NEURAL INFORMATION PROCESSING SYSTEMS

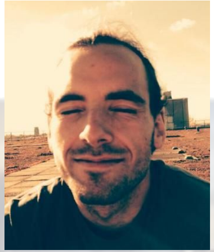
LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

IWR Interdisciplinary Center for Scientific Computing

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

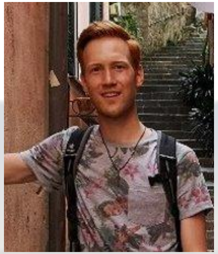
ImageBART: Bidirectional Context with Multinomial Diffusion for Autoregressive Image Synthesis

Patrick Esser*,



Robin Rombach*,

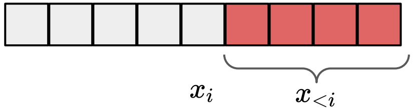
Andreas Blattmann*,



Björn Ommer

*equal contribution

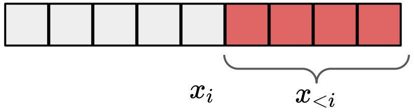
Autoregressive Generative Modeling



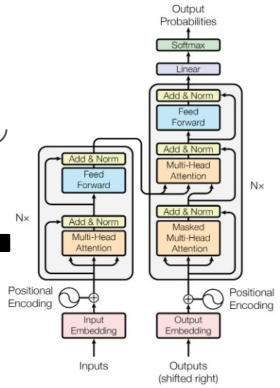
$$p(x) = \prod_i p_\theta(x_i | x_{<i})$$

Sequential Likelihood
Factorization

Autoregressive Generative Modeling



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

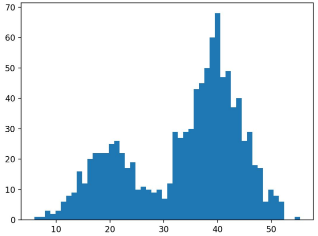


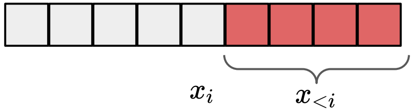
Vaswani et al. ,2017

Sequential Likelihood Factorization

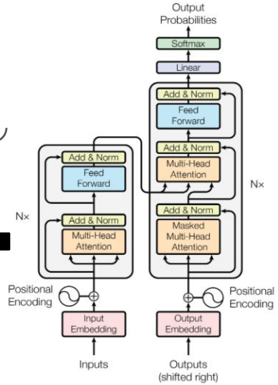
Attention Mechanism

Autoregressive Generative Modeling





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



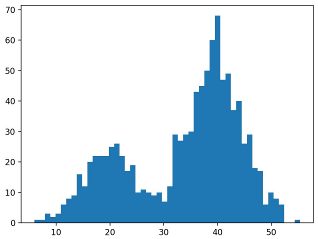
Vaswani et al. ,2017

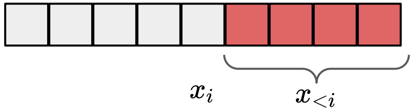
Sequential Likelihood Factorization

Attention Mechanism

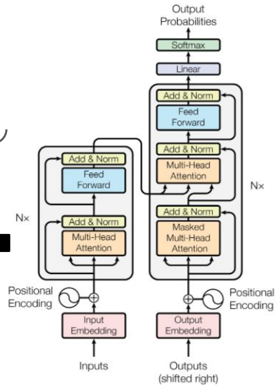
Autoregressive Generative Modeling

van den Oord et al. , 2016





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



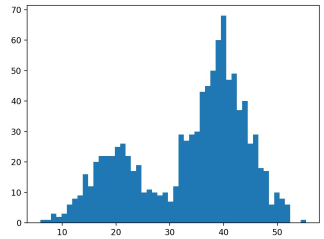
Vaswani et al. ,2017

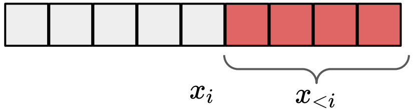
Sequential Likelihood Factorization

Attention Mechanism

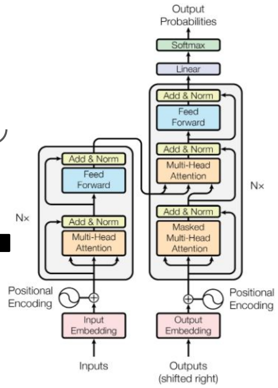
Autoregressive Generative Modeling

van den Oord et al. , 2016





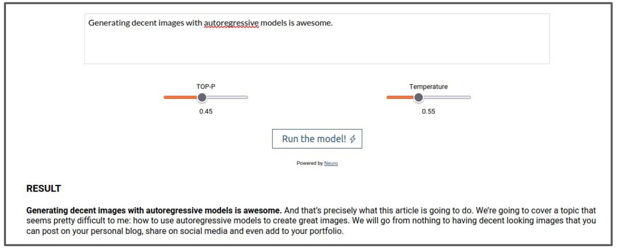
$$p(x) = \prod_i p_\theta(x_i | x_{<i})$$



Vaswani et al. ,2017

Sequential Likelihood Factorization

Attention Mechanism



Generating decent images with autoregressive models is awesome.

TOP-P

0.45

Temperature

0.55

Run the model!

Powered by Neuro

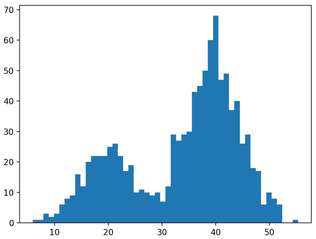
RESULT

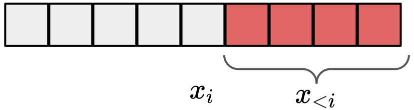
Generating decent images with autoregressive models is awesome. And that's precisely what this article is going to do. We're going to cover a topic that seems pretty difficult to me: how to use autoregressive models to create great images. We will go from nothing to having decent looking images that you can post on your personal blog, share on social media and even add to your portfolio.

https://6b.eleuther.ai/

Autoregressive Generative Modeling

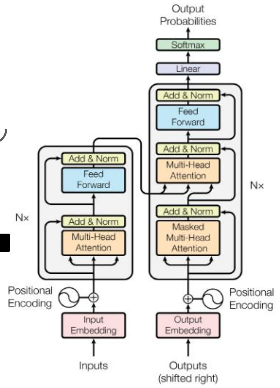
van den Oord et al. , 2016





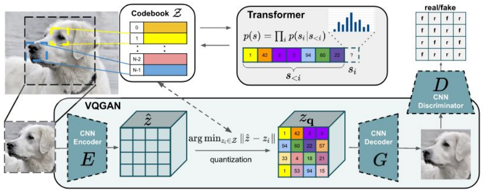
$$p(x) = \prod_i p_\theta(x_i | x_{<i})$$

Sequential Likelihood Factorization

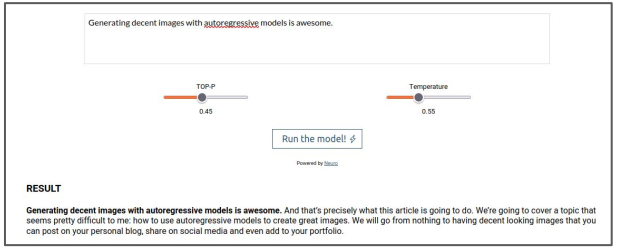


Vaswani et al. ,2017

Attention Mechanism



Esser et al. , 2020



Generating decent images with autoregressive models is awesome.

TOP-P

0.45

Temperature

0.55

Run the model!

Powered by Neuro

RESULT

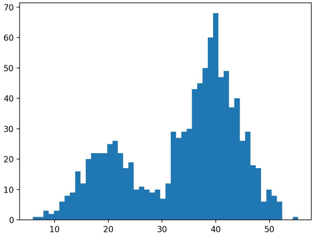
Generating decent images with autoregressive models is awesome. And that's precisely what this article is going to do. We're going to cover a topic that seems pretty difficult to me: how to use autoregressive models to create great images. We will go from nothing to having decent looking images that you can post on your personal blog, share on social media and even add to your portfolio.

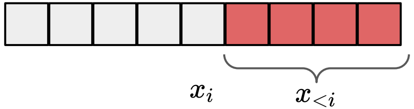
https://6b.eleuther.ai/

Autoregressive Generative Modeling

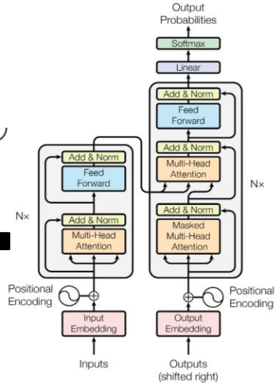
van den Oord et al. , 2016

Esser et al. , 2020

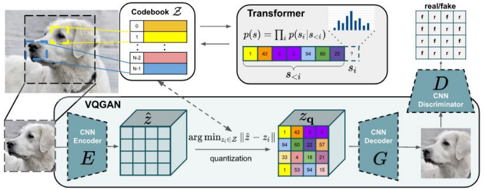




$$p(x) = \prod_i p_\theta(x_i | x_{<i})$$



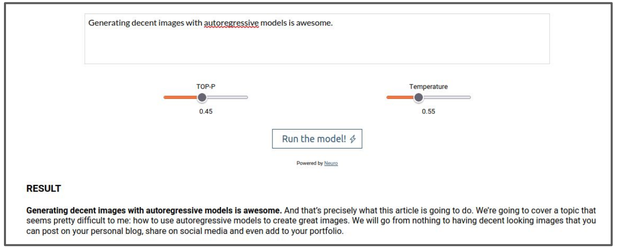
Vaswani et al. ,2017



Esser et al. , 2020

Sequential Likelihood Factorization

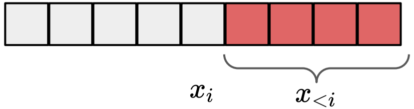
Attention Mechanism



https://6b.eleuther.ai/

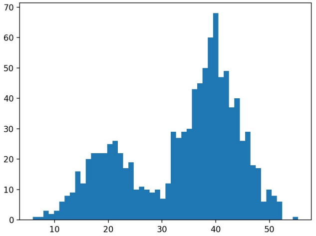
Autoregressive Generative Modeling

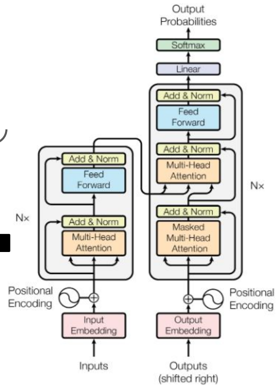
van den Oord et al. , 2016



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

Sequential Likelihood Factorization

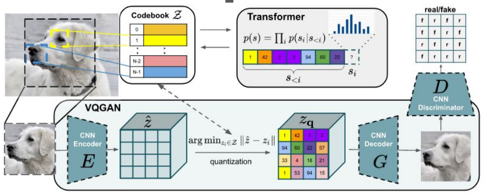




Vaswani et al. ,2017

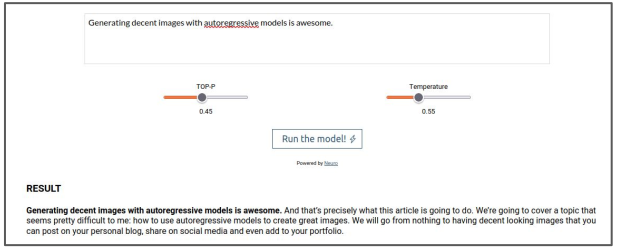
Attention Mechanism

Esser et al. , 2020



Esser et al. , 2020

Yu et al. ,2021



Generating decent images with autoregressive models is awesome.

