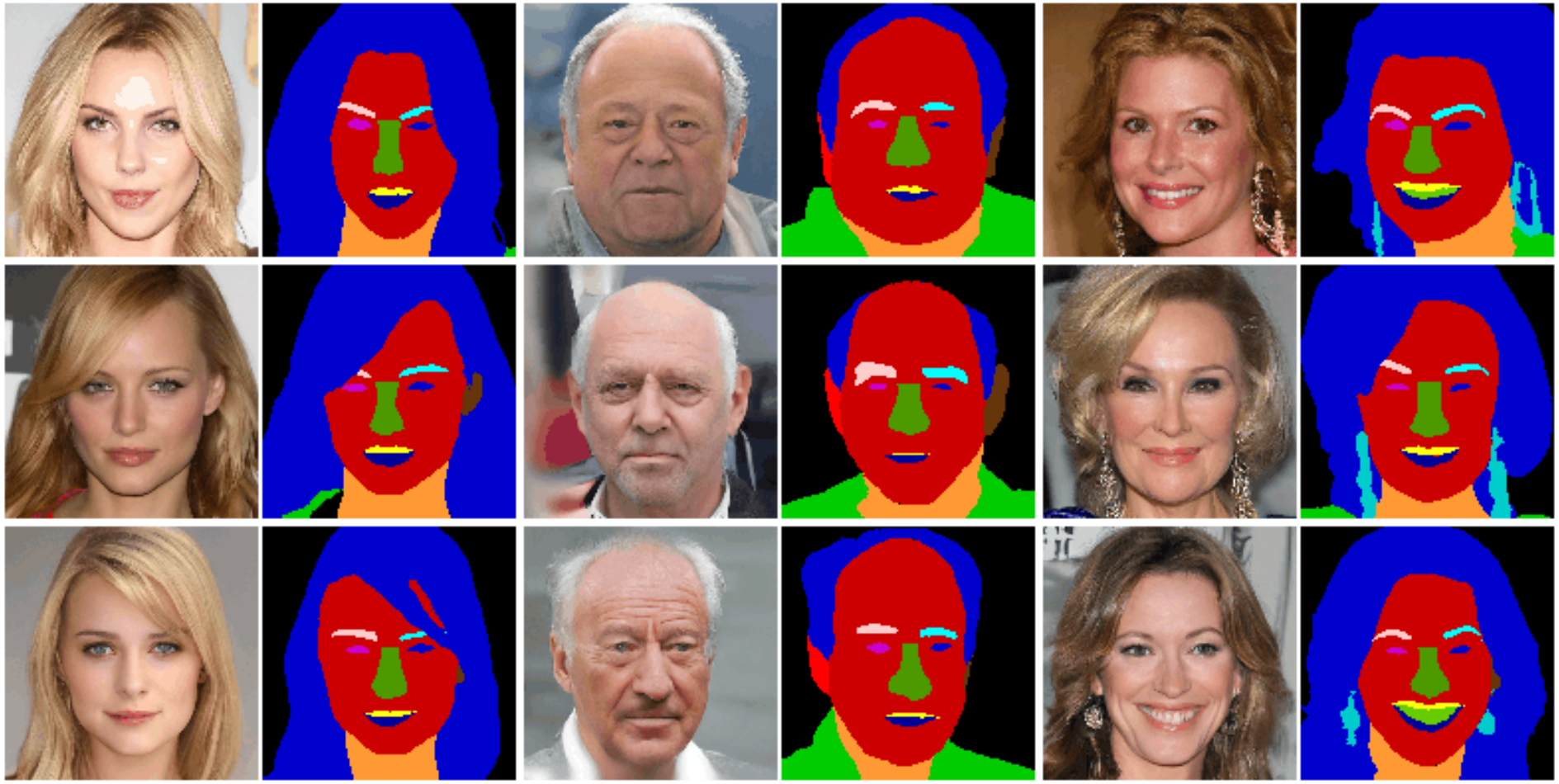
Various Types of Diffusion Models

Presented by Minho Park

KAIST

Final objective: Gaussian-categorical Diffusion Process



"The woman is wearing lipstick. She has blond hair, pointy nose, and oval face."

"This man has bags under eyes, receding hairline, and big nose."

"The woman has rosy cheeks. She is smiling. She wears earrings, and lipstick."

https://pmh9960.github.io/research/GCDP/

Table of Contents

Gaussian Diffusion Process

Denoising Diffusion Probabilistic Models

Jonathan Ho, Ajay Jain, and Pieter Abbeel

UC Berkeley

NeurIPS 2020

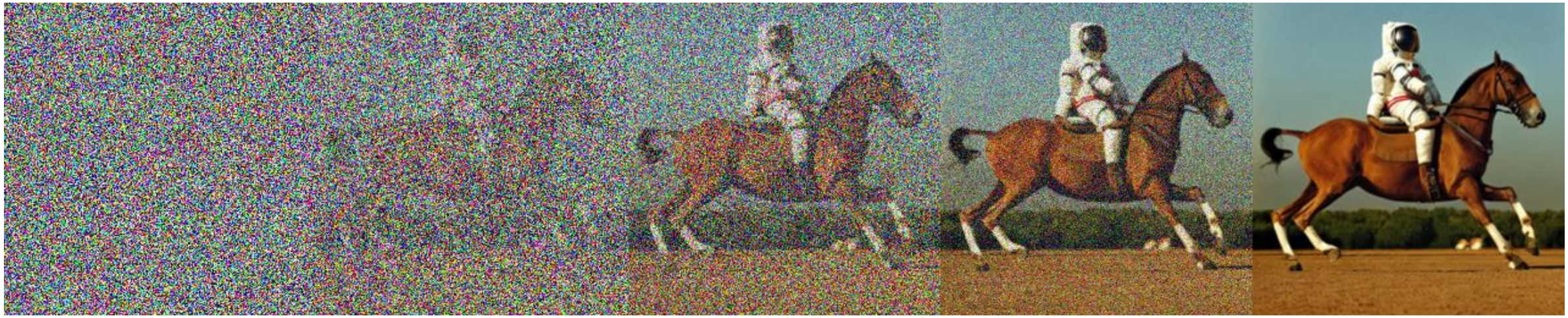
Presented by Minho Park

Overview

- Diffusion models is a generative model that synthesizes an image by **gradually removing noise from a random Gaussian noise which has the same size with the image to generate**.

Random Noise

Generated Image



Denoising Model

Denoising Model

• • •

Denoising Model

Notation

- x_0: Clean image with no noise added
- x_T: Gaussian noise
- $x_1, x_2, \ldots, x_{T-1}$: Noisy images
- $q(x_{0:T})$: Joint distribution of $x_0, x_1, \ldots, x_T$, i.e., $q(x_0, x_1, \ldots, x_T)$.
- **Forward process (diffusion process):** $q(x_t|x_{t-1})$
- **Reverse process (denoising process):** $p_\theta(x_{t-1}|x_t)$

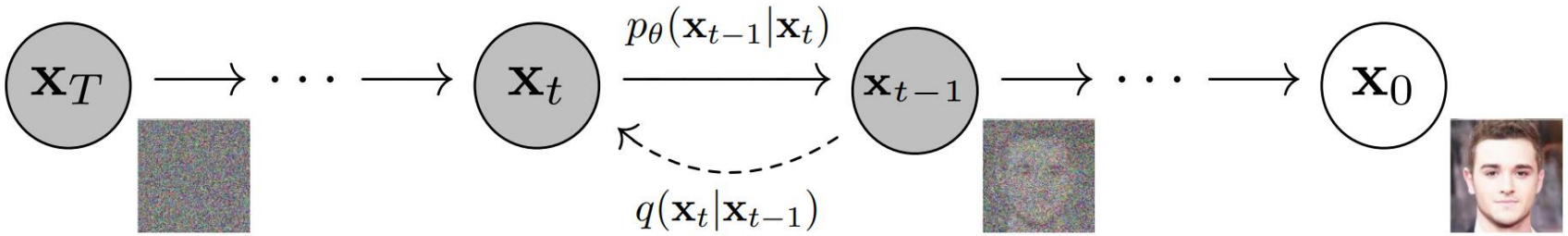


Figure 2: The directed graphical model considered in this work.

