arXiv:2305.04919v1 [cs.HC] 8 May 2023

DiffuseStyleGesture：基于扩散模型的风格化音频驱动伴随语音手势生成

Sicheng Yang[1] , **Zhiyong Wu**[1,4*] , **Minglei Li**[2] , **Zhensong Zhang**[3] ,**Lei Hao**[3] , **Weihong Bao**[1] , **Ming Cheng**[1] , **Long Xiao**[1]1清华大学深圳国际研究生院，清华大学，中国深圳[2]华为云计算技术有限公司，中国深圳[3]华为诺亚方舟实验室，中国深圳[4]香港中文大学，中国香港特别行政区， 中国 yangsc21@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn,{liminglei29, zhangzhensong}@huawei.com

摘要

除了语音之外，手势也是沟通的艺术表现形式。自动生成伴随语音的手势在计算机动画领域备受关注。这是一项具有挑战性的任务，因为手势种类繁多，而且难以使手势的节奏和语义与相应语音相匹配。为了解决这些问题，我们提出了**DiffuseStyleGesture**，一种基于扩散模型的语音驱动手势生成方法。该方法能够根据任意长度的给定语音，生成高质量、与语音匹配、风格化且多样的伴随语音的手势。具体而言，我们在手势扩散流程中引入交叉局部注意力和自注意力，以生成与语音匹配度更高且更逼真的手势。随后，我们采用无分类器引导训练模型，通过插值或外推控制手势风格。此外，我们利用不同的初始手势和噪声提升生成手势的多样性。大量实验表明，我们的方法在语音驱动的手势生成方面优于近期方法。我们的代码、预训练模型和演示可在 https://github.com/YoungSeng/DiffuseStyleGesture 获取。

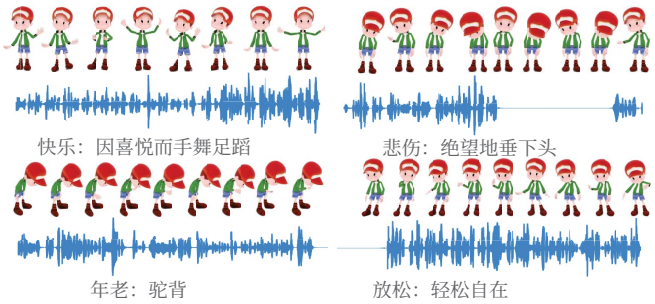


图1：**我们提出的方法在不同类型的语音和风格下生成的手势示例。**论文中使用的所有角色均已公开可用。

1 引言

身体手势和面部表情是人类沟通中传递信息的重要工具[Kucherenko 等， 2021]。自动生成伴随语音的手势是开发电影、游戏、虚拟社交环境以及与社交机器人互动中的逼真化身的关键技术 [Nyatsanga 等， 2023]。伴随语音的手势生成最重要的问题是：1）如何生成与音频节奏和文本语义相匹配的手势；以及 2）如何生成多样化且风格化的手势。近期的手势生成方法可以直接生成以.....为条件的人类手势

中性语音 [Nyatsanga 等，2023年；Yoon等，2022年；Kucherenko等， 2021]。然而，所有这些方法仍然限制了所学分布，因为它们主要采用基于生成对抗网络的方法 [Yoon 等人、 2020], 变分自编码器 [Li等人，2021年a] 或流模型 [Alexanderson 等人，2020年a]。基于生成对抗网络的合成方法会遭遇模式崩溃，导致合成质量较低，尤其是在训练数据中未见过的数据上。使用变分自编码器和流模型的方法则需要在生成质量与多样性之间进行权衡 [Tevet 等人，2022年；Dabral 等人， 2022]。

最近，扩散模型 [Ho 等， 2020] 这类生成方法凭借其高质量和多样化的生成能力，在其他领域取得了令人瞩目的成果，例如图像生成 [Ramesh 等、 2022], 视频生成 [Mei 和 Patel, 以及 2022], 文本生成 [Lin 等，2022]。这些研究展示了基于去噪扩散的模型学习真实数据分布的能力，同时还能提供多样化的采样与操控方式，例如编辑和插值。然而，这些研究并未对依赖时间的序列进行建模，无法解决语音驱动手势等需要时间对齐的问题，而且计算资源开销很大。

为生成高质量、与语音匹配、风格化且多样的伴随语音手势，受近期基于去噪扩散的生成方法进展启发，我们提出了**Dif-fuseStyleGesture**，一种用于音频驱动伴随语音手势生成的灵活、可控且具备时间感知能力的基于去噪扩散模型。生成手势的示例如下

