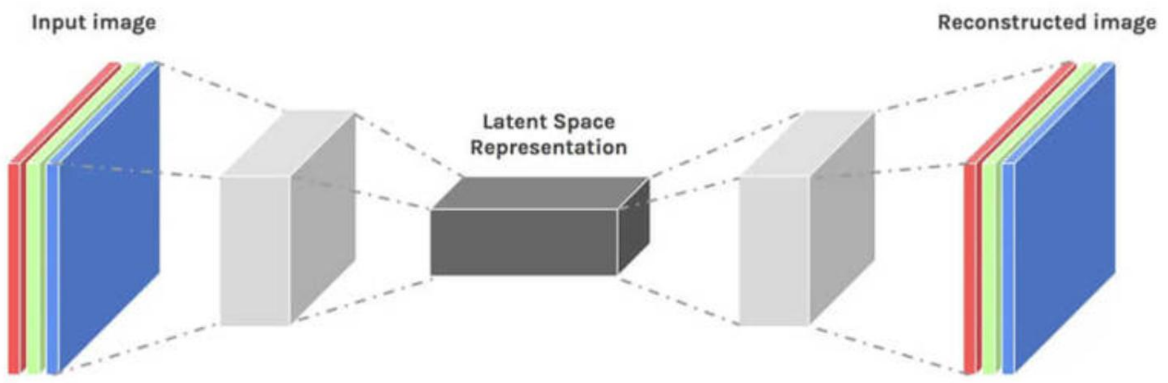
DCTdiff: Intriguing Properties of Image Generative Modeling in the DCT Space

Mang Ning, Mingxiao Li*, Jianlin Su*, Haozhe Jia, Lanmiao Liu, Wenshuo Chen, Martin Beneš,
Albert Ali Salah, Itir Onal Ertugrul

Motivation

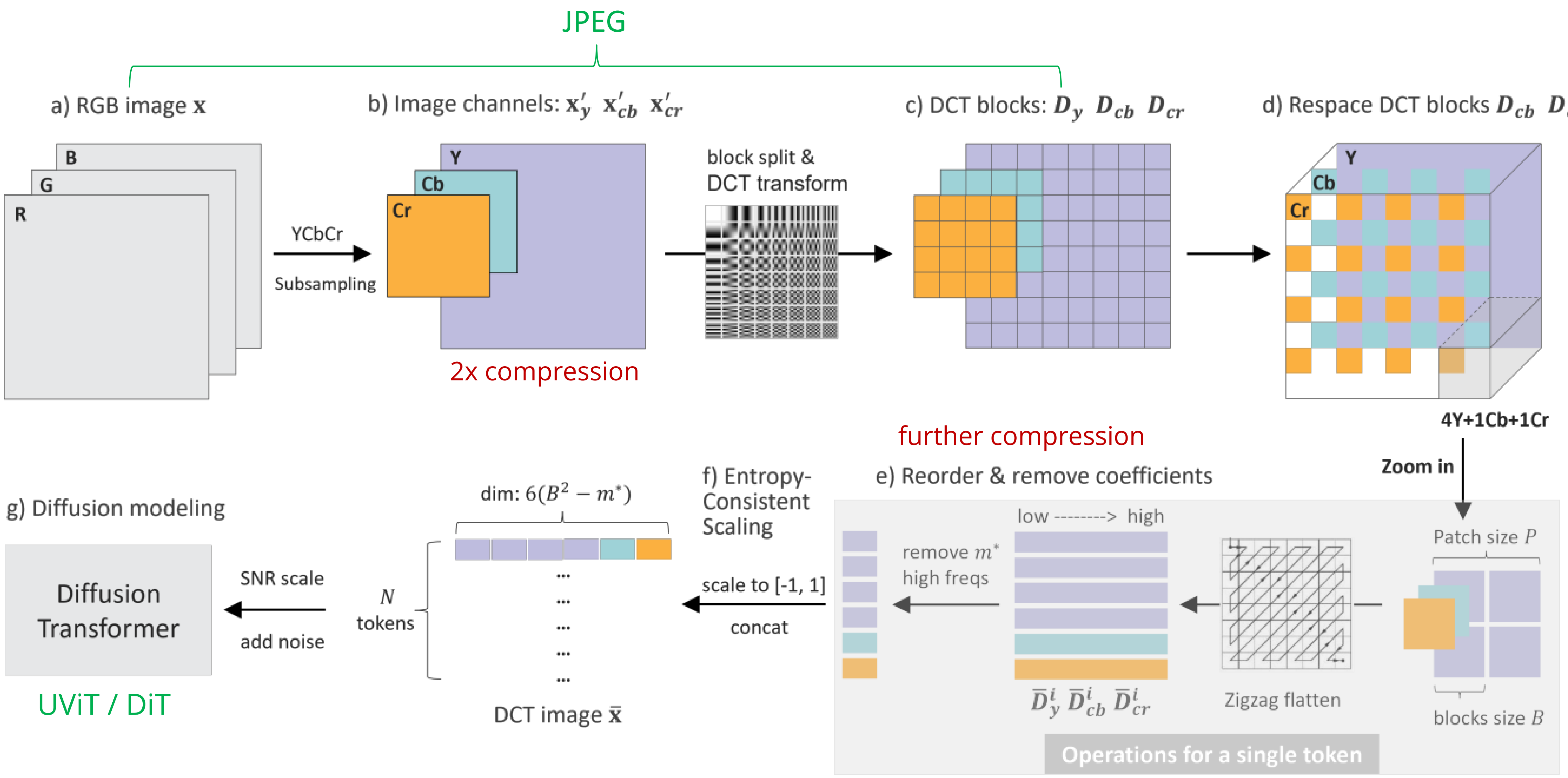
Image modeling in the RGB space vs. images are stored in a compressed form (DCT, DEFLATE)

