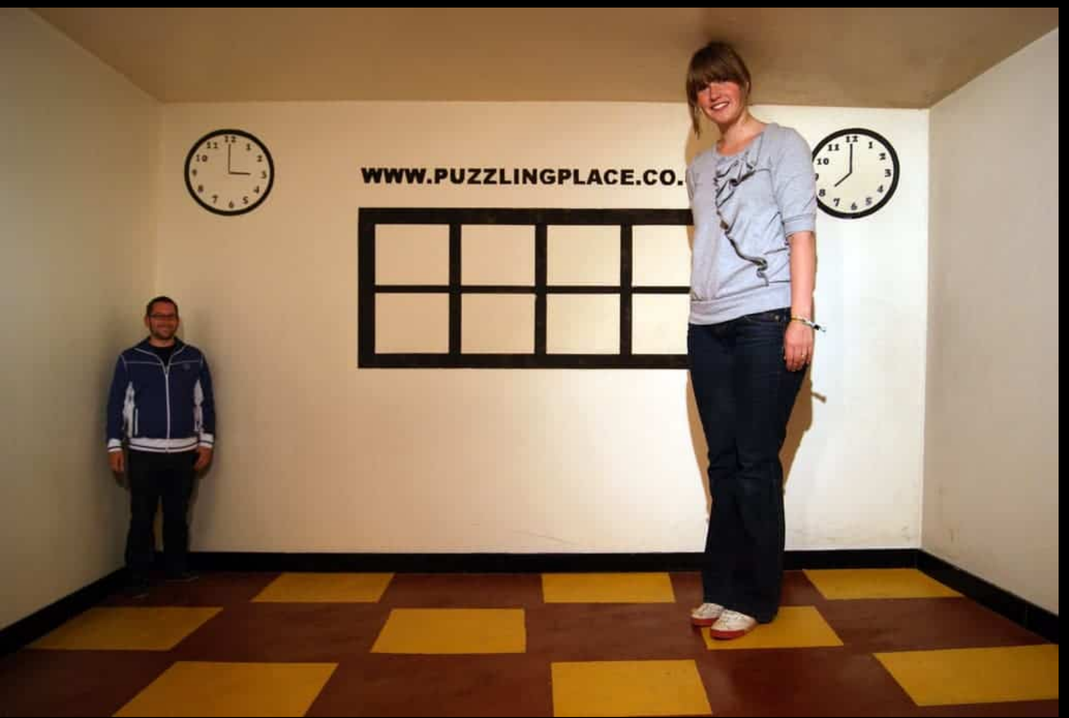
Advanced Computer Vision: Diffusion Models

