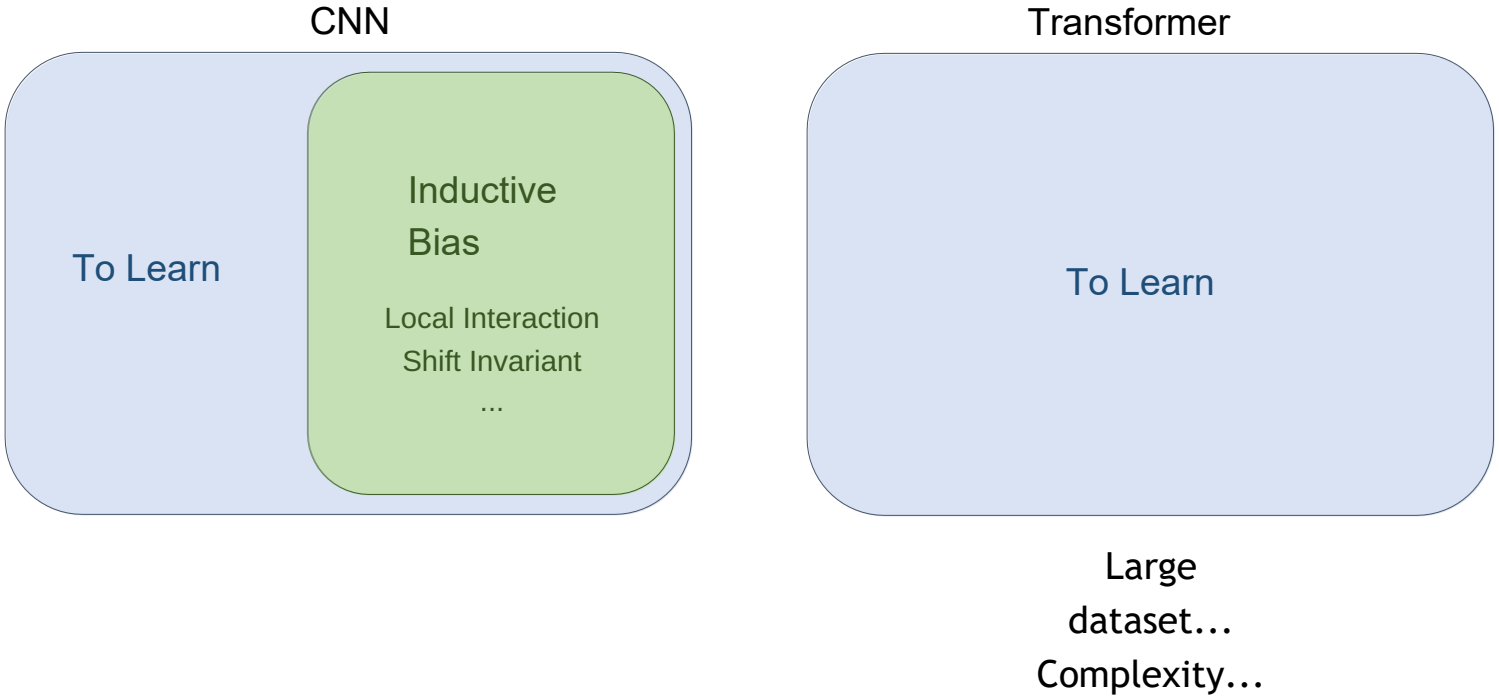
Taming Transformers for High-Resolution Image Synthesis (VQ-GAN)

