Fractal Generative Models

Tianhong Li[1] Qinyi Sun[1] Lijie Fan[2] Kaiming He[1]

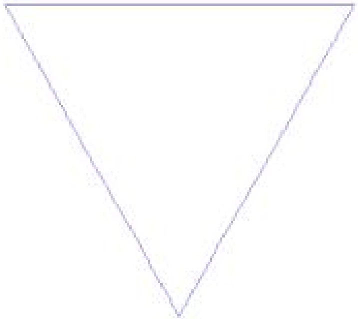
[1]MIT CSAIL [2]Google DeepMind

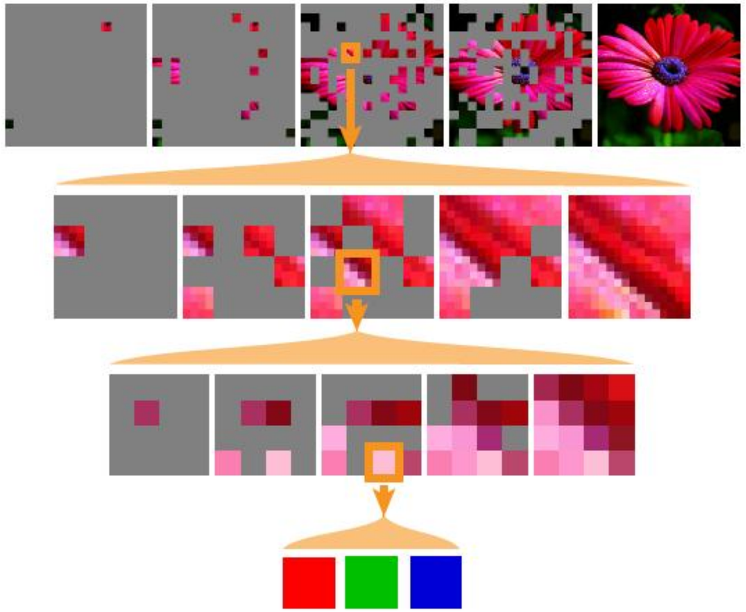
arXiv:2502.17437

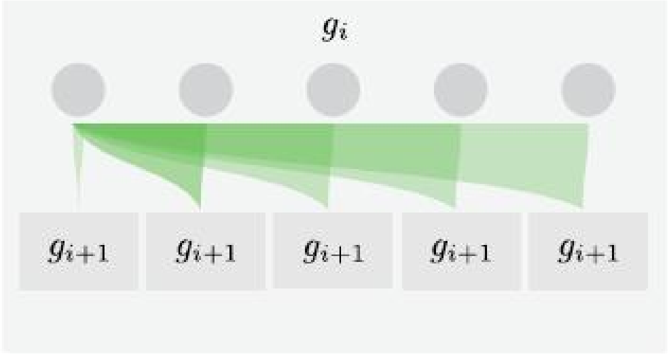
Presenter: Wen Si
2025.3.9

Background: Fractals

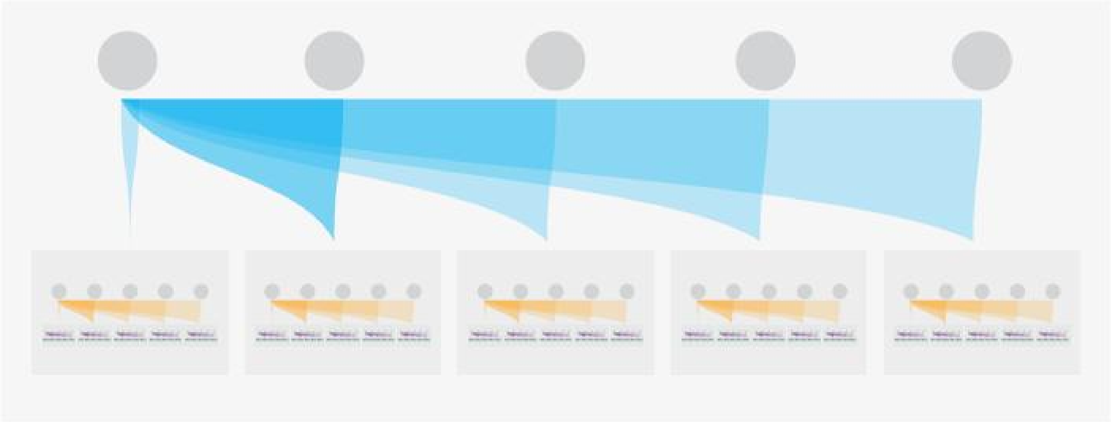
- Self-similarity across different scales
- Recursive generation rule (generator)