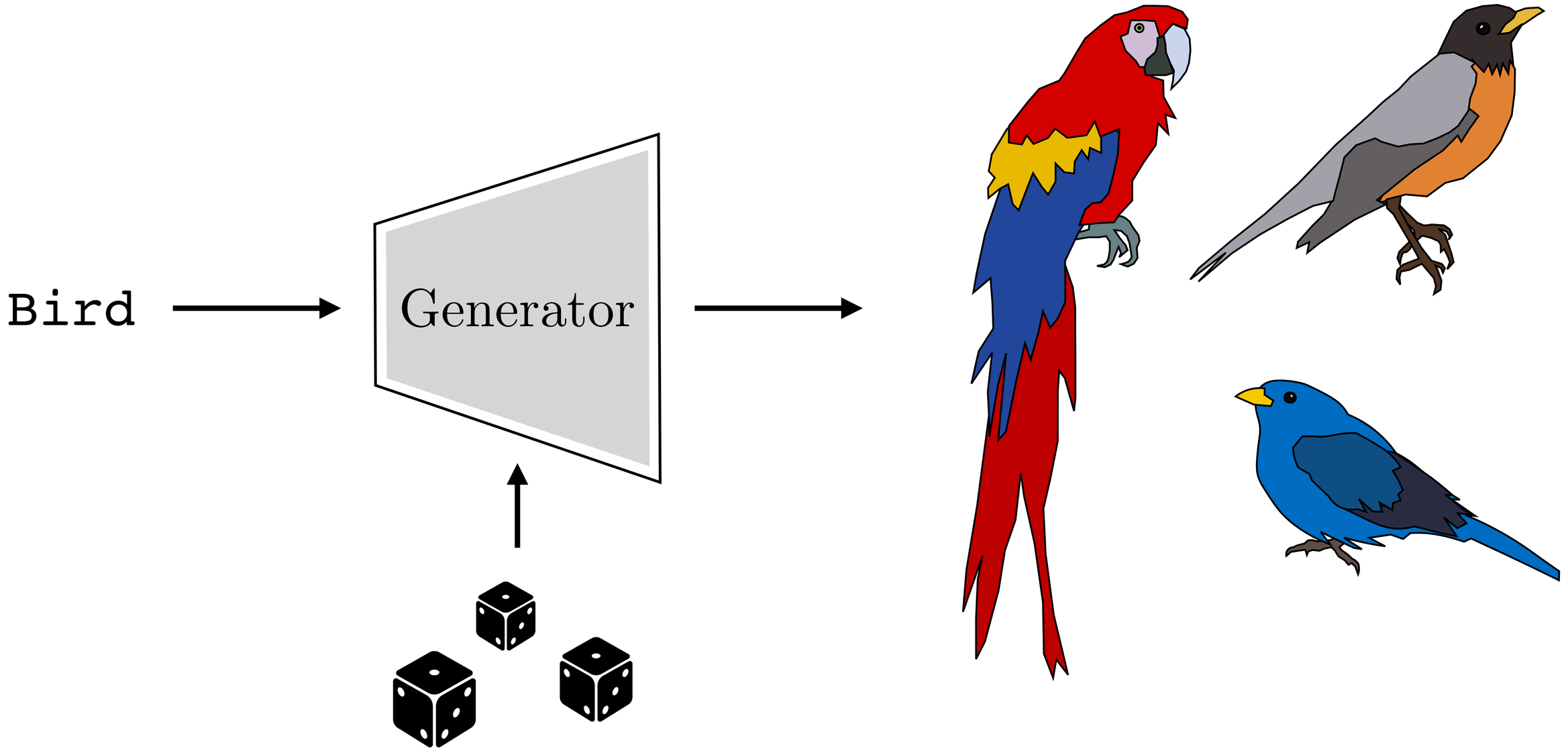
MLMI17

Ayush Tewari



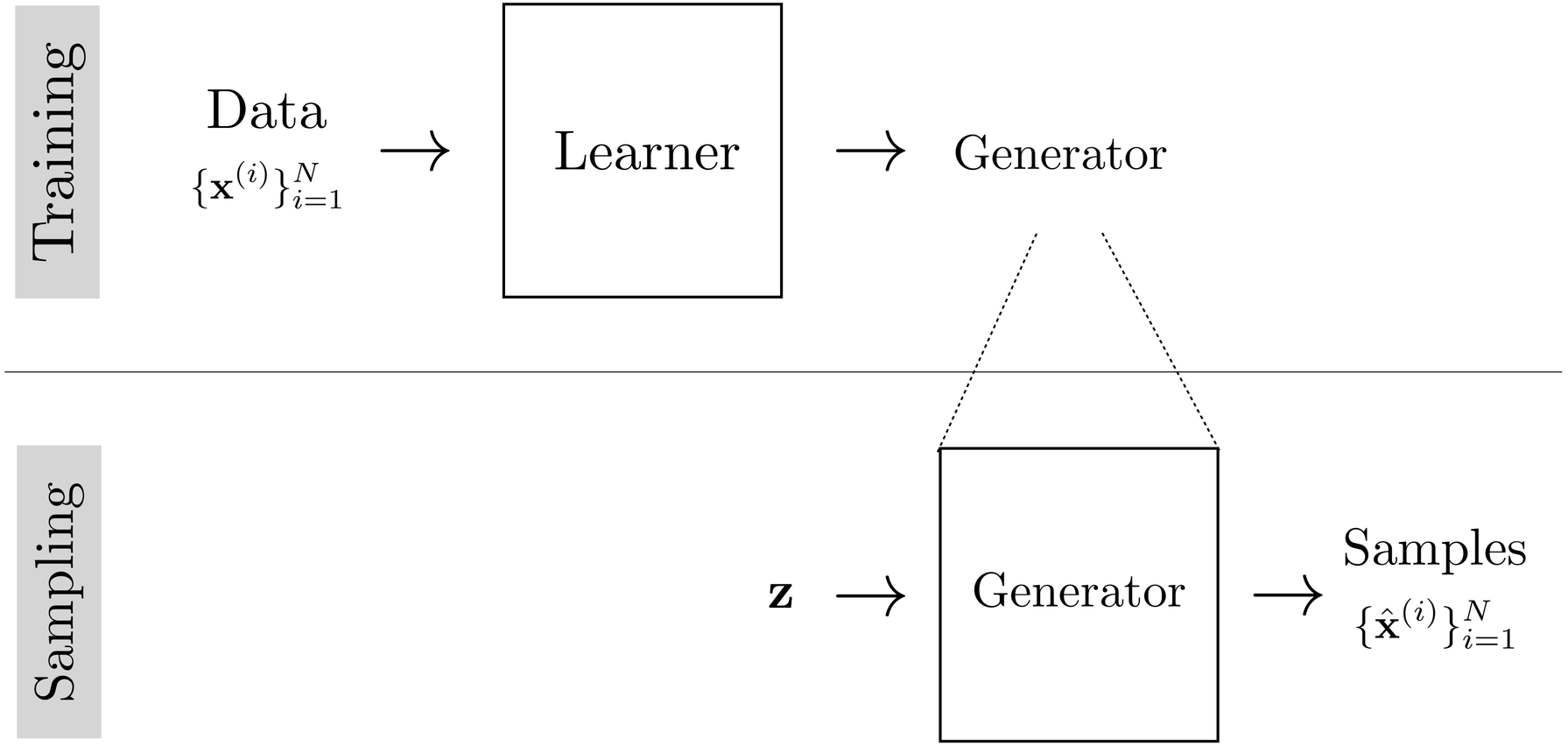
UNIVERSITY OF CAMBRIDGE

Generative models

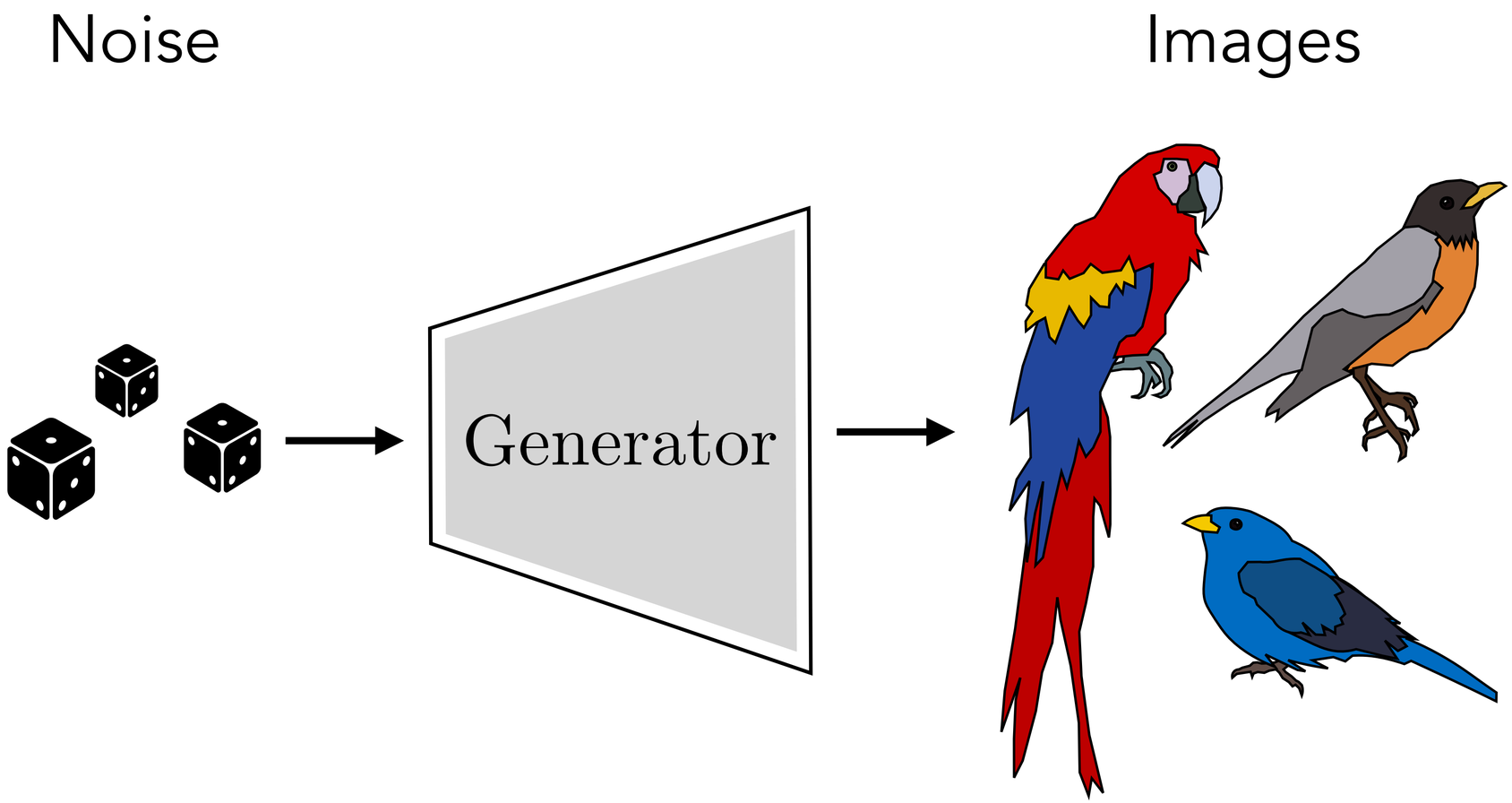


What is a generative model?

