Deep Generative Models

Lecture 14

Roman Isachenko

Moscow Institute of Physics and Technology
Yandex School of Data Analysis

2025, Autumn

Recap of Previous Lecture

$$\mathbb{E}_{p_{\text{data}}(\mathbf{x}(0))}\mathbb{E}_{t\sim U[0,1]}\mathbb{E}_{q(\mathbf{x}(t)|\mathbf{x}(0))}\left\|\mathbf{s}_\theta(\mathbf{x}(t),t)-\nabla_{\mathbf{x}(t)}\log q(\mathbf{x}(t)|\mathbf{x}(0))\right\|_2^2$$

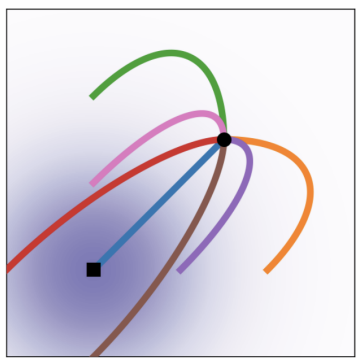
$$p_t(\mathbf{x}_t|\mathbf{x}_1) = q_{1-t}(\mathbf{x}_{1-t}|\mathbf{x}_0 = \mathbf{x}_1)$$

Variance Exploding SDE

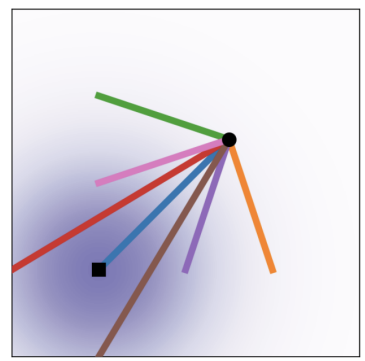
$$p_t(\mathbf{x}_t|\mathbf{x}_1) = \mathcal{N}\left(\mathbf{x}_1, \sigma_{1-t}^2\mathbf{I}\right) \quad \Rightarrow \quad \mathbf{f}(\mathbf{x}_t,\mathbf{x}_1,t) = -\frac{\sigma'_{1-t}}{\sigma_{1-t}}(\mathbf{x}_t-\mathbf{x}_1)$$

Variance Preserving SDE

$$p_t(\mathbf{x}_t|\mathbf{x}_1) = \mathcal{N}\left(\alpha_{1-t}\mathbf{x}_1, (1-\alpha_{1-t}^2)\mathbf{I}\right) \Rightarrow \mathbf{f}(\mathbf{x}_t,\mathbf{x}_1,t) = \frac{\alpha'_{1-t}}{1-\alpha_{1-t}^2}\cdot(\alpha_{1-t}\mathbf{x}_t-\mathbf{x}_1)$$



Diffusion



OT

Lipman Y., et al. Flow Matching for Generative Modeling, 2022

Recap of Previous Lecture

Continuous state space

- **Discrete time** $t \in \{0, 1, \ldots, T\}$ $\Rightarrow$ **DDPM / NCSN**.
- **Continuous time** $t \in [0, 1]$ $\Rightarrow$ **Score-based SDE models**.

Discrete state space

- **Discrete time** $t \in \{0, 1, \ldots, T\}$.
- **Continuous time** $t \in [0, 1]$.

Key advantages of discrete diffusion

- Parallel generation
- Flexible infilling
- Robustness
- Unified framework

Recap of Previous Lecture

Discrete Diffusion Markov Chain

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathrm{Cat}(\mathbf{Q}_t\mathbf{x}_{t-1}),$$

Each $\mathbf{x}_t \in \{0,1\}^K$ is a **one-hot vector** encoding the categorical state (it is just one token).

Transition Matrix

$$[\mathbf{Q}_t]_{ij} = q(x_t = i|x_{t-1} = j), \qquad \sum_{i=1}^{K}[\mathbf{Q}_t]_{ij} = 1.$$

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathrm{Cat}(\mathbf{Q}_{1:t}\mathbf{x}_0), \qquad \mathbf{Q}_{1:t} = \mathbf{Q}_t\mathbf{Q}_{t-1}\cdots\mathbf{Q}_1.$$

- The choice of $\mathbf{Q}_t$ determines how information is erased and what the stationary distribution becomes.
- $\mathbf{Q}_t$ and $\mathbf{Q}_{1:t}$ should be easy to compute for each t.

Austin J. et al. Structured denoising diffusion models in discrete state-spaces, 2021.

Recap of Previous Lecture

Common choices

- **Uniform diffusion**

$$\mathbf{Q}_t = (1-\beta_t)\mathbf{I} + \beta_t \mathbf{U}, \qquad \mathbf{U}_{ij} = \frac{1}{K}.$$

$$\mathbf{Q}_{1:t} = \bar{\alpha}_t \mathbf{I} + (1-\bar{\alpha}_t)\mathbf{U}, \quad \bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s).$$

- **Absorbing diffusion**

$$\mathbf{Q}_t = (1-\beta_t)\mathbf{I} + \beta_t\, \mathbf{e}_m \mathbf{1}^\top.$$

$$\mathbf{Q}_{1:t} = \bar{\alpha}_t\, \mathbf{I} + (1-\bar{\alpha}_t)\, \mathbf{e}_m \mathbf{1}^\top, \qquad \bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s).$$

Austin J. et al. Structured denoising diffusion models in discrete state-spaces, 2021.

Recap of Previous Lecture

Uniform vs. Absorbing Transition Matrix

Aspect	**Uniform Diffusion**	**Absorbing Diffusion**
$\mathbf{Q}_t$	$(1-\beta_t)\mathbf{I} + \beta_t \mathbf{U}$	$(1-\beta_t)\mathbf{I} + \beta_t \mathbf{e}_m \mathbf{1}^\top$
$\mathbf{Q}_{1:t}$	$\bar{\alpha}_t \mathbf{I} + (1-\bar{\alpha}_t)\mathbf{U}$	$\bar{\alpha}_t \mathbf{I} + (1-\bar{\alpha}_t)\mathbf{e}_m \mathbf{1}^\top$
$\mathbf{Q}_{1:\infty}$	$\mathbf{U}$	$\mathrm{Cat}(\mathbf{e}_m)$
Interpretation	Random replacement	Gradual masking of tokens
Application	Image / symbol diffusion	Text diffusion $\approx$ Masked LM

Observation

Both schemes gradually destroy information, but differ in their stationary limit. Absorbing diffusion bridges diffusion and masked-language-model objectives.

