Binary Latent Diffusion

Ze Wang[1], Jiang Wang[2], Zicheng Liu[2], and Qiang Qiu[1]
[1]Purdue University [2]Microsoft Corporation

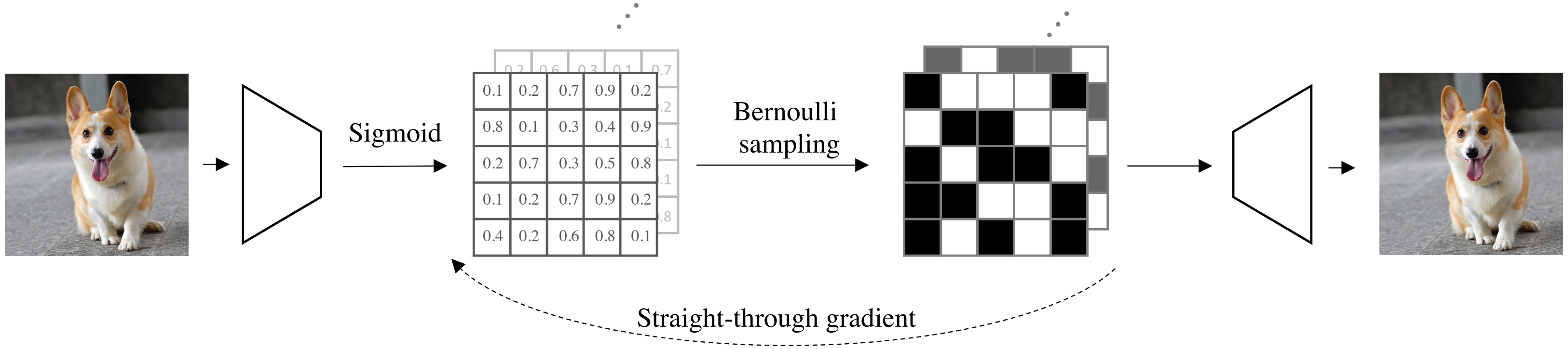
THU-PM-188

Compact yet expressive representations that allow for efficient diffusion (denoising) processes with high image quality

- Compact: low dimension, compact search space
- Expressive: both high quality and coverage

Binary Latent Diffusion

- Binary Auto-encoder
- Multivariate Bernoulli Diffusion
- Unconditional Sampling
- Class-conditional Sampling
- Text-to-Image Generation



Auto-encoder with binary latent space

Binary Latent Diffusion

- Binary Auto-encoder
- Multivariate Bernoulli Diffusion
- Unconditional Sampling
- Class-conditional Sampling
- Text-to-Image Generation

