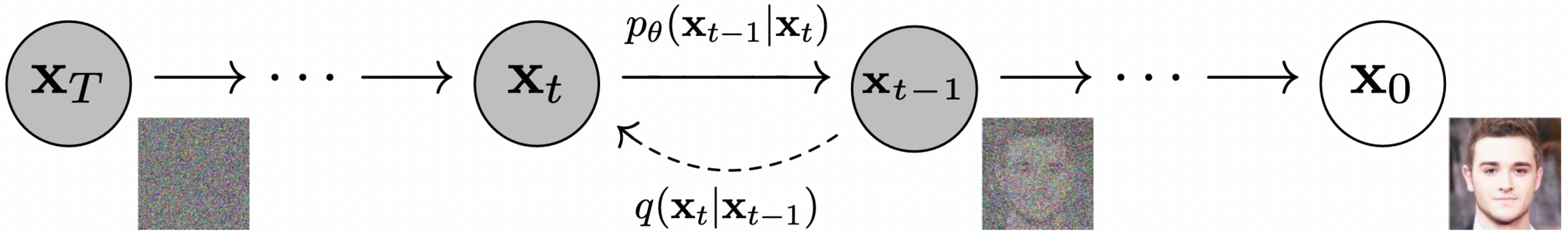
Latent diffusion

The meaning of word “latent” here

An image autoencoder is used to reduce dimensionality, for example, an image with shape [b, 3, 256, 256] is encoded to a latent representation with shape [b, 4, 32, 32].

Example

Forward and reverse process



Single step of forward process

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

$$\text{where} \quad 0 < \beta_1 < \beta_2 < \ldots < \beta_{T-1} < \beta_T < 1$$

$$\text{and} \quad \beta_1 = 10^{-4}, \quad \beta_T = 0.02, \quad T \sim 1000$$

Forward process

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

$$\text{where } \bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i \text{ and } \alpha_t = 1 - \beta_t,$$