Assumption

- Forward and Reverse process are both Markov Chain process.
  - $\forall t' < t-1,\ q(x_t|x_{t-1}, x_{t'}) = q(x_t|x_{t-1})$.
  - $\forall t' > t, \qquad p_\theta(x_{t-1}|x_t, x_{t'}) = p_\theta(x_{t-1}|x_t)$.

- Forward process is Gaussian distribution pre-defined based on α_t.

$$q(x_t|x_{t-1}) \coloneqq \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I)$$

α_t has been scheduled by pre-defined function.

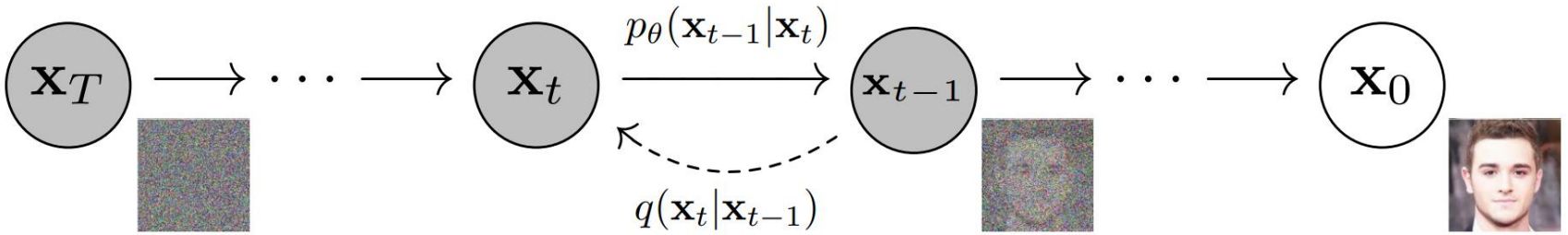


Figure 2: The directed graphical model considered in this work.

Deriving the ELBO

- As other MLE-based generative models, diffusion model is also trained to maximize Log-Likelihood $\mathbb{E}_{x_0 \sim q(x_0)}[\log p_\theta(x_0)]$.
- Following the diffusion modeling, $p_\theta(x_0)$ is transformed into the form of forward $q(x_t|x_{t-1})$ and reverse $p_\theta(x_{t-1}|x_t)$.

$$p_\theta(x_0) = \int p_\theta(x_{0:T}) dx_{1:T}$$ Definition of marginal distribution

$$= \int p_\theta(x_{0:T}) \cdot \frac{q(x_{1:T}|x_0)}{q(x_{1:T}|x_0)} dx_{1:T}$$ Multiply the same term to the numerator and the denominator.

$$= \int p_\theta(x_T) \cdot \frac{\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}{\prod_{t=1}^{T} q(x_t|x_{t-1})} \cdot q(x_{1:T}|x_0) dx_{1:T}$$ Transform following Markov Chain assumption

$$= \mathbb{E}_{x_{1:T} \sim q(x_{1:T}|x_0)} \left[p_\theta(x_T) \cdot \prod_{t=1}^{T} \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right]$$ Transform into the form of Expectation

Sohl-Dickstein et al. "Deep unsupervised learning using nonequilibrium thermodynamics." *ICML*. 2015.

Deriving the ELBO

- Maximize Log-Likelihood $\mathbb{E}_{x_0 \sim q}[\log p_\theta(x_0)]$

$$\mathbb{E}_{x_0 \sim q}[\log p_\theta(x_0)]$$

$$= \int \log\left(p_\theta(x_0)\right) \cdot q(x_0) dx_0$$

$$= \int \log\left(\mathbb{E}_{x_{1:T} \sim q(x_{1:T}|x_0)} \left[p_\theta(x_T) \cdot \prod_{t=1}^{T} \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right] \right) \cdot q(x_0) dx_0$$

Previous slide

$$\geq \int \mathbb{E}_{x_{1:T} \sim q(x_{1:T}|x_0)} \left[\log\left(p_\theta(x_T) \cdot \prod_{t=1}^{T} \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right) \right] \cdot q(x_0) dx_0$$

Jensen's inequality (log is convex)

Evidence Lower BOund (ELBO) or Variational bound

Sohl-Dickstein et al. "Deep unsupervised learning using nonequilibrium thermodynamics." *ICML*. 2015.

Training Objective: Utilization of $q(x_{t-1}|x_t, x_0)$

- Minimize Negative Log-Likelihood by **Minimize Negative ELBO**

$$-(\mathbf{ELBO}) = -\int \mathbb{E}_{x_{1:T}\sim q(x_{1:T}|x_0)}\left[\log\left(p_\theta(x_T)\cdot\prod_{t=1}^{T}\frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})}\right)\right]\cdot q(x_0)dx_0$$

$$= \mathbb{E}_{x_{0:T}\sim q}\left[-\log p_\theta(x_T) - \sum_{t=1}^{T}\log\frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})}\right] \triangleq L$$

Grouping with Expectation

- We aim to transform ELBO into tractable form of KL Divergence such as VAE.
  - Although we define $q(x_t|x_{t-1})$ and $p_\theta(x_{t-1}|x_t)$, we cannot directly use KL Divergence since one is the distribution of x_t and the other is the distribution of x_{t-1}.

Sohl-Dickstein et al. "Deep unsupervised learning using nonequilibrium thermodynamics." *ICML*. 2015.

Training Objective: Utilization of $q(x_{t-1}|x_t, x_0)$

Goal: Transform $p_\theta(x_{t-1}|x_t)$ and $q(x_t|x_{t-1})$ into the form of KL Divergence between $p_\theta(x_{t-1}|x_t)$ and $q(x_{t-1}|x_t, x_0)$.

- Although we do not know $q(x_{t-1}|x_t)$, we can calculate $q(x_{t-1}|x_t, x_0)$ and $q(x_t|x_{t-1})$ using Bayes' Rule when $t \geq 2$.
  - Recap) $q(x_t|x_{t-1}) \coloneqq \mathcal{N}\left(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I\right)$

Bayes' Rule

Remove x_0 (Markov Chain assumption)

$$q(x_{t-1}|x_t, x_0) = \frac{q(x_t|x_{t-1}, x_0)q(x_{t-1}|x_0)}{q(x_t|x_0)} = \frac{q(x_t|x_{t-1})q(x_{t-1}|x_0)}{q(x_t|x_0)}$$

$q(x_t|x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I\right)$
where $\bar{\alpha}_t \triangleq \prod_{s=1}^{t} \alpha_t$

$$= \frac{\frac{1}{\sqrt{1-\alpha_t}\cdot\sqrt{2\pi}}\exp\left(-\frac{\left\|x_t - \sqrt{\alpha_t}x_{t-1}\right\|^2}{2(1-\alpha_t)}\right)\cdot\frac{1}{\sqrt{1-\bar{\alpha}_{t-1}}\cdot\sqrt{2\pi}}\exp\left(-\frac{\left\|x_{t-1} - \sqrt{\bar{\alpha}_{t-1}}x_0\right\|^2}{2(1-\bar{\alpha}_{t-1})}\right)}{\frac{1}{\sqrt{1-\bar{\alpha}_t}\cdot\sqrt{2\pi}}\exp\left(-\frac{\left\|x_t - \sqrt{\bar{\alpha}_t}x_0\right\|^2}{2(1-\bar{\alpha}_t)}\right)}$$

Definition of Gaussian distribution

$$= \mathcal{N}\left(x_{t-1}; \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t, \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t I\right)$$

Form of Gaussian distribution and $\beta_t \coloneqq 1-\alpha_t$

Sohl-Dickstein et al. "Deep unsupervised learning using nonequilibrium thermodynamics." *ICML*. 2015.

