Towards Accurate Image Coding: Improved Autoregressive Image Generation with Dynamic Vector Quantization (Supplementary Material)

Mengqi Huang[1], Zhendong Mao[1, 2]*, Zhuowei Chen[1], Yongdong Zhang[1, 2]
[1]University of Science and Technology of China, Hefei, China;
[2]Institute of Artificial intelligence, Hefei Comprehensive National Science Center, Hefei, China
{huangmq, chenzw01}@mail.ustc.edu.cn, {zdmao, zhyd73}@ustc.edu.cn

1. Detailed Implementations

Architectures of DQ-VAE & DQ-Transformer. DQ-VAE follows the official implementation of VQGAN [2] except for the proposed lightweight Dynamic Coding module. For the hierarchical encoder, we add two residual blocks followed by a down-sampling block to extract each feature map.

DQ-Transformer adopts a stack of causal self-attention blocks [6] for both Content-Transformer and Position-Transformer. We train DQ-Transformer with two different settings, *i.e.*, DQ-Transformer$_b$(base) with 6 layers Position-Transformer and 18 layers Content-Transformer of a total 308M parameters, and DQ-Transformer$_l$(large) with 6 layers Position-Transformer and 42 layers Content-Transformer of a total 608M parameters to demonstrate our scalability. The dimensionality of the DQ-Transformer is all set to be 1024. The number of heads used in the multi-head self-attention is 16. The probability of dropout is all set to be 0.1.

Training Details. MQ-VAE is trained with Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.9$ and the base learning rate is set to be 0.0000045 following [2]. The weight for adversarial loss is set to be 0.75 and the weight for perceptual loss is set to be 1.0. We do not use any other tricks such as the random restart of unused codes proposed in JukeBox [1] to increase the codebook usage, to give a fair comparison with VQGAN [2]. For FFHQ, MQ-VAE is trained for 150 epochs with a linear learning rate warmed up during the first 5 epochs. For ImageNet, MQ-VAE is trained for 50 epochs with a linear learning rate warmed up during the first 0.5 epochs.

The DQ-Transformer is trained using AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.95$. The weight decay is set to be 0.01. We use a cosine learning rate decay schedule with 0.0005 of the initial learning rate. The DQ-Transformer is trained for 100 epochs for FFHQ and ImageNet.

*Zhendong Mao is the corresponding author.

λ	rFID↓
1	3.6288
10	3.5311
100	3.7932

Table 1. Ablations of loss balance hype-parameter λ for MQ-VAE on FFHQ benchmark.

2. More Analysis of DQ-VAE

2.1. Impact of loss balance hyper-parameter λ

We first analyze the impact of the loss balance hype-parameter λ, which is used to balance the budget loss. As shown in Table 1, we conduct ablations on the FFHQ benchmark for MQ-VAE with dual granularities $\mathrm{F} = \{8, 16\}$ and $r_{f=8} = 0.5$. The models are trained for 50 epochs. We find that the reconstruction results are robust to λ and the best reconstruction quality is achieved when $\lambda = 10$. We further show the training and validation curves of budget loss and the ratio of finer granularity ($f = 8$). We find that the network converges very quickly and matches the expectation very well. The budget loss also drops very quickly.

2.2. More Visualization

We provide more visualization of our information-density-based variable-length coding in Figure 2 and Figure 3. We show that our coding map matches VQGAN's error map for both simple and complex images, *i.e.*, important regions are assigned to more codes and unimportant ones are assigned to few codes, leading to better reconstruction quality. We further provide more comparison of dynamic coding with different granularity ratios in Figure 4 and Figure 5. We show that our dynamic coding successfully assigns more codes for important regions with dense information

