One Layer Is Enough: Adapting Pretrained Visual Encoders for Image Generation

Supplementary Material

A. FAE Encoder Structure

We merge the consecutive linear layers in the attention module and use a larger per-head dimension for the encoder. The structure of our encoder is shown in Figure S1.

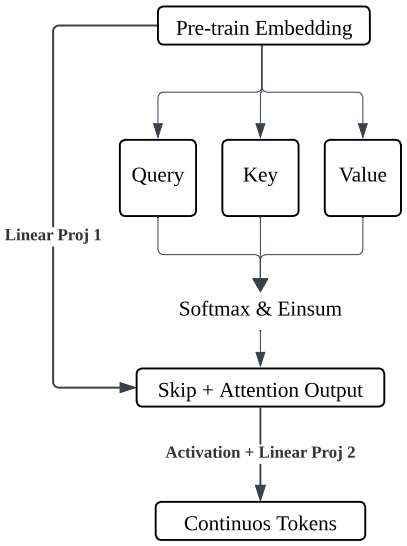


Figure S1. Modified Attention

B. Ablation Results

Model	CFG	64 Epochs	160 Epochs	320 Epochs
32-dim	w/o	2.67	2.02	1.76
	w/	–	1.70	1.52
48-dim	w/o	2.73	2.10	1.88
	w/	–	1.73	1.56
64-dim	w/o	2.86	2.25	1.99
	w/	–	1.76	1.64

Table S1. Ablation results comparing different latent dimension.

Model	Probing	CFG	64 Epochs	160 Epochs	320 Epochs
Single Attention	86.17%	w/o	2.98	2.27	1.98
		w/	–	1.79	1.61
Linear	85.74%	w/o	3.03	2.38	2.07
		w/	–	1.92	1.76
6-Layer Transformer	–	w/o	3.31	2.47	2.13
	–	w/	–	1.84	1.65
Direct Predict	–	w/o	15.37	12.99	12.72
DinoV2	–	w/	–	17.85	16.53

Table S2. Ablation results comparing different encoder structure.

Model	CFG	64 Epochs	160 Epochs	320 Epochs
SiT	w/o	2.98	2.27	1.98
	w/	–	1.79	1.61
SiT + SwiGLU	w/o	3.02	2.26	1.97
	w/	–	1.75	1.60
SiT + SwiGLU + ROPE	w/o	2.86	2.182	1.89
	w/	–	1.78	1.63
SiT + SwiGLU + ROPE + RMSNorm	w/o	2.74	2.15	1.86
	w/	–	1.71	1.55

Table S3. Ablation results comparing LDM model structure

Model	CFG	64 Epochs	160 Epochs	320 Epochs
32-dim, ts=0.7	w/o	2.4087	1.9060	1.6836
	w/	–	1.6786	1.5131
32-dim, ts=0.5	w/o	2.3233	1.8501	1.7086
	w/	–	1.6735	1.5682
32-dim, ts=0.3	w/o	2.3220	1.8800	1.7743
	w/	–	1.7125	1.6227
48-dim, ts=0.5	w/o	2.4329	1.9546	1.6952
	w/	–	1.6797	1.5312
48-dim, ts=0.3	w/o	2.3599	1.9105	1.6911
	w/	–	1.6694	1.5423
64-dim, ts=0.2	w/o	2.4398	1.9549	1.7581
	w/	–	1.7563	1.5402

Table S4. Ablation results comparing different timesteps shift across different token dimension.

C. rFID

Because the encoder training is disentangled with the image reconstruction loss, our rFID and tokenizer reconstruction fidelity lag behind methods such as VA-VAE that directly optimize reconstruction quality.

