Supplementary Material for
DiG: Scalable and Efficient Diffusion Models with Gated Linear Attention

Lianghui Zhu [1,2,◇] Zilong Huang [2 ✉] Bencheng Liao [1] Jun Hao Liew [2] Hanshu Yan [2]
Jiashi Feng [2] Xinggang Wang [1 ✉]

[1] School of EIC, Huazhong University of Science & Technology [2] ByteDance
Code & Models: hustvl/DiG

ImageNet 512 × 512					
Model	FID↓	sFID↓	IS↑	P↑	R↑
DiT-XL/2 (400K)	20.94	6.78	66.30	0.74	0.58
DiG-XL/2 (400K)	**17.36**	**6.12**	**69.42**	**0.75**	**0.63**

Table 1. Comparing the proposed DiG against DiT on Imagenet 512 × 512 benchmark.

1. Comparison on Larger Image Size

We additionally compare the performance of our plain DiG-XL/2 and DiT-XL/2 on ImageNet 512 × 512 class-conditional image generation. As shown in Table 1, DiG-XL/2 outperforms DiT-XL/2 under the same training iterations.

2. Details of Different Scanning Strategies

As mentioned in Section 3.3, traditional multi-path scanning methods of causal modeling often lead to numerical instability and complex extra operations. Algorithm 1, Algorithm 2, and Algorithm 3 present the details of different scanning methods, respectively. "**GLA**" is the scanning operator. "**flip**", "**reshape2d**", "**transpose**", "**flatten**", and "+" are the matrix operators. It can be seen that bidirectional scanning and 4-directional scanning require many extra matrix operations and scanning operations.

3. Additional Visualizations

We also present additional visualizations of DiG-XL/2 with resolutions of 256 × 256 and 512 × 512 in Figure 3- 14.

Algorithm 1: Bidirectional Scanning.

```
Input: token sequence $\mathbf{z}_{in}$ : $(\mathtt{B}, \mathtt{T}, \mathtt{D})$
Output: token sequence $\mathbf{z}_{out}$ : $(\mathtt{B}, \mathtt{T}, \mathtt{D})$
$\mathbf{z}'_{out,1}$ : $(\mathtt{B}, \mathtt{T}, \mathtt{D}) \leftarrow \mathbf{GLA}(z_{in})$
/* Extra Operations */
$\mathbf{z}_{in,2}$ : $(\mathtt{B}, \mathtt{T}, \mathtt{D}) \leftarrow \mathbf{flip}(z_{in})$
$\mathbf{z}'_{out,2}$ : $(\mathtt{B}, \mathtt{T}, \mathtt{D}) \leftarrow \mathbf{GLA}(z_{in,2})$
/* Flip the sequence to ensure the
   same direction as $\mathbf{z}_{out,1}$ */
$\mathbf{z}'_{out,2\to1}$ : $(\mathtt{B}, \mathtt{T}, \mathtt{D}) \leftarrow \mathbf{flip}(\mathbf{z}'_{out,2})$
$\mathbf{z}_{out}$ : $(\mathtt{B}, \mathtt{T}, \mathtt{D}) \leftarrow \mathbf{z}'_{out,1} + \mathbf{z}'_{out,2\to1}$
Return: $\mathbf{z}_{out}$
```

4. Additional Experiments

Experiments for higher-resolutions. We further show qualitative and quantitative results on higher resolutions in Fig. 1 and Table 2, respectively. We follow the previous work [2] to generate higher-resolution images. Table 2 shows that DiG consistently outperforms DiT baseline on higher resolutions, which indicates that DiG presents better scaling up ability to longer sequences. Specifically, at 1536 × 1536 resolution (sequence length **9216**), DiG achieves both better quality (FID 30.51 vs 102.27) and higher efficiency (5.0× faster). Fig. 1 further demonstrates DiG's capability in generating high-quality images at higher resolutions. Overall, our work wanted to show the potential of DiG in dealing with long sequences of image generation, we also leave video generation with extremely long sequences as future work.

Visualizations for Ablations. We have provided quantitative results in Table 3, the 1st experiment of "***Ours.***"

[◇] This work was done when Lianghui Zhu was interning at ByteDance. ✉ Corresponding authors: Xinggang Wang (xgwang@hust.edu.cn) and Zilong Huang (zilong.huang2020@gmail.com)