True data

1.0	0.0	0.0
0.0	1.0	0.0
0.0	1.0	0.0

$\mathbf{z}^{\mathbf{0}}$

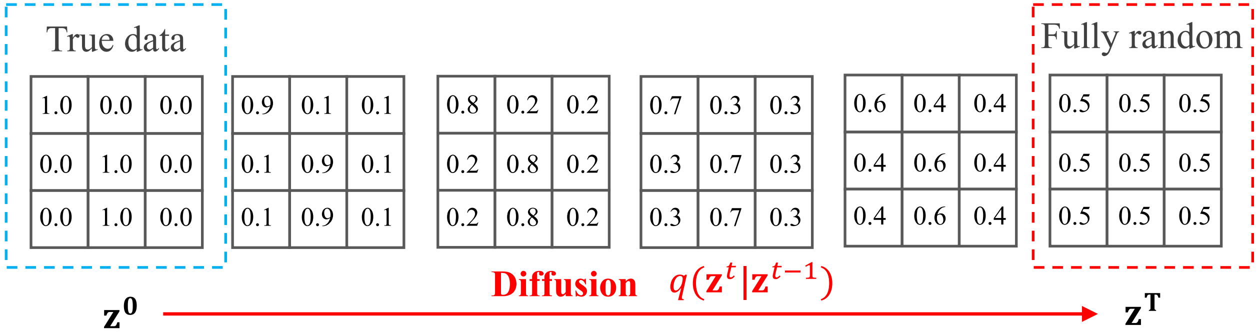
Fully random

0.5	0.5	0.5
0.5	0.5	0.5
0.5	0.5	0.5

$\mathbf{z}^{\mathbf{T}}$

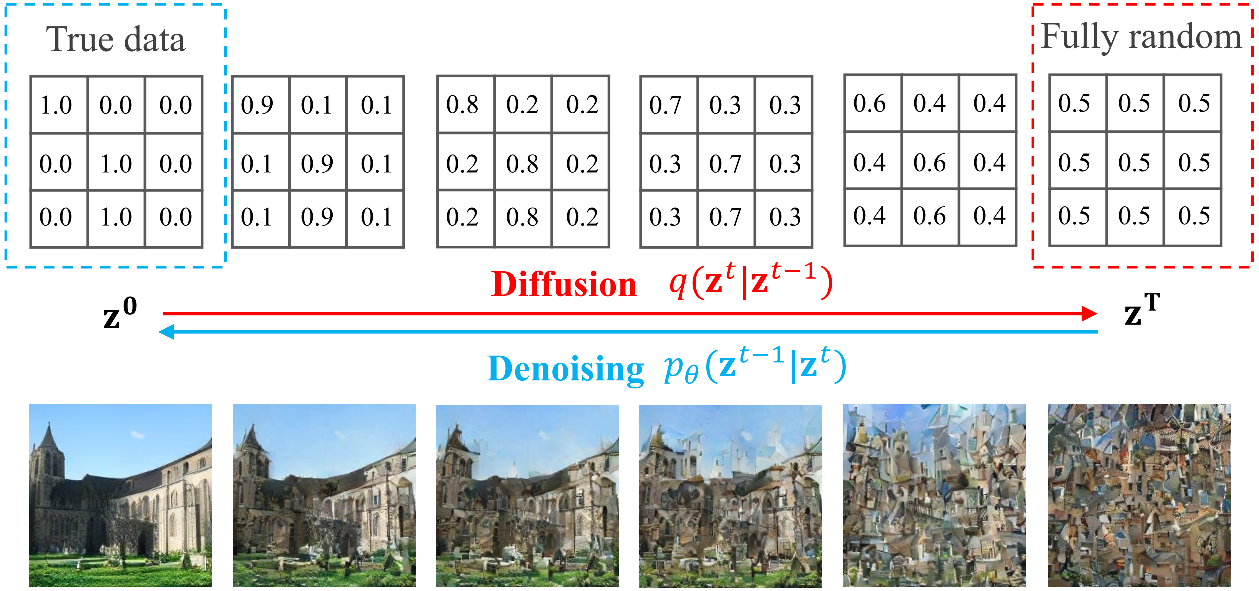
Binary Latent Diffusion

- Binary Auto-encoder
- Multivariate Bernoulli Diffusion
- Unconditional Sampling
- Class-conditional Sampling
- Text-to-Image Generation



Binary Latent Diffusion

- Binary Auto-encoder
- Multivariate Bernoulli Diffusion
- Unconditional Sampling
- Class-conditional Sampling
- Text-to-Image Generation



Reparametrizing the prediction targets

$$p_\theta(\mathbf{z}^{t-1}|\mathbf{z}^t) \longrightarrow p_\theta(\mathbf{z}^0|\mathbf{z}^t) \longrightarrow p_\theta(\mathbf{z}^0 \oplus \mathbf{z}^t|\mathbf{z}^t)$$

$$q(\mathbf{z}^t|\mathbf{z}^{t-1}) = \mathcal{B}(\mathbf{z}^t; \mathbf{z}^{t-1}(1-\beta^t) + 0.5\beta^t)$$

Shared targets across all steps.
Easy step-skipping for sampling.
Faster training convergence and better empirical results.

