Figure 11: **Additional selected samples from our 512×512 and 256×256 resolution DiT-XL/2 models.** We use the ft-EMA VAE decoder and classifier-free guidance scales of 6.0 and 4.0 for the 512×512 and 256×256 models, respectively.

A. Additional Implementation Details

We include detailed information about all of our DiT models in Table 4, including both 256×256 and 512×512 models. In Figure 13, we report DiT training loss curves. Finally, we also include Gflop counts for DDPM U-Net models from ADM and LDM in Table 6.

DiT model details. To embed input timesteps, we use a 256-dimensional frequency embedding [9] followed by a two-layer MLP with dimensionality equal to the transformer's hidden size and SiLU activations. Each adaLN layer feeds the sum of the timestep and class embeddings into a SiLU nonlinearity and a linear layer with output neurons equal to either $4\times$ (adaLN) or $6\times$ (adaLN-Zero) the transformer's hidden size. We use GELU nonlinearities (approximated with tanh) in the core transformer [16].

Classifier-free guidance on a subset of channels. In our experiments using classifier-free guidance, we applied guidance only to the first three channels of the latents instead of all four channels. Upon investigating, we found that three-channel guidance and four-channel guidance give similar results (in terms of FID) when simply adjusting the scale factor. Specifically, three-channel guidance with a scale of $(1 + x)$ appears reasonably well-approximated by four-channel guidance with a scale of $(1 + \frac{3}{4}x)$ (e.g., three-channel guidance with a scale of 1.5 gives an FID-50K of 2.27, and four-channel guidance with a scale of 1.375 gives an FID-50K of 2.20). It is interesting that applying guidance to a subset of elements can still yield good performance; we leave it to future work to explore this phenomenon further.

B. Model Samples

We show samples from our two DiT-XL/2 models at 512×512 and 256×256 resolution trained for 3M and 7M steps, respectively. Figures 1 and 11 show selected samples from both models. Figures 14 through 33 show *uncurated* samples from the two models across a range of classifier-free guidance scales and input class labels (generated with 250 DDPM sampling steps and the ft-EMA VAE decoder). As with prior work using guidance, we observe that larger scales increase visual fidelity and decrease sample diversity.

Model	Image Resolution	Flops (G)	Params (M)	Training Steps (K)	Batch Size	Learning Rate	DiT Block	FID-50K (no guidance)
DiT-S/8	256×256	0.36	33	400	256	1×10^{-4}	adaLN-Zero	153.60
DiT-S/4	256×256	1.41	33	400	256	1×10^{-4}	adaLN-Zero	100.41
DiT-S/2	256×256	6.06	33	400	256	1×10^{-4}	adaLN-Zero	68.40
DiT-B/8	256×256	1.42	131	400	256	1×10^{-4}	adaLN-Zero	122.74
DiT-B/4	256×256	5.56	130	400	256	1×10^{-4}	adaLN-Zero	68.38
DiT-B/2	256×256	23.01	130	400	256	1×10^{-4}	adaLN-Zero	43.47
DiT-L/8	256×256	5.01	459	400	256	1×10^{-4}	adaLN-Zero	118.87
DiT-L/4	256×256	19.70	458	400	256	1×10^{-4}	adaLN-Zero	45.64
DiT-L/2	256×256	80.71	458	400	256	1×10^{-4}	adaLN-Zero	23.33
DiT-XL/8	256×256	7.39	676	400	256	1×10^{-4}	adaLN-Zero	106.41
DiT-XL/4	256×256	29.05	675	400	256	1×10^{-4}	adaLN-Zero	43.01
DiT-XL/2	256×256	118.64	675	400	256	1×10^{-4}	adaLN-Zero	19.47
DiT-XL/2	256×256	119.37	449	400	256	1×10^{-4}	in-context	35.24
DiT-XL/2	256×256	137.62	598	400	256	1×10^{-4}	cross-attention	26.14
DiT-XL/2	256×256	118.56	600	400	256	1×10^{-4}	adaLN	25.21
DiT-XL/2	256×256	118.64	675	2352	256	1×10^{-4}	adaLN-Zero	10.67
DiT-XL/2	256×256	118.64	675	7000	256	1×10^{-4}	adaLN-Zero	9.62
DiT-XL/2	512×512	524.60	675	1301	256	1×10^{-4}	adaLN-Zero	13.78
DiT-XL/2	512×512	524.60	675	3000	256	1×10^{-4}	adaLN-Zero	11.93

