Rethinking the Objectives of Vector-Quantized Tokenizers for Image Synthesis Supplementary Material

Yuchao Gu[1], Xintao Wang[2], Yixiao Ge[2], Ying Shan[2], Mike Zheng Shou[1*]

[1]Show Lab, National University of Singapore [2]ARC Lab, Tencent PCG
https://github.com/TencentARC/BasicVQ-GEN

In this supplementary material, we first present detailed experimental settings in Sec. 1. Next, in Sec. 2, we offer additional visualization and analysis to further our understanding of the observations. We then present more qualitative results of our method in Sec. 3. Finally, we discuss the limitations of our approach and potential directions for future work in Sec. 4.

Input size	Encoder	Decoder
$f1$: 256×256	Conv, c-128 {Residual Block, 128-c} × 2 Downsample Block, 128-c	GN-Swish-Conv, c-3 {Residual Block, 128-c} × 2 + B&D Attn
$f2$: 128×128	{Residual Block, 128-c} × 2 Downsample Block, 256-c	Upsample Block, c-128 {Residual Block, 256-c} × 2 + B&D Attn
$f3$: 64×64	{Residual Block, 256-c} × 2 Downsample Block, 256-c	Upsample Block, c-256 {Residual Block, 256-c} × 2 + B&D Attn
$f4$: 32×32	{Residual Block, 256-c} × 2 Downsample Block, 512-c	Upsample Block, c-256 {Residual Block, 256-c} × 2 + B&D Attn
$f5$: 16×16	{Residual Block, 512-c} × 4 GN-Swish-Conv, 256-c	Upsample Block, c-256 {Residual Block, 512-c} × 4 + B&D Attn Conv, c-512

Table 1. Architecture of SeQ-GAN for 1st phase and 2nd phase tokenizer learning. The residual block consists of GN [12]-Swish [10]-Conv-GN-Swish-Conv. B&D Attn: interleaved block regional and dilated attention [15]; c: channels; f: compression ratio.

Model	#Params	#Blocks	#Heads	Model Dim	Hidden Dim	Dropout	#Tokens
AR/NAR	172M	24	16	768	3072	0.1	256
AR/NAR (Large)	305M	24	16	1024	4096	0.1	256

Table 2. Architecture of autoregressive (AR) and non-autoregressive (NAR) transformers. Both transformers share the same architecture, except for the causal attention used in the AR transformer.

1. Experimental Settings

Tokenizer learning. As shown in Table. 1, SeQ-GAN's architecture is based on VQGAN [4]. However, we modified the architecture in the first learning phase by removing the attention and constructing a convolution-only VQGAN. In the second learning phase, we enhanced the decoder with block regional and dilated attention (B&D Attn) to make attention suitable for high-resolution feature maps. SeQ-GAN has a total of 54.5M and 57.9M parameters for the first and second learning

*Corresponding Author.

phases, respectively. We use the style-based discriminator [9] for training SeQ-GAN, as suggested in VIT-VQGAN [14]. The hyperparameters used for training SeQ-GAN are summarized in Table. 3.

Generative transformer training. The autoregressive (AR) and non-autoregressive (NAR) transformers share the same architecture, except that the AR transformer adopts causal attention. As detailed in Table. 2, the AR/NAR transformers and their large variant have 172M and 305M parameters, respectively. We train both types of generative transformer using the hyperparameters listed in Table. 4. During sampling, we adopt the basic sampling techniques from VQGAN [4] (*i.e.*, top-p sampling [7]) and MaskGIT [2] (*i.e.*, adjusting sample temperature), while excluding the classifier-free guidance [6] and rejection sampling [11] for simplicity.

Detailed settings for observation and ablation experiments. Our observation (see Sec. 3.3 in the manuscript) and ablation experiments (see Sec. 4.4 in the manuscript) are conducted on the ImageNet [3] dataset. We mostly follow the same configurations as the benchmark experiments listed in Table. 3, with the exception that we use a batch size of 64 for SeQ-GAN learning. Based on each VQ tokenizer, we train the generative transformer on ImageNet with a batch size of 64 for 500,000 iterations, while keeping other settings the same as in Table. 4.

For Observation 1 (see Sec. 3.3.1), we evaluate different VQ tokenizers on various transformer configurations: 1) Different parameter sizes, including AR with 172M parameters and AR-Large with 305M parameters. 2) Different types of transformers, including both autoregressive and non-autoregressive transformers with 172M parameters. 3) Different training iterations, including AR-Large and AR-Large-2×, where we add an extra 500,000 iterations to the AR-Large model to investigate whether longer training eliminates the difference in VQ tokenizer.

	ImageNet	FFHQ	Cat	Bedroom	Church
Dataset Statistics					
Training Set	1,281,167	60,000	1,657,266	3,033,042	126,227
Validation Set	50,000	10,000	-	-	-
1st Phase of Tokenizer Learning					
Batch Size	256	64	64	64	32
Iterations	500,000	300,000	26,000	48,000	4,000
Epochs	100	320	1		
Learning Rate	1e-4		5e-5		
LR Decay	Cosine (end_lr=5e-5)		-		
Optimizer	Adam (β_1=0.9, β_2=0.99)				
2nd Phase of Tokenizer Learning					
Batch Size	128	32	64	64	32
Iterations	200,000				
Learning Rate	5e-5				
Optimizer	Adam (β_1=0.5, β_2=0.9)				

Table 3. Experimental setting of training SeQ-GAN on ImageNet [3], FFHQ [8], and LSUN [13]-{Cat, Bedroom, Church}.