Sathvik Reddy

Transformers for Image Synthesis

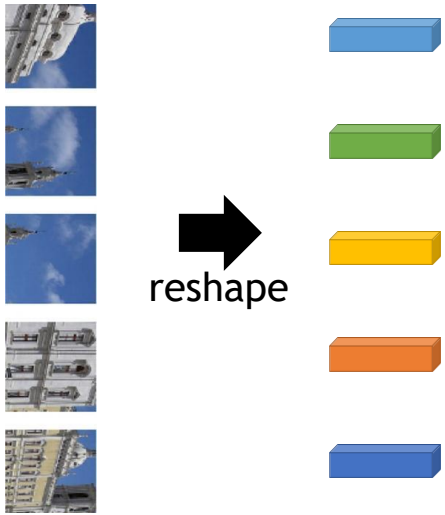
Transformers for Image Synthesis

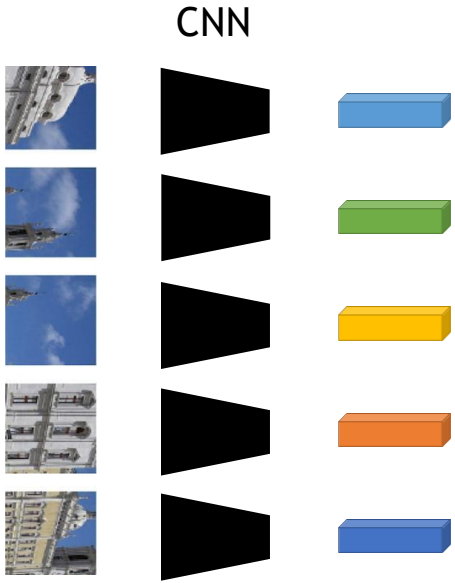
- CNNs vs Transformers



Transformers for Image Synthesis

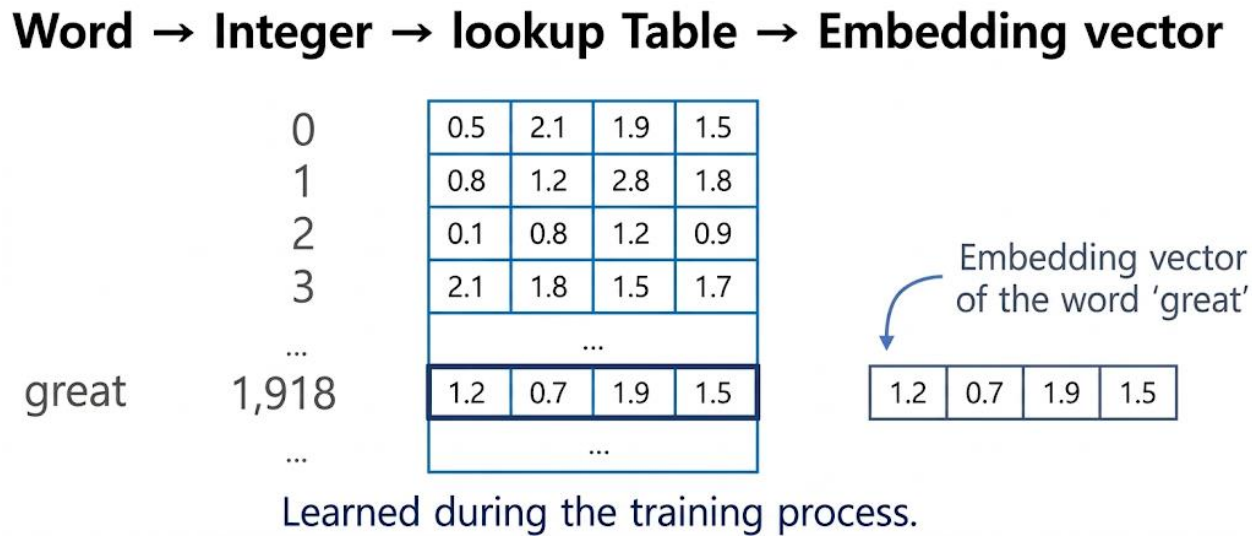
- How to encode Images to Vectors





Transformers for Language

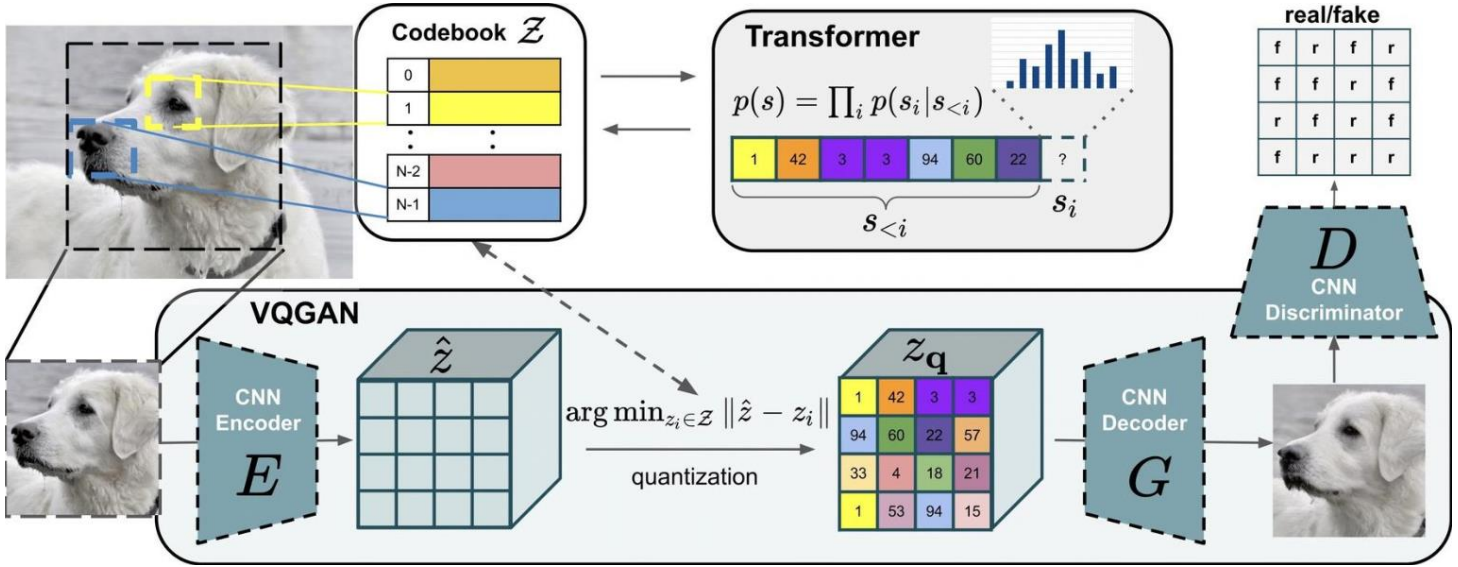