TOP-P

0.45

Temperature

0.55

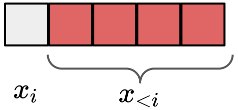
Run the model!

Powered by Neuro

RESULT

Generating decent images with autoregressive models is awesome. And that's precisely what this article is going to do. We're going to cover a topic that seems pretty difficult to me: how to use autoregressive models to create great images. We will go from nothing to having decent looking images that you can post on your personal blog, share on social media and even add to your portfolio.

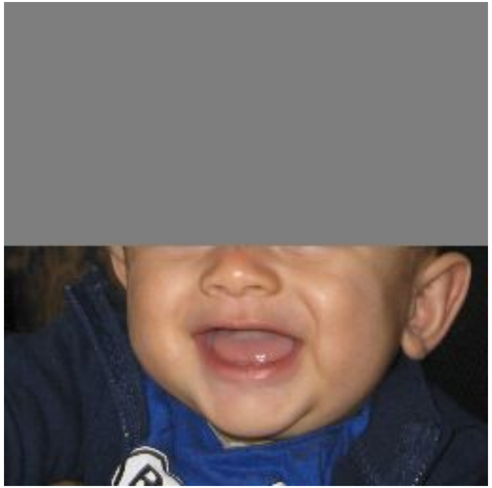
https://6b.eleuther.ai/

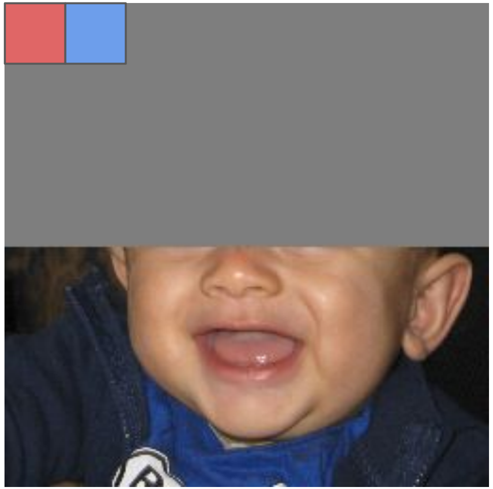


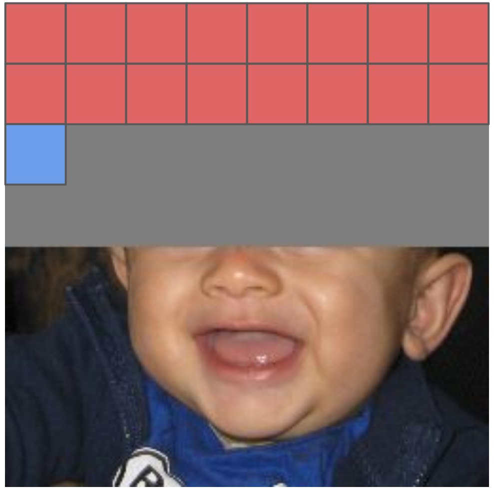
$$p(x) = \prod_i p_\theta(x_i | x_{<i})$$

