FVL2025第四期学习讲座

近年来自回归结构的生成式模型概述

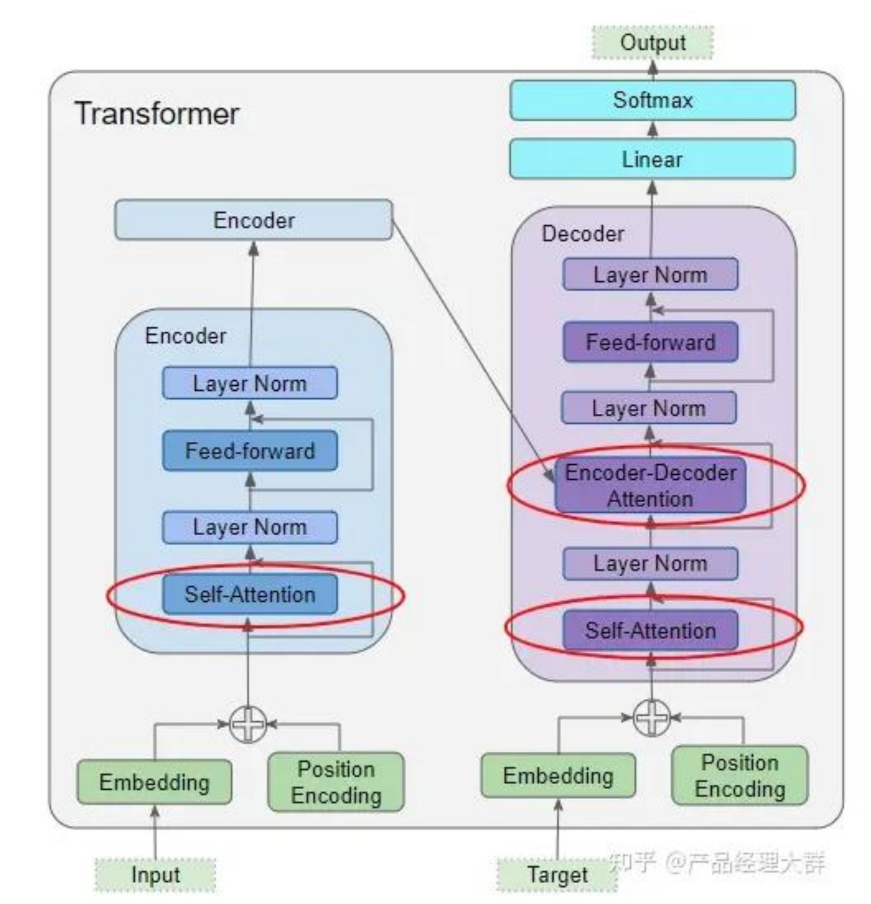
主讲人：吴嘉豪

1

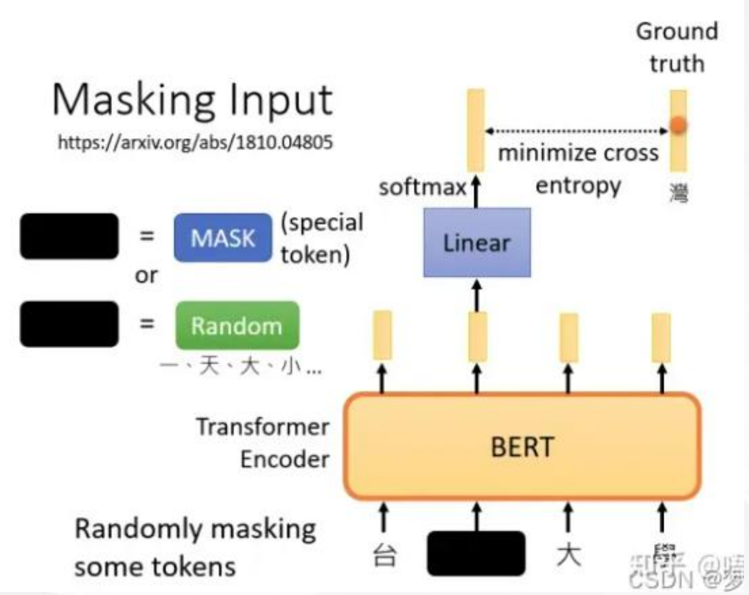
基础知识和背景

1 背景

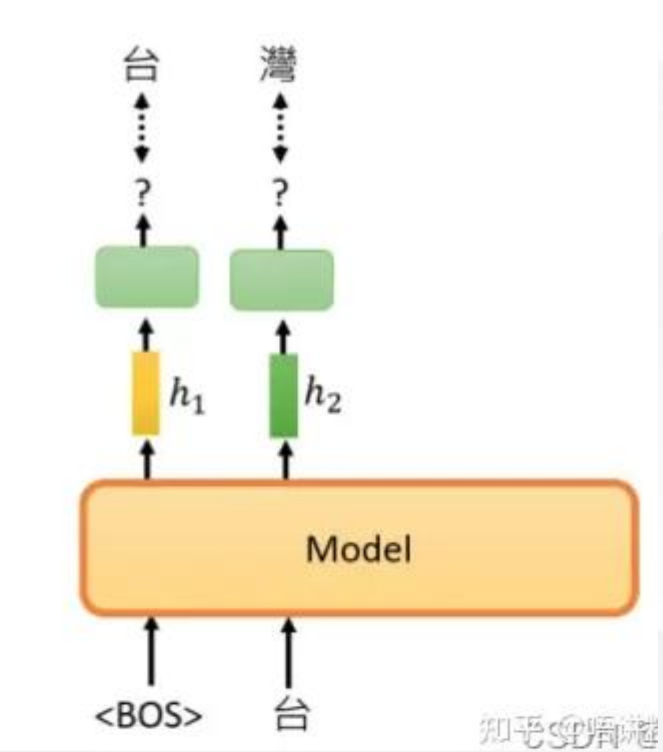
NLP: Transformer -> Bert / Bidirectional/ Encoder-only/MLM ; GPT / Casual / Decoder-only/AR



Attention Is All You Need
Google research 2017



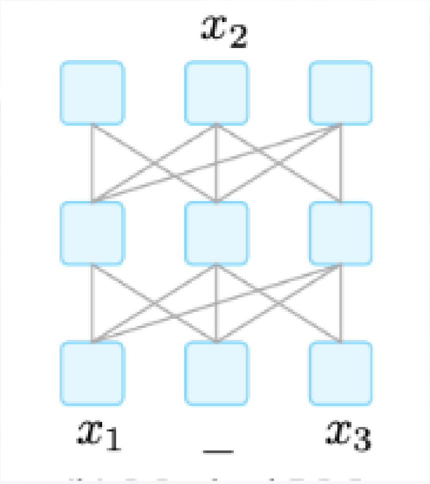
Bidirectional Encoder Representations from Transformers
Google research 2018