	SD-VAE	VA-VAE	FAE 32-dim	FAE 64-dim
rFID	0.73	0.28	0.68	0.66

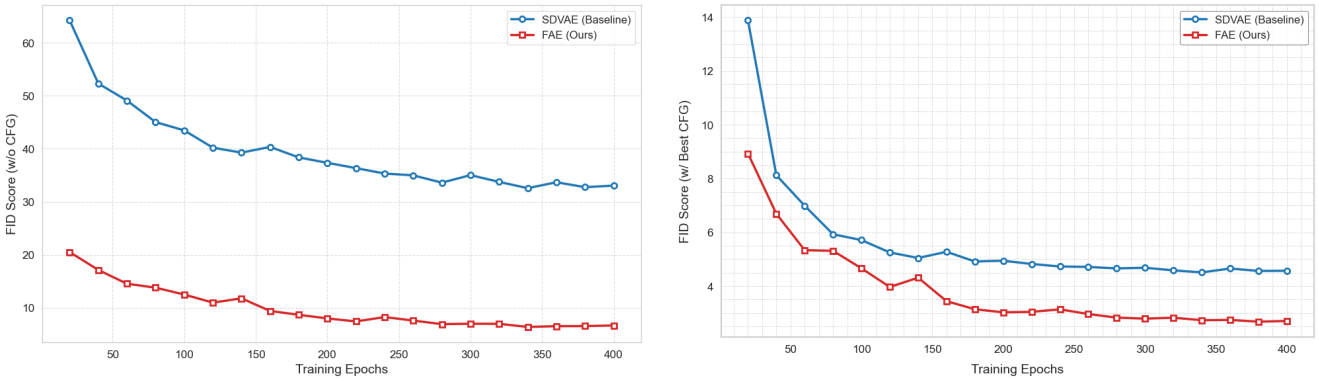
Table S5. Reconstruction rFID comparison.

D. Text-to-Image results on MS-COCO

Model	FID	Type	Training datasets	#Params
DALL-E [37]	~ 28	Autoregressive	DALL-E dataset (250M)	12B
CogView [12]	27.1	Autoregressive	Internal dataset (30M)	4B
LAFITE [60]	26.94	GAN	CC3M (3M)	75M + 151M (TE)
GLIDE [32]	12.24	Diffusion	DALL-E dataset (250M)	3.5B + 1.5B (SR)
Make-A-Scene [14]	11.84	Autoregressive	Union datasets (without MS-COCO) (35M)	4B
DALL-E 2 [38]	10.39	Diffusion	DALL-E dataset (250M)	4.5B + 700M (SR)
Imagen [41]	7.27	Diffusion	Internal dataset (460M) + LAION (400M)	2B + 4.6B (TE) + 600M (SR)
Parti [52]	7.23	Autoregressive	LAION (400M) + FIT (400M) + JFT (4B)	20B + 630M (AE)
Re-Imagen [9]	**6.88**	Diffusion	KNN-ImageText (50M)	2.5B + 750M (SR)
SDVAE+T5 (w/o CFG)	21.25	Diffusion	CC12M	604M
FAE (SigLIPV2)+T5 (w/o CFG)	7.57	Diffusion	CC12M	604M+ 514M (FAE)
FAE (SigLIPV2)+T5 (w/ CFG)	7.11	Diffusion	CC12M	604M+ 514M (FAE)
FAE (DinoV2)+T5 (w/o CFG)	7.47	Diffusion	CC12M	604M+ 514M (FAE)
FAE (DinoV2)+T5 (w/ CFG)	6.90	Diffusion	CC12M	604M+ 514M (FAE)

Table S6. FID results of different models on MS-COCO validation (256×256). All the models are trained on external dataset and zero-shot evaluated on MS-COCO using 30K example.

E. STarFlow Convergence Speed Comparison



(a) Results without CFG.

(b) Results with CFG.

Figure S2. Comparison of STARFlow with SDVAE and the proposed FAE under the same settings.