- Discrete Codebook (Lookup Table)



VQGAN

VQGAN

- Encoder & Decoder

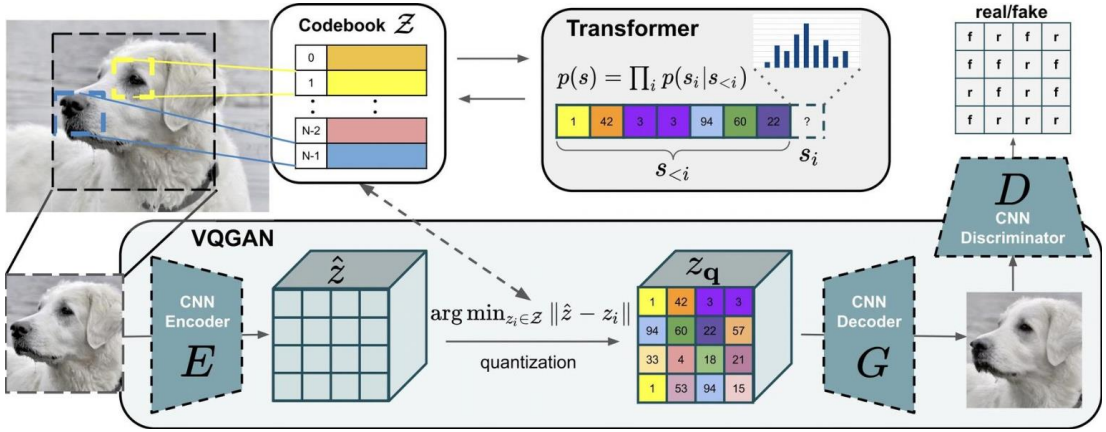


Encoding - Vector Quantization - Reconstruction

VQGAN

- Encoder & Decoder

Reconstruction Loss



$$z_\mathbf{q} = \mathbf{q}(\hat{z}) := \left(\underset{z_k \in \mathcal{Z}}{\arg\min} \|\hat{z}_{ij} - z_k\| \right) \in \mathbb{R}^{h \times w \times n_z}$$