$$\text{also note that } \bar{\alpha}_T \approx 0$$

Training and inference

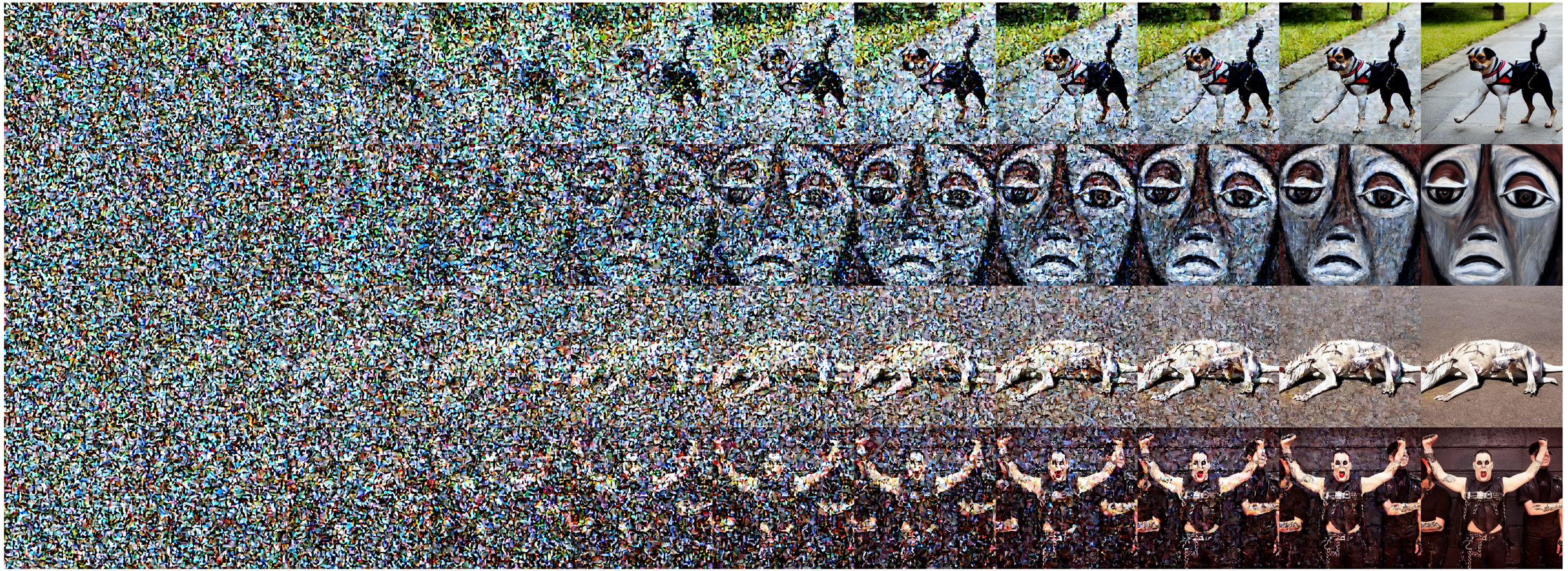
Algorithm 1 Training

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** $\mathbf{x}_0$

Example



Conditional model

$\boldsymbol{\epsilon}_\theta(x_t, t)$ change to $\boldsymbol{\epsilon}_\theta(x_t, t, y)$,