2. More Visualization and Analysis of the Observations

In this section, we present additional visualizations to support our observations and proposed solutions.

First, we train SeQ-GAN with varying semantic ratios α and plot the validation loss curve for each corresponding generative transformer training in Fig. 2. Our results show that a larger semantic ratio α results in lower validation loss, indicating that generative transformers are better able to model the discrete space constructed by VQ tokenizers when more semantics are incorporated.

Next, we employ our proposed visualization pipeline to examine the reconstruction and AR prediction using SeQ-GAN with two different semantic ratios ($\alpha \in 0, 1$). Fig. 3 demonstrates that the generative transformer trained on SeQ-GAN (α=1) is better able to model each instance (*e.g.*, Row (a-c)), the semantic features (*e.g.*, the cat's face in Row (d) and the eagle's beak in Row (e)), and the structure (*e.g.*, the peaked roof in Row (f)). Note that compared to the SeQ-GAN (α=0) in Fig. 3, the reconstruction of the SeQ-GAN (α=1) loses some color fidelity and high-frequency details, leading to similar problems of lost details and spatial distortion in the generation results (see Fig. 1 (1st phase)).

Finally, to address the issue of lost details and spatial distortion resulting from removing shallow layers in $\mathcal{L}_{per}^{\alpha=1}$ during the first phase of tokenizer training, we use a two-phase tokenizer learning approach in our SeQ-GAN. In the second phase, we

	ImageNet	FFHQ	Cat	Bedroom	Church
Dataset Statistics					
Training Set	1,281,167	60,000	1,657,266	3,033,042	126,227
Validation Set	50,000	10,000	-	-	-
Autoregressive Transformer (AR)					
Batch Size	256	32	256	256	64
Iterations	1,500,000	500,000			
Optimizer	AdamW (β_1=0.9, β_2=0.96, weight_decay=1e-2)				
Learning Rate	1e-4				
LR Decay	Exponential (end_lr=5e-6, $start\_iter = 80,000$)				
Top-p Sampling	0.92	0.98			
Non-Autoregressive Transformer (NAR)					
Batch Size	256	32	256	256	64
Iterations	1,500,000	500,000			
Optimizer	AdamW (β_1=0.9, β_2=0.96, weight_decay=1e-2)				
Learning Rate	1e-4				
LR Decay	Linear (end_lr=0, $start\_iter = 50,000$)				
Sampling Temperature	0.45	0.65			

Table 4. Experimental setting of training generative transformers on ImageNet [3], FFHQ [8], and LSUN [13]-{Cat, Bedroom, Church}.

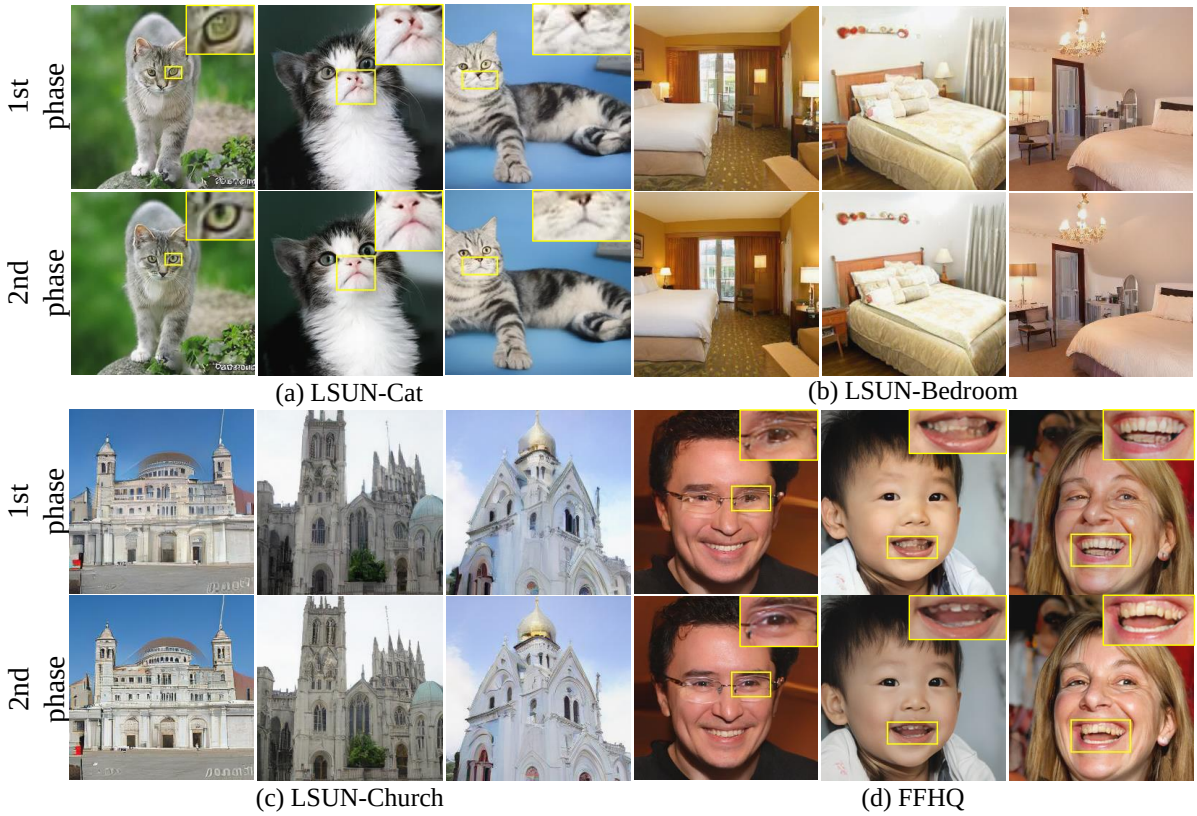

Figure 1. Influence of the 2nd phase tokenizer learning on generation results. (**Zoom in for best view.**)

finetune an enhanced decoder to restore the lost details. To demonstrate the effectiveness of our two-phase tokenizer learning on generation quality, we decode the transformer-sampled indices to image space using the decoder from both SeQ-GAN (1st phase) and SeQ-GAN (2nd phase), and present the generation results in Fig. 1. Our visualization clearly shows that

SeQ-GAN (2nd phase) preserves more details and enhances generation quality compared to SeQ-GAN (1st phase).

