

Outline

① Công trình liên quan

② Mô hình sinh cử chỉ

Các phương pháp

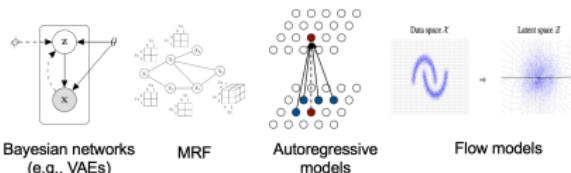
Rule Base & Statistic

- BEAT, Rule-based generation.
- Gesture Controllers

Deep Learning

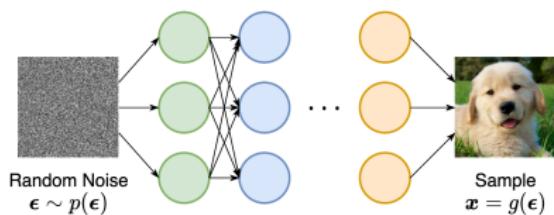
Likelihood-based models:

- **MLP**: Gesticulator
- **RNN**: Speech2AffectiveGestures, HA2G, TransGesture, ..
- **Normalising Flows**:
Text2gestures, Speech2Gesture
- **VAE/VQ-VAE**: Audio2Gestures;



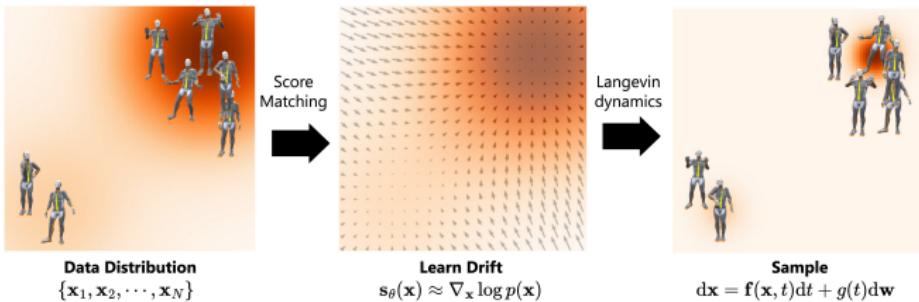
Implicit generative models:

- **GAN**: DiffGAN
- **Diffusion**: Listen denoise action, DiffuseStyleGesture, Taming Diffusion Models.



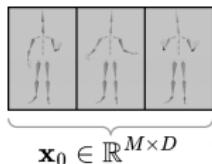
Tư tưởng cơ bản của Diffusion

Dataset $\mathcal{G} = \{\mathbf{x}_i^{M \times D}\}_1^n$, ta chuẩn hoá dữ liệu $\mathbf{x}_i^{M \times D} = \frac{\mathbf{x}_i^{M \times D} - \mu}{\sigma}$



Growth Truth Distribution

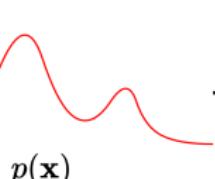
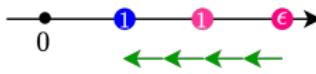
$$\mathbf{x}_0 = q(\mathbf{x})$$



$$\mathbf{x}_0 \in \mathbb{R}^{M \times D}$$

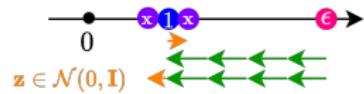
Training (Denoise)

$$\text{Drift } s_\theta(\mathbf{x}) \approx \nabla_{\mathbf{x}} \log p(\mathbf{x})$$



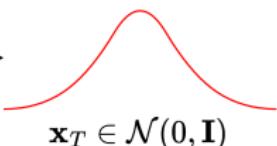
Sampling (Denoise)

$$\mathbf{x}_{\text{sample}} = s_\theta + \sigma_t \cdot \mathbf{z}$$



$$1 \rightarrow T$$

Diffuse



$$f_\theta$$

Công trình liên quan

Ký hiệu

Tham số: Đang huấn luyện θ , đã huấn luyện xong: θ' , $\hat{\mathbf{x}}$: dự đoán

Phân phối chuẩn

$$\mathcal{N}(\textcolor{red}{a}\mathbf{x}, \textcolor{blue}{b}^2)$$

- Một hàm $f(x) = ax + b\epsilon$ với $\epsilon \in \mathcal{N}(0, 1)$ được ký hiệu là $f(x) \sim \mathcal{N}(ax, b^2)$
- Trung bình: $\mu = \textcolor{red}{a}x = \frac{1}{n} \sum_{i=1}^n x_i$
- Phương sai: $\sigma^2 = \textcolor{blue}{b}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$

Xác xuất có điều kiện

$$p(\textcolor{green}{x}|\textcolor{orange}{y})$$

- $p(\textcolor{green}{x}|\textcolor{orange}{y})$ là xác xuất có điều kiện.
- y : xảy ra trước (bên phải)
- x : xảy ra sau y (bên trái)

Đặc điểm của việc học dữ liệu Motion

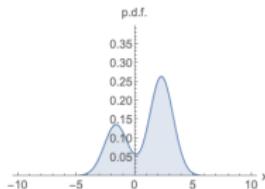
Quan hệ giữa dữ liệu cử chỉ và âm thanh:

- Một đoạn cử có thể bao gồm nhiều âm thanh.
- Mỗi âm thanh có thể tương ứng với nhiều đoạn cử chỉ khác nhau.

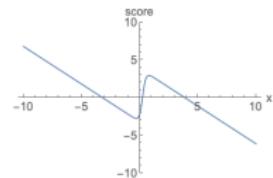
Khó khăn

- Dữ liệu ít, chi phí cao
- Thiếu nhãn và dữ liệu tương ứng giữa âm thanh, cử chỉ.
- Quá trình sinh có thể dễ điều khiển

liệu



Phải chuẩn hoá (diện tích dưới đường cong phải tích phân thành một)

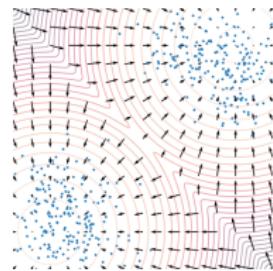


Không cần chuẩn hoá.