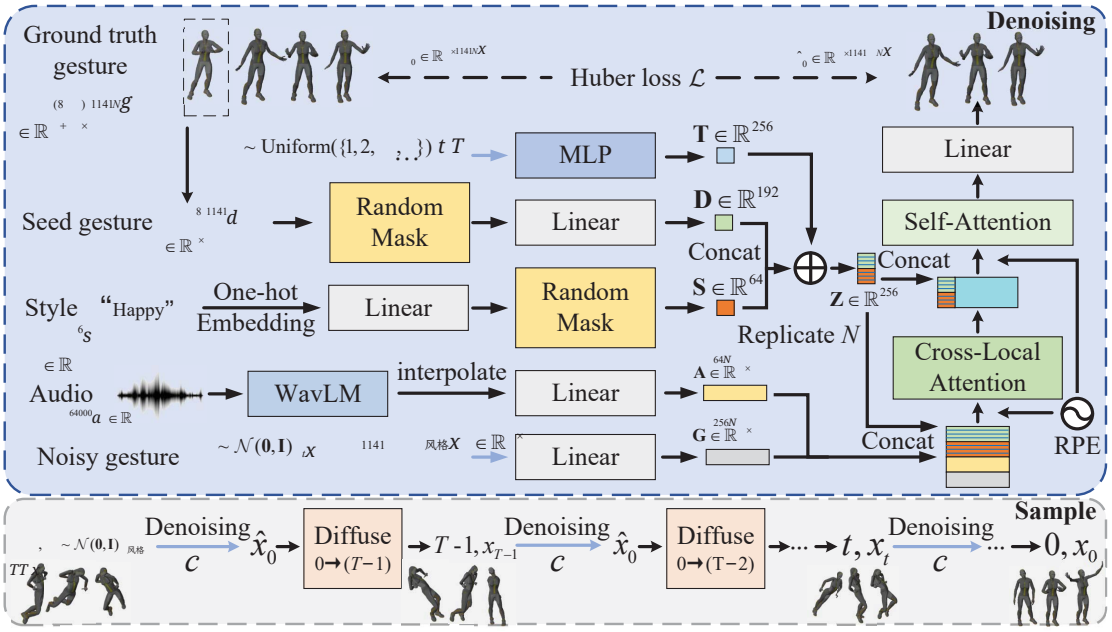


图 2：**(顶部) DiffuseStyleGesture 的去噪模块**。一个加噪步骤 t 以及在该加噪步骤下以 c（包括种子手势 d、风格 s和音频 a）为条件的噪声手势序列 x_t 被输入模型。交叉局部注意力和自注意力能够基于 WavLM 特征更好地捕捉语音与手势之间的相关性。种子手势和风格特征处理流程中的随机掩码有助于模型进行无分类器引导训练，并执行插值或外推，从而实现对生成的手势的高度控制。**(底部) DiffuseStyleGesture 的采样模块**。在每个步骤 t，我们根据相应条件通过去噪过程预测$\hat{x}_0$，然后通过扩散过程将噪声添加到加噪步骤 x_{t-1}。该过程从 $t = T$ 重复，直到 $t = 0$。

见图 1。我们方法的概览见图 2。我们采用基于注意力的架构来捕捉语音与手势之间的时间信息。我们发现，相较于预测噪声，训练模型直接预测信号本身效果更好 [Ramesh 等，2022年；Tevet 等， 2022] 。为使生成的手势与语音更好地对齐，我们还提出一种方法，利用交叉局部注意力捕捉手势和语音的局部信息，再通过自注意力捕捉全局信息，从而根据语音时长生成任意长度的伴随语音手势。此外，我们利用 WavLM 特征 [Chen 等， 2022] 来考虑音频中的语义、情感及其他信息，以提升模型的泛化能力和鲁棒性。最后，我们在训练时使用随机掩码执行无分类器引导 [Ho 和 Salimans， 2022]，从而实现对控制条件的插值与编辑。我们的主要贡献如下：

- 我们将扩散模型扩展为融入时间信息的模型，以实现音频驱动的伴随语音的手势生成。借助扩散模型，我们能够对生成的手势进行高度控制，例如编辑手势的风格、设置初始手势以及生成多样化手势。

- 我们采用交叉局部注意力和全局自注意力来捕捉特征信息，并生成更适合语音的手势。

- 大量实验表明，我们的模型能够生成类人、与语音和风格相匹配的手势，其表现显著优于现有的手势生成方法。

2 相关工作

2.1 语音伴随手势生成

手势生成是一项复杂的任务，需要理解语音、手势及其相互关系。数据驱动方法尝试从人类示范中学习手势表达技能。现有研究主要考虑四种模态：文本 [Alexanderson 等，2021；Yoon 等， 2019]，音频 [Habibie 等，2021；Li等，2021b；Ginosar 等， 2019]，手势动作，以及说话人身份[Yoon 等，2020；Liu 等，2022；Alexanderson 等，2020a]。 [Habibie 等， 2021] 提出了首个联合合成同步的 3D 对话身体和手部手势，以及 3D 面部和头部动画的方法。 [Yi 等， 2022] 使用自编码器处理面部动作，并采用组合式向量量化变分自编码器（VQ-VAE）生成更多样化的手势。 [Xie等， 2022] 引入了一种 VQ-VAE 模型，将姿态序列表示为潜在编码序列，并设计了一种扩散架构，用于生成文本到手语姿态序列。至于个体风格的学习 [Li 等，2021a；Liang 等， 2022]，[Yoon 等， 2020] 提出了首个端到端方法，利用文本、音频和说话人身份三种模态生成伴随语音的手势，

文本、音频和说话人身份三种模态。 [Ahuja 等，2020] 训练一个适用于多个说话人的单一模型，同时学习每位说话人手势的风格嵌入。[Liang 等， 2022] 提出一种语义增强生成方法，用于生成具有语义感知能力的手势。 [Ao 等，2022] 基于语言学理论，解耦语音和动作的低层与高层嵌入。