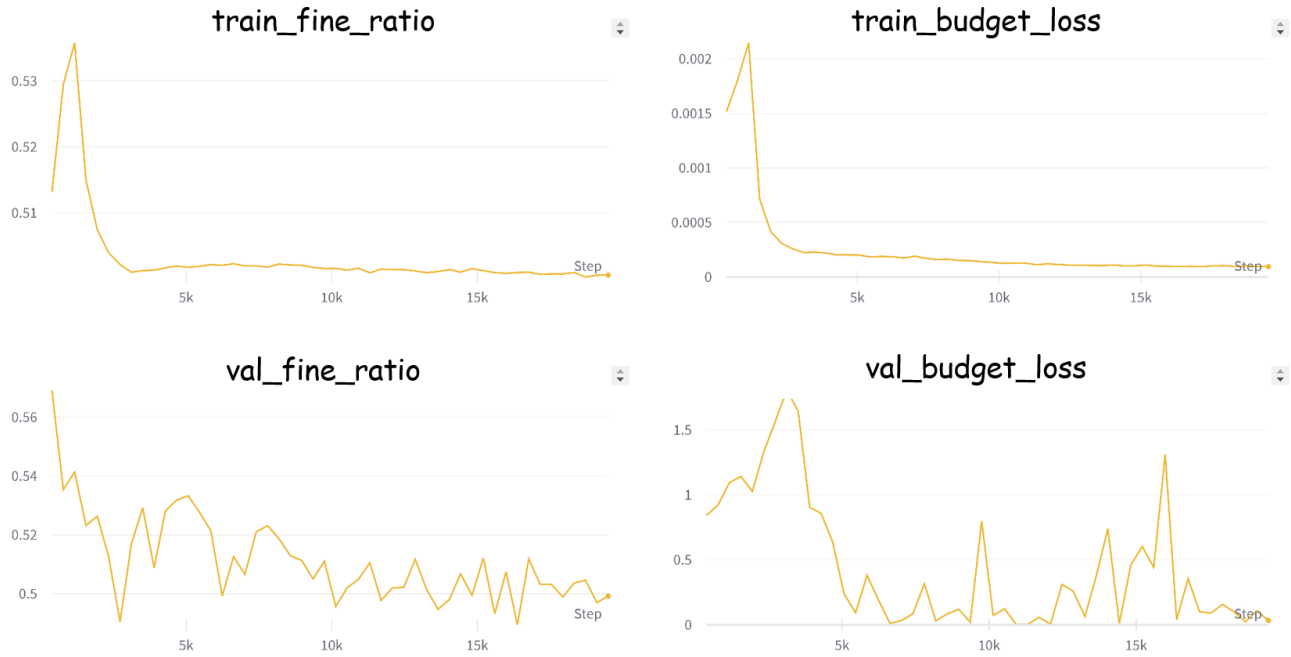

Figure 1. Detailed Training and Validation curves of MQ-VAE with dual granularities $\mathrm{F} = \{8, 16\}$ and $r_{f=8} = 0.5$, $\lambda = 10$.

density for all different granularities ratios.

3. More Analysis of DQ-Transformer

3.1. More results

Following previous works, we conduct more experiments on more datasets such as Celeb-HQ [3] (done by ViT-VQGAN [8]) and LSUN [7] (done by RQ-VAE [4]): on Celeb-HQ we outperform ViT-VQGAN (6.54 *vs*. 7.0); on LUSN-{church} we outperform RQVAE (6.21 *vs*. 7.45).

3.2. Different Methods for Dynamic Prior Learning

Learning the dynamic prior of our proposed DQ-VAE's variable-length coding is a new, challenging but promising task with great value. Except for the DQ-Transformer proposed in the main paper, we also evaluate many other types of structure and model design. To give a comprehensive understanding of our method, we briefly describe two other typical model designs for the interests of other researchers.

Global-Local Model Design. Inspired by [4], we propose to first model the sum of the codes in each image region and then model each code separately. That is, we first model each region globally and then model the codes in each region locally using two different transformers. This design could successfully deal with the irregular code map in each region and naturally support batch training and sampling. However, we find this design only results in very poor FID scores (around 24 on the FFHQ benchmark), and the performance gets even worse when more granularities are adopted.

Raster-Scan Dynamic Model Design. Different from the coarse-to-fine autoregression in the main paper, we also evaluate the traditional raster-scan order autoregression for DQ-VAE. To be specific, we first construct the content sequence in a raster-scan order. As for the position sequence, we construct a two-level position sequence, *i.e*., the first-level position sequence indicates the image region position of each code, and the second-level position sequence indicates the relative position of each code in the image region. We find this model design achieves slightly better generation quality compared to Global-Local Model Design but is still much worse than the proposed coarse-to-fine autoregression in the main paper (around 18 FID score on the FFHQ benchmark). Another vital shortcoming of this design is that it does not support batch sampling. The results are worse than the previous raster-scan autoregression of fixed-length coding because of the challenges of dynamic coding. The results are worse than our proposed coarse-to-fine autoregression which indicates the effectiveness of our coarse-grained to fine-grained generation order.