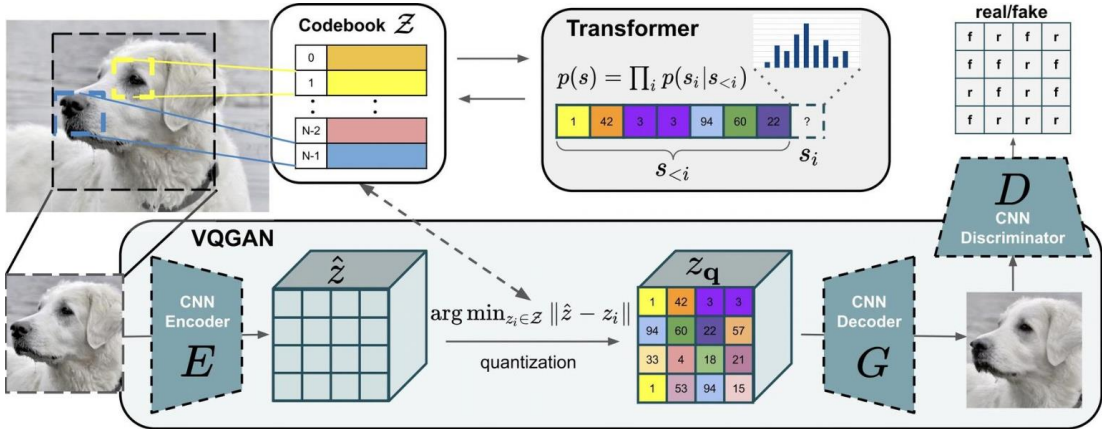
$$\hat{x} = G(z_\mathbf{q}) = G\left(\mathbf{q}(E(x))\right)$$

$$\mathcal{L}_{\text{rec}} = \|x - \hat{x}\|^2$$

VQGAN

- Encoder & Decoder

Losses for Encoder and Codebook

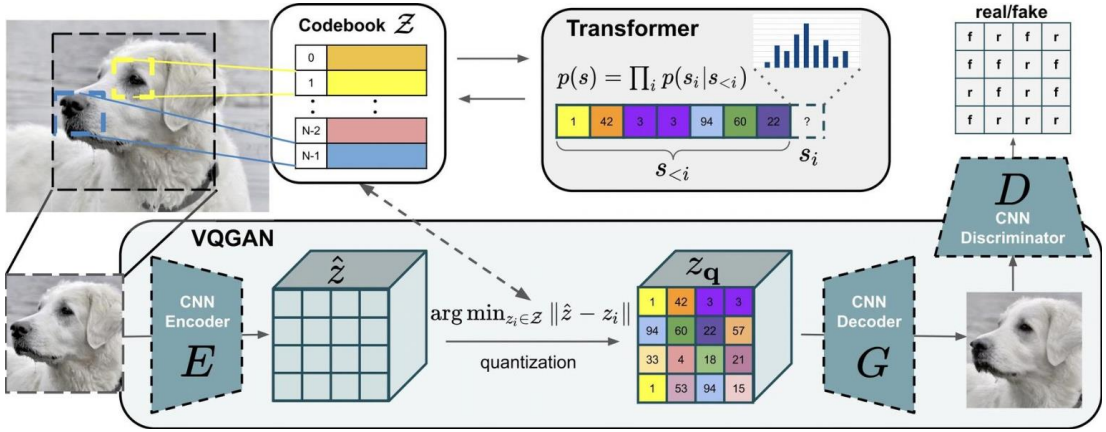


$$\mathcal{L}_{\text{VQ}}(E, G, \mathcal{Z}) = \|x - \hat{x}\|^2 + \|\text{sg}[E(x)] - z_{\mathbf{q}}\|_2^2 + \|\text{sg}[z_{\mathbf{q}}] - E(x)\|_2^2$$

VQGAN

- Encoder & Decoder

Adversarial Loss



$$\mathcal{L}_{\text{GAN}}(\{E, G, \mathcal{Z}\}, D) = [\log D(x) + \log(1 - D(\hat{x}))$$

