A DC-AE Architecture and Training Details

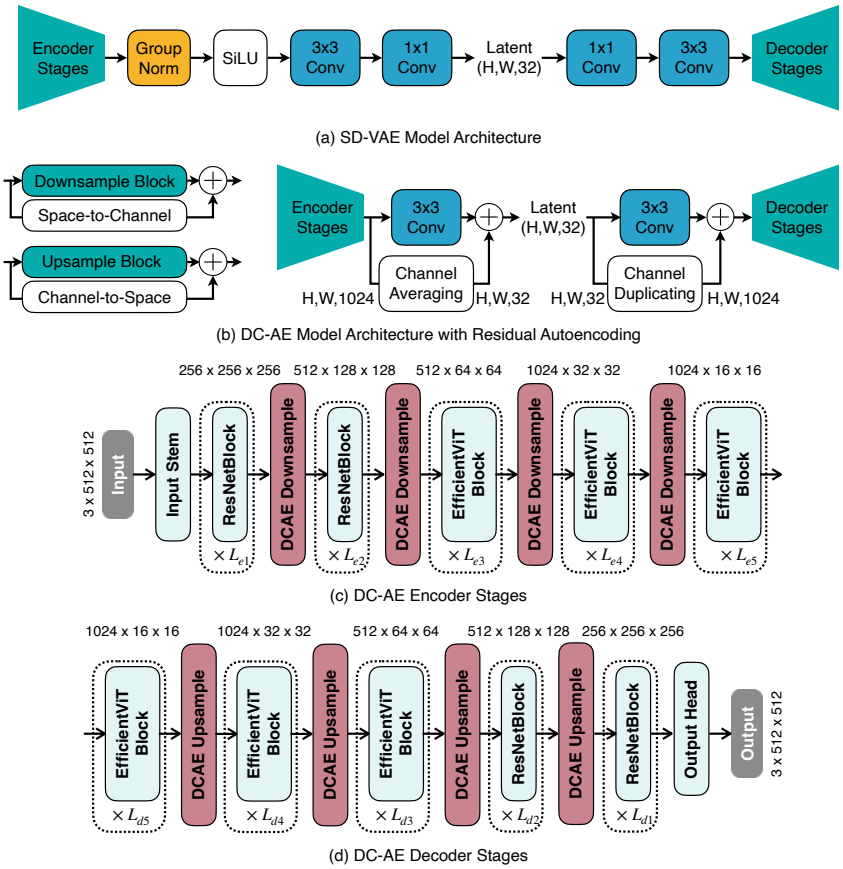

Figure 10: **Detailed Architecture of SD-VAE, DC-AE, DC-AE Encoder, and DC-AE Decoder Stages.**

We present the detailed architecture of SD-VAE, DC-AE, DC-AE encoder, and DC-AE decoder stages in Figure 10 to complement Figure 4.

We use the AdamW optimizer (Loshchilov, 2017) for all training phases.

In phase 1 (low-resolution full training), we use a constant learning rate of 6.4e-5 with a weight decay of 0.1, and AdamW betas of (0.9, 0.999). We use L1 loss and LPIPS loss (Zhang et al., 2018).

In phase 2 (high-resolution latent adaptation), we use a constant learning rate of 1.6e-5, a weight decay of 0.001, and AdamW betas of (0.9, 0.999). We use the same loss as phase 1.

In phase 3 (low-resolution local refinement), we use a constant learning rate of 5.4e-5, and AdamW betas of (0.5, 0.9). We use L1 loss, LPIPS loss (Zhang et al., 2018), and PatchGAN loss (Isola et al., 2017).

B Ablation Study on Training Different Numbers of Layers

Figure 11 presents the ablation study on training different numbers of layers in phase 2 (high-resolution latent adaptation) and phase 3 (low-resolution local refinement).

C Additional Image Reconstruction Results

Table 5 reports the reconstruction results under the low spatial-compression ratio setting. DC-AE delivers slightly better results than SD-VAE under this setting.