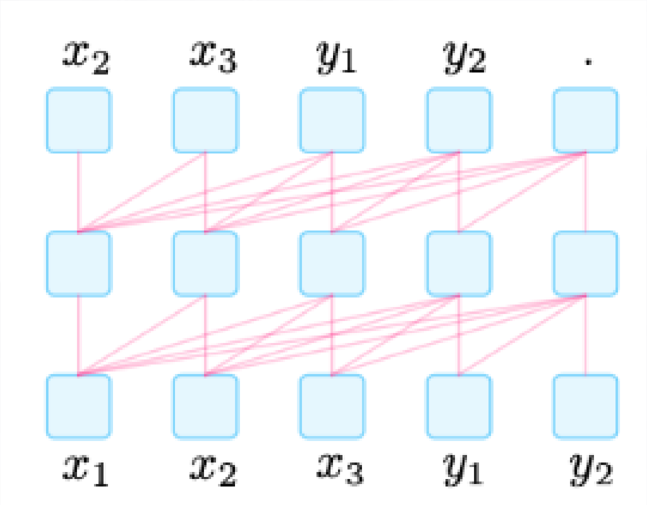


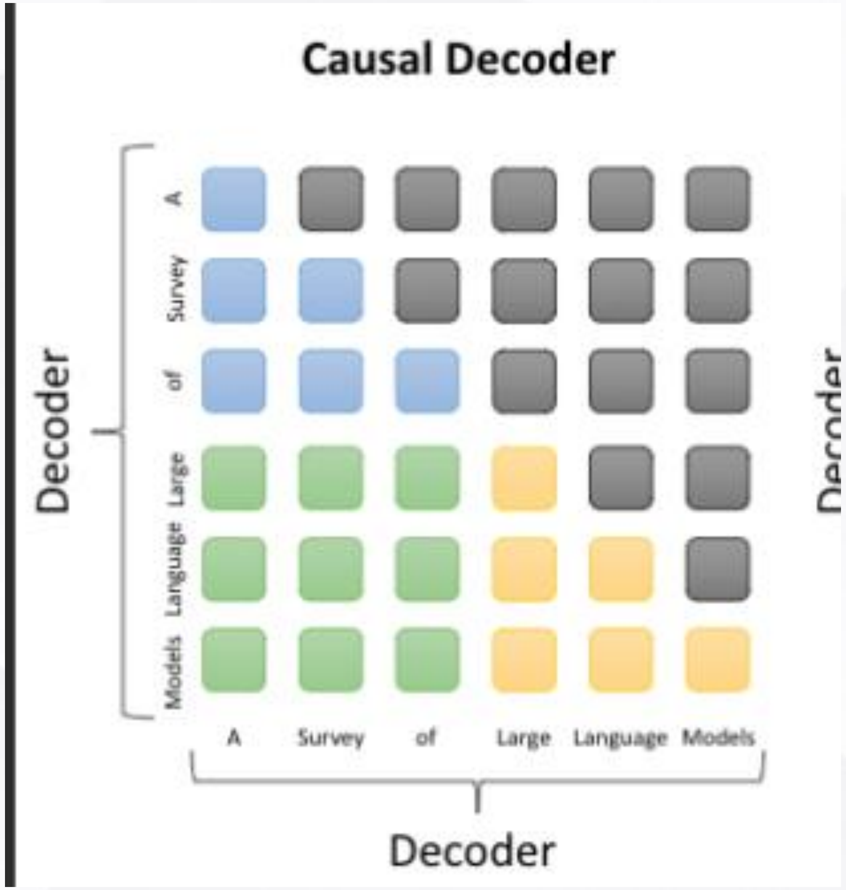
Generative Pre-trained Transformer
OpenAI 2018

NLP: Transformer -> Bert / Bidirectional/ Encoder-only ; GPT / Casual / Decoder-only

Train







test

Extract embedding

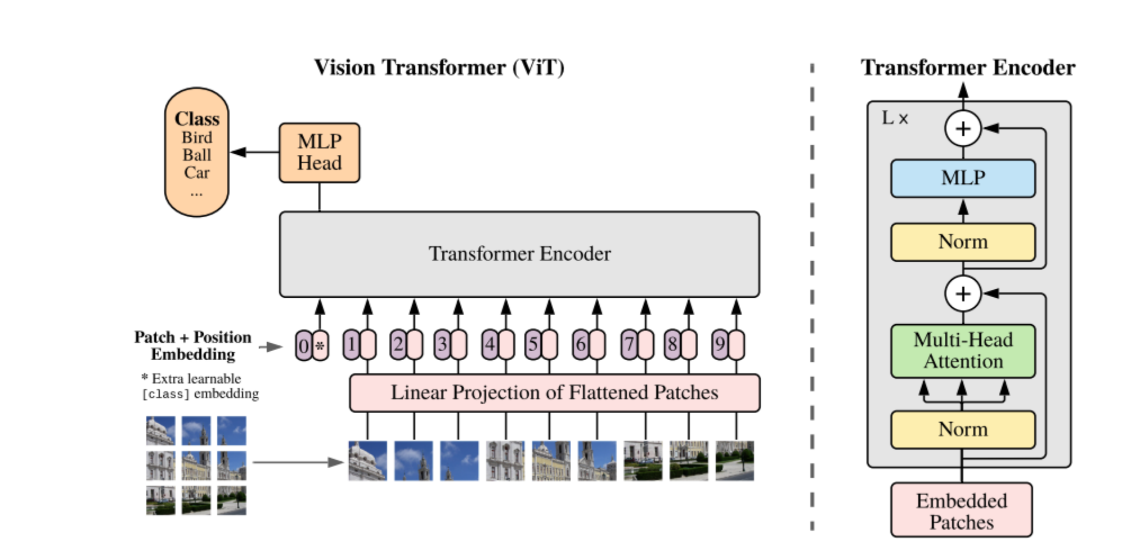
Auto regressive generation

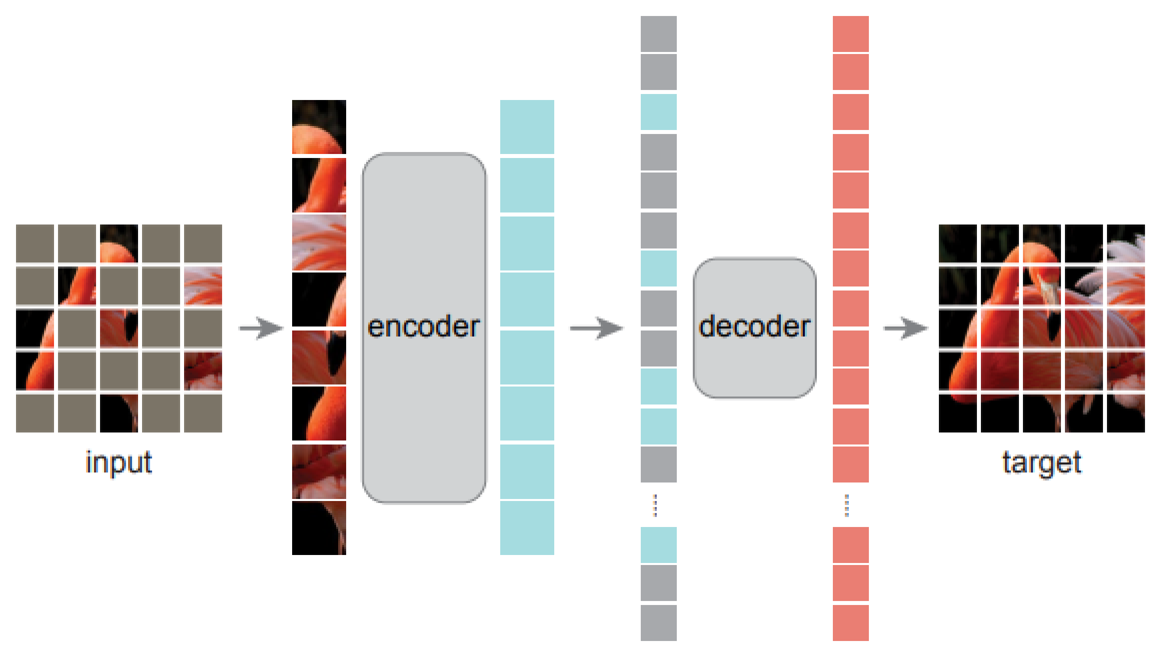
bert

gpt

CV: VIT; MAE

ViT + Bert





An image is worth 16x16 words: Transformers for image recognition at scale
Google research 2020