(a) generator

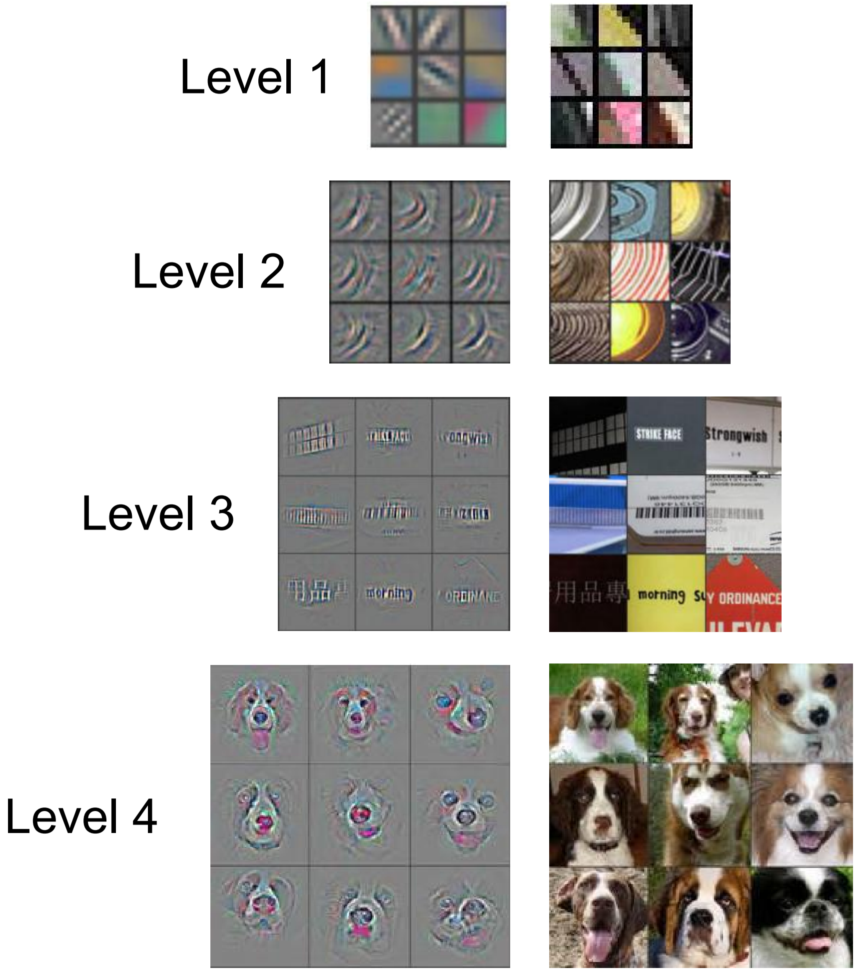


(b) fractal from the generator

Background: Hierarchical Representations

- Break down problems to different scales
- Sharable features across levels
- Deeper layers → Higher levels

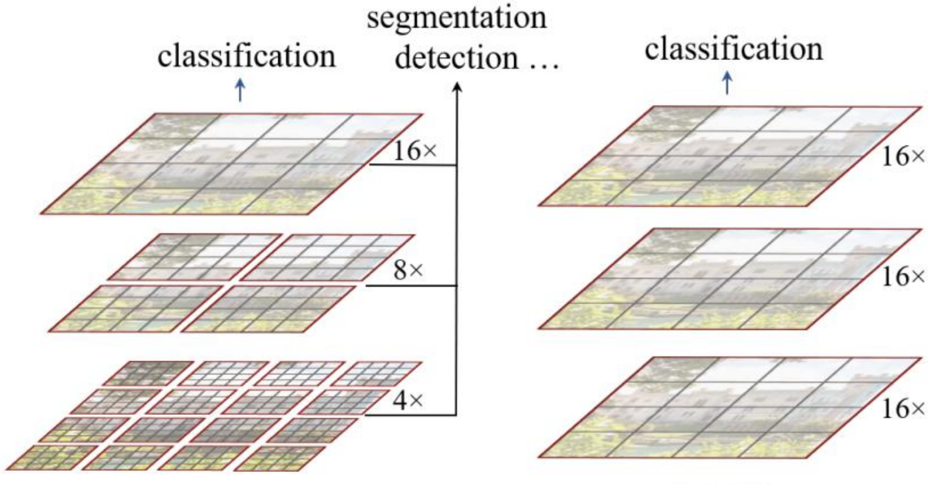
Area	CV	NLP
Low-level Feature	Pixels, Edges	Words, Phrases
Mid-level Feature	Patterns	Syntactic
High-level Feature	Objects	Semantic



Background: Swin Transformers

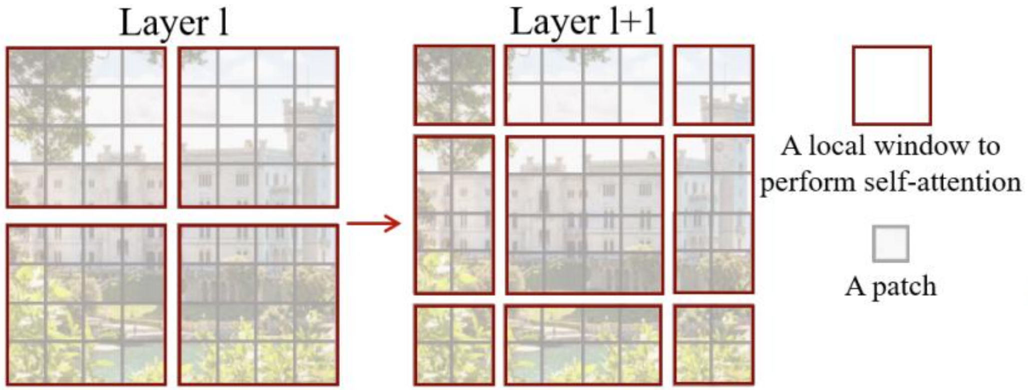
- Localized self-attention
- Linear complexity to image size

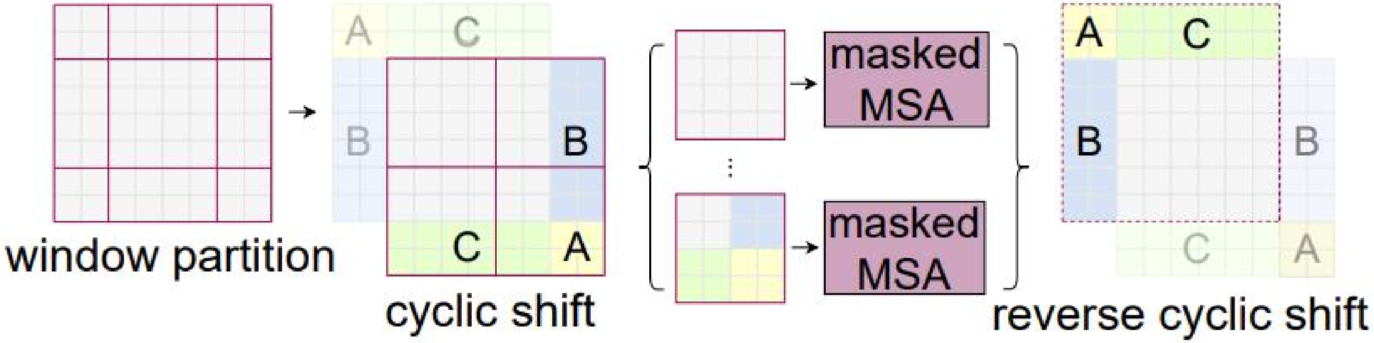
$$\Omega(\text{MSA}) = 4hwC^2 + 2(hw)^2C,$$
$$\Omega(\text{W-MSA}) = 4hwC^2 + 2M^2hwC,$$