Turning into the Form of KL Divergence

$$L \triangleq \mathbb{E}_{x_{0:T}\sim q}\left[-\log p_\theta(x_T) - \sum_{t=1}^{T}\log\frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})}\right]$$

Starting from original ELBO

$$= \mathbb{E}_{x_{0:T}\sim q}\left[-\log p_\theta(x_T) - \sum_{t=2}^{T}\log\frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} - \log\frac{p_\theta(x_0|x_1)}{q(x_1|x_0)}\right]$$

Since $q(x_{t-1}|x_t, x_0)$ is defined when $t \geq 2$, the range of t changes into $t = 2,3,\dots,T$

$$= \mathbb{E}_{x_{0:T}\sim q}\left[-\log p_\theta(x_T) - \sum_{t=2}^{T}\log\left(\frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)}\cdot\frac{q(x_{t-1}|x_0)}{q(x_t|x_0)}\right) - \log\frac{p_\theta(x_0|x_1)}{q(x_1|x_0)}\right]$$

Desired term

Transform into the form of $q(x_{t-1}|x_t, x_0)$

$$= \mathbb{E}_{x_{0:T}\sim q}\left[-\log\frac{p_\theta(x_T)}{q(x_T|x_0)} - \sum_{t=2}^{T}\log\frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} - \log p_\theta(x_0|x_1)\right]$$

Rearrangement

$$= \mathbb{E}_{x_{0:T}\sim q}\left[\underbrace{D_{KL}\big(q(x_T|x_0)\parallel p_\theta(x_T)\big)}_{L_T} + \sum_{t=2}^{T}\underbrace{D_{KL}\big(q(x_{t-1}|x_t, x_0)\parallel p_\theta(x_{t-1}|x_t)\big)}_{L_{t-1}} - \underbrace{\log p_\theta(x_0|x_1)}_{L_0}\right]$$

Turn each loss term into the form of KL Divergence

Sohl-Dickstein et al. "Deep unsupervised learning using nonequilibrium thermodynamics." *ICML*. 2015.

Interpretation of Training Objective

- The objective function can be split into three parts.

$$L \triangleq \mathbb{E}_{x_{0:T}\sim q}\left[\underbrace{D_{KL}\left(q(x_T|x_0) \parallel p_\theta(x_T)\right)}_{L_T} + \sum_{t=2}^{T} \underbrace{D_{KL}\left(q(x_{t-1}|x_t, x_0) \parallel p_\theta(x_{t-1}|x_t)\right)}_{L_{t-1}} - \underbrace{\log p_\theta(x_0|x_1)}_{L_0}\right]$$

- DDPM separately interprets L_T, L_{t-1}, and L_0 and design the model architecture and loss functions.

Interpretation of Training Objective: L_T

$$L_T = D_{KL}\big(q(x_T|x_0) \parallel p_\theta(x_T)\big)$$

- By scheduling α_t, we can make $q(x_T|x_0)$ always follow $\mathcal{N}(0, I)$.
- Since $p_\theta(x_T)$ can also be defined as $\mathcal{N}(0, I)$, we can achieve L_T perfectly as zero.

- $\Rightarrow L_T$ can be ignored during the training.

Interpretation of Training Objective: L_{t-1}

$$L_{t-1} = D_{KL}\big(q(x_{t-1}|x_t, x_0) \parallel p_\theta(x_{t-1}|x_t)\big)$$

- Based on the pre-defined α_t scheduling, $q(x_{t-1}|x_t, x_0)$ is as follows:

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}\left(x_{t-1}; \underbrace{\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{(1-\bar{\alpha}_t)}x_0 + \frac{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)}x_t}_{\tilde{\mu}_t}, \underbrace{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t I}_{\tilde{\sigma}_t}\right)$$

$$\Rightarrow q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t, \tilde{\sigma}_t I)$$

- We model $p_\theta(x_{t-1}|x_t)$ as follows:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\big(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\big)$$

$$\mu_\theta(x_t, t) \to \tilde{\mu}_t \qquad \Sigma_\theta(x_t, t) \to \tilde{\sigma}_t$$

- Since we already know $\tilde{\sigma}_t$, $\Sigma_\theta(x_t, t) \coloneqq \tilde{\sigma}_t I$.

Reparameterization Tricks

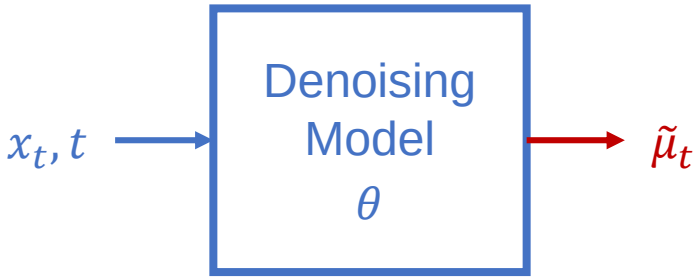
- Denoising model can obtain $\tilde{\mu}_t$ through the following processes:

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}\left(x_{t-1}; \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{(1-\bar{\alpha}_t)}x_0 + \frac{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)}x_t, \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t I\right)$$

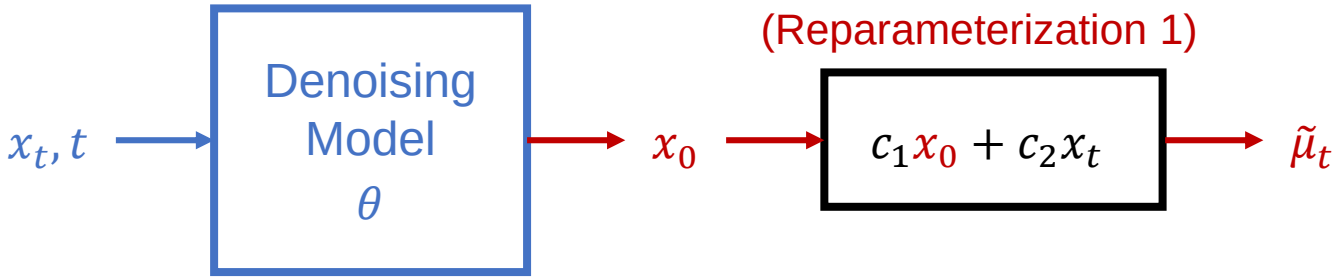
$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t \text{ where } \epsilon_t \sim \mathcal{N}(0, I)$$

$$\Leftrightarrow x_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t}{\sqrt{\bar{\alpha}_t}}$$