Masked Autoencoders Are Scalable Vision Learners
Facebook AI Research 2021

2 方法介绍

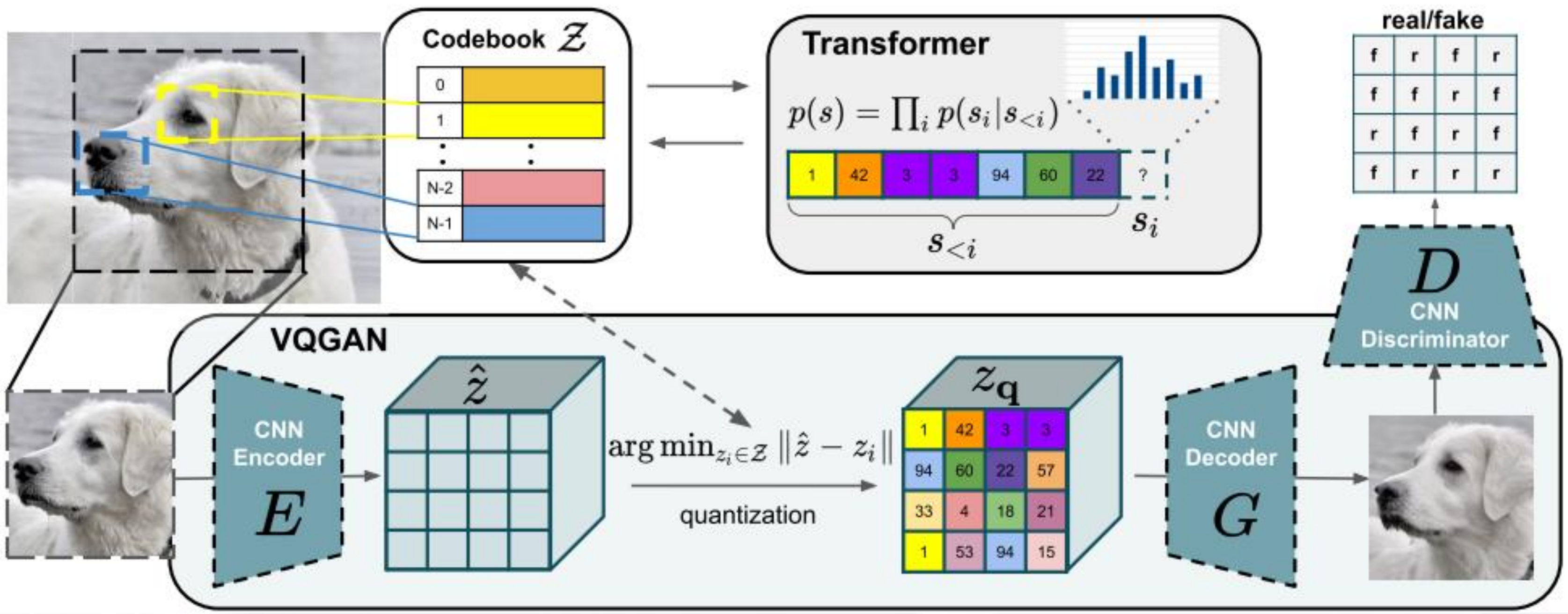
Auto Regressive（AR）

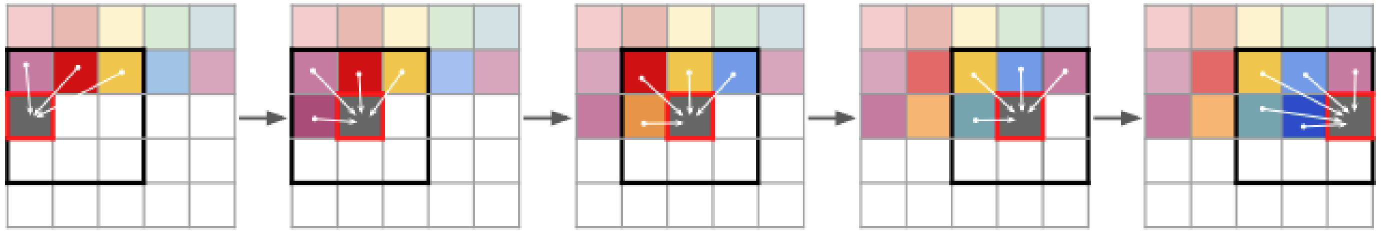
GPT范式

Taming Transformers for High-Resolution Image Synthesis

Heidelberg University CVPR2021

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





Google research

GPT范式
vector-quantized image modeling with improved vqgan （ ICLR2022 ）

Bert范式
Maskgit: Masked generative image transformer（ CVPR2022 ）

Bert 范式
MAGVIT: Masked Generative Video Transformer（ CVPR 2023 ）

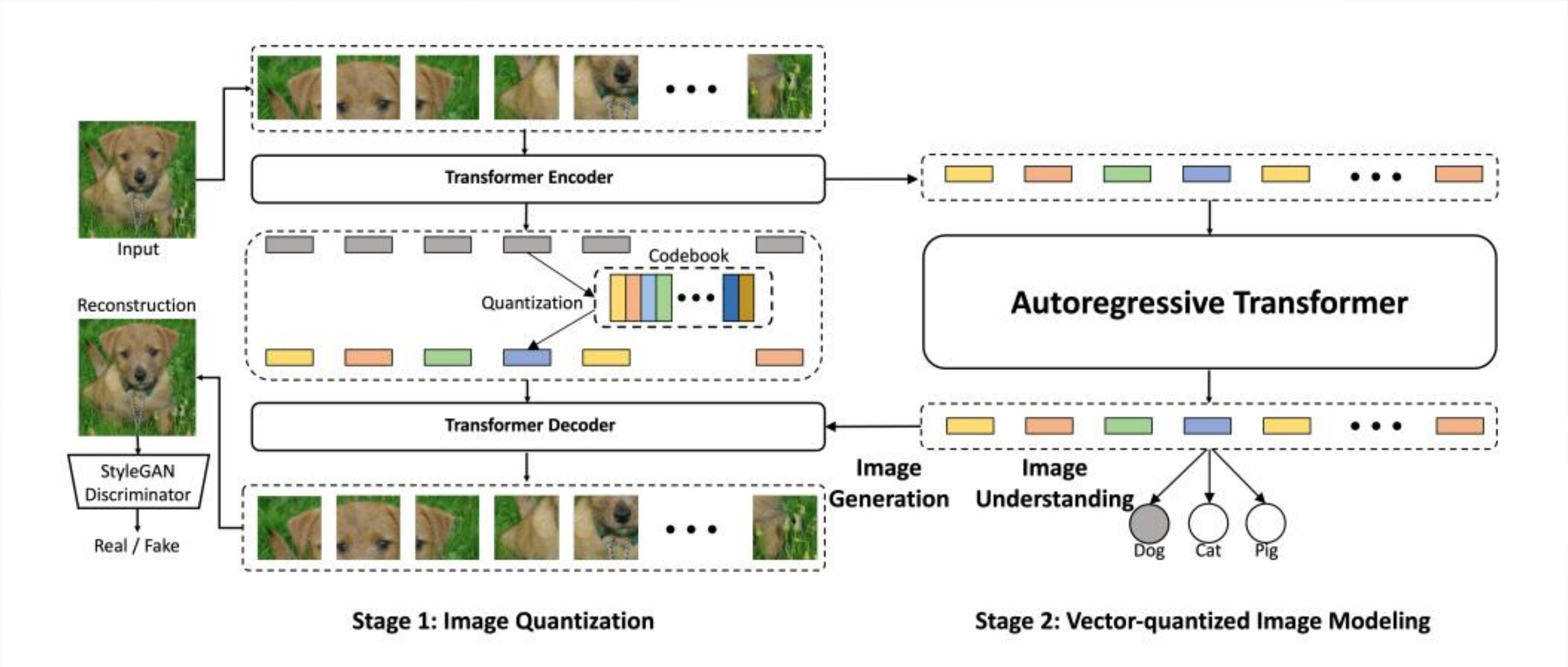
BERT范式 / GPT范式
language model beats diffusion — tokenizer is key to visual generation（ICLR2024）

1 正文

GPT范式
vector-quantized image modeling with improved vqgan
Google research ICLR2022