F. FAE Hyper Parameters

Category	Field	Encoder	Decoder	Pixel Decoder	LDM	MMDiT	MMDiT 384x384
Architecture	Input dim.	16×16×1536	16×16×32	16×16×1536	16×16×32	16×16×32	24×24×64
	Output dim.	16×16×64	16×16×1536	256×256×3	16×16×32	16×16×32	24×24×64
	Hidden dim.	6144	1536	1024	1152	1024	1536
	Num. layers	1	6	24	28	16	24
	MLP Ratio	–	4	4	4	4	4
	Dim. per head	256	64	64	72	64	64
	Num. heads	24	24	16	16	16	24
	Total Params (M)	38.17	170.43	305.36	675.26	603.46	2017.84
Optimization	Training iters	1M		1M	2M	1M	1M
	Batch size	1024		512	512	512	512
	Optimizer	AdamW		AdamW	AdamW	AdamW	AdamW
	Peak LR	1e-4		1e-4	1e-4	1e-4	1e-4
	LR Scheduler	Cosine		Cosine	Constant	Constant	Constant
	Warmup	1000		1000	–	–	–
	(β_1, β_2)	(0.9, 0.999)		(0.9,0.999)	(0.9,0.999)	(0.9,0.999)	(0.9,0.999)
Interpolants	α_t	–		–	1-t	1-t	1-t
	σ_t	–		–	t	t	t
	w_t	–		–	σ_t	σ_t	σ_t
	Training objective	–		–	v-prediction	v-prediction	v-prediction
	Sampler	–		–	Euler-Maruyama (w/o CFG) Euler (w/ CFG)	Euler-Maruyama (w/o CFG) Euler (w/ CFG)	Euler-Maruyama (w/o CFG) Euler (w/ CFG)
	Sampling steps	–		–	250	250	250
	Guidance	–		–	0.9 (t=1∼0.9) 2.5 (t=0.7∼0)	1.5 (t=0.9∼0)	1.5 (t=0.9∼0)

The MMDiT 384×384 and its FAE encoder are only used for generating high quality examples provided in the paper. The three-partitioned CFG are only used for generating Main Results. The main results uses timesteps shift=0.4. All ablation experiments use single cfg scale for $t = 0.7 \sim 0.0$, and the cfg scale is grid searched in 0.1 fineness. The ablation experiments for FAE Model Structure are using SiT. The linear probing results are got from a separate encoder trained with DinoV2 register version with same parameters. And the ablation experiments for Token Dimension and Time Shift are using LightningDiT.

G. Patch Embedding Similarity Maps

We compare the similarity between different patches inside single images. For each triplet of visualizations, the first image shows the similarity map computed from the DINOv2 embeddings, while the second shows the corresponding similarity map derived from our FAE latents. The third image displays the original image patch used as the query. The selected query patches are highlighted with a red rectangle, and darker colors in the similarity maps indicate higher similarity values.

Figure S3. Similarity of a photo of cat.

