Parallelized Autoregressive Visual Generation

Yuqing Wang[1] **Shuhuai Ren**[3] **Zhijie Lin**[2†] **Yujin Han**[1]
Haoyuan Guo[2] **Zhenheng Yang**[2] **Difan Zou**[1] **Jiashi Feng**[2] **Xihui Liu**[1*]
[1]University of Hong Kong [2]ByteDance Seed [3]Peking University

郝飞洋

2025年1月19日

Key Insight

- Parallel generation depends on **visual token dependencies**:
- Tokens with weak dependencies can be generated in parallel,
- while strongly dependent adjacent tokens are difficult to generate together, as their independent sampling may lead to inconsistencies.

Highlight

- develop a **parallel generation strategy** that generates distant tokens with weak dependencies in parallel while maintaining sequential generation for strongly dependent local tokens.
- seamlessly integrated into standard autoregressive models **without modifying the architecture or tokenizer**.

Predict multiple tokens in parallel

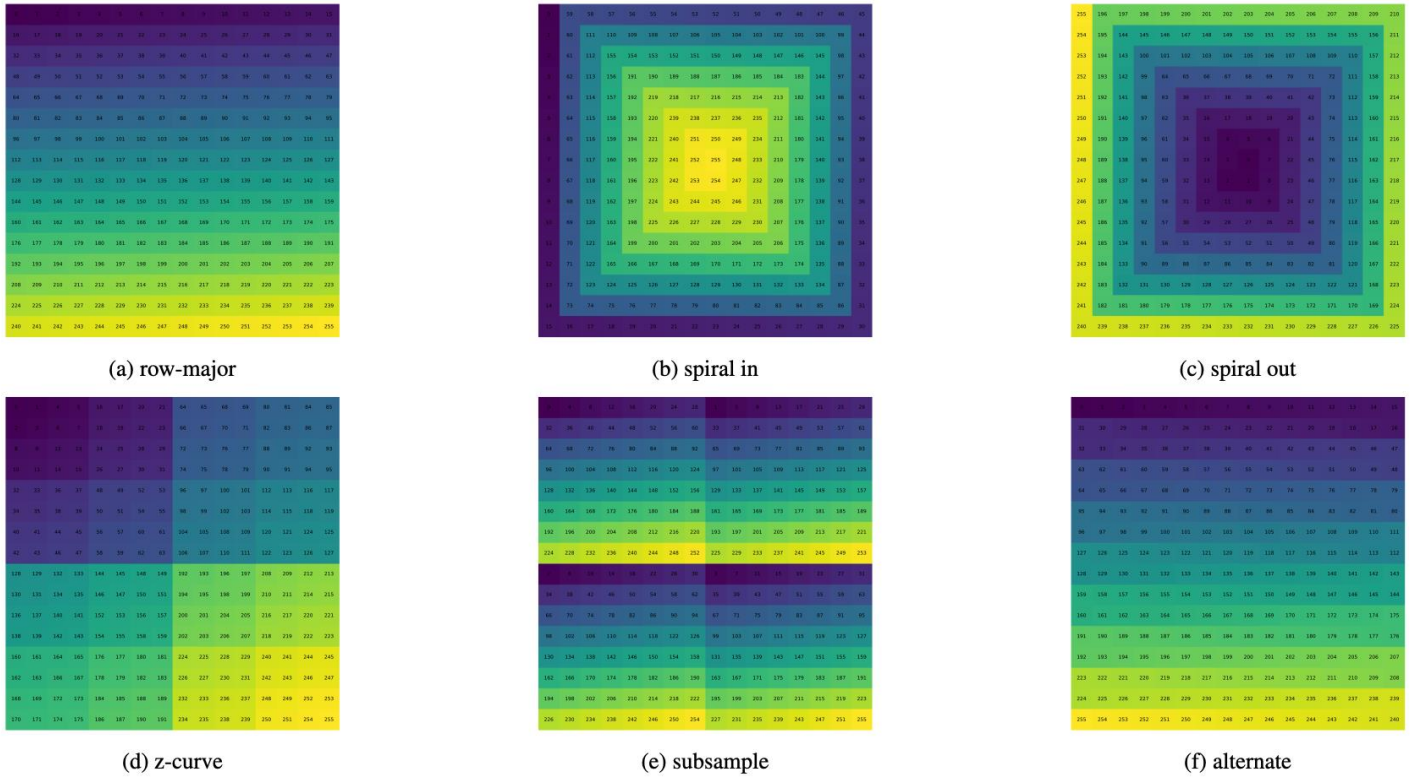
Language modeling

- Speculative decoding推测解码
- Jacobi decoding雅可比解码
- achieve parallel generation through auxiliary draft models or iterative refinement.