where $y -$ any conditional information

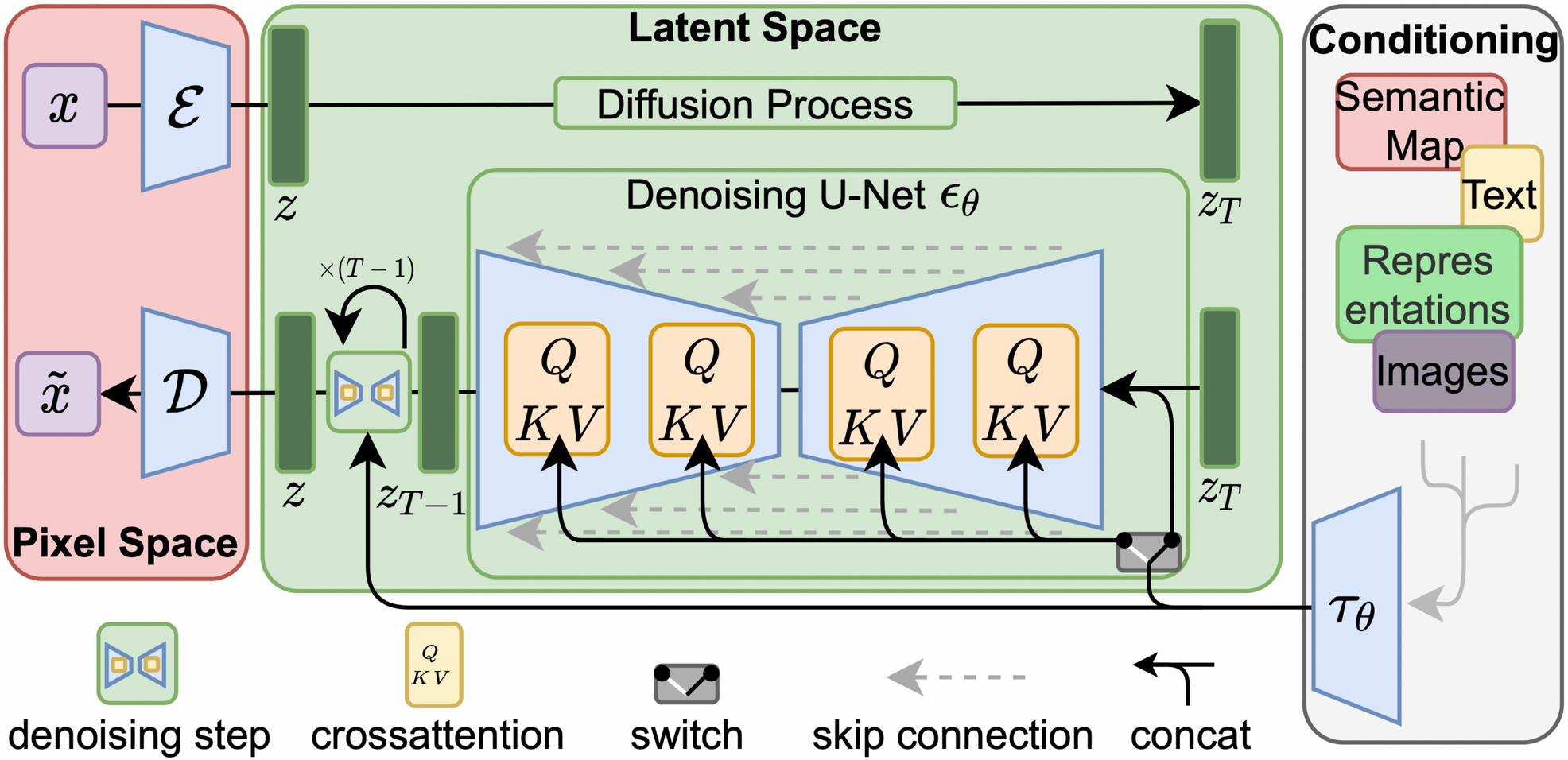
Classifier-free guidance

$$\tilde{\boldsymbol{\epsilon}}_\theta(x_t, t, y) = \boldsymbol{\epsilon}_\theta(x_t, t, y) + s \cdot (\boldsymbol{\epsilon}_\theta(x_t, t, y) - \boldsymbol{\epsilon}_\theta(x_t, t, \emptyset)),$$

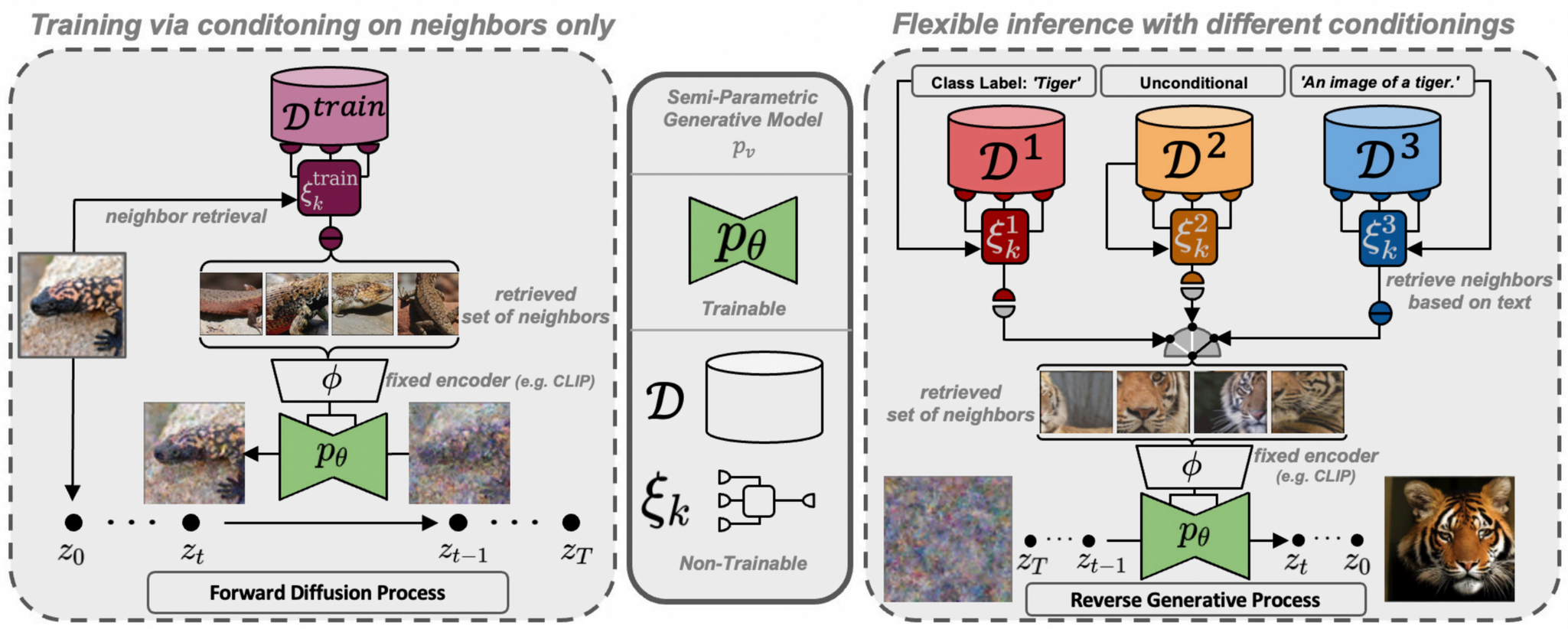
where $s \sim 3.0$ and $\emptyset -$ special "null" label

Example

Architecture



Retrieval-Augmented Diffusion Models



Stable Diffusion

Conditioning is on the last hidden state
of CLIP ViT-L/14 text encoder.
It has shape [b, 77, 768].

Finetuning on FFHQ-512

Finetuning on FFHQ-512 with masks