一些研究采用运动匹配方法生成共语音手势 [Habibie 等，2022年；Zhou 等， 2022]。该方法需要精心设计数据库，而数据库与生成的手势的表现直接相关。匹配长度需要在质量和多样性之间取得平衡。此外，该方法还需要复杂且耗时的人工匹配规则设计。

最近，一个高质量的 3D 手势数据集 ZeroEGGS[Ghorbani 等， 2022] 由带有风格标签的多摄像机视频构建而成。我们的工作使用了该数据集。

2.2 用于动作生成的扩散模型

扩散模型擅长对复杂的数据分布进行建模，并生成生动的动作序列。许多研究将基于扩散的生成模型引入动作领域，并针对人体动作领域仔细调整无分类器扩散生成模型的网络结构，例如基于 Transformer 的方法 [Kim 等，2022年；Tevet 等，2022年；Ren 等，2022年；Zhou 和 Wang， 2022]。[Chang 等，2022b] 设计了一种扩散模型多任务架构，并采用对抗性约束和物理约束进行人体动作合成。 [Dabral 等， 2022] 提出了 Mo-Fusion，用于生成较长、时间上合理且语义准确的动作。物理引导的动作扩散模型 [Yuan 等， 2022] 将物理约束融入扩散过程。

[Ginosar等人提出了一种基于单人语音驱动手势的跨模态转换方法。 [Li等人提出了一种跨条件因果注意力层，以保持生成身体动作的连贯性。[Chang等人（2022a）] 采用局部性约束注意力机制，并在 GENEA 2022 手势生成挑战赛中，在全身层面实现了最佳的手势-语音恰当性。 [Yoon等人 2022]。MotionDiffuse [Zhang等人2022] 利用交叉注意力实现概率映射、逼真合成和多层次操控。在我们的工作中，我们使用交叉局部注意力捕捉手势和语音的局部信息，然后使用自注意力捕捉全局信息，从而使生成的手势与语音更加匹配。

3 我们的方法

3.1 用于手势生成的扩散模型

我们的思路是通过学习从纯噪声开始逐步去噪，利用扩散模型生成手势。 [Ho等人。如图 2 所示，扩散模型由两部分组成：正向过程（扩散过程） q和反向过程（去噪过程） p_θ。

扩散过程。扩散过程 q 被建模为一个马尔可夫加噪过程。我们将生成的手势记为 gesture x，其维度与观测数据 $x_0 \sim q$ 相同。(x_0)， $q(x_0)$ 表示真实数据的分布。根据方差调度方案 β_1， β_2、..... β_T（$0 < \beta_1 < \beta_2 < \cdots < \beta_T < 1$， T 为总时间步数），我们添加高斯噪声

$$q\left(x_t \mid x_{t-1}\right)=\mathcal{N}\left(x_t ; \sqrt{1-\beta_t} x_{t-1}, \beta_t \mathbf{I}\right) \tag{1}$$

逐渐作用于每个时刻的手势 t，若噪声调度设计得当且 T 足够大，

$$q\left(x_{1: T} \mid x_0\right)=\prod_{t=1}^T q\left(x_t \mid x_{t-1}\right) \tag{2}$$

则最终会得到纯噪声。

去噪过程。去噪过程 p_θ 是通过神经网络学习参数 θ 的过程。假设去噪过程也符合高斯分布，即利用时刻 t 的噪声 x_t 来学习 μ_θ、 Σ_θ，那么

$$p_\theta\left(x_{t-1} \mid x_t\right)=\mathcal{N}\left(x_{t-1} ; \mu_\theta\left(x_t, t\right), \Sigma_\theta\left(x_t, t\right)\right) \tag{3}$$

为方便计算，我们假设 $\alpha_t=1-\beta_t$和$\bar{\alpha}_t=\prod_{i=1}^T \alpha_i$。那么时刻 t 的噪声手势 x_t 可以写作

$$q\left(x_t \mid x_0\right)=\mathcal{N}\left(x_t ; \sqrt{\bar{\alpha}_t} x_0,\left(1-\bar{\alpha}_t\right) \mathbf{I}\right) \tag{4}$$

通过最小化真实噪声 ϵ 与预测噪声 ϵ_θ 之间的差异来优化网络（x_t， t）[Ho 等人， 2020]。采样时，我们可以通过固定方差来学习均值μ_θ（x_t， t）$=\frac{1}{\sqrt{\alpha_t}}\left(x_t-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta\left(x_t, t\right)\right)$。

框架。我们的目标是在给定条件 c 的情况下，合成人类手势$x^{1: N}$，其长度为 N。在我们的工作中，我们遵循[Ramesh 等，2022年；Tevet 等， 2022]，预测信号本身，而非预测 ϵ_θ（x_t， t）[Ho 等， 2020]。去噪模块根据输入噪声 x_t、加噪步骤 t 和条件 c重建原始信号 x_0

$$\hat{x}_0=\operatorname{Denoise}\left(x_t, t, c\right) \tag{5}$$

然后，可以通过在训练样例中最小化生成的姿态 $\hat{x}_0$ 与真实标注的人类手势 x_0 之间的 Huber 损失 [Huber， 1992] 来训练去噪模块：

$$\mathcal{L}=E_{x_0 \sim q(x_0 \mid c), t \sim[1, T]}\left[\operatorname{HuberLoss}\left(x_0-\hat{x}_0\right)\right] \tag{6}$$

