

OpenHuman: Sinh cử chỉ dựa trên âm thanh

The synthesis system for conversational gestures based on emotion and semantics

Thanh Hoang-Minh

Giảng viên hướng dẫn:
TS. Lý Quốc Ngọc



Faculty of Information Technology
VNUHCM-University of Science

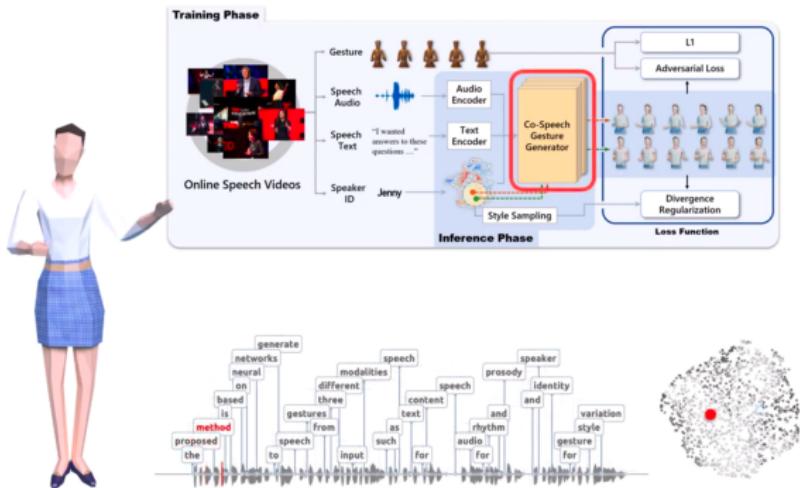
November 7, 2024

Outline

- ① Minh họa
- ② Giới thiệu
- ③ Công trình liên quan
- ④ Mô hình sinh cử chỉ
- ⑤ Thực nghiệm
- ⑥ Kết luận

Minh họa quá trình sinh cử chỉ

- Sinh cử chỉ (Gesture Generation) là gì?



ACM CCS: • Human-centered computing →
Human computer interaction (HCI).



Demo Gesture
Generation: [Youtube](#)

Outline

- ① Minh họa
- ② Giới thiệu
- ③ Công trình liên quan
- ④ Mô hình sinh cử chỉ
- ⑤ Thực nghiệm
- ⑥ Kết luận

Động lực nghiên cứu

Text-base Deep Learning

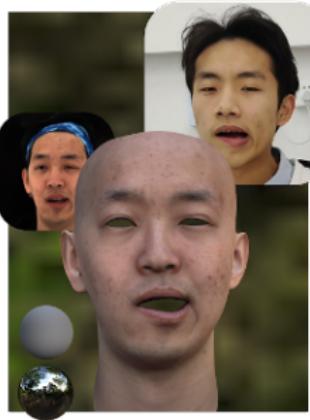
- ChatGPT, Alexa, Character.AI,..
- Text to speech, text to text,..

Text/audio to Realistic Digital Human

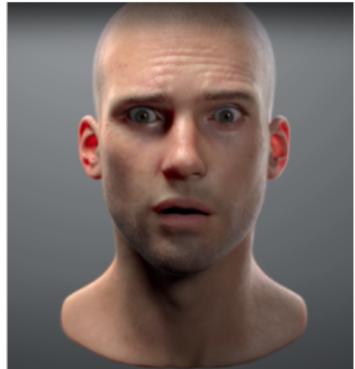
- Video-base (HeyGen, Midjourney,...)
- Rendering-base:
 - Character model (Gaussian Splatting)
 - Character animation: (3D keypoints)



(a)



(b)



Minh họa người kỹ
thuật số siêu thật
(realistic digital human)

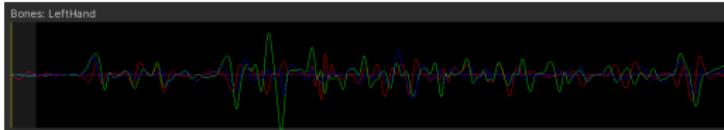
Dữ liệu bài toán

Dữ liệu

- Dữ liệu chuyển động thu nhận từ cảm biến.
- Âm thanh tương ứng với cử chỉ.

