Spectral Image Tokenizer

Supplementary Material

A. Ablation study

We conduct an ablation study to evaluate the effects of our design decisions. Starting from the model denoted "SIT-4" in Tab. 1, we train SIT using various number of scales, sequences lengths, vocabulary sizes, wavelet families, and using scale-causal attention on the encoder and/or decoder. Tab. 5 shows the results.

We note that while increasing the vocabulary size or sequence length improves reconstruction, it is at the cost of making the task harder for generation so these tokenizers did not actually lead to better generative models. This has also been observed by prior and concurrent work [4, 19, 36, 54].

Another interest result is that the simple Haar wavelet performs better than the ones that are widely used for compression: LeGall5/3 and CDF9/7. We hypothesize it's related to the larger filter supports, since our results show worse reconstruction the larger the support. It could be due to increased padding and leakage of information between neighboring patches; the same trend is observed for both reflective and zero padding. Some related work with wavelets in diffusion models and GANs also use Haar [17, 35], but they did not report ablations.

B. Implementation details

B.1. Multiscale reconstruction

For the multiscale image reconstruction experiments in Sec. 5.1, we train SIT following ViT-VQGAN [57] closely. We use the "small" encoder and "small" decoder as described in ViT-VQGAN [57]. Each transformer layer comprises a layer norm, self-attention, residual connection, followed by an MLP with layer norm, two dense layers, and a second residual connection. The "small" configuration has 8 layers, 8 heads, feature dimension of 512 that increases to 2048 in the MLP hidden layer (the hidden dimension), then back to 512. Learnable positional embedding is added to the transformer input. The codebook size is 8192 for both codebooks (approximation and details).

The tokenizer is trained for $500\,000$ steps with a learning rate that linearly increases up to 1×10^{-4} during the first 10% steps, then decays following a cosine schedule. We use batch size 256 and L2 weight decay with a 1×10^{-4} factor. The loss components are weighted as follows: 1.0 for L2, 0.1 for perceptual, 0.1 for adversarial, and 0.25 for codebook commitment.

B.2. Coarse-to-fine text-to-image generation

For the text-to-image generation experiments in Sec. 5.2, we train SIT models with a "small" encoder and "large" decoder as defined in ViT-VQGAN [57] and used by Parti [58]. The "small" transformer has 8 layers with 8 heads, feature dimension 512 and hidden dimension 2048. The "large" transformer has 32 layers with 16 heads, feature dimension 1280 and hidden dimension 5120. Differently from Parti, we train this tokenizer once from scratch instead of in two stages; Parti first trains a "small" encoder and decoder, then freezes the "small" encoder and trains "large" decoder.

For the generative AR-SIT, we follow the smallest architecture presented in Parti [58], with 12 layers in the text encoder and 12 decoder layers, 16 heads, 1024 feature dimensions and 4096 hidden dimensions. The text conditioning is through cross-attention. We use classifier-free guidance (CFG) [21] scale of 3.0. No reranking is used.

AR-SIT is trained for $500\,000$ steps with a learning rate that linearly increases up to 4.5×10^{-4} during the first 10% steps, then decays exponentially. We use batch size 256 and the loss is the just the softmax cross-entropy.

B.3. Class-conditional generation

For the class-conditional image generation fair comparison experiments in Sec. 5.5, we again follow ViT-VQGAN closely and train SIT-4 with "small" encoder and "small" decoder as defined in Sec. B.1, and AR-SIT-4 is based on VIM-Base with 24 transformer layers, 16 heads, 1536 model dimension, 6144 hidden dimension, and dropout ratio of 0.1. The model is trained for $500\,000$ steps with a learning rate that linearly increases up to 4.5×10^{-4} during the first 10% steps, then decays exponentially. We use batch size 1024. No CFG or reranking is used.

AR-SIT-4* modifies both the tokenizer and autoregressive models and training. The main architecture improvements are the use of GeGLU activations on all MLPs and axial 2D RoPE for both the tokenizer and AR model. The tokenizer has a latent dimension of 8, vocabulary size of $16\,384$ for the approximation and details codebooks, is trained for $300\,000$ steps with a batch size of 128, constant learning rate of 1×10^{-4}, L2 weight decay of 0.05, and loss weights of 1.0 for L2, 1.0 for LPIPS, 0.5 for adversarial, 0.25 for codebook commitment. The adversarial training starts at $20\,000$ steps. The AR model has 24 transformer layers, 16 heads, 1024 model dimension, 4096 hidden dimension, dropout ratio of 0.1, and is trained for 1.2M steps with batch size 256 and learning rate of 1×10^{-4}. We weight

Input resolution	FID	IS	PSNR
32×32	6.19	32.46	19.05
16×16	7.53	31.75	16.74

Table 4. Image upsampling metrics on MS-COCO [27].

the loss such that each scale has a weight 4x smaller than the previous and found that it improves performance slightly.