3.3. Impact of Different Content & Position Layers

We analyze the impact of different content and position layers for DQ-Transformer in Table 2. We find that more layers for content are more important than more layers for the position. The reason is that we find the position dis-

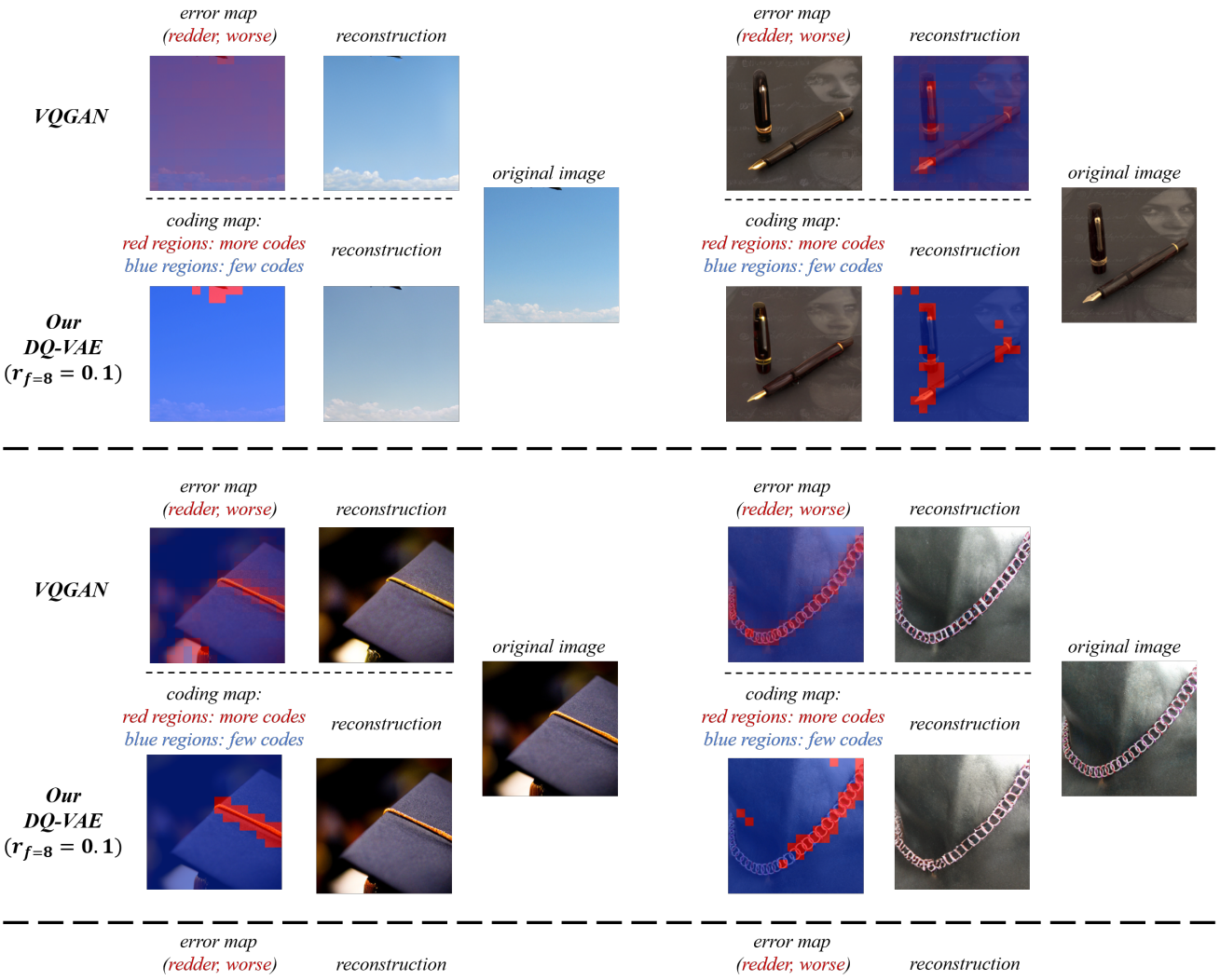

Figure 2. Visualization of the variable-length coding of our DQ-VAE, where ***our coding map exactly matches the error map of VQGAN*** and therefore leads to better reconstruction quality, *i.e.*, the ***information-dense regions*** where VQGAN has ***higher reconstruction error*** are assigned to ***more codes***, while ***information-sparse regions*** where VQGAN has ***lower reconstruction error*** are assigned to ***few codes***.