Table 4: **Details of all DiT models.** We report detailed information about every DiT model in our paper. Note that FID-50K here is computed *without* classifier-free guidance. Parameter and flop counts exclude the VAE model which contains 84M parameters across the encoder and decoder. For both the 256×256 and 512×512 DiT-XL/2 models, we never observed FID saturate and continued training them as long as possible. Numbers reported in this table use the ft-MSE VAE decoder.

C. Additional Scaling Results

Impact of scaling on metrics beyond FID. In Figure 12, we show the effects of DiT scale on a suite of evaluation metrics—FID, sFID, Inception Score, Precision and Recall. We find that our FID-driven analysis in the main paper generalizes to these metrics—across the board, scaled-up DiT models are more compute-efficient and model Gflops are highly-correlated with performance. Notably, Inception Score and Precision benefit heavily from increased scale.

Impact of scaling on training loss. We also examine the impact of scale on training loss in Figure 13. Increasing DiT model Gflops (via transformer size or number of input tokens) causes the training loss to decrease more rapidly and saturate at a lower value. This phenomenon is consistent with trends observed with language models, where scaled-up transformers demonstrate improved loss as well as performance on downstream evaluation suites [26].

D. VAE Decoder Ablations

We used off-the-shelf, pre-trained VAEs across our experiments. The VAE models (ft-MSE and ft-EMA) are fine-tuned versions of the original LDM "f8" model (only the decoder weights are fine-tuned). We monitored metrics for our scaling analysis in Section 5 using the ft-MSE decoder, and we used the ft-EMA decoder for our final metrics reported in Tables 2 and 3. In this section, we ablate three different choices of the VAE decoder; the original one used by LDM and the two fine-tuned decoders used by Stable Diffusion.

Class-Conditional ImageNet 256×256, DiT-XL/2-G (cfg=1.5)					
Decoder	FID↓	sFID↓	IS↑	Precision↑	Recall↑
original	2.46	5.18	271.56	0.82	0.57
ft-MSE	2.30	4.73	276.09	0.83	0.57
ft-EMA	2.27	4.60	278.24	0.83	0.57

Table 5: **Decoder ablation.** We tested different pre-trained VAE decoder weights available at https://huggingface.co/stabilityai/sd-vae-ft-mse. Different pre-trained decoder weights yield comparable results on ImageNet 256×256.

Diffusion U-Net Model Complexities				
Model	Image Resolution	Base Flops (G)	Upsampler Flops (G)	Total Flops (G)
ADM	128×128	307	-	307
ADM	256×256	1120	-	1120
ADM	512×512	1983	-	1983
ADM-U	256×256	110	632	742
ADM-U	512×512	307	2506	2813
LDM-4	256×256	104	-	104
LDM-8	256×256	57	-	57

Table 6: **Gflop counts for baseline diffusion models that use U-Net backbones.** Note that we only count Flops for DDPM components.

Because the encoders are identical across models, the decoders can be swapped-in without retraining the diffusion model. Table 5 shows results; XL/2 outperforms all prior diffusion models even when using the LDM decoder.