Austin J. et al., Structured denoising diffusion models in discrete state-spaces, 2021.

Recap of Previous Lecture

ELBO

$$\mathcal{L}_{\phi,\theta}(\mathbf{x}) = \mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_0)} \log p_\theta(\mathbf{x}_0|\mathbf{x}_1) - \mathrm{KL}\big(q(\mathbf{x}_T|\mathbf{x}_0)\|p(\mathbf{x}_T)\big) - \\ - \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)} \mathrm{KL}\big(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)\|p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)\big)}_{\mathcal{L}_t}$$

Discrete conditioned reverse distribution

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathrm{Cat}\left(\frac{\mathbf{Q}_t\mathbf{x}_t \odot \mathbf{Q}_{1:t-1}\mathbf{x}_0}{\mathbf{x}_t^\top \mathbf{Q}_{1:t}\mathbf{x}_0}\right).$$

- Both $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ and $q(\mathbf{x}_t|\mathbf{x}_0)$ are known analytically from the forward process.
- The reverse process $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is a learned categorical distribution:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathrm{Cat}\big(\boldsymbol{\pi}_\theta(\mathbf{x}_t, t)\big).$$

Austin J. et al., Structured denoising diffusion models in discrete state-spaces, 2021.

Recap of Previous Lecture

ELBO term

$$\mathcal{L}_t = \mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)} \mathrm{KL}\big(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \,\|\, p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)\big).$$

Austin J. et al., Structured denoising diffusion models in discrete state-spaces, 2021.

Recap of Previous Lecture

ELBO term

$$\mathcal{L}_t = \mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)}\mathrm{KL}\big(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \,\|\, p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)\big).$$

Categorical KL

$$\mathrm{KL}\big(\mathrm{Cat}(\mathbf{q}) \,\|\, \mathrm{Cat}(\mathbf{p})\big) = \sum_{k=1}^{K} q_k \log \frac{q_k}{p_k} = \mathrm{H}(\mathbf{q}, \mathbf{p}) - \mathrm{H}(\mathbf{q}),$$

Austin J. et al., Structured denoising diffusion models in discrete state-spaces, 2021.

Recap of Previous Lecture

ELBO term

$$\mathcal{L}_t = \mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)}\mathrm{KL}\big(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \,\|\, p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)\big).$$

Categorical KL

$$\mathrm{KL}\big(\mathrm{Cat}(\mathbf{q}) \,\|\, \mathrm{Cat}(\mathbf{p})\big) = \sum_{k=1}^{K} q_k \log \frac{q_k}{p_k} = \mathrm{H}(\mathbf{q}, \mathbf{p}) - \mathrm{H}(\mathbf{q}),$$

- $\mathrm{H}\big(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)\big)$ is a constant w.r.t. θ.
- $\mathrm{H}(\mathbf{q}, \mathbf{p}) = -\sum_k q_k \log p_k$ is a **cross-entropy loss**.

Austin J. et al., Structured denoising diffusion models in discrete state-spaces, 2021.

Recap of Previous Lecture

ELBO term

$$\mathcal{L}_t = \mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)}\mathrm{KL}\big(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \,\|\, p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)\big).$$

Categorical KL

$$\mathrm{KL}\big(\mathrm{Cat}(\mathbf{q}) \,\|\, \mathrm{Cat}(\mathbf{p})\big) = \sum_{k=1}^{K} q_k \log \frac{q_k}{p_k} = \mathrm{H}(\mathbf{q}, \mathbf{p}) - \mathrm{H}(\mathbf{q}),$$

- $\mathrm{H}\big(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)\big)$ is a constant w.r.t. $\boldsymbol{\theta}$.
- $\mathrm{H}(\mathbf{q}, \mathbf{p}) = -\sum_k q_k \log p_k$ is a **cross-entropy loss**.

Therefore, minimizing $\mathcal{L}_t$ w.r.t. $\boldsymbol{\theta}$ is equivalent to minimizing

$$\mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)}\mathrm{H}\Big(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0),\ p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)\Big).$$

Austin J. et al., Structured denoising diffusion models in discrete state-spaces, 2021.

Outline

1. Discrete Diffusion
   - Absorbing Diffusion
2. Latent Space Models
   - Score-Based Models
   - Autoregressive Models
3. The Worst Course Overview

Outline

Outline

Absorbing Diffusion: Forward Process

Let's restrict to the case of absorbing transition matrix. At each step t:

- with probability $(1 - \beta_t)$ a token is kept;
- with probability β_t it is replaced by the mask token m.

$$\mathbf{Q}_t = (1 - \beta_t)\mathbf{I} + \beta_t \, \mathbf{e}_m \mathbf{1}^\top, \qquad \bar{\alpha}_t = \prod_{s=1}^{t} (1 - \beta_s).$$

$$\mathbf{Q}_{1:t} = \bar{\alpha}_t \, \mathbf{I} + (1 - \bar{\alpha}_t) \, \mathbf{e}_m \mathbf{1}^\top.$$

Each position is either still clean or already masked:

$$q(\mathbf{x}_t | \mathbf{x}_0) = \begin{cases} \bar{\alpha}_t, & \mathbf{x}_t = \mathbf{x}_0 \\ 1 - \bar{\alpha}_t, & \mathbf{x}_t = \mathbf{e}_m \\ 0, & \text{otherwise.} \end{cases}$$

Sahoo S. et al. Simple and Effective Masked Diffusion Language Models, 2024

NOT READY

Absorbing / Masked Diffusion: Sequence View

Consider a sequence $\mathbf{x}_0 = (x_0^1, \ldots, x_0^L)$.

Independent masking across positions

Because the forward chain factorizes over positions,

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \prod_{\ell=1}^{L} q(x_t^\ell \mid x_0^\ell),$$

and for each position ℓ:

$$q(x_t^\ell = x_0^\ell \mid \mathbf{x}_0) = \bar{\alpha}_t, \qquad q(x_t^\ell = m \mid \mathbf{x}_0) = 1 - \bar{\alpha}_t.$$

- At small t, most tokens remain clean; a few are masked.
- As $t \to T$, almost all tokens become m and $q(\mathbf{x}_T)$ is concentrated on the fully masked sequence.
- This gives a **multi-step masking schedule**, instead of BERT's single-step masking.

