香 港 大 學
THE UNIVERSITY OF HONG KONG

CVPR Nashville JUNE 11-15, 2025

Parallelized Autoregressive Visual Generation

CVPR 2025 Highlight🔥

Yuqing Wang, Shuhuai Ren, Zhijie Lin, Yujin Han
Haoyuan Guo, Zhenheng Yang, Difan Zou, Jiashi Feng, Xihui Liu

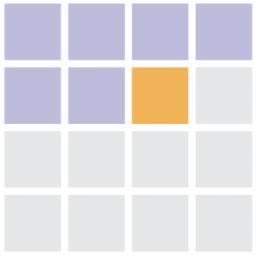
University of Hong Kong, ByteDance Seed, Peking University

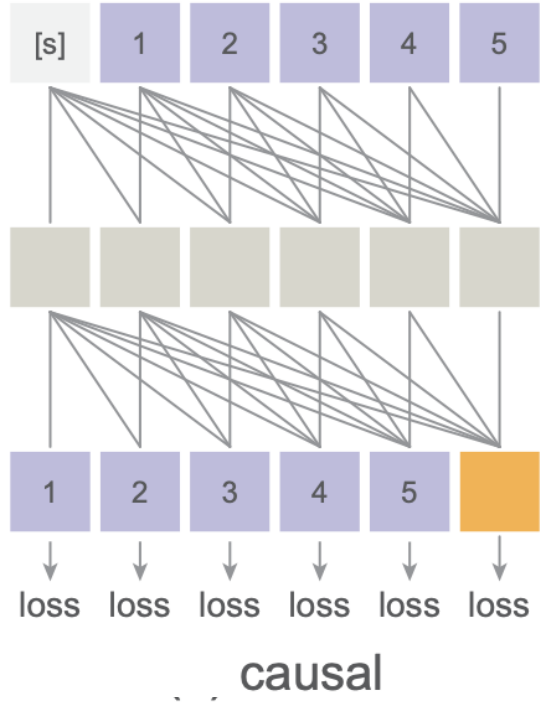
ByteDance

PEKING UNIVERSITY 1898

Motivation

- Visual autoregressive generation is a promising approach to achieve unified multimodal modeling - for both generation and understanding across text and vision
- Autoregressive models suffer from slow inference speed due to their sequential token-by-token prediction process





Li, Tianhong, et al. “Autoregressive image generation without vector quantization.” , NeurIPS 2024

Latency Comparison

Latency: 12.41s

```
-1/-1->0->-1 [21] -1/-1/-1->0->-1 [22] -1/-1/-1->0->-1 [23] -1/-1/-1->0->-1 [2
4] -1/-1/-1->0->-1 [25] -1/-1/-1->0->-1 [26] -1/-1/-1->0->-1 [27] -1/-1/-1->0-
>-1 [28] -1/-1/-1->0->-1 [29] -1/-1/-1->0->-1 [30] -1/-1/-1->0->-1 [31] -1/-1/
-1->0->-1
dc61-p18-t10-n032:19711:20235 [0] NCCL INFO P2P Chunksize set to 131072
dc61-p18-t10-n032:19711:20235 [0] NCCL INFO Connected all rings
dc61-p18-t10-n032:19711:20235 [0] NCCL INFO Connected all trees
dc61-p18-t10-n032:19711:20235 [0] NCCL INFO 32 coll channels, 0 nvls channels,
 32 p2p channels, 32 p2p channels per peer
dc61-p18-t10-n032:19711:20235 [0] NCCL INFO comm 0xb64cee0 rank 0 nranks 1 cud
aDev 0 nvmlDev 0 busId 14000 commId 0x5c89da2bc4d51b8b - Init COMPLETE
Total number of images that will be sampled: 50000
  0%|                                                  | 0/50000 [00:00<?, ?it/s]
```

Traditional Autoregressive Generation(LlamaGen)

Latency: 1.31s

```
-1/-1->0->-1 [21] -1/-1/-1->0->-1 [22] -1/-1/-1->0->-1 [23] -1/-1/-1->0->-1 [2
4] -1/-1/-1->0->-1 [25] -1/-1/-1->0->-1 [26] -1/-1/-1->0->-1 [27] -1/-1/-1->0-
>-1 [28] -1/-1/-1->0->-1 [29] -1/-1/-1->0->-1 [30] -1/-1/-1->0->-1 [31] -1/-1/
-1->0->-1
dc61-p18-t10-n032:14858:15154 [0] NCCL INFO P2P Chunksize set to 131072
dc61-p18-t10-n032:14858:15154 [0] NCCL INFO Connected all rings
dc61-p18-t10-n032:14858:15154 [0] NCCL INFO Connected all trees
dc61-p18-t10-n032:14858:15154 [0] NCCL INFO 32 coll channels, 0 nvls channels,
 32 p2p channels, 32 p2p channels per peer
dc61-p18-t10-n032:14858:15154 [0] NCCL INFO comm 0x97ea730 rank 0 nranks 1 cud
aDev 0 nvmlDev 0 busId 14000 commId 0x635c2904f5220450 - Init COMPLETE
Total number of images that will be sampled: 50000
  0%|                                                  | 0/50000 [00:00<?, ?it/s]
```