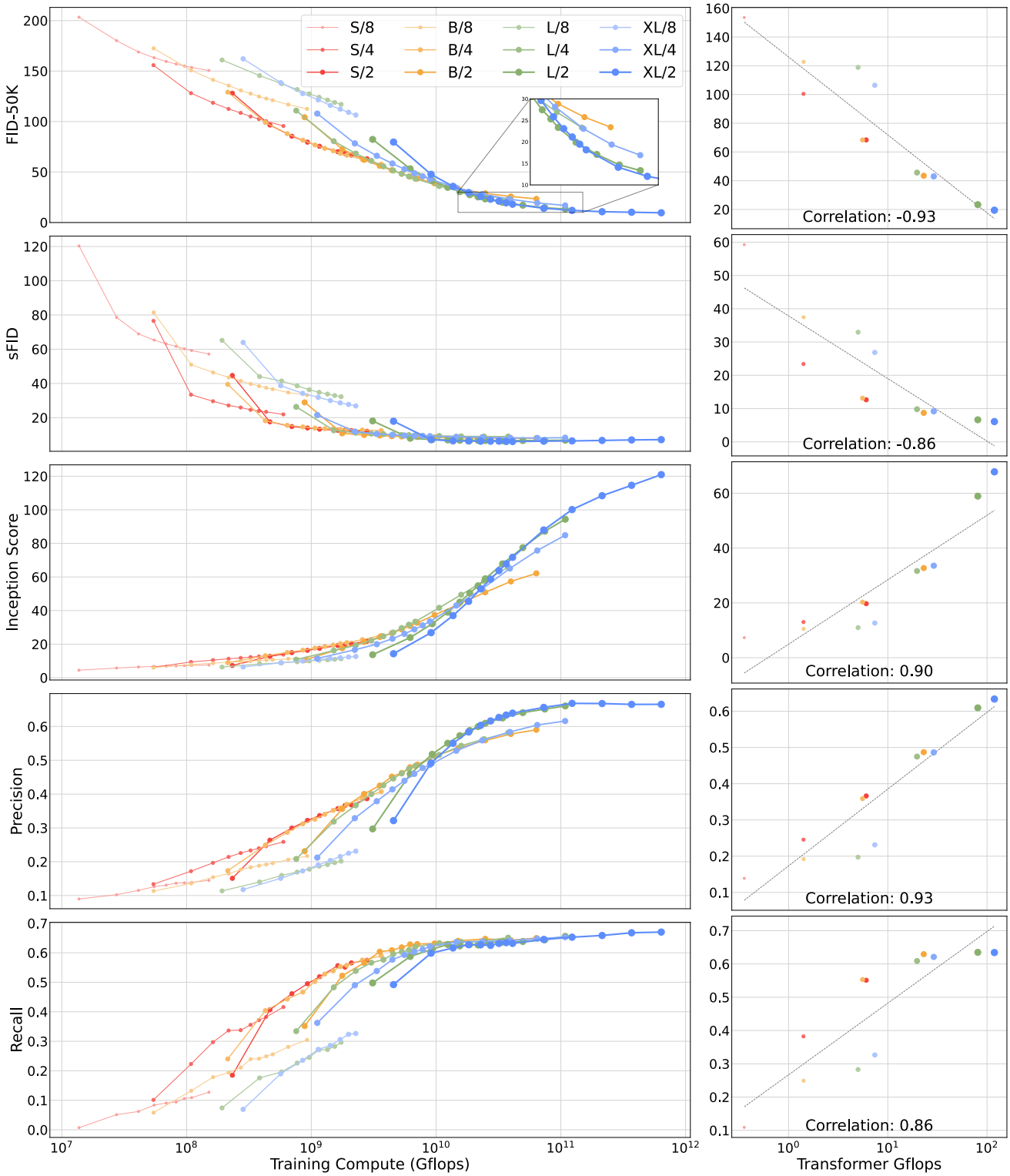

Figure 12: **DiT scaling behavior on several generative modeling metrics.** *Left:* We plot model performance as a function of total training compute for FID, sFID, Inception Score, Precision and Recall. *Right:* We plot model performance at 400K training steps for all 12 DiT variants against transformer Gflops, finding strong correlations across metrics. All values were computed using the ft-MSE VAE decoder.