Austin J. et al., Structured denoising diffusion models in discrete state-spaces, 2021.

Posterior in Absorbing Diffusion: Unmask vs Stay Masked

Recall the general discrete posterior

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

For the absorbing process we can obtain a closed-form expression.

Case 1: $x_t = x_0$ (token not yet masked)

Because the mask is absorbing, we cannot go from mask back to a clean token:

$$q(x_{t-1} = x_0 \mid x_t = x_0, x_0) = 1.$$

If we observe $x_t = x_0$, we know the token has **never been masked** up to time t.

Austin J. et al., Structured denoising diffusion models in discrete state-spaces, 2021.

Posterior in Absorbing Diffusion: Unmask vs Stay Masked

Case 2: $x_t = m$ (token is masked)

Now x_{t-1} could be:

- already masked at $t-1$ and stayed masked, or
- still clean (x_0) at $t-1$ and masked only at step t.

Using the forward marginals,

$$q(x_{t-1} = x_0 \mid x_t = m, x_0) = \frac{\bar{\alpha}_{t-1}\,\beta_t}{1-\bar{\alpha}_t},$$

$$q(x_{t-1} = m \mid x_t = m, x_0) = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t},$$

and all other states have probability 0.

Posterior in Absorbing Diffusion: Interpretation

Unmask vs stay masked

When $x_t = m$,

$$q(x_{t-1} \mid x_t = m, x_0) = \underbrace{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}}_{\text{already masked}} \delta_{x_{t-1}=m} + \underbrace{\frac{\bar{\alpha}_{t-1}\beta_t}{1-\bar{\alpha}_t}}_{\text{just masked}} \delta_{x_{t-1}=x_0}.$$

- The posterior is a simple binary choice:
  - **stay masked**: keep $x_{t-1} = m$,
  - **unmask**: revert to the original symbol x_0.
- The reverse model $p_{\theta}(x_{t-1} \mid x_t)$ learns, at masked positions, how likely it is to *unmask* vs *stay masked*.
- This is exactly the semantic of **iterative infilling**: tokens start from mask and are gradually turned into meaningful symbols.

ELBO Term for Absorbing Diffusion

Recall the per-timestep ELBO term

$$\mathcal{L}_t = \mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)}\mathrm{KL}\big(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \,\|\, p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)\big).$$

Categorical KL $\Rightarrow$ cross-entropy

As before,

$$\mathrm{KL}\big(\mathrm{Cat}(\mathbf{q}) \,\|\, \mathrm{Cat}(\mathbf{p})\big) = \mathrm{H}(\mathbf{q}, \mathbf{p}) - \mathrm{H}(\mathbf{q}),$$

and the entropy term $\mathrm{H}(\mathbf{q})$ does not depend on $\boldsymbol{\theta}$.
Therefore minimizing $\mathcal{L}_t$ is equivalent (w.r.t. $\boldsymbol{\theta}$) to

$$\mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)}\mathrm{H}\Big(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0), p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)\Big).$$

- For absorbing diffusion, $q(x_{t-1}^\ell \mid x_t^\ell, x_0^\ell)$ is supported only on $\{x_0^\ell, m\}$.
- This makes the target distribution extremely simple, and opens the door to a much simpler training loss.

Austin J. et al., Structured denoising diffusion models in discrete state-spaces, 2021.

Simplified Training: Predict Clean Token at Masked Positions

Key observation

For absorbing diffusion:

- If $x_t^\ell \neq m$, then $x_t^\ell = x_0^\ell$ and the posterior $q(x_{t-1}^\ell \mid x_t^\ell, x_0^\ell)$ is a delta at x_0^ℓ.
- If $x_t^\ell = m$, the posterior is a binary distribution over $\{m, x_0^\ell\}$.

The informative supervision is concentrated at **masked positions**.

Austin, J. et al. Structured denoising diffusion models in discrete state-spaces, 2021.
See also recent refinements in arXiv:2406.07524.

Simplified Training: Predict Clean Token at Masked Positions

Practical training objective

In practice we parameterize the model to predict x_0 from $(\mathbf{x}_t, t)$:

$$p_{\boldsymbol{\theta}}(x_0^\ell \mid \mathbf{x}_t, t) = \mathrm{Cat}\big(\boldsymbol{\pi}_\theta(\mathbf{x}_t, t)^\ell\big),$$

and minimize a time-conditioned cross-entropy:

$$\mathcal{L}_{\mathsf{mask}}(\theta) = \mathbb{E}_{t,\,\mathbf{x}_0,\,\mathbf{x}_t \sim q(\mathbf{x}_t|\mathbf{x}_0)} \sum_{\ell=1}^{L} w_t\, \mathbb{I}\{x_t^\ell = m\} \big[-\log p_{\boldsymbol{\theta}}(x_0^\ell \mid \mathbf{x}_t, t)\big].$$

- w_t – optional weighting over timesteps (e.g., uniform over t).
- We apply cross-entropy only at positions where the input token is masked.

Absorbing Diffusion as Multi-step Masked LM

- Forward process: gradually replace tokens by a mask m according to a diffusion schedule $\{\beta_t\}$.
- Reverse process: starting from an all-mask sequence, iteratively **unmask** positions by predicting clean tokens x_0 from $(\mathbf{x}_t, t)$.
- Training: time-conditioned masked language modeling objective on masked positions:

$$(\mathbf{x}_0, t) \;\mapsto\; \mathbf{x}_t \sim q(\mathbf{x}_t \mid \mathbf{x}_0), \quad \text{predict } x_0^\ell \text{ wherever } x_t^\ell = m.$$