1. ViT
2. Low dimension look up
3. L2 norm

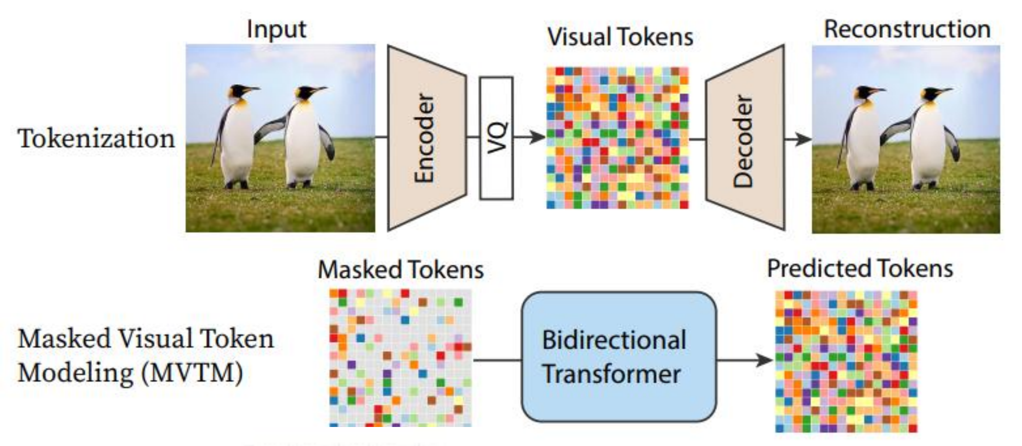
$$\|\ell_2(z_e(x)) - \ell_2(e_j)\|_2^2$$



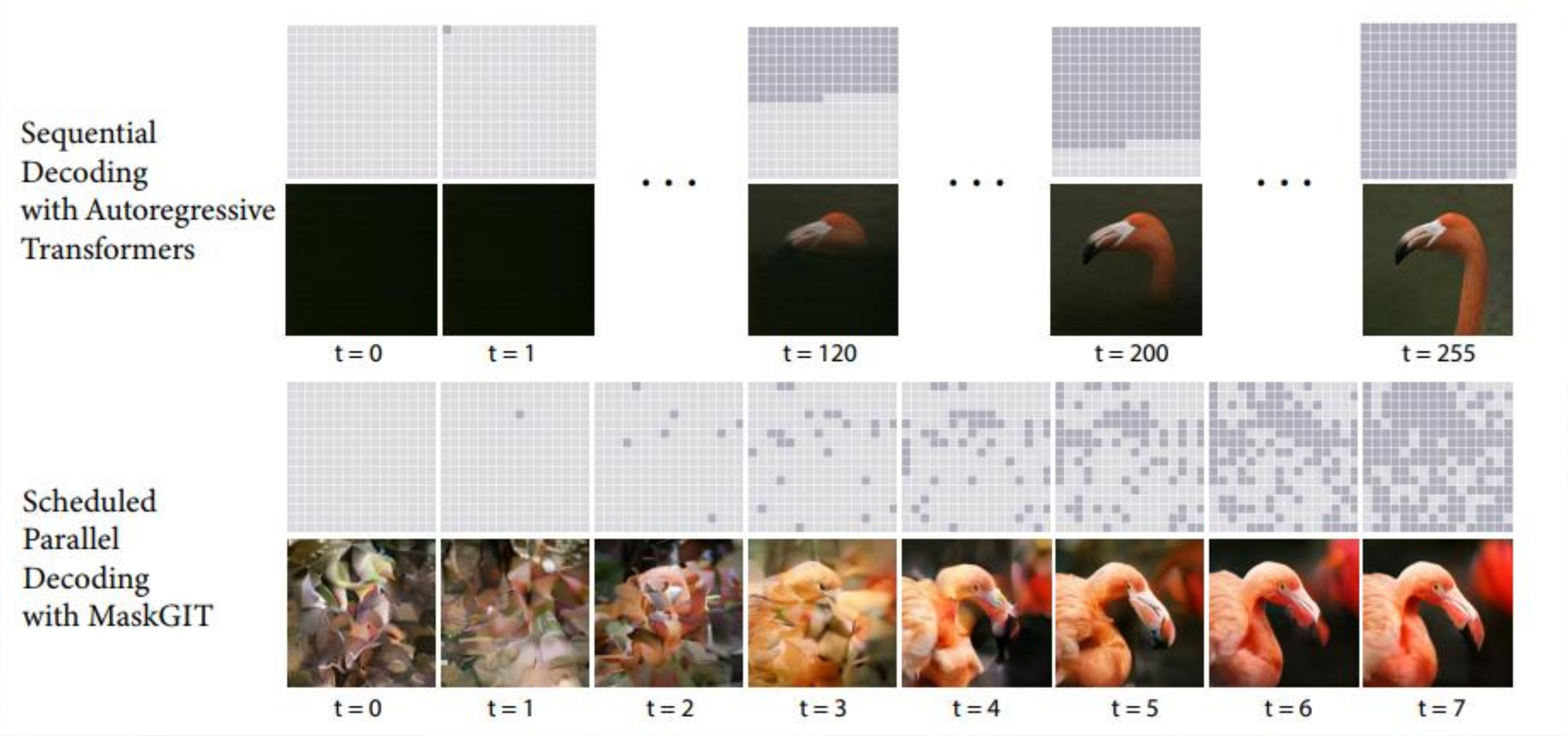
1 正文

Mask Image Modeling（MIM）

Bert范式
Maskgit: Masked generative image transformer
Google research CVPR2022



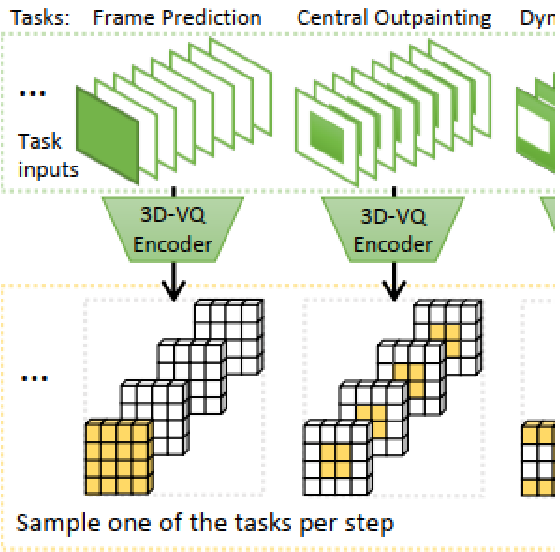
$$\mathcal{L}_{\text{mask}} = - \mathop{\mathbb{E}}_{\mathbf{Y} \in \mathcal{D}} \Big[\sum_{\forall i \in [1,N], m_i = 1} \log p(y_i | Y_{\overline{\mathbf{M}}}) \Big],$$