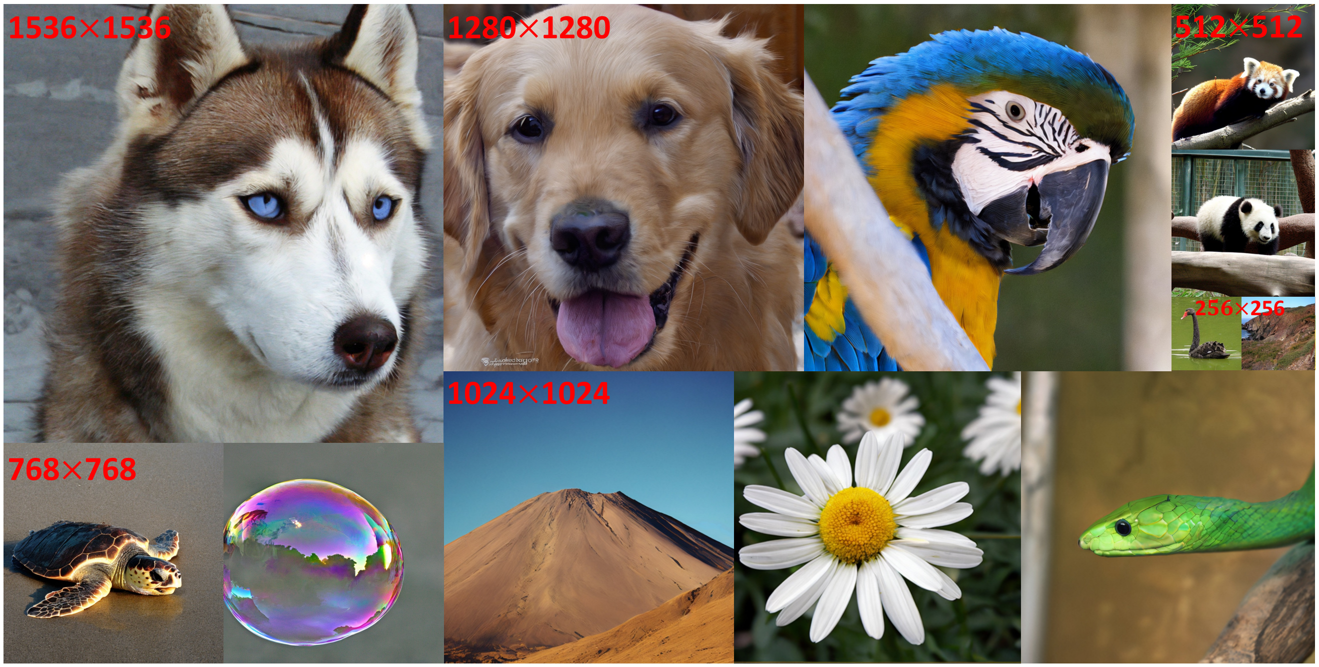

Figure 1. Image generation with the proposed Diffusion Gated Linear Attention Transformers (DiG). We use upsample guidance [2] to generate images from 768 × 768 to 1536 × 1536 resolution. Noting the sequence length of 1536 × 1536 resolution is $\left(\frac{1536}{8\times 2}\right)^2 = 9216$.

Algorithm 2: 4-directional Scanning.

```
Input: token sequence $\mathbf{z}_{in}$ : $(\mathtt{B}, \mathtt{T}, \mathtt{D})$
Output: token sequence $\mathbf{z}_{out}$ : $(\mathtt{B}, \mathtt{T}, \mathtt{D})$
$\mathbf{z}'_{out,1}$ : $(\mathtt{B}, \mathtt{T}, \mathtt{D}) \leftarrow \mathbf{GLA}(\mathbf{z}_{in})$
/* Extra Operations */
$\mathbf{z}_{in,2}$ : $(\mathtt{B}, \mathtt{T}, \mathtt{D}) \leftarrow \mathbf{flip}(\mathbf{z}_{in})$
$\mathbf{z}'_{out,2}$ : $(\mathtt{B}, \mathtt{T}, \mathtt{D}) \leftarrow \mathbf{GLA}(\mathbf{z}_{in,2})$
/* Flip the sequence to ensure the
   same direction as $\mathbf{z}_{out,1}$ */
$\mathbf{z}'_{out,2\to 1}$ : $(\mathtt{B}, \mathtt{T}, \mathtt{D}) \leftarrow \mathbf{flip}(\mathbf{z}'_{out,2})$
$\mathbf{z}_{in,3}$ : $(\mathtt{B}, \mathtt{T}, \mathtt{D}) \leftarrow$
  $\mathbf{flatten}(\mathbf{transpose}(\mathbf{reshape2d}(\mathbf{z}_{in})))$
$\mathbf{z}'_{out,3}$ : $(\mathtt{B}, \mathtt{T}, \mathtt{D}) \leftarrow \mathbf{GLA}(\mathbf{z}_{in,3})$
$\mathbf{z}'_{out,3\to 1}$ : $(\mathtt{B}, \mathtt{T}, \mathtt{D}) \leftarrow$
  $\mathbf{flatten}(\mathbf{transpose}(\mathbf{reshape2d}(\mathbf{z}'_{out,3})))$
$\mathbf{z}_{in,4}$ : $(\mathtt{B}, \mathtt{T}, \mathtt{D}) \leftarrow \mathbf{flip}(\mathbf{z}_{in,3})$
$\mathbf{z}'_{out,4}$ : $(\mathtt{B}, \mathtt{T}, \mathtt{D}) \leftarrow \mathbf{GLA}(\mathbf{z}_{in,4})$
$\mathbf{z}'_{out,4\to 1}$ : $(\mathtt{B}, \mathtt{T}, \mathtt{D}) \leftarrow$
  $\mathbf{flatten}(\mathbf{transpose}(\mathbf{reshape2d}(\mathbf{flip}(\mathbf{z}'_{out,3}))))$
$\mathbf{z}_{out}$ : $(\mathtt{B}, \mathtt{T}, \mathtt{D}) \leftarrow$
  $\mathbf{z}'_{out,1} + \mathbf{z}'_{out,2\to 1} + \mathbf{z}'_{out,3\to 1} + \mathbf{z}'_{out,4\to 1}$
Return: $\mathbf{z}_{out}$
```

Algorithm 3: Block-by-Block Scanning.

```
Input: token sequence $\mathbf{z}_{in}$ : $(\mathtt{B}, \mathtt{T}, \mathtt{D})$
Output: token sequence $\mathbf{z}_{out}$ : $(\mathtt{B}, \mathtt{T}, \mathtt{D})$
$\mathbf{z}'_{out}$ : $(\mathtt{B}, \mathtt{T}, \mathtt{D}) \leftarrow \mathbf{GLA}(z_{in})$
/* Extra Operations */
if $l$ % $2 == 0$ then
    $\mathbf{z}_{out}$ : $(\mathtt{B}, \mathtt{T}, \mathtt{D}) \leftarrow \mathbf{flip}(\mathbf{z}'_{out})$
end
else
    $\mathbf{z}_{out}$ : $(\mathtt{B}, \mathtt{T}, \mathtt{D}) \leftarrow$
      $\mathbf{flatten}(\mathbf{transpose}(\mathbf{reshape2d}(\mathbf{z}'_{out})))$
end
Return: $\mathbf{z}_{out}$
```

Method	512×512	768×768	1024×1024	1280×1280	1536×1536
DiT	21.35	24.54	28.78	54.03	102.27
DiG	18.29	19.83	20.42	23.75	30.51

Table 2. FID-10K comparisons between DiT and DiG.

demonstrates only the unidirectional scanning leads to poor performance, i.e., a FID-50K score of 175.84. The DW-Conv2D provides local awareness, which can further improve the FID from 69.28 (2nd experiment) to 63.84 (4th experiment). We also provide qualitative results in Fig. 2, which shows removing local awareness leads to distortion of details, i.e., the waves are not real enough. If we further remove multi-path scanning and turn it into unidirectional scanning, the performance will drop further.

