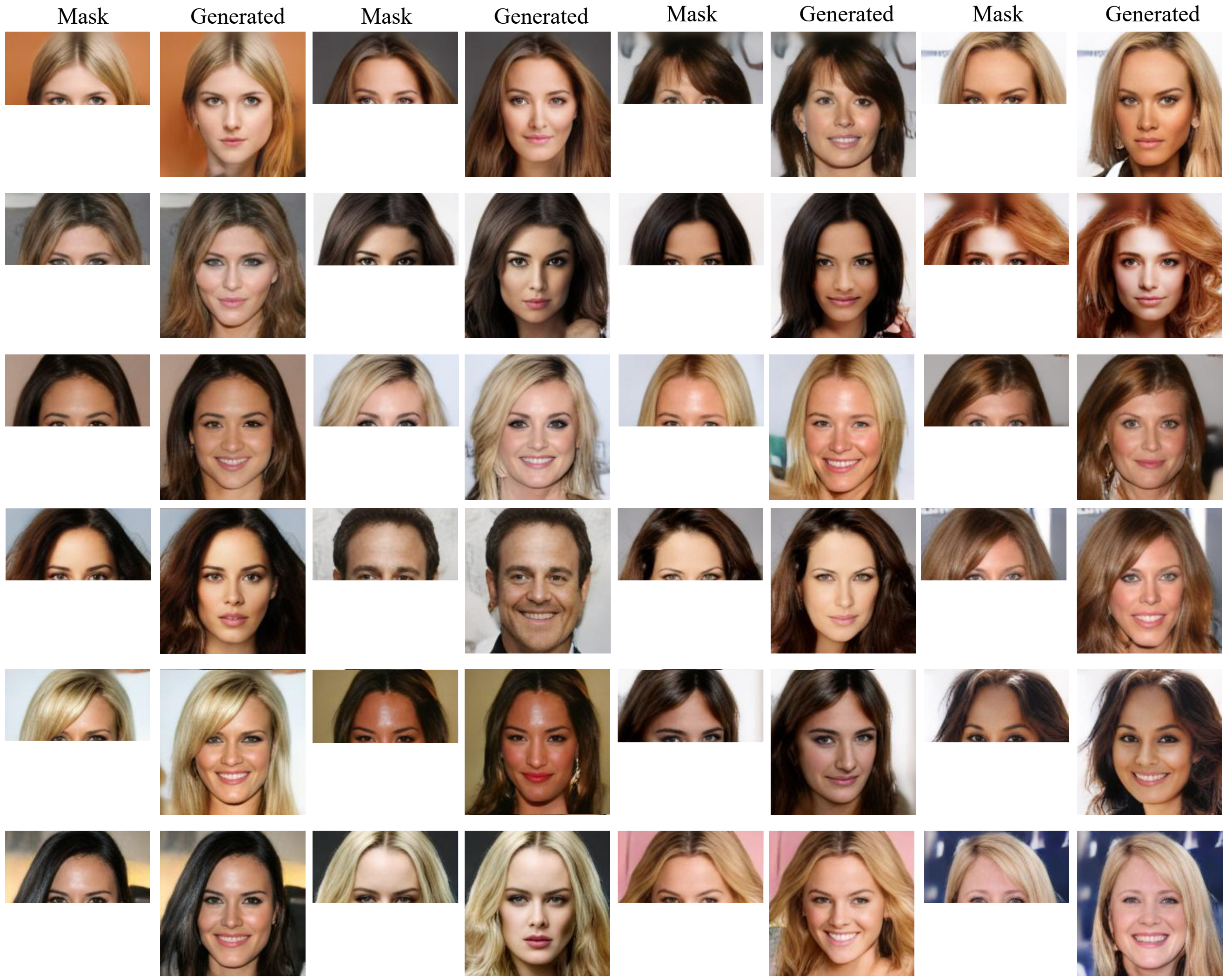


Figure 2. image completion on CelebA-HQ

3.4. Results on ImageNet

In Table 2, we take VQGAN and SPAE as baselines and conduct performance comparison on ImageNet. From results, we find that our VQCT is superior over SPAE, which is because we focus on transferring codebook instead of freezing codebook.