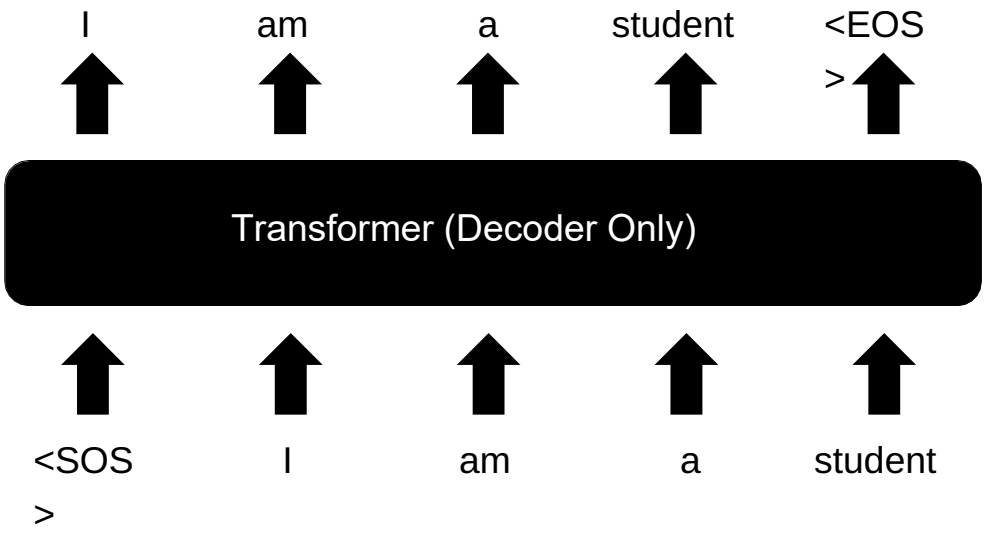
$$\mathcal{Q}^* = \arg\min_{E,G,\mathcal{Z}} \max_D \mathbb{E}_{x \sim p(x)} \Big[\mathcal{L}_{\text{VQ}}(E, G, \mathcal{Z}) + \lambda \mathcal{L}_{\text{GAN}}(\{E, G, \mathcal{Z}\}, D) \Big]$$

$$\lambda = \frac{\nabla_{G_L}[\mathcal{L}_{\text{rec}}]}{\nabla_{G_L}[\mathcal{L}_{\text{GAN}}] + \delta}$$

VQGAN

- Transformer – Unconditional Generation

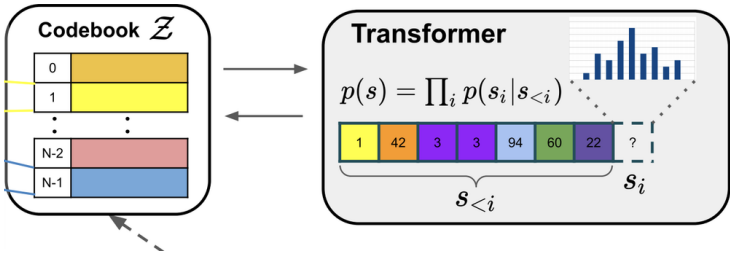
Sentence generation in language (next-word prediction)



VQGAN

- Transformer – Unconditional Generation

Unconditional Image Synthesis



Initial: Pick any Code Vector using RandInt

$$s_{ij} = k \text{ such that } (z_{\mathbf{q}})_{ij} = z_k$$

$$p(s) = \prod_i p(s_i|s_{<i})$$

$$\mathcal{L}_{\text{Transformer}} = \mathbb{E}_{x \sim p(x)} \left[-\log p(s) \right]$$