3.2 基于注意力的语音驱动手势生成模型

去噪模块中的特征处理。如图 2 所示，手势根据加噪步骤 t、噪声手势 x_t 和条件 c（包括音频 a、风格 s和种子手势 d）生成。每种特征的处理流程如下：

- 加噪步骤：在训练过程中，加噪步骤 t 从 $\{1, 2, \ldots, T\}$的均匀分布中采样，并采用与 [Vaswani 等人相同的位置编码， 2017], 然后通过多层感知机（MLP）映射到维度为 256 的空间 $\mathbf{T}$。

• 噪声手势：在训练过程中， x_t 是与真实手势 x_0 维度相同的噪声手势，通过从标准正态分布 $\mathcal{N}(0, \mathbf{I})$中采样获得。采样时，初始噪声手势$x_T$从标准正态分布中采样，另一个 x_t， $t < T$则是上一个加噪步骤的结果。然后通过线性层将维度调整为 256，得到 **G** 。

• 音频：所有音频录音均降采样至 16 kHz，并使用预训练模型 WavLM Large [Chen 等人， 2022]提取特征。我们使用线性插值，在时间维度上将 WavLM 特征和手势 x_0 对齐至 20 帧/秒，然后使用线性层将特征维度降至 64，形成最终的音频特征序列**A**。

• 风格：手势风格由独热向量表示，其中选定风格对应的向量仅有一个元素非零，并通过线性层映射至 64 维空间**S** 。

•种子手势：种子手势有助于在连续合成之间实现平滑过渡 [Yoon 等人，2020]。有关真实手势片段g $\in\mathbb{R}^{(8+N)\times 1141}$的更多详情，请参阅我们的补充材料。手势片段$g$ 的前 8 帧用作种子手势 d ，其余 N帧用作真实手势 x_0 ，以计算损失$\mathcal{L}$。然后，我们使用线性层将种子手势d 的特征维度映射至 192 维空间**D** 。我们生成的手势时长为 4 秒，并将手势动画重采样至 20 帧/秒，然后 N = 80。

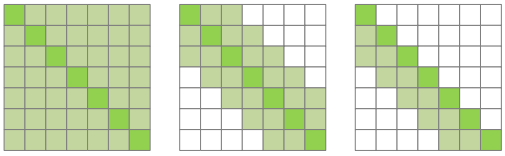
去噪模块中的模型。 我们采用基于注意力的架构进行去噪。在利用长程相关性之前，建议先利用局部上下文构建表征 [Rae 和 Razavi，2020]。

我们将种子手势**D** 与风格**S** 拼接在一起，形成一个 256 维向量，然后添加加噪步骤**T** 的信息，得到**Z**。我们的网络接收向量**Z** ，并将其副本堆叠成序列特征，以与音频和手势特征的时间线对齐；随后将它们与音频**A** 和手势**G** 拼接，作为交叉局部注意力网络的输入。我们提出的用于伴随语音的手势生成的交叉局部注意力基于 Routing Transformer [Roy 等， 2021], 该方法表明，局部注意力对于构建中间表征至关重要，如图 3(c) 所示。

之后，我们将交叉局部注意力的输出与**Z** 拼接，并将其输入自注意力网络，如图 3(a) 所示。自注意力机制类似于 Transformer [Vaswani 等， 2017] 编码器，用于确定数据序列元素之间的计算依赖关系，其实现方式如下：

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{M}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T + \mathbf{M}}{\sqrt{C}}\right)\mathbf{V} \tag{7}$$

其中，**Q**、**K**、**V** 分别表示输入中的查询、键和值，**M** 则是掩码，用于确定注意力模式的类型。我们采用相同的相对位置编码（ RPE）



(a) full self-注意力　(b) 滑动窗口-向下注意力　(c) 跨局部注意力

图 3：我们在实验中使用的不同注意力模式，其中 (a) 和 (c) 是模型中使用的注意力机制，(b) 是第 4.3 节中进行比较的一种模式。行表示输出，列表示输入。彩色方块突出显示了每行输出对应的相关元素。

机制，正如 [Kitaev 等人所述， 2020] 从而使手势转换中的时间效应保持不变。最后，自注意力的输出经过一个线性层后，被映射回与 x_0 相同的维度。

从图 3 中可以看出，通过简单调整相应的掩码 **M**，即可实现这些不同的注意力机制。我们还在 Longformer [Beltagy 等人提出的方法中尝试了滑动窗口注意力， 2020], 相关结果将在第 4.3 节中进行分析。