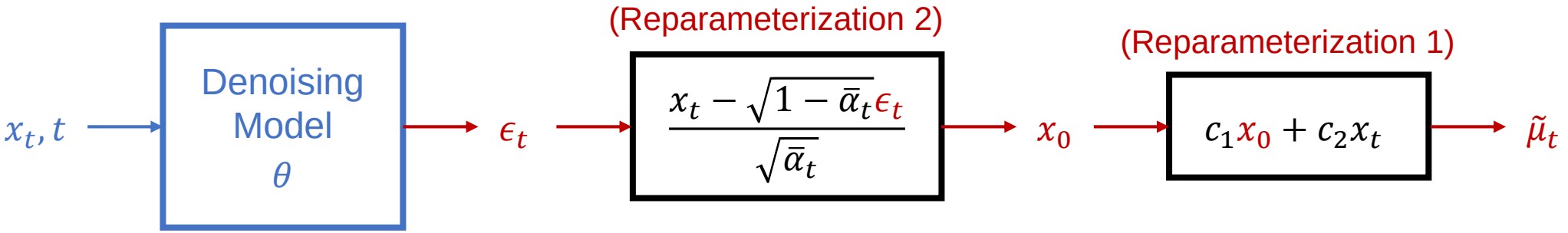
1) Directly predict $\tilde{\mu}_t$



2) Predict the ground-truth x_0

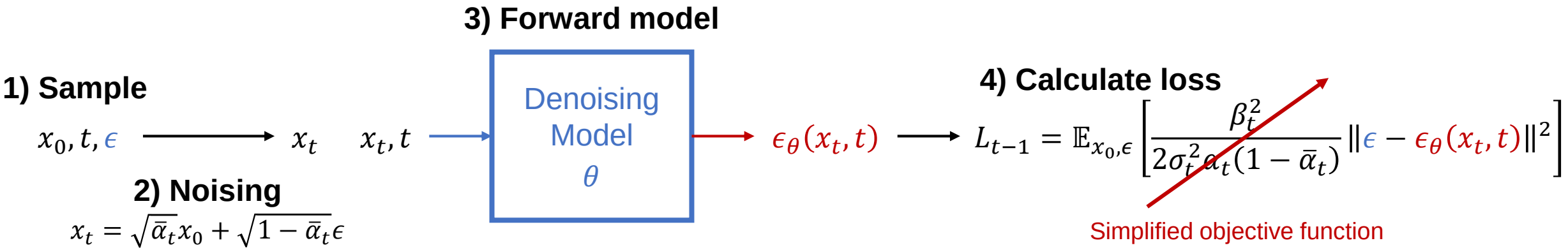


3) Predict the direction ϵ_t



Training Process: L_{t-1}

1. Sample an image x_0 from training data, the degree of noise t from 0 to 1000, and noise ϵ from $\mathcal{N}(0, I)$, respectively.
2. Synthesize noisy image x_t using the sampled x_0, t, and ϵ, based on $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t$
3. Given x_t and t as input, the denoising model θ and predict the added noise ϵ.
4. Train the denoising model θ based on the pre-defined loss function using the difference between the predicted $\epsilon_\theta(x_t, t)$ and the real added noise ϵ.



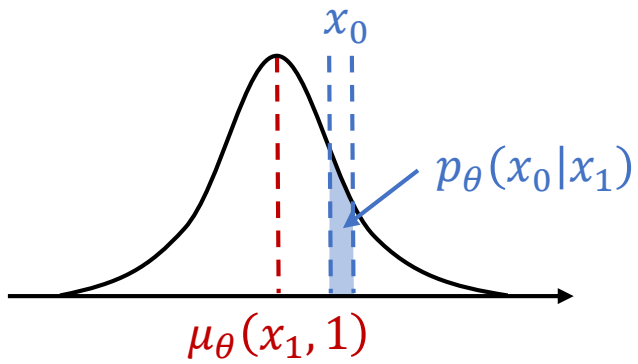
Interpretation of Training Objective: L_0

$$L_0 = \mathbb{E}_{x_{0:T}\sim q}[-\log p_\theta(x_0|x_1)]$$

- Given x_1 and t as inputs, the model is trained to maximize likelihood $p_\theta(x_0|x_1)$ to make the predicted x_0 close to the real x_0.
- Since the real x_0 consists of integers in $\{0, 1, \ldots, 255\}$, which are then normalized between -1 and 1, we compute the likelihood of such a discrete variable.

$$p_\theta(\mathbf{x}_0|\mathbf{x}_1) = \prod_{i=1}^{D} \int_{\delta_-(x_0^i)}^{\delta_+(x_0^i)} \mathcal{N}(x; \mu_\theta^i(\mathbf{x}_1, 1), \sigma_1^2)\, dx$$

$$\delta_+(x) = \begin{cases} \infty & \text{if } x = 1 \\ x + \frac{1}{255} & \text{if } x < 1 \end{cases} \qquad \delta_-(x) = \begin{cases} -\infty & \text{if } x = -1 \\ x - \frac{1}{255} & \text{if } x > -1 \end{cases}$$



Summary: Training and Sampling

- Unlike the training procedure, inference cannot be performed in a single step.
  - At inference time, we first sample an image of the pure Gaussian noise and gradually denoise it (similar to the inference of autoregressive models).

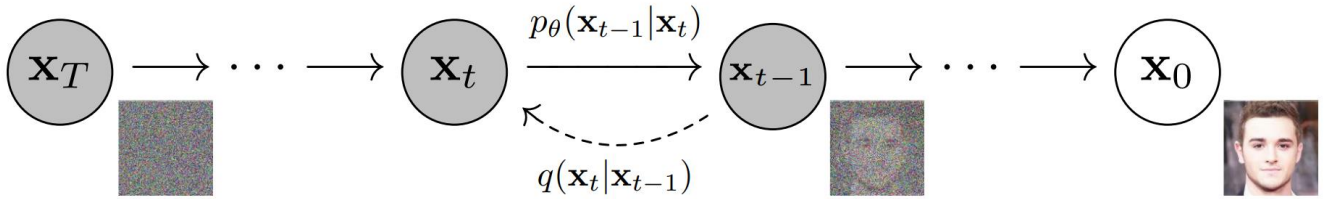


Figure 2: The directed graphical model considered in this work.

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
6: **until** converged

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

Quantitative/Qualitative Results

- A diffusion model achieves a better or comparable performance compared to the existing GAN-based approaches.

Table 1: CIFAR10 results. NLL measured in bits/dim.

Model	IS	FID	NLL Test (Train)
Conditional			
EBM [11]	8.30	37.9	
JEM [17]	8.76	38.4	
BigGAN [3]	9.22	14.73	
StyleGAN2 + ADA (v1) [29]	**10.06**	**2.67**	
Unconditional			
Diffusion (original) [53]			≤ 5.40
Gated PixelCNN [59]	4.60	65.93	3.03 (2.90)
Sparse Transformer [7]			**2.80**
PixelIQN [43]	5.29	49.46	
EBM [11]	6.78	38.2	
NCSNv2 [56]		31.75	
NCSN [55]	8.87 ± 0.12	25.32	
SNGAN [39]	8.22 ± 0.05	21.7	
SNGAN-DDLS [4]	9.09 ± 0.10	15.42	
StyleGAN2 + ADA (v1) [29]	$\mathbf{9.74} \pm 0.05$	3.26	
Ours (L, fixed isotropic $\mathbf{\Sigma}$)	7.67 ± 0.13	13.51	≤ 3.70 (3.69)
Ours (L_{simple})	9.46 ± 0.11	**3.17**	≤ 3.75 (3.72)

Figure 3: LSUN Church samples. FID=7.89

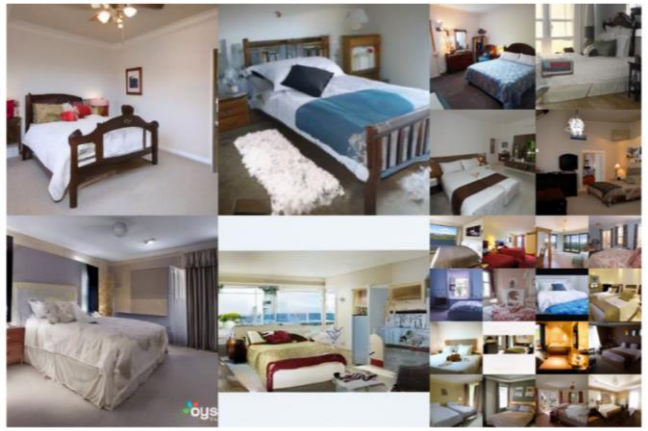


Figure 4: LSUN Bedroom samples. FID=4.90

Categorical Diffusion Process

Argmax Flows and Multinomial Diffusion: Learning Categorical Distributions

---

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, Max Welling

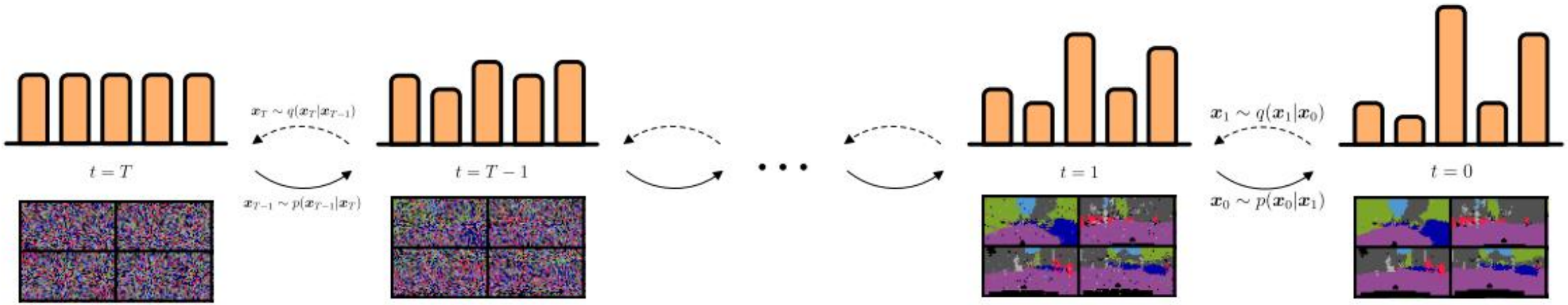
University of Amsterdam

NeurIPS 2021

Presented by Minho Park

Categorical Diffusion Process

- They define the multinomial diffusion process (categorical diffusion process) using a categorical distribution that has a β_t **chance of resampling a category uniformly.**
- $q(x_t|x_{t-1}) = \mathcal{C}(x_t; (1-\beta_t)x_{t-1} + \beta_t/K)$ Forward process of categorical diffusion process