Parallelized Autoregressive Generation (16x)

Key Insight

Parallel autoregressive generation depends **on visual token dependencies**

Independent sampling of highly dependent tokens leads to inconsistent predictions

Strongly dependent adjacent tokens are difficult to generate together

Distant tokens with weak dependencies can be generated in parallel

Main Idea

Generate **distant tokens with weak dependencies in parallel** while maintaining **sequential generation for strongly dependent local tokens**

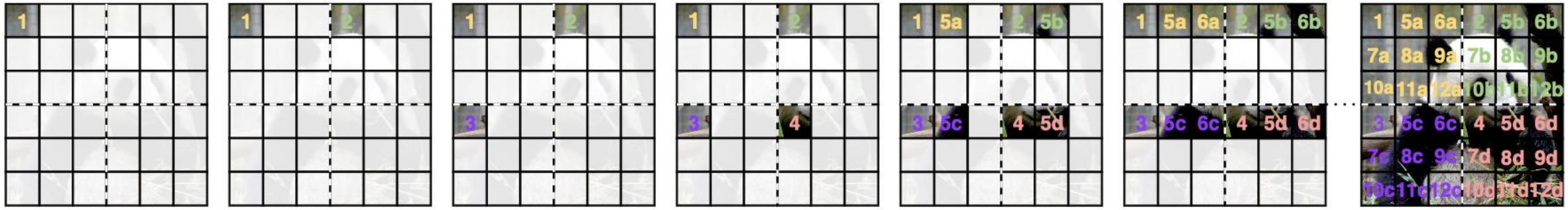


Figure 3. **Illustration of our non-local parallel generation process.** Stage 1: sequential generation of initial tokens (1-4) for each region (separated by dotted lines) to establish global structure. Stage 2: parallel generation at aligned positions across different regions (e.g., 5a-5d), then moving to next aligned positions (6a-6d, 7a-7d, etc.) for parallel generation. Same numbers indicate tokens generated in the same step, and letter suffix (a,b,c,d) denotes different regions .