tribution is very easy to learn since the position loss drops very quickly to nearly zero during training. Moreover, previous autoregressive models [4, 5, 9, 10] always face the overfitting problem while we observe no overfitting for DQ-Transformer, which indicates that our dynamic coding provides a more general and effectiveness discrete representations for images.

Content Layer	Position Layer	FID↓
6	18	4.91
12	12	5.32
18	6	6.08

Table 2. Impact of Different Content & Position Layers on FFHQ benchmark.

References

[1] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020. 1

[2] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 1

[3] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. 2017. 2

[4] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and

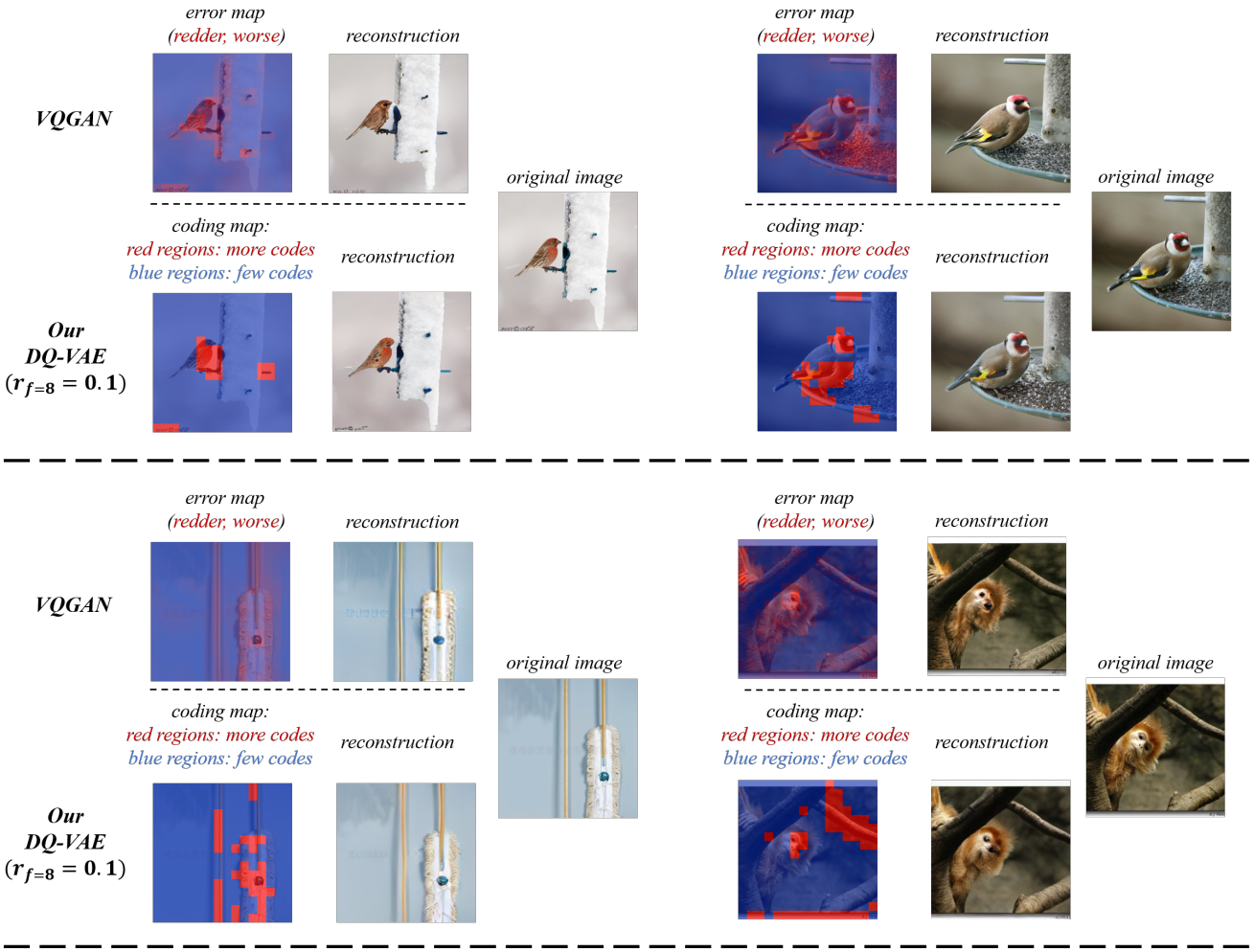

Figure 3. Visualization of the variable-length coding of our DQ-VAE, where ***our coding map exactly matches the error map of VQGAN*** and therefore leads to better reconstruction quality, *i.e.*, the ***information-dense regions*** where VQGAN has ***higher reconstruction error*** are assigned to ***more codes***, while ***information-sparse regions*** where VQGAN has ***lower reconstruction error*** are assigned to ***few codes***.

Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11523–11532, 2022. 2, 3

[5] Zhicong Tang, Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. Improved vector quantized diffusion models. *arXiv preprint arXiv:2205.16007*, 2022. 3

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[7] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *computer science*, 2015. 2

[8] J. Yu, X. Li, J. Y. Koh, H. Zhang, R. Pang, J. Qin, A. Ku, Y. Xu, J. Baldridge, and Y. Wu. Vector-quantized image modeling with improved vqgan. 2021. 2

[9] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021. 3

[10] Chuanxia Zheng, Long Tung Vuong, Jianfei Cai, and Dinh Phung. Movq: Modulating quantized vectors for high-fidelity image generation. *arXiv preprint arXiv:2209.09002*, 2022. 3