Recap the Objective Function

- Objective function of the diffusion process is

$$L \triangleq \mathbb{E}_{x_{0:T}\sim q}\left[D_{KL}\big(q(x_T|x_0) \parallel p_\theta(x_T)\big) + \sum_{t=2}^{T} D_{KL}\big(q(x_{t-1}|x_t, x_0) \parallel p_\theta(x_{t-1}|x_t)\big) - \log p_\theta(x_0|x_1)\right].$$

  - We do not incorporate Gaussian properties when formulating the objective function.

- We also need $q(x_{t-1}|x_t, x_0)$ for training the categorical diffusion process.

Categorical posterior $q(x_{t-1}|x_t, x_0)$

- Forward process: $q(x_t|x_{t-1}) = \mathcal{C}(x_t; (1-\beta_t)x_{t-1} + \beta_t/K)$
- Express the probability of any x_t directly: $q(x_t|x_0) = \mathcal{C}(x_t; \bar{\alpha}_t x_0 + (1-\bar{\alpha}_t)/K)$
- Then, the categorical posterior can be computed in closed-form:

$$q(x_{t-1}|x_t, x_0) = \mathcal{C}(x_{t-1}; Z[(1-\beta_t)x_t + \beta_t/K] \odot [\bar{\alpha}_{t-1}x_0 + (1-\bar{\alpha}_{t-1})/K])$$
$$= \mathcal{C}\left(x_{t-1}; \Theta_{\text{post}}(x_t, x_0)\right)$$

- Detailed proofs for each step are provided in A.1.1 in our paper.

Training Objective

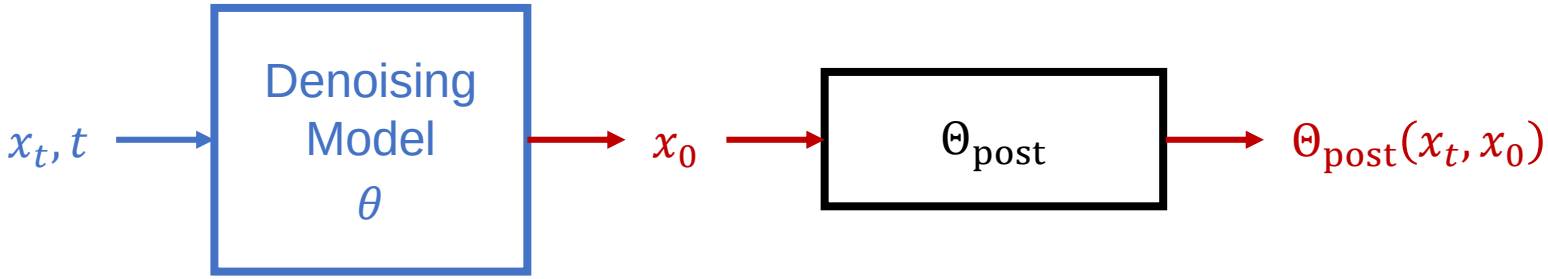
- Since the categorical denoising model can be defined as we reparametrize the denoising model to predict the ground-truth x_0

$$p_\theta(x_{t-1}|x_t) = \mathcal{C}\big(x_{t-1}; \Theta_\theta(x_t)\big),$$
$$= \mathcal{C}\left(x_{t-1}; \Theta_{\text{post}}\big(x_t, \hat{x}_0(x_t, t; \theta)\big)\right)$$

- The KL divergence loss between $q(x_{t-1}|x_t, x_0)$ and $p_\theta(x_{t-1}|x_t)$ can be represented as KL divergence loss between two PMF.

$$D_{KL}\big(q(x_{t-1}|x_t, x_0) \parallel p_\theta(x_{t-1}|x_t)\big) = D_{KL}\left(\Theta_{\text{post}}(x_t, x_0) \parallel \Theta_{\text{post}}\big(x_t, \hat{x}_0(x_t, t; \theta)\big)\right)$$

2) Predict the ground-truth x_0



Quantitative/Qualitative Results

- The experiments have been conducted on language modelling tasks and learning image segmentation maps unconditionally.

text8

```
 that the role of tellings not be required also action characters passe
d on constitution ahmad a nobilitis first be closest to the cope and dh
ur and nophosons she criticized itm specifically on august one three mo
vement and a renouncing local party of exte

nt is in this meant the replicat today through the understanding elemen
t thinks the sometimes seven five his final form of contair you are lot
ur and me es to ultimately this work on the future all all machine the
silon words thereis greatly usaged up not t
```

Cityscapes

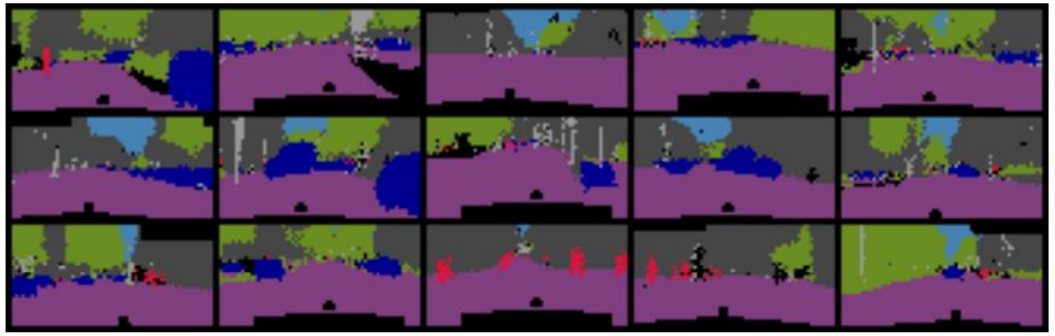


Table 3: Comparison of different methods on `text8` and `enwik8`. Results are reported in negative log-likelihood with units bits per character (bpc) for `text8` and bits per raw byte (bpb) for `enwik8`.