Table 2. Comparison on ImageNet. “*” is the results from [34].

Method	VQGAN*	SPAE*	Our VQCT
FID ↓	4.04	3.60	**1.75**

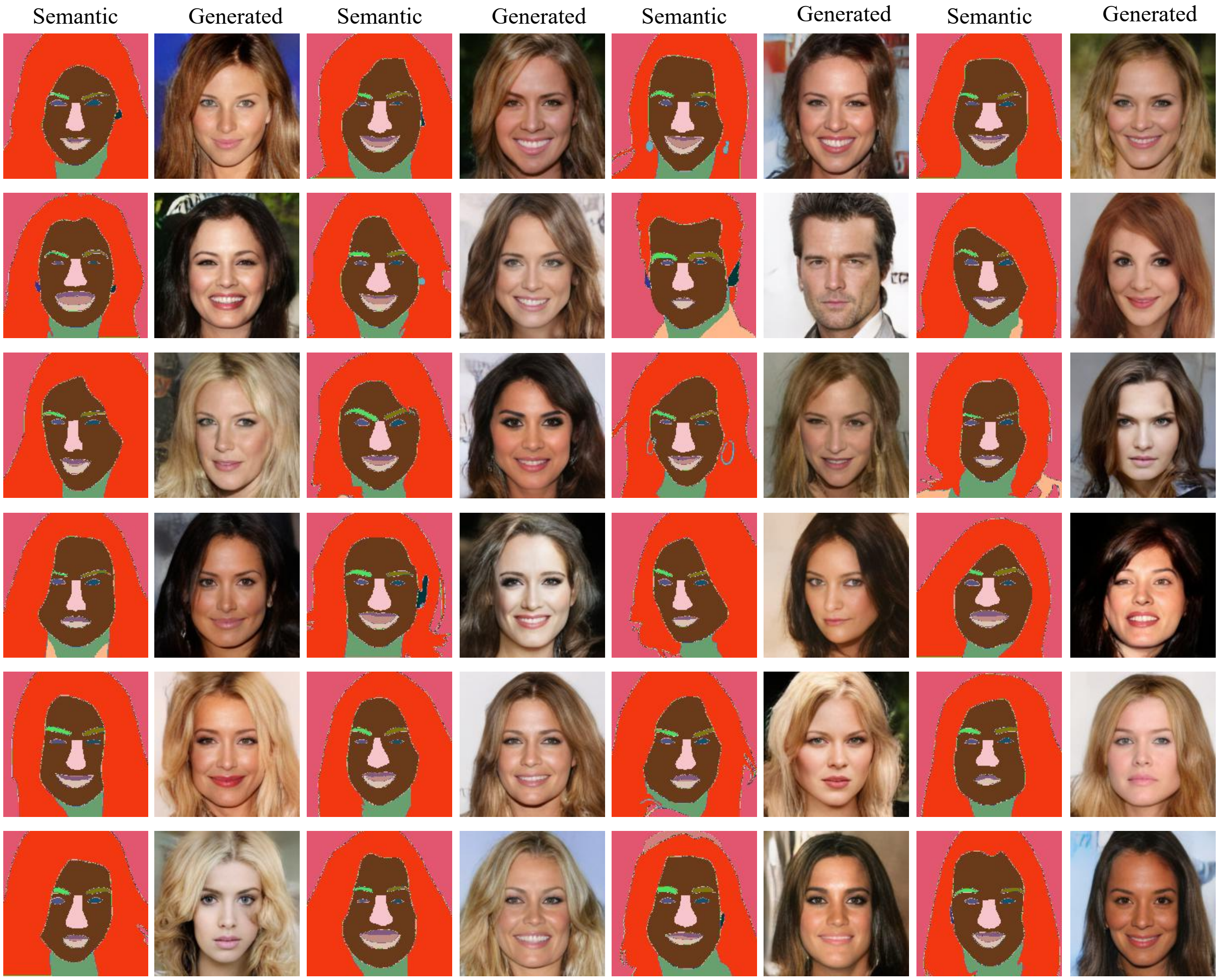


Figure 3. More semantic image synthesis on CElebA-HQ