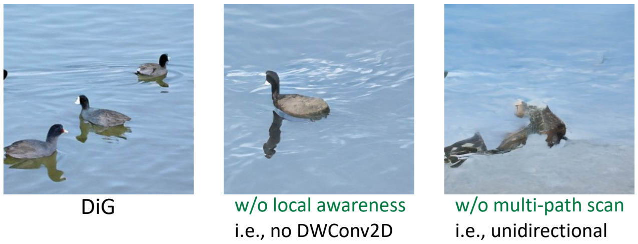

Figure 2. Ablation of unidirectional scanning and local awareness.

5. Discussion about Vision/Diffusion Mamba

Vision Mamba [3] and Diffusion Mamba (DiS) [1] are pioneer works of state-space models in vision and diffusion fields. While they focus on bidirectional state space models and lack of local awareness, our DiG proposes the SREM for block-wise scanning control and local awareness. With comparable FID scores (Table 2), DiG further presents higher efficiency for scaling up, as shown in Table 1, Fig. 2, and Fig. 3.

References

[1] Zhengcong Fei, Mingyuan Fan, Changqian Yu, and Junshi Huang. Scalable diffusion models with state space backbone. *arXiv preprint arXiv:2402.05608*, 2024. 3

[2] Juno Hwang, Yong-Hyun Park, and Junghyo Jo. Upsample guidance: Scale up diffusion models without training. *arXiv preprint arXiv:2404.01709*, 2024. 1, 2

[3] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *International Conference on Machine Learning*, pages 62429–62442. PMLR, 2024. 3

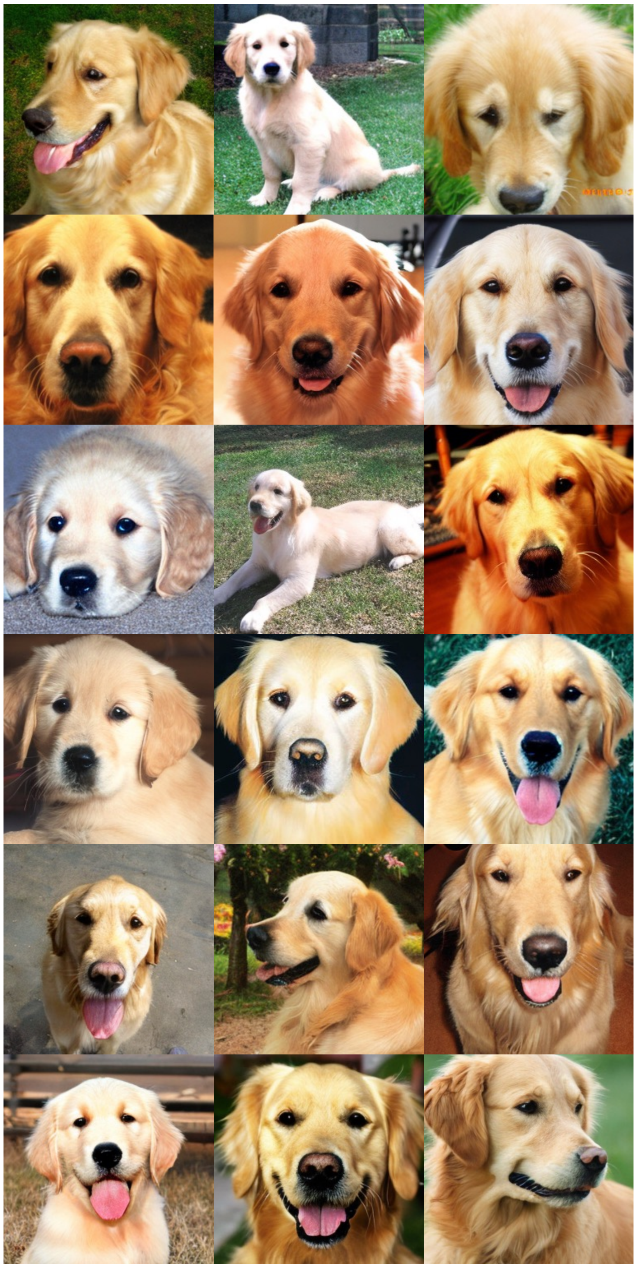

Figure 3. **Uncurated** 256×256 **DiG-XL/2 samples**.
Classifier-free guidance scale = 4.0
Class label = “golden retriever” (207)