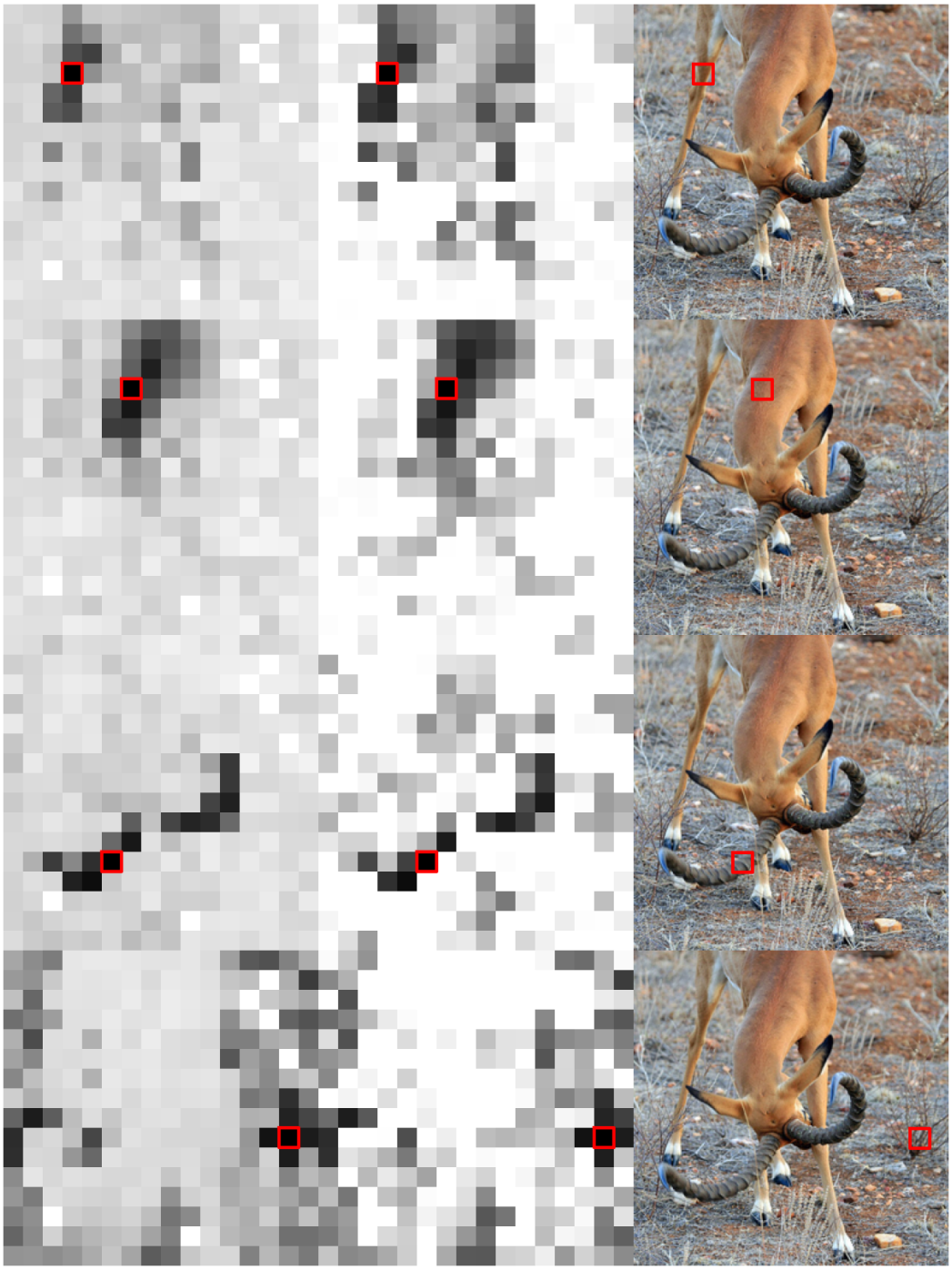


Figure S4. Similarity of a photo of impala.

H. Matching most Similar patch pair across two images

Our latents retain the cross-image patch–matching behavior characteristic of DINOv2: semantically corresponding regions across different images are still reliably matched using cosine similarity in the latent space. This suggests that FAE preserves fine-grained, part-level semantics rather than only coarse global information.

We first identify animal-related patches using K-Means clustering. From these, we randomly select patches in the first image and match each one to the patch in the second image with the highest cosine similarity. For each example, 16 patch pairs are selected and visualized.

Figure S5. Matching most similar patch pair from two photo of bird.