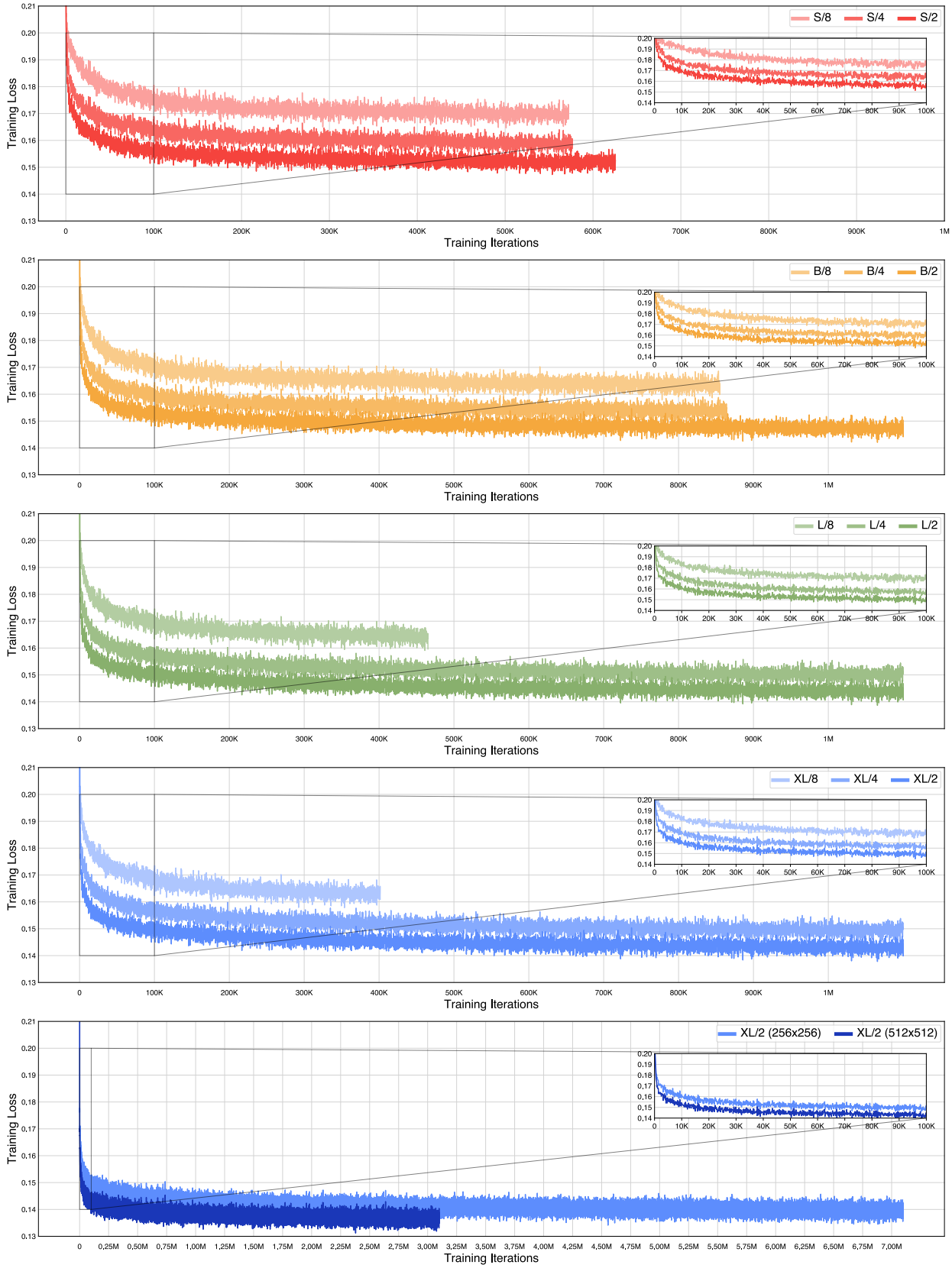

Figure 13: **Training loss curves for all DiT models.** We plot the loss over training for all DiT models (the sum of the noise prediction mean-squared error and $\mathcal{D}_{KL}$). We also highlight early training behavior. Note that scaled-up DiT models exhibit lower training losses.

Figure 14: **Uncurated** 512×512 **DiT-XL/2 samples.**
Classifier-free guidance scale = 4.0
Class label = "arctic wolf" (270)

Figure 15: **Uncurated** 512×512 **DiT-XL/2 samples.**
Classifier-free guidance scale = 4.0
Class label = "volcano" (980)

Figure 16: **Uncurated** 512×512 **DiT-XL/2 samples.**
Classifier-free guidance scale = 4.0
Class label = "husky" (250)

Figure 17: **Uncurated** 512×512 **DiT-XL/2 samples.**
Classifier-free guidance scale = 4.0
Class label = "sulphur-crested cockatoo" (89)