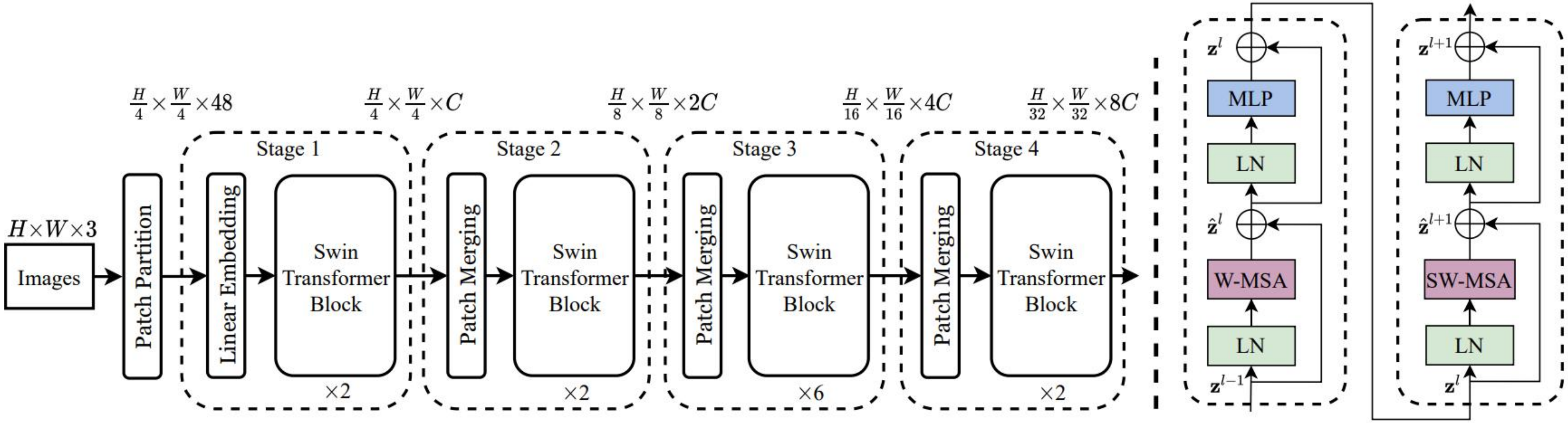
(a) Swin Transformer (ours)

(b) ViT



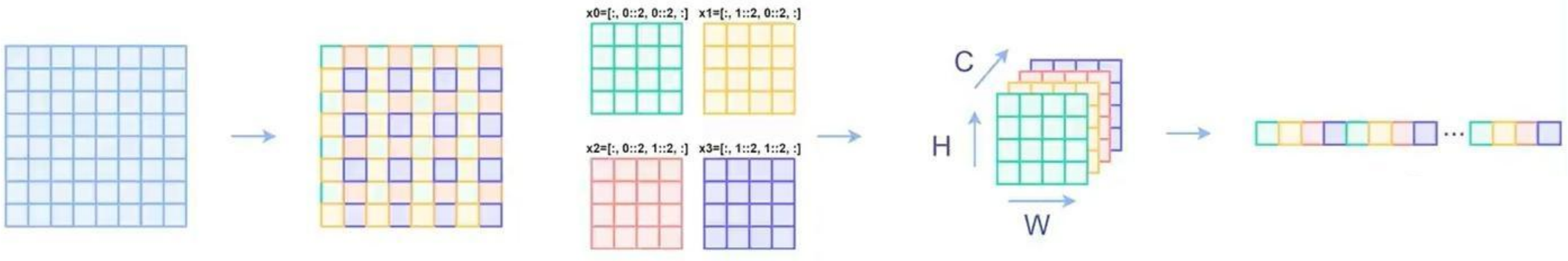


Background: Swin Transformers



(a) Architecture

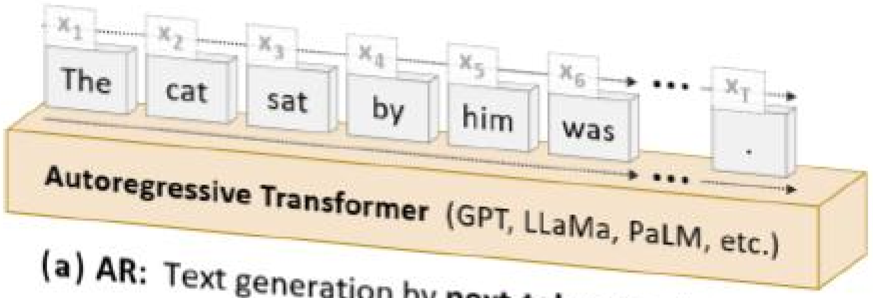
(b) Two Successive Swin Transformer Blocks



Background: Autoregressive Models (AR)