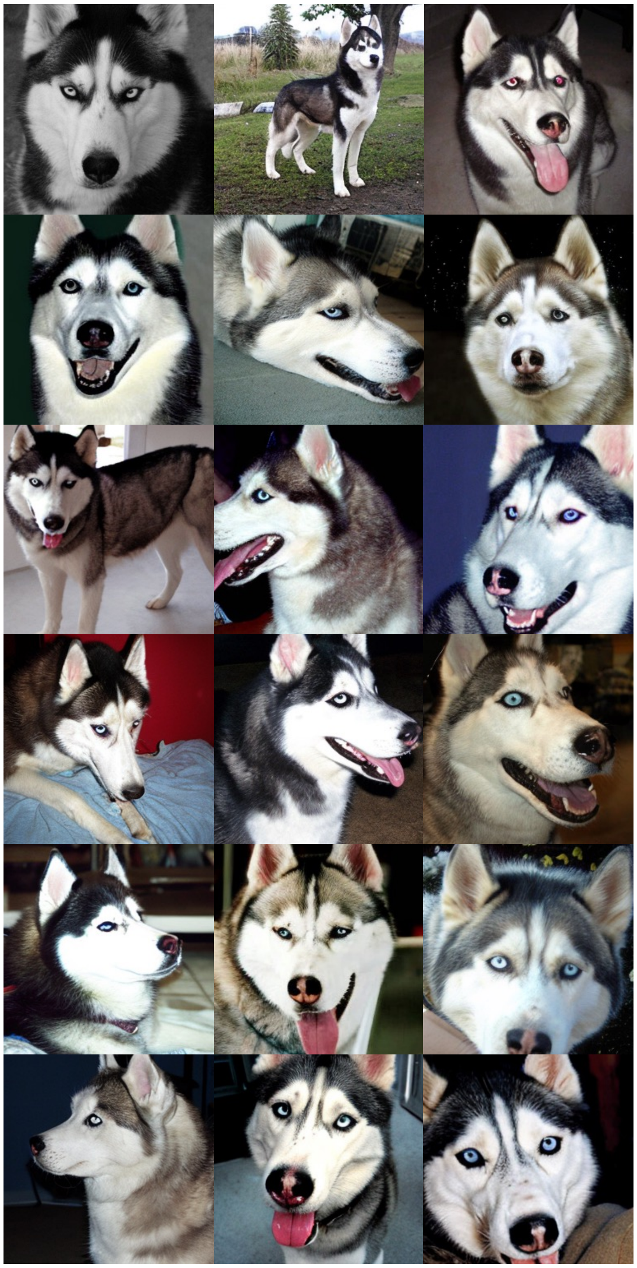

Figure 4. **Uncurated** 256×256 **DiG-XL/2 samples**.
Classifier-free guidance scale = 4.0
Class label = “husky” (250)

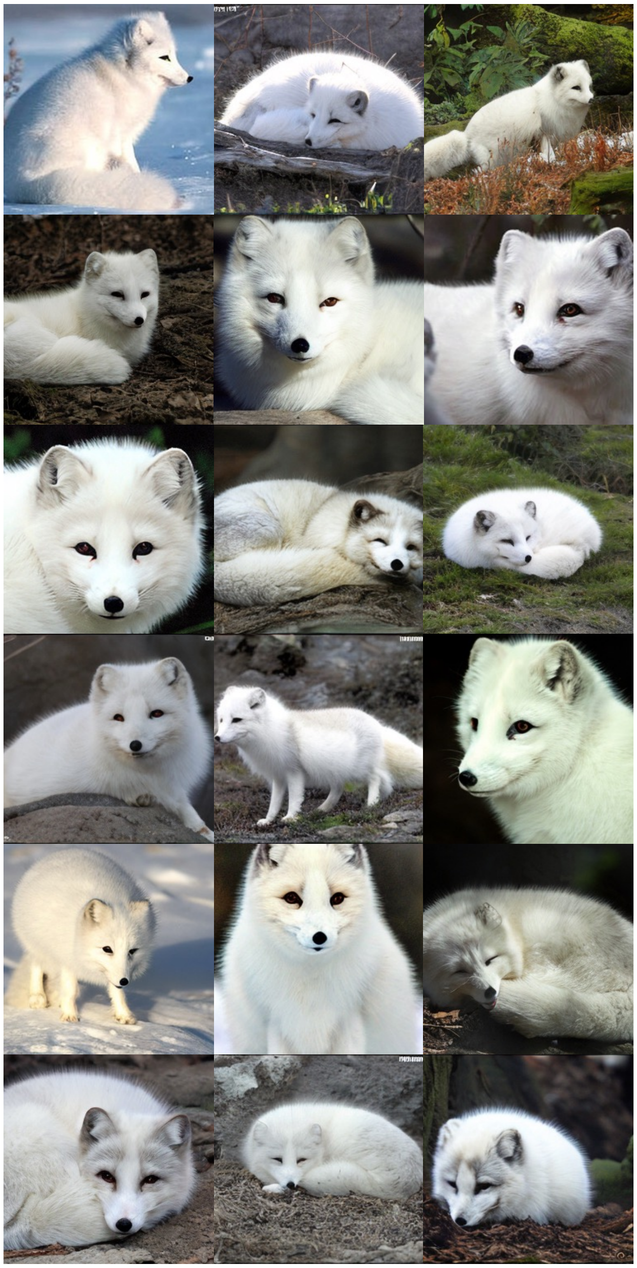

Figure 5. **Uncurated** 256×256 **DiG-XL/2 samples**.
Classifier-free guidance scale = 4.0
Class label = “arctic fox” (279)