Kiến trúc khung xương

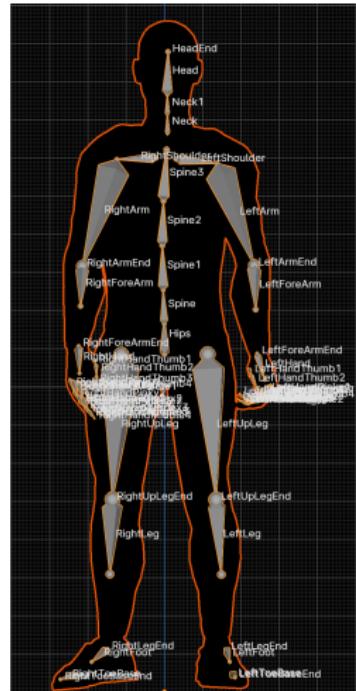
- Chuỗi chuyển động khung xương 60fps
 - Skeleton: 75 bones: $B = \{b_i\}^{75}$
(Head, Spine, Hips, LeftArm, RightArm...)
 - Dữ liệu BVH: $b_i = \{r_x, r_y, r_z\}$



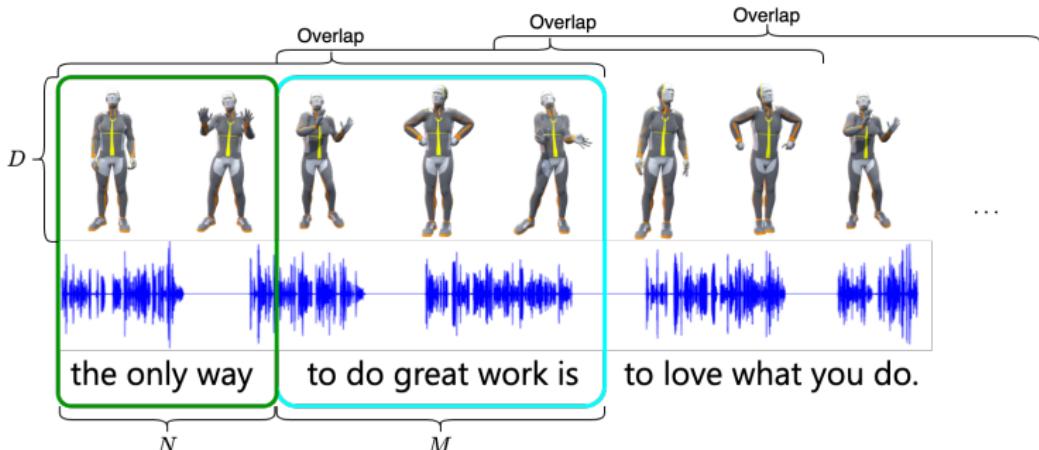
- Dữ liệu sau xử lý: $g \in \mathbb{R}^D$, với $D = 1141$
 $g = [p, r, p', r', p_{joins}, r_{joins}, p'_{joins}, r'_{joins}, d_{gaze}]$

Loại bài toán: Hồi quy (regression)

- Cho trước N khung hình (frame) cử chỉ, dự đoán M khung hình tiếp theo tương ứng với âm thanh a .



Phát biểu bài toán



Input

- Chuỗi cử chỉ khởi tạo: $g \in \mathbb{R}^{N \times D}$
- Chuỗi âm thanh: $a_{\text{raw}}^{\text{length}}$ sample rate 16000 trong 4s: $a \in \mathbb{R}^{6400}$
- Cảm xúc: $e \in \mathbb{R}^6$ (Happy, Sad, Neutral, Angry, Old, Funny)