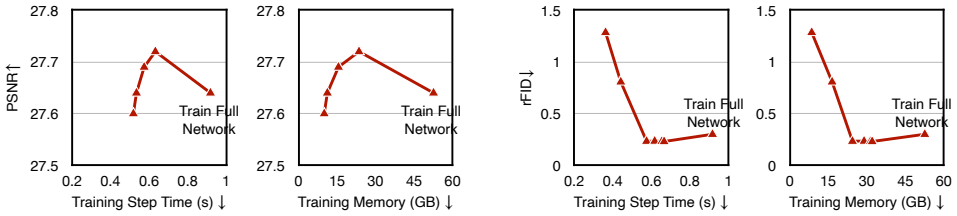

Figure 11: **Ablation Study on Training Different Numbers of Layers in Phase 2 (Left) and Phase 3 (Right).**

ImageNet 256×256	Latent Shape	Autoencoder	rFID ↓	PSNR ↑	SSIM ↑	LPIPS ↓
f8c4	32×32×4	SD-VAE [40]	0.63	24.99	0.71	0.063
		DC-AE	**0.46**	**25.46**	**0.73**	**0.057**

Table 5: **Image Reconstruction Results under the Low Spatial-Compression Ratio Setting.**

D Latent Scaling and Shifting Factors

Following the common practice (Rombach et al., 2022; Peebles & Xie, 2023; Bao et al., 2023; Esser et al., 2024; Labs, 2024; Chen et al., 2024b;a), we normalize the latent space of our autoencoders to apply to latent diffusion models. Given a dataset, we compute the root mean square of the latent features and use its multiplicative inverse as the scaling factor for our autoencoders. We do not use the shifting factor for our autoencoders.

E Diffusion Model Architecture Details

In addition to existing UViT models, we scaled the model up to 1.6B parameters, with a depth of 28, a hidden dimension of 2048, and 32 heads. We denote this model as UViT-2B.

F Diffusion Sampling Hyperparameters

For the DiT models, we use the DDPM (Ho et al., 2020) sampler from the DiT (Peebles & Xie, 2023) codebase with 250 sampling steps and a guidance scale of 1.3.

For the UViT models, we use the DPMSolver (Lu et al., 2022a) sampler with 30 sampling steps and a guidance scale of 1.5.

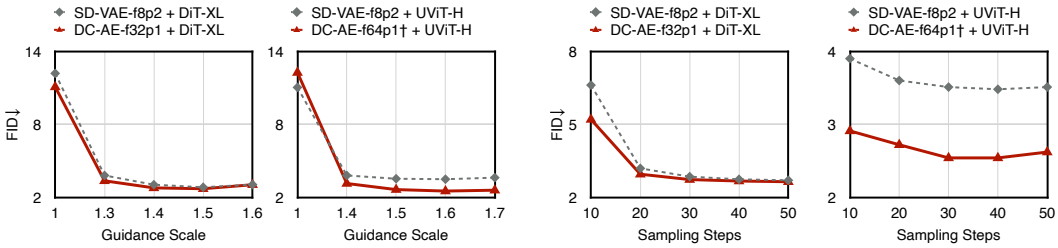

Figure 12: **Ablation Study on Diffusion Sampling Hyperparameters.** We use the DPMSolver sampler for both DiT-XL and UViT-H. DC-AE provides significant speedup over the baseline models while maintaining the generation performance under different diffusion sampling hyperparameters.

G High-Resolution Image Generation Results

Apart from ImageNet 512×512, we also test our models for higher-resolution image generation. As shown in Table 6, we have a similar finding where DC-AE-f32p1 achieves better FID than SD-VAE-f8p2 for all cases.