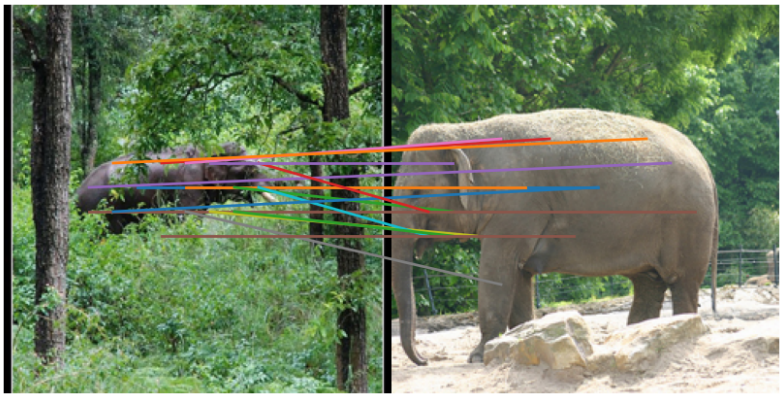


Figure S6. Matching most similar patch pair from two photo of elephant.

I. Text-to-Image Prompts of Figure 5

The prompts for the text to image examples in **Figure 5** are:
”a wooden bench under a large oak tree with warm sunlight streaming through branches and fallen leaves scattered below”,
”a panoramic mountain ridge under soft morning clouds”,
”a wooden arrow sign reading ’north trail’ pointing into the woods”,
”an alpine lake surrounded by steep cliffs, water perfectly still except for faint circular ripples, floating pollen creating tiny shimmering patterns on the surface”,
”a window sticker reading ’welcome’”,
”a snow leopard walking across snowy slope, faint pawprints trailing behind”,
”a winter woodland with heavy snow draped asymmetrically across branches, faint animal tracks weaving between tree shadows under pale blue light”,
”a sticky note attached to a monitor reading ’finish draft by 5 pm’”,
”a small shop window sign reading ’local goods’”,
”a tortoise lumbering through sunlit shrubs, shell etched with age patterns”,
”a wooden produce crate stamped ’farm fresh’ positioned in a sunlit garden shed beside tools”,
”a rocky mountain meadow scattered with boulders and wildflowers under bright daylight”