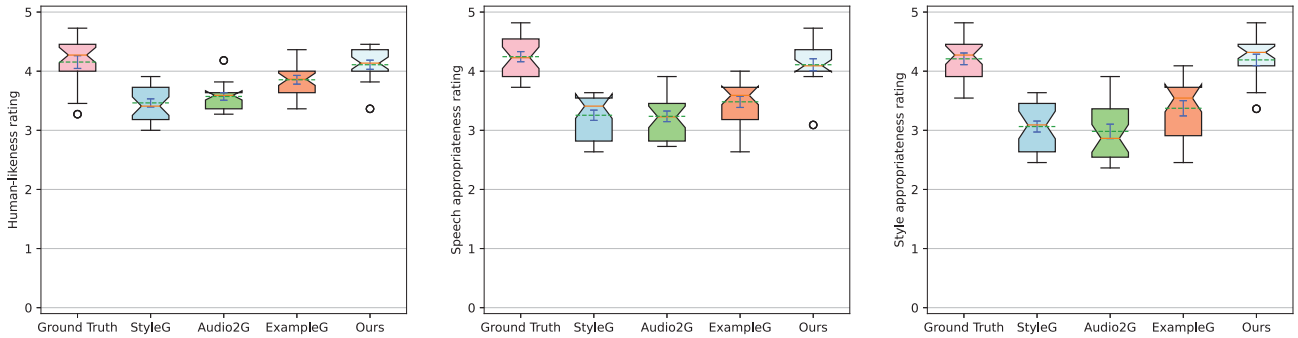
示例模块。 最终的语音伴随手势通过拼接若干个长度为 N 的片段得到。第一个片段的种子手势可以通过从数据集中随机采样一个手势生成，也可以将其设置为平均手势。随后，其他片段的种子手势则取自前一个片段所生成手势的最后 8 帧。对于每个片段，在每个加噪步骤 t 中，我们使用扩散过程预测干净手势 $\hat{x}_0$=Denoise(x_t, t, c)，并根据公式（1）向加噪步骤 x_{t-1} 添加噪声。该过程从 $t = T$ 开始重复执行，直至达到 x_0 （图 2 底部）。

3.3 风格可控的手势生成

由于该算法根据控制条件生成手势，因此控制条件不仅可以是音频 a ，也可以是风格 s或种子手势 d等。如图 2 所示，我们参考了 [Ho 和 Salimans，2022年；Tevet 等人，2022] 并在种子手势 d 和风格 s 特征处理流程中添加随机掩码，以进行无分类器学习，从而实现对不同条件的精准控制。随后，在训练过程中，结合条件模型 Denoise $(x_t,\ t,\ c_1)$ $c_1 = [d, s, a]$ 与无条件模型 Denoise $(x_t,\ t,\ c_2)$ $c_2 = [\varnothing, \varnothing, a]$ 的预测结果，即可实现手势生成的无分类器引导，如公式（8）所示。

$$\hat{x}_{0\gamma,c_1,c_2} = \gamma\,\text{Denoise}\,(x_t, t, c_1) + (1-\gamma)\,\text{Denoise}\,(x_t, t, c_2) \tag{8}$$

在实践中，去噪模块通过使用伯努利掩码随机屏蔽 10% 的样本，学习条件分布和无条件分布。随后，对于风格 s 条件，我们可以在采样时使用 γ对这两种变体进行插值甚至外推，从而生成受风格控制的手势，如 $c_1 = [d, s_1, a]$, $c_2 = [d, s_2, a]$ 公式（8）所示。有关数据集和实现细节等训练详情，请参阅我们的补充材料。



(a) 拟人度评分的箱线图。(b) 手势-语音适切性评分的箱线图。(c) 手势-风格适切性评分的箱线图。

图4：可视化展示不同模型在不同维度上的平均意见分（MOS）比较结果的箱线图。箱体从数据的第一下四分位数（Q1）延伸至第三上四分位数（Q3）。红线表示中位数。凹口表示中位数周围95%的置信区间（CI）。当CI小于Q1或大于Q3时，凹口会延伸至箱体之外，呈现独特的“翻转”外观。图中还分别以绿色虚线和蓝色竖线标出了均值及其95%的CI。

4 实验

4.1 与现有方法的比较

我们将提出的模型与 StyleGestures[Alexanderson 等人，2020年b]，Audio2Gestures [Li 等人，2021年b]，ExampleGestures [Ghorbani等人， 2022]进行了比较。目前，语音驱动手势缺乏与人类主观感知一致的客观指标 [Yoon 等人，2022年；Kucherenko 等人，2021年；Alexanderson 等人， 2022]，甚至弗雷歇手势距离（FGD）也不例外 Yoon 等人，2020年；Dabral 等人， 2022]，因此，我们所有实验的评分均通过人类主观评估完成。我们从三个维度开展评估。前两个维度沿用 GENEA 中的评估方式 [Yoon 等人， 2022]，评估拟人度和手势-语音恰当性。第三个维度是手势风格恰当性。

用户研究。 为了了解我们方法在真实场景中的视觉表现，我们针对各个对比方法生成的手势序列以及真实动作捕捉数据开展了用户研究。评估片段的长度介于11秒和51秒之间，平均长度为31.6秒。需要注意的是，这里用于主观评估的片段手势时长比GENEA [Yoon 等人的 2022] 评估（8至10秒）更长，因为更长的时间段能够产生更加明显且更令人信服的适切性结果 [Yang 等人， 2022]。参与者以1分为间隔，从5分到1分进行评分，并分别以“优秀”“良好”“一般”“较差”和“糟糕”作为标签（从最好到最差）。有关用户研究的更多细节请参见补充材料。