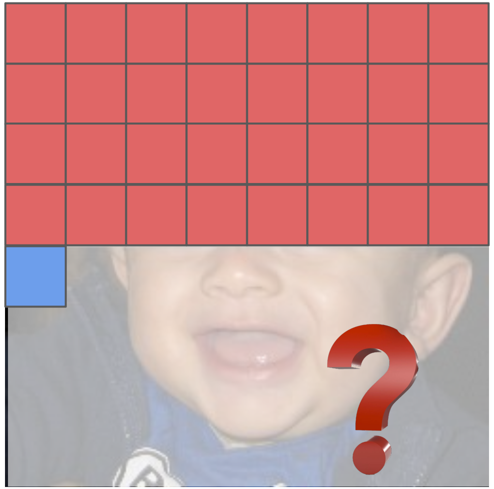
Sequential Likelihood

Factorization

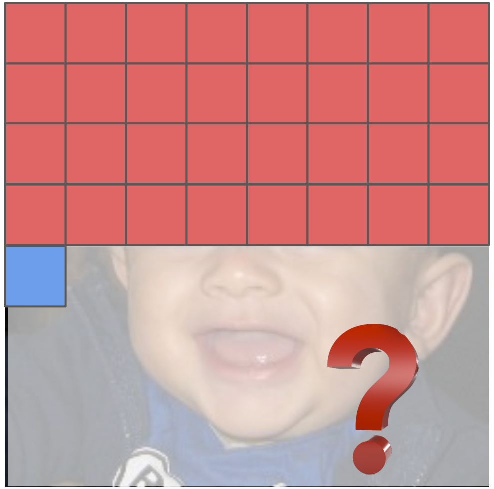




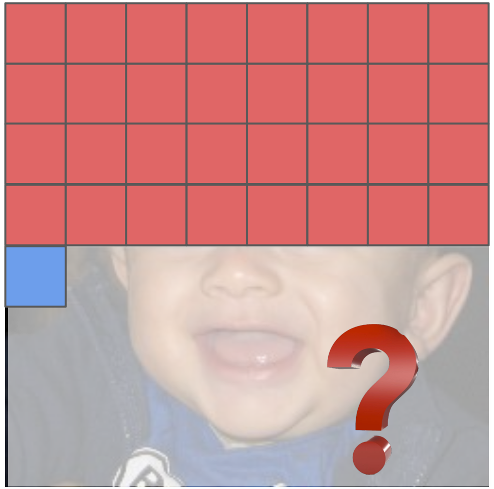




No global image representation

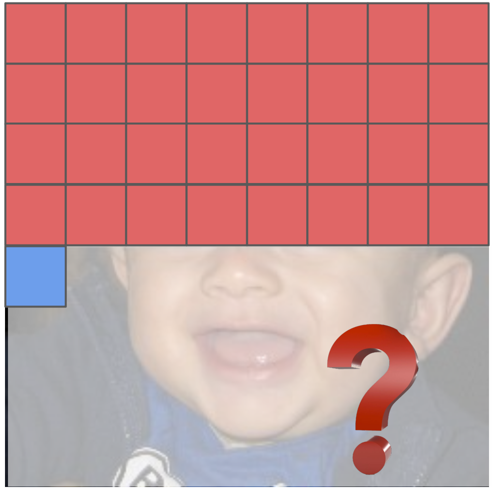


No global image representation



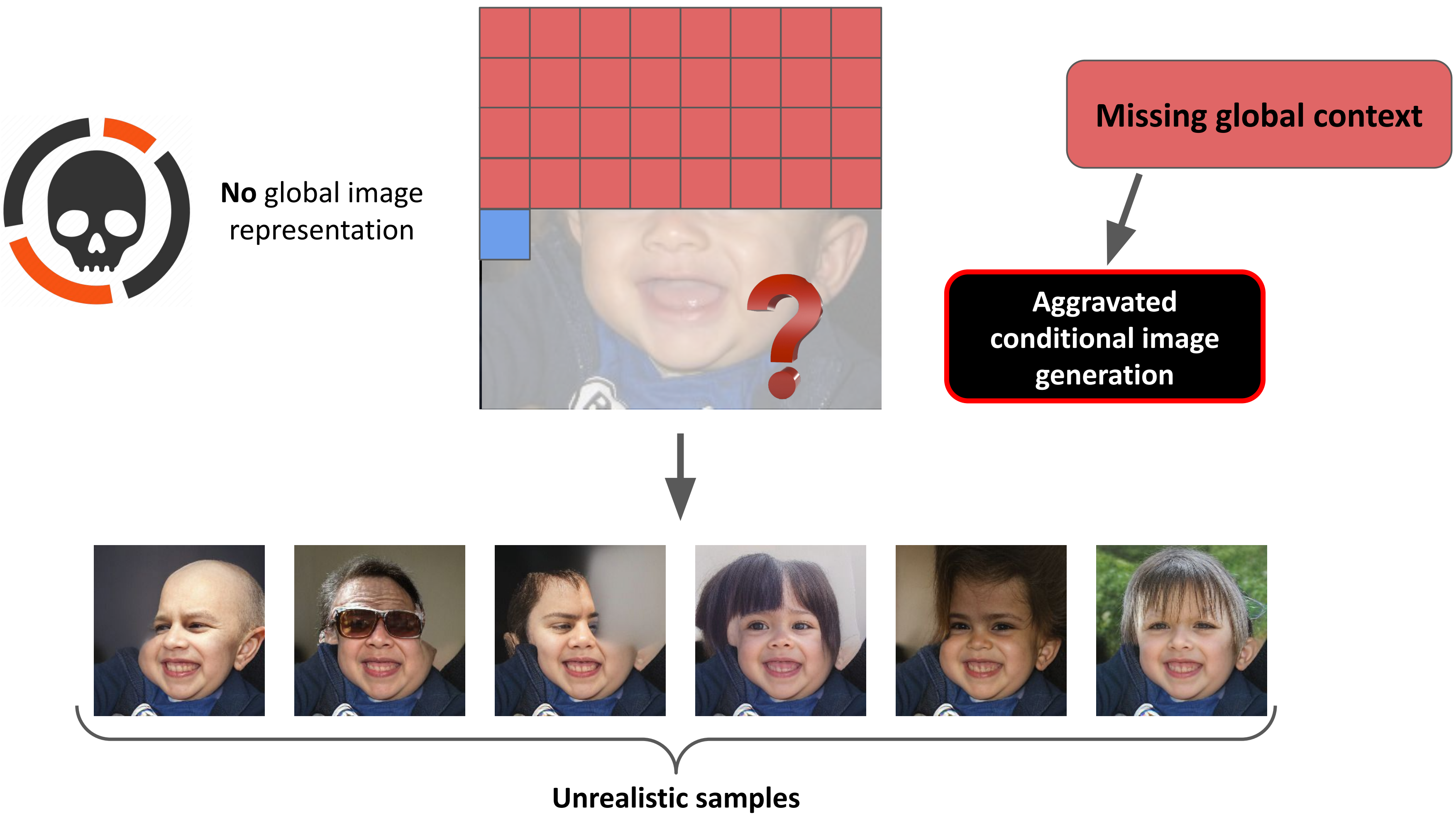
Unrealistic samples

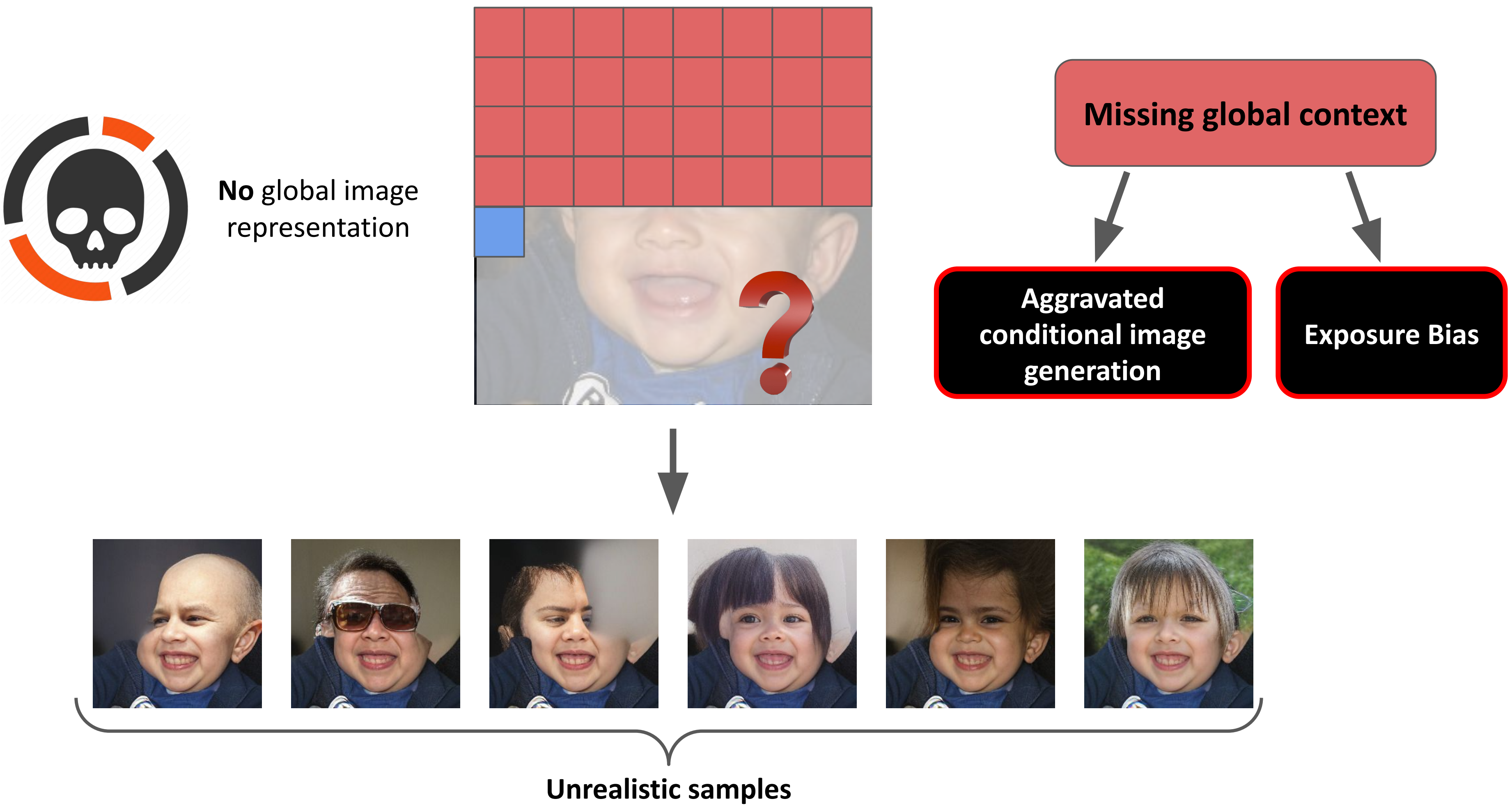
No global image representation



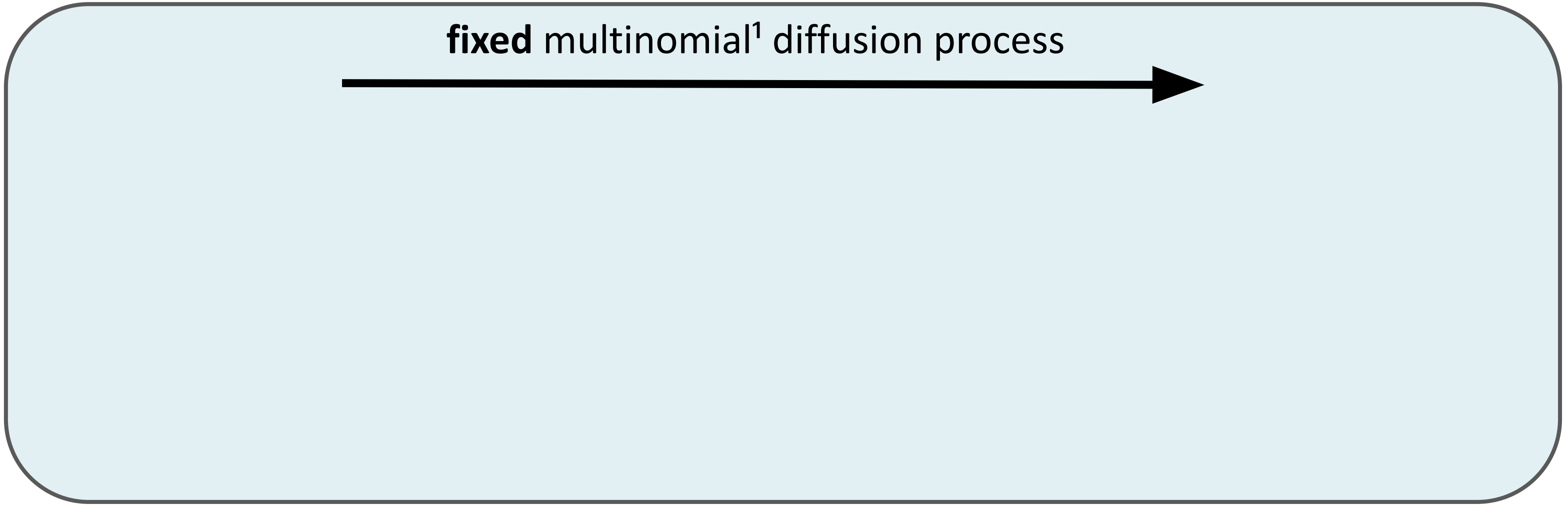
Missing global context

Unrealistic samples

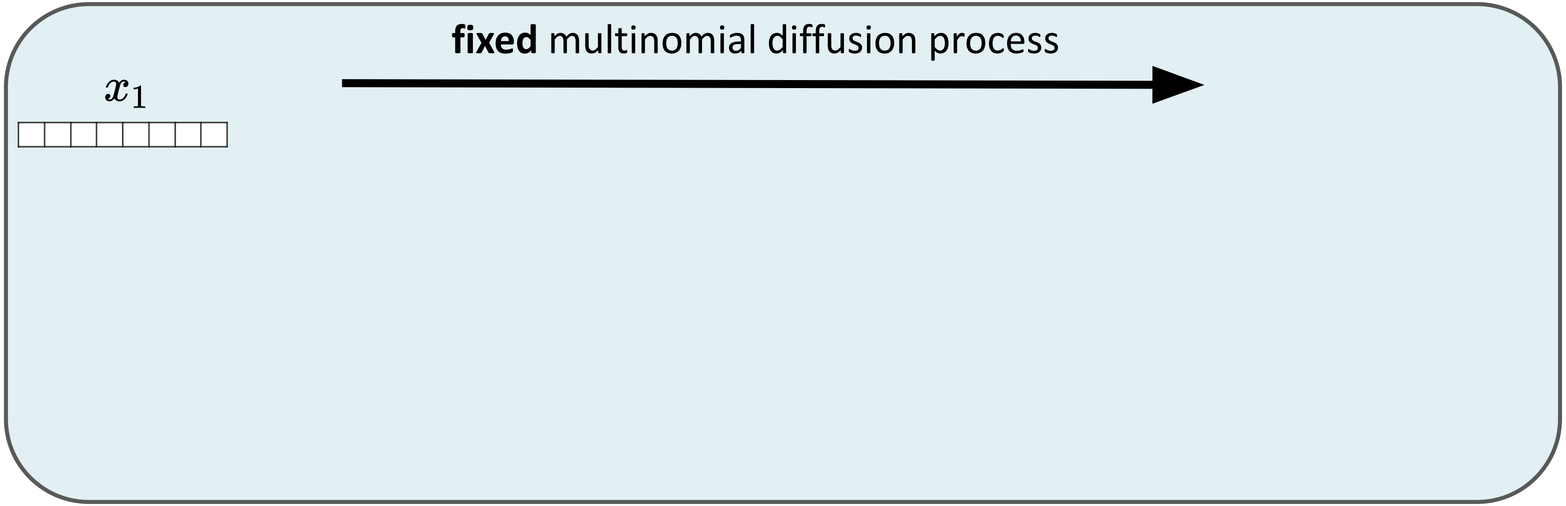

No global image representation
Missing global context
Aggravated conditional image generation
Unrealistic samples

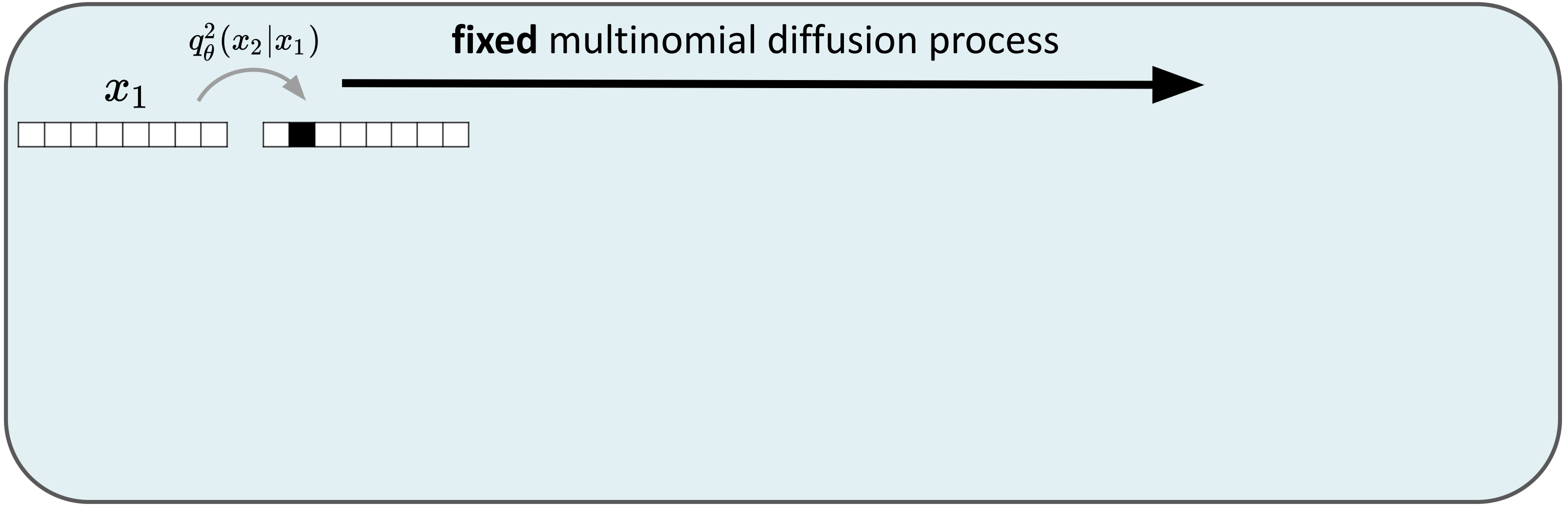

No global image representation
Missing global context
Aggravated conditional image generation
Exposure Bias
Unrealistic samples

How can we incorporate global context for autoregressive modeling?