Visual domain

- MaskGIT employ non-autoregressive paradigms with masked modeling strategies.
- VAR achieves faster speed through next-scale prediction that requires specially designed multiscale tokenizers and longer token sequences.

来自RAR的结果



(a) row-major (b) spiral in (c) spiral out

(d) z-curve (e) subsample (f) alternate

Figure 6. **Different scan orders for a** 16×16 **grid (256 tokens).** The number indicates the token's indices in the scanning order.

scan order	FID↓	IS↑	Precision↑	Recall↑
row-major	2.18	269.7	0.83	0.55
spiral in	2.50	256.1	0.84	0.54
spiral out	2.46	256.6	0.84	0.54
z-curve	2.29	262.7	0.83	0.55
subsample	2.39	258.0	0.84	0.54
alternate	2.48	270.9	0.84	0.53

Table 3. **Effect of different scan orders RAR-L converges to.** We mainly consider 6 different scan orders (row major, spiral in, spiral out, z-curve, subsample, alternate) as studied in [22]. Our default setting is marked in gray. A visual illustration of different scan orders are available in the appendix.

Design Principles

- For visual tokens, dependencies naturally decrease with spatial distance - tokens from distant regions typically have weaker correlations than adjacent ones.
- The initial tokens of each regions are particularly crucial as they jointly determine the global image structure.

Comparison of Generation Strategies

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.

3 Key Design Principles

- generate **initial tokens** for each region **sequentially** to establish proper global structure;
- maintain **sequential** generation **within local regions** where dependencies are strong;
- enable **parallel** generation **across regions** where dependencies are weak through proper token organization.