$p(\mathbf{x})$
probability density



$\nabla_{\mathbf{x}} \log p(\mathbf{x})$
score function



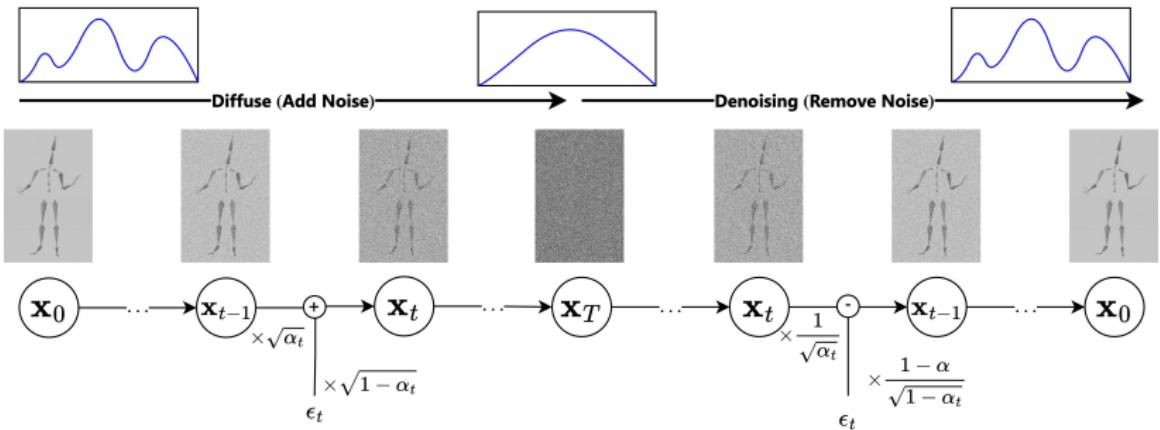
score function vs
probability density

Diffuse và Denoise thông thường

Diffuse: Cho $\{\alpha_t \in (0, 1)\}_{t=1}^T$ và $\alpha_1 > \alpha_2 > \dots > \alpha_T$.

Với nhiễu: $\epsilon_0, \epsilon_1, \dots, \epsilon_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$; $\epsilon_i \neq \epsilon_j$ ($\forall i, j \in T$), nhiễu ϵ_t cố định.

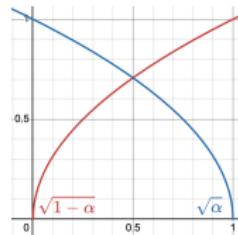
$$\mathbf{x}_t = \sqrt{\alpha_t} \cdot \mathbf{x}_{t-1} + \sqrt{1 - \alpha_t} \cdot \epsilon \quad (1)$$



Denoise

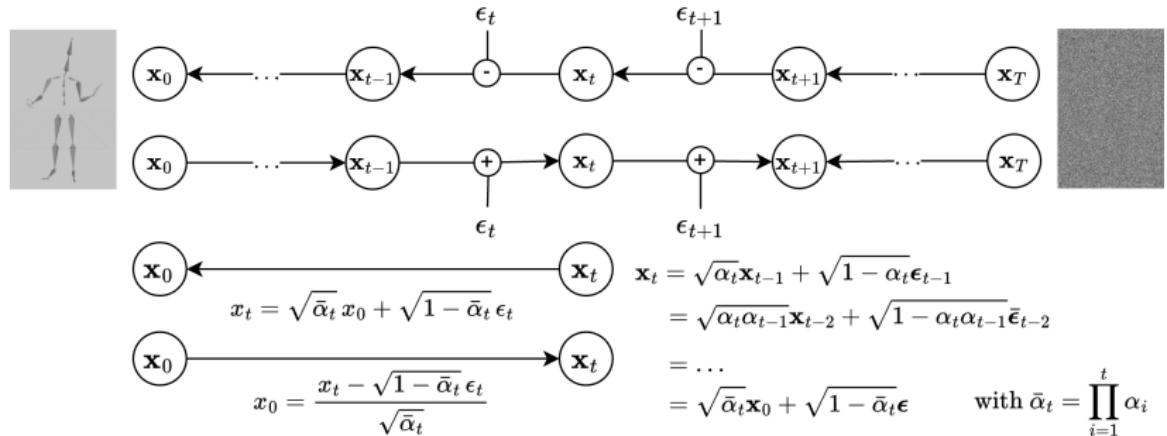
$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} (\mathbf{x}_t - \sqrt{1 - \alpha_t} \cdot \epsilon) \quad (2)$$

Công trình liên quan



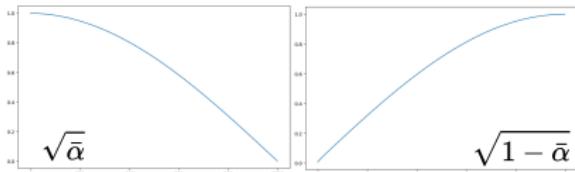
Quan hệ \mathbf{x}_t và \mathbf{x}_{t-1}

Ta có thể suy ra \mathbf{x}_t từ \mathbf{x}_0 và ngược lại. Với $\epsilon_{t-1}, \epsilon_{t-2}, \dots \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$



- $\bar{\alpha}_1 > \dots > \bar{\alpha}_T$
- Tổng hai nhiễu cũng là nhiễu:

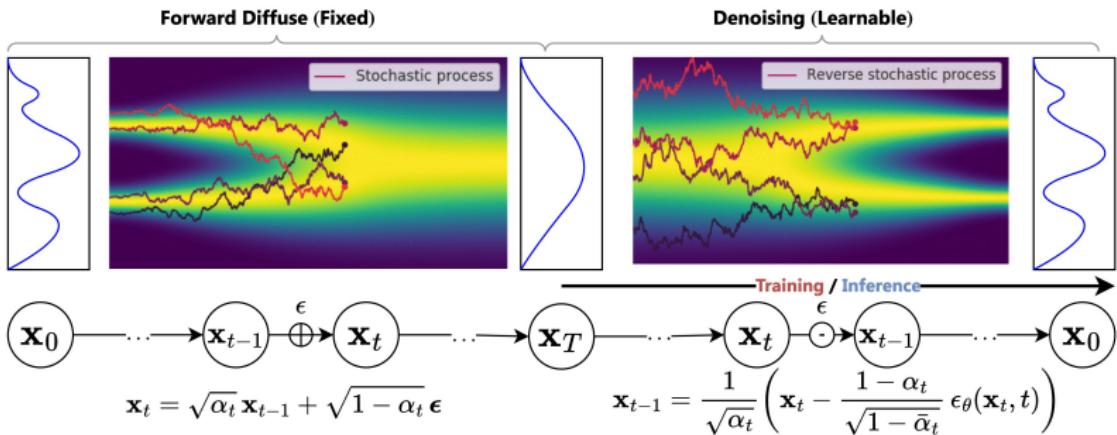
$$\mathcal{N}(\mathbf{0}, \sigma_1^2 \mathbf{I}) + \mathcal{N}(\mathbf{0}, \sigma_2^2 \mathbf{I}) = \mathcal{N}(\mathbf{0}, (\sigma_1^2 + \sigma_2^2) \mathbf{I})$$



$q(\mathbf{x}_t | \mathbf{x}_{t-1})$ và $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)$ trong DDPM

DDPM (Denoising Diffusion Probabilistic Model)

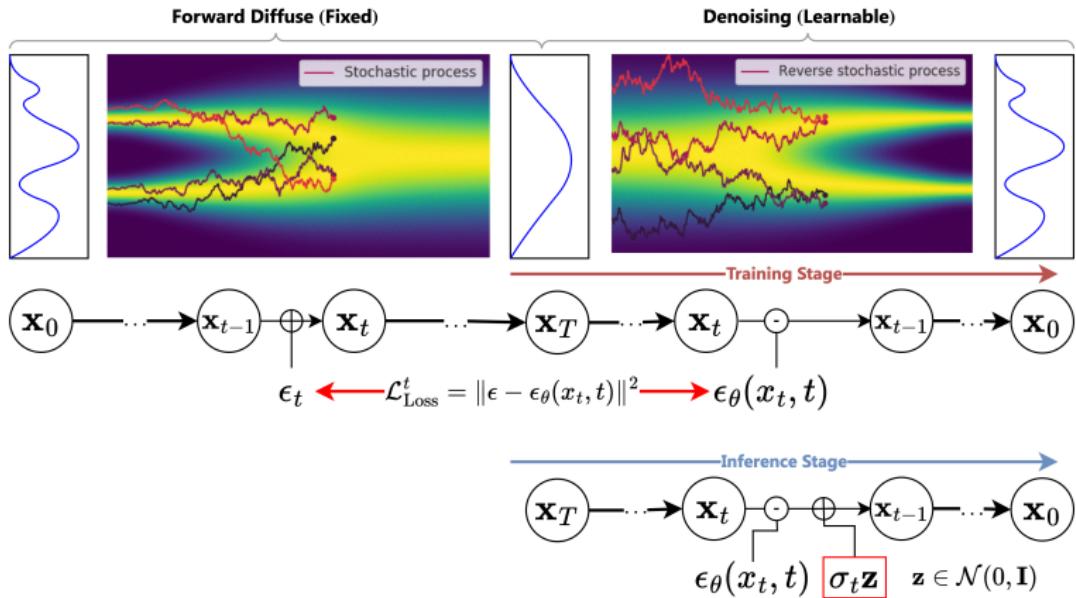
$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t} \mathbf{x}_{t-1}, (1 - \alpha_t) \mathbf{I}) \rightarrow p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)) \rightarrow$$



- $q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t} \mathbf{x}_{t-1}, (1 - \alpha_t) \mathbf{I})$
- $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t))$

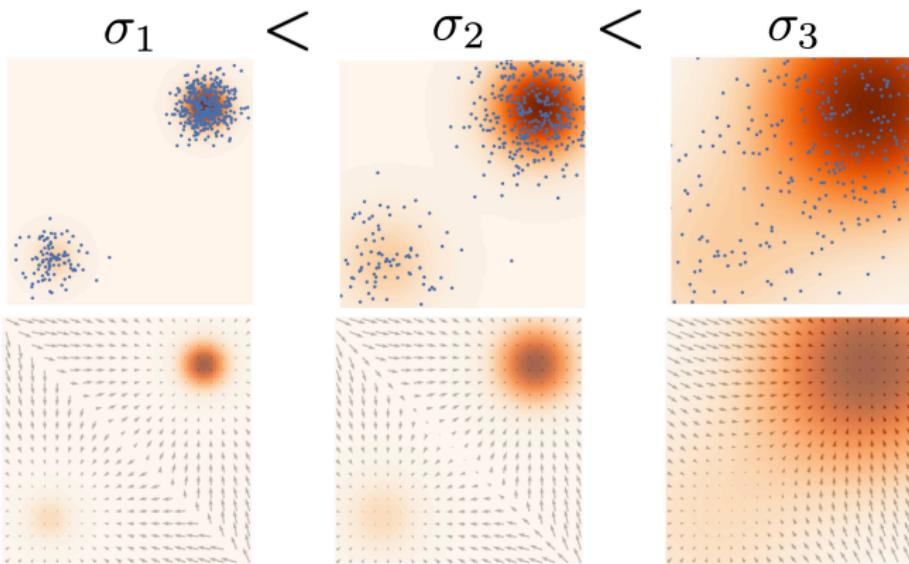
Vanilla Diffusion với ϵ Objective

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t} \mathbf{x}_{t-1}, (1 - \alpha_t) \mathbf{I}) \rightarrow p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)) \rightarrow$$