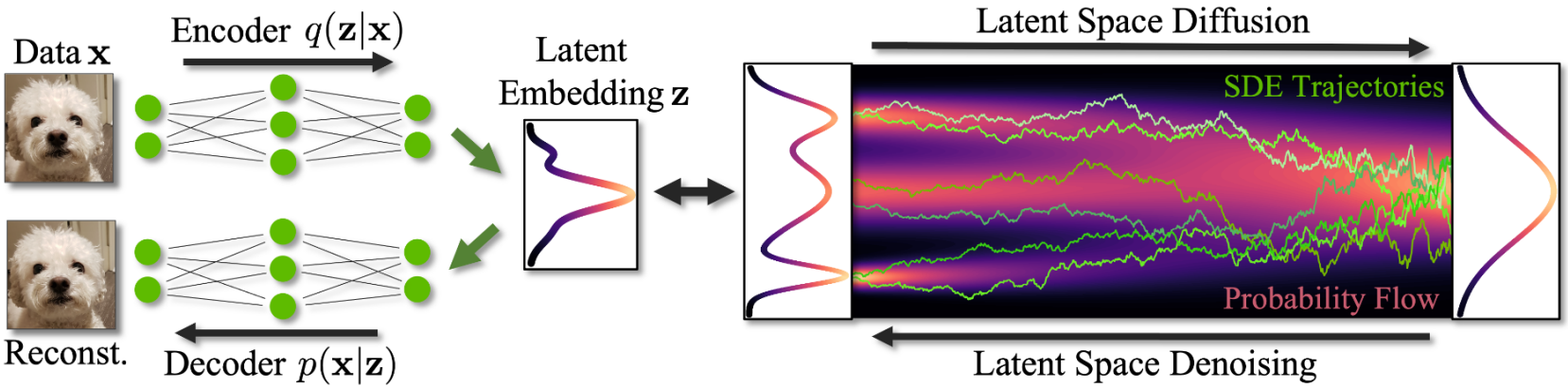
- This perspective makes absorbing diffusion feel very close to BERT-style masked LMs, but with:
  - a **multi-step** corruption schedule,
  - explicit modeling of the full reverse Markov chain.

Outline

Outline

Latent Space Models

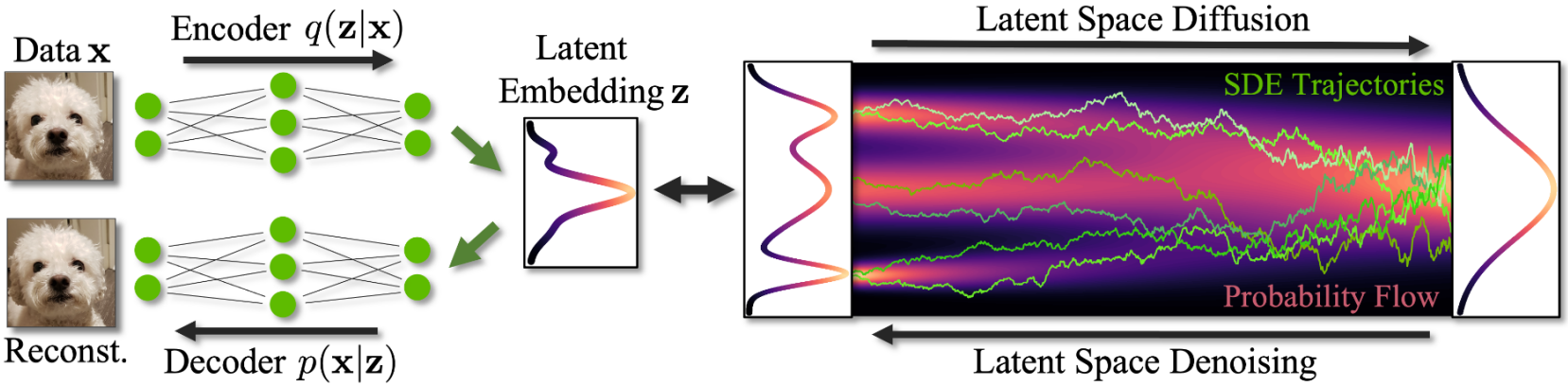
Score-Based Models (Diffusion)



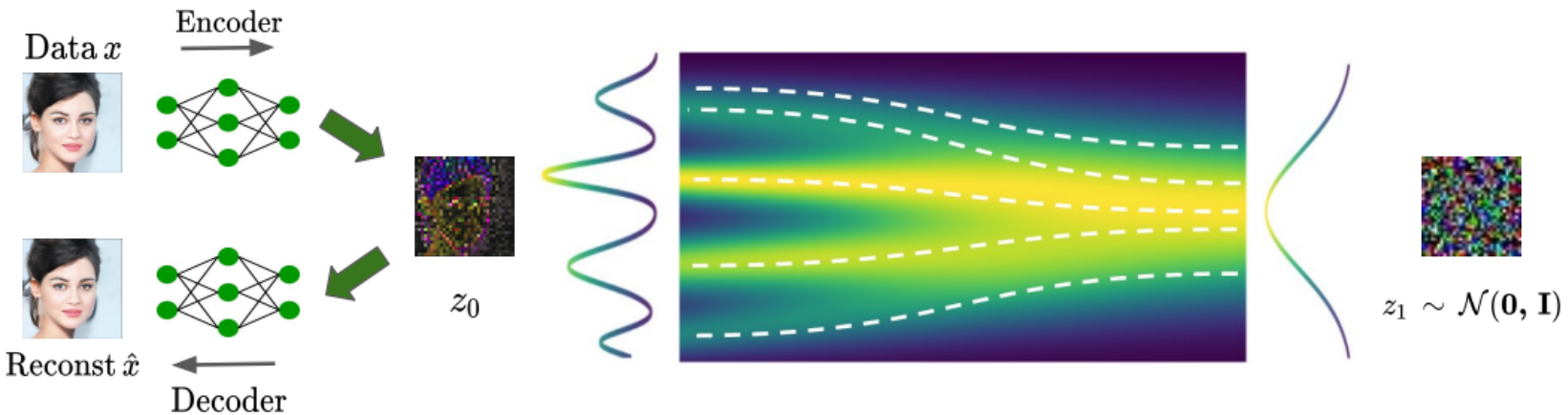
Dao Q. et al. Flow Matching in Latent Space, 2023
NeurIPS 2023 Tutorial: Latent Diffusion Models: Is the Generative AI Revolution Happening in Latent Space?

Latent Space Models

Score-Based Models (Diffusion)



Flow Matching



Dao Q. et al. Flow Matching in Latent Space, 2023
NeurIPS 2023 Tutorial: Latent Diffusion Models: Is the Generative AI Revolution Happening in Latent Space?