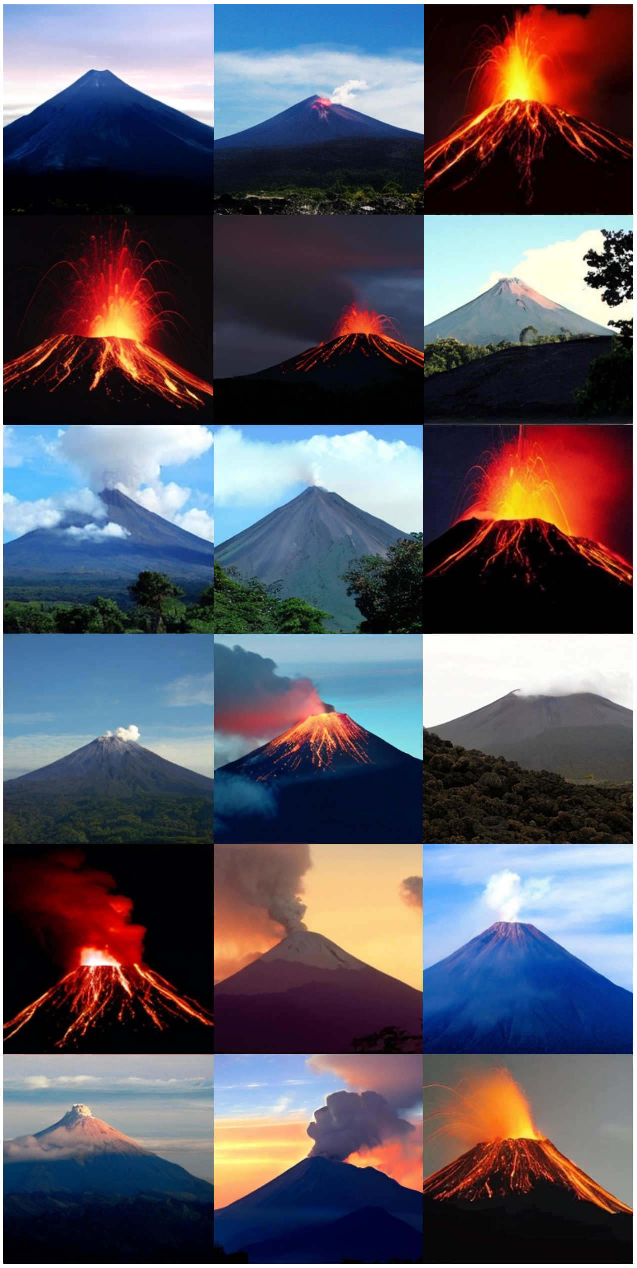

Figure 6. **Uncurated** 256×256 **DiG-XL/2 samples**.
Classifier-free guidance scale = 4.0
Class label = “volcano” (980)

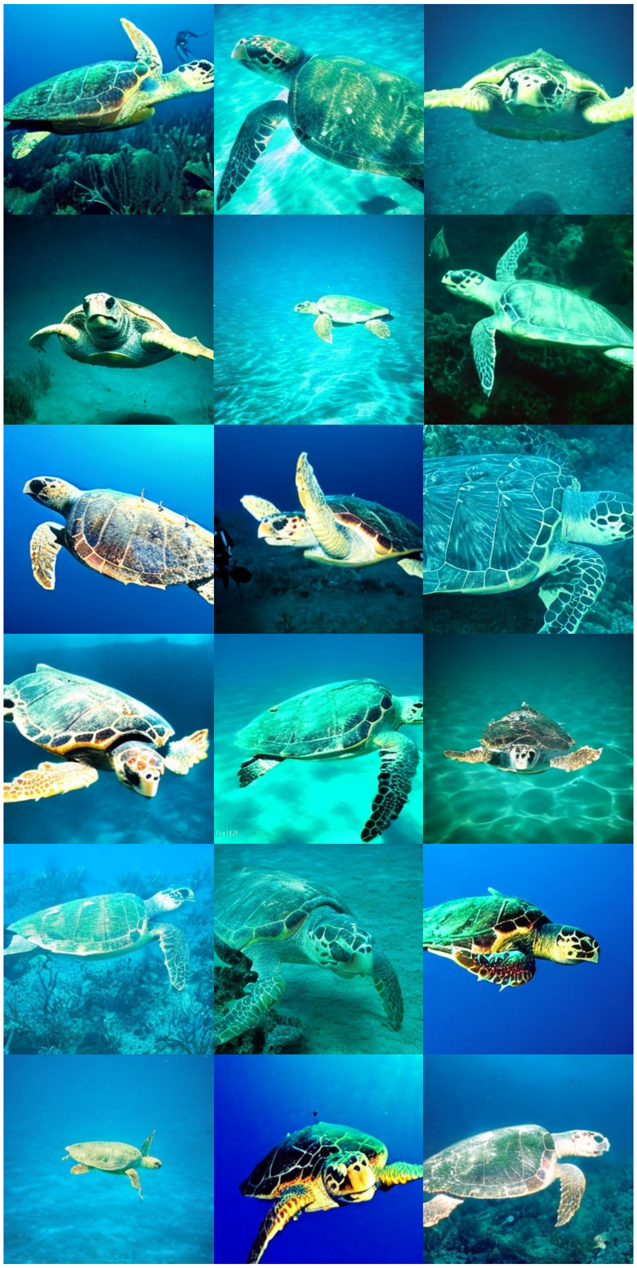

Figure 7. **Uncurated** 256×256 **DiG-XL/2 samples**.
Classifier-free guidance scale = 4.0
Class label = “loggerhead sea turtle” (33)