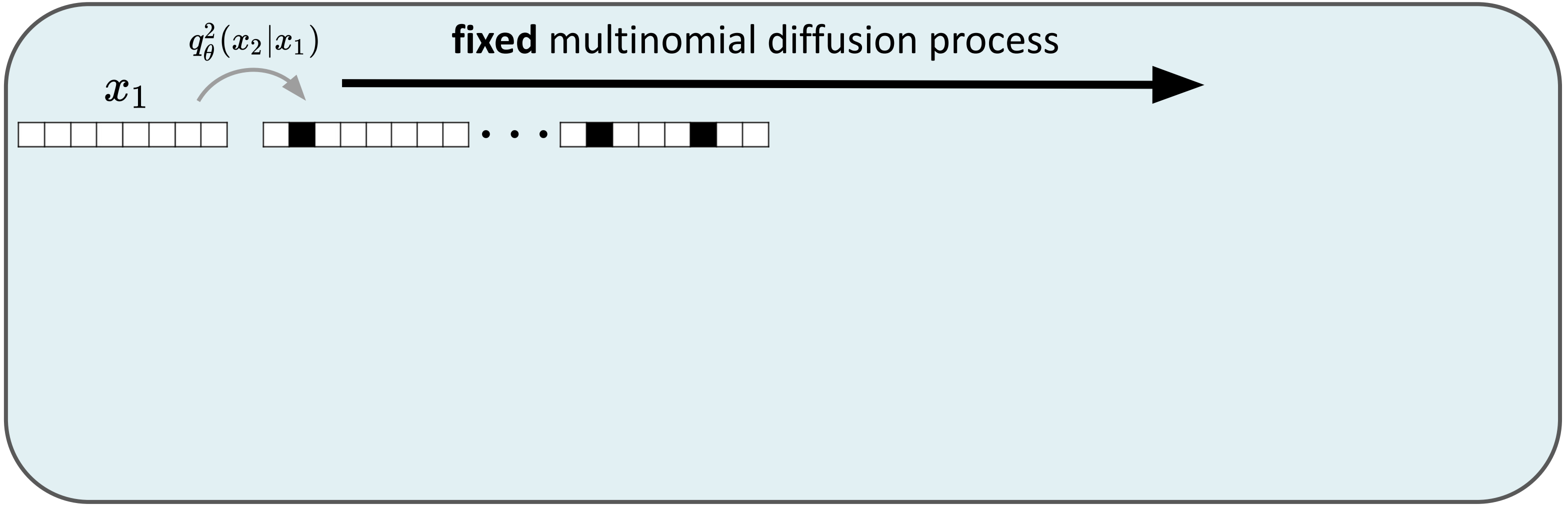


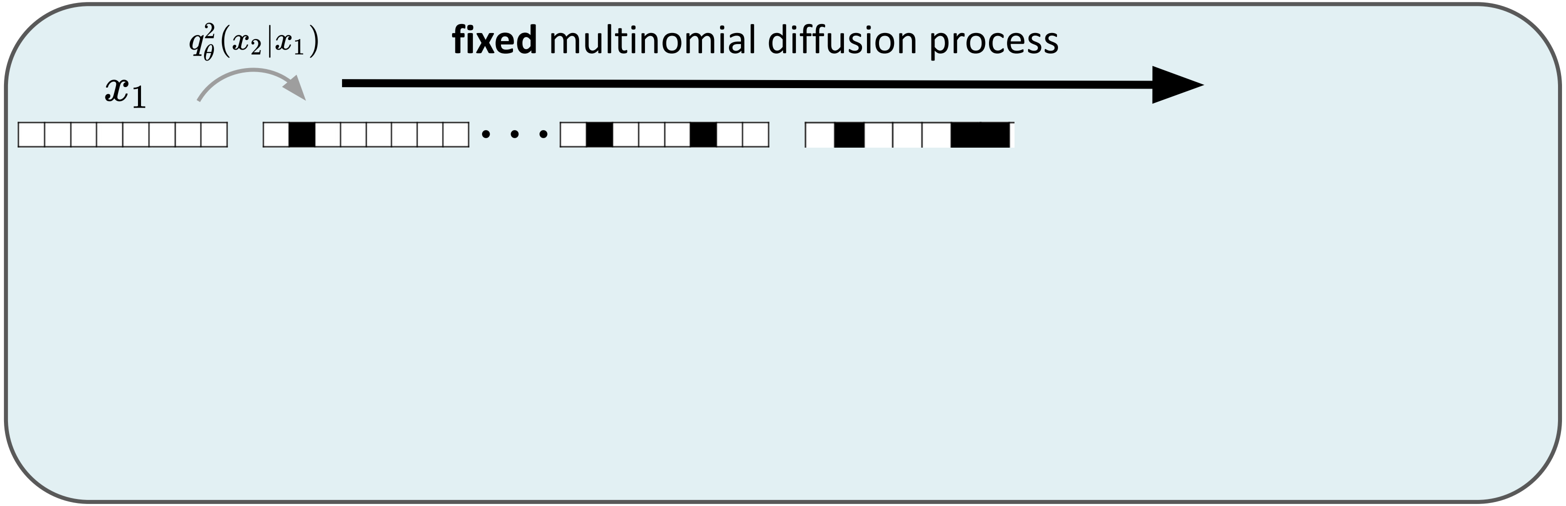
[1] see e.g.: *Argmax Flows and Multinomial Diffusion: Learning Categorical Distributions,* Nielsen et al, 2021

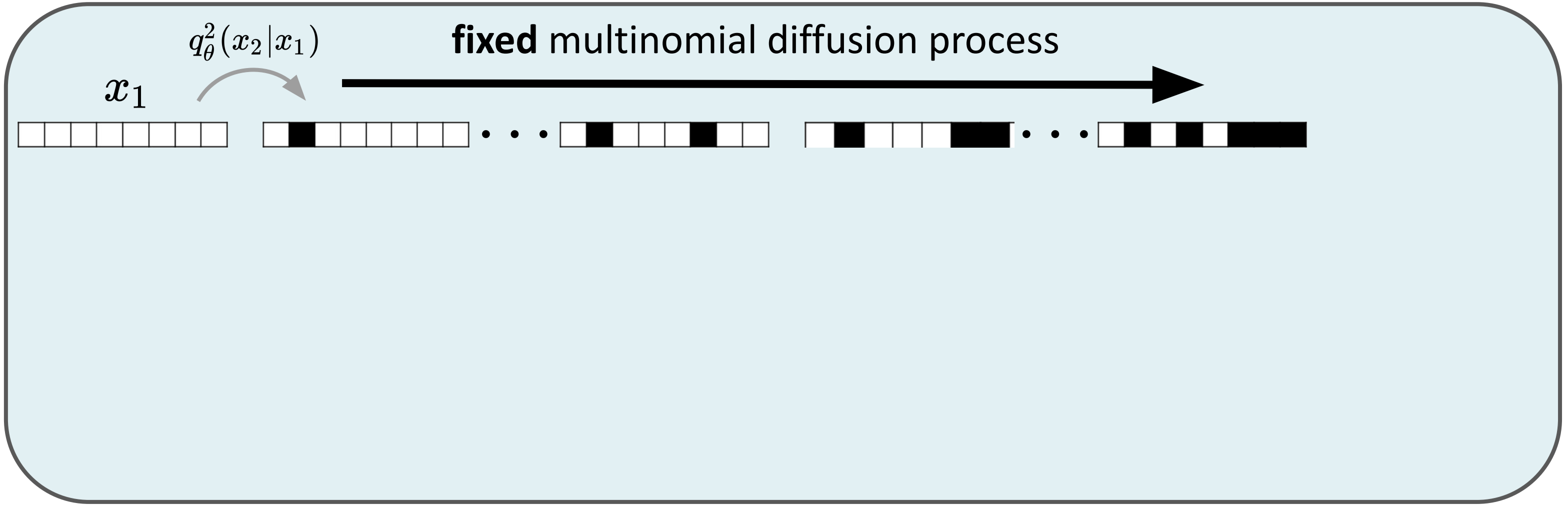

fixed multinomial diffusion process
x_1

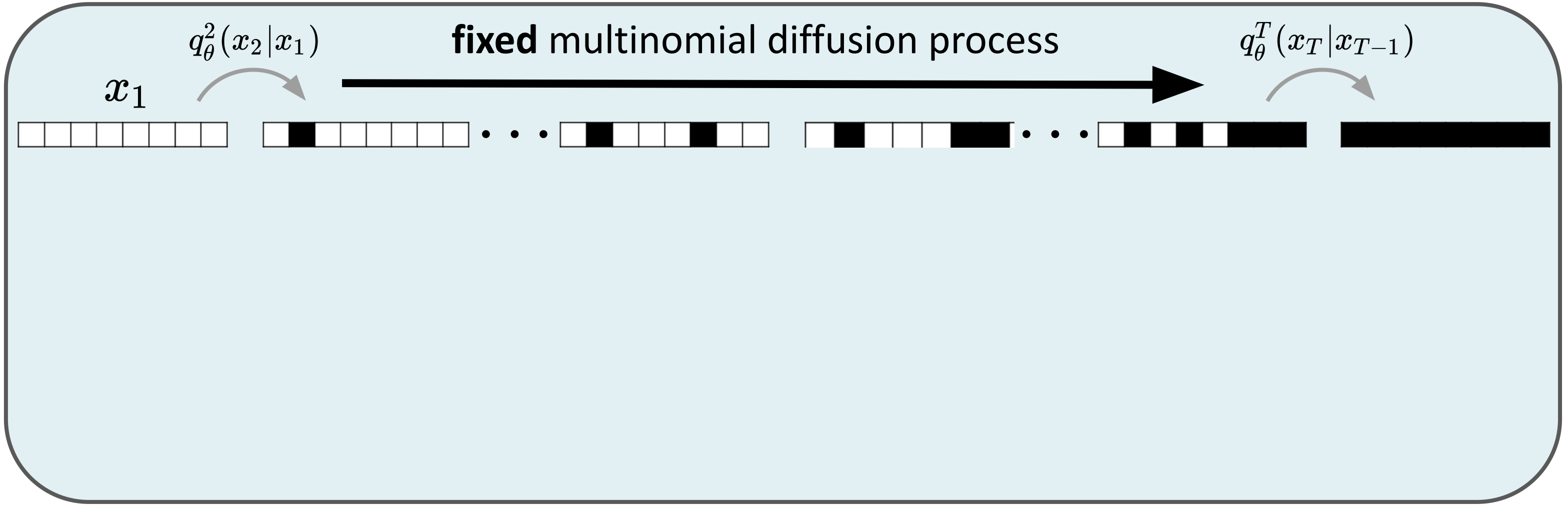


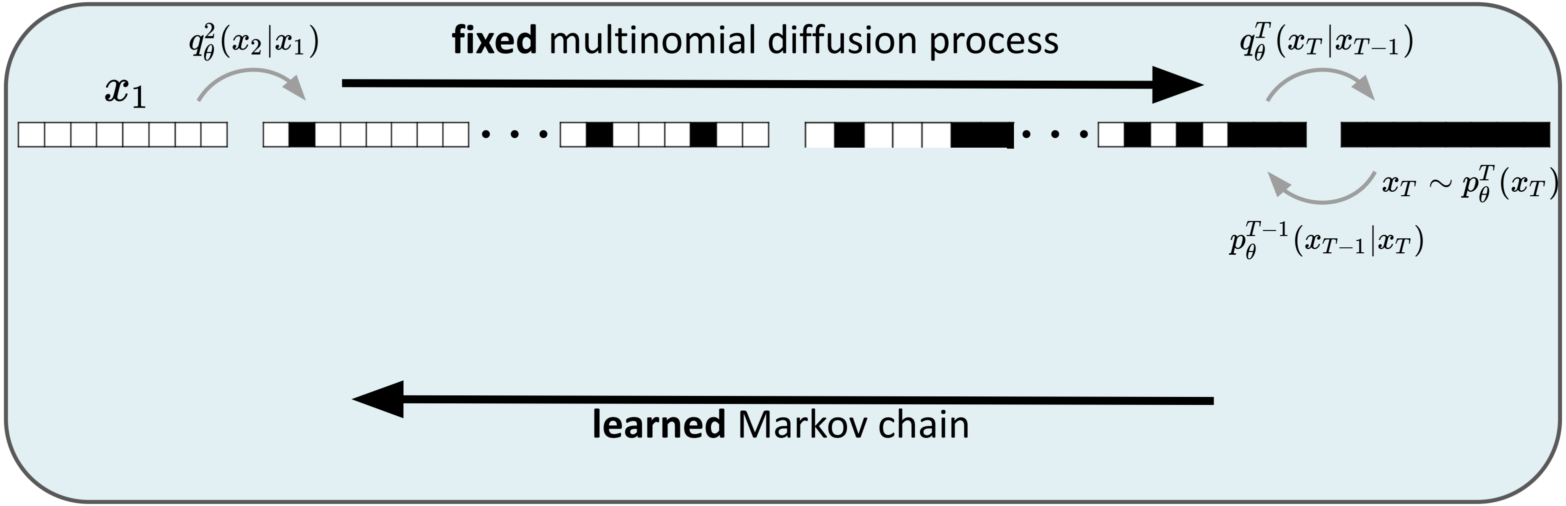
$q_\theta^2(x_2|x_1)$
fixed multinomial diffusion process
x_1