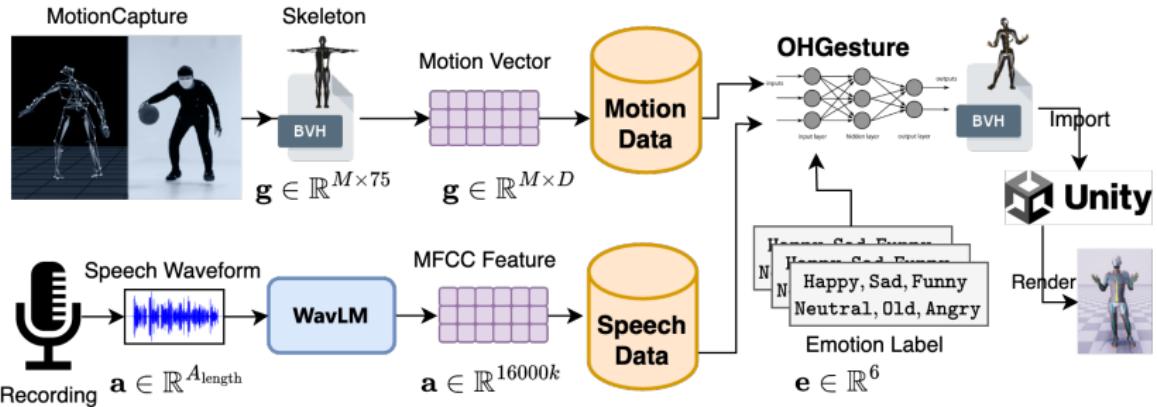
Output

- Chuỗi cử chỉ dự đoán: $\hat{x} \in \mathbb{R}^{M \times D}$

Groud Truth

- Chuỗi cử chỉ gốc: $x \in \mathbb{R}^{M \times D}$

Khung chương trình



Training

- $f_\theta(g_0, a, e)$

Inference

- $x' = f_{\theta'}$

Outline

- ① Minh họa
- ② Giới thiệu
- ③ Công trình liên quan
- ④ Mô hình sinh cử chỉ
- ⑤ Thực nghiệm
- ⑥ Kết luận

Các phương pháp

Rule Base & Statistic

- BEAT, Rule-based generation.
- Gesture Controllers

Phase Manifold :

- DeepPhase , Walk the Dog

Deep Learning

- **MLP**: Gesticulator
- **RNN**: Speech2AffectiveGestures, HA2G, TransGesture, ..
- **Normalising Flows**: Text2gestures, Speech2Gesture;; **GAN**: DiffGAN
- **VAE/VQ-VAE**: Audio2Gestures;
- **Diffusion**: Listen denoise action, DiffuseStyleGesture, Taming Diffusion Models.

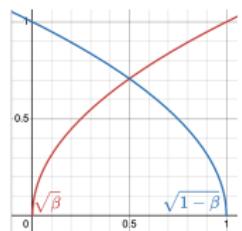
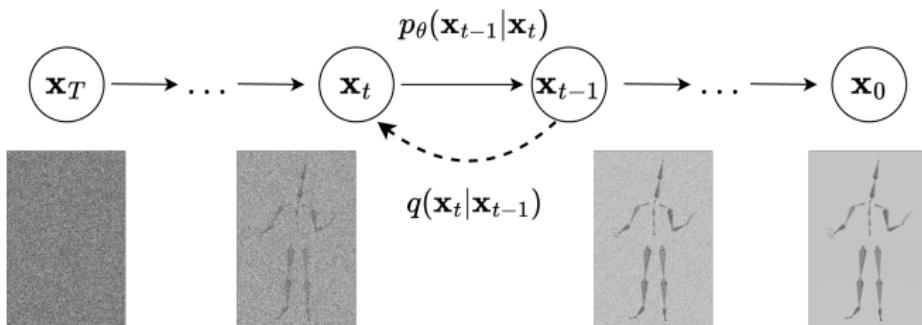
Baseline: Motion Diffusion Model (MDM)

Overview Diffusion Model

Forward Diffusion Process (Fixed)

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

hay $\mathbf{x}_{t-1} = \frac{1}{\sqrt{1-\beta_t}} (\mathbf{x}_t - \sqrt{\beta_t} \cdot \epsilon)$



Reverse Diffusion Process

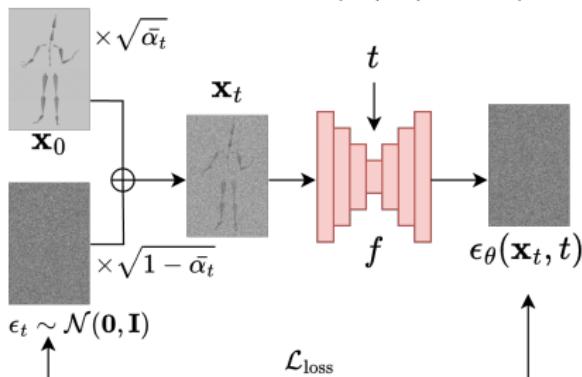
$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{1 - \beta_t}} \left(\mathbf{x}_t - \sqrt{\beta_t} \cdot \epsilon_\theta(\mathbf{x}_t, t) \right) + \beta_t \cdot \sigma_t \mathbf{z} \quad (2)$$