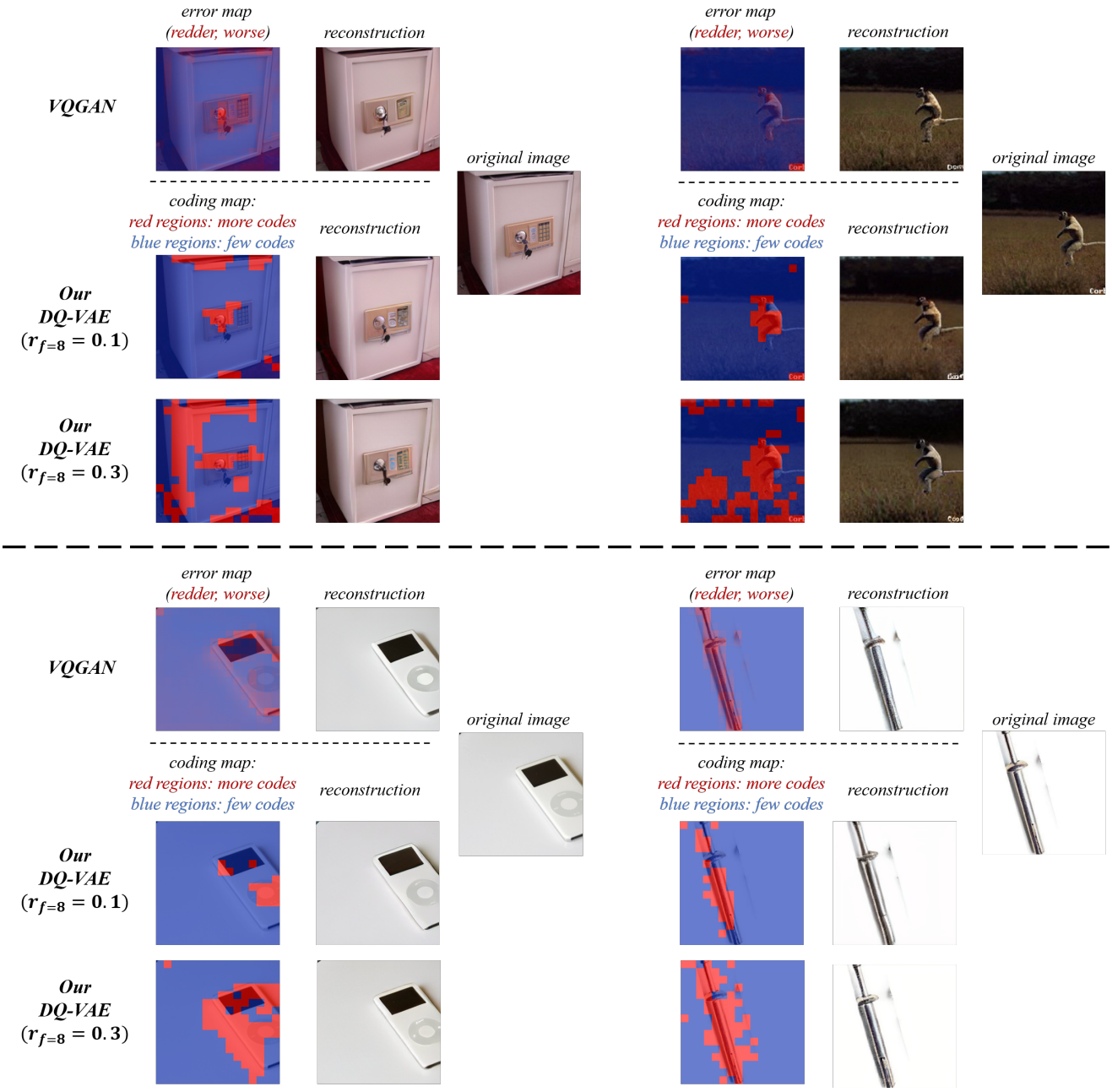

Figure 4. Visualization of the variable-length coding of our DQ-VAE, where ***our coding map exactly matches the error map of VQGAN*** and therefore leads to better reconstruction quality, *i.e.*, the ***information-dense regions*** where VQGAN has ***higher reconstruction error*** are assigned to ***more codes***, while ***information-sparse regions*** where VQGAN has ***lower reconstruction error*** are assigned to ***few codes***.