拟人度、语音恰当性和风格恰当性的平均意见分（平均意见分）报告于图4中。如果两个箱体的缺口不重叠，我们可以将其视为两个分布存在显著差异的有力证据 [McGill 等， 1978]。在拟人度、手势-语音以及手势风格恰当性方面，我们的方法显著优于对比的最先进方法，甚至在这三个维度上都取得了与真实标注具有竞争力的结果。Ac-

根据参与者的反馈，我们生成的手势“在语义上更加相关”、“更加自然”，并且“符合风格”，但与真实标注相比，我们的方法存在“脚部滑动”问题。这是非物理驱动的动作生成系统中一个常见的问题，可以通过后处理来解决 [Ghorbani等人，2022年；Luvizon 等人， 2023]。

4.2 手势可控性

风格控制。 假设中性音频不会影响手势风格，那么通过设置公式（8）中的 $\gamma = 1$ 和 s，我们就能利用中性语音生成风格化手势。我们从测试集中选取了两个包含中性音频的语音片段，分别生成六种风格化手势。图 5 展示了使用 t-SNE 方法对相同中性音频下不同输入风格 s 所生成的手势 $\hat{x}_t$ 进行可视化的结果。

图中还绘制了相应风格生成的人体骨架。可以发现，对于“年老”风格，腰部和膝盖弯曲幅度更大，双手基本位于膝盖或腰部；对于“悲伤”风格，头部低垂，双手位置较低；对于“放松”风格，臀部前倾，站姿较为放松；对于“愤怒”风格，双手快速上下摆动。需要注意的是，尽管从骨架的风格可视化结果来看，“中性”与“开心”风格手势之间的差异仍然相当明显，例如“开心”风格的手部位置更高、振幅更大，但它们在 t-SNE 图中几乎重叠。在我们的分析中，这是因为所谓的中性语音仍包含 WavLM 特征中的情绪和语义等信息。通过调整风格强度 γ，可以进一步控制音频 a 与风格 s 对风格的影响平衡。

风格编辑。 为了进一步分析风格强度与语音所暗示风格之间的关系，我们选择“开心”和“年长”两种风格，并在公式 8 中设置 $\gamma = 1$ 和 3。此外，为了比较结果，我们在公式 8 中设置 $\gamma = 0.5$，以插值生成不同风格。

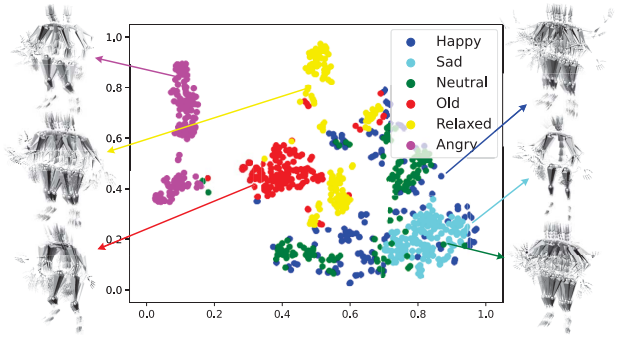


图 5：不同风格手势的 tSNE 可视化结果，以及与相应风格的骨骼手势对应的阴影图。例如，对于“老年”手势，其腰部和膝盖弯曲得更多，双手基本放在膝盖或腰部。

其他参数保持不变，并绘制了一个时长为 12 秒、FPS = 1 的手势生成结果，如图 6 所示。

如图 6 所示，使用 $\gamma = 3$ 的“快乐”风格时，身体旋转幅度和手部动作幅度都最大，手部位置也最高；相反，使用 $\gamma = 3$ 的“老年”风格时，腰部弯曲幅度最大，双手几乎没有抬起，整个动作序列也几乎没有变化。至于另外三种结果，其风格强度介于上述两者之间，并且从上到下，风格逐渐由快乐转为老年。得益于我们的模型架构，即使这些风格不同，生成的手势与语音仍更加协调。值得注意的是，当我们对 $\gamma = 0.5$ 使用“快乐”和“老年”风格时，结果更接近于对 $\gamma = 1$ 使用“快乐”风格，而“老年”风格几乎无法察觉。这一观察进一步验证了先前的发现：用于测试的“中性”语音中嵌入了“快乐”风格。这一发现很有用，例如，我们在实验中发现，如果希望控制“快乐”语音生成“悲伤”手势，$\gamma = 1$基本不起作用，因为模型可以从语音中学到快乐风格。由于语音风格与手势风格之间存在这种耦合，设置更大的 γ 就能更好地编辑风格。因此，我们可以通过风格强度生成原始数据集中不存在的手势（例如，以“悲伤”风格为“快乐”语音生成的手势）。

用户研究。 此外，我们希望进一步探究风格强度与拟人度、语音恰当性之间的关系，因此开展了一项用户研究。为了避免语音中的风格影响参与者的评分，与之前一样，我们仅控制中性语音的风格强度，然后要求参与者对上述三个维度进行评分。ExampleGesture [Ghorbani 等人， 2022] 也能够控制基于同一段语音生成不同风格的手势。因此，我们选择它作为基线模型。由于此处生成的手势并不存在于数据集中，因此使用具有中性风格的原始中性语音作为参照——