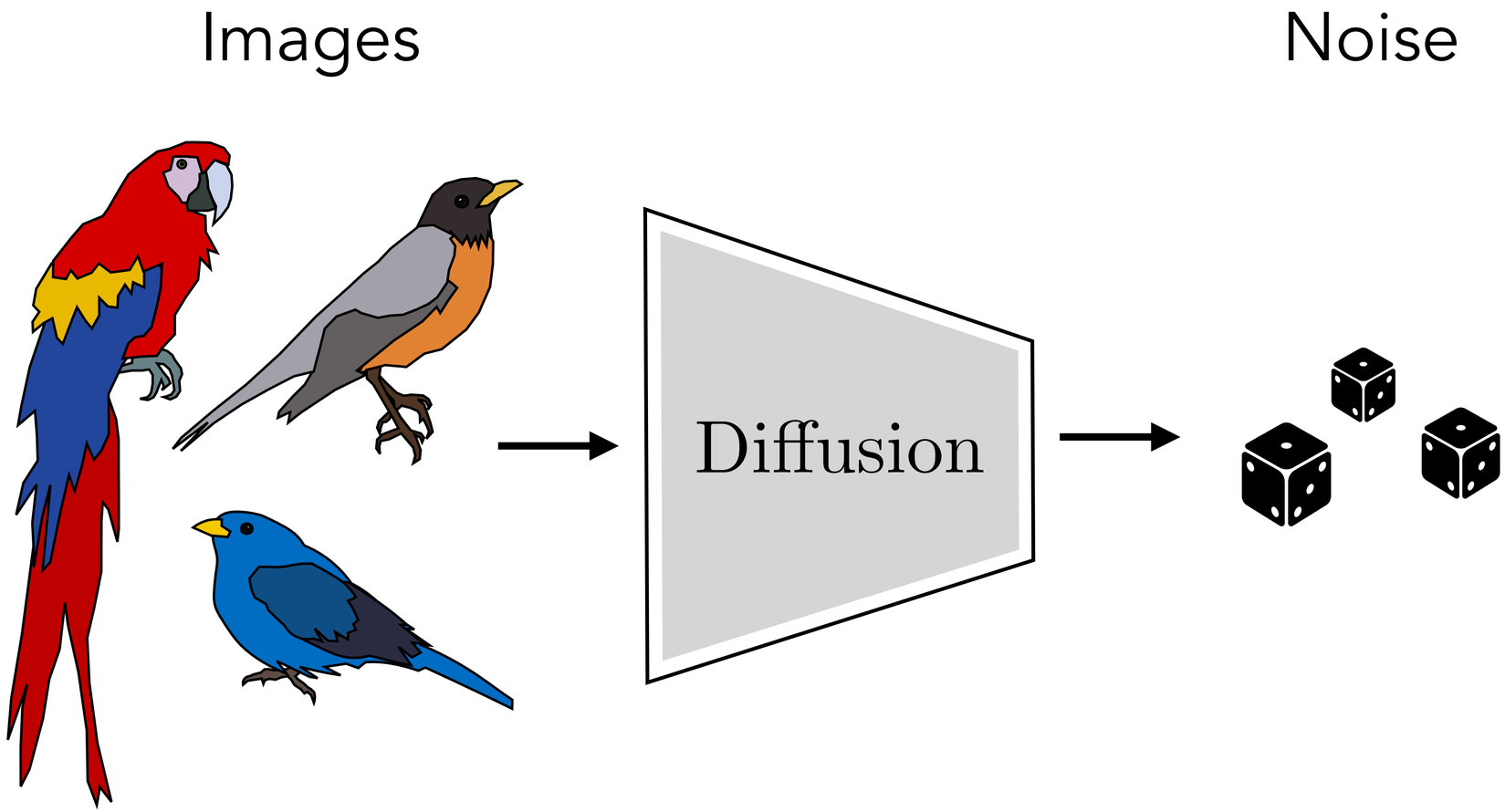
- **An algorithm that generates data**



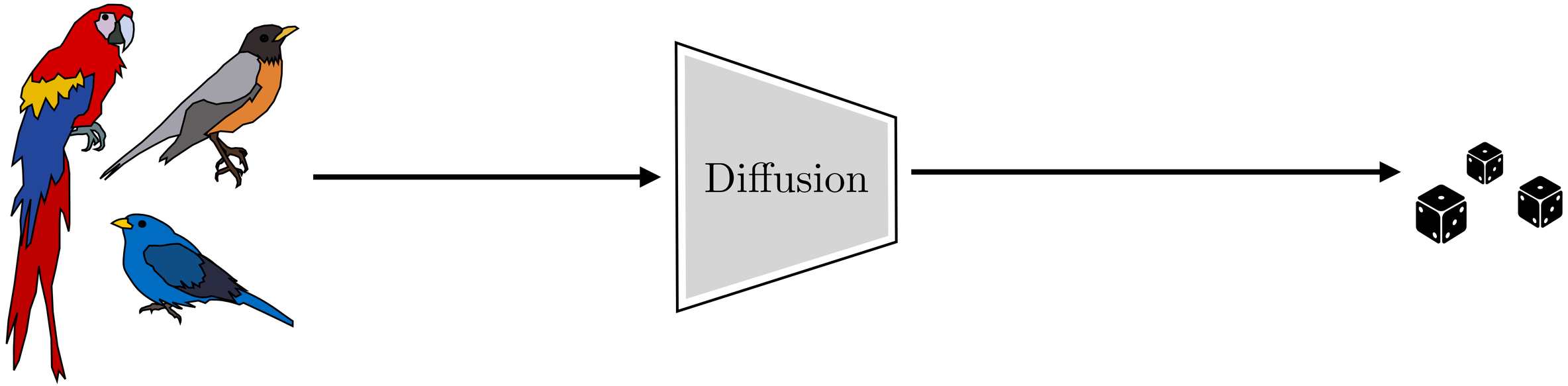
Diffusion Generative Models: Noise -> Data

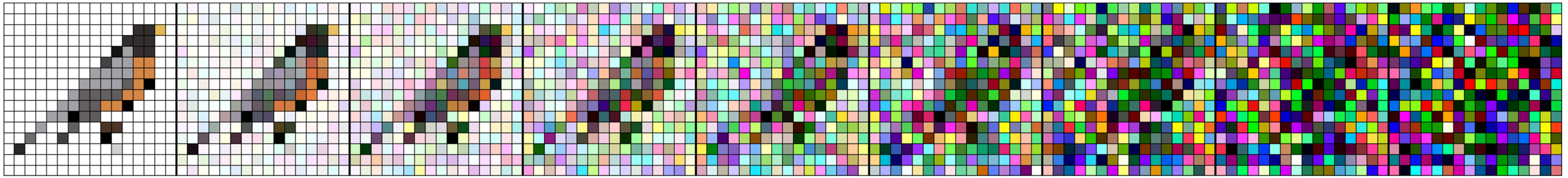


Diffusion Generative Models: Data -> Noise



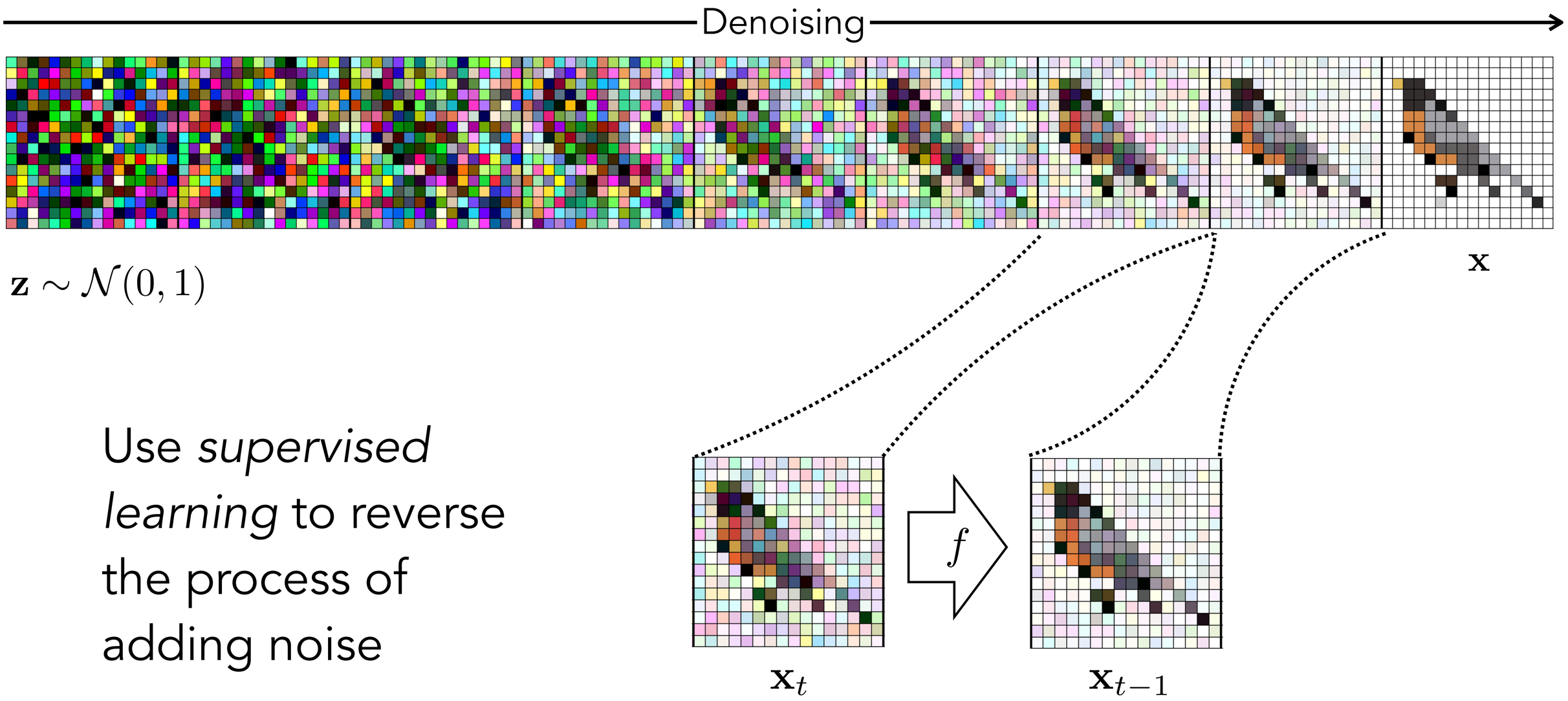
Diffusion Generative Models: Data -> Noise