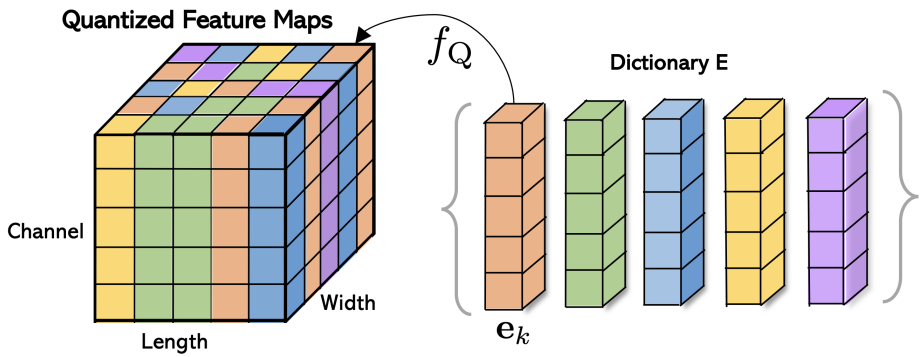
Outline

Vector Quantized VAE (VQ-VAE)

Define a dictionary space $\{\mathbf{e}_k\}_{k=1}^K$, where $\mathbf{e}_k \in \mathbb{R}^C$ and K is the dictionary's size.

$$\mathbf{z}_q = \mathbf{q}(\mathbf{z}) = \mathbf{e}_{k^*}$$

Here $k^* = \arg\min_k \|\mathbf{z} - \mathbf{e}_k\|$.



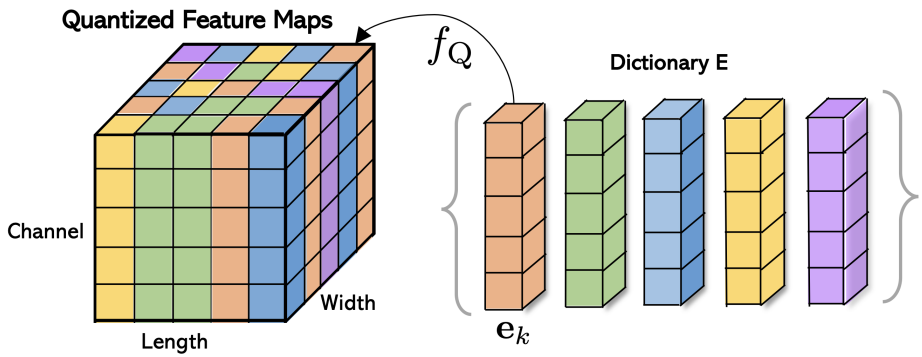
Zhao Y. et al. Feature Quantization Improves GAN Training, 2020
Oord A., Vinyals O., Kavukcuoglu K. Neural Discrete Representation Learning, 2017

Vector Quantized VAE (VQ-VAE)

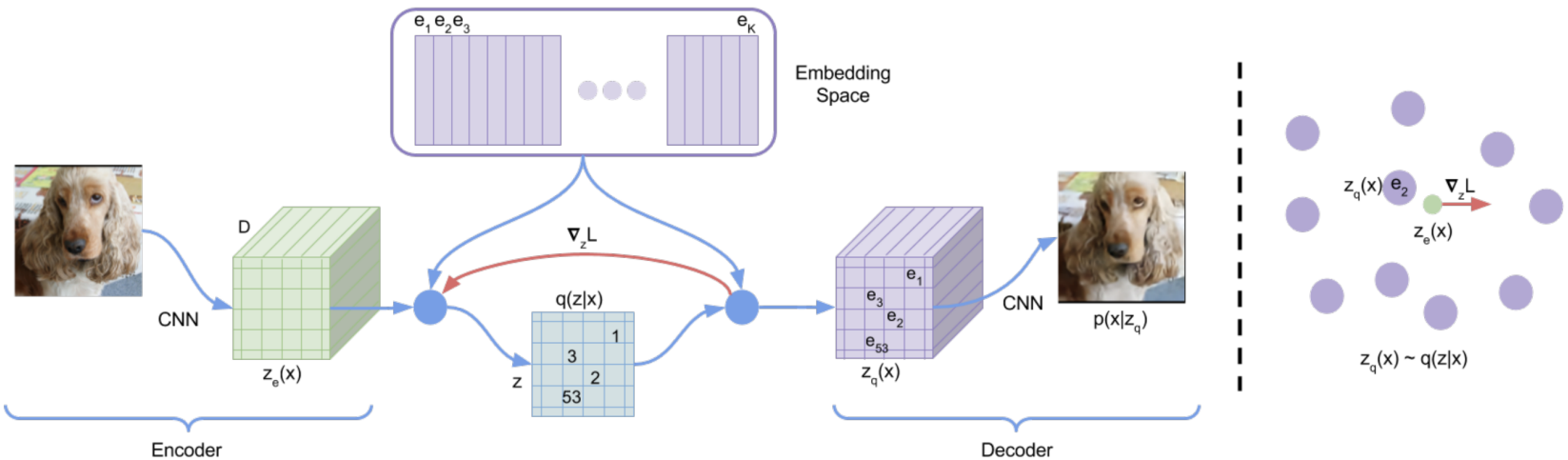
Define a dictionary space $\{\mathbf{e}_k\}_{k=1}^K$, where $\mathbf{e}_k \in \mathbb{R}^C$ and K is the dictionary's size.

$$\mathbf{z}_q = \mathbf{q}(\mathbf{z}) = \mathbf{e}_{k^*}$$

Here $k^* = \arg\min_k \|\mathbf{z} - \mathbf{e}_k\|$.