J. STARFlow Examples

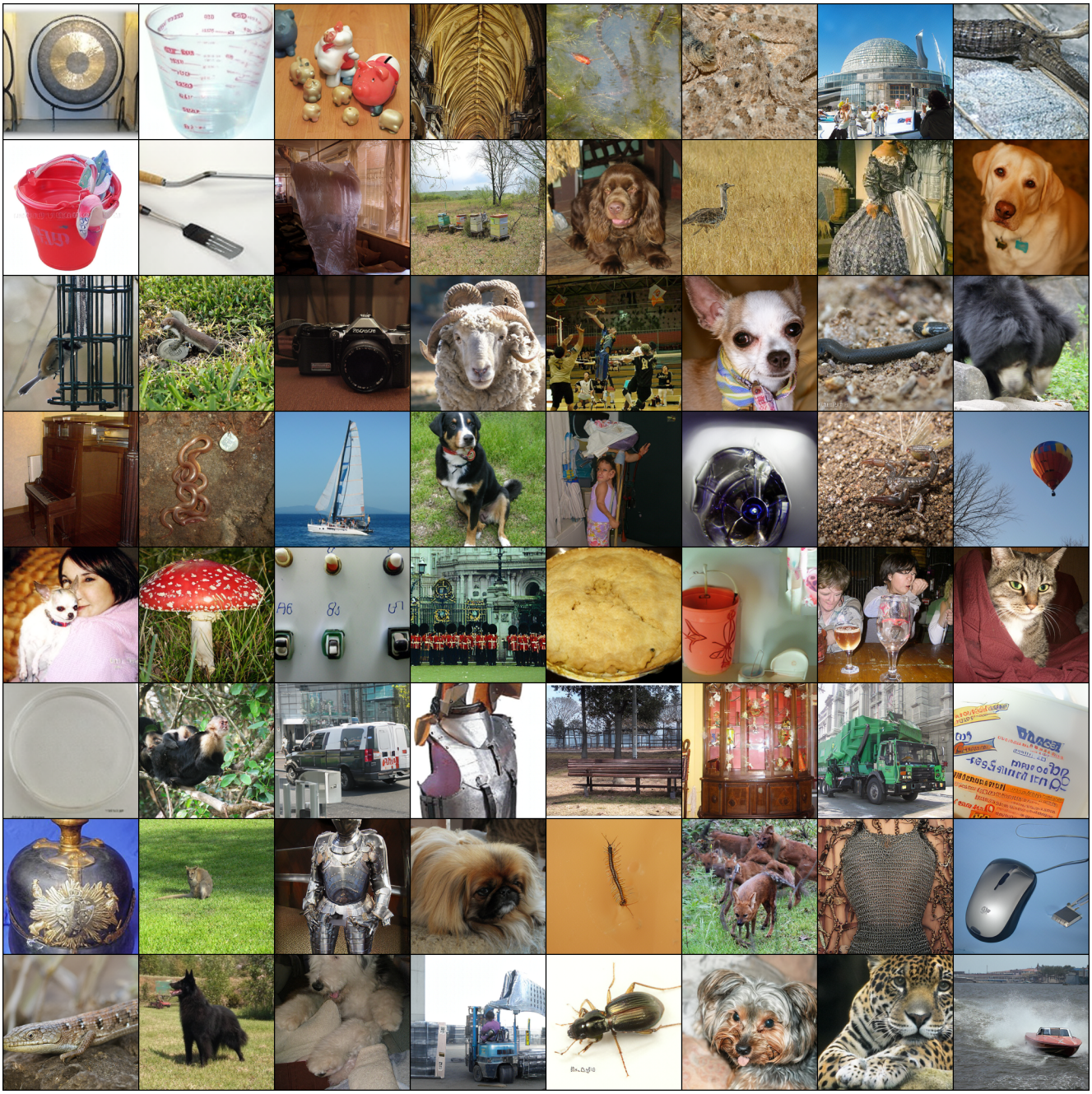


Figure S7. Random samples of ImageNet 256x256 generated by STarFlow model.