Diffusion: Just add noise

Diffusion Generative Models

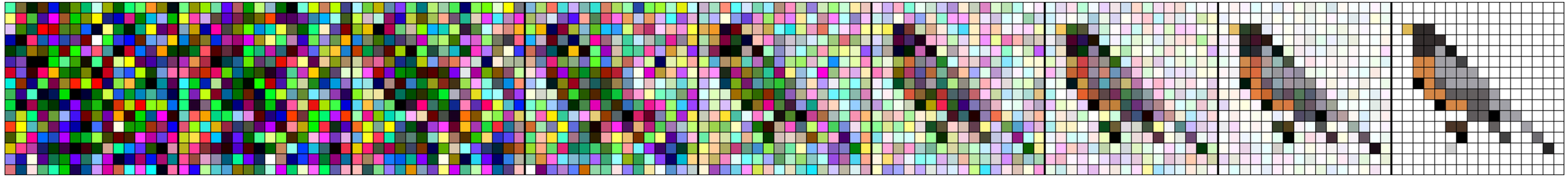


Use *supervised learning* to reverse the process of adding noise

Diffusion Generative Models

$\mathbf{z} \sim \mathcal{N}(0, 1)$

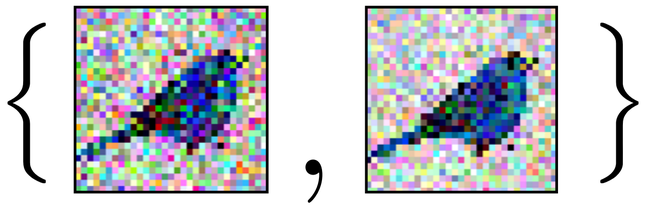
$\mathbf{x}$



Denoising

Training data

$\mathbf{x}_t$ $\mathbf{x}_{t-1}$

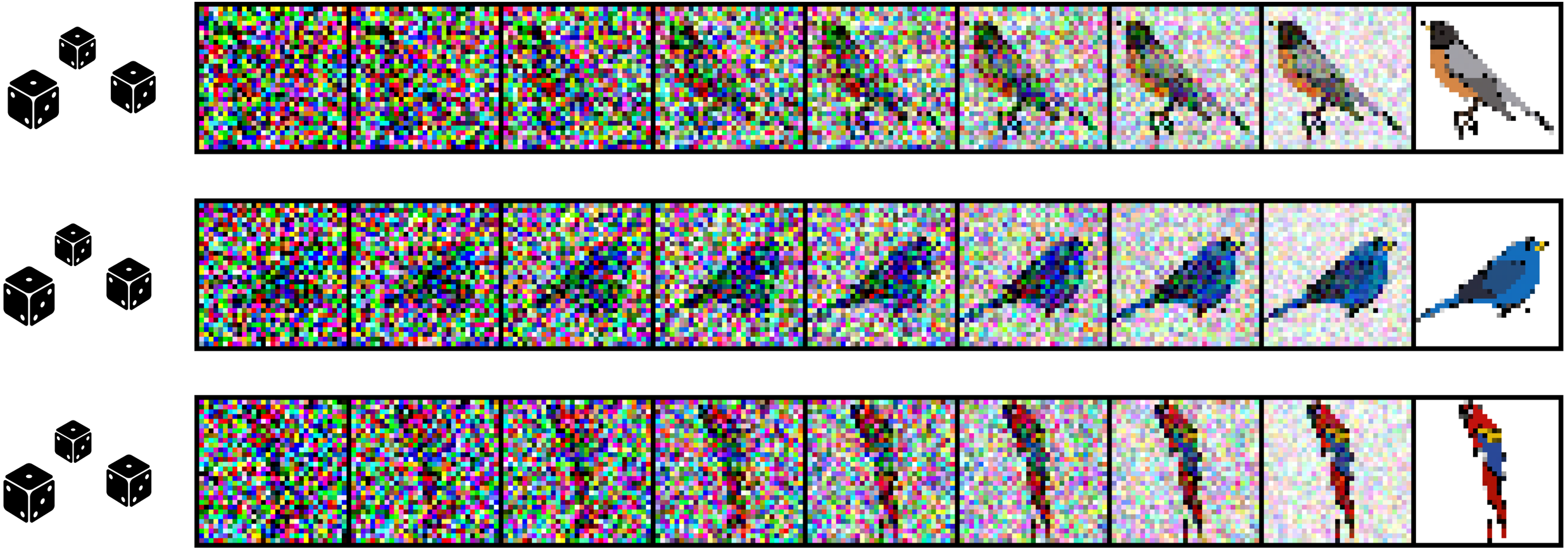


$$\arg\min_{f \in \mathcal{F}} \sum_{i=1}^{N} \mathcal{L}(f(\mathbf{x}_t), \mathbf{x}_{t-1})$$

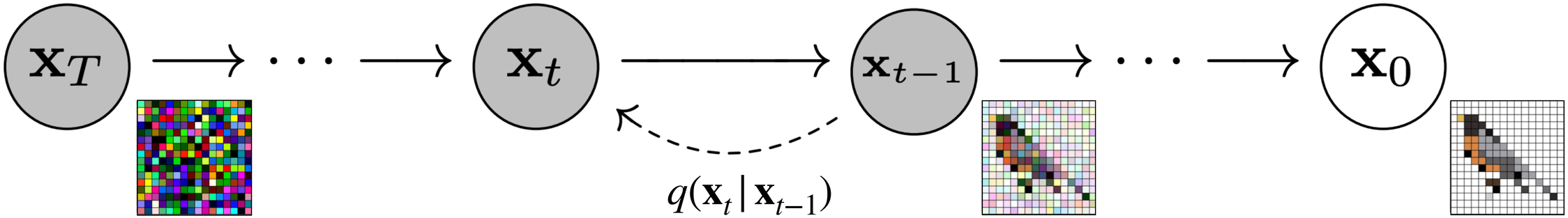
Converts generative modeling into a bunch of supervised prediction problems

Diffusion Generative Models

Different noise samples (dice rolls) result in different images



Diffusion Generative Models - Forward Diffusion



$$\epsilon \sim \mathcal{N}(\mathbf{O}, \mathbf{I})$$

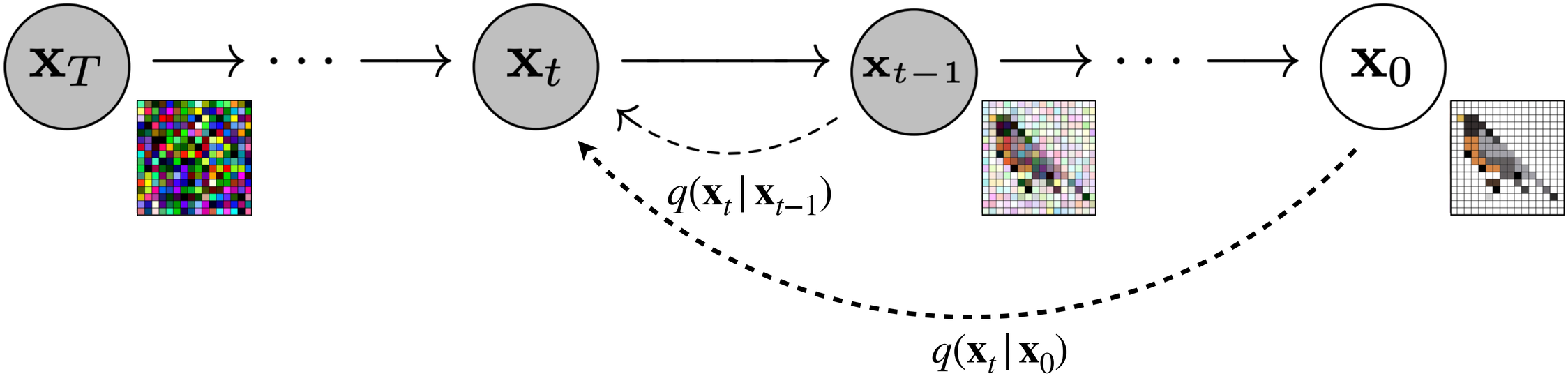
$$\mathbf{x}_t = \sqrt{(1-\beta_t)}\,\mathbf{x}_{t-1} + \sqrt{\beta_t}\,\epsilon$$

$$\beta_1, \ldots, \beta_T$$

variance schedule: controls how much noise is added at any timestep

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\,\mathbf{x}_{t-1}, \beta_t\,\mathbf{I})$$

