ImageBART: Bidirectional Context with Multinomial Diffusion for Autoregressive Image Synthesis

Patrick Esser[1,2]*   Robin Rombach[1,2]*   Andreas Blattmann[1,2]*   Björn Ommer[1,2]
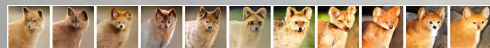
[1]Ludwig Maximilian University, Munich   [2]Heidelberg University, HCI/IWR
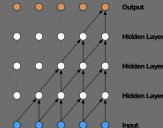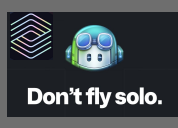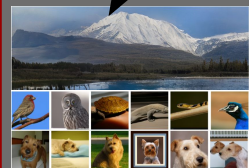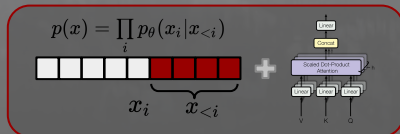
*equal contribution

NEURAL INFORMATION PROCESSING SYSTEMS

Autoregressive Models are powerful ...

$$p(x) = \prod_i p_\theta(x_i \mid x_{<i})$$

x_i $x_{<i}$

Scaled Dot-Product Attention

Don't fly solo.

Output
Hidden Layer
Hidden Layer
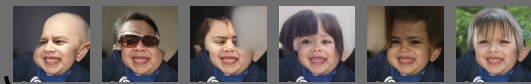Hidden Layer
Input

... but have no explicit global image representation

Conditional Generation
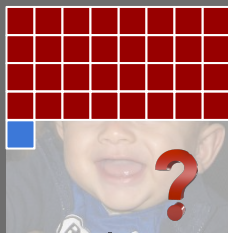
Exposure Bias

→ unrealistic modifications

Global Context by Reverting a Multinomial Diffusion Process (MDP)

fixed multinomial diffusion process

x_1 $q_\theta^0(x_2 \mid x_1)$ $q_\theta^T(x_T \mid x_{T-1})$

$p_\theta^1(x_1 \mid x_2)$ $x_T \sim p_\theta^T(x_T)$ $p_\theta^{T-1}(x_{T-1} \mid x_T)$

Discrete Representation Learning (VQGAN)

q_θ^1 CNN Encoder

p_θ^0 CNN Decoder

Transformer (AR) Decoder   Transformer Encoder

generation   **learned** Markov chain   **ImageBART**

CNN Decoder

compression
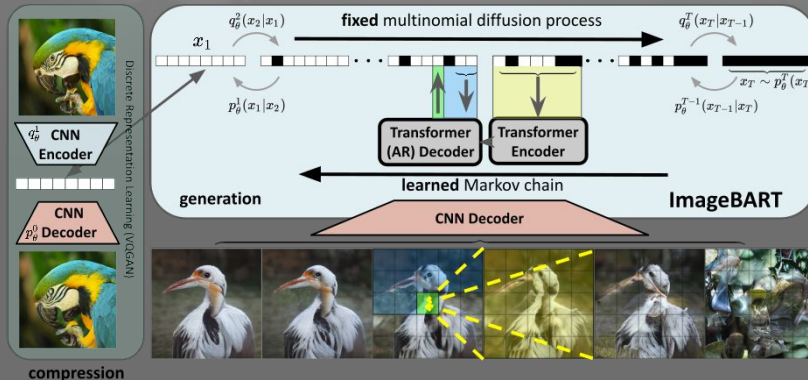
- Sequence of independent AR models reverts fixed MDP
- Global context based on preceding representation
- Efficiently learned in latent space of *VQGAN*

✓ **High-fidelity and high-resolution image synthesis**

✓ **Flexible conditional AR models**

High-Fidelity Image Synthesis
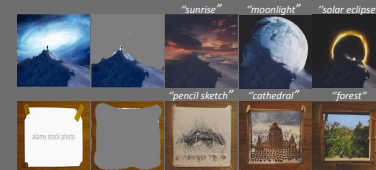
Upper Half Completion with Global Context

Masked   VQGAN+T (AR)   ImageBART

Image Editing without retraining ...

... unconditional

... conditioned on text prompts ...

"sunrise"   "moonlight"   "solar eclipse"

"pencil sketch"   "cathedral"   "forest"

... or class labels.

leopard   doberman   arctic fox

lorikeet   zebra   tiger