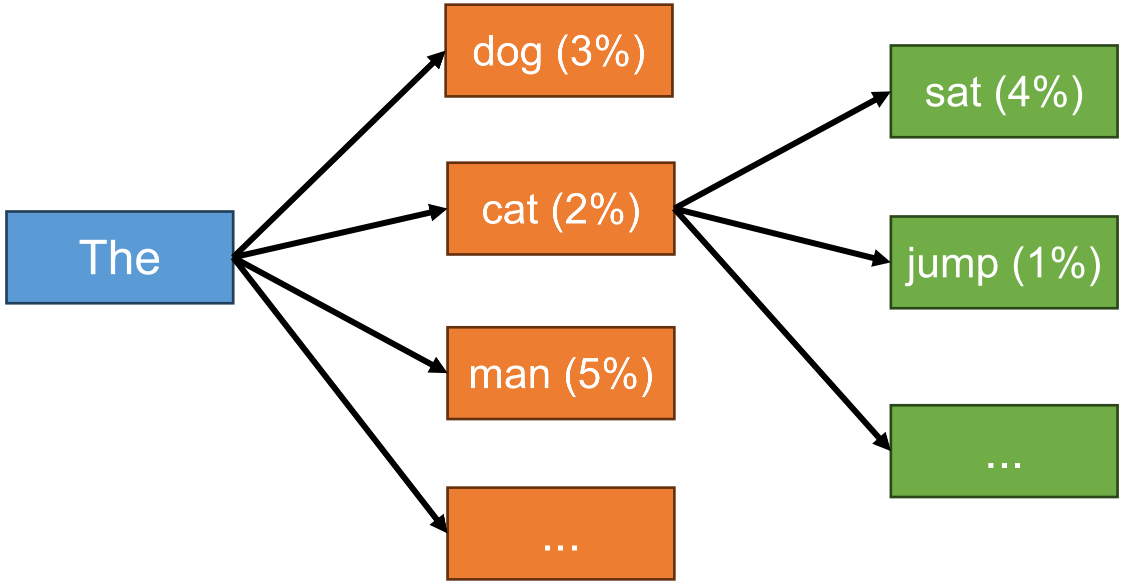
- Predicts next token's probability distribution
- Compatible with sequenced data
- Non-directional token dependency assumption

$$p(x_1, x_2, \ldots, x_T) = \prod_{t=1}^{T} p(x_t \mid x_1, x_2, \ldots, x_{t-1}).$$



(a) AR: Text generation by **next-token** prediction

- Images are bidirectional
- Needs alignment across scales



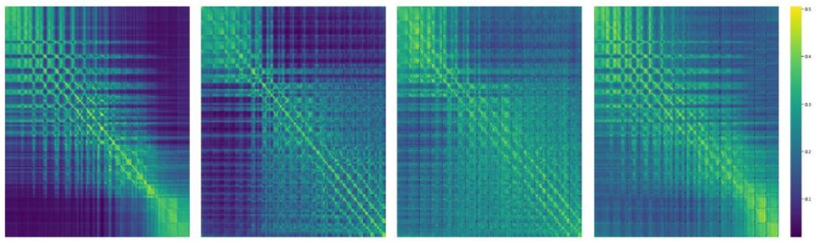
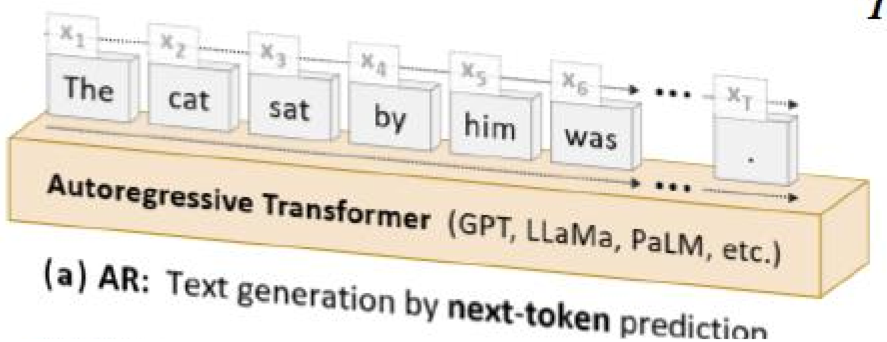


Figure 9: **Token dependency plotted.** The normalized heat map of attention scores in the last self-attention layer of VQGAN encoder is visualized. 4 random 256×256 images from ImageNet validation set are used.

Background: Visual Autoregressive Modeling (VAR)

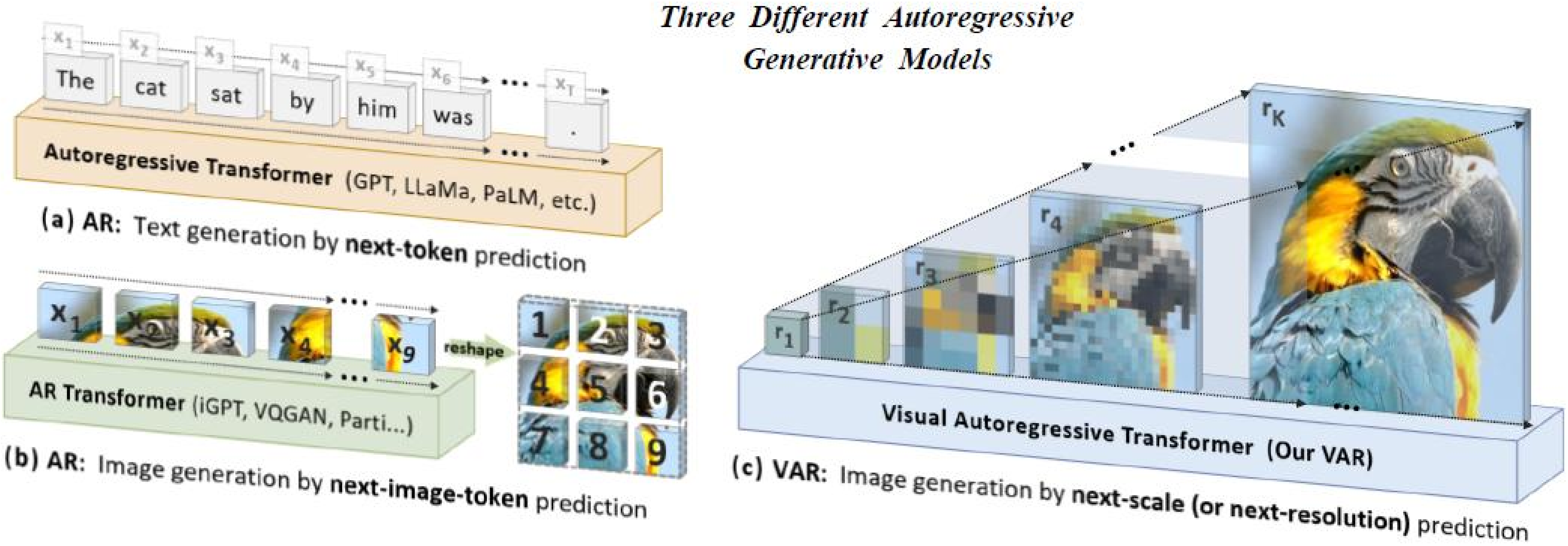
- Problem: Image data is non-sequential
- Modify AR, predict next token → predict next scale