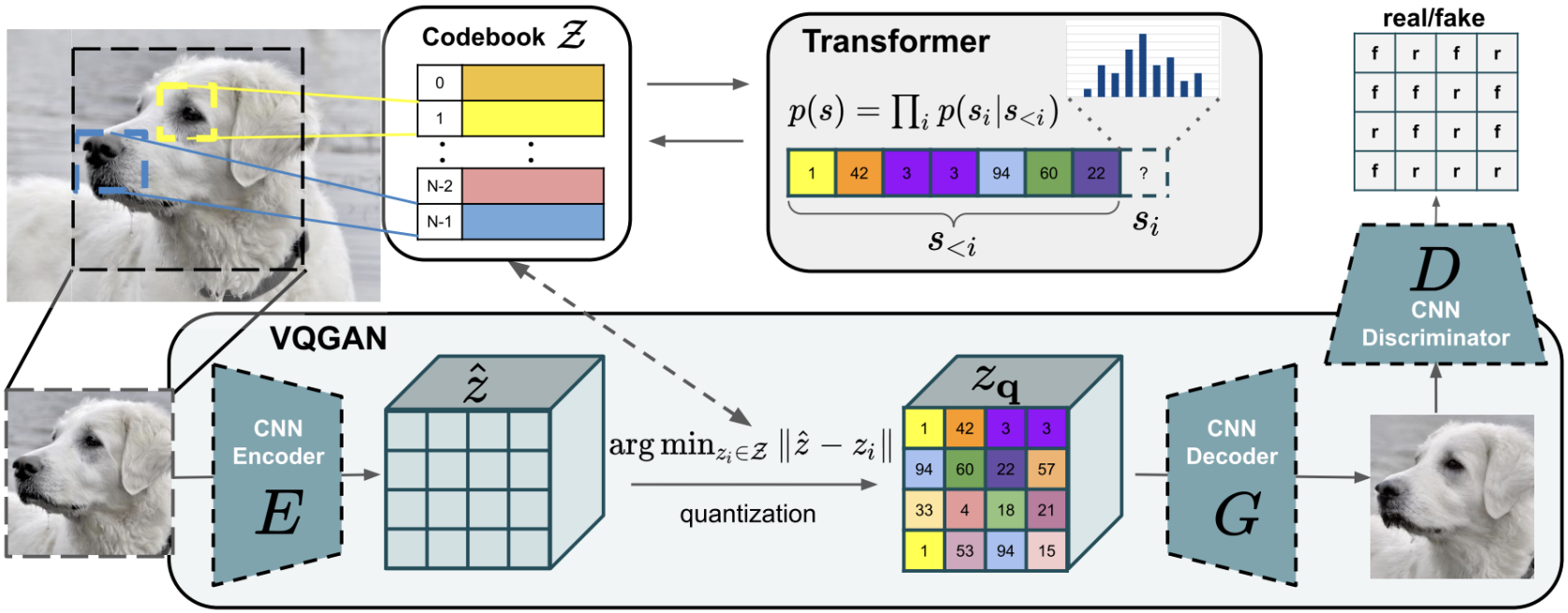
$$\mathcal{L}_{\phi,\theta}(\mathbf{x}) = \log p_\theta(\mathbf{x}|\mathbf{z}_q) - \log K$$



Zhao Y. et al. Feature Quantization Improves GAN Training, 2020
Oord A., Vinyals O., Kavukcuoglu K. Neural Discrete Representation Learning, 2017

Vector Quantized GAN

- We use a VQ-VAE model and its objective.
- We add an adversarial loss between generated and real images to further improve the visual quality of reconstructions.



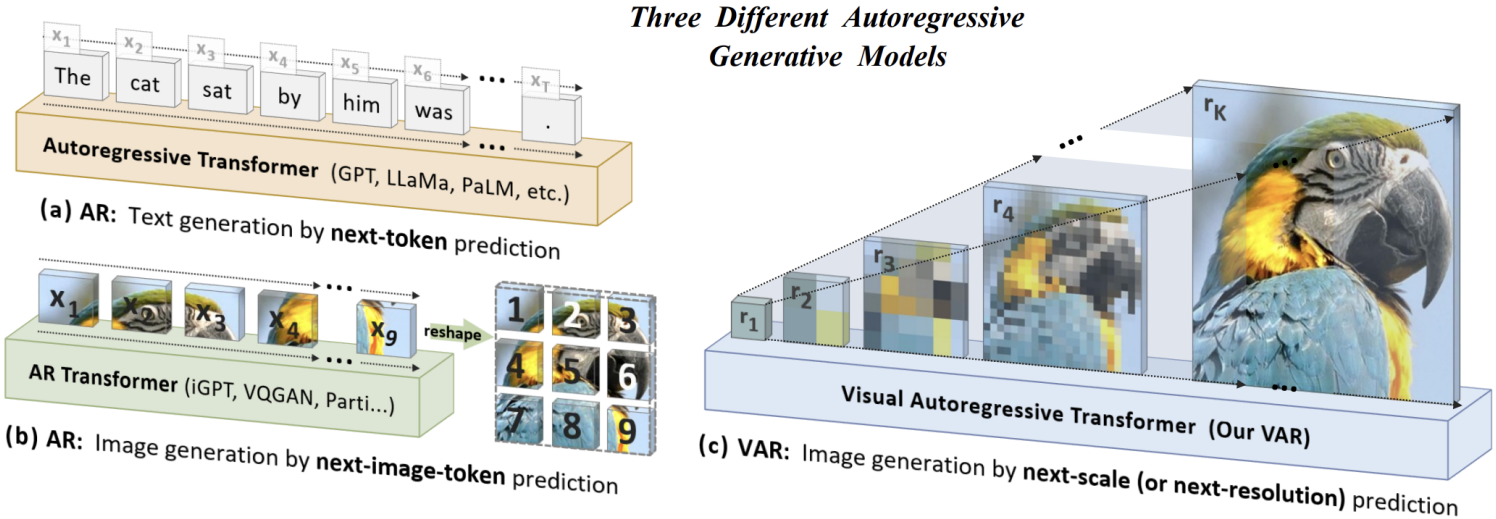
Esser P. et al. Taming Transformers for High-Resolution Image Synthesis, 2020

LlamaGen: Pure Autoregression

- Use a VQ-GAN encoder for mapping images into the discrete latent space (codebook vectors).
- Train a pure autoregressive model (Llama-based) in the latent space.
- Use the VQ-GAN decoder to map discrete tokens back to image space.