High resolution (≥256x256) image generation relies on Latent Diffusion (SD3, Flux, Imagen3, DALLE-3)



Can we perform image modeling in a (near) lossless compression space ? → DCT space

Architecture of DCTdiff



Results (faster & better)

Low-resolution (from 32x32 to 128x128)

NFE	Model	Euler ODE solver (DDIM sampler)				DPM-Solver			
		CIFAR-10	CelebA 64	ImageNet 64	FFHQ 128	CIFAR-10	CelebA 64	ImageNet 64	FFHQ 128
100	UViT	6.23	1.99	10.65	13.87	5.80	**1.57**	10.07	9.18
	DCTdiff	**5.02**	**1.91**	**8.69**	**8.22**	**5.28**	1.71	**9.73**	**6.25**
50	UViT	7.88	3.50	15.05	26.26	5.82	**1.58**	10.09	9.20
	DCTdiff	**5.21**	**2.24**	**8.70**	**9.99**	**5.30**	1.72	**9.78**	**6.28**
20	UViT	21.48	31.09	52.10	87.68	6.19	**1.73**	10.25	9.21
	DCTdiff	**6.81**	**3.84**	**21.88**	**24.88**	**5.54**	1.84	**9.85**	**7.29**
10	UViT	81.67	224.21	166.63	209.69	26.65	**4.37**	13.27	14.26
	DCTdiff	**12.45**	**67.78**	**129.93**	**161.05**	**9.10**	5.29	**12.38**	**12.87**

High-resolution (256x256, 512x512)

NFE	Model	Dataset		
		FFHQ 256	FFHQ 512	AFHQ 512
100	UViT (latent)	**4.26**	10.89	10.86
	DCTdiff	5.08	**7.07**	**8.76**
50	UViT (latent)	**4.29**	10.94	10.86
	DCTdiff	5.18	**7.09**	**8.87**
20	UViT (latent)	**4.74**	11.31	11.94
	DCTdiff	6.35	**8.04**	**10.05**
10	UViT (latent)	13.29	23.61	28.31
	DCTdiff	**12.05**	**19.67**	**21.05**