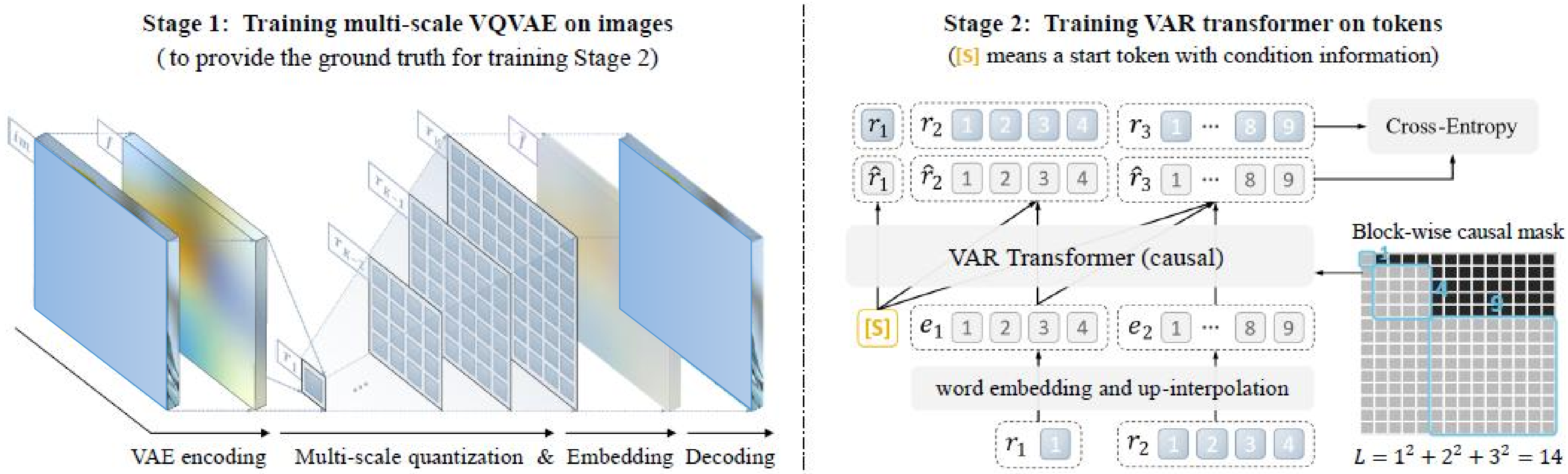
Three Different Autoregressive Generative Models

Background: Visual Autoregressive Modeling (VAR)

- Problem: Image data is non-sequential
- Modify AR, predict next token → predict next scale



Background: Visual Autoregressive Modeling (VAR)



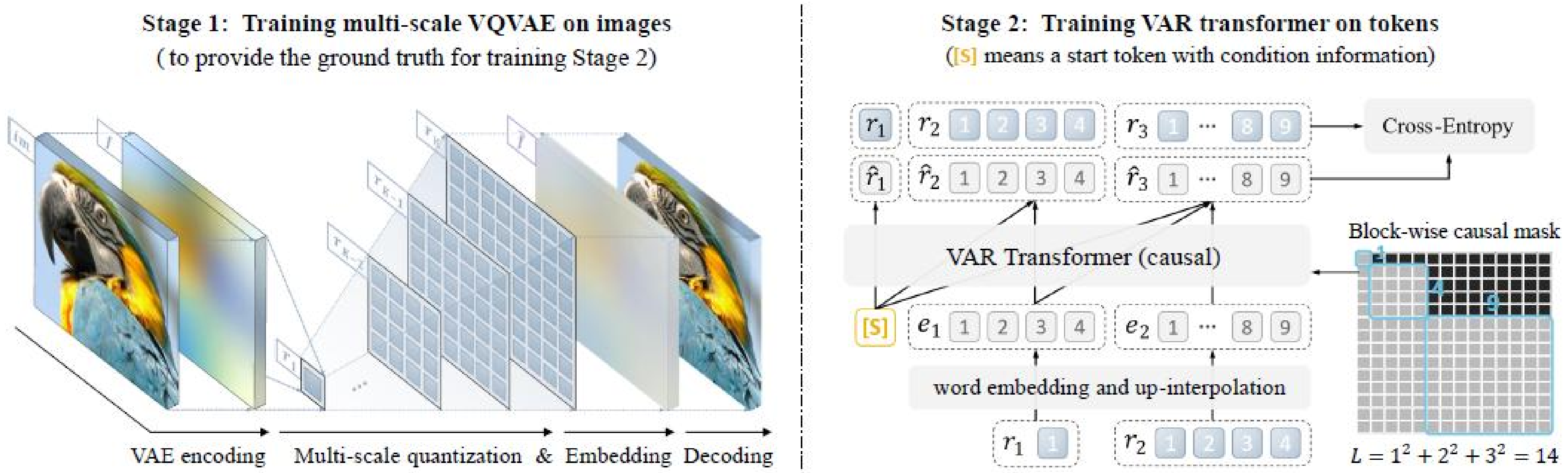
- Complexity of AR:

$$\sum_{i=1}^{n^2} i^2 = \frac{1}{6}n^2(n^2+1)(2n^2+1) \sim \mathcal{O}(n^6).$$

- Complexity of VAR:

$$\sum_{i=1}^{k} n_i^2 = \sum_{i=1}^{k} a^{2\cdot(k-1)} = \frac{a^{2k}-1}{a^2-1}. \qquad \sum_{k=1}^{\log_a(n)+1} \left(\frac{a^{2k}-1}{a^2-1}\right)^2 \sim \mathcal{O}(n^4).$$

Background: Visual Autoregressive Modeling (VAR)



- Complexity of AR:

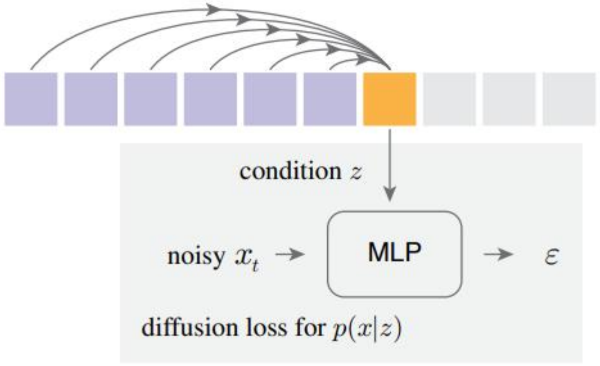
$$\sum_{i=1}^{n^2} i^2 = \frac{1}{6} n^2 (n^2+1)(2n^2+1) \sim \mathcal{O}(n^6).$$

- Complexity of VAR:

$$\sum_{i=1}^{k} n_i^2 = \sum_{i=1}^{k} a^{2\cdot(k-1)} = \frac{a^{2k}-1}{a^2-1}. \qquad \sum_{k=1}^{\log_a(n)+1} \left(\frac{a^{2k}-1}{a^2-1}\right)^2 \sim \mathcal{O}(n^4).$$

Background: Masked Autoregressive Model (MAR)

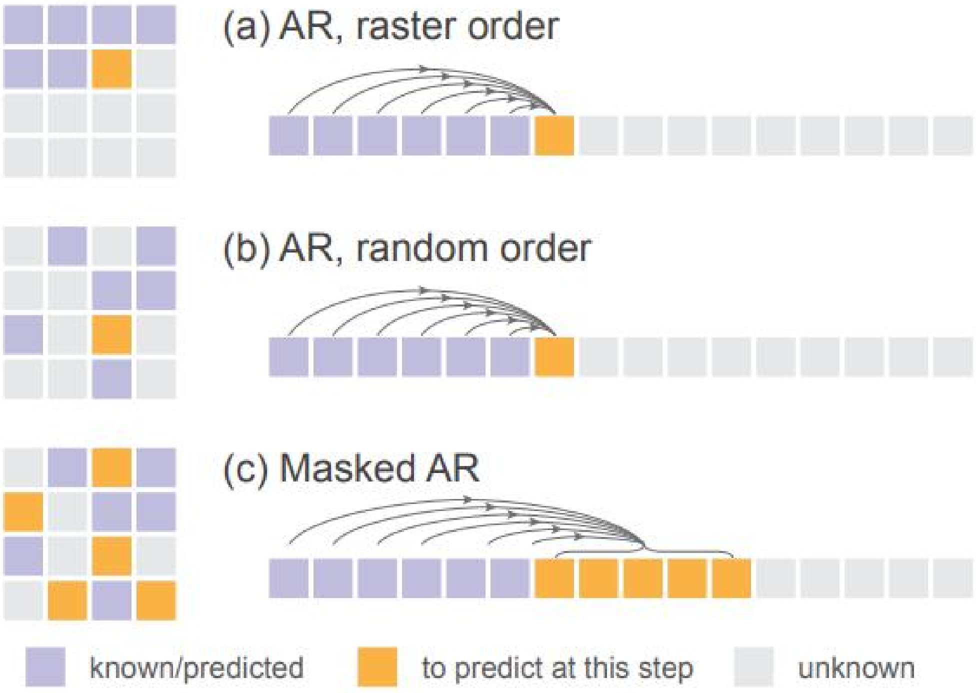
Modularization



- Modularize diffusion models as atomic building blocks
- Use diffusion loss to predict tokens

$$p(x^1, ..., x^n) = p(X^1, ..., X^K) = \prod_k^K p(X^k \mid X^1, ..., X^{k-1}).$$

$$X^k = \{x^i, x^{i+1} ..., x^j\}$$



- Predict multiple tokens simultaneously
- Random order masks
- Bidirectional attention

Method: Fractal Generative Models

- **Motivation**
  - To construct more advanced generative models from existing modules
- **Intuition**
  - Fractal structures exist in biological neural networks
  - Fractals are complex patterns that emerge from simple, recursive rules
- **Autoregressive Model as Fractal Generator**
  - Can handle k^n tokens with n layers, sequence length k manageable
  - Reduces computational cost; captures intrinsic hierarchical structure
  - Compatible with all divide-and-conquer-able data

Method: FGM VS Previous Methods

- AR
  - Linear, causal
  - Next token prediction

$$p(x_1, x_2, \ldots, x_T) = \prod_{t=1}^{T} p(x_t \mid x_1, x_2, \ldots, x_{t-1}).$$

- VAR
  - Divide feature map to scales
  - Next scale prediction

$$p(r_1, r_2, \ldots, r_K) = \prod_{k=1}^{K} p(r_k \mid r_1, r_2, \ldots, r_{k-1}),$$

- MAR
  - Modularized, masked
  - Next set-of-tokens prediction

$$p(x^1, \ldots, x^n) = p(X^1, \ldots, X^K) = \prod_{k}^{K} p(X^k \mid X^1, \ldots, X^{k-1}).$$