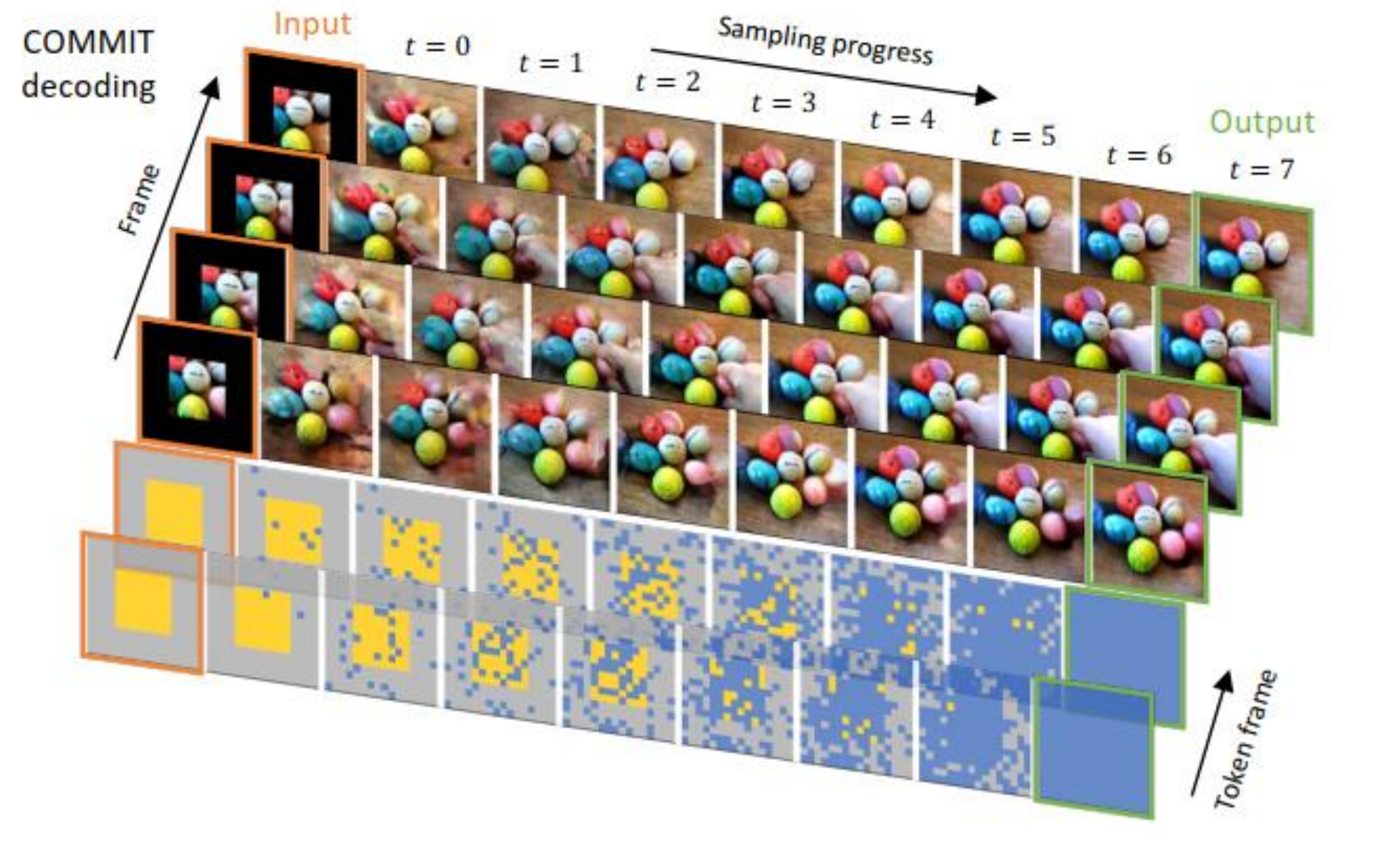


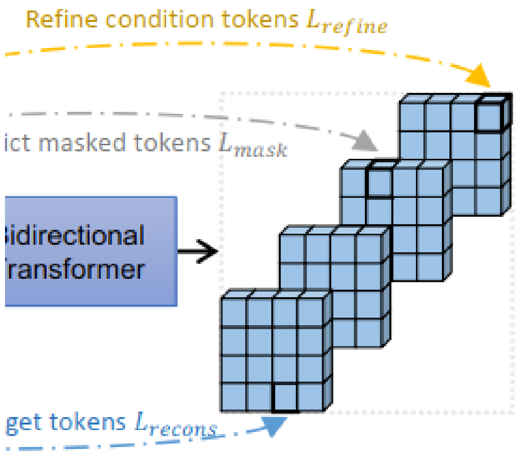
Bert 范式

MAGVIT: Masked Generative Video Transformer

Google research CVPR 2023







1 正文

Lookup-Free Quantization（LFQ）

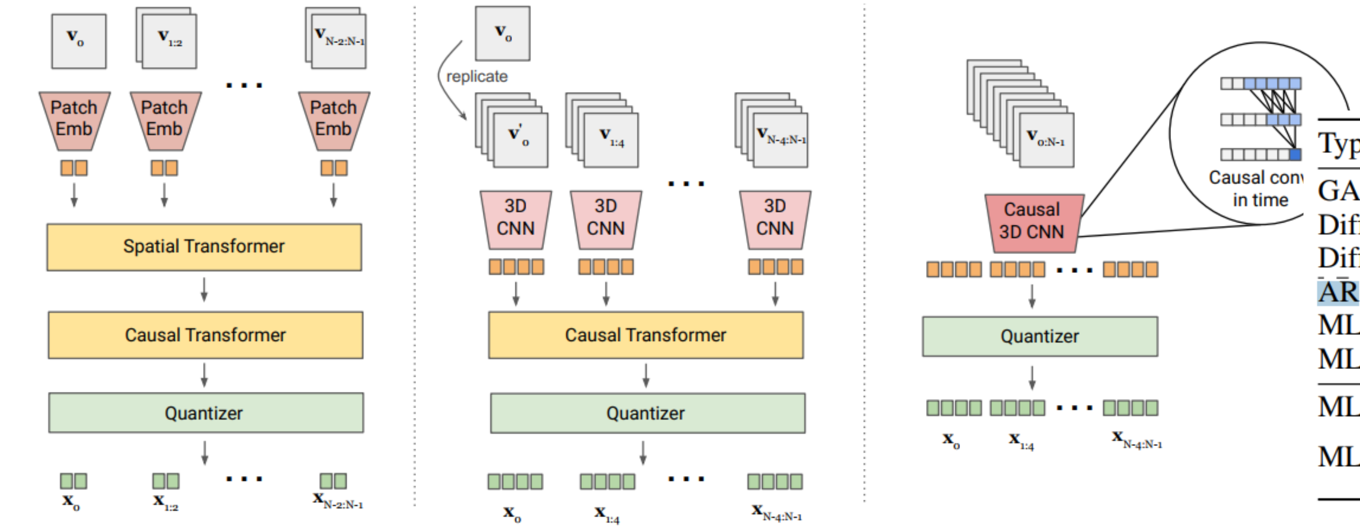
BERT范式 / GPT范式

language model beats diffusion — tokenizer is key to visual generation

Google research ICLR2024

$$q(z_i) = \mathrm{sign}(z_i) = -\mathbb{1}\{z_i \le 0\} + \mathbb{1}\{z_i > 0\}.$$

$$\mathcal{L}_{entropy} = \mathbb{E}[H(q(\mathbf{z}))] - H[\mathbb{E}(q(\mathbf{z}))].$$



(a) C-ViViT (b) C-ViViT + MAGVIT (c) Causal 3D CNN

Figure 2: **Causal tokenizer architecture comparison**. The decoders, which are omitted from the figure, employ an architecture that is symmetric to the encoder. See detailed architecture diagram in the Appendix.