Forward Diffusion Process

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

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

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



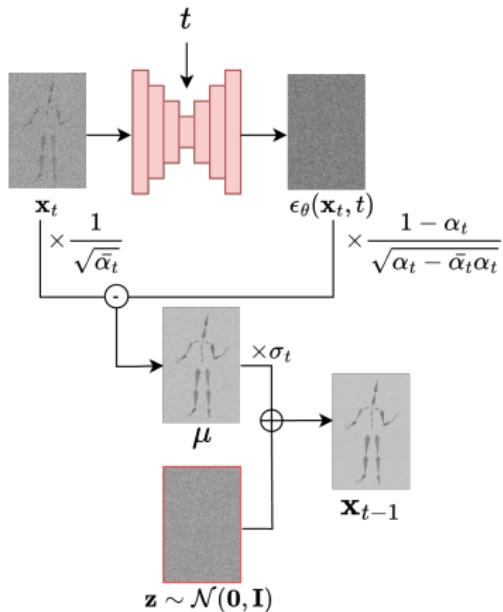
Algorithm Training

-
- 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
 $\nabla_\theta \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2$
 - 6: **until** converged
-

Reverse Diffusion Process

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

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



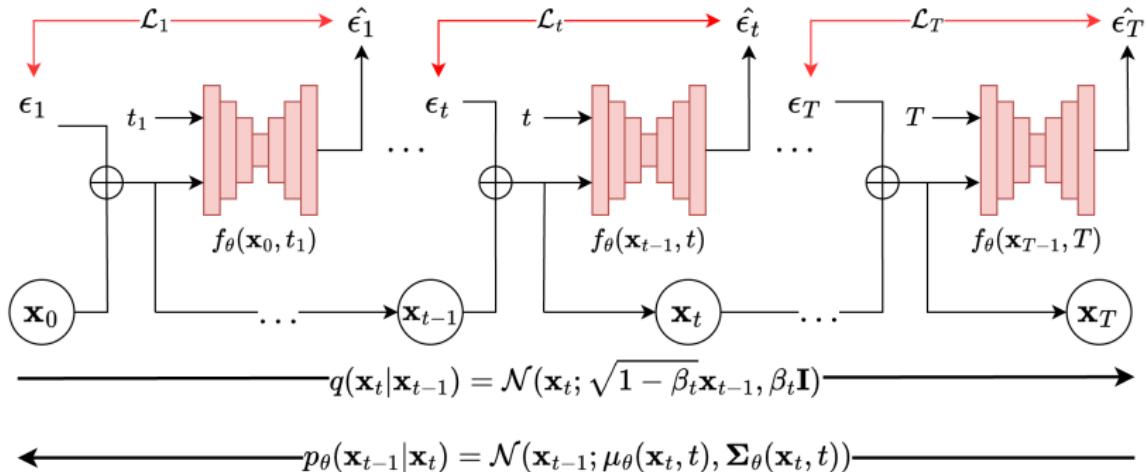
Algorithm Sampling

- 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

Hàm Loss \mathcal{L}

Hàm loss: $\mathcal{L} = \sum_{t=1}^T \mathcal{L}_t$

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



Đặc điểm của việc học dữ liệu Motion

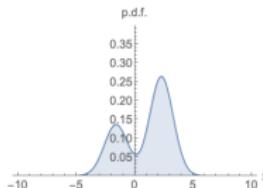
Quan hệ giữa dữ liệu cử chỉ và âm thanh:

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

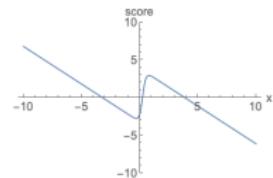
Khó khăn

- 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)

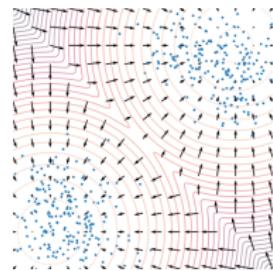


Không cần chuẩn hoá.