Diffusion Generative Models - Forward Diffusion



$$\epsilon \sim \mathcal{N}(\mathbf{O}, \mathbf{I})$$

$$\mathbf{x}_t = \sqrt{(1-\beta_t)}\,\mathbf{x}_{t-1} + \sqrt{\beta_t}\,\epsilon$$

$$\beta_1, \ldots, \beta_T$$

variance schedule: controls how much noise is added at any timestep

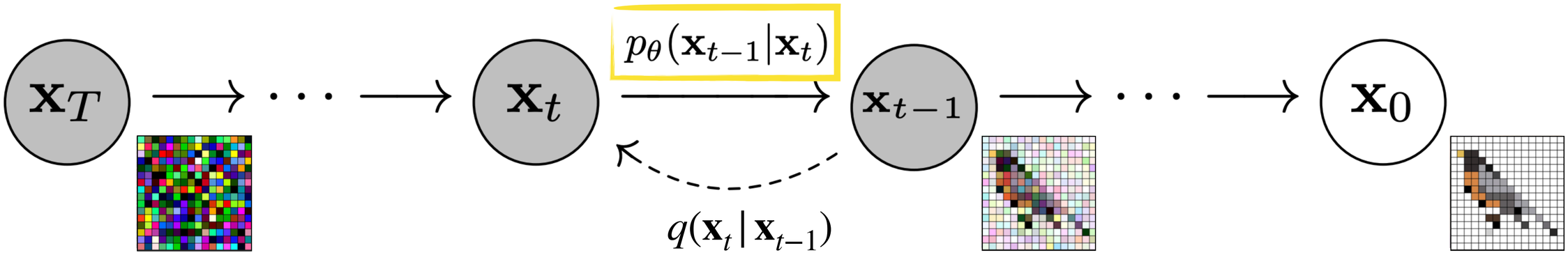
$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t ; \sqrt{1-\beta_t}\,\mathbf{x}_{t-1}, \beta_t\,\mathbf{I})$$

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t ; \sqrt{\bar{\alpha}_t}\,\mathbf{x}_{t-1}, (1-\alpha_t)\,\mathbf{I})$$

$$\alpha_t = 1 - \beta_t$$

$$\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$$

Diffusion Generative Models - Reverse Diffusion



$q(\mathbf{x}_t|\mathbf{x}_t)$ is known, $p(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is not

We use a network with parameters θ to learn $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$

$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_{t-1}, t),\ \sigma^2\mathbf{I})$ — Assumption!

$\mu_\theta = a\mathbf{x}_t - b\epsilon_\theta$
Predict $\epsilon_\theta = \text{net}(\mathbf{x_t}, t)$
Loss = $c\,||\epsilon_\theta - \epsilon||^2$

DDPM

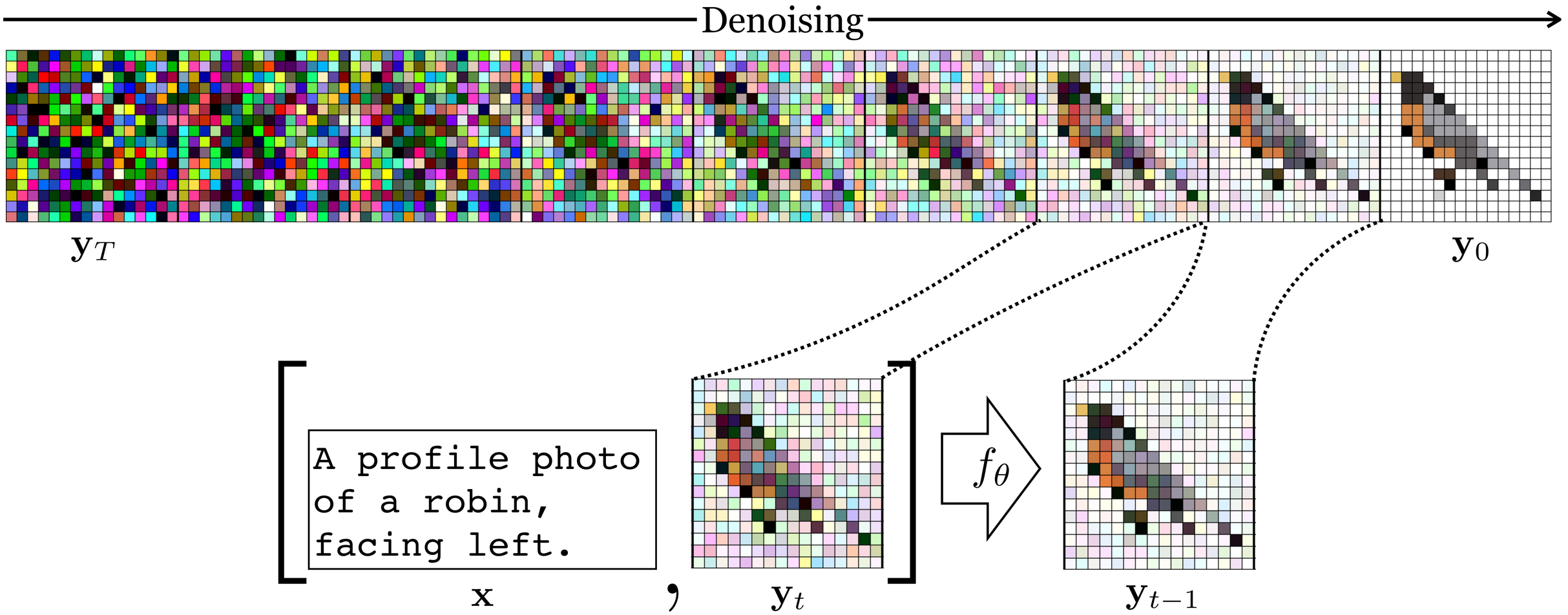
Algorithm 1 Training

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
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\underbrace{\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}}_{\mathbf{x_t}}, t) \right\|^2$$
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: $\mathbf{x}_{t-1} = \underbrace{\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right)}_{\mu_\theta} + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

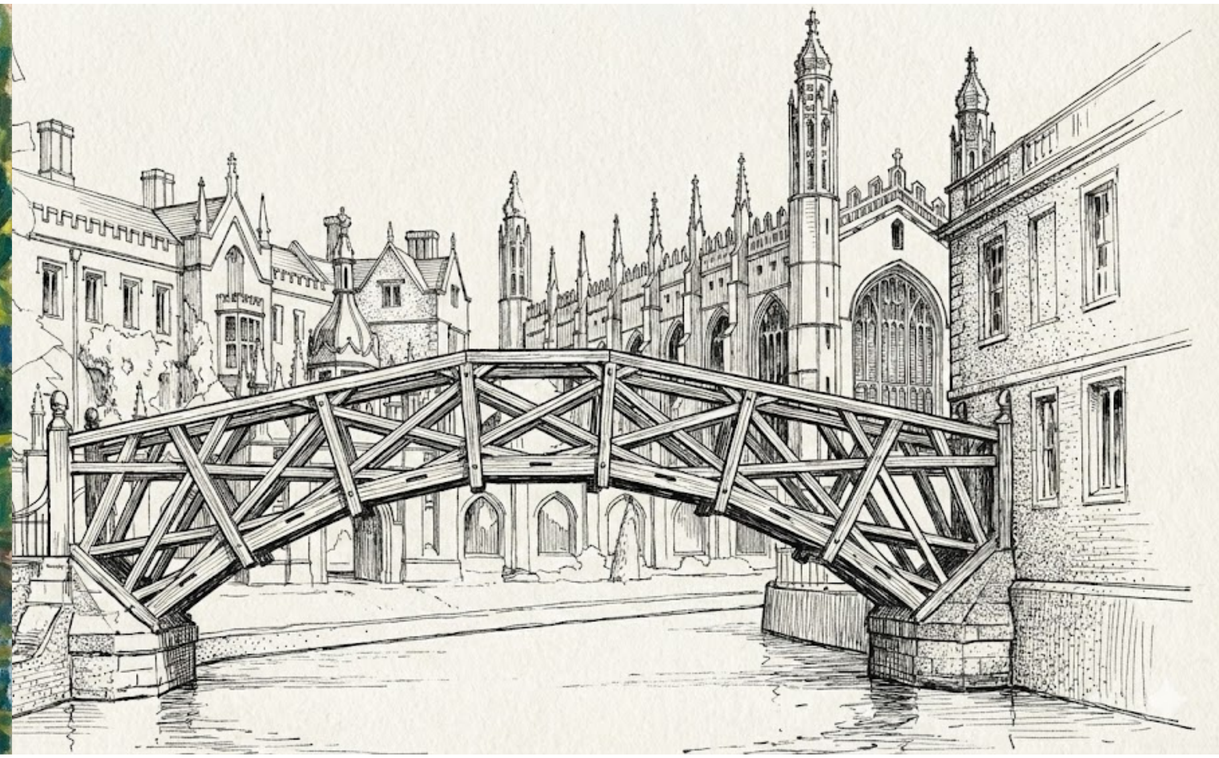
Conditional Diffusion Models



For example: DALL-E 2 [Ramesh et al. 2022], Stable Diffusion [Rombach*, Blattman* et al. 2022]

Conditional Diffusion Models

Artistic Renditions of Cambridge (Gemini)