B.4. Metrics

Throughout our experiments we report the following metrics:

- FID - Fréchet Inception Distance [20] estimates the distance between the distribution of generated and ground truth features obtained from a pre-trained inception model, by assuming such distributions are Gaussians and applying the Fréchet distance.
- IS - Inception Score (IS) [41] measures the entropy of a pre-trained classifier on the generated images, where low entropy is expected for good quality images.
- LPIPS - Learned Perceptual Image Patch Similarity [61] measures the distance between visual features reconstructed and ground truth images, where the features come from a pre-trained model.
- PSNR - Peak Signal-to-Noise Ratio is a pixel-wise similarity metric, the negative log of the mean-squared error.

FID and IS evaluate the quality of generated images without a corresponding ground truth, while LPIPS and PSNR are used when we have a ground truth, for example when evaluating the tokenizer.

C. Additional results

C.1. Multiscale reconstruction

Fig. 7 shows reconstruction samples at multiple resolutions, for the experiments described in Sec. 5.1.

C.2. Text-guided upsampling

In this section, we show additional results to the text-guided image upsampling experiments from Sec. 5.3. We evaluate the same task on the more challenging case where we upsample a 16×16 input to 256×256. Tab. 4 shows the metrics and Fig. 8 shows some examples.

C.3. Additional editing results

Fig. 6 shows additional results for the text-guided image editing experiment described in Sec. 5.4.

C.4. Class-conditional generated samples

Fig. 9 shows samples of class-conditional 256×256 ImageNet generation, and Fig. 10 shows 512×512 ImageNet generations as described in Sec. 5.5.

model	num scales	seq len	vocab	wavelet	SCE	SCD	LPIPS ↓	PSNR ↑	FID ↓	IS ↑
ViT-VQGAN	-	1024	8192	-	-	-	0.164	23.76	1.20	194.6
SIT-4	4	1024	8192	Haar	X	X	0.143	24.01	1.20	199.5
SIT-5	5	1280	8192	Haar	X	X	0.135	24.48	0.97	202.3
SIT-4	4	1024	8192	Haar	X	✓	0.166	23.51	1.45	191.3
SIT-4	4	1024	8192	Haar	✓	✓	0.184	23.41	1.97	179.8
SIT-5	5	1280	8192	LeGall5/3	X	X	0.145	24.35	1.10	198.0
SIT-5	5	1280	8192	CDF9/7	X	X	0.147	24.31	1.14	198.1
SIT-5	5	1280	4096	Haar	X	X	0.155	23.66	1.33	193.9
SIT-5	5	1280	16384	Haar	X	X	0.134	24.62	0.91	203.0
SIT-2	2	2048	8192	Haar	X	X	0.097	26.21	0.77	212.0
SIT-3	3	768	8192	Haar	X	X	0.186	23.12	2.04	176.8
SIT-2	2	512	8192	Haar	X	X	0.244	20.93	4.30	141.2
SIT-6	6	384	8192	Haar	X	X	0.273	20.72	6.25	129.9
SIT-5	5	320	8192	Haar	X	X	0.283	20.51	7.76	118.0

Table 5. Ablation study. We report ImageNet reconstruction metrics for SIT variations, following the ViT-VQGAN protocol from Tab. 1. We evaluate the effect of number of scales, sequence length, vocabulary size, and wavelet family. We also evaluate the scale-causal attention on the encoder (SCE) and decoder (SCD), while it generally reduces reconstruction accuracy, it enables the various multiscale properties demonstrated in the text.

Figure 6. Additional results for text-guided image editing on MS-COCO [27]. Each triplet shows the given image, its reconstruction given only the coefficients used to start the generation, and the edited image after generating the whole sequence. The guiding prompt is shown under each triplet.