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



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

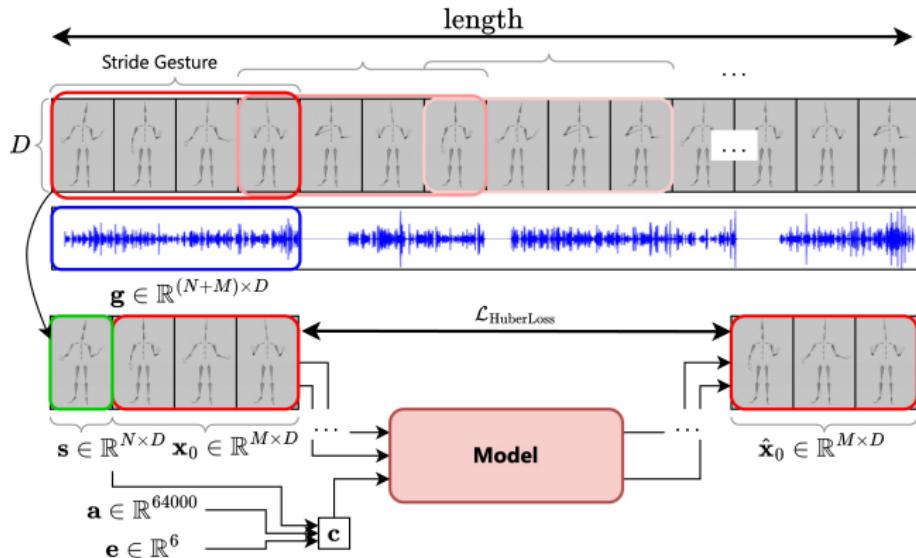


score function vs
probability density

Outline

- ① Minh họa
- ② Giới thiệu
- ③ Công trình liên quan
- ④ Mô hình sinh cử chỉ
- ⑤ Thực nghiệm
- ⑥ Kết luận

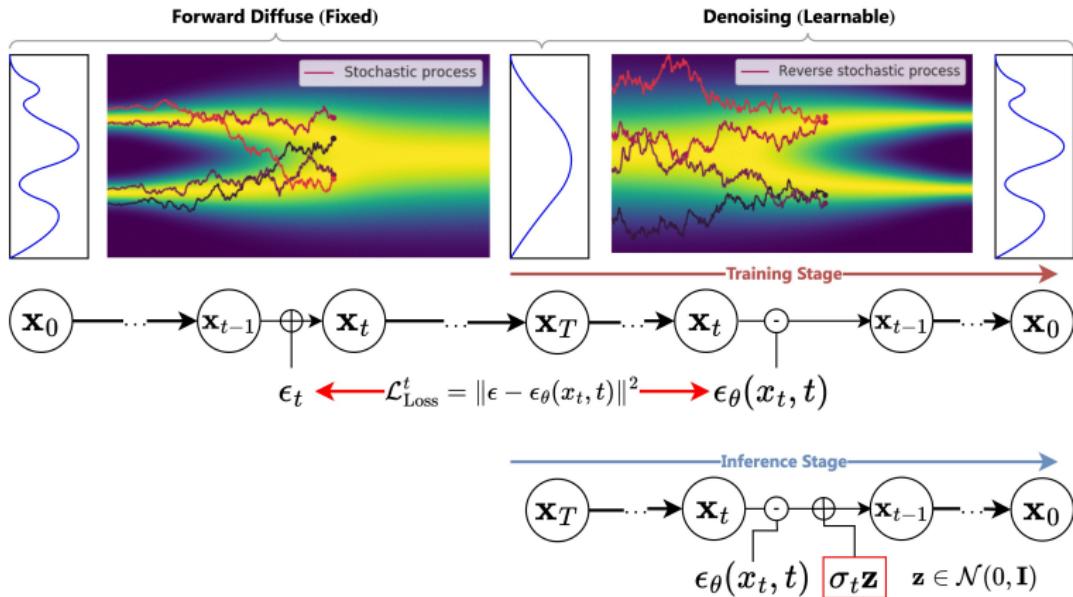
Tổng quan về bài toán



Cho chuỗi cử chỉ $g \in \mathbb{R}^{(N+M) \times D}$, với $s \in \mathbb{R}^{N \times D}$ và đoạn cử chỉ nhãn $x_0 \in \mathbb{R}^{M \times D}$ với $D = 1141$, $N = 8$, $M = 80$. Chuỗi âm thanh $a \in \mathbb{R}^{64000}$ (4s: 16000 sample rate). Cảm xúc $e \in \mathbb{R}^6$.

