Addressing Representation Collapse in Vector Quantized Models with One Linear Layer

Supplementary Material

8. Appendix

8.1. Experimental Configurations

Tab. 5 provides the experimental configurations for both image and audio modalities utilized in this study. For the image modality, the input size is specified as $128 \times 128 \times 3$. The batch size for images is set at 256. The model is trained for a total of 50 epochs. Each image is represented with a quantized sequence length of 16×16, dividing the input data into a grid of tokens. In terms of optimization, the AdamW optimizer is employed with a constant learning rate of $1e-4$, and no warmup epochs are implemented. The commitment coefficient for images is set to 1.0. The adversarial coefficient for this modality is established at 0.1, affecting the training dynamics in the context of adversarial methodologies. Regarding data augmentation, a random horizontal flip is applied to the image inputs, enhancing the robustness of the model.

The audio input size is defined as $24,000 \times 1$, reflecting a one-dimensional audio signal sampled at a rate of $24,000$ Hz (1 second). The batch size for audio data is set at 64. The model undergoes a training duration of 50 epochs. The optimization settings remain consistent, utilizing the AdamW optimizer and a constant learning rate of $1e-4$ with no warmup epochs. The commitment coefficient for audio is set to 1000.0 and the adversarial coefficient is set at 1.0, which is the same as WavTokenizer.

Config	Image	Audio
inputs	pixels	window size
input size	128 × 128 × 3	24,000 × 1
batch size	256	64
training epochs	50	50
quantized sequence length	16 × 16	75
optimization		
optimizer	AdamW	AdamW
learning rate	1e-4	1e-4
learning rate schedule	constant	constant
warmup epochs	0	0
commitment coefficient	1.0	1000.0
adversarial coefficient	0.1	1.0
data augmentation		
random horizontal flip	true	false

Table 5. Experimental configurations on image and audio.

8.2. Loss Curve

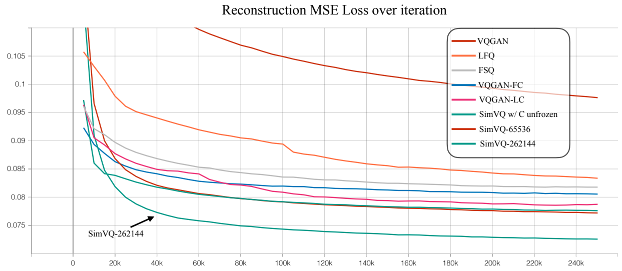

Figure 6. The loss curve over epochs of different models on the validation dataset.

8.3. Codebook Distribution

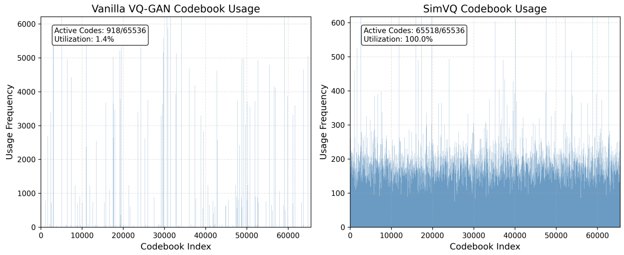

Figure 7. The frequency of codebook on ImageNet validation set.

8.4. Qualitative Cases

We provide image and audio cases of SimVQ with various codebook sizes below.