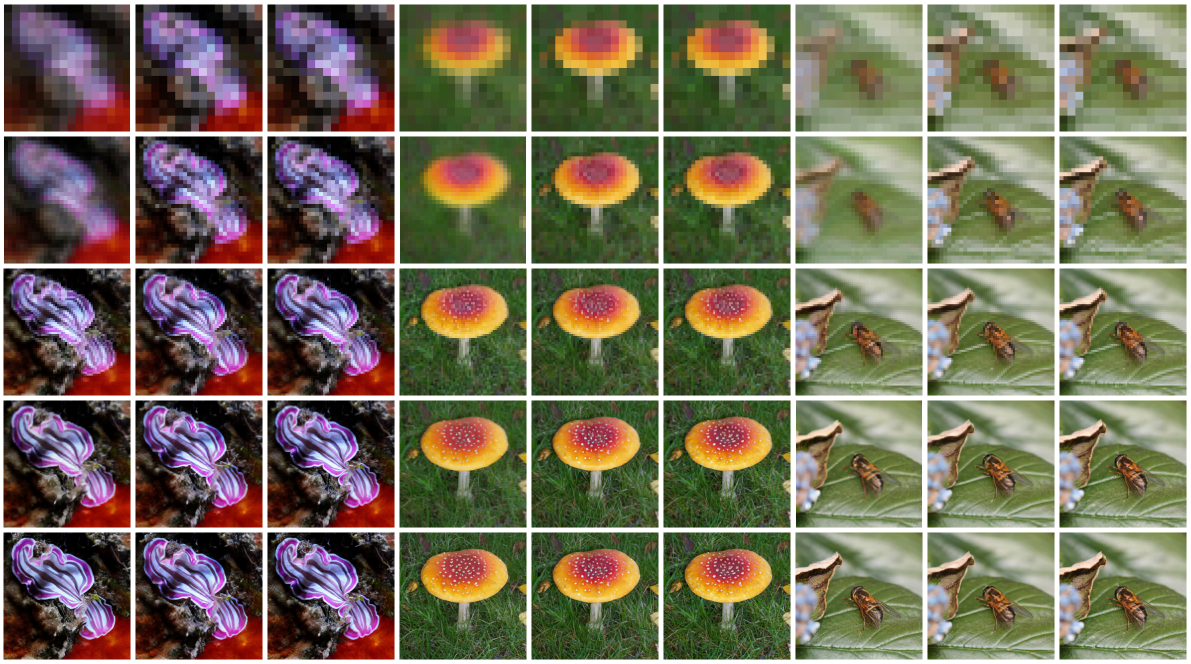

Figure 7. Multiscale reconstruction on ImageNet. Each triplet shows reconstruction from the ViT-VQGAN baseline [57], our SIT-SC-5 (Spectral Image Tokenizer with Scale-Causal attention and 5 scales), and the ground truth. Each row shows $4\times$ as many pixel inputs as the previous, with the first row corresponding to 16×16 resolution, and the last to 256×256. Our method is naturally multiresolution, significantly outperforming the baseline on lower resolutions even when trained only on 256×256 inputs, while achieving similar accuracy on higher resolutions.

Figure 8. Additional text-guided image upsampling results. Here we consider the more challenging task of upsampling from 16×16 to 256×256. Each triplet shows the given 16×16 image, our 256×256 reconstruction and the ground truth.

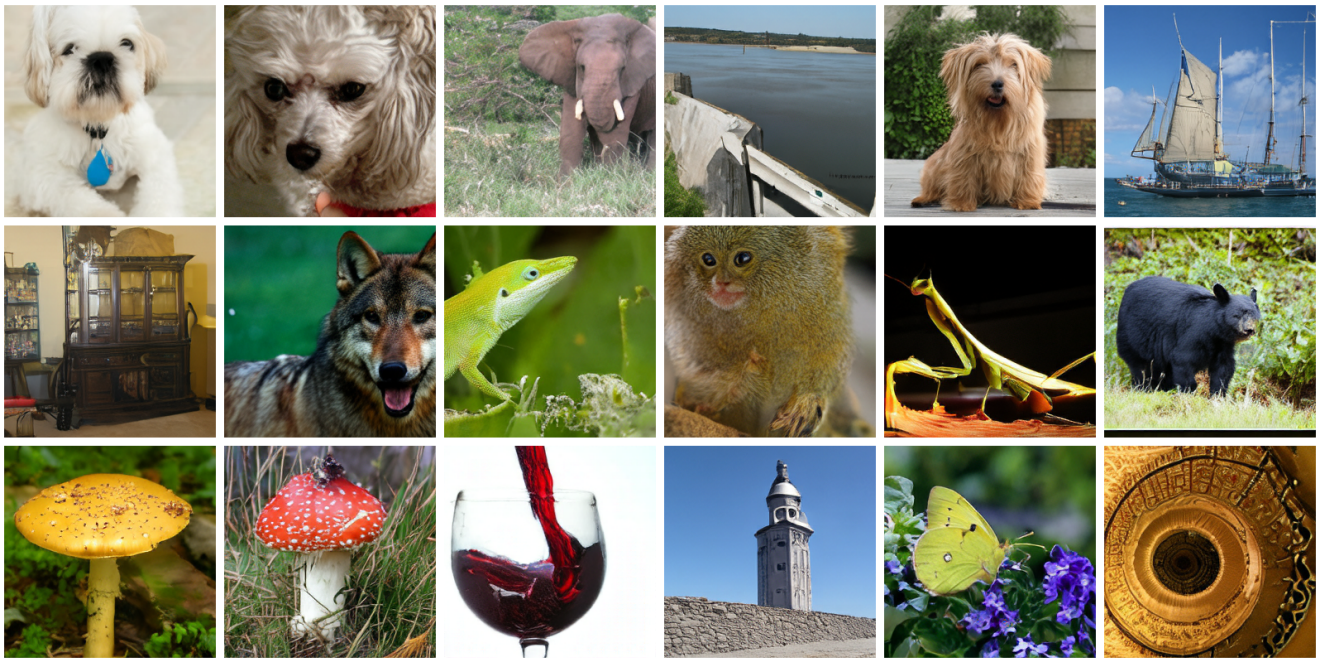

Figure 9. Samples of class-conditional generation with AR-SIT-4 on 256×256 ImageNet.

Figure 10. Samples of class-conditional generation with AR-SIT-5* on 512×512 ImageNet.