Non-Local Parallel Generation

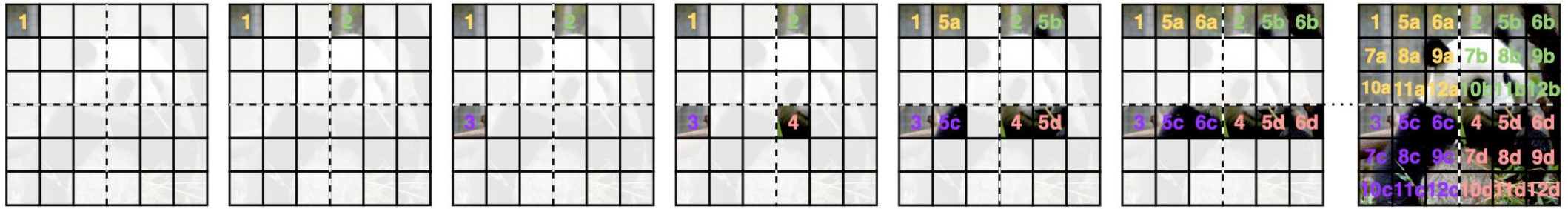
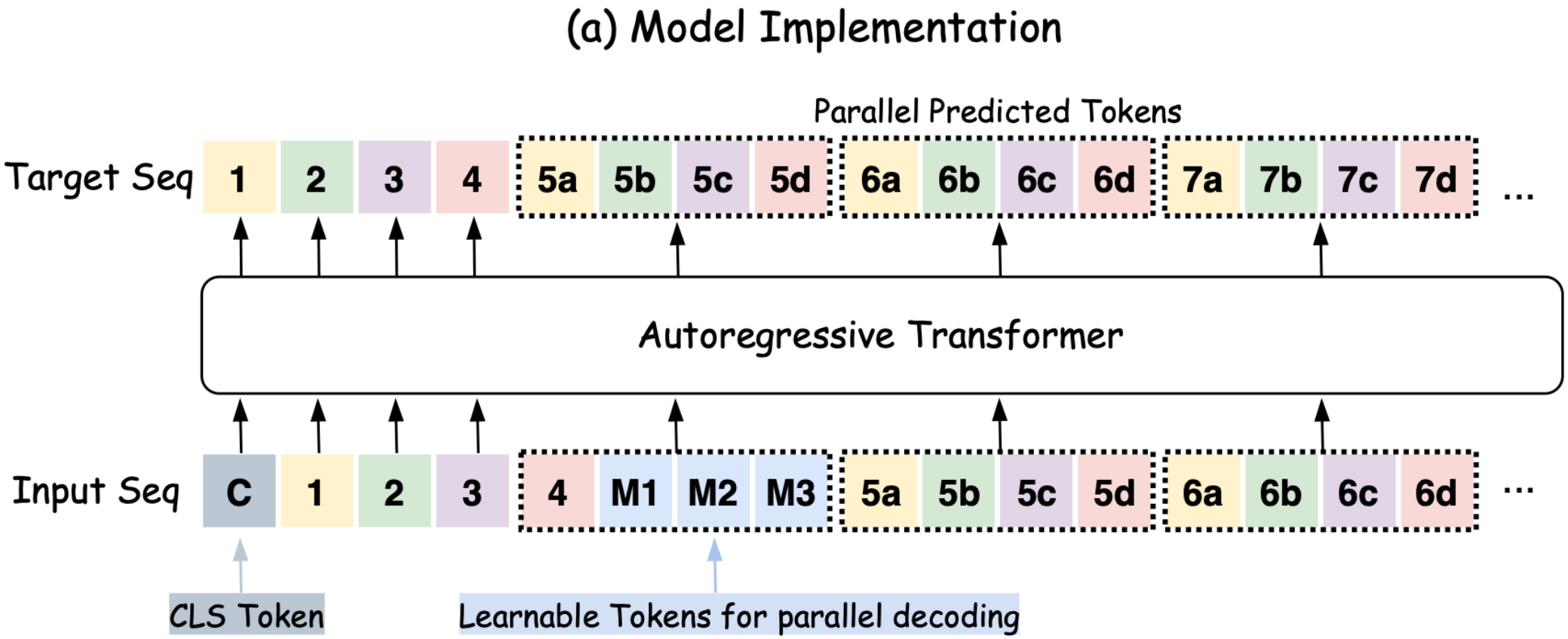
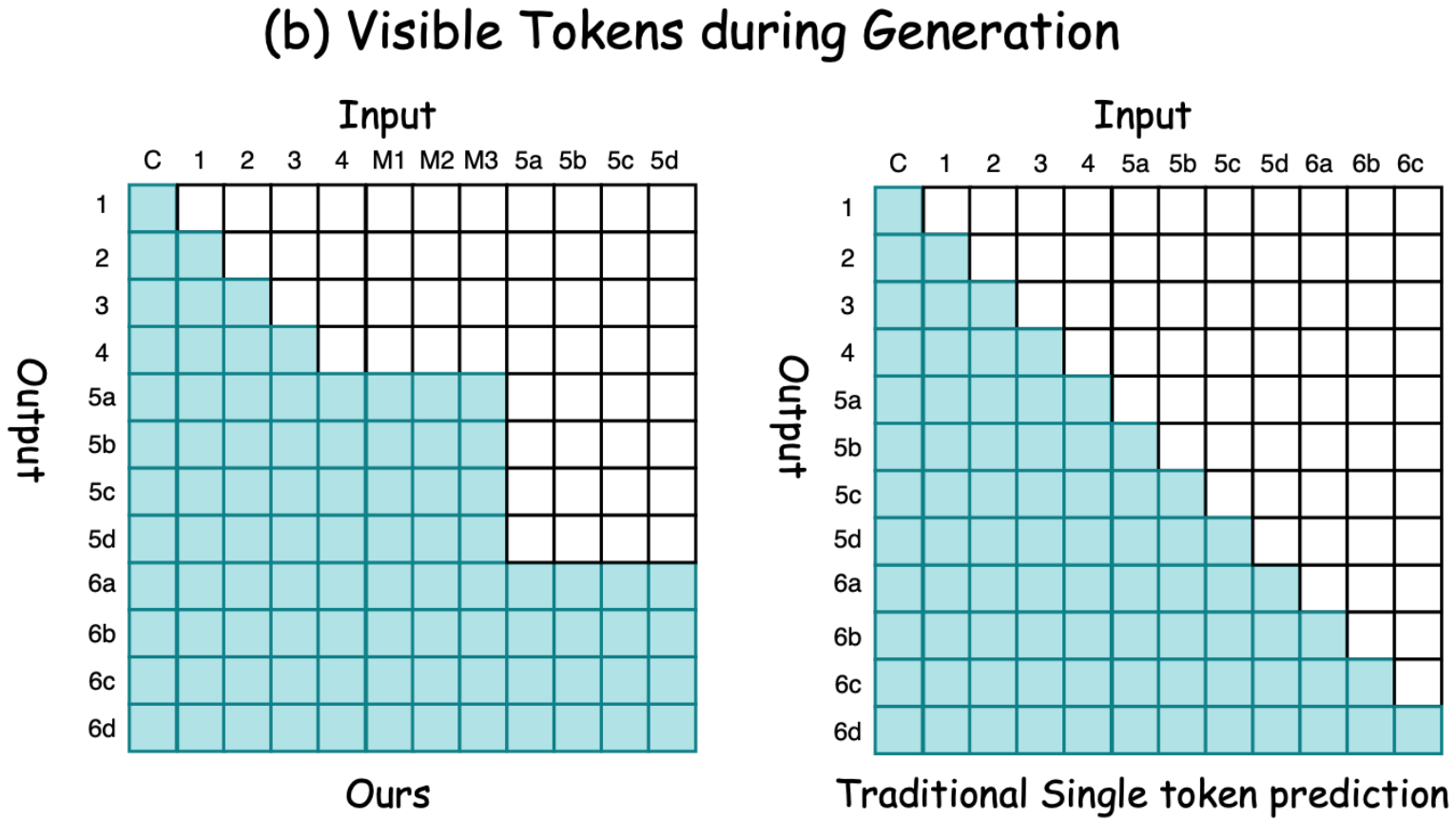