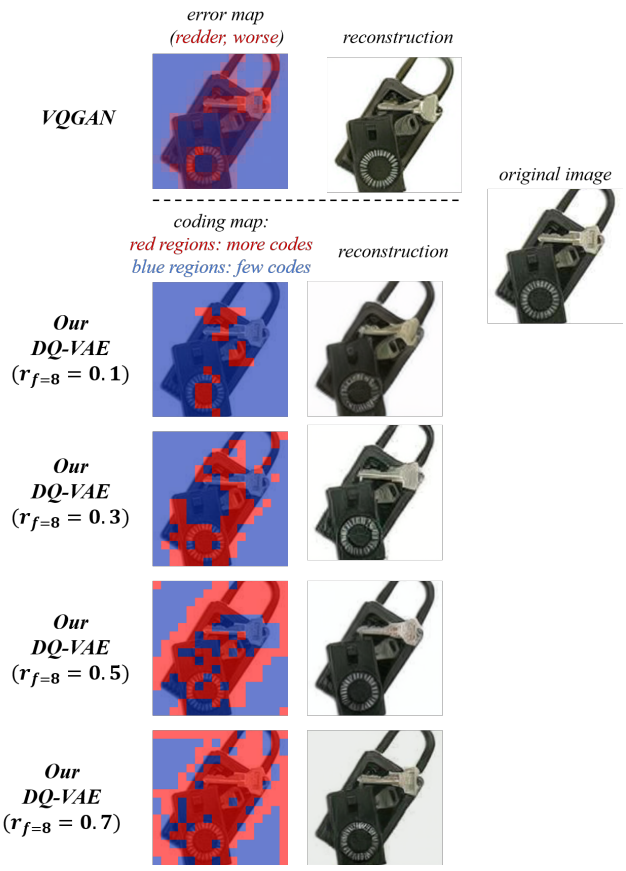

Figure 5. Visualization of the variable-length coding of our DQ-VAE, where ***our coding map exactly matches the error map of VQGAN*** and therefore leads to better reconstruction quality, *i.e.*, the ***information-dense regions*** where VQGAN has ***higher reconstruction error*** are assigned to ***more codes***, while ***information-sparse regions*** where VQGAN has ***lower reconstruction error*** are assigned to ***few codes***.