

Table 1: **Generative model family comparison on class-conditional ImageNet 256x256.** ***** ... or ... indicate lower or higher values are better. Metrics include Fréchet inception distance (FID), inception score (IS), precision (Pre) and recall (rec). **#Step**: the number of model runs needed to generate an image. Wall-clock inference time relative to VAR is reported. Models with the suffix **--re** used rejection sampling. **·**: taken from MaskGIT [17].

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]	3.0	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.) [58]	4.49	215.5	—	—	502M	20	1.9 [51]
AR	VQVAE-2+ [68]	1.11	~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-d16	3.30	274.4	0.84	0.51	310M	10	0.4
VAR	VAR-d20	2.57	302.6	0.83	0.56	600M	10	0.5
VAR	VAR-d24	2.09	312.9	0.82	0.59	1.0B	10	0.6
VAR	VAR-d30	1.92	323.1	0.82	0.59	2.0B	10	1
VAR	VAR-d30-re (validation on 50)	1.73	350.2	0.82	0.60	2.0B	10	1

****Overall comparison.** In comparison with existing generative approaches including generative adversarial networks (GAN), diffusion models (Diff.), BERT-style masked-prediction models (Mask.), and GPT-style autoregressive models (AR), our visual autoregressive (VAR) establishes a new model class. As shown in Tab. 1, VAR not only achieves the best FID/IS but also demonstrates remarkable speed in image generation. VAR also maintains decent precision and recall, confirming its semantic consistency. These advantages hold true on the 512x512 synthesis benchmark, as detailed in Tab. 2. Notably, VAR significantly advances traditional AR capabilities. To our knowledge, this is the first time of autoregressive models outperforming Diffusion transformers, a milestone made possible by VAR’s resolution of AR limitations discussed in Section 3.

****Efficiency comparison.** Conventional autoregressive (AR) models [30, 68, 92, 50] suffer a lot from the high computational cost, as the number of image tokens is quadratic to the image resolution. A full autoregressive generation of n^2 tokens requires $O(n^2)$ decoding iterations and $O(n^4)$ total computations. In contrast, VAR only requires $O(\log(n))$ iterations and $O(n^4)$ total computations. The wall-clock time reported in Tab. 1 also provides empirical evidence that VAR is around 20 times faster than VQGAN and ViT-VQGAN even with more model parameters, reaching the speed of efficient GAN models which only require 1 step to generate an image.

Table 2: **ImageNet 512x512 conditional generation.** ***** **·**: quoted from MaskGIT [17]. **--s**: a single shared AdaLN layer is used due to resource limitation.

Type	Model	FID↓	IS↑	Time
GAN	BigGAN [13]	8.43	177.9	—
Diff.	ADM [26]	23.24	101.0	—
Diff.	DiT-XL/2 [63]	3.04	240.8	81
Mask.	MaskGIT [17]	7.32	156.0	0.5†
AR	VQGAN [30]	26.52	66.8	25†
VAR	VAR-d36-s	2.63	303.2	1

****Compared with popular diffusion transformer.** The VAR model surpasses the recently popular diffusion model Stable Diffusion Transformer (DiT), which serves as the precursor to the latest Stable-Diffusion 3 [29] and SORA [14], in multiple dimensions: 1) In image generation diversity and quality