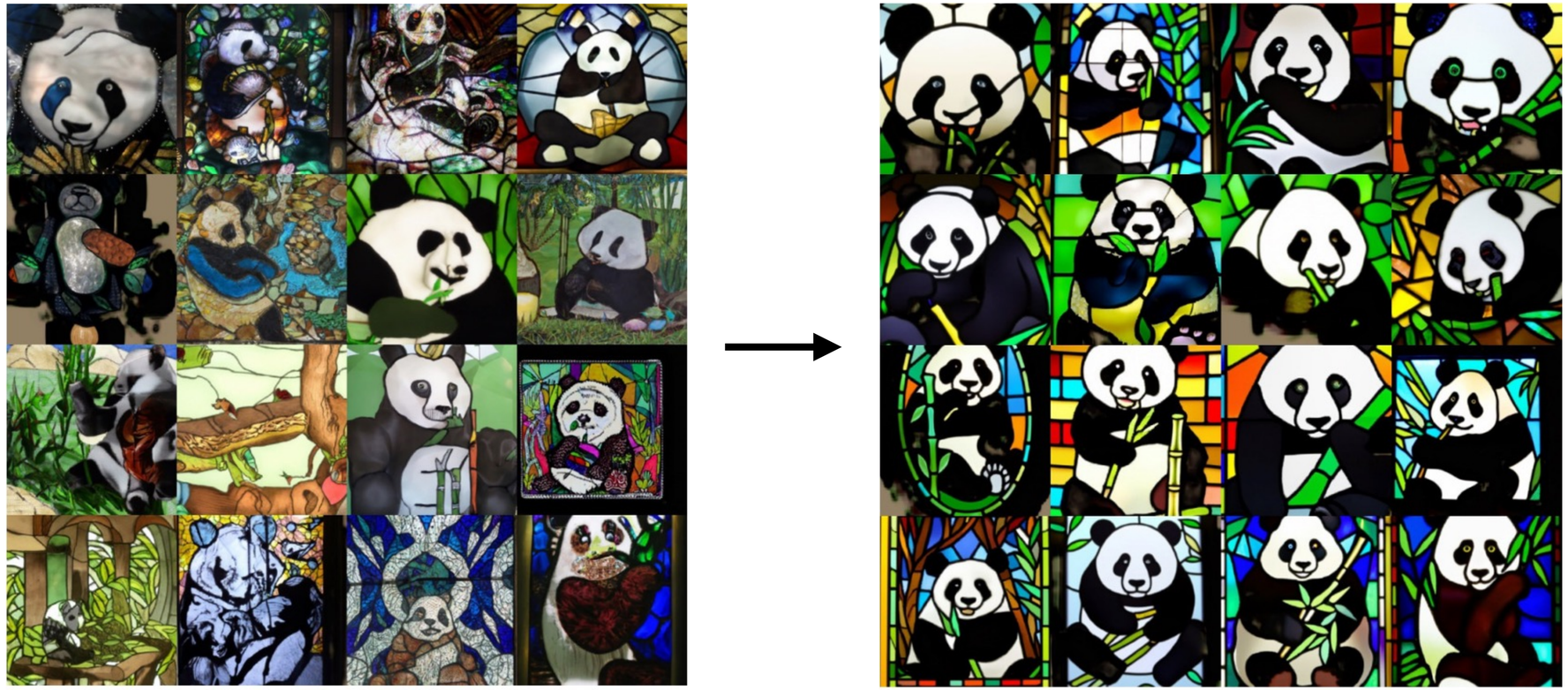


Classifier-free Guidance

Unconditional denoiser: $\epsilon_\theta(\mathbf{x}_t, t)$

Conditional denoiser: $\epsilon_\theta(\mathbf{x}_t, t, c)$, e.g., c = "A stained glass window of a panda eating."

Classifier-free guidance: $\tilde{\epsilon}_t = \epsilon_\theta(\mathbf{x}_t, t, c) + (1 - w)(\epsilon_\theta(\mathbf{x}_t, t, c) - \epsilon_\theta(\mathbf{x}_t, t))$



GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, Nichol et al., ICML 2022

Classifier-free Guidance

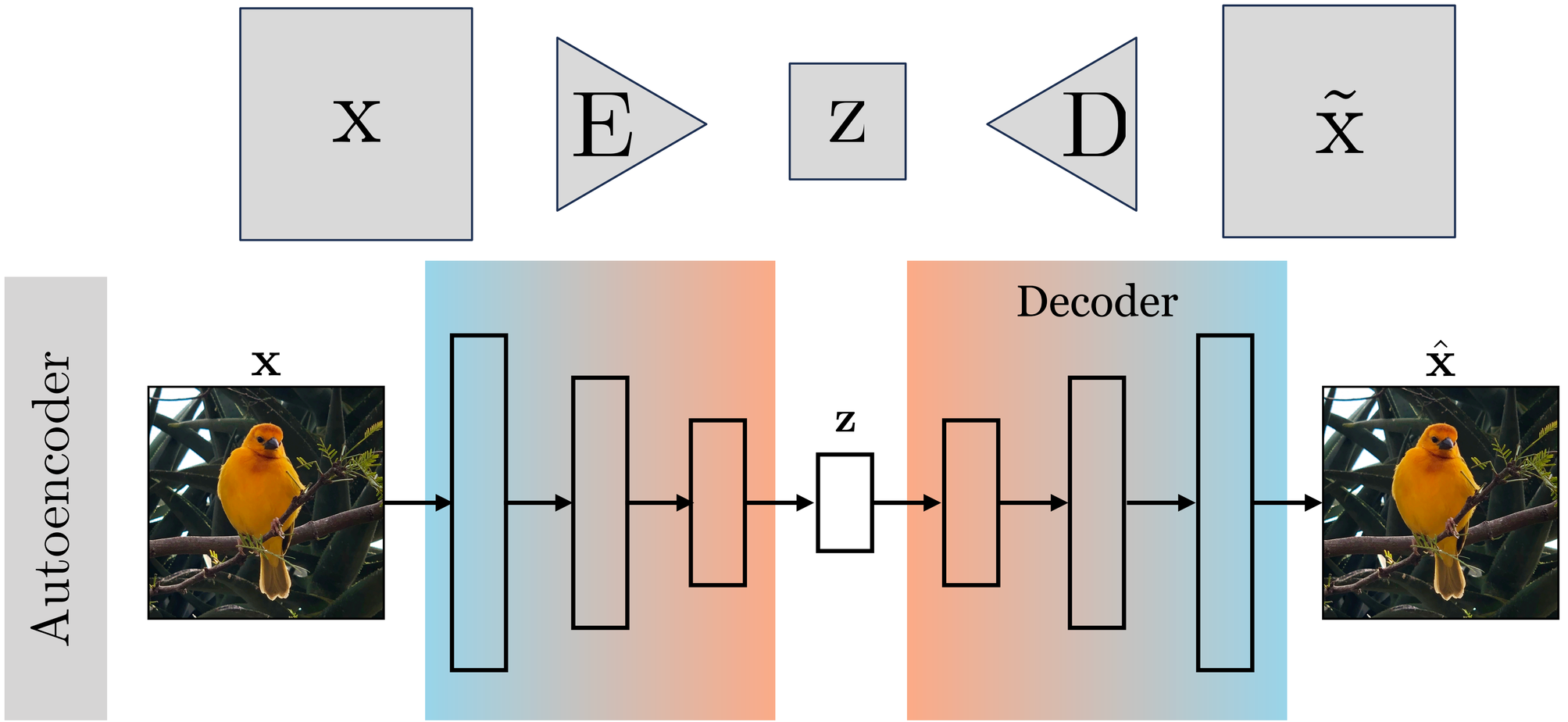
$$\tilde{\epsilon}_t = \epsilon_\theta(\mathbf{x}_t, t, c) + (1 - w)(\epsilon_\theta(\mathbf{x}_t, t, c) - \epsilon_\theta(\mathbf{x}_t, t))$$

Figure 2: The effect of guidance on a mixture of three Gaussians, each mixture component representing data conditioned on a class. The leftmost plot is the non-guided marginal density. Left to right are densities of mixtures of normalized guided conditionals with increasing guidance strength.