ENCE。结果如图 7 所示。

结果表明，在 $\gamma = 1$ 时，我们的模型在结果的手势风格恰当性方面与 ExampleGesture 相近，而拟人度和语音恰当性

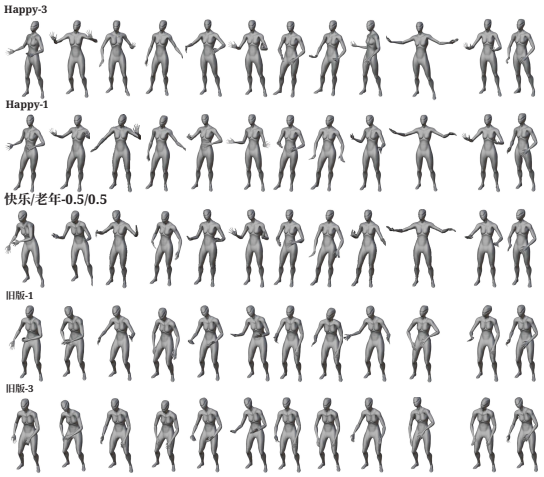


图 6：风格编辑与插值结果。从上到下，身体扭转幅度和手部动作逐渐减小，手部位置逐渐降低。尽管风格发生变化，生成的手势仍能与不同风格的语音保持良好的匹配。

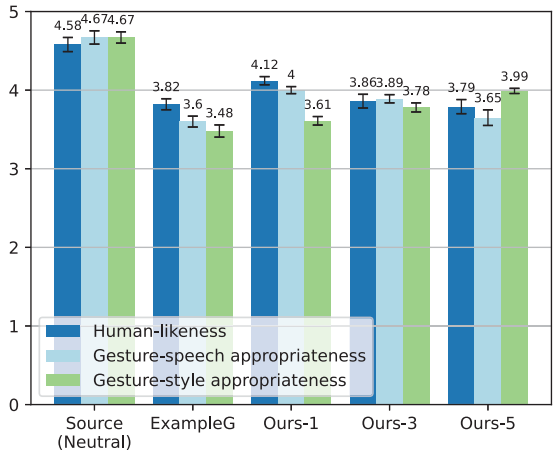


图 7：三个维度的平均意见分结果及 95% 置信区间。“Ours-γ”表示我们模型的风格控制强度为 γ。总体而言，我们的模型显著优于 ExampleGesture，并且可以编辑风格的强度。参数 γ 增加时，另外两个分数会相应降低。

则显著优于 ExampleGesture。与此同时，随着 γ 增大，风格恰当性显著提升，但另外两个维度的分数有所下降。这也符合直觉：如果“老年”风格的强度过高，手部几乎不会抬起，整个动作序列的振幅也很小，因此看起来不够拟人，与语音也不够恰当。我们还发现，与直接生成与语音对应的风格相比，生成风格控制的结果（图7）有所退化（图4）。我们认为，通过控制一种语音风格来生成另一种手势风格，本身就是一项“困难且相互冲突”的任务，因为语音风格和手势风格仍然彼此相关、相互耦合。

生成多样化手势。由于我们的模型架构，

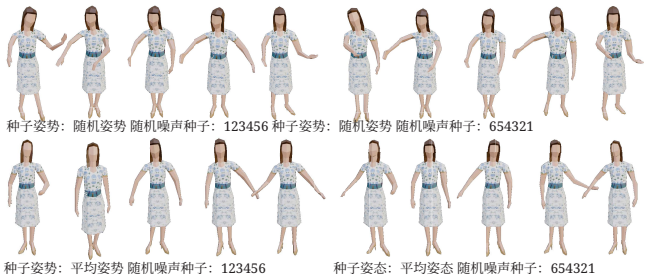


图8：生成的手势多样性的可视化。人们会在不同状态下的不同时刻做出不同的伴随语音的手势。与真实的人一样，对于相同的语音，我们的方法能够通过使用不同的种子手势或不同的噪声手势，生成不同的手势。

Name	人类相似度 ↑	手势-语音恰当性 ↑
真实标注	4.15 ± 0.11	4.25 ± 0.09
Ours	$\mathbf{4.11 \pm 0.08}$	$\mathbf{4.11 \pm 0.10}$
– WavLM	4.05 ± 0.10	3.91 ± 0.11
–交叉局部注意力	3.76 ± 0.09	3.51 ± 0.15
– 自注意力	3.55 ± 0.13	3.08 ± 0.10
–注意力 +门控循环单元	3.10 ± 0.11	2.98 ± 0.14
+前向注意力	3.75 ± 0.15	3.23 ± 0.24

表1：消融研究结果。’–’ 表示未使用的模块，’+’ 表示额外添加的模块。粗体表示最佳指标。

即使语音和风格相同，不同的噪声手势和不同的种子手势也可能生成不同的结果，如图8所示。这与真实人类语音相同，会产生与初始位置相关的多样化伴随语音的手势。我们之前的分析是针对风格维度进行的。请注意，模型还会在处理种子手势时加入随机掩码，因此也可以对不同的种子手势进行插值和外推，以控制生成不同且多样化的初始位置手势。

4.3 消融研究

此外，我们还进行了消融研究，以探究模型中不同组件对性能的影响。由于手势风格恰当性可以通过参数进行控制，并且会影响另外两个维度，为便于比较，我们将 γ 设为1，仅对拟人度和语音恰当性进行评分。消融研究的结果汇总于表1中。相关研究的可视化对比也可参见补充视频。我们探究了以下组件的有效性：（1）WavLM 特征；（2）局部注意力；（3）局部注意力模式；（4）自注意力；（5）注意力。我们分别针对这五个组件进行了实验。