$$p(\{x^i, x^{i+1} \ldots, x^j\} \mid x^1, \ldots, x^{i-1})$$

- FGM
  - Modularize entire models
  - Next model prediction

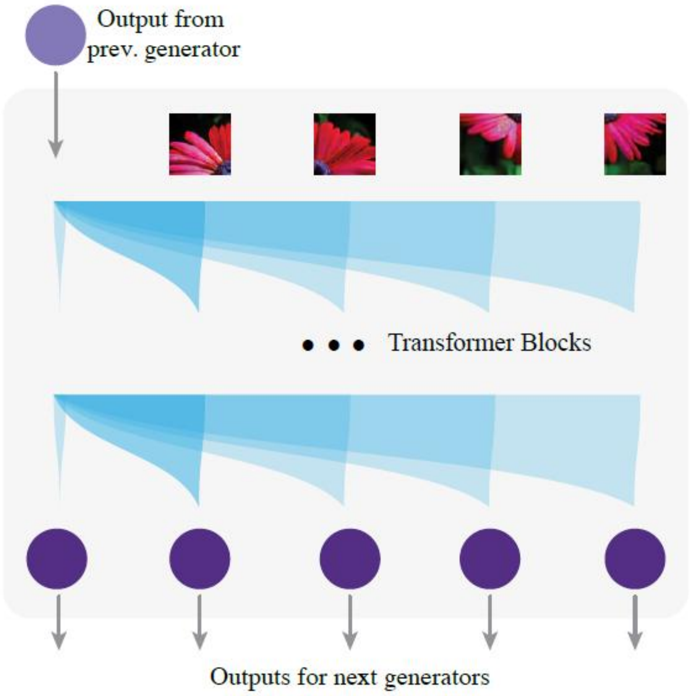
$$N = k^n, \quad n = \log_k(N)$$

$$p(x_1, \ldots, x_{k^n}) = \prod_{i=1}^{k} p(x_{(i-1)\cdot k^{n-1}+1}, \ldots, x_{i\cdot k^{n-1}} | x_1, \ldots, x_{(i-1)\cdot k^{n-1}}).$$

Method: An Image Generation Instantiation

- **Task**: pixel-by-pixel image generation
  - Challenge: high dimensionality and complexity of raw image data
  - Importance: element-by-element generation with non-sequential data

- **Architecture**:
  - Divide last layer outputs to patches
  - Embed patches to a sequence
  - Feed the sequence to transformer blocks
  - Forward the embeddings to the next layer
  - 1^{st} layer: 16×16



Method: Computational Cost Reduction

		image resolution	
		64×64×3	256×256×3
seq. len.	g_1	256	256
	g_2	16	16
	g_3	3	16
	g_4	-	3
#layers	g_1	32	32
	g_2	8	8
	g_3	3	4
	g_4	-	1
hidden dim	g_1	1024	1024
	g_2	512	512
	g_3	128	256
	g_4	-	64
#params (M)	g_1	403	403
	g_2	25	25
	g_3	0.6	3
	g_4	-	0.1
#GFLOPs	g_1	215	215
	g_2	208	208
	g_3	15	419
	g_4	-	22

- **256×256 to 64×64** (number of GFLOPs)**:**
  - 16 times big, only 2 times slow
- **Compared to VAR** (256×256, last layer)**:**
  - VAR: full attention across the image

$$(256 \times 256)^2 = 4{,}294{,}967{,}296$$

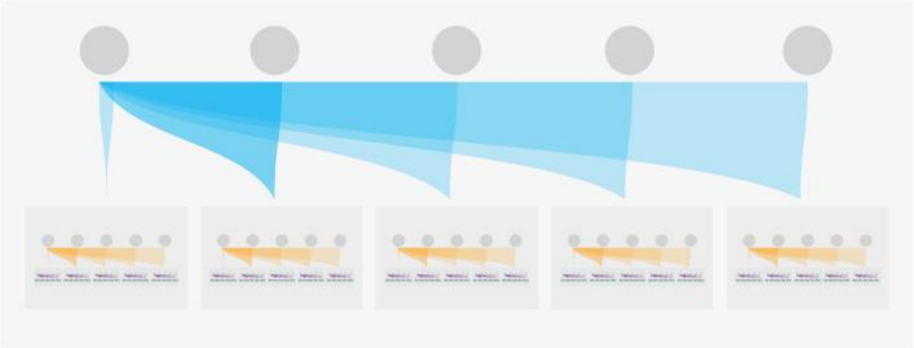
  - FGM: only in 4×4 patches

$$(64 \times 64) \times (4 \times 4)^2 = 1{,}048{,}576$$

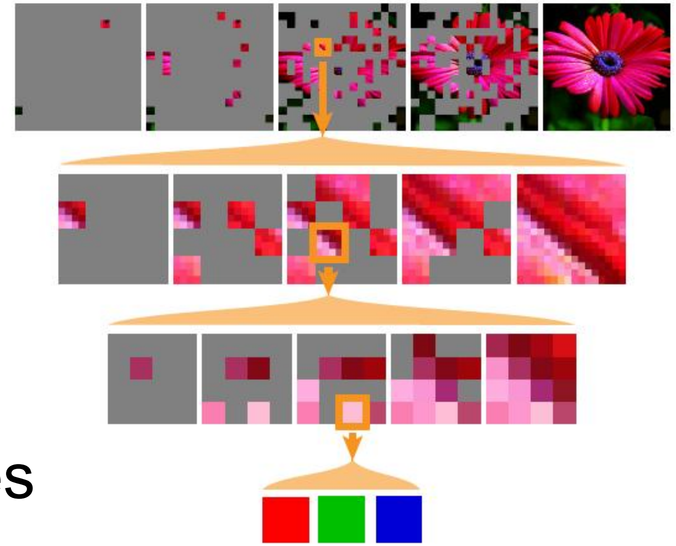
    - 4096 times fast!

Method: Implementation

- **Training**
  - End-to-end training on raw image pixels
  - Go through fractal architecture breadth-first
  - Each model produces a set of outputs
  - Last layer predicts RGB channels
- **Generation**
  - Pixel-by-pixel
  - Go through fractal architecture depth-first
  - Generator captures interdependence between patches



(b) fractal from the generator



Method: Implementation

Experiments: Likelihood Estimation

	Seq Len				
	g_1	g_2	g_3	#GFLOPs	NLL↓
AR, full-length	12288	-	-	29845	N/A
MAR, full-length	12288	-	-	29845	N/A
AR, 2-level	4096	3	-	5516	3.34
MAR, 2-level	4096	3	-	5516	3.36
FractalAR (3-level)	256	16	3	438	3.14
FractalMAR (3-level)	256	16	3	438	3.15

	type	NLL↓
iDDPM (Nichol & Dhariwal, 2021)	diffusion	3.53
VDM (Kingma et al., 2021)	diffusion	3.40
FM (Lipman et al., 2022)	diffusion†	3.31
NFDM (Bartosh et al., 2024)	diffusion	3.20
PixelRNN (van den Oord et al., 2016b)	AR	3.63
PixelCNN (van den Oord et al., 2016a)	AR	3.57
Sparse Transformer (Child et al., 2019)	AR	3.44
Routing Transformer (Roy et al., 2021)	AR	3.43
Combiner AR (Ren et al., 2021)	AR	3.42
Perceiver AR (Hawthorne et al., 2022)	AR	3.40
MegaByte (Yu et al., 2023)	AR	3.40
FractalAR	fractal	**3.14**
FractalMAR	fractal	3.15