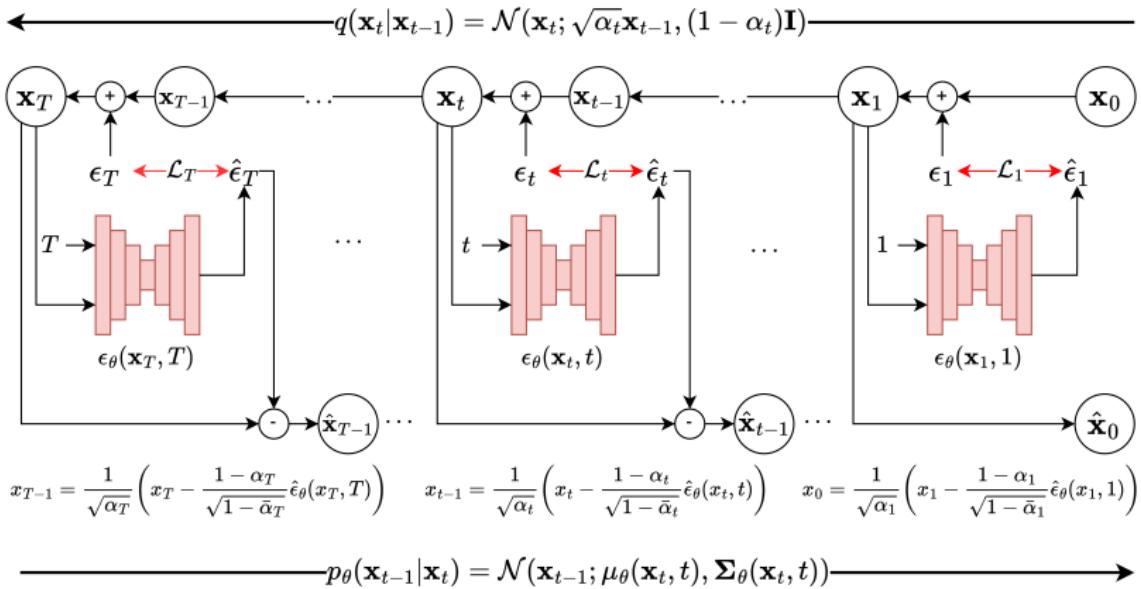
Điều khiển σ_t bằng với Langevin dynamics

Trong quá trình Denoise ($T \rightarrow 0$) $\sigma_T > \dots > \sigma_2 > \sigma_1$, ta sẽ giảm σ_t để giảm dần nhiễu, để mô hình có thể hội tụ ở những vùng có mật độ xác xuất cao.



Training DDPM

Hàm loss: $\mathcal{L} = \sum_{t=1}^T \mathcal{L}_t$. $\mathcal{L}_t = \mathbb{E}_{\mathbf{x}_0, \epsilon_t \sim \mathcal{N}(0, I), t} [\|\epsilon_t - \epsilon_\theta(\mathbf{x}_t, t)\|^2]$



Các bước huấn luyện với DDPM

- ❶ Tính sẵn các giá trị $\sqrt{\alpha_t}$, $\sqrt{1 - \alpha_t}$ và $\sqrt{\bar{\alpha}_t}$ ở mọi bước $t : 1 \rightarrow T$.
 $\{\alpha_t \in (0, 1)\}_{t=1}^T$, $\alpha_1 < \alpha_2 < \dots < \alpha_T$
- ❷ Lấy nhãn x_0 từ phân bố của dữ liệu đã chuẩn hóa
- ❸ Random nhiễu ϵ_t ở mọi bước $t : 1 \rightarrow T$, với $\forall t : \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
- ❹ Gây nhiễu (forward) x_0 để thu được x_t ở mọi bước $t : 1 \rightarrow T$

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}_t$$

- ❺ for all t , lấy t **ngẫu nhiên** $t \sim [1, T]$
- ❻ Cho \mathbf{x}_t và t vào mô hình để dự đoán nhiễu $\hat{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$
- ❼ Đạo hàm để cập nhật trọng số $\nabla_{\theta_t} \|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2$

$$\mathcal{L}_t = \mathbb{E}_{t \sim [1, T], \mathbf{x}_0, \boldsymbol{\epsilon}_t} \left[\|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}_t, t)\|^2 \right]$$

- ❽ Quay lại bước 6 cho đến khi hội tụ để thu được θ'

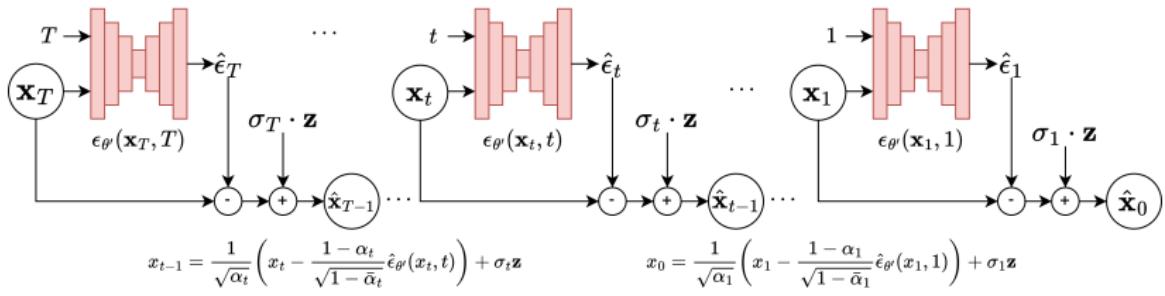
Sampling DDPM

Hàm sampling:

- $\mathbf{x}_T \in \mathcal{N}(0, \mathbf{I})$

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \hat{\epsilon}_{\theta'}(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$$

- Diffusion: $\mathcal{L}_{\text{loss}} = \mathbb{E}_{\mathbf{x}_0, \epsilon_t \sim \mathcal{N}(0, I), t} [\|\epsilon_t - \epsilon_{\theta}(\mathbf{x}_t, t)\|^2]$



$$p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_{\theta}(\mathbf{x}_t, t), \Sigma_{\theta}(\mathbf{x}_t, t)) \longrightarrow$$

Các bước lấy mẫu với DDPM

- ① Bắt đầu với nhiễu: $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$
- ② Các giá trị $\sqrt{\alpha_t}$, $\sqrt{1 - \alpha_t}$ và $\sqrt{\bar{\alpha}_t}$ có được từ bước huấn luyện
- ③ Tính hệ số điều chỉnh nhiễu σ_t từ α_t ở mọi bước $t: 1 \rightarrow T$
$$\sigma_t = \sqrt{\frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}}(1 - \alpha_t)$$
- ④ for all t , lấy t **tuần tự** $t \sim [T, \dots, 1]$
- ⑤ Random nhiễu $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$
- ⑥ Đưa \mathbf{x}_t vào để suy luận nhiễu $\epsilon_{\theta'} = \epsilon_{\theta'}(\mathbf{x}_t, t)$
- ⑦ Dùng nhiễu dự đoán để trừ đi \mathbf{x}_t ở bước t

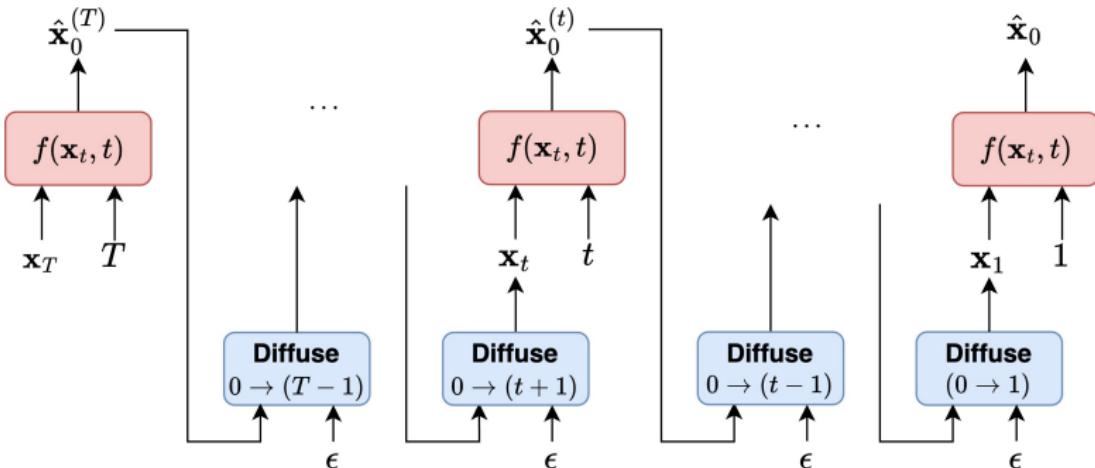
$$\mu = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_{\theta'}(\mathbf{x}_t, t) \right)$$

- ⑧ Cộng thêm một lượng nhiễu $\hat{\mathbf{x}}_{t-1} = \mu + \sigma_t \mathbf{z}$
- ⑨ Khi $t = 1$ ta thu được $\hat{\mathbf{x}}_0$ từ quá trình khử nhiễu

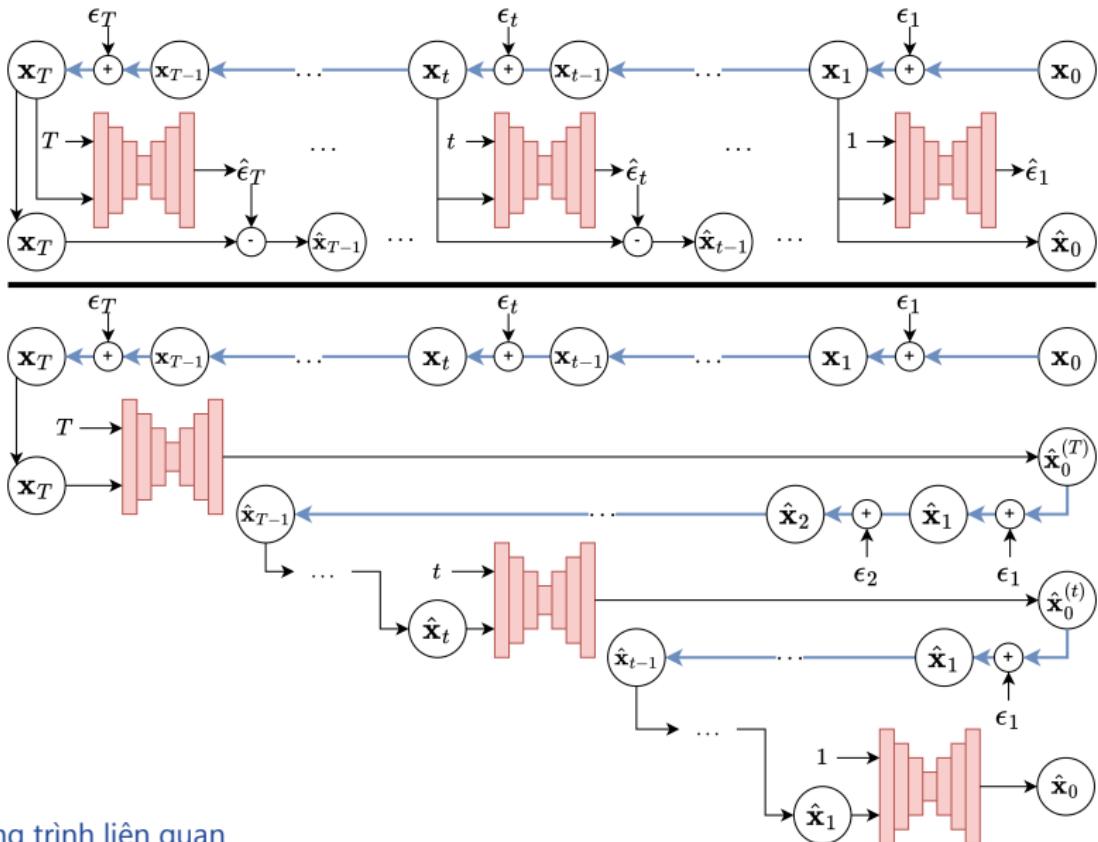
Cải tiến của với x_0 Objective (DALLE-2)