FFHQ 1024×1024 (Unconditional) & MJHQ 1024×1024 (Class-Conditional)										
Diffusion Model	Autoencoder	Patch Size	NFE	Throughput (image/s) ↑ Training	Inference	Latency (ms) ↓	Memory (GB) ↓	FFHQ FID ↓ w/o CFG	MJHQ FID ↓ w/o CFG	w/ CFG
DiT-S [38]	SD3-VAE-f8 [9]	2	250	83	1.63	3554	41.4	46.28	109.43	103.02
	Flux-VAE-f8 [20]	2	250	83	1.63	3554	41.4	59.15	143.16	139.06
	SDXL-VAE-f8 [39]	2	250	84	1.67	3530	41.2	16.82	49.00	39.21
	Asym-VAE-f8 [58]	2	250	84	1.67	3530	41.2	17.10	48.30	38.35
	SD-VAE-f8 [40]	2	250	84	1.67	3530	41.2	16.98	48.05	38.19
		4	250	470	11.13	632	10.7	23.81	60.94	51.29
	DC-AE-f32	1	250	475	11.15	634	10.7	13.65	34.35	27.20
	DC-AE-f32‡	1	250	475	11.15	634	10.7	**11.39**	**28.36**	**21.89**
	DC-AE-f64	1	250	**2085**	**50.26**	**230**	**3.1**	26.88	61.30	53.38

MapillaryVistas 2048×2048 (Unconditional)								
Diffusion Model	Autoencoder	Patch Size	NFE	Throughput (image/s) ↑ Training	Inference	Latency (ms) ↓	Memory (GB) ↓	MapillaryVistas FID ↓ w/o CFG
DiT-S [38]	SD-VAE-f8 [40]	4	250	84	1.64	3561	41.4	69.50
	DC-AE-f64	1	250	**459**	**10.91**	**639**	**11.0**	**59.55**

Table 6: **1024×1024 and 2048×2048 Image Generation Results.** ‡ represents the model is trained with 4× batch size (i.e., 256 → 1024).

H Image Generation Results with Other Evaluation Metrics

Table 7 presents a comprehensive evaluation of different diffusion models and autoencoders on ImageNet 512×512. The evaluation metrics include FID (Martin et al., 2017), inception score (IS) (Salimans et al., 2016), precision, recall (Kynkäänniemi et al., 2019), and CMMD (Jayasumana et al., 2024). Our DC-AE consistently delivers significant efficiency improvements while maintaining the generation performance under different evaluation metrics.

I Additional Samples

In Figure 13 and 14, we provide additional image reconstruction samples produced by SD-VAE and DC-AE. Reconstructed images by DC-AE demonstrate better visual qualities than SD-VAE's reconstructed images, especially for the f64 and f128 autoencoders. Some samples are cropped for better visualization of details like human faces and small texts.

In Figure 15 and Figure 16, we show randomly generated samples on ImageNet 512×512 and MJHQ-30K 512×512 by the diffusion models using our DC-AE.