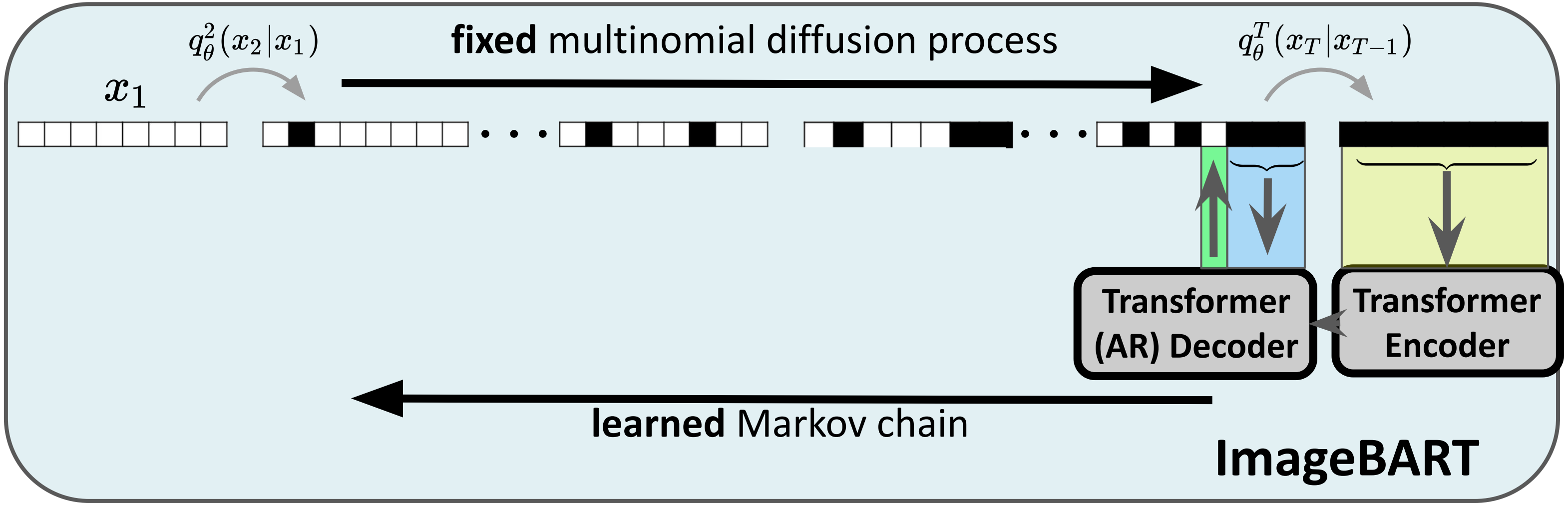

$q_\theta^2(x_2|x_1)$
fixed multinomial diffusion process
x_1

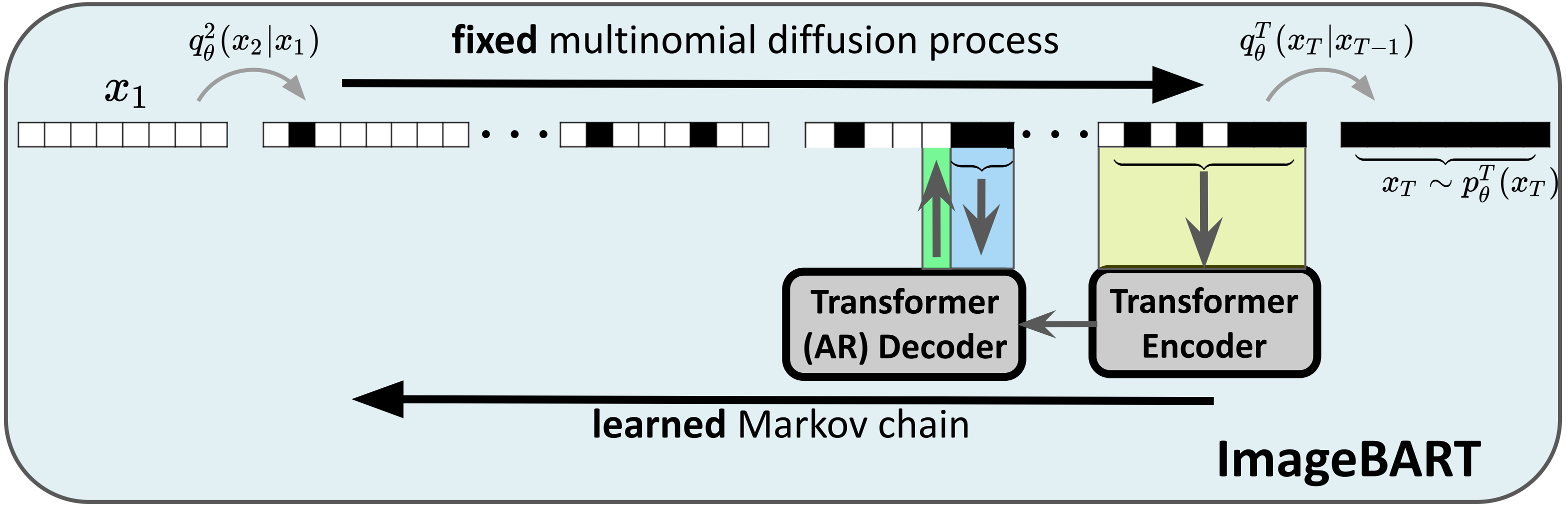

$q_\theta^2(x_2|x_1)$
fixed multinomial diffusion process
x_1


$q_\theta^2(x_2|x_1)$
fixed multinomial diffusion process
x_1

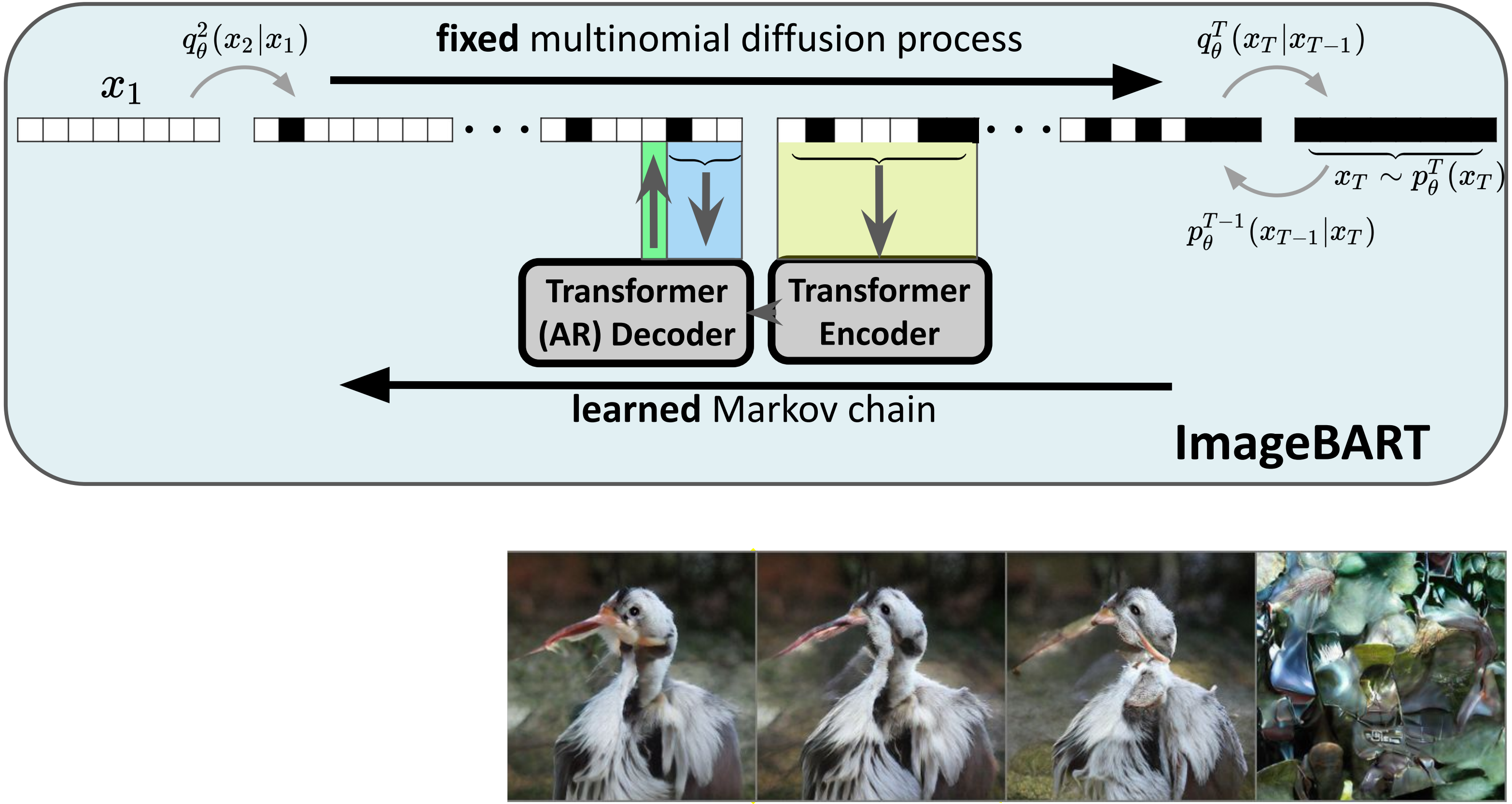

$q_\theta^2(x_2|x_1)$
fixed multinomial diffusion process
$q_\theta^T(x_T|x_{T-1})$
x_1


$q_\theta^2(x_2|x_1)$
fixed multinomial diffusion process
$q_\theta^T(x_T|x_{T-1})$
x_1
$x_T \sim p_\theta^T(x_T)$
$p_\theta^{T-1}(x_{T-1}|x_T)$
learned Markov chain