Những điểm cải tiến của x_0 so với ϵ objective

- Thay vì hàm $f_\theta(\mathbf{x}_t, t)$ dự đoán nhiễu ϵ_t thì $f_\theta(\mathbf{x}_t, t)$ dự đoán \mathbf{x}_0 .
- Sau khi có \mathbf{x}_0 thì ta thêm nhiễu ϵ đã có từ trước (từ forward process) để được $\hat{\mathbf{x}}_{t-1}$
- Tiếp tục cho đến khi được $\hat{\mathbf{x}}_0$



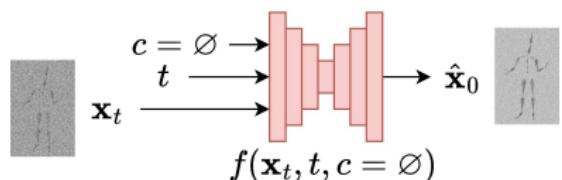
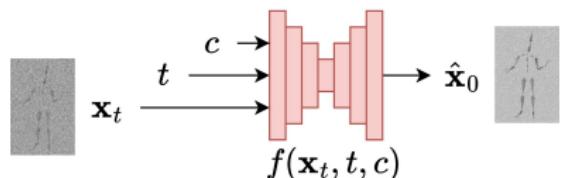
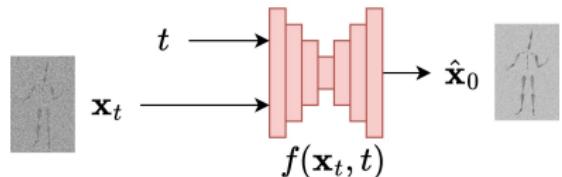
So sánh ϵ objective và x_0 objective



Công trình liên quan

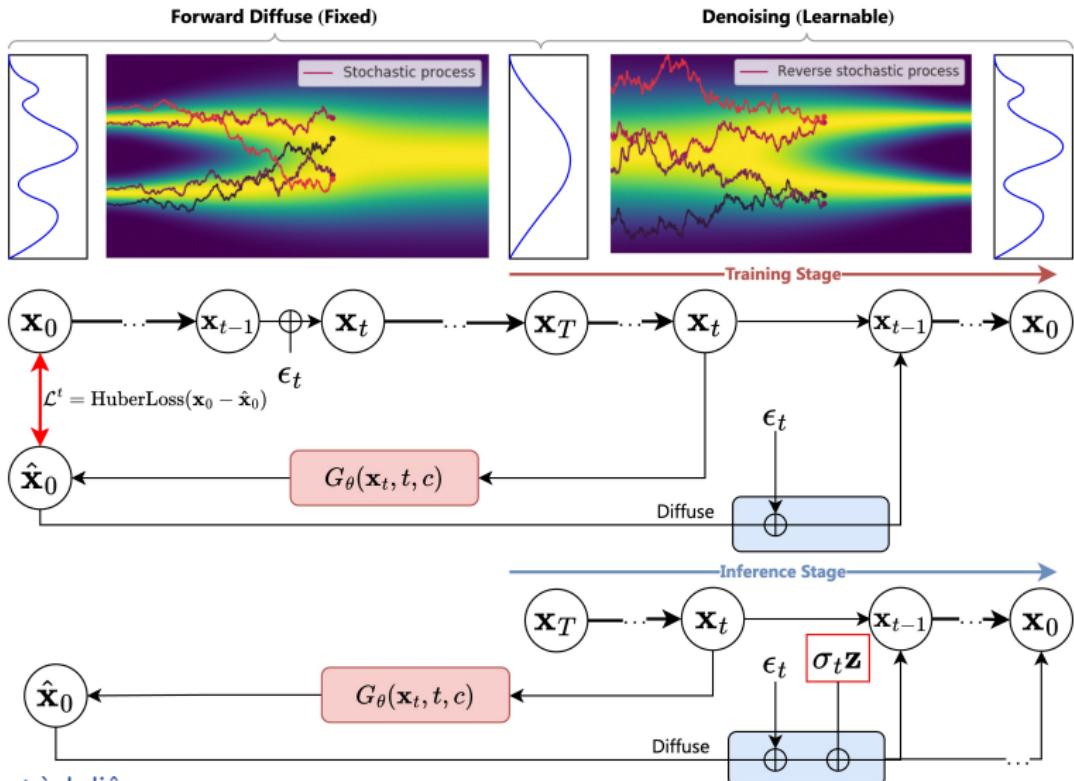
Classifier-Free Guidance

Để có thể học được có điều kiện
(condition c):



Condition Diffusion với \mathbf{x}_0 Objective

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t} \mathbf{x}_{t-1}, (1 - \alpha_t) \mathbf{I}) \rightarrow p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t, c), \Sigma_\theta(\mathbf{x}_t, t, c)) \rightarrow$$



