Formula Size Test

Einstein's famous equation is $E = mc^2$ where E is energy, m is mass, and c is the speed of light.

Newton's second law states that $F = ma$ where F is force, m is mass, and a is acceleration.

The quadratic formula is $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ which solve

More complex example: The Gaussian integral is $\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$ which is fundament

ssive Modeling: Scalable Image Generation via Next-Scale Prediction

Keyu Tian[1,2], Yi Jiang[2,†], Zehuan Yuan[2,*], Bingyue Peng[2], Liwei Wang[1,*]

g University [2]Bytedance Inc

u.edu.cn, jiangyi.enjoy@bytedance.com,
yuanzehuan@bytedance.com, bingyue.peng@bytedance.com, wanglw@pku.edu.cn

Try and explore our online demo at: https://var.vision

Codes and models: https://github.com/FoundationVision/VAR

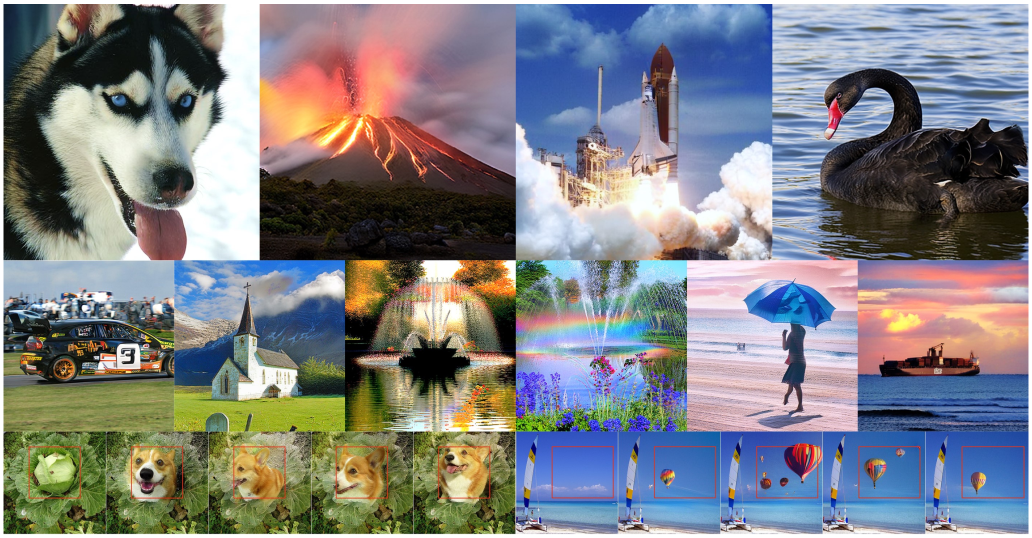

Figure 1: **Generated samples from Visual AutoRegressive (VAR) transformers trained on ImageNet**. We show 512×512 samples (top), 256×256 samples (middle), and zero-shot image editing results (bottom).

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 distributions fast and can generalize well: VAR, for the *first time*, makes GPT-style AR models surpass diffusion transformers in image generation. On ImageNet 256×256 benchmark, VAR significantly 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 20× faster inference speed. It is also empirically verified that VAR outperforms the Diffusion Transformer (DiT) in multiple dimensions including image quality, inference speed, data efficiency, and scalability. Scaling up VAR models exhibits 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 results 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 exploration of AR/VAR models for visual generation and unified learning.

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