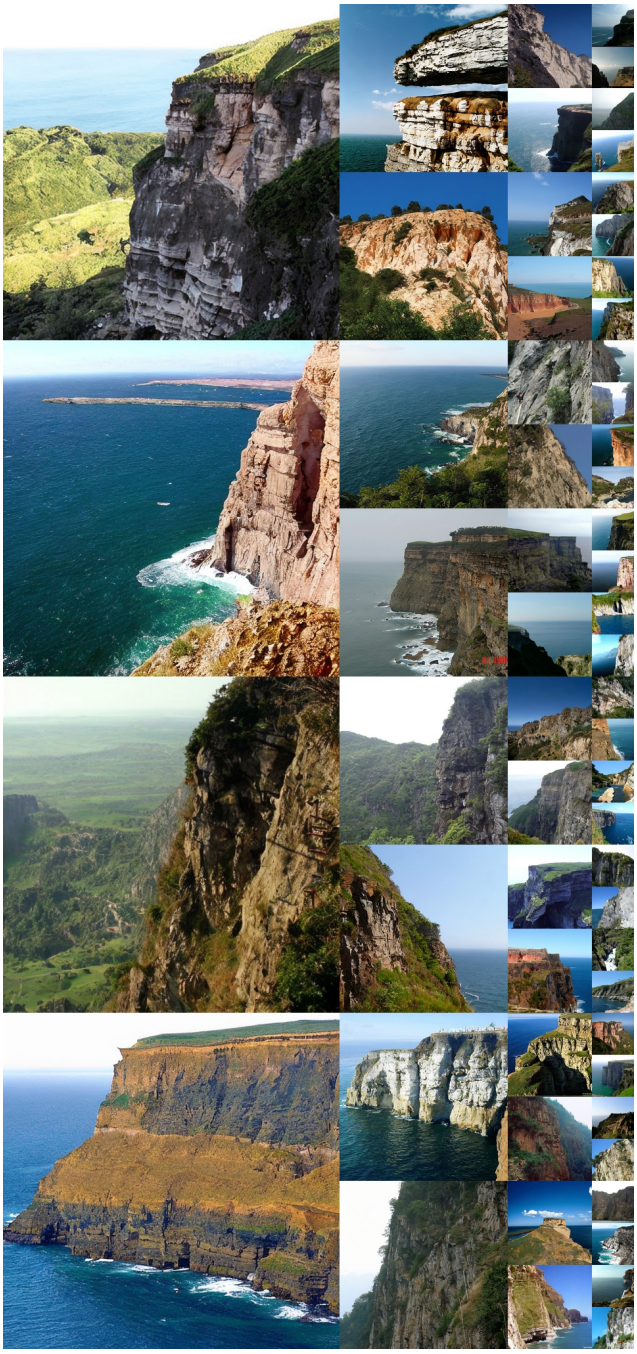

Figure 18: **Uncurated** 512×512 **DiT-XL/2 samples.**
Classifier-free guidance scale = 4.0
Class label = "cliff drop-off" (972)

Figure 19: **Uncurated** 512×512 **DiT-XL/2 samples.**
Classifier-free guidance scale = 4.0
Class label = "balloon" (417)

Figure 20: **Uncurated** 512×512 **DiT-XL/2 samples.**
Classifier-free guidance scale = 4.0
Class label = “lion” (291)

Figure 21: **Uncurated** 512×512 **DiT-XL/2 samples.**
Classifier-free guidance scale = 4.0
Class label = “otter” (360)