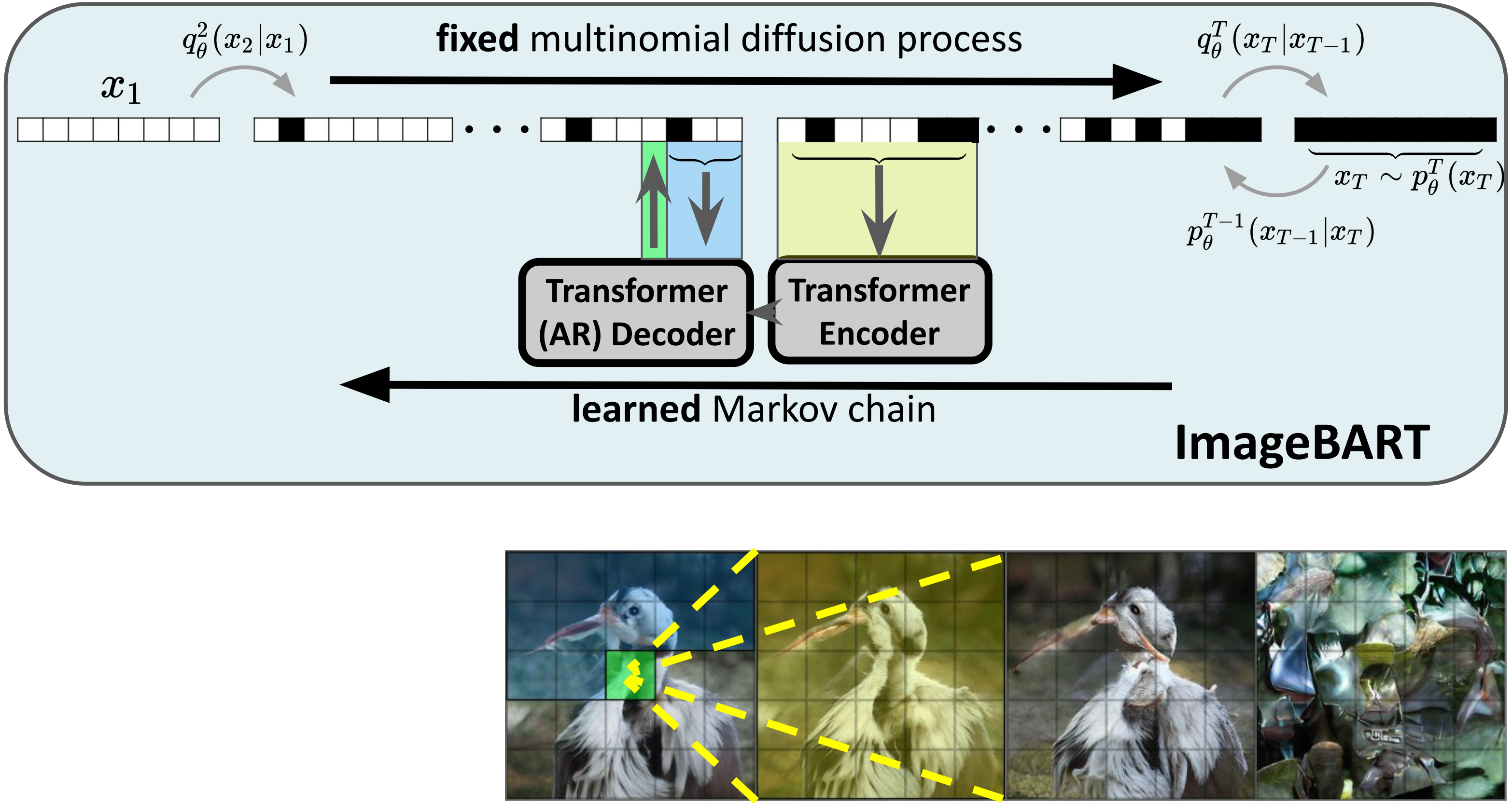

$q_\theta^2(x_2|x_1)$
fixed multinomial diffusion process
$q_\theta^T(x_T|x_{T-1})$
x_1
Transformer (AR) Decoder
Transformer Encoder
learned Markov chain
ImageBART

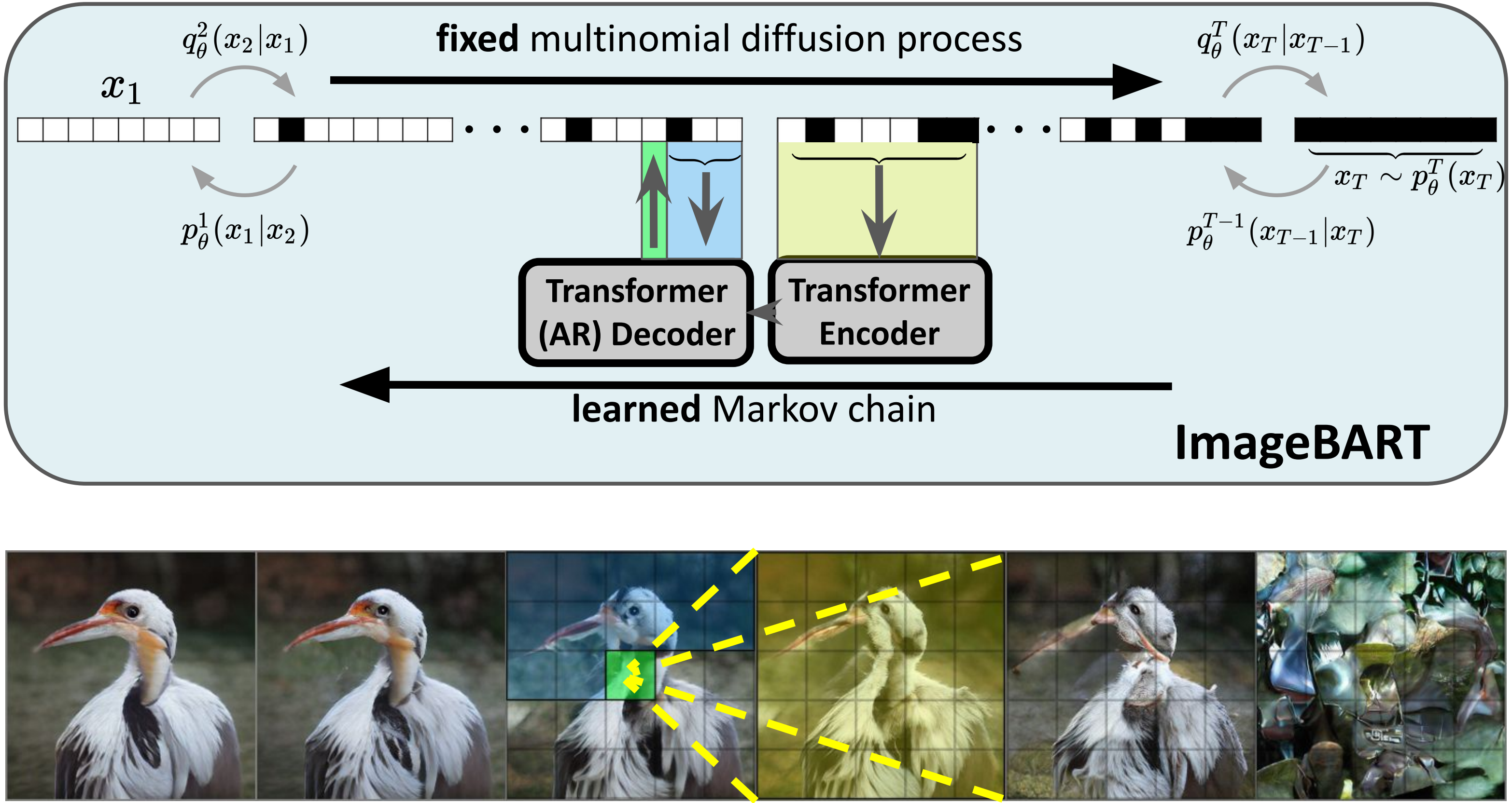


$q_\theta^2(x_2|x_1)$
fixed multinomial diffusion process
$q_\theta^T(x_T|x_{T-1})$
x_1
$x_T \sim p_\theta^T(x_T)$
Transformer (AR) Decoder
Transformer Encoder
learned Markov chain
ImageBART

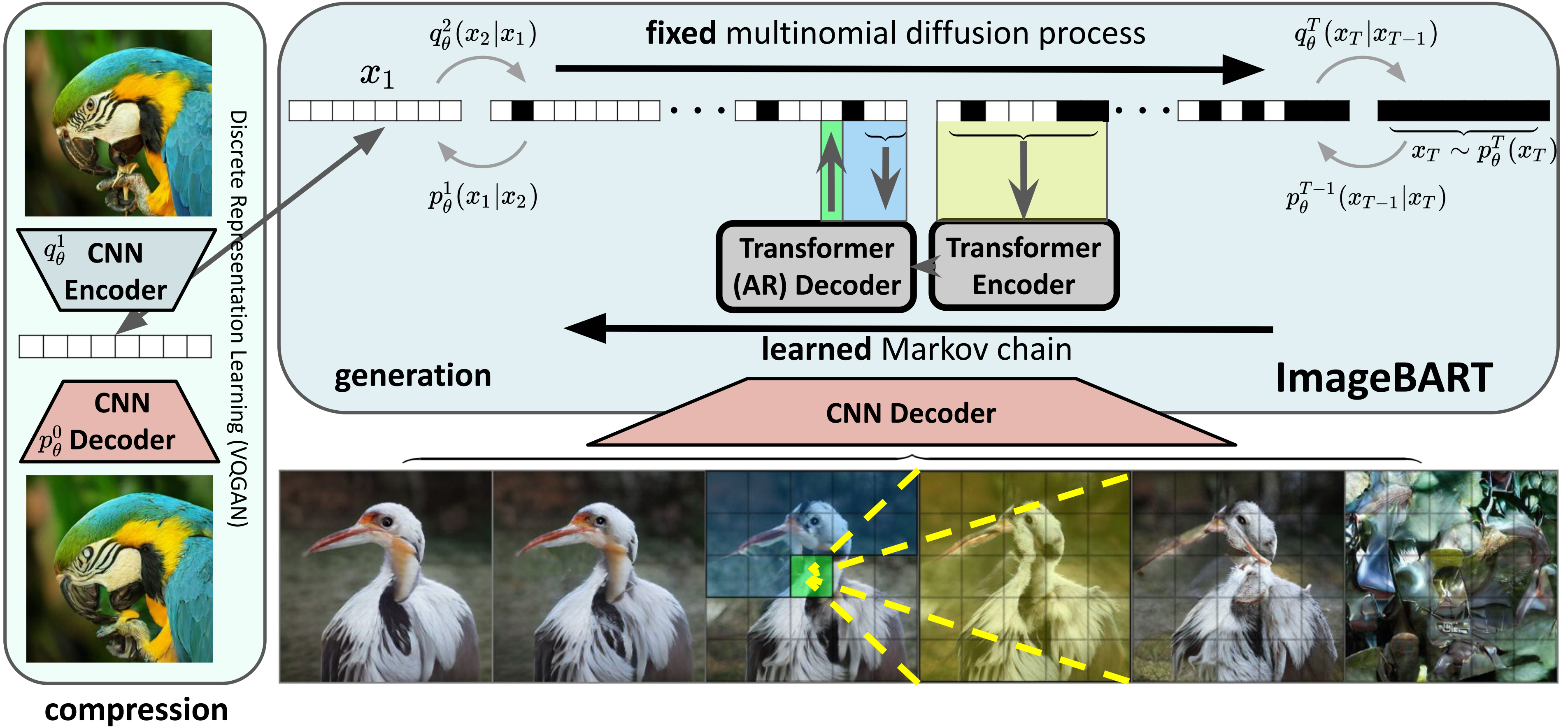


$q_\theta^2(x_2|x_1)$
fixed multinomial diffusion process
$q_\theta^T(x_T|x_{T-1})$
x_1
$x_T \sim p_\theta^T(x_T)$
$p_\theta^{T-1}(x_{T-1}|x_T)$
Transformer (AR) Decoder
Transformer Encoder
learned Markov chain
ImageBART