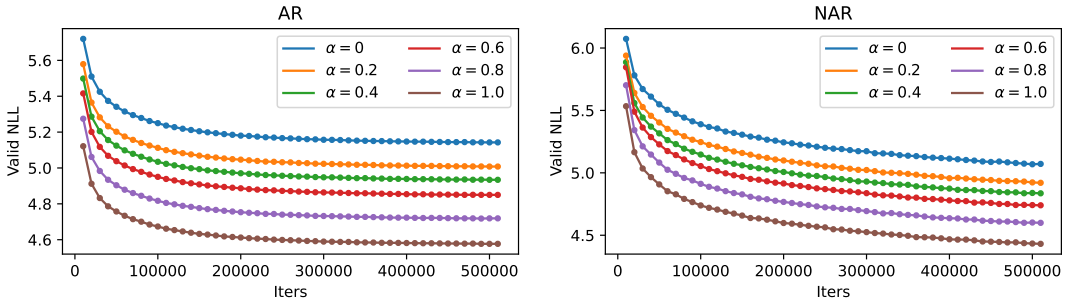

Figure 2. Validation loss curves of generative transformers training on ImageNet. Generative transformers are built upon the SeQ-GAN tokenizers with different semantic ratios (α).

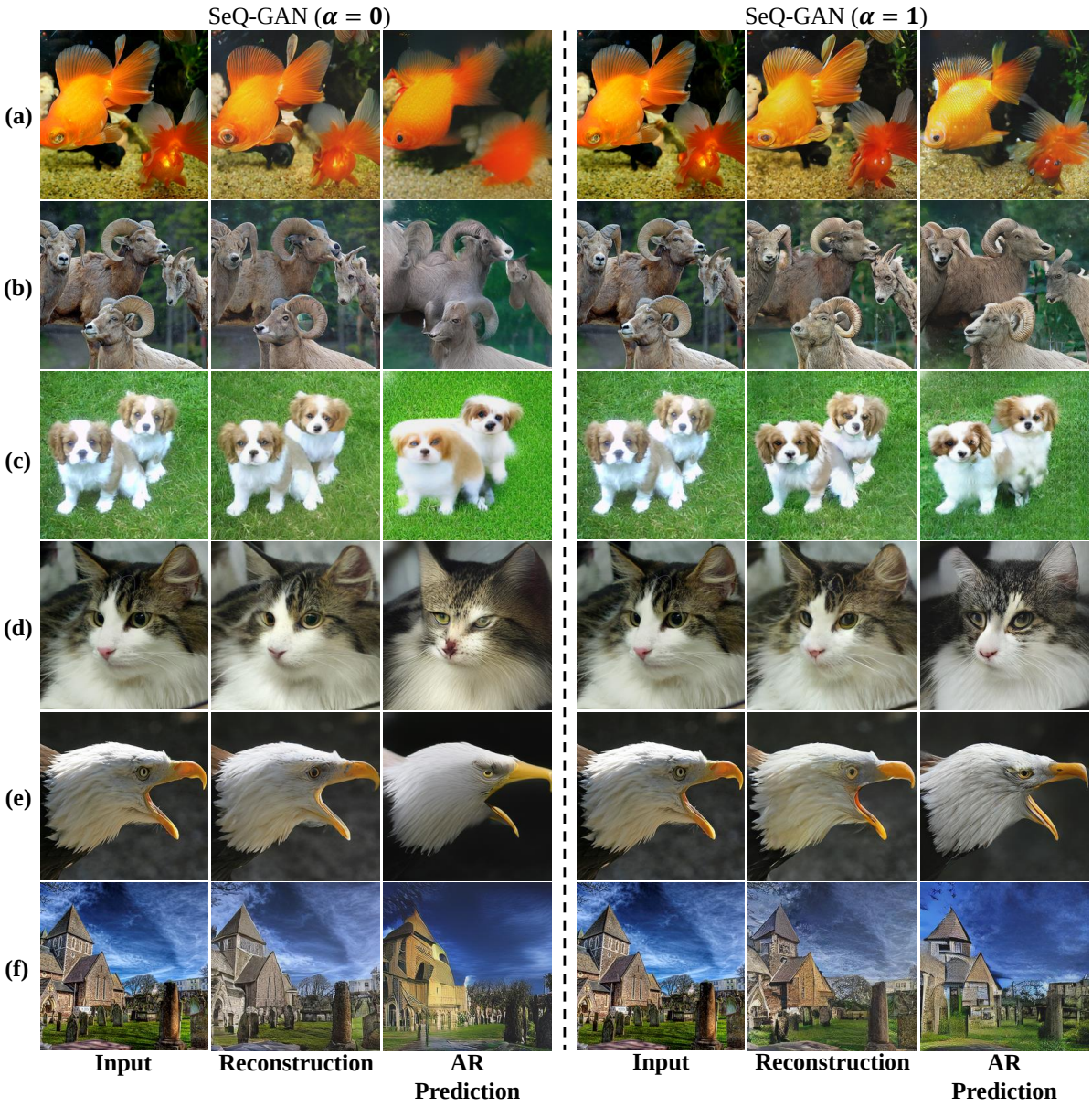

Figure 3. Visual comparison of the influence of SeQ-GAN with different semantic ratios ($\alpha \in 0, 1$) on image reconstruction and AR prediction. Compared to SeQ-GAN (α=0), the AR transformer built on SeQ-GAN (α=1) better models each instance (*e.g.*, Row (a-c)), semantic features (*e.g.*, the cat's face in Row (d) and the eagle's beak in Row (e)), and structure (*e.g.*, the peaked roof in Row (f)).