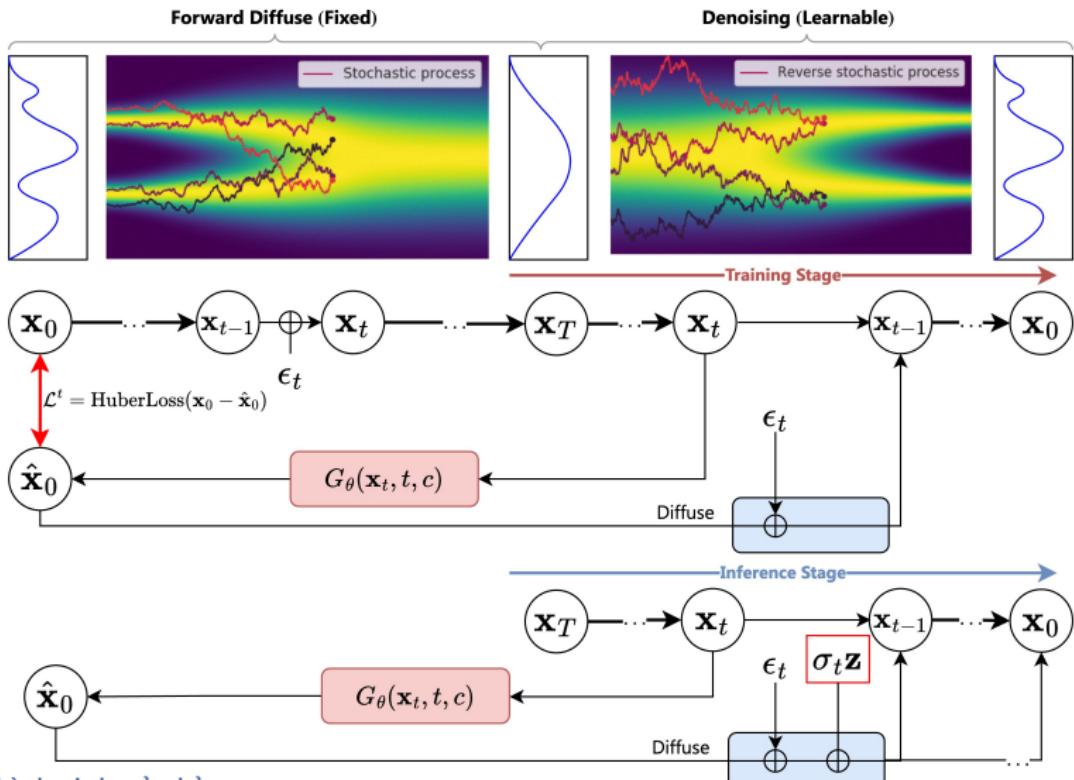
Vanilla Diffusion với ϵ Objective

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



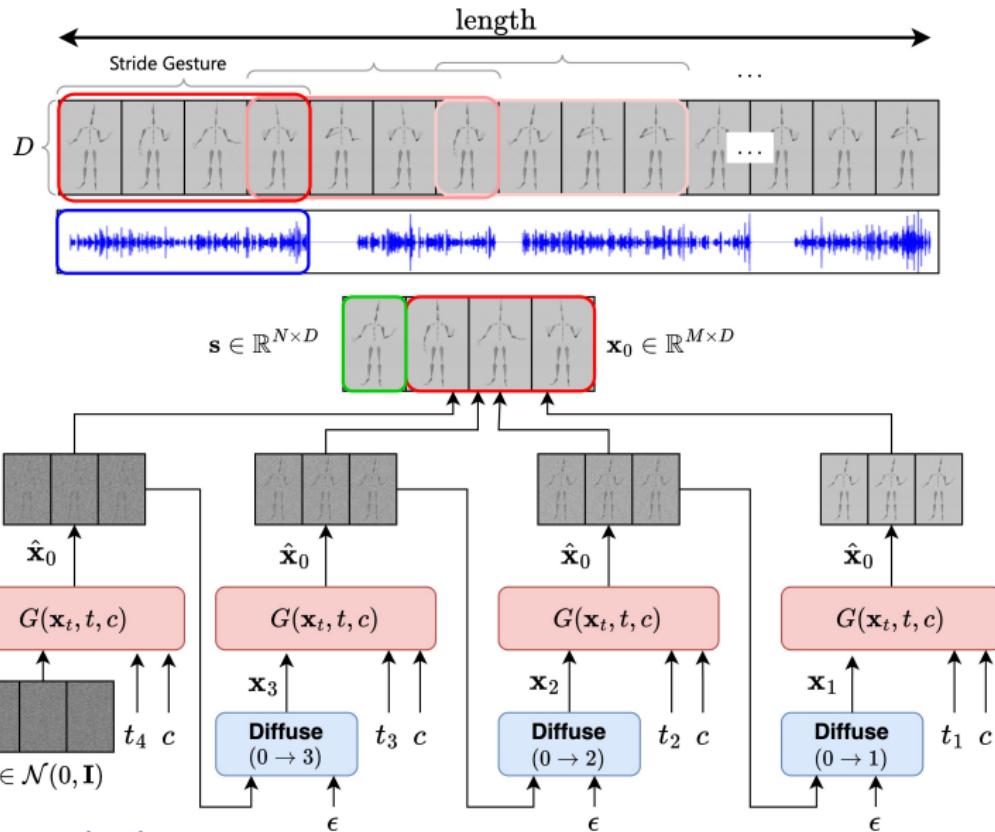
Condition Diffusion với \mathbf{x}_0 Objective