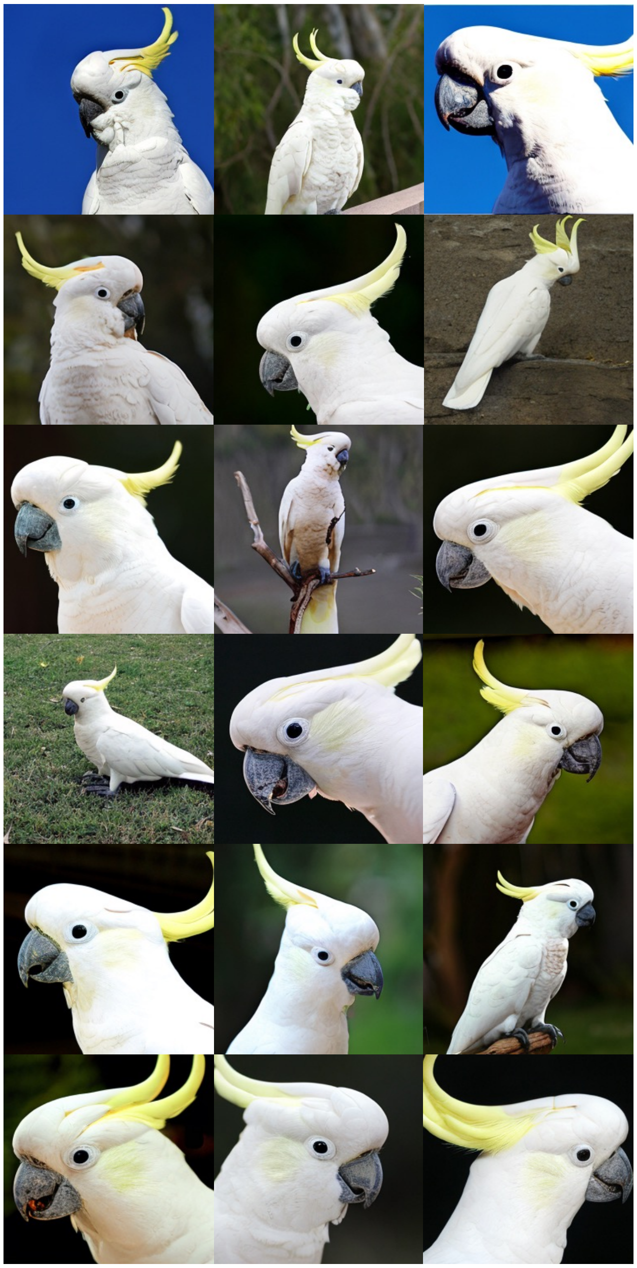

Figure 8. **Uncurated** 256×256 **DiG-XL/2 samples**.
Classifier-free guidance scale = 4.0
Class label = “sulphur-crested cockatoo” (89)

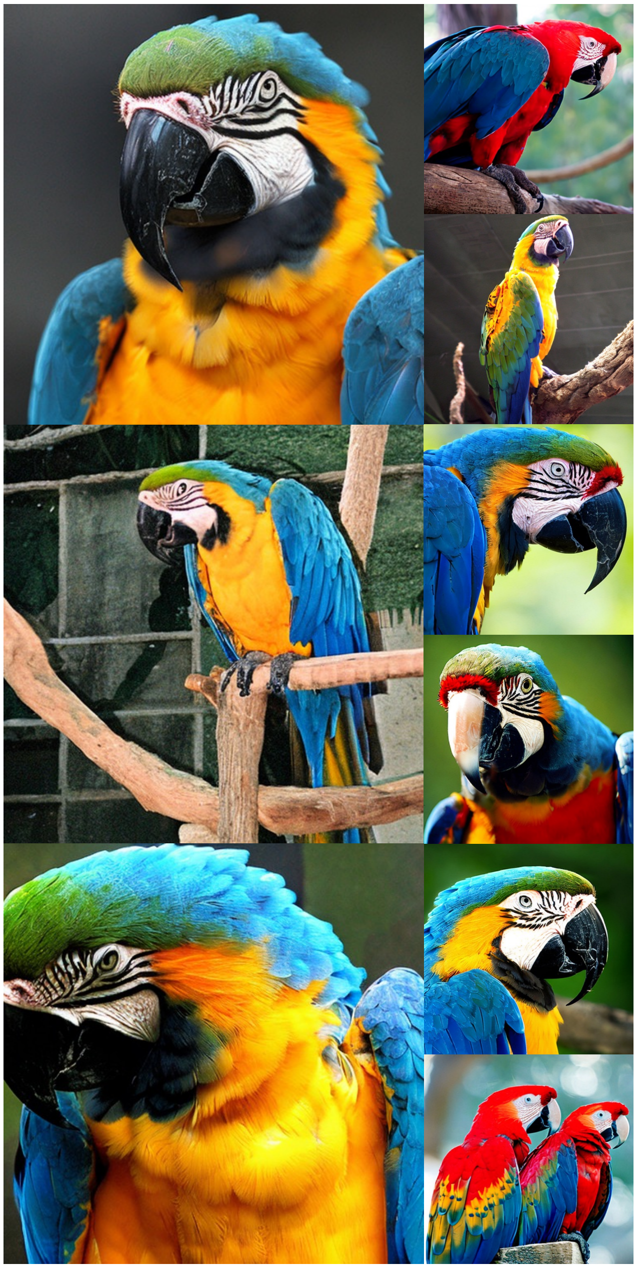

Figure 9. **Uncurated** 512 × 512 **DiG-XL/2 samples**.
Classifier-free guidance scale = 4.0
Class label = “macaw” (88)

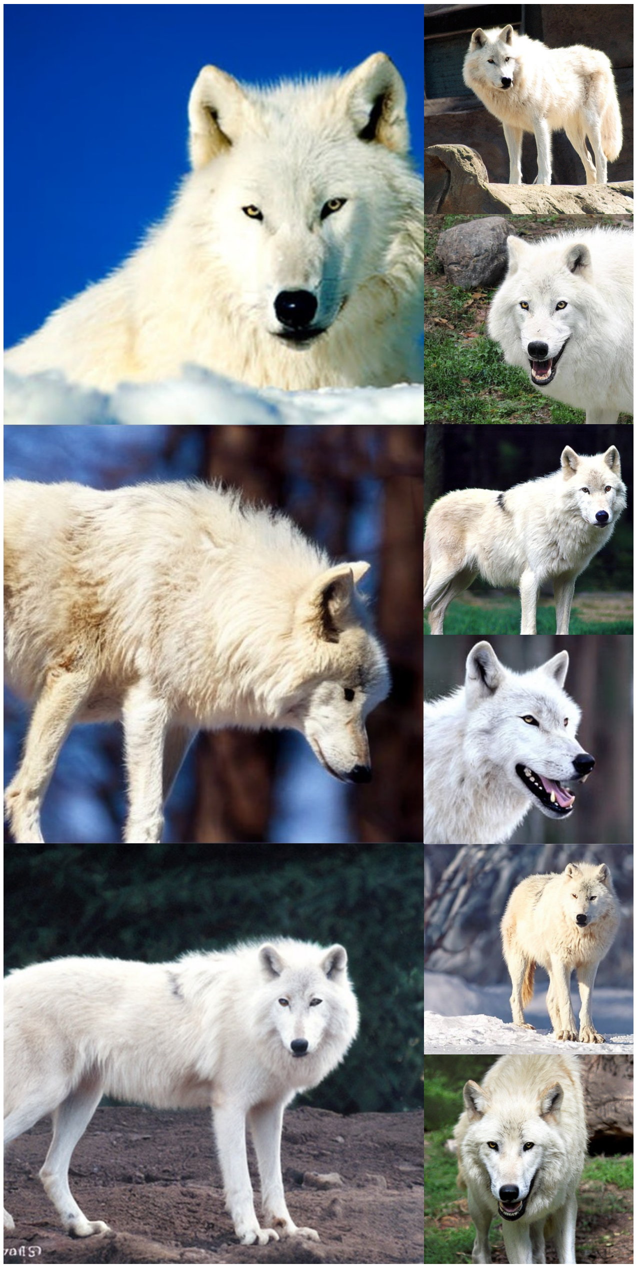

Figure 10. **Uncurated** 512 × 512 **DiG-XL/2 samples**.
Classifier-free guidance scale = 4.0
Class label = “arctic wolf” (270)