3. More Qualitative Results

We provide qualitative comparisons to BigGAN [1], VQGAN [4] and MaskGIT [2] in Fig. 4, Fig. 5 and Fig. 6. For MaskGIT and BigGAN, the samples are extracted from the paper and for VQGAN, we use their pre-generated samples in the official codebase[1]. Our SeQ-GAN+NAR produces results with better quality and diversity than previous methods. From the uncurated results in Fig. 7, Fig. 8, Fig. 9 and Fig. 10, our SeQ-GAN+NAR can generate images with high quality and diversity on unconditional image generation.

[1] https://github.com/CompVis/taming-transformers

4. Limitation and Future Work

4.1. Future Work

Our observation 1 indicates that the quality of the discrete latent space in VQ-based generative models cannot be directly assessed by reconstruction fidelity, as the reconstruction and generation have different optimization goals. Thus, future work could design more intuitive methods to evaluate the quality of the discrete latent space.

In addition, observation 2 highlights the importance of semantics in the discrete latent space for visual synthesis. We have kept our approach simple by controlling the semantic ratio through the modification of the perceptual loss. Future works on VQ tokenizers can explore more effective ways to balance semantic compression and details preservation. For example, contrastive learning may improve the semantics compression of VQ tokenizers.

4.2. Limitation

Our method has the limitation inherited from likelihood-based generative models. While techniques such as rejection sampling [11] and classifier-free guidance [6] can be used to filter out samples with bad shapes and improve sample quality in conditional image generation, there are few sampling techniques available for improving unconditional image generation. Classifier-based metrics such as FID tend to focus more on textures than overall shapes, and thus may not be consistent with human perception, as pointed out in [5]. To address this, StyleGAN2 introduces the perceptual path length (PPL) metric [8], which is more related to the shape quality of samples. StyleGAN2 regularizes the GAN training to favor lower PPL. Although generative transformers with the SeQ-GAN tokenizer can achieve a better FID than StyleGAN2 in unconditional image generation, some samples still have poor overall shapes (as seen in the uncurated samples in Fig. 8). Therefore, an interesting area for future research is to investigate the design of sampling techniques for unconditional image generation in likelihood-based generative models to improve overall shape quality.