1. Train with both conditional and unconditional objectives
2. Evaluate denoising network with and without conditional input (2x cost)
3. Combine the results

Classifier-Free Diffusion Guidance [Ho and Salimans 2022]

Latent Diffusion



Images are very high dimensional. Latent spaces can compress the dimensions significantly.
Two stage training :(

Latent Diffusion

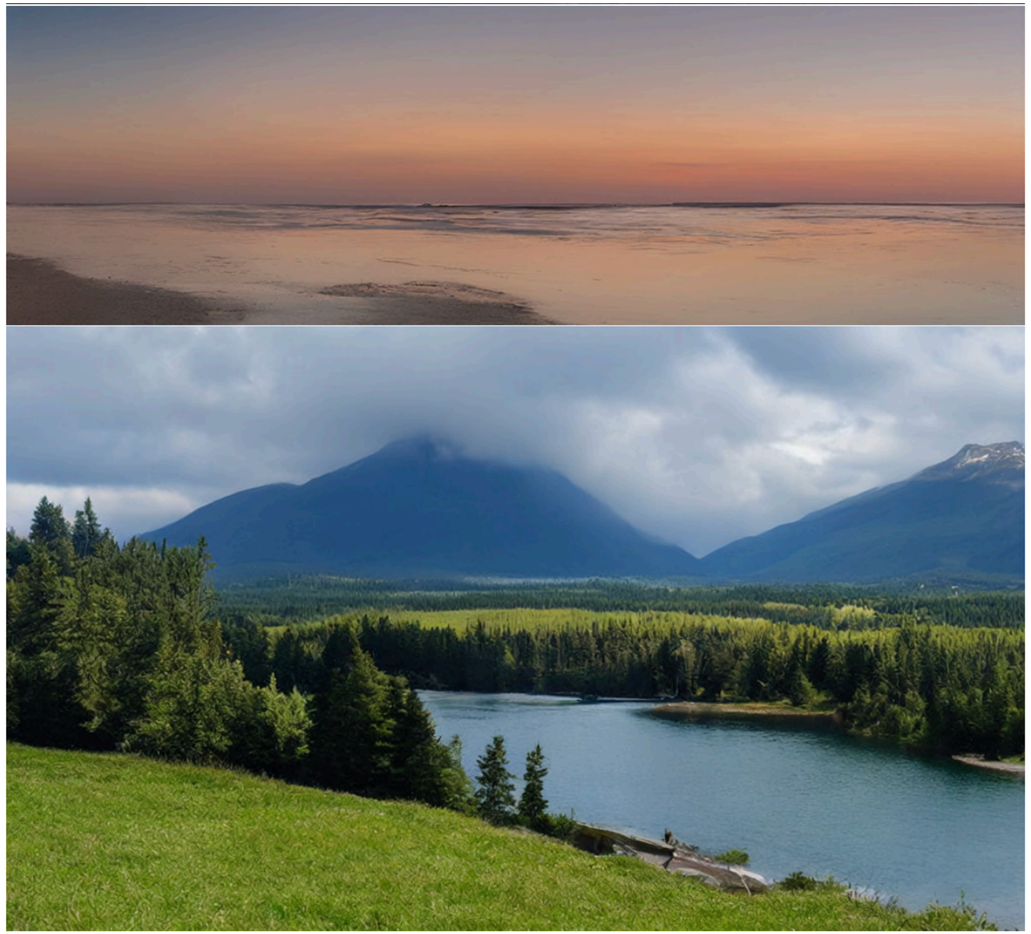
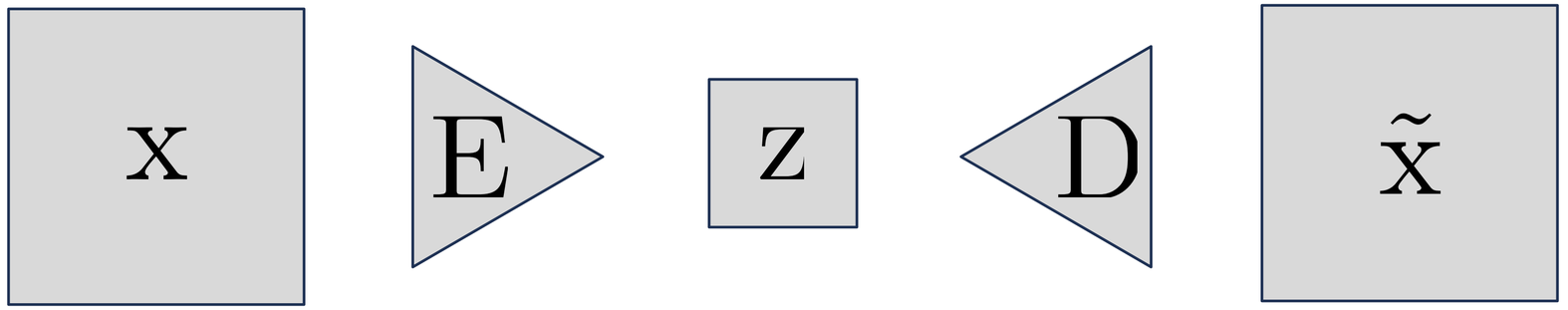
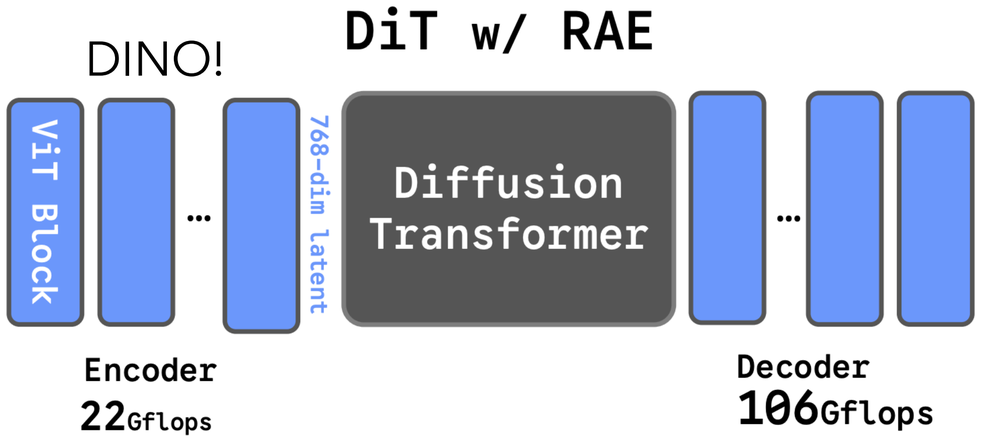


Figure 12. Convolutional samples from the semantic landscapes model as in Sec. 4.3.2, finetuned on 512^2 images.

High-Resolution Image Synthesis with Latent Diffusion Models. [Rombach et al., 2022]

Latent Diffusion

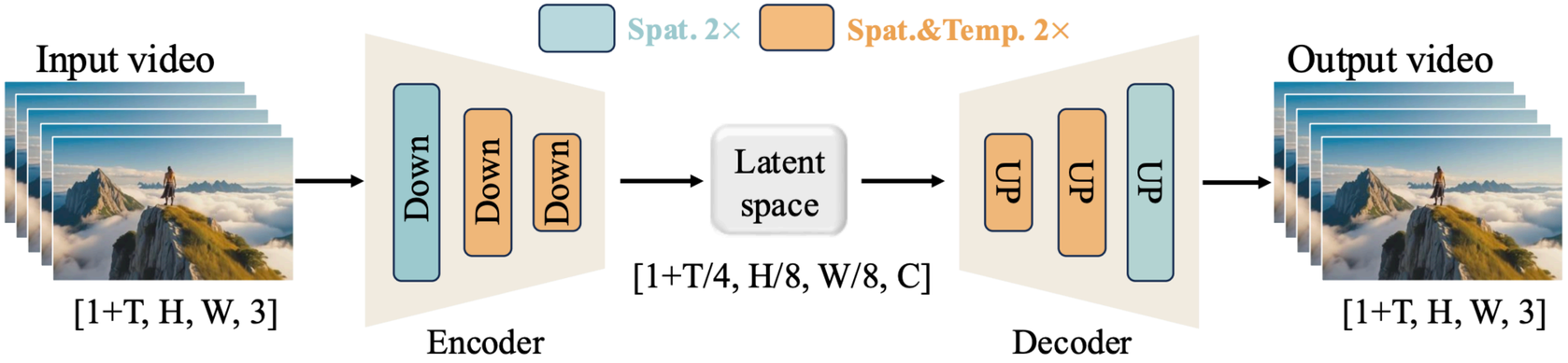




Latent spaces can be more “structured” and “semantic” — easier distribution to model.