$q_\theta^2(x_2|x_1)$
fixed multinomial diffusion process
$q_\theta^T(x_T|x_{T-1})$
x_1
$x_T \sim p_\theta^T(x_T)$
$p_\theta^{T-1}(x_{T-1}|x_T)$
Transformer (AR) Decoder
Transformer Encoder
learned Markov chain
ImageBART



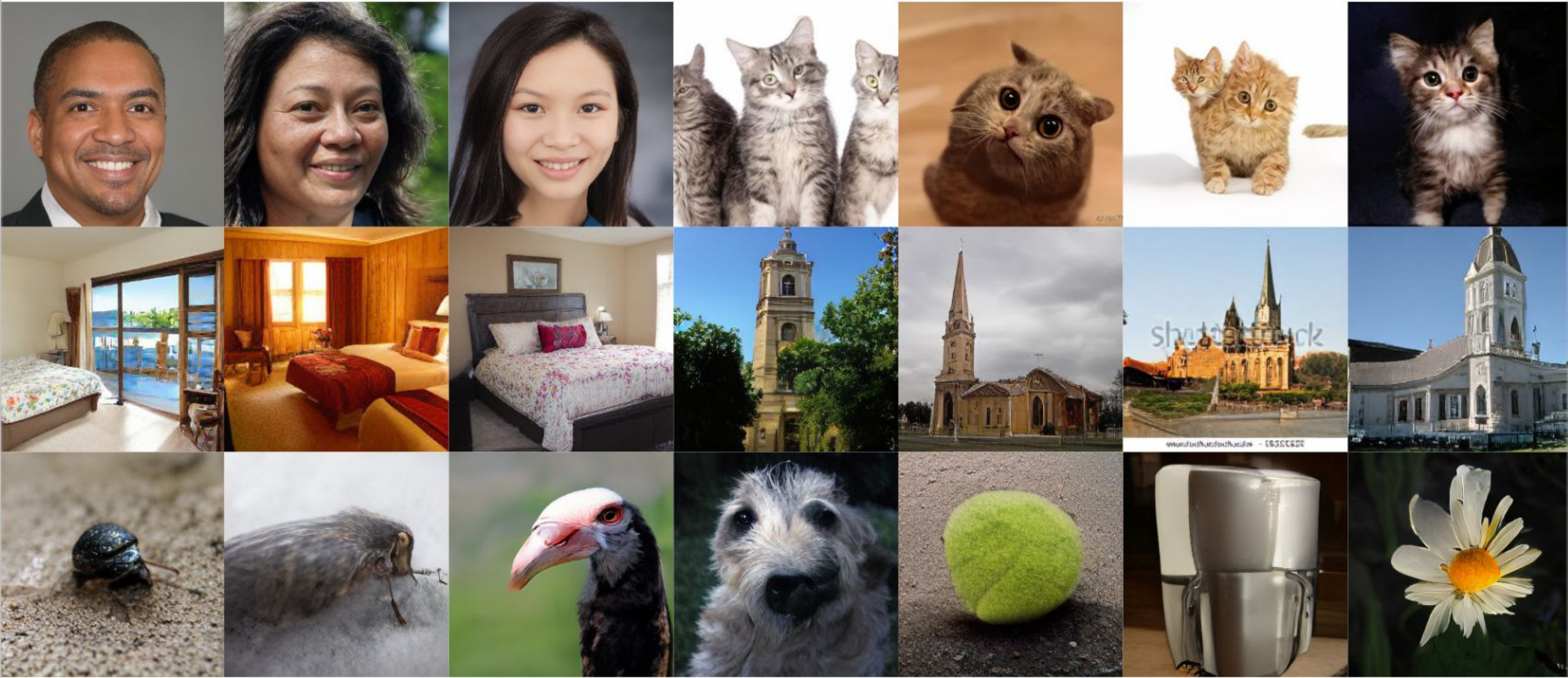
fixed multinomial diffusion process
$q_\theta^2(x_2|x_1)$
$q_\theta^T(x_T|x_{T-1})$
x_1
$p_\theta^1(x_1|x_2)$
$x_T \sim p_\theta^T(x_T)$
$p_\theta^{T-1}(x_{T-1}|x_T)$
Transformer (AR) Decoder
Transformer Encoder
learned Markov chain
ImageBART



Discrete Representation Learning (VQGAN)
q_θ^1 CNN Encoder
CNN p_θ^0 Decoder
compression
$q_\theta^2(x_2|x_1)$
fixed multinomial diffusion process
$q_\theta^T(x_T|x_{T-1})$
x_1
$p_\theta^1(x_1|x_2)$
$x_T \sim p_\theta^T(x_T)$
$p_\theta^{T-1}(x_{T-1}|x_T)$
Transformer (AR) Decoder
Transformer Encoder
learned Markov chain
generation
ImageBART
CNN Decoder

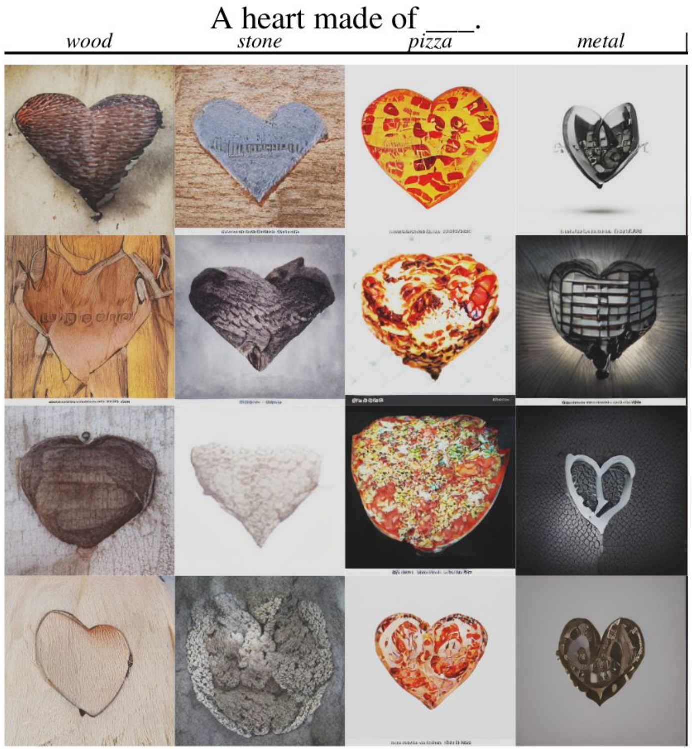
Basic Image Synthesis

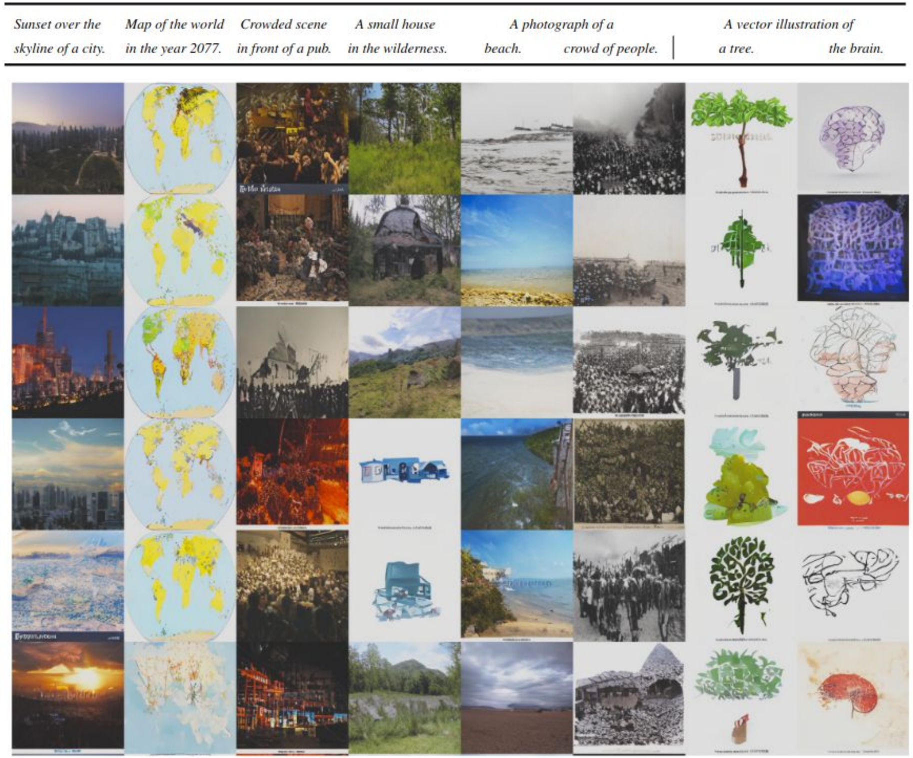
FFHQ, LSUN, ImageNet, ...