Model type	Model	text8 (bpc)	enwik8 (bpb)
ARM	64 Layer Transformer (Al-Rfou et al., 2019)	1.13	1.06
	TransformerXL (Dai et al., 2019)	1.08	0.99
VAE	AF/AF⋆ (AR) (Ziegler and Rush, 2019)	1.62	1.72
	IAF / SCF⋆ (Ziegler and Rush, 2019)	1.88	2.03
	CategoricalNF (AR) (Lippe and Gavves, 2020)	1.45	-
Generative Flow	Argmax Flow, AR (ours)	1.39	1.42
	Argmax Coupling Flow (ours)	1.82	1.93
Diffusion	Multinomial Text Diffusion (ours)	1.72	1.75

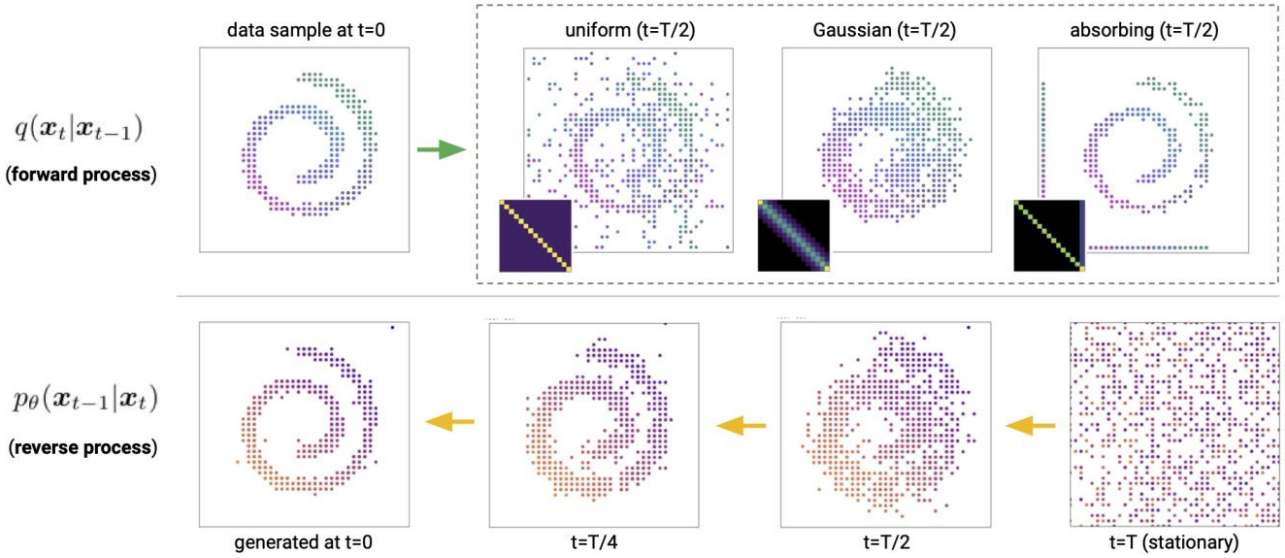
⋆ Results obtained by running code from the official repository for the `text8` and `enwik8` datasets.

Table 4: Performance of different dequantization methods on squares and cityscapes dataset, in bits per pixel, lower is better.

Cityscapes	ELBO	IWBO
Round / Unif. (Uria et al., 2013)	1.010	0.930
Round / Var. (Ho et al., 2019)	0.334	0.315
Argmax / Softplus thres. (ours)	**0.303**	**0.290**
Argmax / Gumbel dist. (ours)	0.365	0.341
Argmax / Gumbel thres. (ours)	**0.307**	**0.287**
Multinomial Diffusion (ours)	0.305	

Variation of the corruption

- Discrete Denoising Diffusion Probabilistic Models: generalize the multinomial diffusion model by going beyond corruption processes with uniform transition probabilities.
- They propose and compare the 1) uniform, 2) absorbing, 3) discretized Gaussian, and 4) Token embedding distance transition.



a)	T = 0	The great brown fox hopped over the lazy dog.
	T = 10	The great [MASK] fox hopped over [MASK] lazy dog.
	T = 20	The [MASK][MASK] [MASK] ship over [MASK] lazy the.
	T = 25	[MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK]
b)	T = 0	The great brown fox hopped over the lazy dog.
	T = 10	The vast black fox hopping over the lazy cat.
	T = 20	Their vast tripped this jumping upon walked organizations.
	T = 25	Bunk scamper tripped this Sanchez walked organizations.

Austin et al. "Structured denoising diffusion models in discrete state-spaces." *NeurIPS.* 2021.

Gaussian-Categorical Diffusion Process

Learning to Generate Semantic Layouts for Higher Text-Image Correspondence in Text-to-Image Synthesis