Diffusion Transformers with Representation Autoencoders. [Zheng et al., 2025]

Video Diffusion - Wan



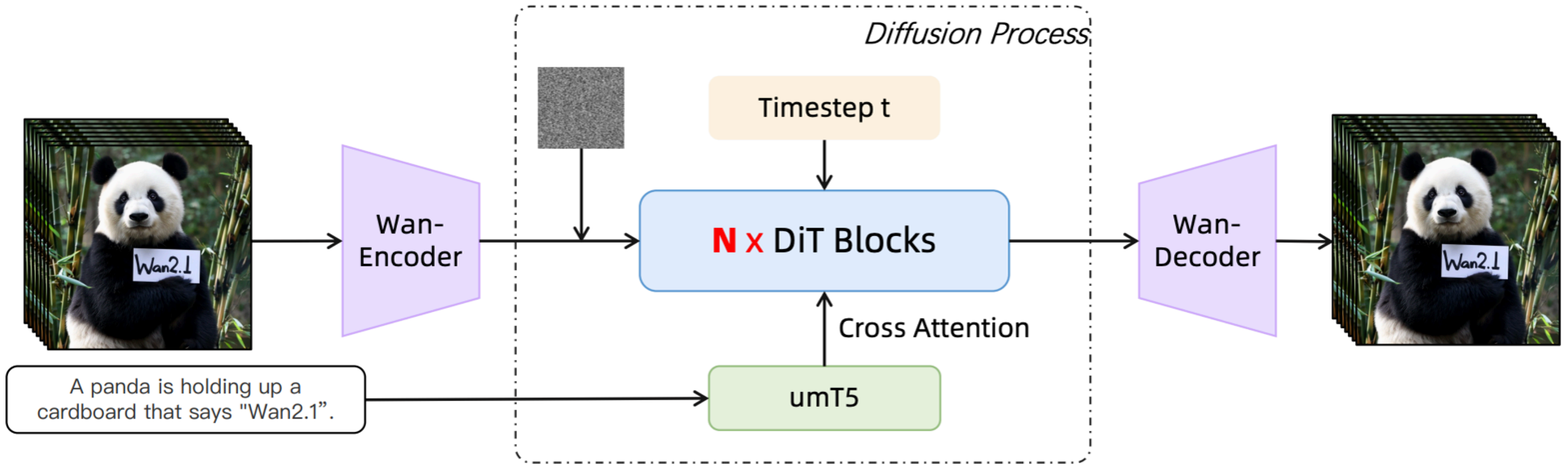


Figure 9: Architecture of the Wan.

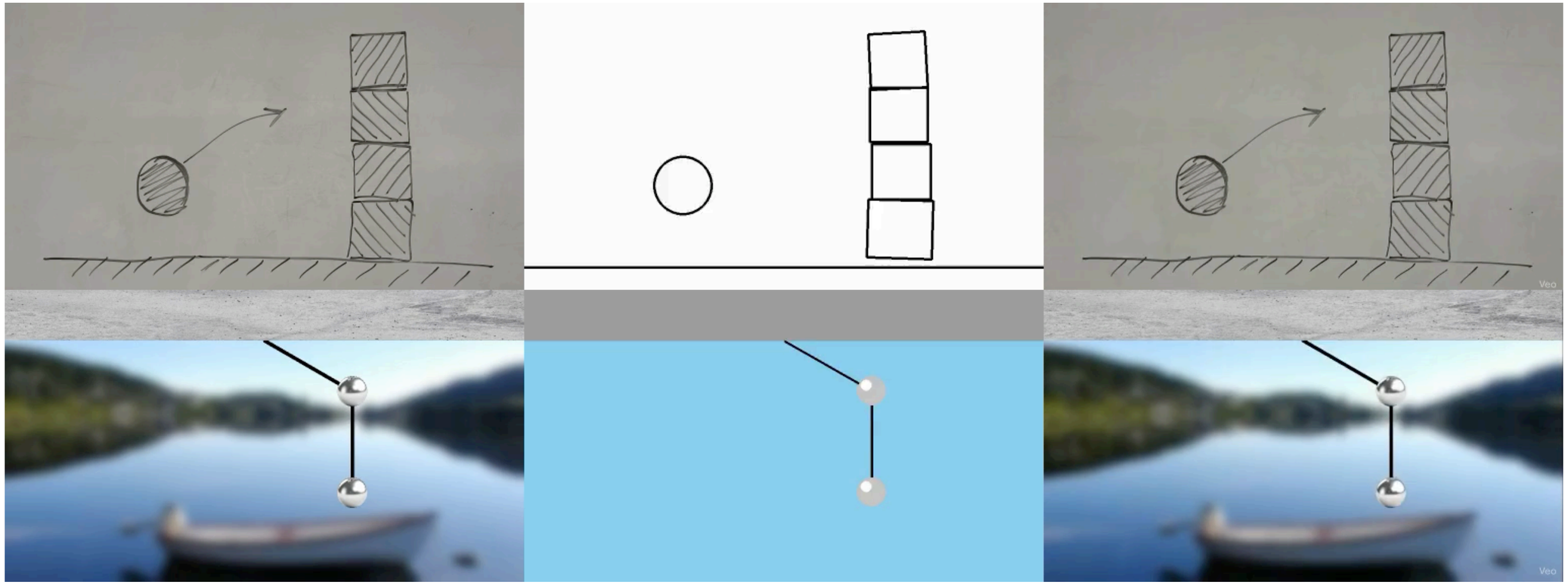
Wan: Open and Advanced Large-Scale Video Generative Models [Wan Team, 2025]

Veo3 Prompt: The scene explodes with the raw, visceral, and unpredictable energy of a hardcore off-road rally, captured with a dynamic, almost found-footage or

Veo3 Prompt: The scene explodes with the raw, visceral, and unpredictable energy of a hardcore off-road rally, captured with a dynamic, almost found-footage or

Generative Models

- Unbelievable progress in the last years
- Several open questions still remain
  - Training and inference are both very slow
  - Still far from perfect, especially for videos



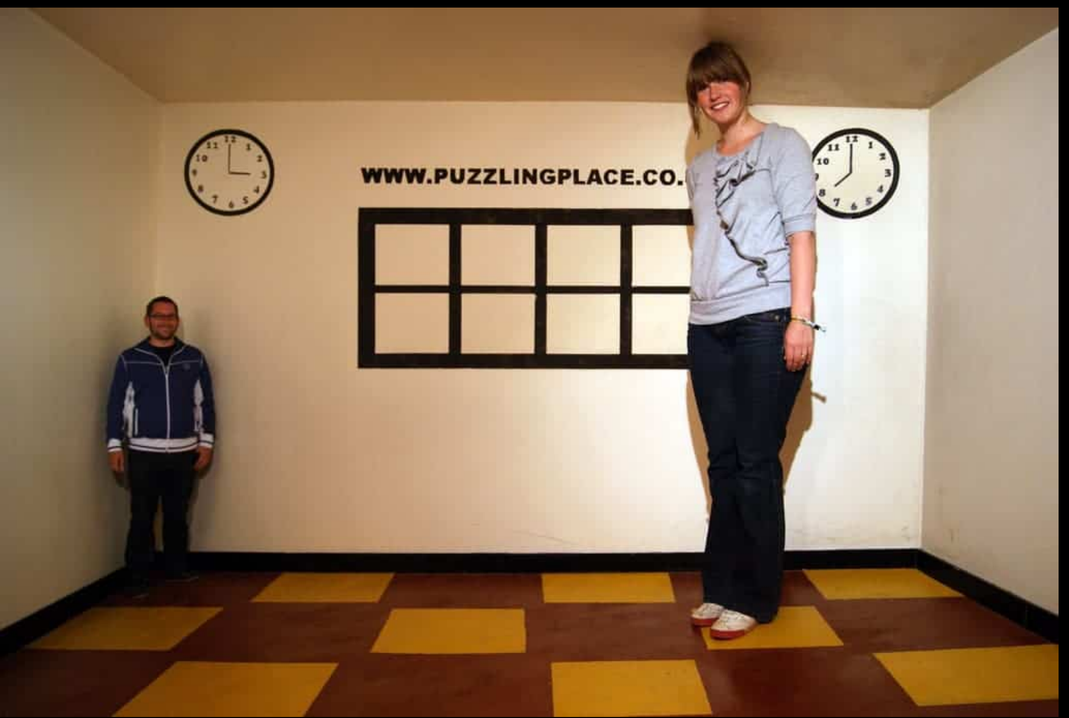
Generative Models

- Unbelievable progress in the last years
- Several open questions still remain
  - Training and inference are both very slow
  - Still far from perfect, especially for videos
  - How can they be used to solve other downstream tasks?

Advanced Computer Vision: Diffusion Models

MLMI17

Ayush Tewari



UNIVERSITY OF CAMBRIDGE