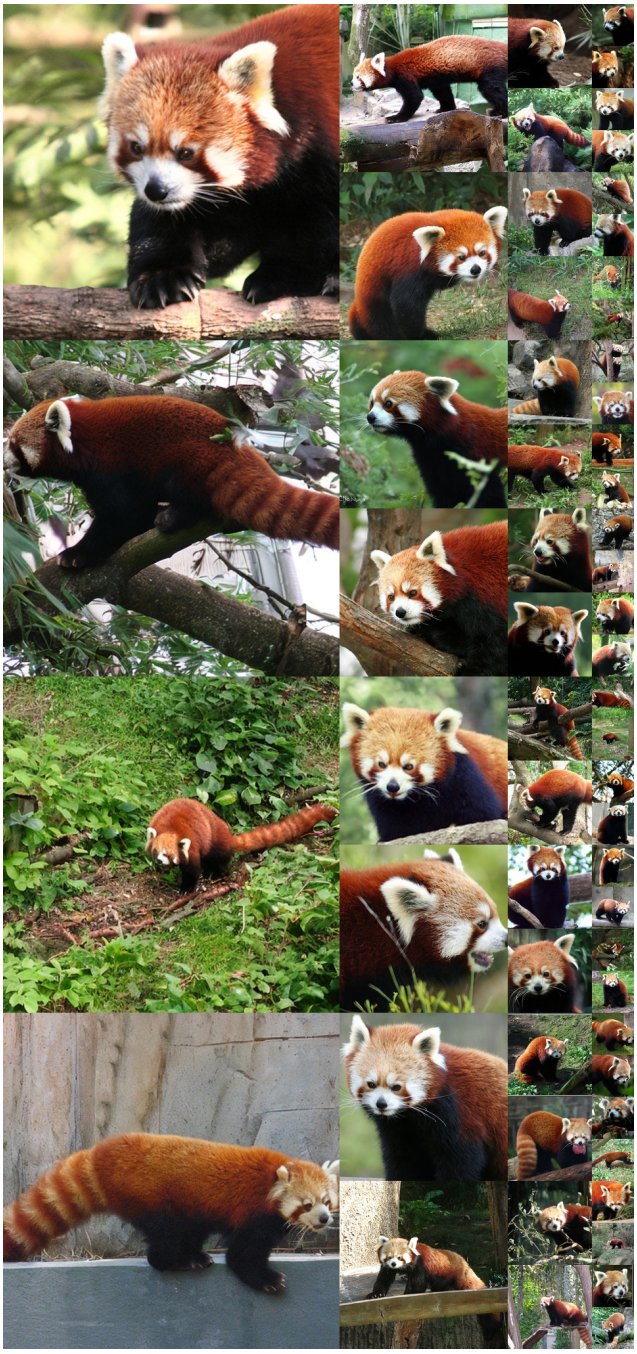

Figure 22: **Uncurated** 512×512 **DiT-XL/2 samples.**
Classifier-free guidance scale = 2.0
Class label = “red panda” (387)

Figure 23: **Uncurated** 512×512 **DiT-XL/2 samples.**
Classifier-free guidance scale = 2.0
Class label = “panda” (388)