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



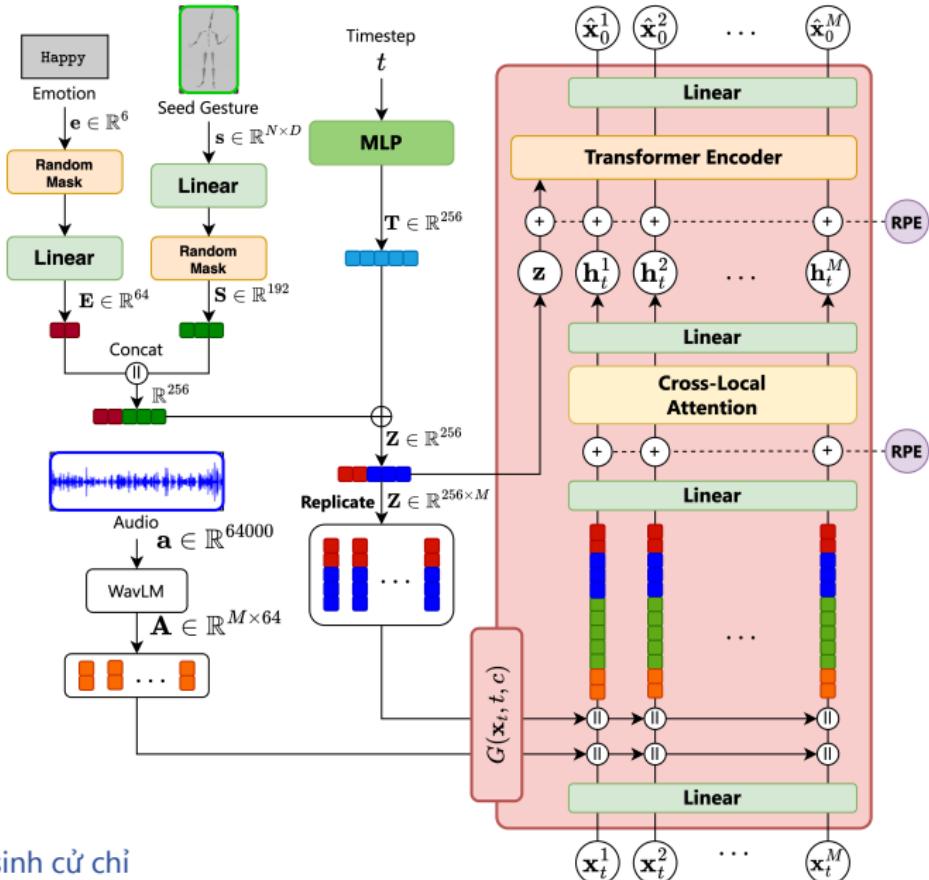
Mô hình sinh cử chỉ

Tổng quan phương pháp



Mô hình sinh cử chỉ

Kiến trúc OHGesture



Mô hình sinh cử chỉ

Cross-Local Attention and Self-Attention

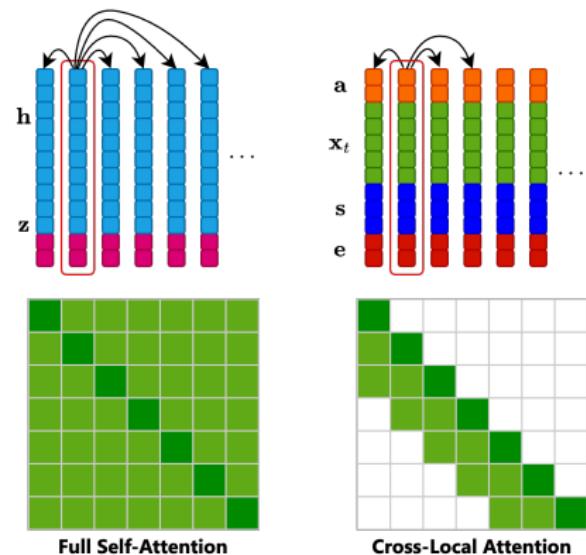
Cross-Local Attention

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

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

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

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

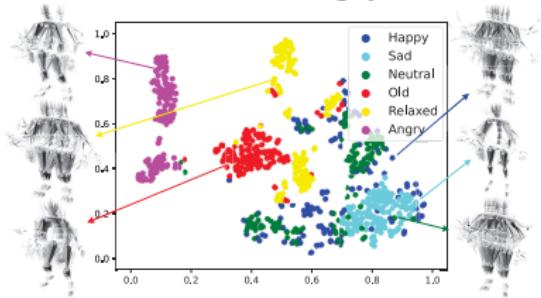
Outline

- ① Minh họa
- ② Giới thiệu
- ③ Công trình liên quan
- ④ Mô hình sinh cử chỉ
- ⑤ Thực nghiệm
- ⑥ Kết luận

Dataset

ZeroEGGs Dataset

- Happy, Sad, Neutral
- Old, Relaxed, Angry



Độ đo

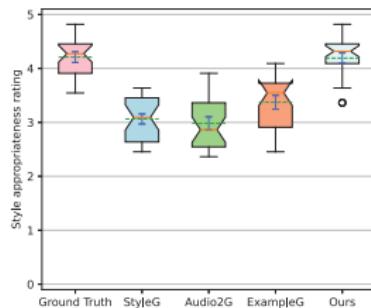
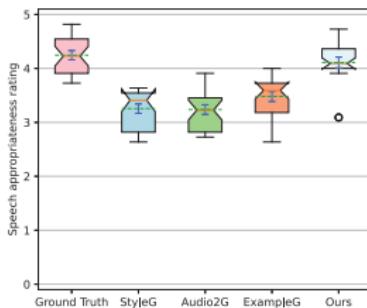
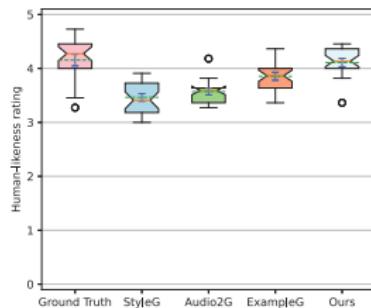
Các độ đo trong Gesture Generation:

- Human-likeness
- Gesture-Speech Appropriateness
- Gesture-style Appropriateness
- FID (Fréchet Inception Distance):