At Inference: Sampling according to the Softmax distribution

VQGAN

- Transformer – Unconditional Generation

High Resolution Image Generation

$$H \times W \text{ to } h = H/2^m \times w = W/2^m$$

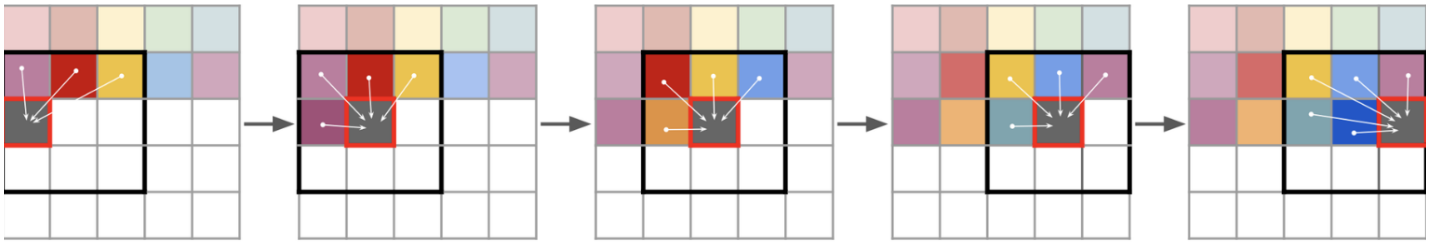
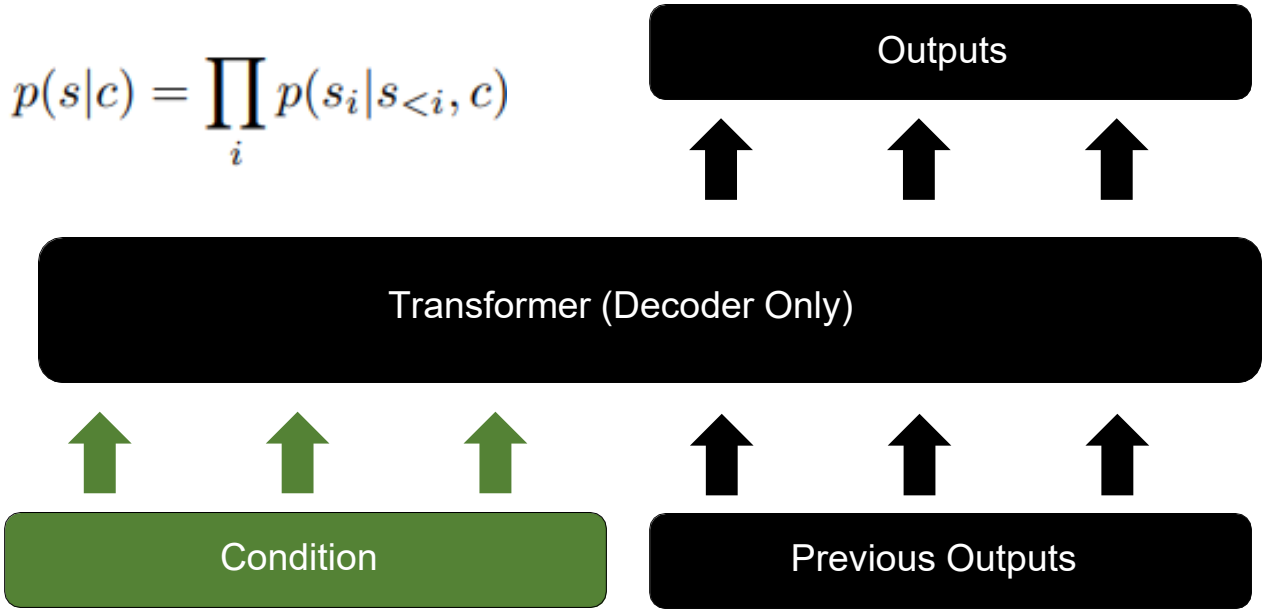


Figure 3. Sliding attention window.

VQGAN

- Transformer – Conditional Generation

$$p(s|c) = \prod_i p(s_i|s_{<i}, c)$$



Experiments

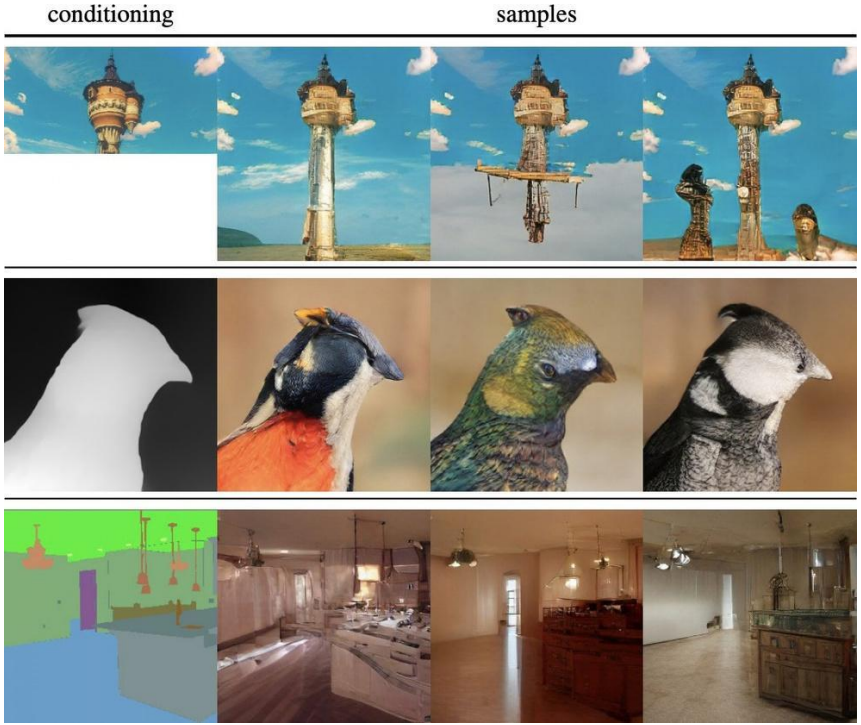
Experiments

1.Does the transformer make a difference?