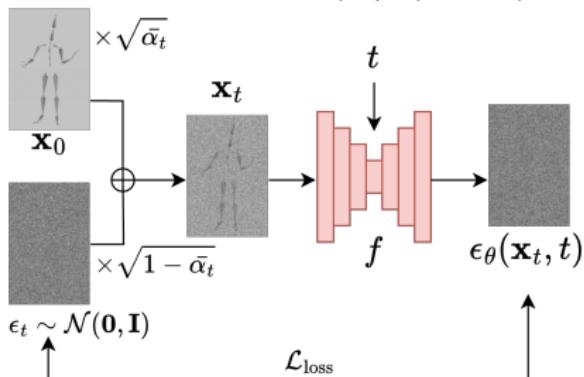
Công trình liên quan

Forward Diffusion Process

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

$$\begin{aligned}\mathbf{x}_t &= \sqrt{\alpha_t} \mathbf{x}_{t-1} + \sqrt{1 - \alpha_t} \boldsymbol{\epsilon}_{t-1} \\ &= \sqrt{\alpha_t \alpha_{t-1}} \mathbf{x}_{t-2} + \sqrt{1 - \alpha_t \alpha_{t-1}} \bar{\boldsymbol{\epsilon}}_{t-2} \\ &= \dots \\ &= \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}\end{aligned}$$

$$\begin{aligned}q(\mathbf{x}_t | \mathbf{x}_{t-1}) &= \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t} \mathbf{x}_{t-1}, (1 - \alpha_t) \mathbf{I}) \\ \rightarrow q(\mathbf{x}_t | \mathbf{x}_0) &= \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I})\end{aligned}$$



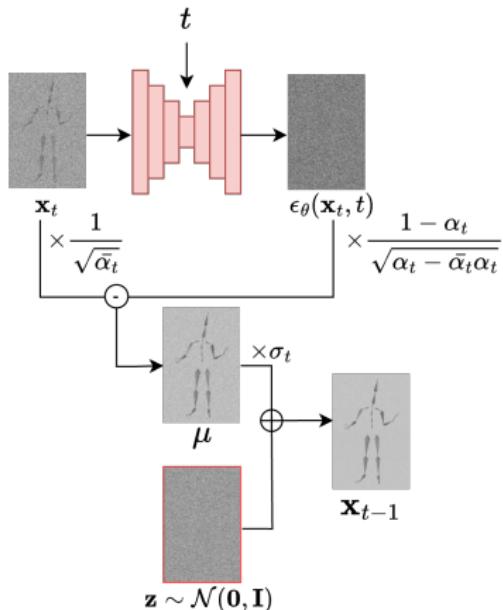
Algorithm Training

- repeat
- $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
- $t \sim \text{Uniform}(\{1, \dots, T\})$
- $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
- Take gradient descent step on
 $\nabla_\theta \|\boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{x}_t, t)\|^2$
- until converged

Reverse Diffusion Process

$$p_\theta(\mathbf{x}_T) = \mathcal{N}(0, \mathbf{I}) \quad p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_t; \mu_\theta(\mathbf{x}_t), \sigma_t^2 \mathbf{z})$$

$$\mu_\theta(\mathbf{x}_t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right)$$



Algorithm Sampling

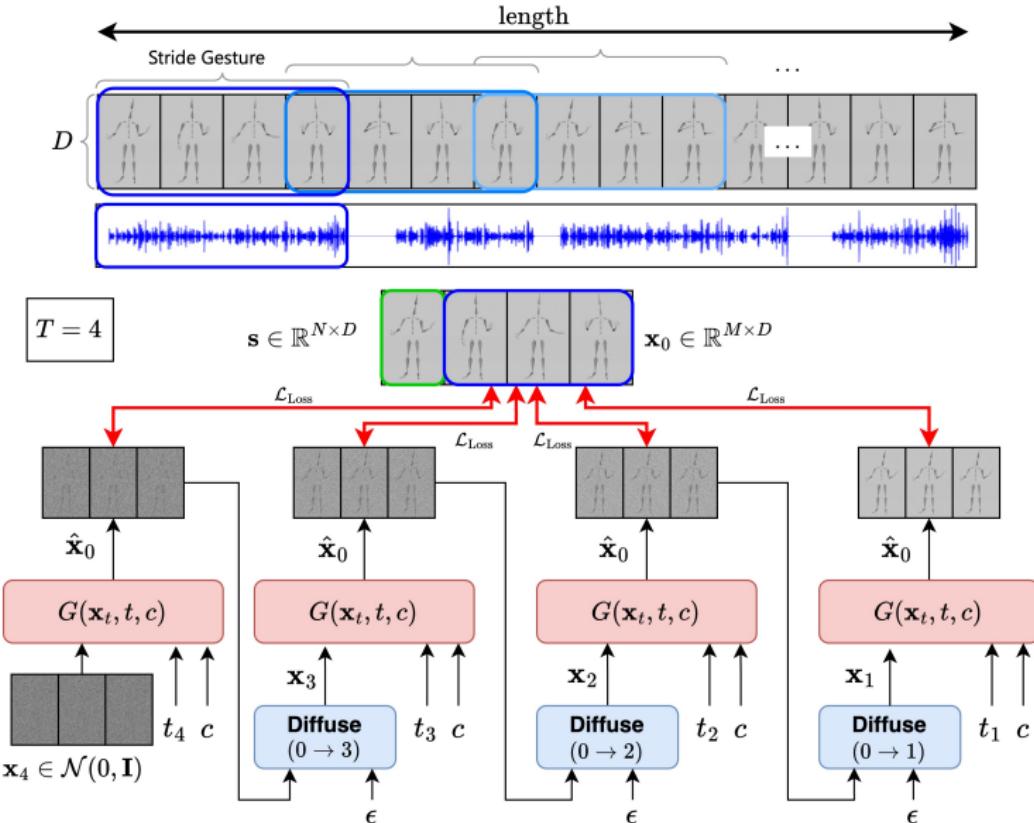
- 1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
 - 2: **for** $t = T, \dots, 1$ **do**
 - 3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
 - 4: $\mu = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}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