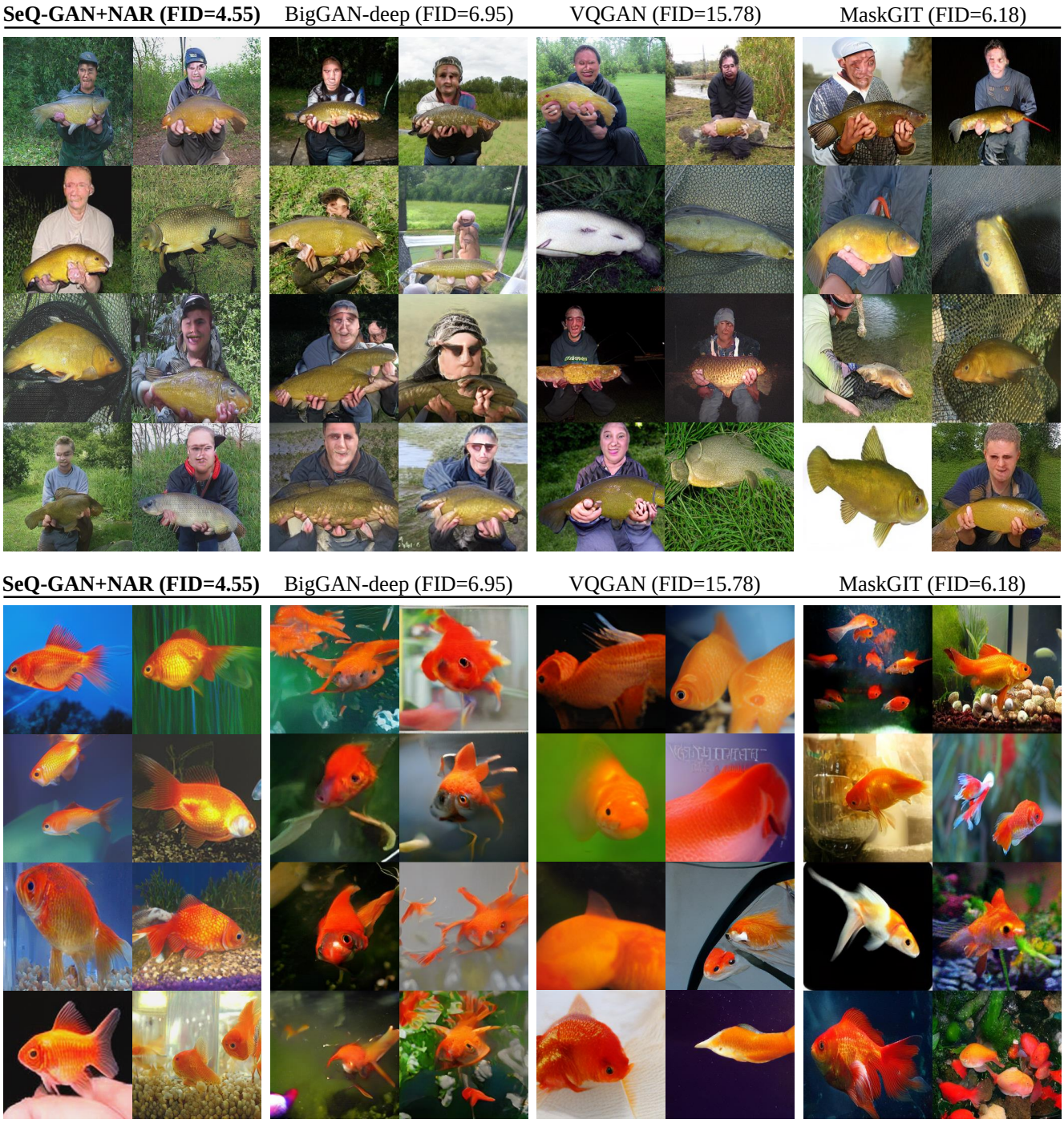

Figure 4. Qualitative comparison with BigGAN [1], VQGAN [4] and MaskGIT [2] on the class 0 (tench) and class 1 (glodfish) of ImageNet [3].

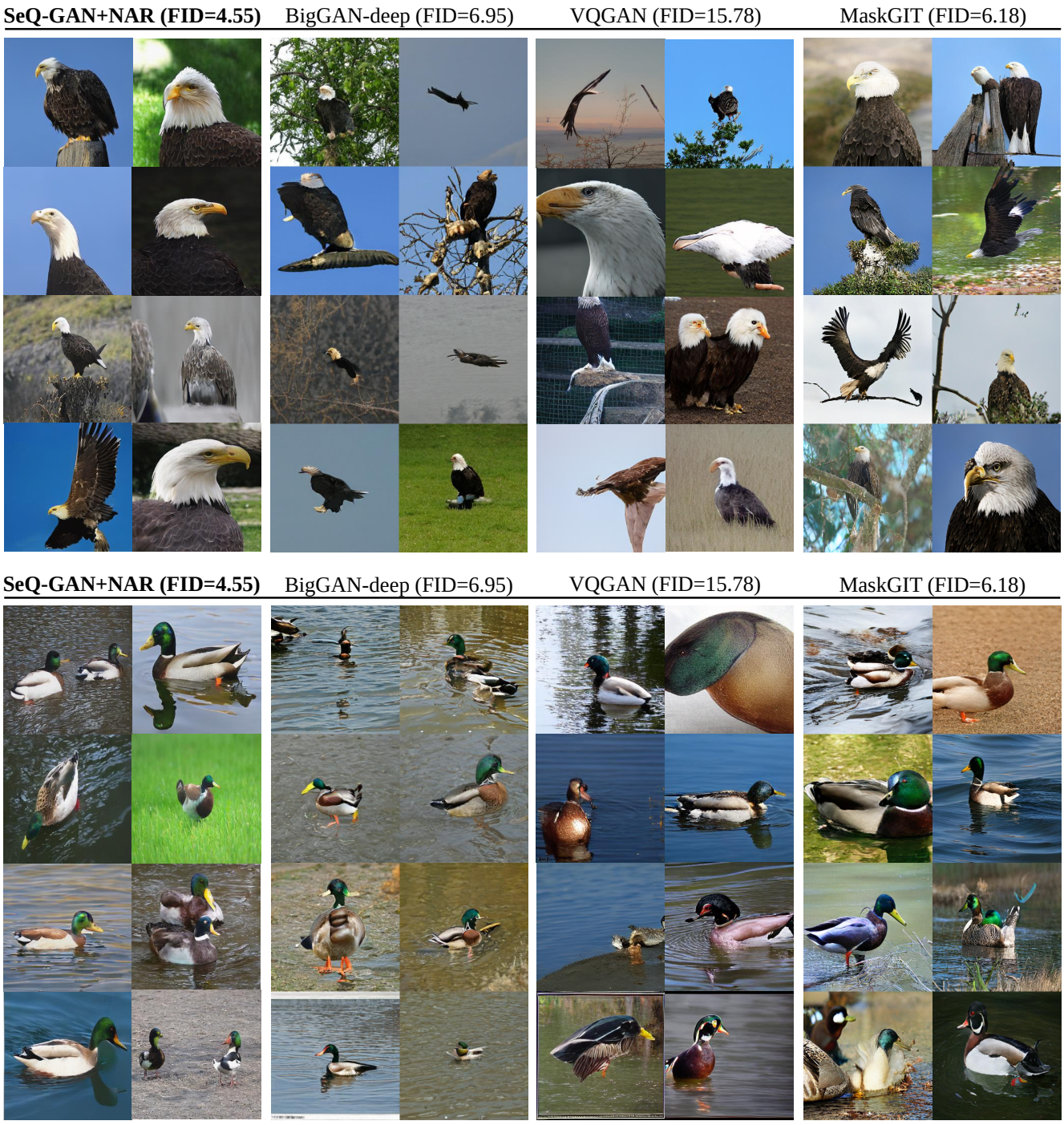

Figure 5. Qualitative comparison with BigGAN [1], VQGAN [4] and MaskGIT [2] on the class 22 (bald eagle) and class 97 (drake) of ImageNet [3].