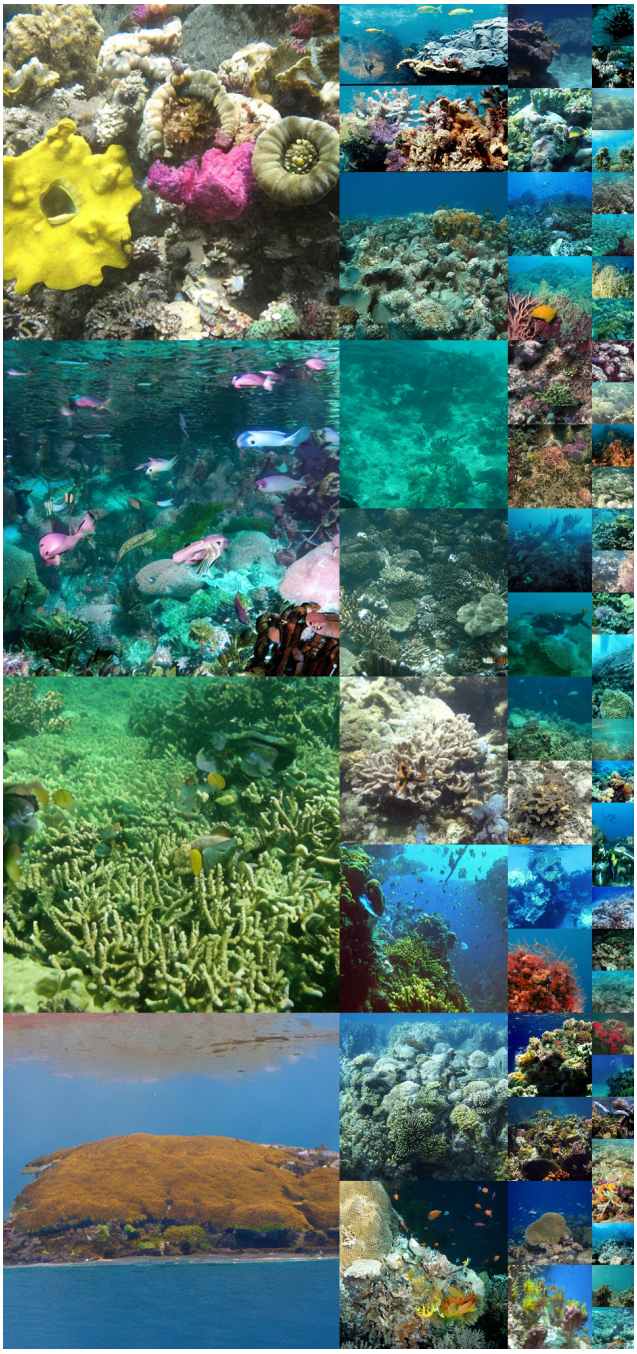

Figure 24: **Uncurated** 512×512 **DiT-XL/2 samples.**
Classifier-free guidance scale = 1.5
Class label = “coral reef” (973)

Figure 25: **Uncurated** 512×512 **DiT-XL/2 samples.**
Classifier-free guidance scale = 1.5
Class label = “macaw” (88)

Figure 26: **Uncurated** 256×256 **DiT-XL/2 samples.**
Classifier-free guidance scale = 4.0
Class label = “macaw” (88)

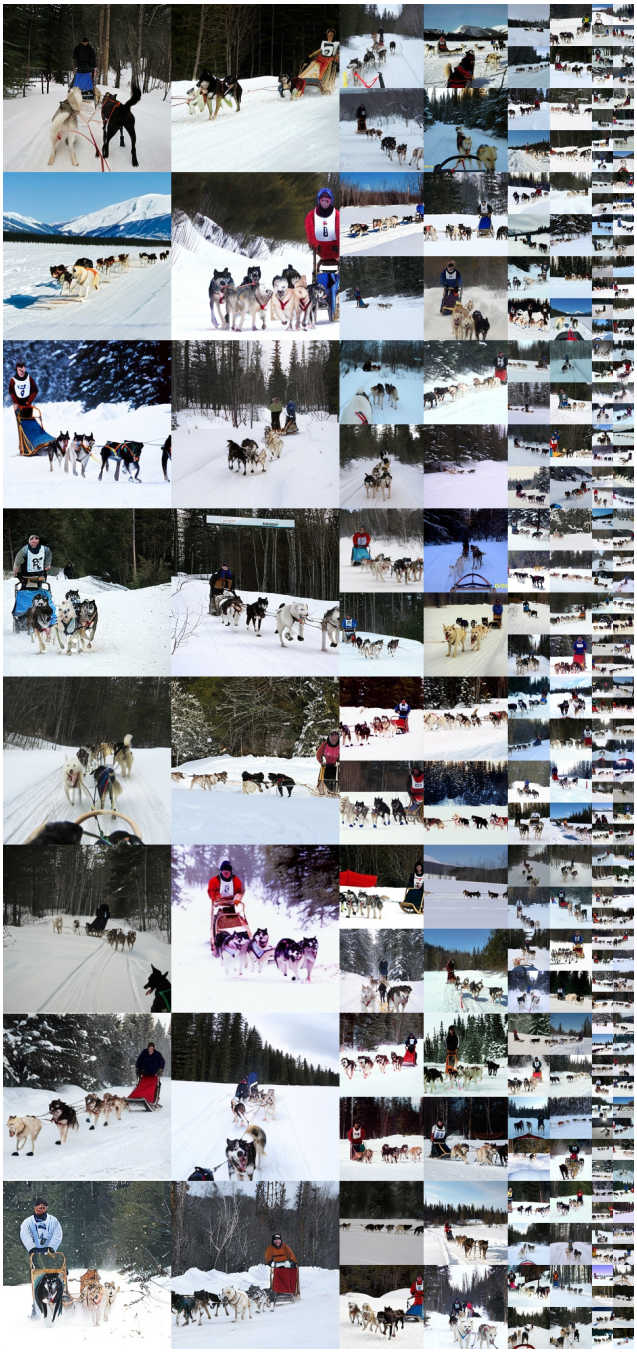

Figure 27: **Uncurated** 256×256 **DiT-XL/2 samples.**
Classifier-free guidance scale = 4.0
Class label = “dog sled” (537)