Figure 3. **Illustration of our non-local parallel generation process.** Stage 1: sequential generation of initial tokens (1-4) for each region (separated by dotted lines) to establish global structure. Stage 2: parallel generation at aligned positions across different regions (e.g., 5a-5d), then moving 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 .

Framework Implementation



Group-wise Bi-directional Attention with Global Autoregression



Extension to Video Generation

- While we also explored parallel generation along the temporal dimension, we found it less effective than spatial parallelization.
- This is because temporal dependencies exhibit stronger sequential characteristics that are fundamental to video coherence, making them less suitable for parallel prediction compared to spatial relationships.

Image Generation

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

Table 2. **Class-conditional image generation on ImageNet 256×256 benchmark**. "↓" or "↑" indicate lower or higher values are better. "-re" means using rejection sampling. PAR-4× and PAR-16× means generating 4 and 16 tokens per step in parallel, respectively.

tokenizer	type	generator	#params	FID↓	IS↑	Pre.↑	Rec.↑
VQ [50]	Diff.	LDM-8 [50]	258M	7.76	209.5	0.84	0.35
VAE [50]	Diff.	LDM-4 [50]	400M	3.60	247.7	0.87	0.48
VAE [51]	Diff.	UViT-L/2 [6]	287M	3.40	219.9	0.83	0.52
		UViT-H/2 [6]	501M	2.29	263.9	0.82	0.57
		DiT-L/2 [45]	458M	5.02	167.2	0.75	0.57
		DiT-XL/2 [45]	675M	2.27	278.2	0.83	0.57
		SiT-XL [40]	675M	2.06	270.3	0.82	0.59
		DiMR-XL/2R [37]	505M	1.70	289.0	0.79	0.63
		MDTv2-XL/2 [25]	676M	1.58	314.7	0.79	0.65
VQ [10]	Mask.	MaskGIT [10]	177M	6.18	182.1	-	-
VQ [73]	Mask.	TiTok-S-128 [73]	287M	1.97	281.8	-	-
VQ [72]	Mask.	MAGVIT-v2 [72]	307M	1.78	319.4	-	-
VQ [65]	Mask.	MaskBit [65]	305M	1.52	328.6	-	-
VAE [36]	MAR	MAR-B [36]	208M	2.31	281.7	0.82	0.57
		MAR-L [36]	479M	1.78	296.0	0.81	0.60
		MAR-H [36]	943M	1.55	303.7	0.81	0.62
VQ [58]	VAR	VAR-d30 [58]	2.0B	1.92	323.1	0.82	0.59
		VAR-d30-re [58]	2.0B	1.73	350.2	0.82	0.60
VQ [22]	AR	GPT2 [22]	1.4B	15.78	74.3	-	-
		GPT2-re [22]	1.4B	5.20	280.3	-	-
VQ [69]	AR	VIM-L [69]	1.7B	4.17	175.1	-	-
		VIM-L-re [69]	1.7B	3.04	227.4	-	-
VQ [39]	AR	Open-MAGVIT2-B [39]	343M	3.08	258.3	0.85	0.51
		Open-MAGVIT2-L [39]	804M	2.51	271.7	0.84	0.54
		Open-MAGVIT2-XL [39]	1.5B	2.33	271.8	0.84	0.54
VQ [52]	AR	LlamaGen-L [52]	343M	3.80	248.3	0.83	0.51
		LlamaGen-XL [52]	775M	3.39	227.1	0.81	0.54
		LlamaGen-XXL [52]	1.4B	3.09	253.6	0.83	0.53
		LlamaGen-3B [52]	3.1B	3.05	222.3	0.80	0.58
		LlamaGen-L-384 [52]	343M	3.07	256.1	0.83	0.52
		LlamaGen-XL-384 [52]	775M	2.62	244.1	0.80	0.57
		LlamaGen-XXL-384 [52]	1.4B	2.34	253.9	0.80	0.59
		LlamaGen-3B-384 [52]	3.1B	2.18	263.3	0.81	0.58
VQ [10]	AR	RAR-B (ours)	261M	1.95	290.5	0.82	0.58
		RAR-L (ours)	461M	1.70	299.5	0.81	0.60
		RAR-XL (ours)	955M	1.50	306.9	0.80	0.62
		RAR-XXL (ours)	1.5B	**1.48**	326.0	0.80	0.63