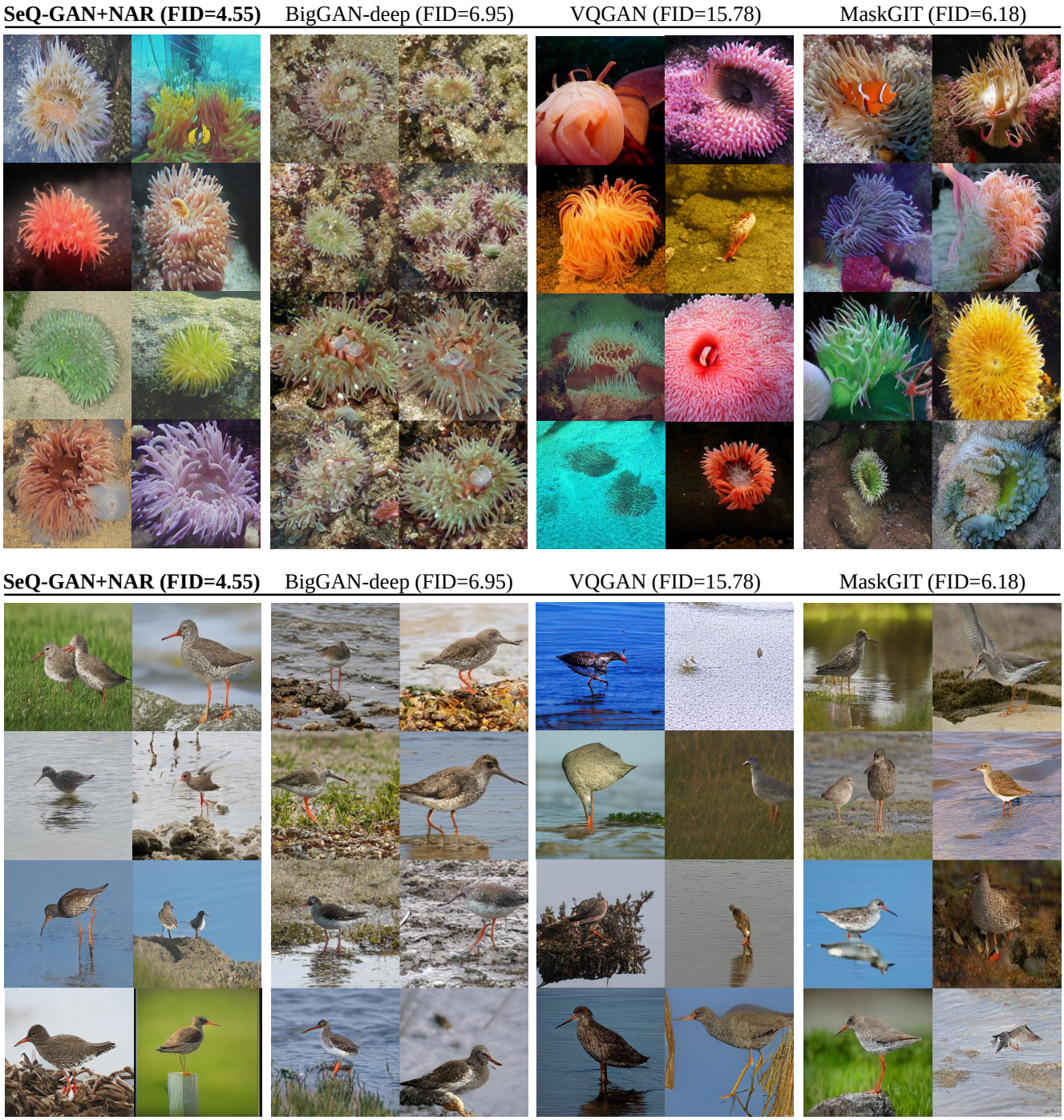

Figure 6. Qualitative comparison with BigGAN [1], VQGAN [4] and MaskGIT [2] on the class 108 (sea anemone) and class 141 (redshank) of ImageNet [3].