K. Extra Examples from Siglip2 MMDiT 384 Model

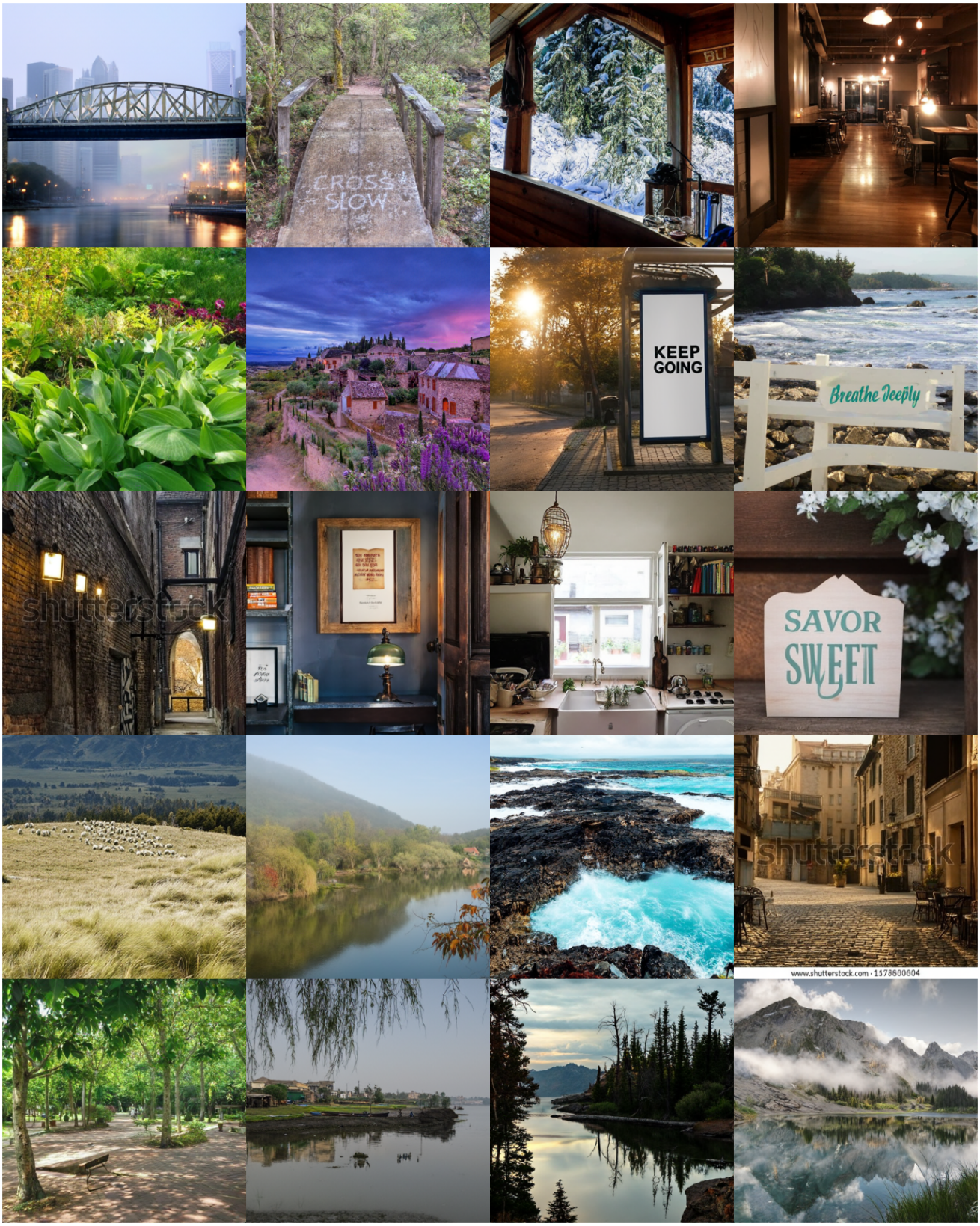


Figure S8. Random samples of Siglip2 MMDiT 384×384 Model.

The prompts for the text to image examples are:
”a foggy bridge spans a calm river reflecting muted city lights.”,
”a forest bridge plank etched: ‘cross slow’. water murmurs beneath the boards.”,
”a hiking lodge interior has a carved plaque: ‘rest; the mountains will wait.’ snow drifts past the windows.”,
”a small café interior glows softly as candles flicker on wooden tables.”,
”a quiet backyard garden with warm sun patterns filtering through leaves.”,
”a hillside dotted with tiny cottages beneath a lavender evening sky.”,
”a bus stop poster saying: ‘keep going’. golden morning light warms the street.”,
”a seaside cabin porch sign saying: ‘breathe deeply’. waves pulse against the rocks below.”,
”a lantern-lit alleyway glowing softly between old brick walls.”,
”a quiet library alcove features a framed message: ‘seek answers, but also seek the calm between them.’ warm lamplight glows against tall wooden shelves.”,
”a kitchen window frames sun-washed herbs, bowls, and warm shelves.”,
”an orchard bench plaque reading: ‘savor sweet’. blossoms float in warm breezes.”,
”a windy hillside covered in tall dry grass, each stalk catching light differently, distant sheep forming small irregular white clusters”,
”a river winds beside a sleepy village, reflecting pale morning skies and drifting willow branches.”,
”a stormy coastline where winds whip through rugged rocks and dark water.”,
”a quiet european street lined with stone buildings glows under early dawn light as café chairs sit empty on cobblestones.”,
”a quiet university quad filled with shaded benches and tree-lined paths.”,
”a calm river flows beside a small town, reflecting the pale sky while fishermen prepare their nets and willow branches trail across the water’s surface.”,
”a serene lake mirrors the surrounding pines and layered mountain ridges during early dawn.”,
”a remote mountain lake reflects drifting clouds and jagged peaks.”,