Figure 28: **Uncurated** 256×256 **DiT-XL/2 samples.**
Classifier-free guidance scale = 4.0
Class label = "arctic fox" (279)

Figure 29: **Uncurated** 256×256 **DiT-XL/2 samples.**
Classifier-free guidance scale = 4.0
Class label = "loggerhead sea turtle" (33)

Figure 30: **Uncurated** 256×256 **DiT-XL/2 samples.**
Classifier-free guidance scale = 2.0
Class label = "golden retriever" (207)

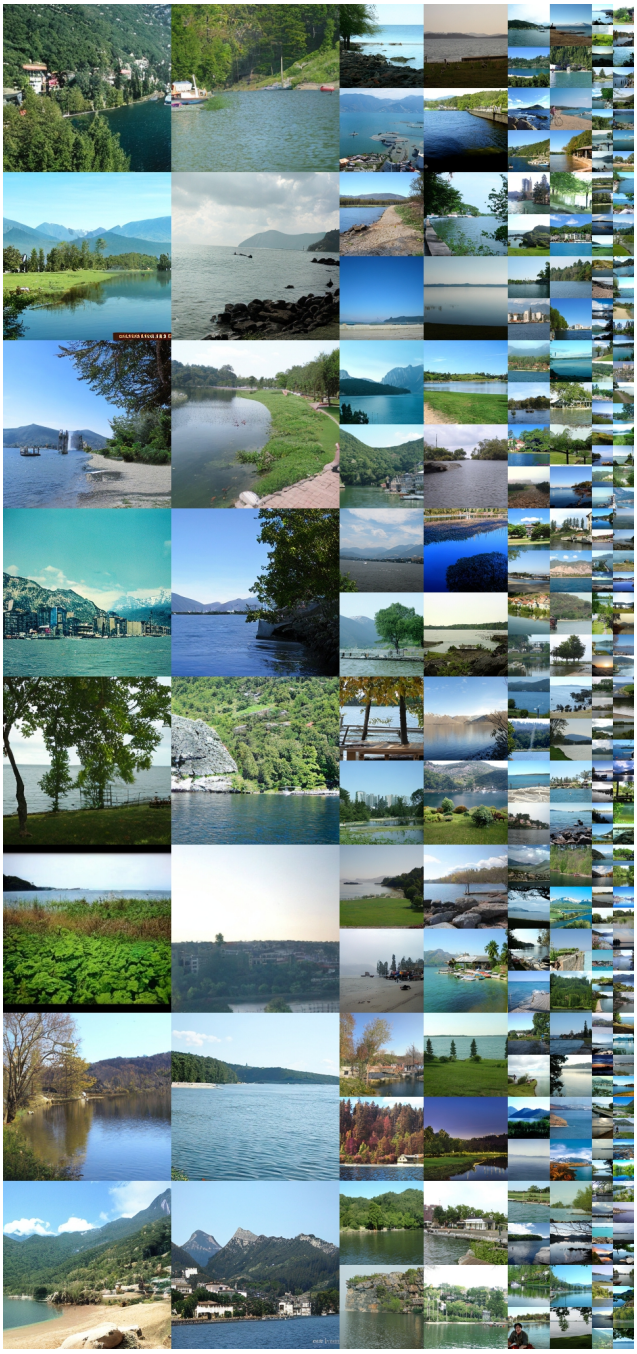

Figure 31: **Uncurated** 256×256 **DiT-XL/2 samples.**
Classifier-free guidance scale = 2.0
Class label = "lake shore" (975)

Figure 32: **Uncurated** 256×256 **DiT-XL/2 samples.**
Classifier-free guidance scale = 1.5
Class label = "space shuttle" (812)

Figure 33: **Uncurated** 256×256 **DiT-XL/2 samples.**
Classifier-free guidance scale = 1.5
Class label = "ice cream" (928)