Type	Method	K600 FVD↓	UCF FVD↓	#Params	#Steps
GAN	TrIVD-GAN-FP (Luc et al., 2020)	25.7±0.7			1
Diffusion	Video Diffusion (Ho et al., 2022c)	16.2±0.3		1.1B	256
Diffusion	RIN (Jabri et al., 2023)	10.8		411M	1000
AR-LM + VQ	TATS (Ge et al., 2022)		332±18	321M	1024
MLM + VQ	Phenaki (Villegas et al., 2022)	36.4±0.2		227M	48
MLM + VQ	MAGVIT (Yu et al., 2023a)	9.9±0.3	76±2	306M	12
MLM + LFQ	non-causal baseline	11.6±0.6		307M	12
MLM + LFQ	*MAGVIT-v2 (this paper)*	5.2±0.2		307M	12
		4.3±0.1	**58**±3		24

Tokenizer	FVD↓	#Params	#Steps
MAGVIT (Yu et al., 2023a)	265	306M	1024
MAGVIT-v2 (this paper)	**109**	840M	1280

1 正文

何凯明 DeepMind&MIT

MAR 范式（MAE+AR）
Autoregressive Image Generation without Vector Quantization（ NIPS 2024 ）

MAR / AR范式
Fractal Generative Models

MAR 范式（MAE+

Autoregressive Ima

DeepMind&MIT NI

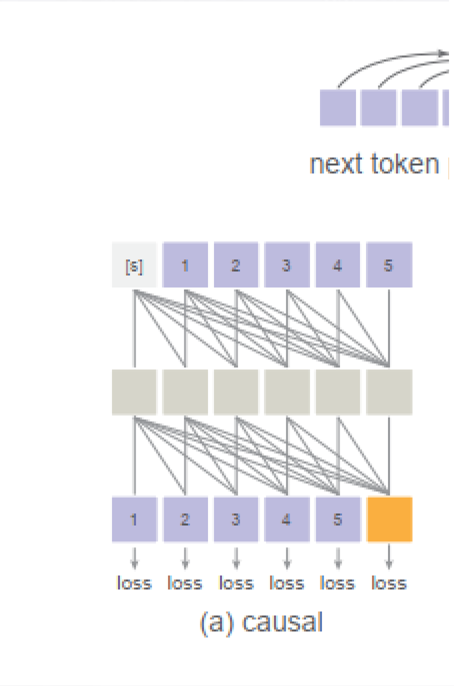
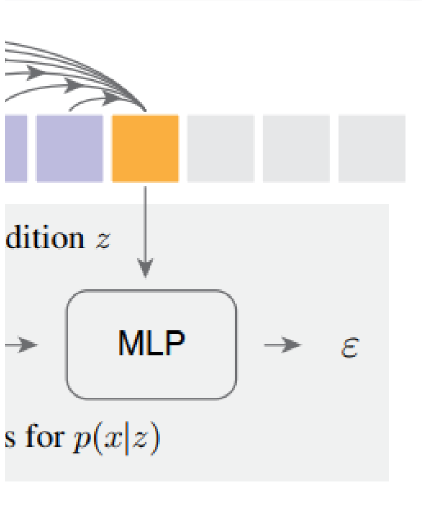


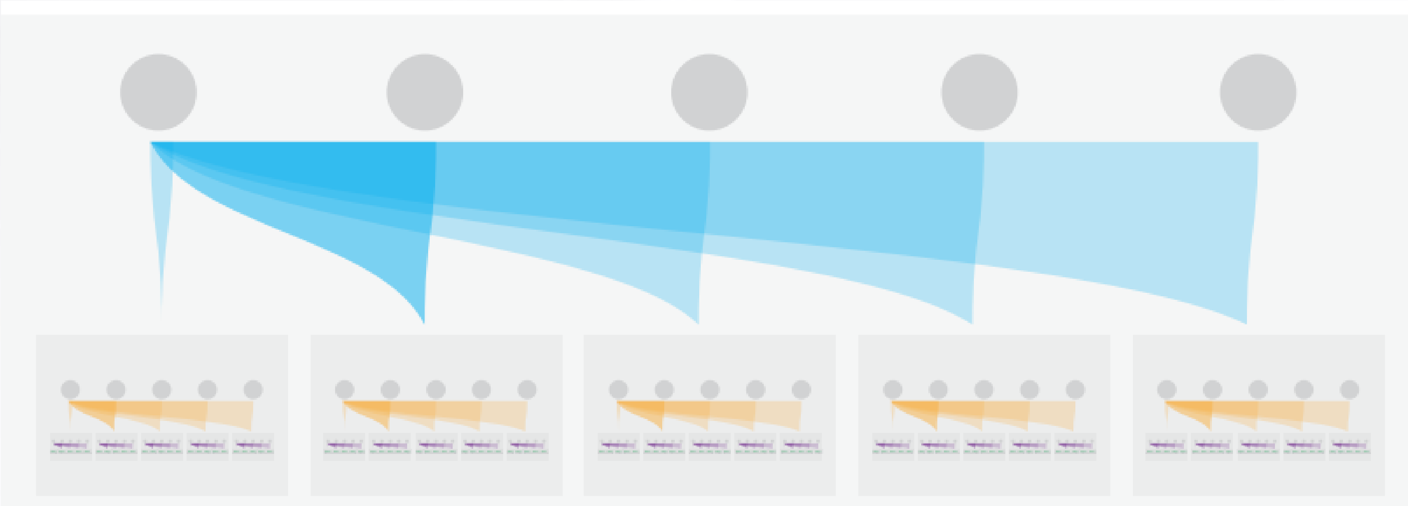
Table 4: **System-level comparison** on ImageNet 256×256 conditional generation. Diffusion Loss enables Masked Autoregression to achieve leading results in comparison with previous systems.
†: LDM operates on continuous-valued tokens, though this result uses a quantized tokenizer.

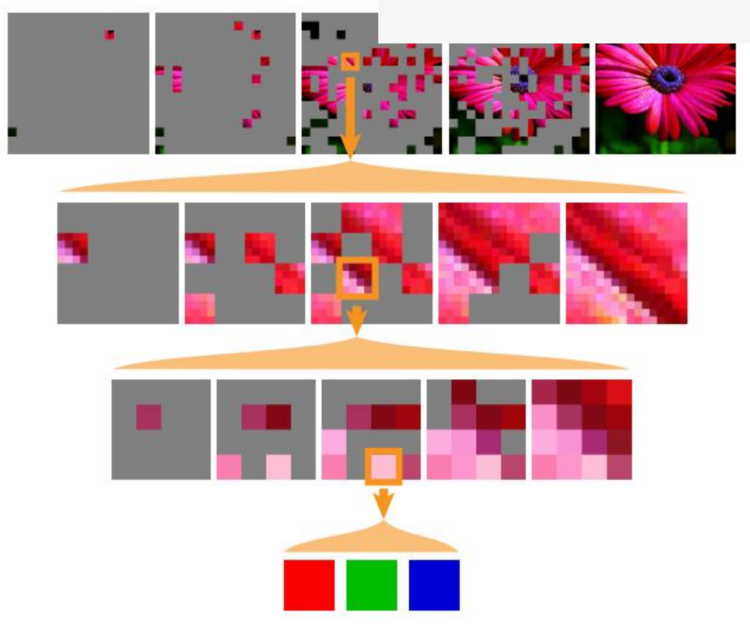
		w/o CFG				w/ CFG			
	#params	FID↓	IS↑	Pre.↑	Rec.↑	FID↓	IS↑	Pre.↑	Rec.↑
pixel-based									
ADM [10]	554M	10.94	101.0	0.69	0.63	4.59	186.7	0.82	0.52
VDM++ [26]	2B	2.40	225.3	-	-	2.12	267.7	-	-
vector-quantized tokens									
Autoreg. w/ VQGAN [13]	1.4B	15.78	78.3	-	-	-	-	-	-
MaskGIT [4]	227M	6.18	182.1	0.80	0.51	-	-	-	-
MAGE [29]	230M	6.93	195.8	-	-	-	-	-	-
MAGVIT-v2 [55]	307M	3.65	200.5	-	-	1.78	**319.4**	-	-
continuous-valued tokens									
LDM-4† [42]	400M	10.56	103.5	0.71	0.62	3.60	247.7	0.87	0.48
U-ViT-H/2-G [2]	501M	-	-	-	-	2.29	263.9	0.82	0.57
DiT-XL/2 [37]	675M	9.62	121.5	0.67	0.67	2.27	278.2	0.83	0.57
DiffiT [19]	-	-	-	-	-	1.73	276.5	0.80	0.62
MDTv2-XL/2 [14]	676M	5.06	155.6	0.72	0.66	1.58	314.7	0.79	0.65
GIVT [48]	304M	5.67	-	0.75	0.59	3.35	-	0.84	0.53
MAR-B, Diff Loss	208M	3.48	192.4	0.78	0.58	2.31	281.7	0.82	0.57
MAR-L, Diff Loss	479M	2.60	221.4	0.79	0.60	1.78	296.0	0.81	0.60
MAR-H, Diff Loss	943M	**2.35**	**227.8**	0.79	0.62	**1.55**	303.7	0.81	0.62



