
Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction

Keyu Tian^{1, 2}, Yi Jiang^{2, ,}, Zehuan Yuan^{2, *}, Bingyue Peng², Liwei Wang¹,

¹Peking University ²Bytedance Inc

keyutian@stu.pku.edu.cn, jiangyi.enjoy@bytedance.com,
yuanzehuan@bytedance.com, pengby@pku.edu.cn

Try and explore our online demo at: <https://var.vision>
models: <https://github.com/FoundationVision/VAR>

Codes and



Figure 1: Generated samples from Visual AutoRegressive (VAR) transformers trained on ImageNet. We show 5x512 samples (top), 256x256 samples (middle), and zero-shot image editing results (bottom).

Abstract

We present Visual AutoRegressive modeling (VAR), a new generation paradigm that redefines the autoregressive learning on images as coarse-to-fine ·next-scale prediction· or ·next-resolution prediction·, diverging from the standard raster-scan ·next-token prediction·. This simple, intuitive methodology allows autoregressive (AR) transformers to learn visual distribution s fast and can generalize well: VAR, for the *first time*, makes GPT-style AR models surpass s diffusion transformers in image generation. On ImageNet 256x256 benchmark, VAR signifi cantly improve AR baseline by improving Fréchet inception distance (FID) from 18.65 to 1.73 , inception score (IS) from 80.4 to 350.2, with 20x faster inference speed. It is also empirically verified that VAR outperforms the Diffusion Transformer (DIT) in multiple dimensions includ ing image quality, inference speed, data efficiency, and scalability. Scaling up VAR models e xhibits clear power-law scaling laws similar to those observed in LLMs, with linear correlation coefficients near -0.998 as solid evidence. VAR further showcases zero-shot generalization ability in downstream tasks including image in-painting, out-painting, and editing. These resul ts suggest VAR has initially emulated the two important properties of LLMs: **Scaling Laws** and **zero-shot** generalization. We have released all models and codes to promote the exp loration of AR/VAR models for visual generation and unified learning.

*Corresponding authors: wanglw@pku.edu.cn, yuanzehuan@bytedance.com; :- project lead