Less training cost

Dataset	Model	# Parameters	GFLOPs	Training steps
CelebA 64	UViT	44M	11	400k
	DCTdiff	44M	11	**250k**
FFHQ 128	UViT	44M	11	750k
	DCTdiff	44M	11	**300k**
FFHQ 256	UViT (latent)	131M + 84M	169	**200k**
	DCTdiff	131M	**133**	300k
AFHQ 512	UViT (latent)	131M + 84M	575	225k
	DCTdiff	131M	**133**	**225k**

23% cost

Property: Frequency Prioritization

Generative tasks:

- **RGB:** which pixel is more important than another pixel ?
- **DCT:** low-frequency signal contributes more to the image quality than a high-frequency signal

$$\mathbb{E}_t \lambda(t) \mathbb{E}_{\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_t} [\boldsymbol{H}(B) || \boldsymbol{s}_{\boldsymbol{\theta}}(\bar{\mathbf{x}}_t, t) - \nabla_{\bar{\mathbf{x}}_t} \log P_{0t}(\bar{\mathbf{x}}_t | \bar{\mathbf{x}}_0) ||_2^2]$$

reweigting

Discriminative tasks

- **DCT:**
  - High frequencies (medical image analysis, forgery detection)
  - Low frequencies (scene recognition, action recognition)

Property: Significant Compression

Generative tasks

- DCT enables flexible and domain-agnostic compression
- rFID = 0.5 as near lossless compression
- 4x compression on 256*256
- 7x compression on 512*512

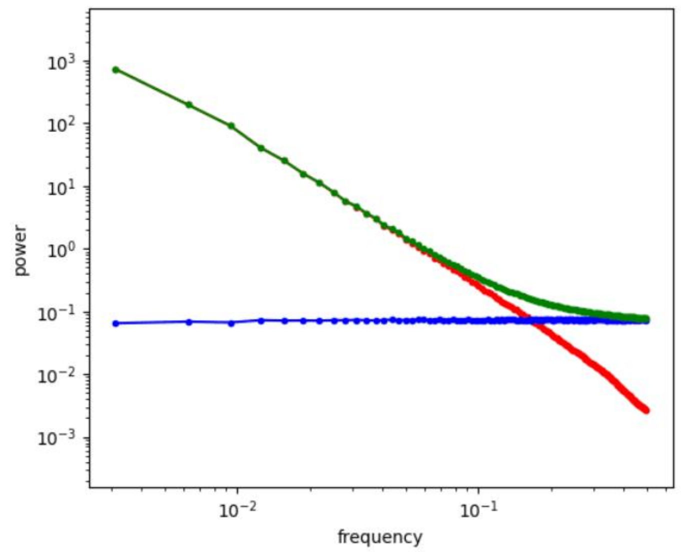
Dataset	Block size	m	rFID	Compression ratio
FFHQ 256×256	4	7	0.19	3.56
		8	**0.49**	**4.00**
		9	0.96	4.57
FFHQ 512×512	8	44	0.23	6.40
		46	**0.48**	**7.11**
		48	1.18	8.00