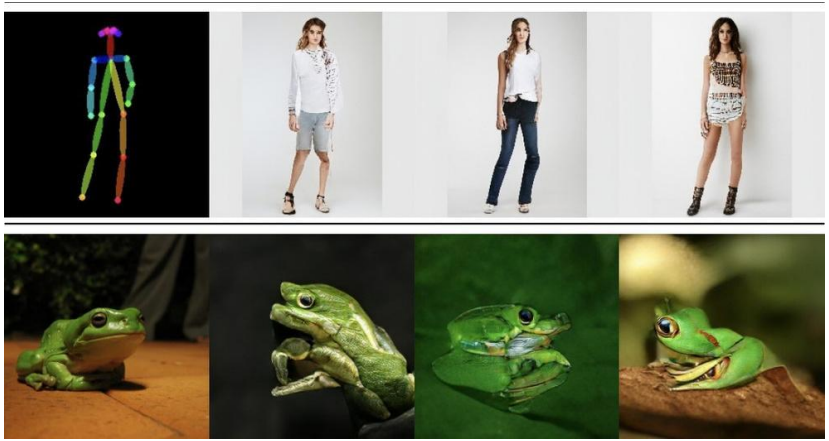
	Negative Log-Likelihood (NLL)		
Data / # params	Transformer *P-SNAIL steps*	Transformer *P-SNAIL time*	PixelSNAIL *fixed time*
RIN / 85M	**4.78**	4.84	4.96
LSUN-CT / 310M	**4.63**	4.69	4.89
IN / 310M	**4.78**	4.83	4.96
D-RIN / 180 M	**4.70**	4.78	4.88
S-FLCKR / 310 M	**4.49**	4.57	4.64