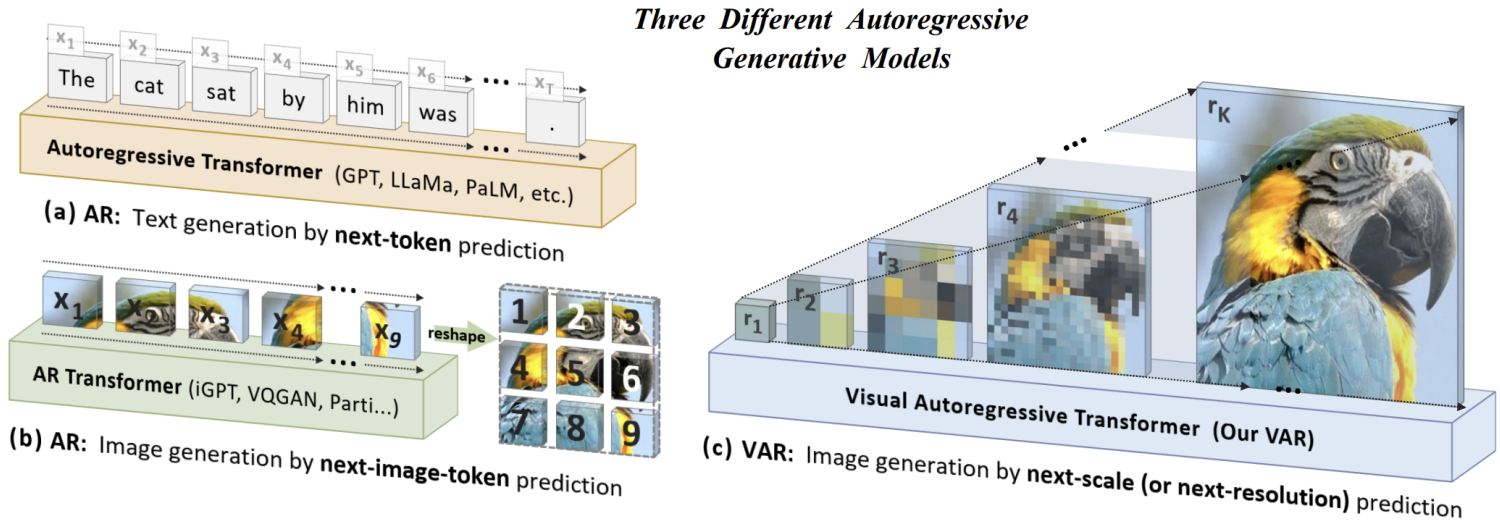
Sun P. et al. Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation, 2024

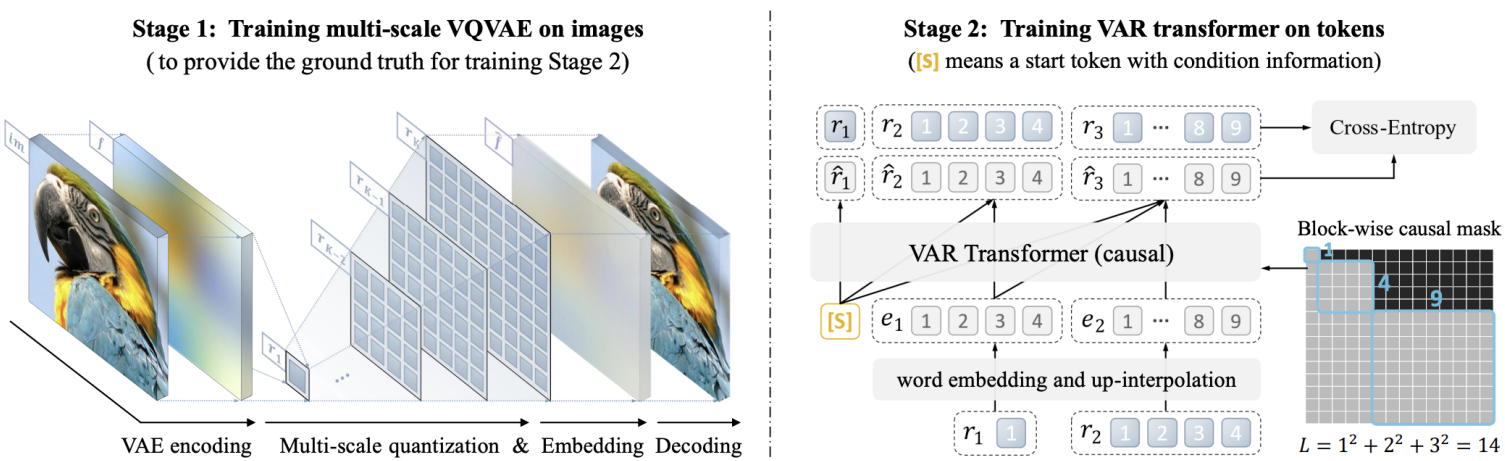
Visual Autoregressive Modeling (VAR)



Tean K. et al. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction, 2024

Visual Autoregressive Modeling (VAR)