$$\text{FID} = \|\mu_{\text{real}} - \mu_{\text{fake}}\|^2 + \text{Tr} \left(\Sigma_{\text{real}} + \Sigma_{\text{fake}} - 2 (\Sigma_{\text{real}} \Sigma_{\text{fake}})^{\frac{1}{2}} \right)$$

- Mean opinion scores (MOS)

Kết quả



Name	Human likeness ↑	Gesture-speech appropriateness ↑
Ground Truth	4.15 ± 0.11	4.25 ± 0.09
Ours	4.11 ± 0.08	4.11 ± 0.10
– WavLM	4.05 ± 0.10	3.91 ± 0.11
– Cross-local attention	3.76 ± 0.09	3.51 ± 0.15
– Self-attention	3.55 ± 0.13	3.08 ± 0.10
– Attention + GRU	3.10 ± 0.11	2.98 ± 0.14
+ Forward attention	3.75 ± 0.15	3.23 ± 0.24

Outline

- ① Minh họa
- ② Giới thiệu
- ③ Công trình liên quan
- ④ Mô hình sinh cử chỉ
- ⑤ Thực nghiệm
- ⑥ Kết luận

Kết luận

- Mô hình **Diffusion**, ánh xạ xác suất làm tăng sự đa dạng đồng thời cho phép tạo ra các cử chỉ có chất lượng cao, giống con người.
- Mô hình của **OHGesture** tổng hợp các cử chỉ sao cho phù hợp với nhịp điệu âm thanh và ngữ nghĩa văn bản dựa trên Cross-Local Attention và Self-Attention.
- Sử dụng phương pháp **Classifier-free Guidance**, có thể điều khiển các điều kiện như cảm xúc, cử chỉ khởi tạo, có thể nội suy để suy luận ra giữa các cảm xúc khác nhau.

References