$$p_\theta(\mathbf{z}^{t-1}|\mathbf{z}^t) = q(\mathbf{z}^{t-1}|\mathbf{z}^t, \mathbf{z}^0 = \mathbf{0})p_\theta(\mathbf{z}^0 = \mathbf{0}|\mathbf{z}^t) + q(\mathbf{z}^{t-1}|\mathbf{z}^t, \mathbf{z}^0 = \mathbf{1})p_\theta(\mathbf{z}^0 = \mathbf{1}|\mathbf{z}^t)$$

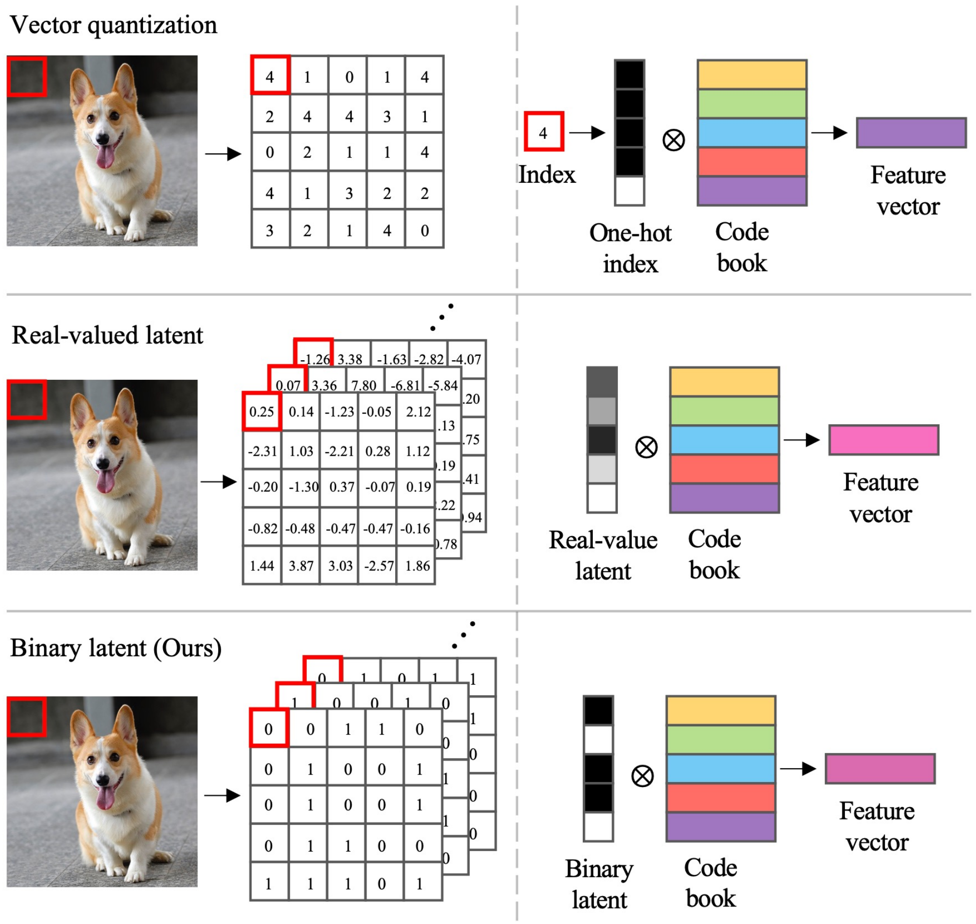
$\oplus$ element-wise logic XOR
Sampler now predicts the flipping of each element, which can be considered as the residual.

Enables classifier-free guidance for discrete distribution.