MAR / AR范式
Fractal Generative Models
DeepMind&MIT 2025

		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





	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

MAR / AR范式

Fractal Generative Models

DeepMind&MIT 2025

第一层GPT的输入是条件＋image的16x16patch序列（L_C+HW, C），输出是HW个token（去掉了最后一个token对应的输出，所以是(HW,C)）

第二层GPT的输入是上一层GPT的输出+patch的4x4sub patch序列（1+16，c），输出同理是16个token

第三层GPT的输入是上一层GPT的输出+sub patch的每个pixel（1+16，c），输出同理是16个token

第四层GPT的输入是上一层GPT的输出+每个pixel的RGB值（1+3，c），输出是3个token

最后的3个token各自过全连接变成256维的分类输出，与对应的GT算交叉熵损失

1 正文

字节 2024

Bert范式
An Image is Worth 32 Tokens for Reconstruction and Generation

GPT范式
Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation

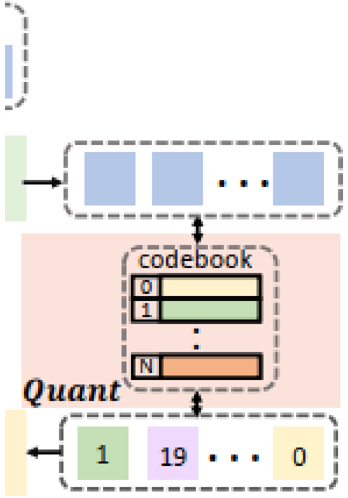
VAR范式
Visual **Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction（NIPS2024 best paper）**

bert范式

An Image is Worth 32 Tokens for Reconstruction and Generation

Table 1: **ImageNet-1K** 256 × 256 **generation results evaluated with ADM [16].** †: Trained on OpenImages [35] ‡: Trained on OpenImages, LAION-Aesthetics/-Humans [56]. P: generator's parameters. S: sampling steps. T: throughput as samples per seconds on A100 with float32 precision.

tokenizer	#tokens	codebook size	rFID↓	generator	gFID↓	P↓	S↓	T↑
diffusion-based generative models								
Taming-VQGAN† [55]	1024	16384	1.14	LDM-8 [55]	7.76	258M	200	-
VAE† [55]	4096×3	-	0.27	LDM-4 [55]	3.60	400M	250	0.4
VAE [57]‡	1024×4	-	0.62	UViT-L/2 [4]	3.40	287M	50	1.1
				UViT-H/2 [4]	2.29	501M	50	0.6
				DiT-XL/2 [49]	2.27	675M	250	0.6
transformer-based generative models								
Taming-VQGAN [19]	256	1024	7.94	Taming-Transformer [19]	15.78	1.4B	256	7.5
RQ-VAE [36]	256	16384	3.20	RQ-Transformer [36]	8.71 7.55	1.4B 3.8B	64	16.1 9.7
MaskGIT-VQGAN [9]	256	1024	2.28	MaskGIT-ViT [9]	6.18	**177M**	**8**	50.5
ViT-VQGAN [65]	1024	8192	1.28	VIM-Large [65]	4.17	1.7B	1024	0.3
TiTok-L-32	32	4096	2.21	MaskGIT-ViT [9]	2.77	**177M**	**8**	**101.6**
TiTok-B-64	64	4096	1.70	MaskGIT-ViT [9]	2.48	**177M**	**8**	89.8
TiTok-S-128	128	4096	1.71	MaskGIT-UViT-L [9, 4]	2.50 **1.97**	287M	**8** 64	53.3 7.8



GPT范式

Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation

Type	Model	#Para.	FID↓	IS↑	Precision↑	Recall↑
GAN	BigGAN [Brock et al. 2018]	112M	6.95	224.5	0.89	0.38
	GigaGAN [Kang et al. 2023]	569M	3.45	225.5	0.84	0.61
	StyleGan-XL [Sauer et al. 2022]	166M	2.30	265.1	0.78	0.53
Diffusion	ADM [Dhariwal & Nichol 2021]	554M	10.94	101.0	0.69	0.63
	CDM [Ho et al. 2022b]	–	4.88	158.7	–	–
	LDM-4 [Rombach et al. 2022]	400M	3.60	247.7	–	–
	DiT-XL/2 [Peebles & Xie 2023]	675M	2.27	278.2	0.83	0.57
Mask.	MaskGIT [Chang et al. 2022]	227M	6.18	182.1	0.80	0.51
	MaskGIT-re [Chang et al. 2022]	227M	4.02	355.6	–	–
AR	VQGAN [Esser et al. 2021]	227M	18.65	80.4	0.78	0.26
	VQGAN [Esser et al. 2021]	1.4B	15.78	74.3	–	–
	VQGAN-re [Esser et al. 2021]	1.4B	5.20	280.3	–	–
	ViT-VQGAN [Yu et al. 2021]	1.7B	4.17	175.1	–	–
	ViT-VQGAN-re [Yu et al. 2021]	1.7B	3.04	227.4	–	–
	RQTran. [Lee et al. 2022]	3.8B	7.55	134.0	–	–
	RQTran.-re [Lee et al. 2022]	3.8B	3.80	323.7	–	–
AR	LlamaGen-B (cfg=2.00)	111M	5.46	193.61	0.83	0.45
	LlamaGen-L (cfg=2.00)	343M	3.07	256.06	0.83	0.52
	LlamaGen-XL (cfg=1.75)	775M	2.62	244.08	0.80	0.57
	LlamaGen-XXL (cfg=1.75)	1.4B	2.34	253.90	0.80	0.59
	LlamaGen-3B (cfg=1.65)	3.1B	2.18	263.33	0.81	0.58
	LlamaGen-3B (cfg=1.75)	3.1B	2.32	280.10	0.82	0.56
	LlamaGen-3B (cfg=2.00)	3.1B	2.81	311.59	0.84	0.54

VAR范式

Visual **Autoregressive M** ... **NIPS2024 best paper）**

Algorithm 2: Mu ... **cale VQVAE Encoding**

1 **Inputs:** multi-sca ... ;

2 **Hyperparameters** ... ps K, resolutions

$(h_k, w_k)_{k=1}^{K}$;

3 $\hat{f} = 0$;

4 **for** $k = 1, \cdots, K$

5 $r_k = \text{queue\_p}$ ... $(f, h_k, w_k))$;

6 $z_k = \text{lookup}($ ... $, r_k)$;

7 $z_k = \text{interpola}$ ... $)$;

8 $\hat{f} = \hat{f} + \phi_k(z$ ... $, h_K, w_K)$;

9 $\hat{im} = \mathcal{D}(\hat{f})$;

10 **Return:** reconstru ... kens R;