- More layers, less cost, better NLL
- Outperforms previous AR models by a margin

Experiments: Generation Quality

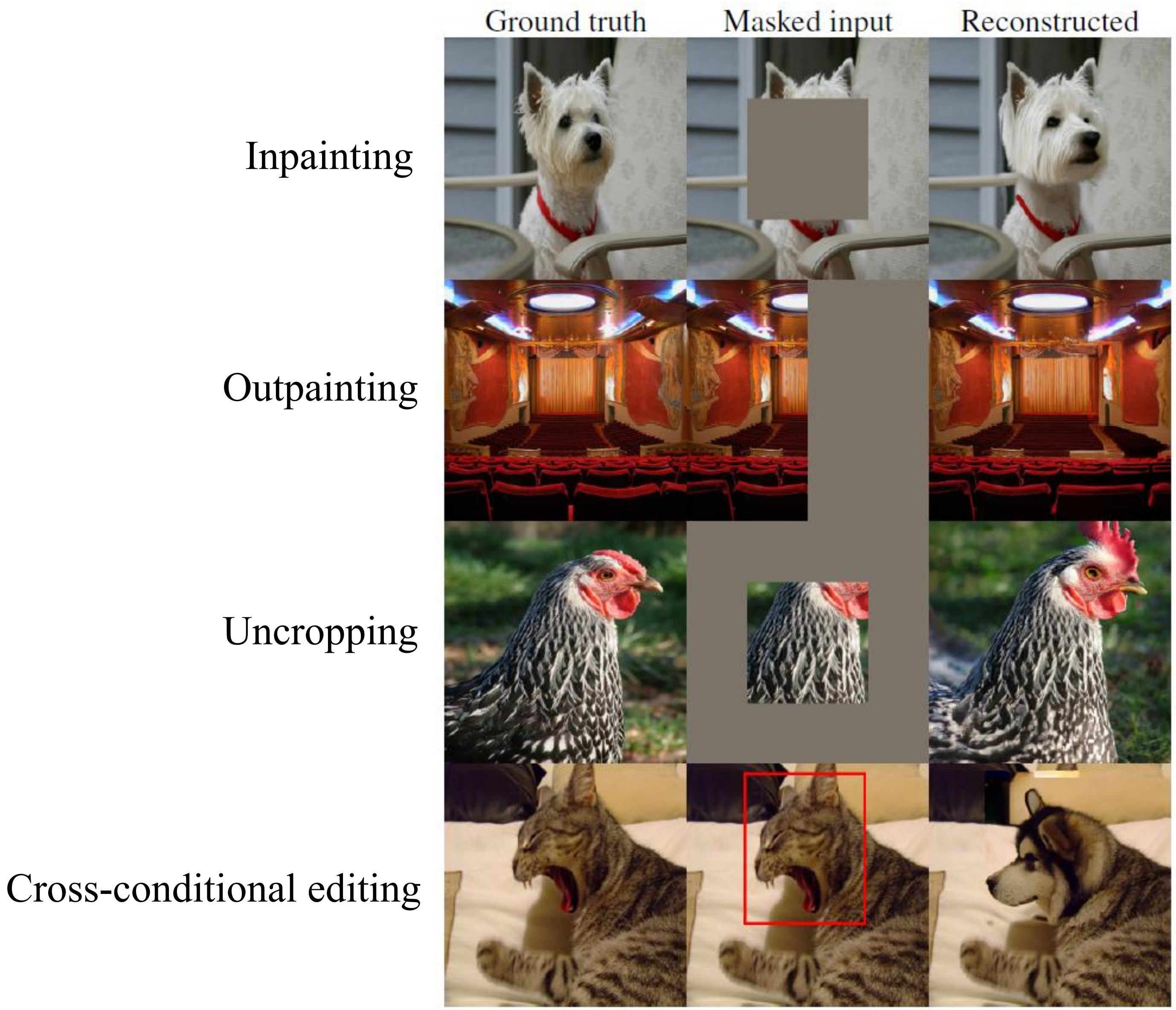
	type	#params	FID↓	IS↑	Pre.↑	Rec.↑
BigGAN-deep	GAN	160M	6.95	198.2	**0.87**	0.28
GigaGAN	GAN	569M	3.45	225.5	0.84	**0.61**
StyleGAN-XL	GAN	166M	2.30	265.1	0.78	0.53
ADM	diffusion	554M	4.59	186.7	0.82	0.52
Simple diffusion	diffusion	2B	3.54	205.3	-	-
VDM++	diffusion	2B	2.12	267.7	-	-
SiD2	diffusion	-	**1.38**	-	-	-
JetFormer	AR+flow	2.8B	6.64	-	0.69	0.56
FractalMAR-B	fractal	186M	11.80	274.3	0.78	0.29
FractalMAR-L	fractal	438M	7.30	334.9	0.79	0.44
FractalMAR-H	fractal	848M	6.15	**348.9**	0.81	0.46

- Strong IS and Precision
- Weak FID and Recall
- More params improves

- Already larger than GANs
- The only pixel-by-pixel model

Experiments: Generation Quality

Experiments: Conditional Pixel-by-pixel Prediction



Conclusion and discussion

- Proposes a new type of model structure: fractal generative models
- Reduces computational costs and accelerates training significantly
- Effective on pixel-by-pixel image generation
- Simple and widely applicable
- Discussion / Limitations:
  - Just an accelerated version of VAR?
  - Computational optimization requires more proof
  - Limited innovation on architecture

Thanks for listening!

Presenter: Wen Si
2025.3.9