---

Minho Park*, Jooyeol Yun*, Seunghwan Choi, Jaegul Choo

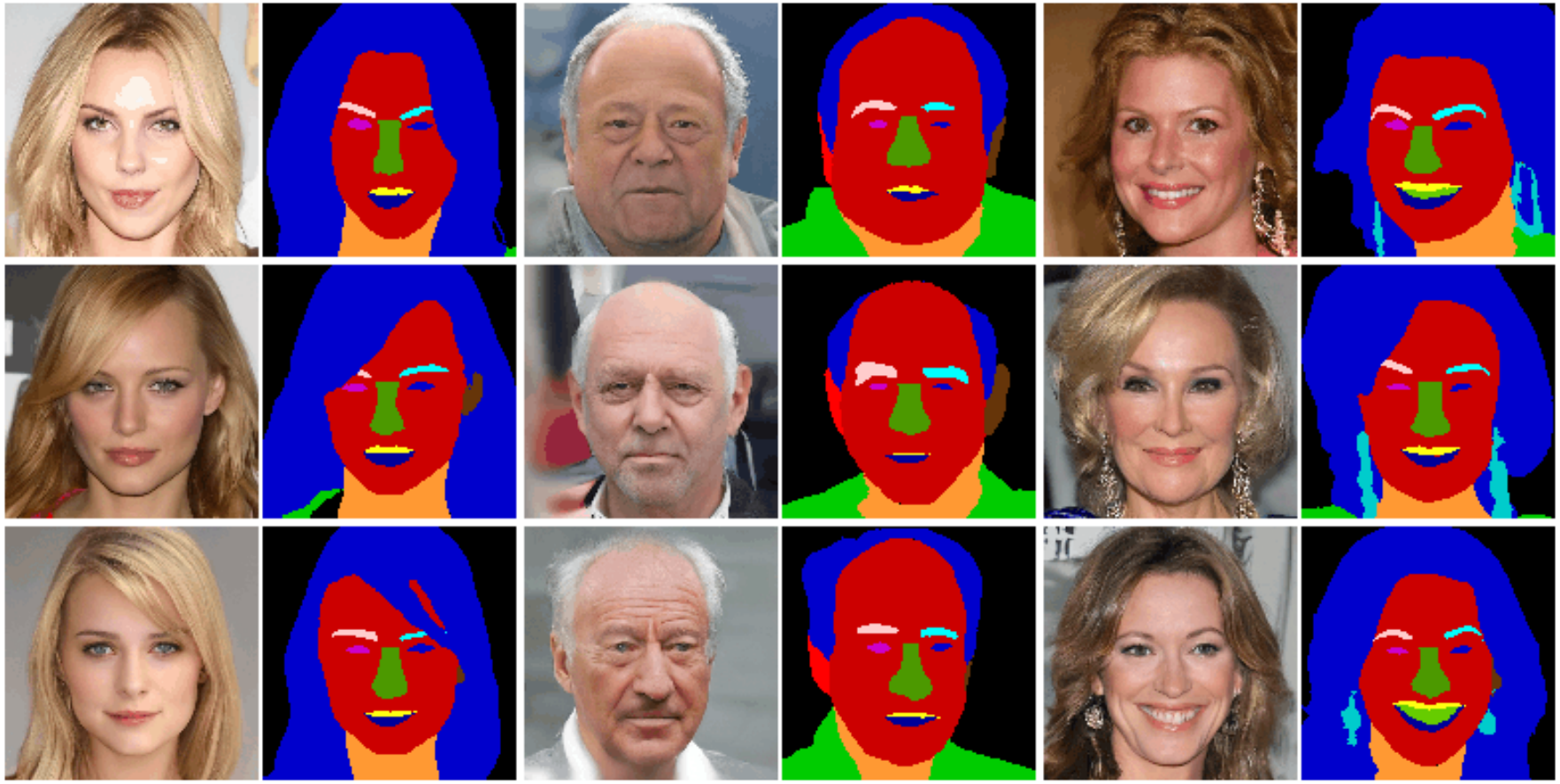
KAIST

ICCV 2023

Presented by Minho Park

Overview

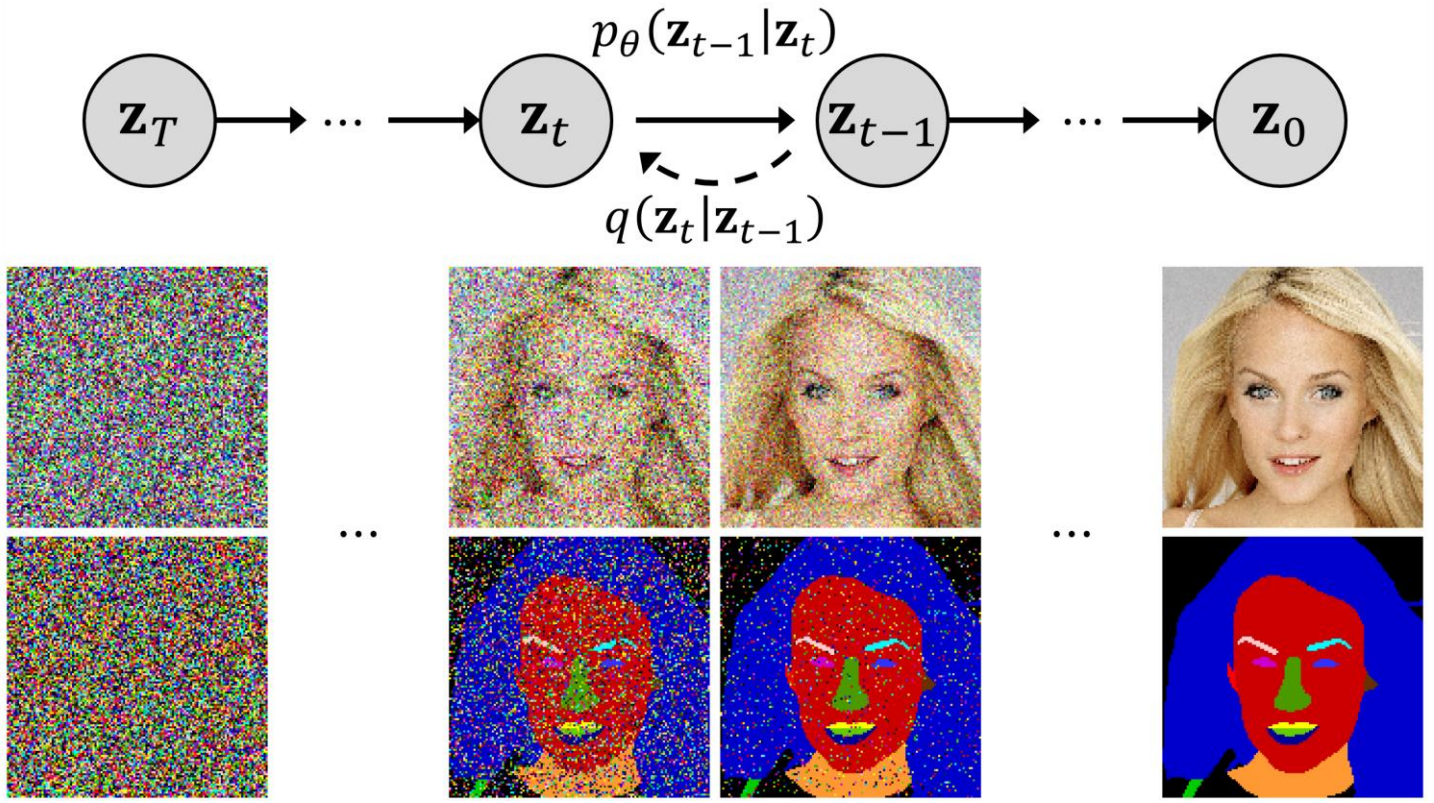
- We propose Gaussian-categorical diffusion process for higher text-image correspondence in text-to-image synthesis.



https://pmh9960.github.io/research/GCDP/

Gaussian-categorical diffusion process

- Gaussian-categorical diffusion process: jointly modeling the images (continuous and ordinal) and labels (discrete and categorical) for generating semantic segmentation datasets.



Gaussian-categorical diffusion process

- Gaussian-categorical distribution

$$(X, Y) \sim \mathcal{NC}(\mathbf{x}, \mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Theta}) = \left(\prod_{i=1}^{M} \boldsymbol{\Theta}_{i,\mathbf{y}_i}\right) (2\pi)^{-\frac{N}{2}} |\boldsymbol{\Sigma}_{\mathbf{y}}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{y}})^{\top} \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} (\mathbf{x}-\boldsymbol{\mu}_{\mathbf{y}})\right)$$

$p(\mathbf{y}) = \mathcal{C}(\mathbf{y}; \boldsymbol{\Theta})$

$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{\mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{y}})$

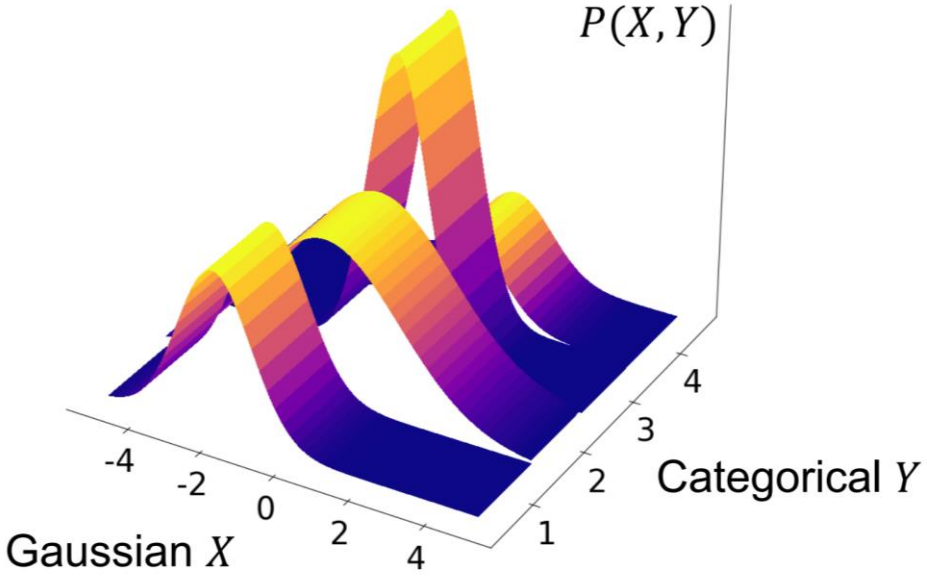
$X = [X_1, X_2, \dots, X_N] \in \mathbb{R}^N$

$Y = [Y_1, Y_2, \dots, Y_M] \in \{1,2, \dots, K\}^M \subset \mathbb{R}^M$

For example, $N = (\#\text{ of pixels}) \times 3$ and $M = (\#\text{ of pixels})$

$\boldsymbol{\mu} \in \mathbb{R}^{S \times N}, \boldsymbol{\Sigma} \in \mathbb{R}^{S \times N \times N}, \boldsymbol{\Theta} \in \mathbb{R}^{M \times K}$

$S = K^M$, possible number of the categorical variable



Visualization of a Gaussian-categorical distribution with a single variable ($N = 1, M = 1, K = 4$)

Gaussian-categorical diffusion process

- We define the forward process of image-layout pairs under the Markov chain assumption as

$$q\left(\mathbf{z}_t \mid \mathbf{z}_{t-1}\right) := \mathcal{NC}\left(\mathbf{z}_t; \left[\boldsymbol{\mu}_{t|t-1}\right]_{\times S}, \left[\boldsymbol{\Sigma}_{t|t-1}\right]_{\times S}, \boldsymbol{\Theta}_{t|t-1}\right), \tag{10}$$

Independent noise with the categorical variable.

Independent noise with the Gaussian variable.

$$\begin{aligned}
\boldsymbol{\mu}_{t|t-1} &:= \sqrt{1-\beta_t^{\mathcal{N}}}\mathbf{x}_{t-1}, \\
\boldsymbol{\Sigma}_{t|t-1} &:= \beta_t^{\mathcal{N}}\boldsymbol{I}, \\
\boldsymbol{\Theta}_{t|t-1} &:= (1-\beta_t^{\mathcal{C}})\mathbf{y}_{t-1} + \beta_t^{\mathcal{C}}/K,
\end{aligned}$$

where $\beta^{\mathcal{C}}$ and $\beta^{\mathcal{N}}$ are predefined noise schedules. We use the notation $[\mathbf{v}]_{\times S}$ to indicate row-wise duplication of a vector $\mathbf{v}$ (*i.e.*, $[\mathbf{v}, \mathbf{v}, ..., \mathbf{v}]^T$).