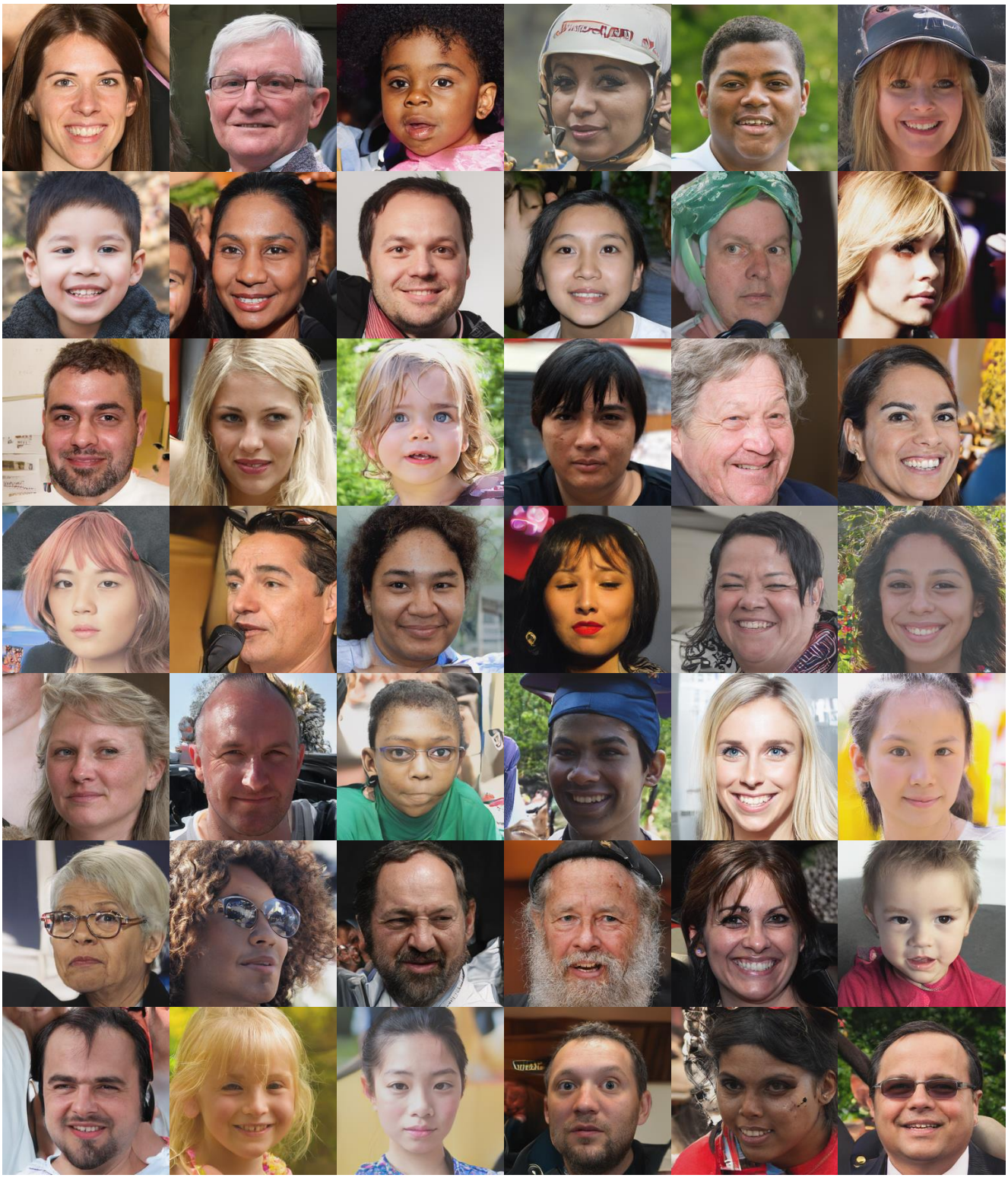

Figure 7. Uncurated set of samples of SeQ-GAN+NAR on 256×256 FFHQ.

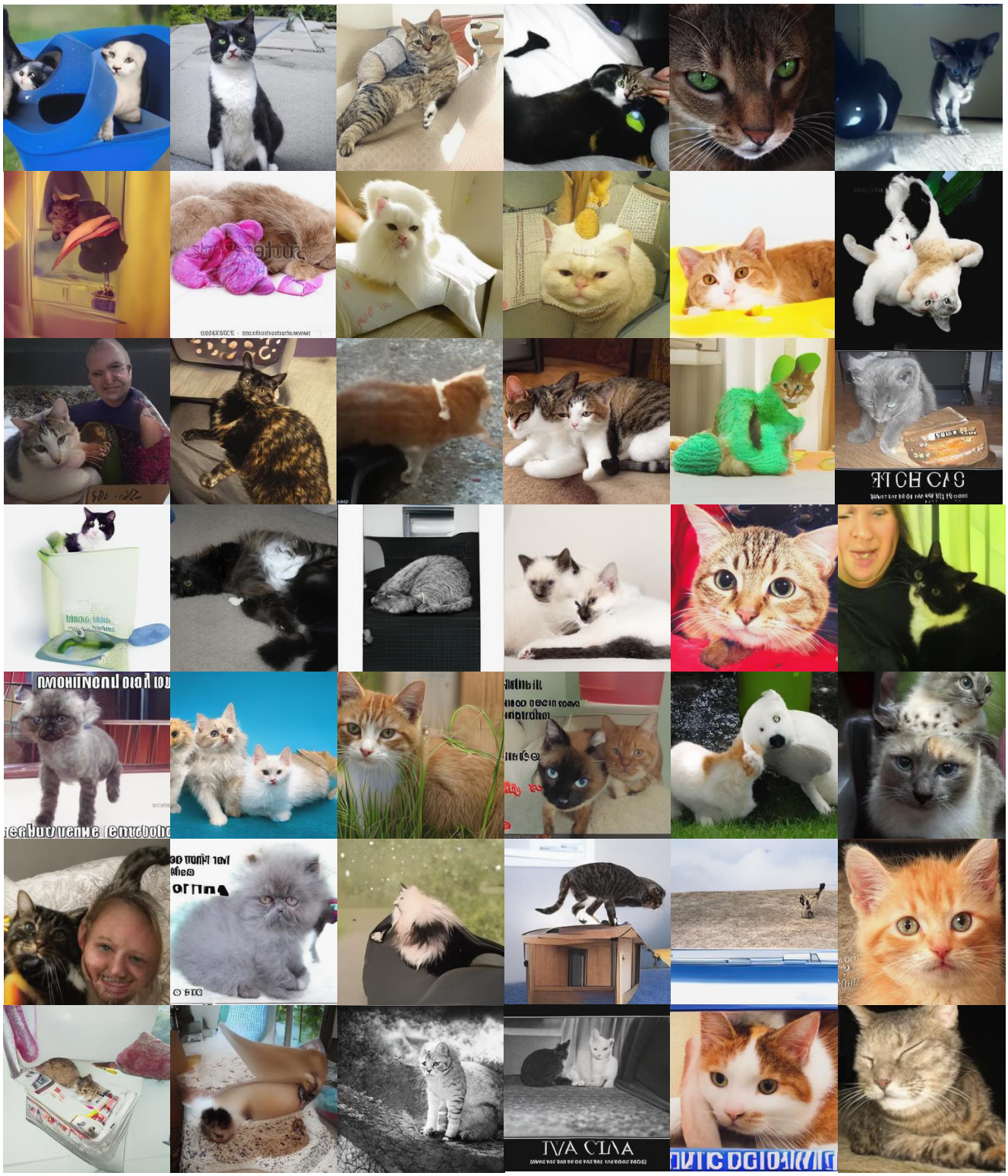

Figure 8. Uncurated set of samples of SeQ-GAN+NAR on 256×256 LSUN cat.