Experiments

2.Conditional Image Synthesis



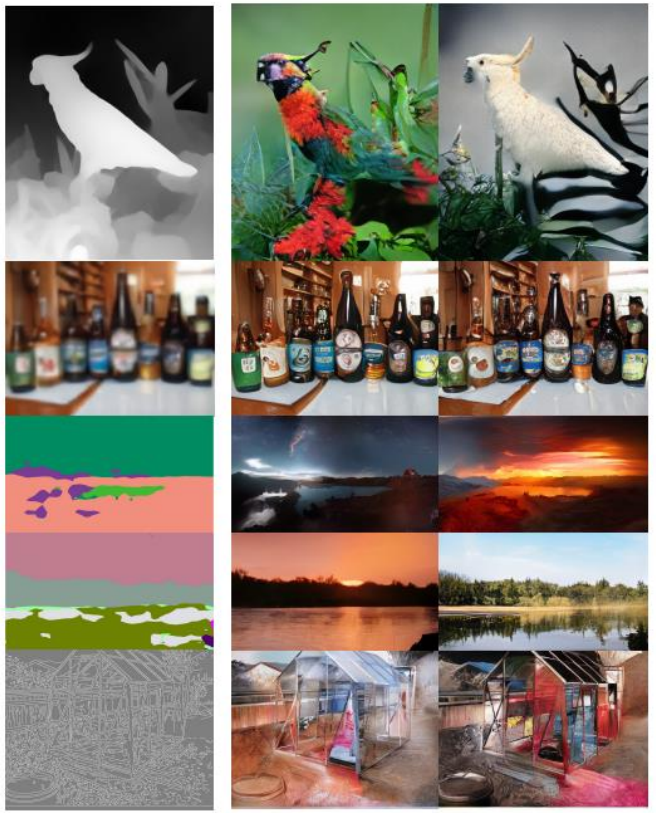
256×256



Experiments

2.Conditional Image Synthesis

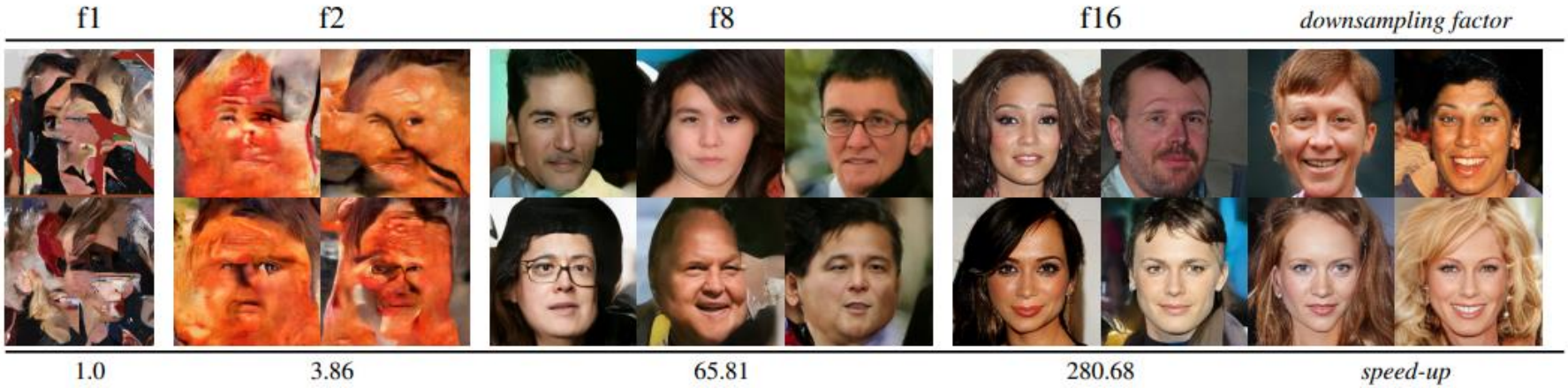
From semantic layouts on S-FLCKR.



High-resolution synthesis from various conditions.

Experiments

3.Importance of Context-rich Vocabulary



$$H \times W \text{ into } H/f \times W/f$$

Experiments

4.Benchmanrking the results

Dataset	**ours**	SPADE [53]	Pix2PixHD (+aug) [75]	CRN [9]
COCO-Stuff	22.4	22.6/23.9(*)	111.5 (54.2)	70.4
ADE20K	35.5	33.9/35.7(*)	81.8 (41.5)	73.3

semantic image synthesis

CelebA-HQ 256 × 256		FFHQ 256 × 256	
Method	FID ↓	Method	FID ↓
GLOW [37]	69.0	VDVAE ($t = 0.7$) [11]	38.8
NVAE [69]	40.3	VDVAE ($t = 1.0$)	33.5
PIONEER (B.) [23]	39.2 (25.3)	VDVAE ($t = 0.8$)	29.8
NCPVAE [1]	24.8	VDVAE ($t = 0.9$)	28.5
VAEBM [77]	20.4	*VQGAN*+P.SNAIL	21.9
Style ALAE [56]	19.2	BigGAN	12.4
DC-VAE [54]	15.8	**ours** (k=300)	9.6
ours (k=400)	10.2	U-Net GAN (+aug) [66]	10.9 (7.6)
PGGAN [31]	8.0	StyleGAN2 (+aug) [34]	3.8 (3.6)

face image synthesis

Model	acceptance rate	FID	IS
mixed k, $p = 1.0$	1.0	17.04	70.6 ± 1.8
$k = 973$, $p = 1.0$	1.0	29.20	47.3 ± 1.3
$k = 250$, $p = 1.0$	1.0	15.98	78.6 ± 1.1
$k = 973$, $p = 0.88$	1.0	15.78	74.3 ± 1.8
$k = 600$, $p = 1.0$	0.05	5.20	280.3 ± 5.5
mixed k, $p = 1.0$	0.5	10.26	125.5 ± 2.4
mixed k, $p = 1.0$	0.25	7.35	188.6 ± 3.3
mixed k, $p = 1.0$	0.05	5.88	304.8 ± 3.6
mixed k, $p = 1.0$	0.005	6.59	402.7 ± 2.9
DCTransformer [48]	1.0	36.5	n/a
VQVAE-2 [61]	1.0	∼31	∼45
VQVAE-2	n/a	∼10	∼330
BigGAN [4]	1.0	7.53	168.6 ± 2.5
BigGAN-deep	1.0	6.84	203.6 ± 2.6
IDDPM [49]	1.0	12.3	n/a
ADM-G, no guid. [15]	1.0	10.94	100.98
ADM-G, 1.0 guid.	1.0	4.59	186.7
ADM-G, 10.0 guid.	1.0	9.11	283.92
val. data	1.0	1.62	234.0 ± 3.9

Class Conditional synthesis

Experiments

Figure 8. Samples from our class-conditional ImageNet model trained on 256×256 images.

Experiments

- Unconditional Generation

Strengths and Weaknesses

Strengths

- Combines the inductive biases of CNNs and the ‘expressivity’ of transformers
- High-Resolution image synthesis

Weaknesses

- Autoregressive Bottleneck: token-by-token generation
- Quantization inherently causes some information loss

Questions?