Binary Latent Diffusion

- Binary Auto-encoder
- Multivariate Bernoulli Diffusion
- Unconditional Sampling
- Class-conditional Sampling
- Text-to-Image Generation

Comparing to VQ and LDM from a dictionary perspective



One-hot selection of dictionary atoms

Large codebook, small patch size, to cover sufficiently diverse patterns

Linear combination of dictionary atoms

Binary combination of dictionary atoms

Binary Latent Diffusion

- Binary Auto-encoder
- Multivariate Bernoulli Diffusion
- Unconditional Sampling
  - Temperature
  - Results
  - Speed
  - High resolution
  - Conditional inpainting
  - NN
- Class-conditional Sampling
- Text-to-Image Generation

$$f_\theta(z^t, t) = \sigma(T_\theta(z^t, t)/\tau)$$

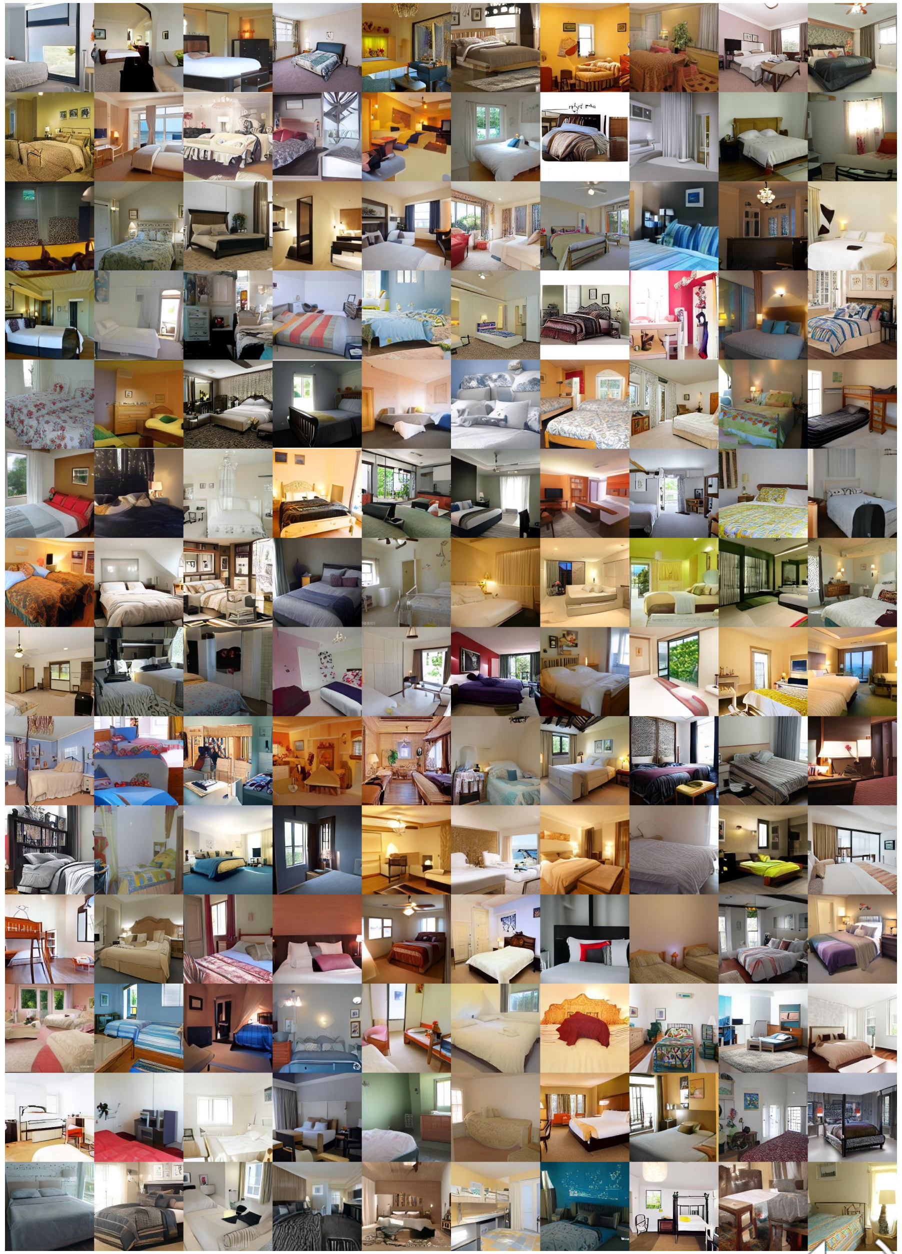
f_θ Sampling function
σ Sigmoid function
T_θ Transformer parametrizing the sampling function
τ Sampling temperature (effective only during sampling), decides diversity