Outline

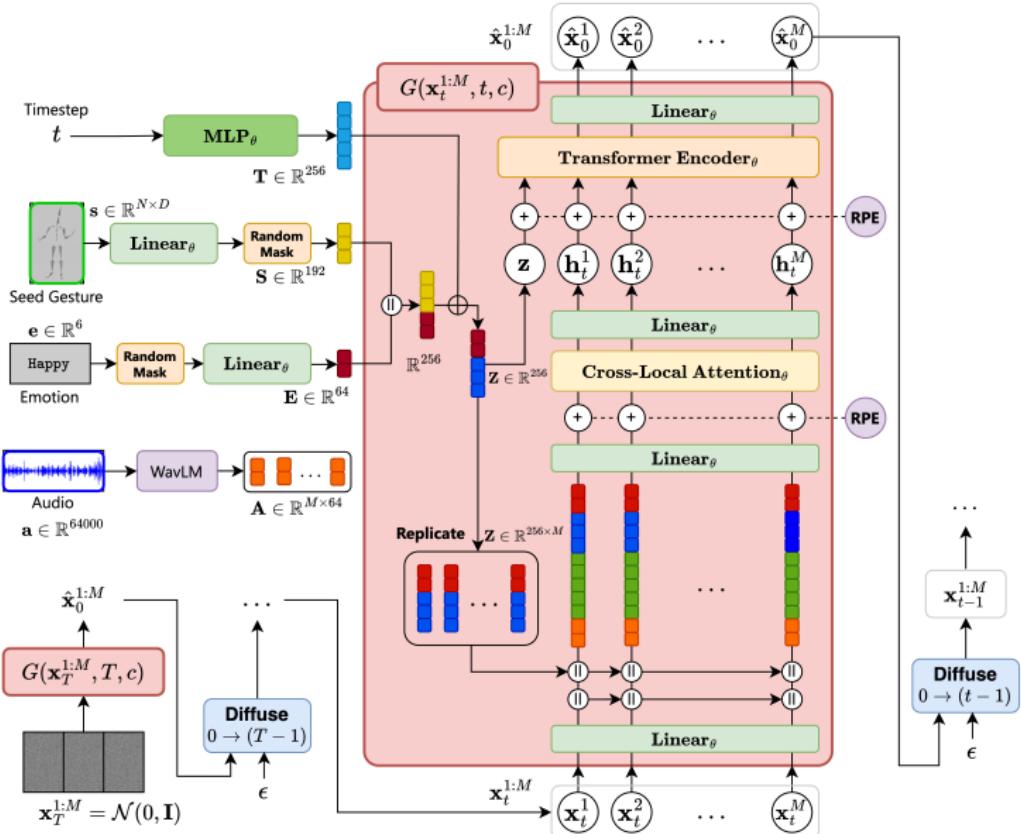
① Công trình liên quan

② Mô hình sinh cử chỉ

Tổng quan phương pháp



Kiến trúc OHGesture



Mô hình sinh cử chỉ

23

Cross-Local Attention and Self-Attention

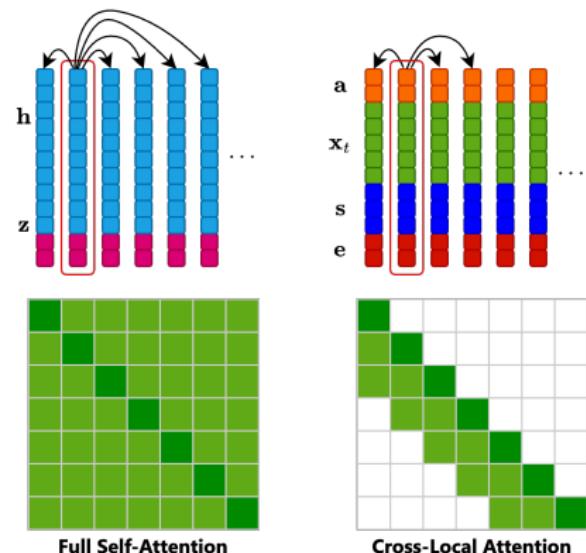
Cross-Local Attention

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{M}) = \text{softmax} \left(\frac{\mathbf{Q}\mathbf{K}^T + \mathbf{M}}{\sqrt{C}} \right) \mathbf{V}$$

Self-Attention

$$f_{\text{MultiHead}}(\mathbf{x}) = \text{concat} (\mathbf{H}_1, \mathbf{H}_2, \dots, \mathbf{H}_h) \mathbf{W}_O$$

$$\mathbf{H}_i = \text{softmax} \left(\frac{\mathbf{Q}_i \mathbf{K}_i^T}{\sqrt{d_k}} \right) \mathbf{V}_i$$



Emotion-controllable Gesture Generation

$$\hat{\mathbf{x}}_0 = G(\mathbf{x}_t, t, c) \quad (3)$$

Classifier-free guidance

- Unconditional: $c_1 = [\emptyset, \emptyset, a]$
- Conditional: $c_1 = [s, e, a]$:

$$\hat{\mathbf{x}}_{0\gamma, c_1, c_2} = \gamma G(\mathbf{x}_t, t, c_1) + (1 - \gamma) G(\mathbf{x}_t, t, c_2) \quad (4)$$

Emotion-controllable: $c_1 = [s, e_1, a]$, $c_2 = [s, e_2, a]$.

Huber Loss

$$\mathcal{L} = E_{\mathbf{x}_0 \sim q(\mathbf{x}_0 | c), t \sim [1, T]} [\text{HuberLoss}(\mathbf{x}_0 - \hat{\mathbf{x}}_0)] \quad (5)$$

$$\mathcal{L}_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for } |y - f(x)| \leq \delta, \\ \delta \cdot |y - f(x)| - \frac{1}{2}\delta^2 & \text{otherwise.} \end{cases}$$