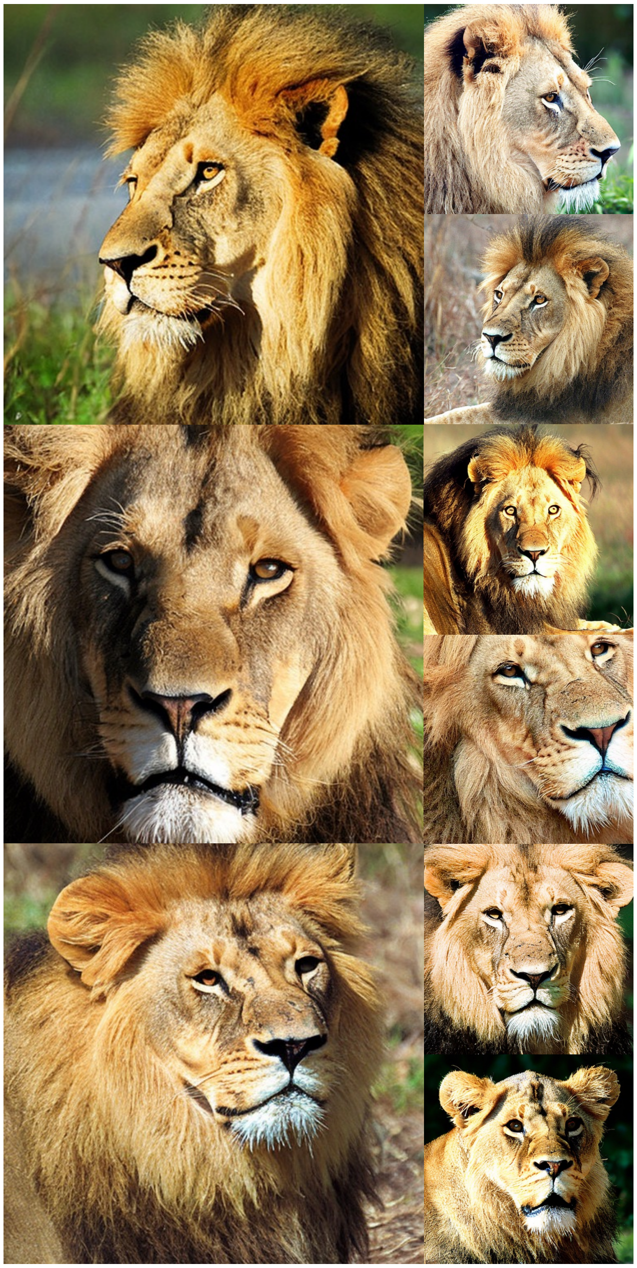

Figure 11. **Uncurated** 512 × 512 **DiG-XL/2 samples**.
Classifier-free guidance scale = 4.0
Class label = “lion” (291)

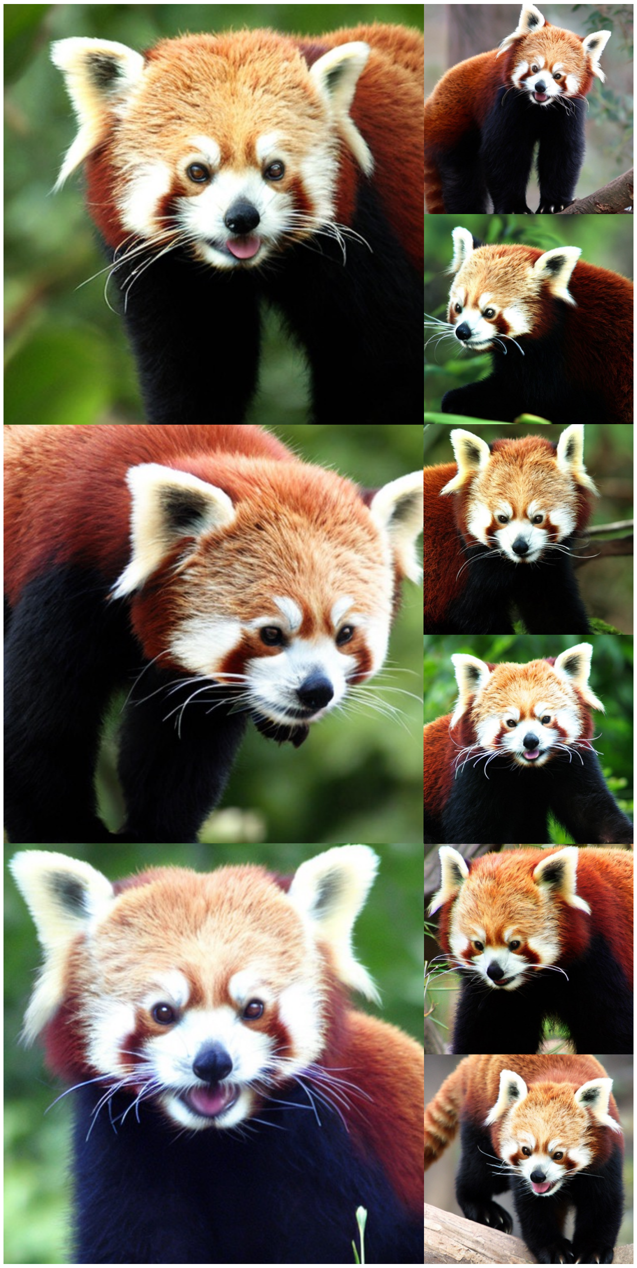

Figure 12. **Uncurated** 512 × 512 **DiG-XL/2 samples**.
Classifier-free guidance scale = 4.0
Class label = “red panda” (387)