Temperature low → **Temperature high**

Binary Latent Diffusion

- Binary Auto-encoder
- Multivariate Bernoulli Diffusion
- Unconditional Sampling
  - Temperature
  - Results
  - Speed
  - High resolution
  - Conditional inpainting
  - NN
- Class-conditional Sampling
- Text-to-Image Generation

LSUN Churches

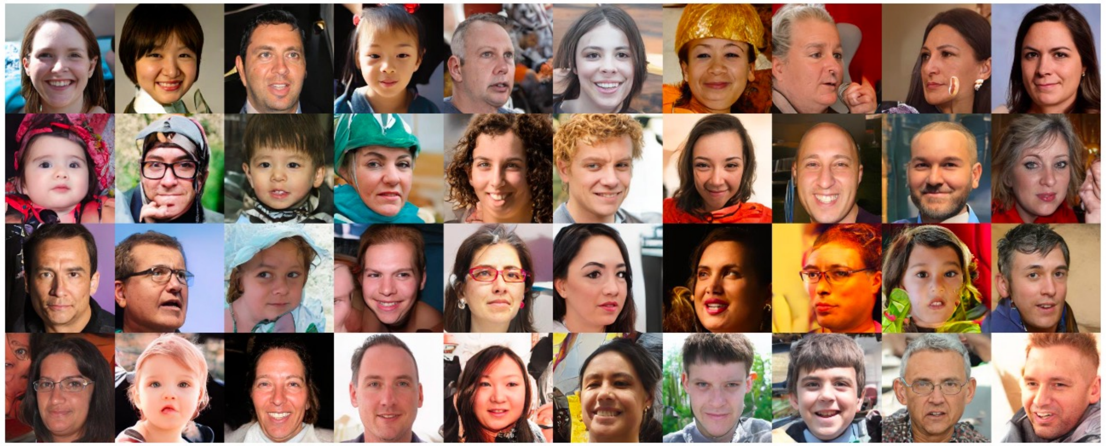


LSUN Bedrooms

Higher speed, comparable results

Methods	StyleGAN-2	Absorbing	LDM
s/sample	0.04	3.40	15.68

Methods	DDPM	Ours 64s	Ours 16s
s/sample	63.85	0.82	0.20



(a) Absorbing diffusion.

(b) Latent diffusion.

(c) Ours.

High-resolution image generation in one shot

1024x1024 images generated with 32x32 latent. 32x downsampling ratio.