Table 4. **ImageNet-1K** 256 × 256 **generation results evaluated with ADM [18].** "type" refers to the type of the generative model, where "Diff." and "Mask." stand for diffusion models and masked transformer models, respectively. "VQ" denotes discrete tokenizers and "VAE" stands for continuous tokenizers. "-re" stands for rejection sampling. "-384" denotes for generating images at resolution 384 and resize back to 256 for evaluation, as is used in [52].

AR	PAR-L-4×	343M	3.76	218.9	0.84	0.50
	PAR-XL-4×	775M	2.61	259.2	0.82	0.56
	PAR-XXL-4×	1.4B	2.35	263.2	0.82	0.57
	PAR-3B-4×	3.1B	2.29	255.5	0.82	0.58
	PAR-XXL-16×	1.4B	3.02	270.6	0.81	0.56
	PAR-3B-16×	3.1B	2.88	262.5	0.82	0.56

Video Generation

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	-	-
	MAGVIT-v2 [67]	840M	58	-	-
AR	CogVideo [17]	9.4B	626	-	-
	TATS [12]	321M	332	-	-
	OmniTokenizer [60]	650M	191	5120	336.70
	MAGVIT-v2-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

Table 3. **Comparison of class-conditional video generation methods on UCF-101 benchmark.** FVD measures generation quality, where lower values (↓) indicate better performance. PAR-1× represents our token-by-token baseline, while PAR-4× and PAR-16× indicate our parallel generation variants with different speedup ratios, achieving competitive FVD scores with significantly reduced generation steps and wall-clock time.

Ablation Study

	FID↓	IS↑	steps↓
w/o	3.67	221.36	144
w	**2.61**	259.17	147

(a) **Importance of initial sequential token generation.** Sequential generation of initial tokens improves FID by 1.06 with negligible step increase.

n	FID↓	IS↑	steps↓
1	**2.34**	253.90	576
4	2.35	263.24	147
16	3.02	270.57	51

(b) **Number of parallel predicted tokens** (PAR-XXL). n=1 is the token-by-token baseline. n=4 reduces steps by 4× with similar FID (2.35 vs. 2.34), while n=16 reduces steps by 11.3× at the cost of 0.67 FID.

attn	FID↓	IS↑	steps↓
causal	3.64	228.08	147
full	**2.61**	259.17	147

(c) **Attention pattern between parallel tokens.** Full attention allows complete context access from previous parallel groups (vs. causal attention's limited access), bringing 1.03 FID improvement.

order	pattern	FID↓	IS↑	steps↓
raster	one	2.62	244.08	576
distant	one	2.64	262.72	576
raster	multi	5.64	265.46	147
distant	multi	**2.61**	259.17	147

(d) **Comparison of different scan orders under single-token and multi-token prediction.** Our region-based distant ordering shows similar performance with raster scan in single-token setting, but significantly outperforms in multi-token prediction (2.61 vs. 5.64 FID).

Params	FID↓	IS↑	steps
343M	3.76	218.92	147
775M	2.61	259.17	147
1.4B	2.35	263.24	147
3.1B	**2.29**	255.46	147

(e) **Scaling of model size** (4× parallel). Generation quality steadily improves with more parameters, from 343M (FID 3.76) to 3.1B (FID 2.29).

Table 4. **Ablation studies on image generation model designs.**