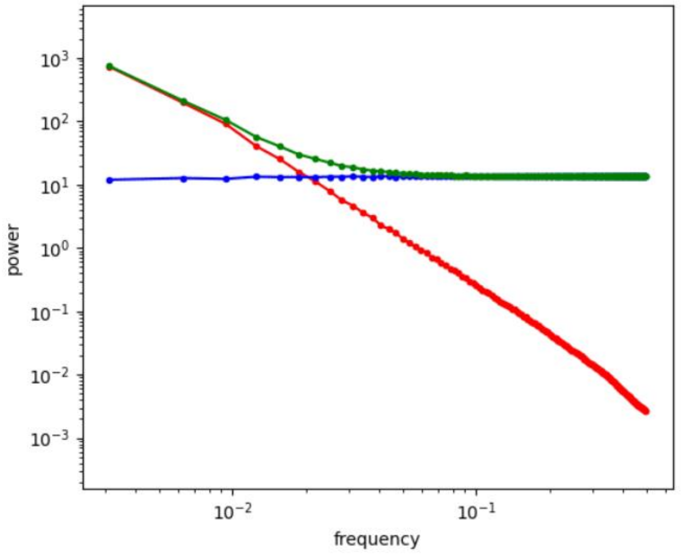
Discriminative tasks

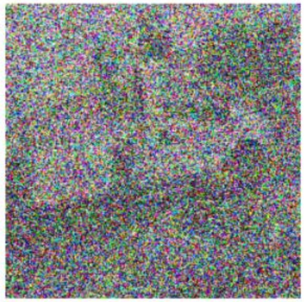
- Higher compression is possible

NFE	Model	Dataset	
		FFHQ 512	AFHQ 512
100	UViT (latent)	10.89	10.86
	DCTdiff	**7.07**	**8.76**
50	UViT (latent)	10.94	10.86
	DCTdiff	**7.09**	**8.87**
20	UViT (latent)	11.31	11.94
	DCTdiff	**8.04**	**10.05**
10	UViT (latent)	23.61	28.31
	DCTdiff	**19.67**	**21.05**

Property: Image Diffusion Is Spectral Autoregression





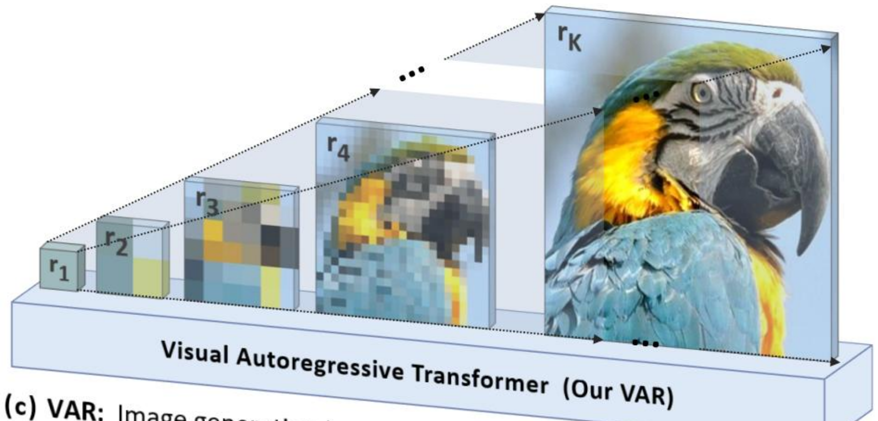


We provide a formal proof:

Theorem 5.1. *Consider a diffusion model described by* $\mathrm{d}\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, t)\mathrm{d}t + g(t)\mathrm{d}\boldsymbol{w}_t$. *Let* ω *denote the frequency,* $\hat{\boldsymbol{x}}_0(\omega)$ *and* $\hat{\boldsymbol{x}}_t(\omega)$ *represent the Fourier transform of the pixel image* $\boldsymbol{x}_0$ *and* $\boldsymbol{x}_t$*, respectively. The averaged power spectral density of the noisy image* $\boldsymbol{x}_t$ *satisfies:*

$$\mathbb{E}\left[|\hat{\boldsymbol{x}}_t(\omega)|^2\right] = |\hat{\boldsymbol{x}}_0(\omega)|^2 + \int_0^t |g(s)|^2 \mathrm{d}s \qquad (11)$$

in which $|\hat{\boldsymbol{x}}_0(\omega)|^2$ *is the power spectral density of the image* $\boldsymbol{x}_0$ *and natural images have the power-law:* $|\hat{\boldsymbol{x}}_0(\omega)|^2 = K|\omega|^{-\alpha}$ *(Ruderman, 1997)(*K *and* α *are constants). Meanwhile,* $\int_0^t |g(s)|^2 \mathrm{d}s$ *is independent of frequency* ω *and appears as a horizontal line in the spectral density graph.*



(c) VAR: Image generation by **next-scale (or next-resolution)** prediction

Takeaways

- Image modeling in the DCT space is efficient (512x512 generation without VAE)
- DCT space is underexplored, and has promising directions
  - Spectral bias in NN
  - Image → Video
  - Representation learning (MIM)
  - Network architecture (MoE)

Samples generated by DCTdiff trained on AFHQ 512×512