Diffusion Model	Autoencoder	Patch Size	NFE	Inference Throughput	FID ↓		Inception Score ↑		Precision ↑		Recall ↑		CMMD ↓	
					w/o CFG	w/ CFG	w/o CFG	w/ CFG	w/o CFG	w/ CFG	w/o CFG	w/ CFG	w/o CFG	w/ CFG
UViT-S [1]	SD3-VAE-f8 [9]	2	30	49.73	164.34	143.82	6.07	7.53	0.06	0.09	0.31	0.39	3.13	2.94
	Flux-VAE-f8 [20]	2	30	49.73	106.07	84.73	13.39	17.71	0.28	0.37	0.39	0.42	1.90	1.67
	SDXL-VAE-f8 [39]	2	30	49.85	51.03	26.38	27.58	56.72	0.57	0.74	0.58	0.50	1.35	1.05
	Asym-VAE-f8 [58]	2	30	49.85	52.68	25.14	30.22	65.27	0.58	0.74	0.62	0.51	1.09	0.80
	SD-VAE-f8 [40]	2	30	49.85	51.96	24.57	30.37	65.73	0.57	0.74	0.64	0.52	1.23	0.91
	SD-VAE-f16 [40]	2	30	214.68	76.86	44.22	21.38	43.35	0.43	0.62	0.60	0.55	1.83	1.46
	SD-VAE-f32 [40]	1	30	214.72	70.23	38.63	23.07	47.72	0.46	0.64	0.58	0.56	1.71	1.36
	DC-AE-f32	1	30	214.17	46.12	18.08	34.82	84.73	0.59	0.76	0.66	0.56	1.00	0.70
	DC-AE-f64	1	30	896.23	67.30	35.96	24.55	52.86	0.44	0.64	0.60	0.56	1.44	1.14
	DC-AE-f64†	1	30	896.23	61.84	30.63	27.28	61.76	0.47	0.67	0.63	0.56	1.35	1.04
DiT-XL [38]	Flux-VAE-f8 [20]	2	250	0.83	27.35	8.72	53.09	130.20	0.68	0.83	0.61	0.48	0.54	0.30
	Asym-VAE-f8 [58]	2	250	0.85	11.39	2.97	108.70	241.10	0.75	0.83	0.65	0.53	0.37	0.20
	SD-VAE-f8 [40]	2	250	0.85	12.03	3.04	105.25	240.82	0.75	0.84	0.64	0.54	0.43	0.25
	DC-AE-f32	1	250	4.03	9.56	2.84	117.49	226.98	0.75	0.82	0.64	0.55	0.34	0.22
	DC-AE-f32‡	1	250	4.03	6.88	2.41	141.07	263.56	0.76	0.82	0.63	0.56	0.29	0.18
UViT-H [1]	Flux-VAE-f8 [20]	2	30	5.82	30.91	12.63	56.72	127.93	0.64	0.76	0.59	0.49	0.50	0.31
	Asym-VAE-f8 [58]	2	30	5.85	11.36	3.51	124.24	249.21	0.75	0.82	0.61	0.53	0.32	0.20
	SD-VAE-f8 [40]	2	30	5.85	11.04	3.55	125.08	250.66	0.75	0.82	0.61	0.53	0.39	0.26
	DC-AE-f32	1	30	27.03	9.83	2.53	121.91	255.07	0.76	0.83	0.65	0.54	0.34	0.20
	DC-AE-f64	1	30	111.77	13.96	3.01	99.20	229.16	0.73	0.83	0.64	0.53	0.50	0.31
	DC-AE-f64†	1	30	111.77	12.26	2.66	109.20	239.82	0.73	0.82	0.67	0.57	0.43	0.27
UViT-2B [1]	Flux-VAE-f8 [20]	2	30	2.58	25.03	10.12	74.04	161.29	0.67	0.78	0.58	0.51	0.38	0.24
	Asym-VAE-f8 [58]	2	30	2.62	9.87	3.62	131.95	258.63	0.76	0.83	0.59	0.52	0.30	0.19
	SD-VAE-f8 [40]	2	30	2.62	9.73	3.57	132.86	260.50	0.76	0.83	0.59	0.52	0.37	0.24
	DC-AE-f32	1	30	11.08	8.13	2.30	135.44	272.73	0.76	0.82	0.66	0.56	0.30	0.17
	DC-AE-f64	1	30	45.55	7.78	2.47	138.11	280.49	0.77	0.84	0.63	0.54	0.35	0.22
	DC-AE-f64†	1	30	45.55	6.50	2.25	152.35	293.45	0.77	0.83	0.65	0.56	0.31	0.19
MAGVIT-v2 [51]	-	-	-	-	3.07	1.91	213.1	324.3	-	-	-	-	-	-
EDM2-XXL [17]	-	-	-	-	1.91	1.81	-	-	-	-	-	-	-	-
MAR-L [24]	-	-	-	-	2.74	1.73	205.2	279.9	-	-	-	-	-	-
SiT-XL [33]	DC-AE-f32	1	-	-	7.47	2.41	131.37	237.71	0.77	0.82	0.65	0.58	0.36	0.23
USiT-H	DC-AE-f32	1	-	-	3.80	1.89	174.58	252.35	0.78	0.82	0.64	0.60	0.24	0.18
USiT-2B	DC-AE-f32	1	-	-	2.90	1.72	187.68	248.10	0.79	0.82	0.63	0.61	0.21	0.17

Table 7: **Class-Conditional Image Generation Results on ImageNet 512×512 with More Evaluation Metrics.** † represents the model is trained for 4× training iterations (i.e., 500K → 2,000K iterations). ‡ represents the model is trained with 4× batch size (i.e., 256 → 1024). 'NFE' denotes the number of functional evaluations. The NFEs for SiT (Ma et al., 2024a) and USiT models are left blank as they use an adaptive-step evaluation scheduler.