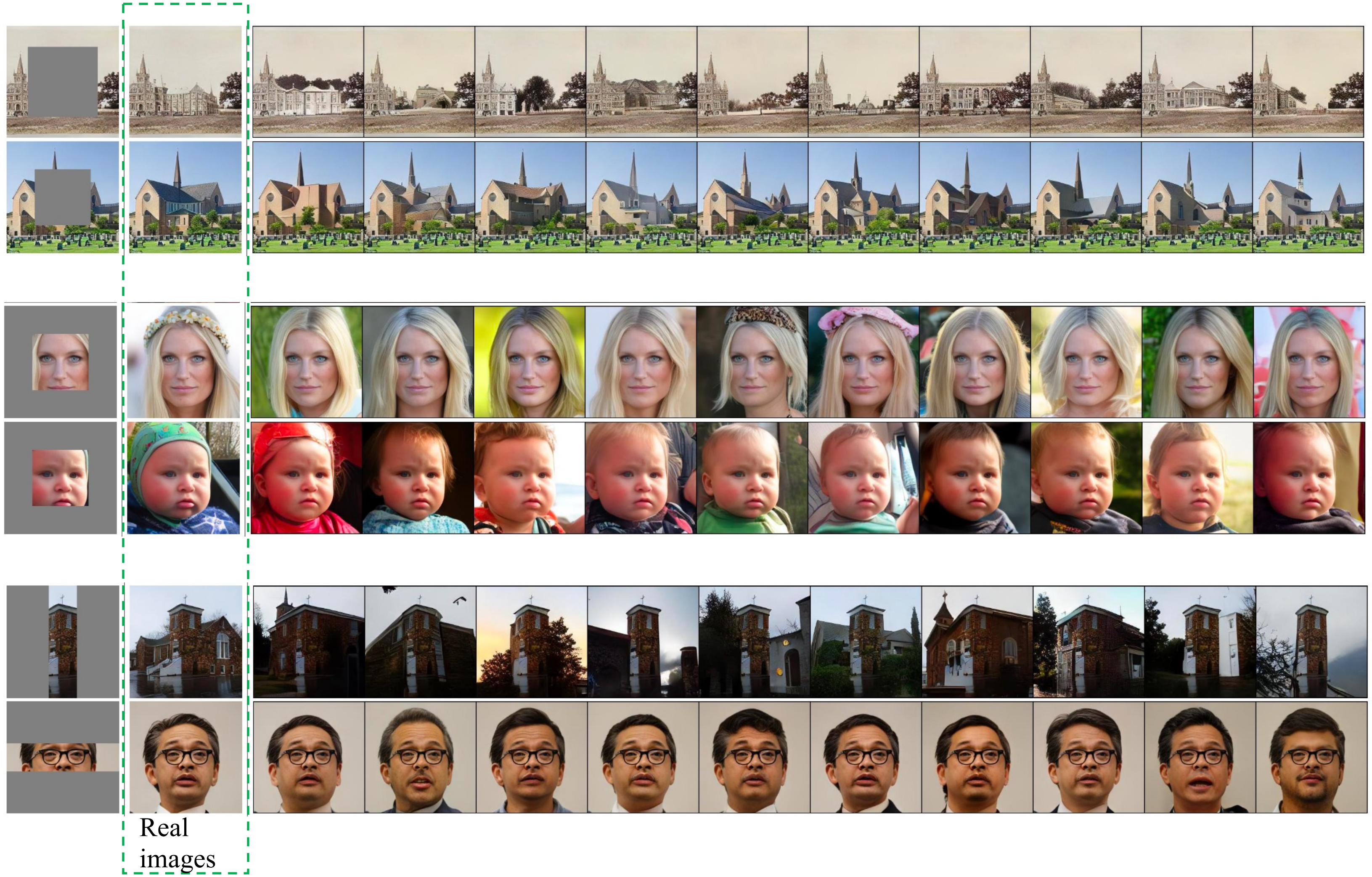
FFHQ 1024 x 1024

CelebA-HQ 1024 x 1024

- Binary Auto-encoder
- Multivariate Bernoulli Diffusion
- Unconditional Sampling
  - Temperature
  - Results
  - Speed
  - High resolution
  - Conditional inpainting
  - NN
- Class-conditional Sampling
- Text-to-Image Generation

Conditional sampling after unconditional training

with different scales and patterns of masking.



Top-10 Nearest Neighbors in the Training Datasets

The models are not overfitting to the training data.

Generated Nearest neighbors

Binary Latent Diffusion

- Binary Auto-encoder
- Multivariate Bernoulli Diffusion
- Unconditional Sampling
- Class-conditional Sampling
- Text-to-Image Generation

Standard sampling:

$$f_\theta(z^t, t, c) = \sigma(T_\theta(z^t, t, c)/\tau)$$

Sampling with Classifier-free guidance:

$$f_\theta(z^t, t, c) = \sigma\big((1+\omega)T_\theta(z^t, t, c) - \omega T_\theta(z^t, t)\big)$$

Higher ω, higher image quality but lower diversity

$\omega = 0$

$\omega = 2.5$

$\omega = 10.0$

Thank you!