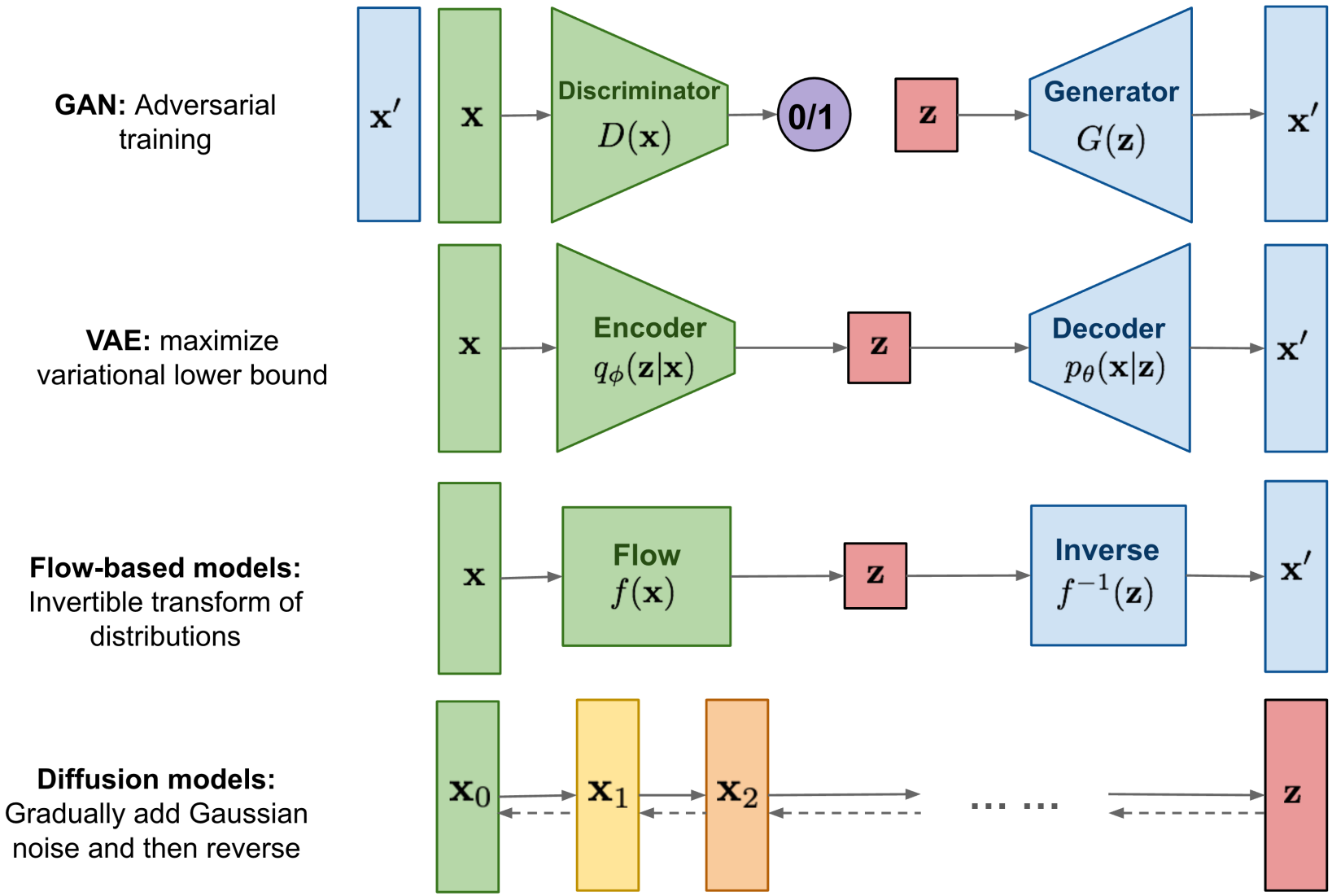


Tean K. et al. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction, 2024

Outline

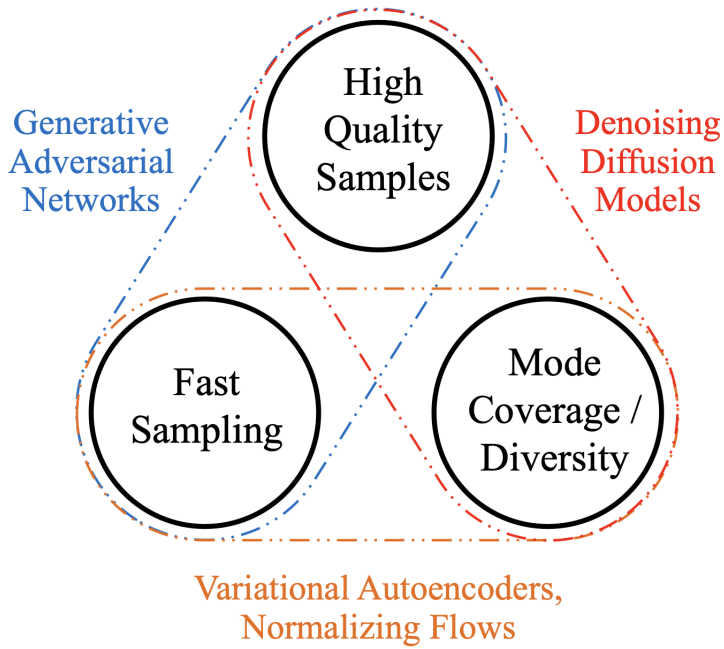
1. Discrete Diffusion
   - Absorbing Diffusion
2. Latent Space Models
   - Score-Based Models
   - Autoregressive Models
3. The Worst Course Overview

The Worst Course Overview :)



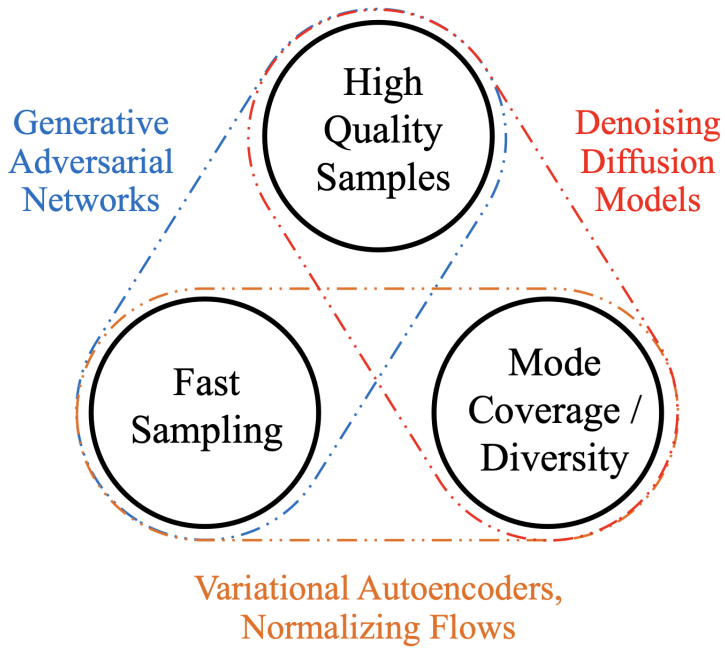
Weng L. What are Diffusion Models?, blog post, 2021

The Worst Course Overview :)



Xiao Z., Kreis K., Vahdat A. Tackling the generative learning trilemma with denoising diffusion GANs, 2021
Simon J.D. Prince. Understanding Deep Learning, 2023

The Worst Course Overview :)



Model	Efficient	Sample quality	Coverage	Well-behaved latent space	Disentangled latent space	Efficient likelihood
GANs	✓	✓	✗	✓	?	n/a
VAEs	✓	✗	?	✓	?	✗
Flows	✓	✗	?	✓	?	✓
Diffusion	✗	✓	?	✗	✗	✗

Xiao Z., Kreis K., Vahdat A. Tackling the generative learning trilemma with denoising diffusion GANs, 2021
Simon J.D. Prince. Understanding Deep Learning, 2023

Summary

- Most state-of-the-art generative models are latent variable models with either continuous or discrete latent spaces.