Figure 13: **Additional Autoencoder Image Reconstruction Samples.**

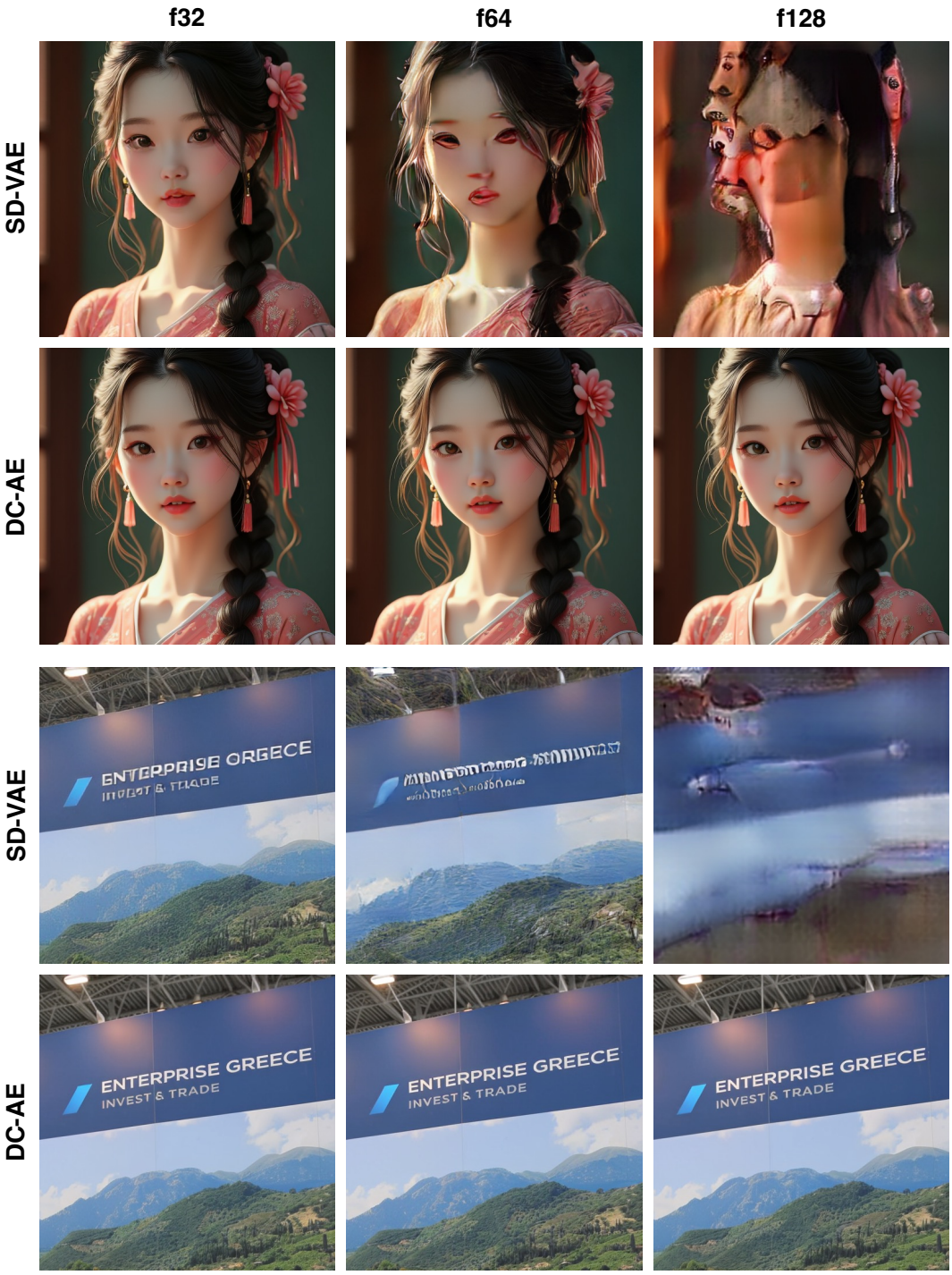

Figure 14: **Additional Autoencoder Image Reconstruction Samples.**