Ablation

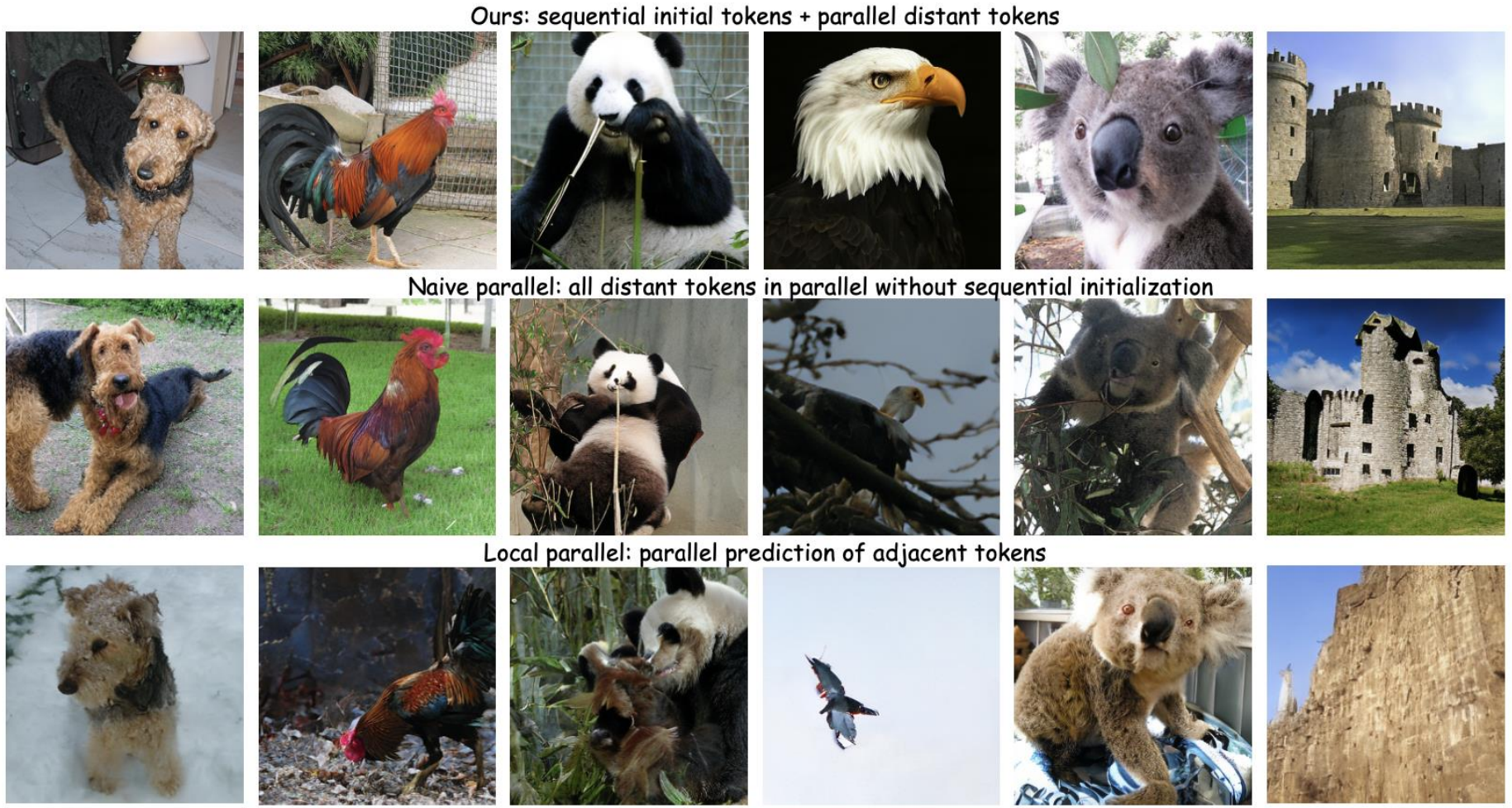


Figure 5. **Qualitative comparison of parallel generation strategies. Top:** Our method with sequential initial tokens followed by parallel distant token prediction produces high-quality and coherent images. **Middle:** Direct parallel prediction without sequential initial tokens leads to inconsistent global structures. **Bottom:** Parallel prediction of adjacent tokens results in distorted local patterns and broken details.

Main Results

Achieves a **3.6×** speedup with comparable quality and up to **9.5×** speedup with minimal quality degradation across **both image and video generation** tasks

Main Results

Achieves a **3.6×** speedup with comparable quality and up to **9.5×** speedup with minimal quality degradation across **both image and video generation** tasks

Type	Model	#Para.	FID↓	IS↑	Precision↑	Recall↑	Steps	Time(s)↓
GAN	BigGAN [3]	112M	6.95	224.5	0.89	0.38	1	–
	GigaGAN [19]	569M	3.45	225.5	0.84	0.61	1	–
	StyleGan-XL [40]	166M	2.30	265.1	0.78	0.53	1	0.08
Diffusion	ADM [10]	554M	10.94	101.0	0.69	0.63	250	44.68
	CDM [16]	–	4.88	158.7	–	–	8100	–
	LDM-4 [38]	400M	3.60	247.7	–	–	250	–
	DiT-XL/2 [34]	675M	2.27	278.2	0.83	0.57	250	11.97
Mask	MaskGIT [5]	227M	6.18	182.1	0.80	0.51	8	0.13
VAR	VAR-d30 [49]	2B	1.97	334.7	0.81	0.61	10	0.27
MAR	MAR [25]	943M	1.55	303.7	0.81	0.62	64	28.24
AR	VQGAN [11]	227M	18.65	80.4	0.78	0.26	256	5.05
	VQGAN [11]	1.4B	15.78	74.3	–	–	256	5.05
	VQGAN-re [11]	1.4B	5.20	280.3	–	–	256	6.38
	ViT-VQGAN [64]	1.7B	4.17	175.1	–	–	1024	>6.38
	ViT-VQGAN-re [64]	1.7B	3.04	227.4	–	–	1024	>6.38
	RQTran. [23]	3.8B	7.55	134.0	–	–	256	5.58
	RQTran.-re [23]	3.8B	3.80	323.7	–	–	256	5.58
AR	LlamaGen-L [47]	343M	3.07	256.1	0.83	0.52	576	12.58
	LlamaGen-XL [47]	775M	2.62	244.1	0.80	0.57	576	18.66
	LlamaGen-XXL [47]	1.4B	2.34	253.9	0.80	0.59	576	24.91
	LlamaGen-3B [47]	3.1B	2.18	263.3	0.81	0.58	576	12.41
AR	PAR-L-4×	343M	3.76	218.9	0.84	0.50	147	3.38
	PAR-XL-4×	775M	2.61	259.2	0.82	0.56	147	4.94
	PAR-XXL-4×	1.4B	2.35	263.2	0.82	0.57	147	6.84
	PAR-3B-4×	3.1B	2.29	255.5	0.82	0.58	147	3.46
	PAR-XXL-16×	1.4B	3.02	270.6	0.81	0.56	51	2.28
	PAR-3B-16×	3.1B	2.88	262.5	0.82	0.56	51	1.31

Image Generation Performance

Type	Method	#Param	FVD↓	Steps	Time(s)
Diffusion	VideoFusion [29]	N/A	173	-	-
	Make-A-Video [41]	N/A	81.3	-	-
	HPDM-L [42]	725M	66.3	-	-
Mask.	MAGVIT [66]	306M	76	-	-
	MAGVITv2 [67]	840M	58	-	-
AR	CogVideo [17]	9.4B	626	-	-
	TATS [12]	321M	332	-	-
	OmniTokenizer [60]	650M	191	5120	336.70
	MAGVITv2-AR [67]	840M	109	1280	-
AR	PAR-1×	792M	94.1	1280	43.30
	PAR-4×	792M	99.5	323	11.27
	PAR-16×	792M	103.4	95	3.44

Video Generation Performance