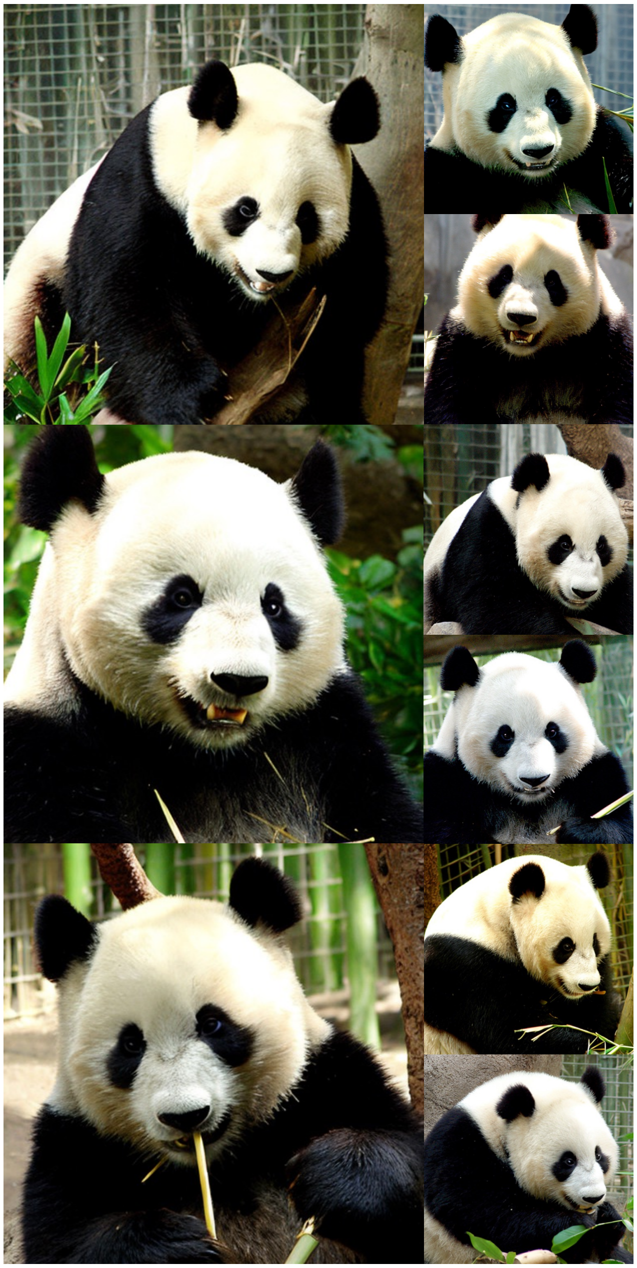

Figure 13. **Uncurated** 512×512 **DiG-XL/2 samples**.
Classifier-free guidance scale = 4.0
Class label = “panda” (388)

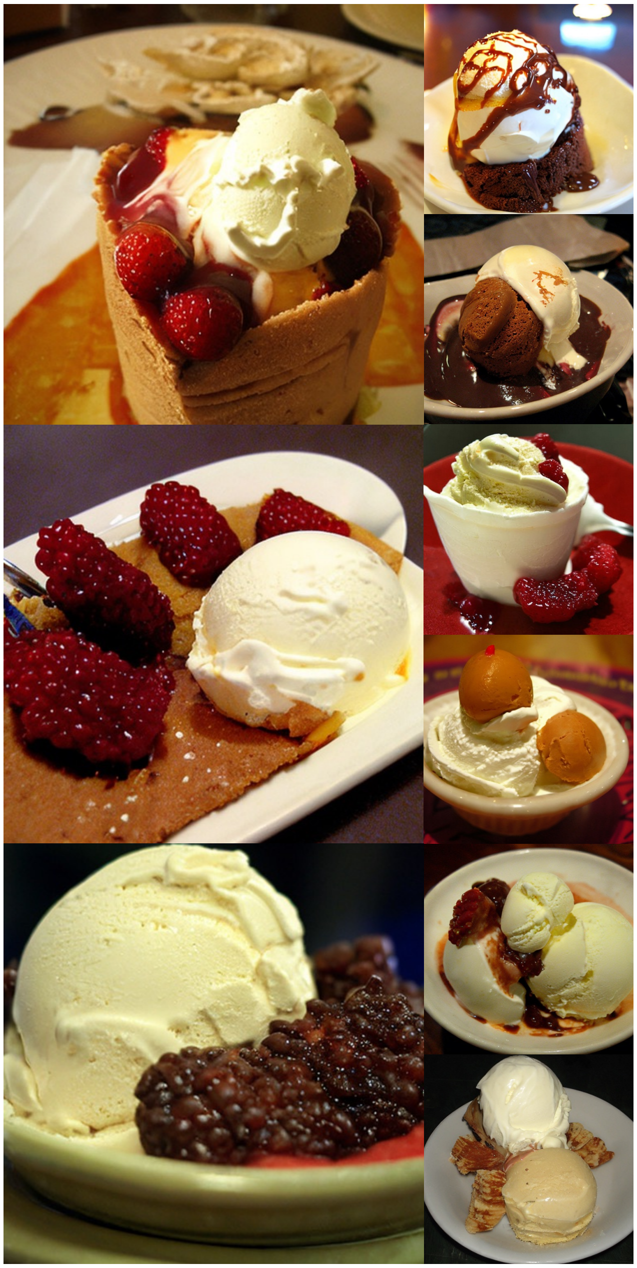

Figure 14. **Uncurated** 512×512 **DiG-XL/2 samples**.
Classifier-free guidance scale = 4.0
Class label = “ice cream” (928)