Figure 9. Uncurated set of samples of on 256×256 LSUN church.

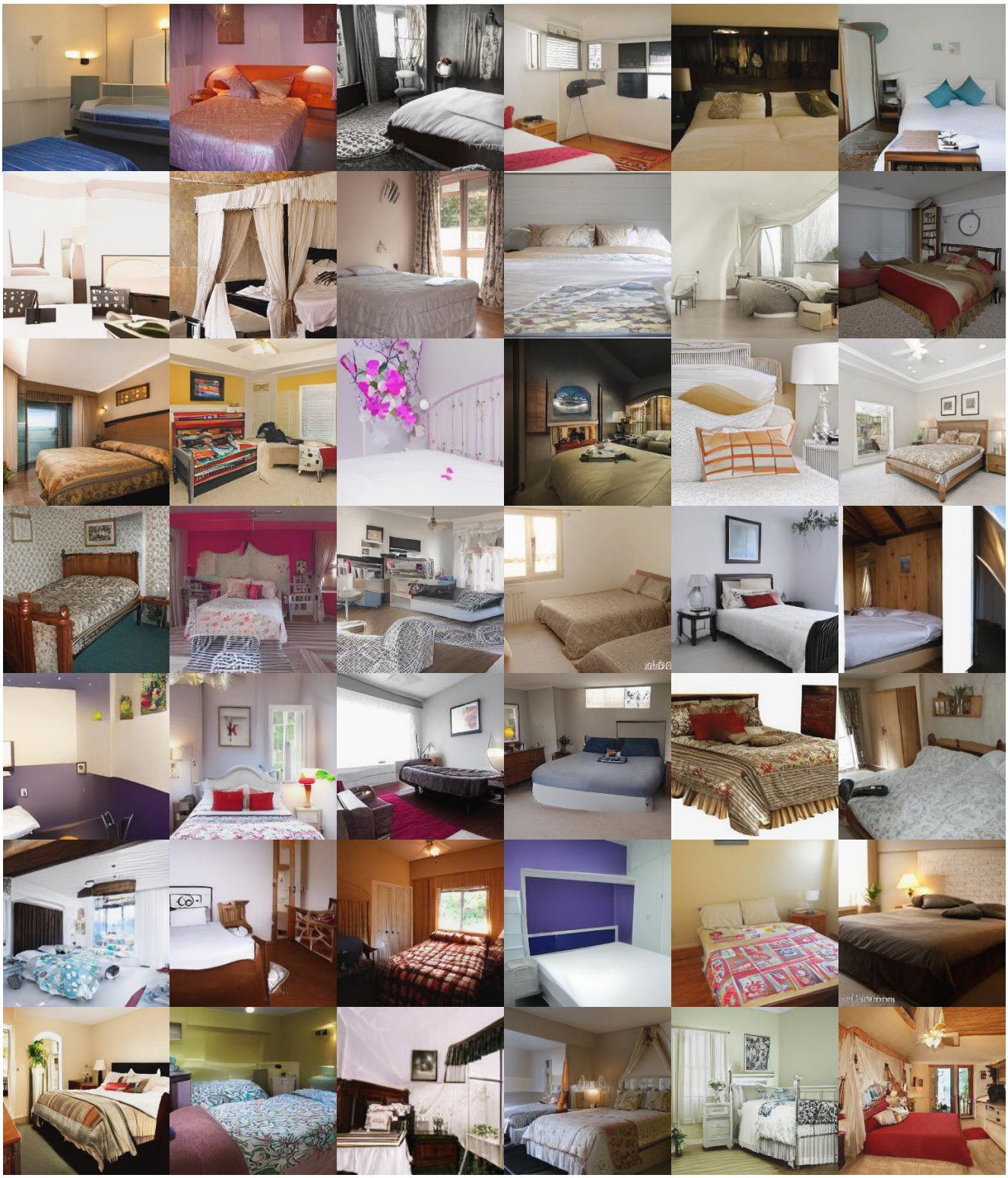

Figure 10. Uncurated set of samples of SeQ-GAN+NAR on 256×256 LSUN bedroom.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 5, 7, 8, 9

[2] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *CVPR*, pages 11315–11325, 2022. 2, 5, 7, 8, 9

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 2, 3, 7, 8, 9

[4] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 1, 2, 5, 7, 8, 9

[5] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 6

[6] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 6

[7] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019. 2

[8] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 2, 3, 6

[9] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020. 2

[10] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017. 1

[11] 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*, pages 8821–8831. PMLR, 2021. 2, 6

[12] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, pages 3–19, 2018. 1

[13] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 2, 3

[14] 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. 2

[15] Long Zhao, Zizhao Zhang, Ting Chen, Dimitris Metaxas, and Han Zhang. Improved transformer for high-resolution gans. *NeurIPS*, 34:18367–18380, 2021. 1