Type	Model	FID↓	IS↑	Pre↑	Rec↑	#Para	#Step	Time
GAN	BigGAN [13]	6.95	224.5	**0.89**	0.38	112M	1	–
GAN	GigaGAN [42]	3.45	225.5	0.84	**0.61**	569M	1	–
GAN	StyleGan-XL [74]	2.30	265.1	0.78	0.53	166M	1	0.3 [74]
Diff.	ADM [26]	10.94	101.0	0.69	0.63	554M	250	168 [74]
Diff.	CDM [36]	4.88	158.7	–	–	–	8100	–
Diff.	LDM-4-G [70]	3.60	247.7	–	–	400M	250	–
Diff.	DiT-L/2 [63]	5.02	167.2	0.75	0.57	458M	250	31
Diff.	DiT-XL/2 [63]	2.27	278.2	0.83	0.57	675M	250	45
Diff.	L-DiT-3B [3]	2.10	304.4	0.82	0.60	3.0B	250	>45
Diff.	L-DiT-7B [3]	2.28	316.2	0.83	0.58	7.0B	250	>45
Mask.	MaskGIT [17]	6.18	182.1	0.80	0.51	227M	8	0.5 [17]
Mask.	RCG (cond.) [51]	3.49	215.5	–	–	502M	20	1.9 [51]
AR	VQVAE-2† [68]	31.11	~45	0.36	0.57	13.5B	5120	–
AR	VQGAN† [30]	18.65	80.4	0.78	0.26	227M	256	19 [17]
AR	VQGAN [30]	15.78	74.3	–	–	1.4B	256	24
AR	VQGAN-re [30]	5.20	280.3	–	–	1.4B	256	24
AR	ViTVQ [92]	4.17	175.1	–	–	1.7B	1024	>24
AR	ViTVQ-re [92]	3.04	227.4	–	–	1.7B	1024	>24
AR	RQTran. [50]	7.55	134.0	–	–	3.8B	68	21
AR	RQTran.-re [50]	3.80	323.7	–	–	3.8B	68	21
VAR	VAR-*d*16	3.30	274.4	0.84	0.51	310M	10	0.4
VAR	VAR-*d*20	2.57	302.6	0.83	0.56	600M	10	0.5
VAR	VAR-*d*24	2.09	312.9	0.82	0.59	1.0B	10	0.6
VAR	VAR-*d*30	1.92	323.1	0.82	0.59	2.0B	10	1
VAR	VAR-*d*30-re	**1.73**	**350.2**	0.82	0.60	2.0B	10	1
	(validation data)	1.78	236.9	0.75	0.67			

腾讯 2024

GPT范式
OPEN-MAGVIT2: AN OPEN-SOURCE PROJECT TOWARD DEMOCRATIZING AUTO-REGRESSIVE VISUAL GENERATION

GPT范式
Taming Scalable Visual Tokenizer for Autoregressive Image Generation

1 正文

GPT范式

OPEN-MAGVIT2: A UAL GENERATION

Image

Reconstruction

Type	Model	#Para.	FID↓	IS↑	Precision↑	Recall↑
Diffusion	ADM (Dhariwal & Nichol, 2021)	554M	10.94	101.0	0.69	0.63
	CDM (Ho et al., 2022)	–	4.88	158.7	–	–
	LDM-4 (Rombach et al., 2022a)	400M	3.60	247.7	–	–
	DiT-XL/2 (Peebles & Xie, 2023)	675M	2.27	278.2	0.83	0.57
AR	VQGAN (Esser et al., 2021)	227M	18.65	80.4	0.78	0.26
	VQGAN (Esser et al., 2021)	1.4B	15.78	74.3	–	–
	VQGAN-re (Esser et al., 2021)	1.4B	5.20	280.3	–	–
	ViT-VQGAN (Yu et al., 2022)	1.7B	4.17	175.1	–	–
	ViT-VQGAN-re (Yu et al., 2022)	1.7B	3.04	227.4	–	–
	RQTran. (Lee et al., 2022)	3.8B	7.55	134.0	–	–
	RQTran.-re (Lee et al., 2022)	3.8B	3.80	323.7	–	–
VAR	VAR-d16 (Tian et al., 2024)	310M	3.30	274.4	0.84	0.51
	VAR-d20 (Tian et al., 2024)	600M	2.57	302.6	0.83	0.56
	VAR-d24 (Tian et al., 2024)	1.0B	2.09	312.9	0.82	0.59
	VAR-d30 (Tian et al., 2024)	2.0B	1.92	323.1	0.82	0.59
AR	LlamaGen-L* (Sun et al., 2024)	343M	3.07	256.06	0.83	0.52
	LlamaGen-XL* (Sun et al., 2024)	775M	2.62	244.08	0.80	0.57
	LlamaGen-XXL* (Sun et al., 2024)	1.4B	2.34	253.90	0.80	0.59
	LlamaGen-L (Sun et al., 2024)	343M	3.80	248.28	0.83	0.51
	LlamaGen-XL (Sun et al., 2024)	775M	3.39	227.08	0.81	0.54
	LlamaGen-XXL (Sun et al., 2024)	1.4B	3.09	253.61	0.83	0.53
	Open-MAGVIT2-B	343M	3.08	258.26	0.85	0.51
	Open-MAGVIT2-L	804M	2.51	271.70	0.84	0.54
	Open-MAGVIT2-XL	1.5B	2.33	271.77	0.84	0.54

$$q = \arg\min_{C_k \in C} ||z - C_k|| \in \mathbb{R}^D,$$
$$z_q = z + \text{sg}[q - z],$$

GPT范式

Taming Scalable Visual Tokenizer for Autoregressive Image Generation

Specifically, we first perform dot product between the given visual feature z and all code embeddings as logits and get probabilities (soft one-hot) by softmax function.

$$\text{logits} = [z^T C_1, z^T C_2, \cdots, z^T C_K]^T \in \mathbb{R}^K, \quad (3)$$
$$\text{Ind}_{\text{soft}} = \text{softmax}(\text{logits}), \quad (4)$$
$$\text{Ind}_{\text{hard}} = \text{One-Hot}(\text{argmax}(\text{Ind}_{\text{soft}})). \quad (5)$$

Then we copy the gradients of soft one-hot categorical distribution to hard one-hot index:

$$\text{Ind} = \text{Ind}_{\text{hard}} - \text{sg}[\text{Ind}_{\text{soft}}] + \text{Ind}_{\text{soft}}. \quad (6)$$

Given the index, the quantized feature is obtained by:

$$z_q = \text{Ind}^T C. \quad (7)$$

Method	Token Type	Tokens	Ratio	Train Resolution	Codebook Size	Codebook Dim	rFID↓	LPIPS↓	Codebook Usage↑
VQGAN [6]	2D	16 × 16	16	256 × 256	1,024	256	7.94	–	44%
VQGAN [6]	2D	16 × 16	16	256 × 256	16,384	256	4.98	0.2843	5.9%
VQGAN* [6]	2D	16 × 16	16	256 × 256	16,384	256	3.98	0.2873	5.3%
SD-VQGAN [20]	2D	16 × 16	16	256 × 256	16,384	8	5.15	–	–
MaskGIT [3]	2D	16 × 16	16	256 × 256	1,024	256	2.28	–	–
LlamaGen [24]	2D	16 × 16	16	256 × 256	16,384	256	9.21	–	0.29%
LlamaGen [24]	2D	16 × 16	16	256 × 256	16,384	8	2.19	0.2281	97%
VQGAN-LC [36]	2D	16 × 16	16	256 × 256	16,384	8	3.01	0.2358	99%
VQGAN-LC [36]	2D	16 × 16	16	256 × 256	100,000	8	2.62	0.2212	99%
MaskBit [30]	2D	16 × 16	16	256 × 256	16,384	0	1.61	–	–
Open-MAGVIT2 [16]	2D	16 × 16	16	256 × 256	16,384	0	1.58	0.2261	100%
Open-MAGVIT2 [16]	2D	16 × 16	16	256 × 256	262,144	0	1.17	0.2038	100%
IBQ (Ours)	2D	16 × 16	16	256 × 256	16,384	256	1.37	0.2235	96%
IBQ (Ours)	2D	16 × 16	16	256 × 256	262,144	256	**1.00**	**0.2030**	84%
Titok-L [33]	1D	32	–	256 × 256	4,096	16	2.21	–	–
Titok-B [33]	1D	64	–	256 × 256	4,096	16	1.70	–	–
Titok-S [33]	1D	128	–	256 × 256	4,096	16	1.71	–	–

Method	Codebook size	Codebook dim	Transformer scale	rFID↓	LPIPS↓	gFID↓	IS↑
LFQ	16,384	0	342M	1.58	0.2261	3.40	228.03
IBQ	16,384	256	342M	1.37	0.2235	2.88	254.73

Table 5. Performance comparison with LFQ.

THANKS

感谢聆听