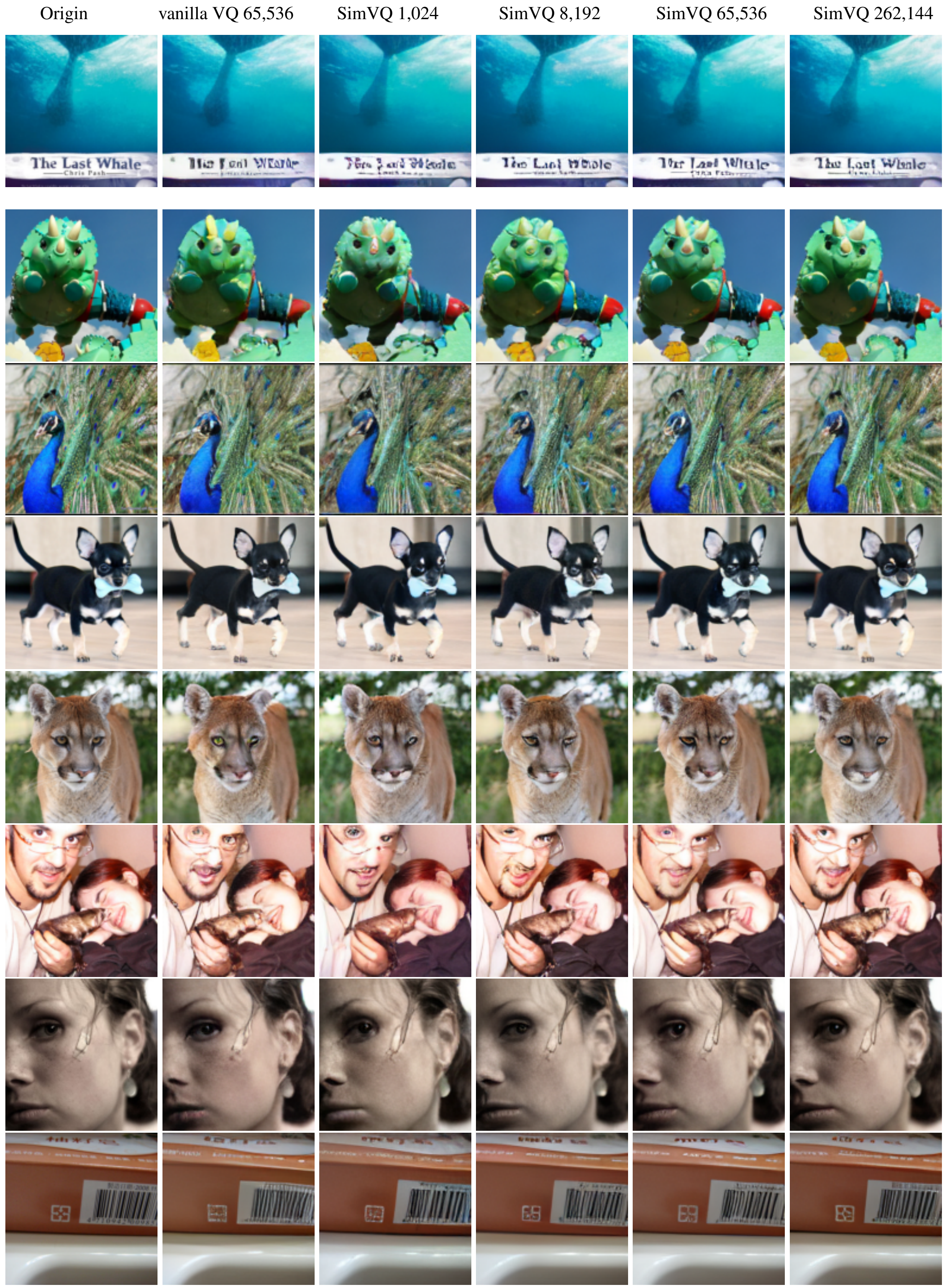

Figure 8. Image reconstruction samples with different codebook sizes.

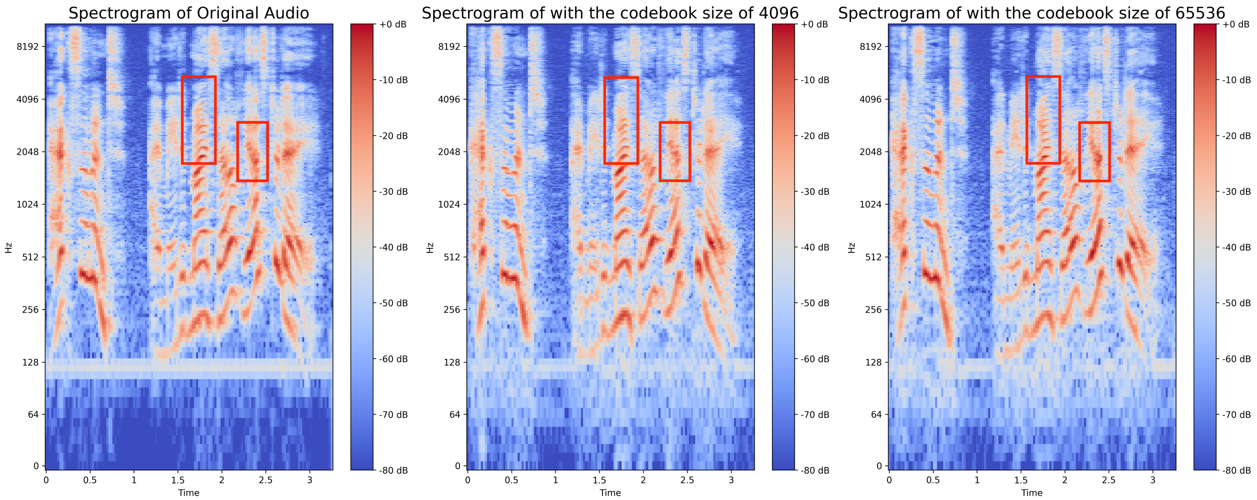

Figure 9. The spectrogram of audio reconstruction samples with different codebook sizes.

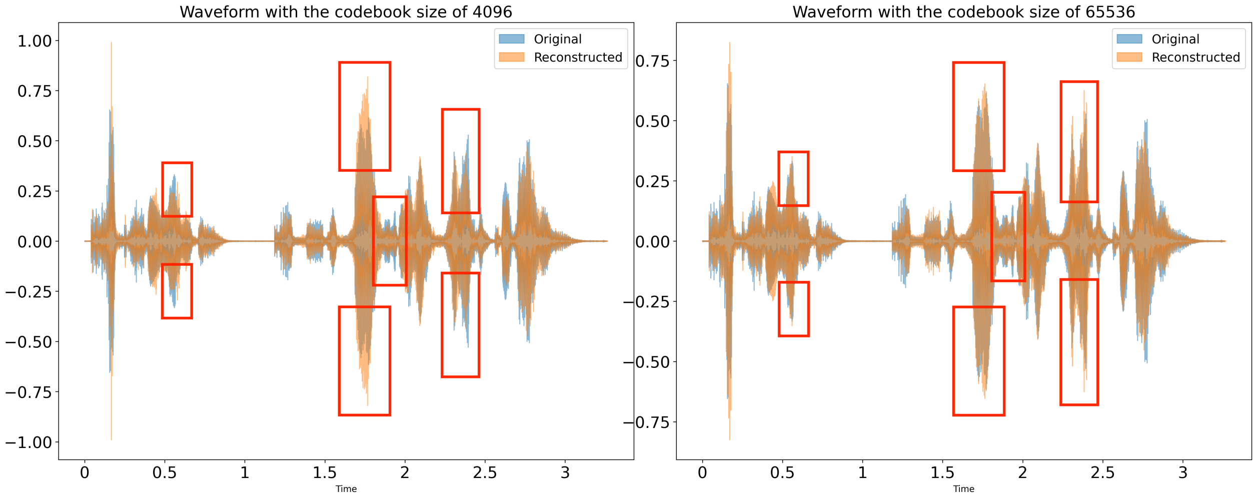

Figure 10. The waveform of audio reconstruction samples with different codebook sizes.