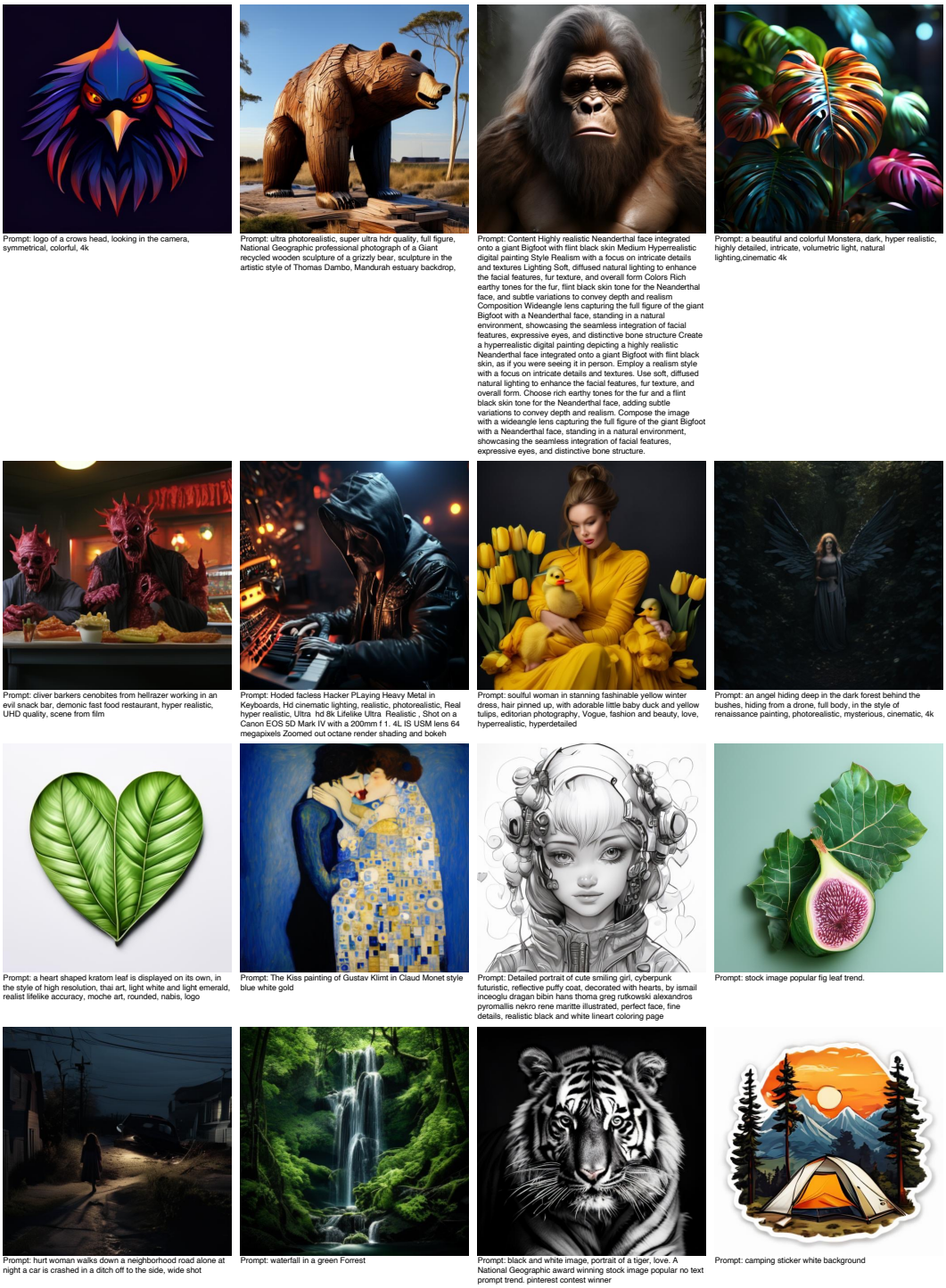

Figure 15: **Random 512×512 Text-to-Image Samples.** Prompts are randomly drawn from MJHQ-30K (Li et al., 2024a).

Figure 16: **Random Generated Samples on ImageNet 512×512.**