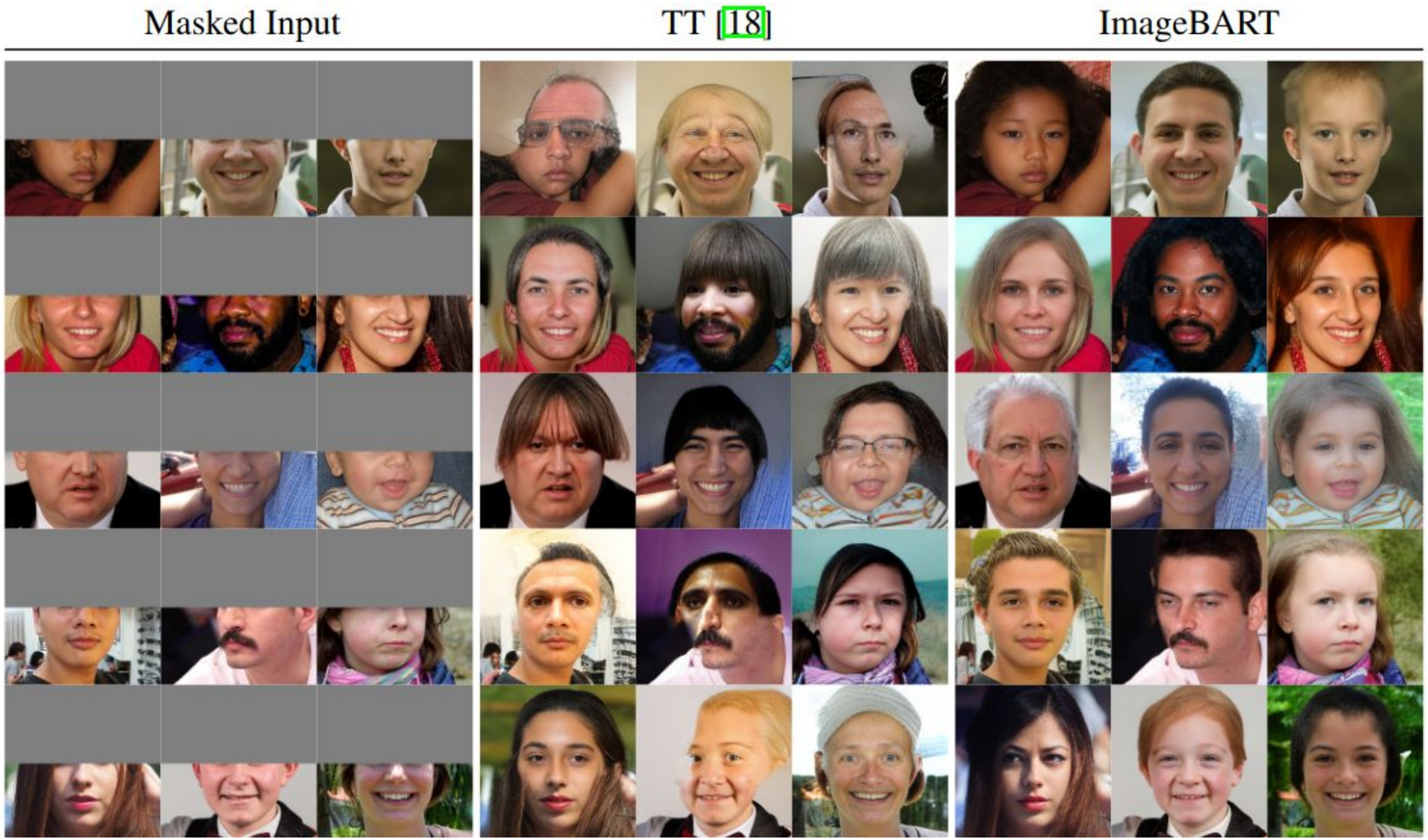
Conditional Image Synthesis

Txt2Img (Conceptual Captions),





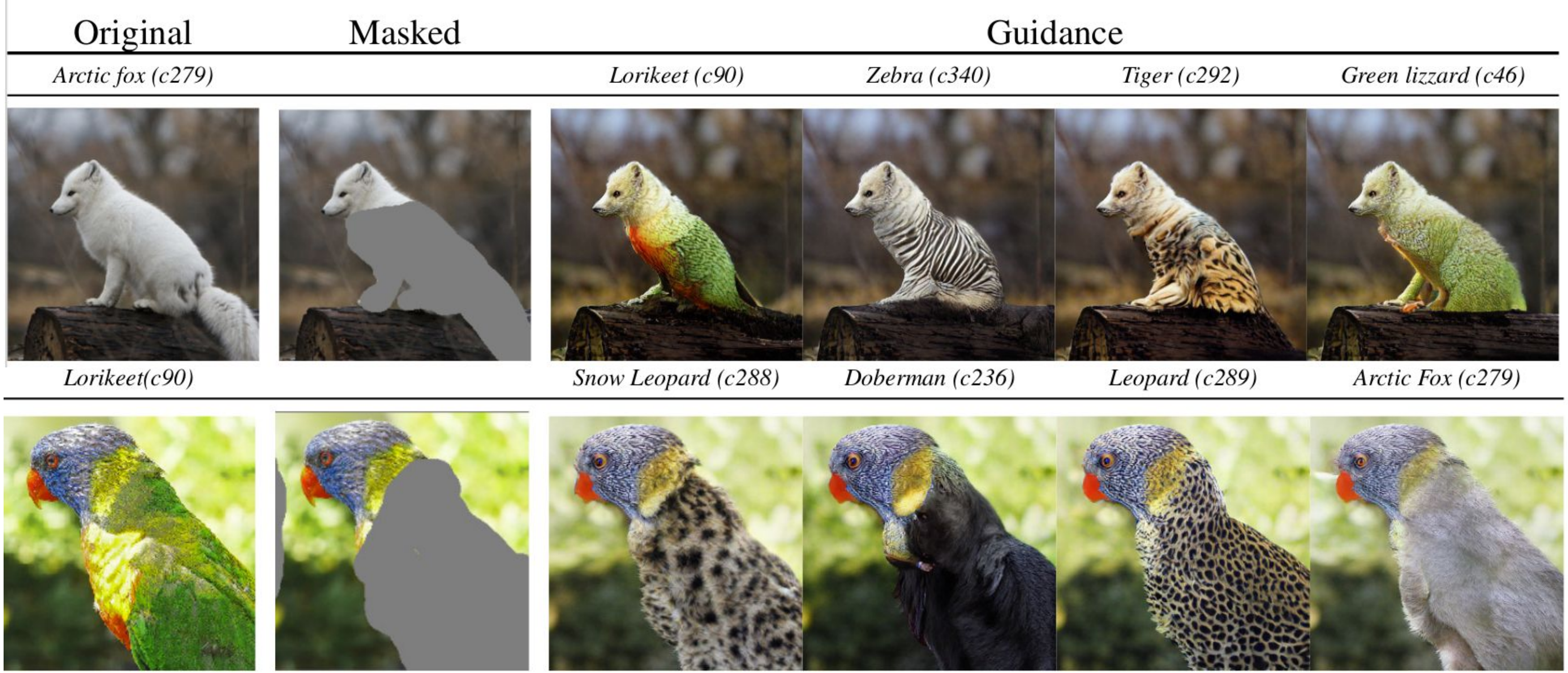
Global Context for Autoregressive Image Synthesis



Arbitrary Image Completion (a.k.a. Inpainting)

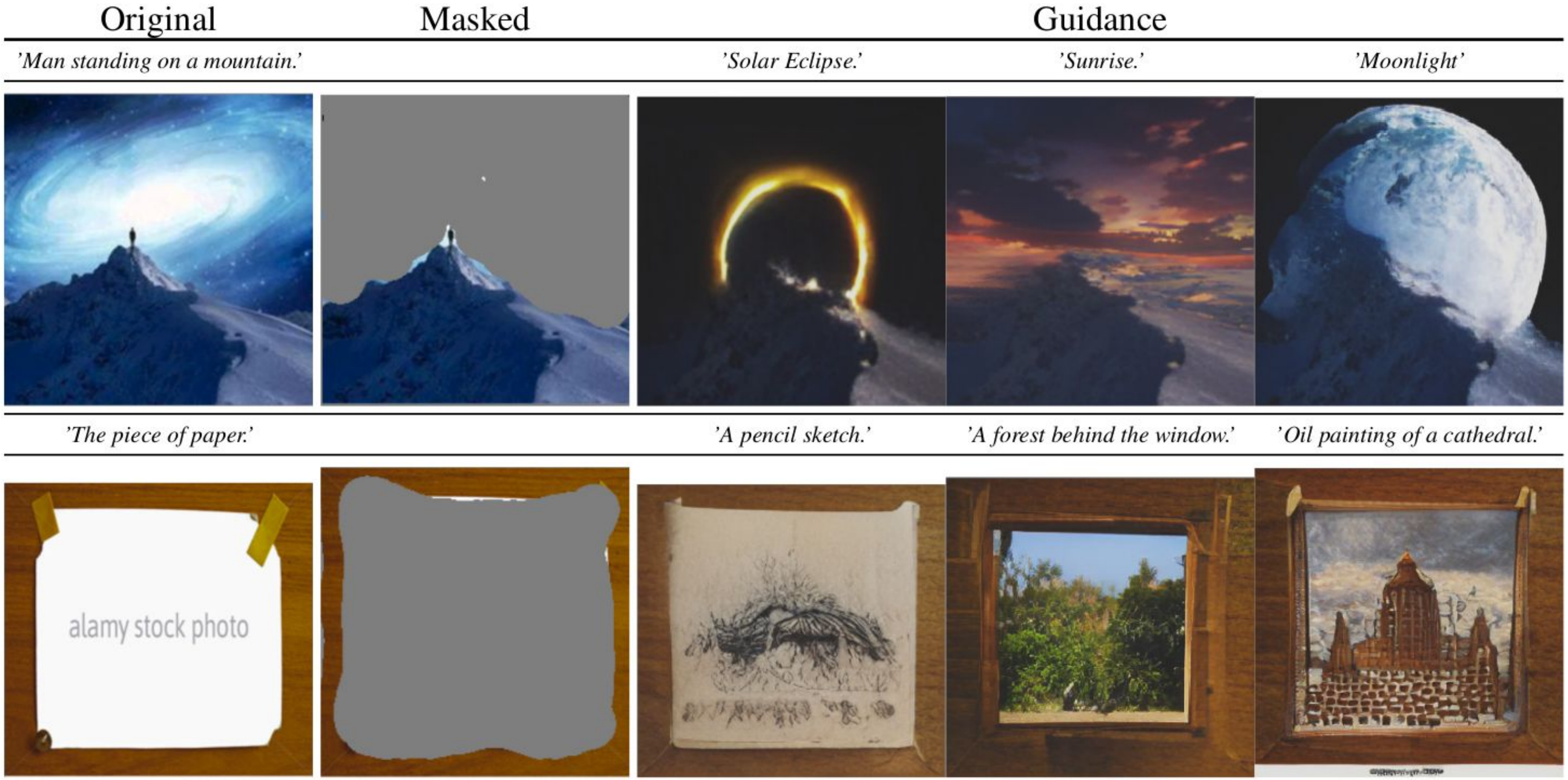
Controllable Inpainting/Modification

e.g. via class labels...



Controllable Inpainting/Modification

...or text



Do we really need more steps?

Unconditional Generation		
method	FID ↓	IS ↑
TT ($T = 2$)	12.44	3.98 ± 0.07
ImageBART ($T = 3$)	12.55	3.98 ± 0.07
ImageBART ($T = 5$)	$\boxed{10.69}$	4.27 ± 0.05
ImageBART ($T = 9$)	10.81	$\boxed{4.49 \pm 0.05}$

Upper Half Completion		
method	FID ↓	IS ↑
TT ($T = 2$)	11.80	4.48 ± 0.10
ImageBART ($T = 3$)	9.25	4.49 ± 0.13
ImageBART ($T = 5$)	6.87	4.81 ± 0.13
ImageBART ($T = 9$)	$\boxed{6.64}$	$\boxed{4.86 \pm 0.15}$

Code and pretrained models at

https://github.com/CompVis/imagebart

THE FOXCHAIN

Thanks!

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

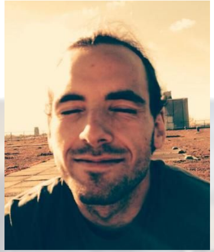
NEURAL INFORMATION PROCESSING SYSTEMS

IWR Interdisciplinary Center for Scientific Computing

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

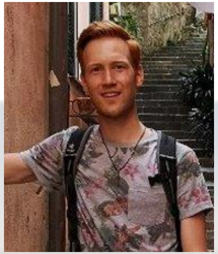
ImageBART: Bidirectional Context with Multinomial Diffusion for Autoregressive Image Synthesis

Patrick Esser*,



Robin Rombach*,

Andreas Blattmann*,



Björn Ommer

*equal contribution

Arbitrary Image Completion (a.k.a. Inpainting)