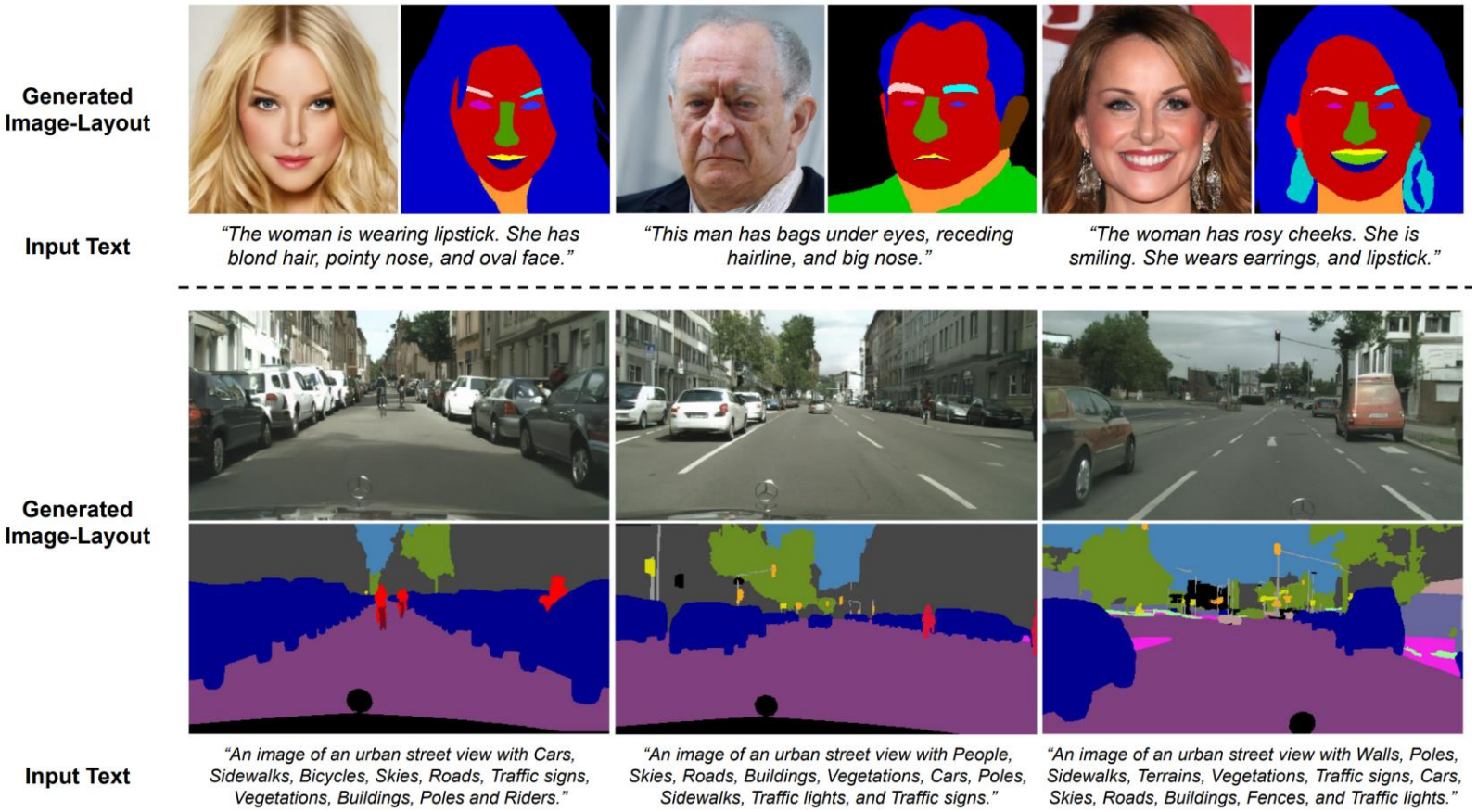
We derive the following objective. Detailed proof for each step are provided in Appendix A.1 of the paper.

$$D_{KL}(q\left(\mathbf{z}_{t-1} \mid \mathbf{z}_t, \mathbf{z}_0\right) \parallel p_\theta\left(\mathbf{z}_{t-1} \mid \mathbf{z}_t\right)) = \mathbb{E}_q\left[\frac{1}{2\sigma_t^2}\left\|\tilde{\boldsymbol{\mu}}_t - \tilde{\boldsymbol{\mu}}_\theta(\mathbf{z}_t)\right\|^2\right] + D_{\mathrm{KL}}(\tilde{\boldsymbol{\Theta}}_t \parallel \boldsymbol{\Theta}_\theta(\mathbf{z}_t)) + C,$$

Objective of the Gaussian Categorical diffusion process

Generating image-layout pairs

- The proposed diffusion process effectively generate the image-layout pairs.

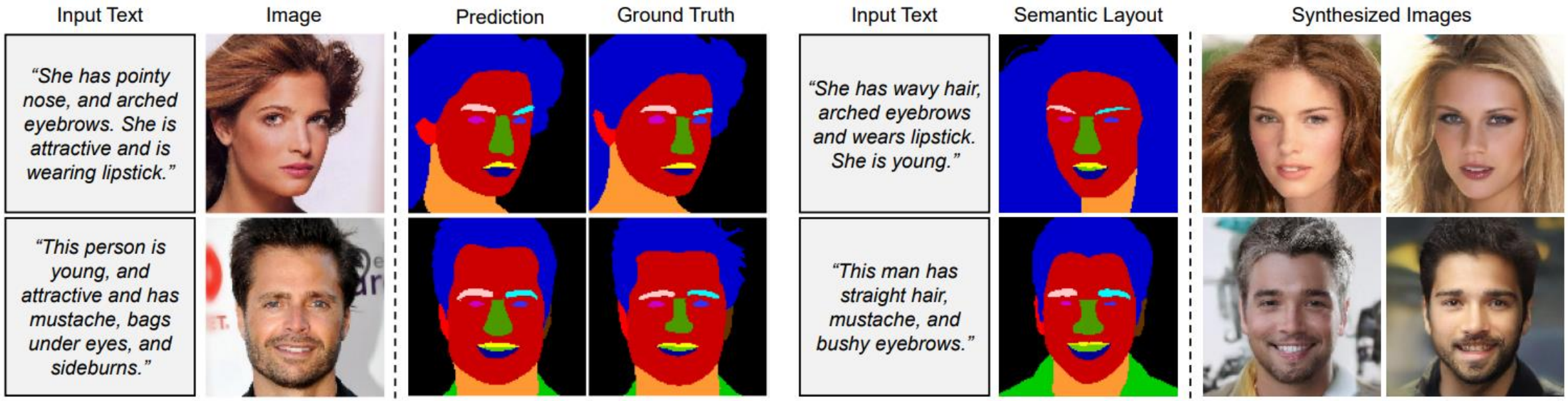


Methods	FID ↓	mIoU ↑	FSD ↓
GANformer [20]	24.86	-	481.5
DatasetDDPM [3]	55.38	33.88	90.31
Semantic Palette [22]	52.13	53.17	48.29
Ours	**20.36**	**65.80**	**42.22**

Table 1. Image-layout alignment and FID of different Image-layout generation approaches for scene generation in the Cityscapes [8] dataset.

Additional advantage of joint generative models

- Text-guided cross-modal outpainting with the joint diffusion model.
  - Since we model the joint distribution $p(x, y)$, we can derive the conditional distribution i.e., $p(x|y)$ or $p(y|x)$.



(a) Text-guided Image-to-Layout Generation

(b) Text-guided Layout-to-Image Generation

Thank You

Q&A