用户研究。 表1中的结果表明，当我们不使用 WavLM 特征，而改用梅尔频率倒谱系数（MFCC）的前 13 个系数时，两个维度的得分均有所下降，其中语音恰当性尤为明显。这是因为预训练 WavLM 模型提取的特征包含更多信息，例如语义和情绪，有助于生成相应的手势。没有交叉局部注意力时，两个维度的得分都大幅下降。由于许多手势生成步骤只涉及短程相关性，局部注意力能够更好地捕捉局部信息，这与 [Rae 和 Razavi 2020]的观察一致。仅依赖长序列全局信息的自注意力效果较差。移除自注意力后，拟人度和手势-语音恰当性都下降得更多，这表明自注意力比局部注意力更重要，因为语音和手势之间存在固有的异步性，而且仅从一段局部窗口（近半秒）的语音中难以学习到足够的手势信息。不使用注意力时，我们用基于门控循环单元的模型替代，结果在所有模型中最差，进一步说明了注意力机制的有效性。此外，我们还尝试使用图3(b)中的注意力结构，发现效果更差。添加前向注意力与我们模型中使用的交叉局部注意力之间，唯一的区别是生成手势时会额外查看未来窗口中的语音信息。这一发现令人振奋；尽管语音和手势之间存在固有的异步性，但这在某种程度上表明，手势与当下和过去的一小段时间更相关，而与未来的一小段时间关系不大。换句话说，人们更可能先说“你好”再挥手，而不是先挥手再说“你好”。也有可能不同的人有不同的风格，而这个数据集只有一名演员，这一点还需要进一步研究。

5 讨论与结论

本文提出了 DiffuseStyleGesture，这是一种基于扩散模型的音频驱动协同手势生成方法。DiffuseStyleGesture 展现出三大主要优势：1）基于扩散模型，概率映射在提高多样性的同时，能够生成高质量、类人化的手势。2）我们的模型通过交叉局部注意力机制和自注意力机制合成的手势，能够与音频节奏和文本语义相匹配。3）采用无分类器引导训练方法，我们能够操控特定条件，即风格和初始手势，并执行插值或外推，从而实现对生成手势的高度可控性。主观评估结果表明，在音频驱动协同手势生成的时序任务中，我们的模型优于现有方法，并展现出更出色的风格操控能力。本研究仍有进一步改进的空间。例如，解决扩散方法采样步骤过多、耗时较长的问题，以便将其应用于实时系统，是我们未来的研究方向。

致谢

本研究得到中国国家自然科学基金（62076144）、深圳市科技计划（WDZC20220816140515001、JCYJ20220818101014030）以及深圳市下一代交互媒体创新技术重点实验室（ZDSYS20210623092001004）的资助。

参考文献

[Ahuja 等人，2020] Chaitanya Ahuja、Dong Won Lee、Yukiko I. Nakano 和 Louis-Philippe Morency。《伴随语音的手势动画风格迁移：一种多说话人条件混合方法》。载于《计算机视觉—ECCV》，《计算机科学讲义》第 12363 卷，第 248–265 页，2020年。[Alexanderson 等人，2020a] Simon Alexanderson、Gus- tav Eje Henter、Taras Kucherenko 和 Jonas Beskow。《使用归一化流实现风格可控的语音驱动手势合成》。《计算机图形论坛》，39(2)：487–496，2020年。[Alexanderson 等人，2020b] Simon Alexanderson、Gus- tav Eje Henter、Taras Kucherenko 和 Jonas Beskow。《使用归一化流实现风格可控的语音驱动手势合成》。《计算机图形论坛》，39(2)：487–496，2020年。[Alexanderson 等人，2021] Simon Alexanderson、´Eva Sz´ekely、Gustav Eje Henter、Taras Kucherenko 和 Jonas Beskow。《根据文本生成连贯的自发语音和手势》。CoRR，abs/2101.05684，2021年。[Alexanderson 等人，2022] Simon Alexanderson、Rajmund Nagy、Jonas Beskow 和 Gustav Eje Henter。《聆听、去噪、行动！使用扩散模型进行音频驱动的动作合成》。CoRR，abs/2211.09707，2022年。[Ao 等人，2022] Tenglong Ao、Qingzhe Gao、Yuke Lou、Baoquan C hen 和 Libin Liu。《节奏型手势生成器：基于层次化神经嵌入的节奏感知伴随语音手势合成》。《ACM 图形学汇刊》，41(6)：209:1–209:19，2022年。[Beltagy 等人，2020] Iz Beltagy、Matthew E. Peters 和 Ar- man Cohan。《Longformer：长文档 Transformer》。CoRR，abs/2004.05150，2020年。[Chang 等人，2022a] Che-Jui Chang、Sen Zhang 和 Mub- basir Kapadia。《IVI 实验室参加 2022 年 GENEA 挑战赛的方案——一种基于 Tacotron2、采用局部约束注意力机制的伴随语音手势生成方法》。载于《2022年国际多模态交互会议论文集》，ICMI ’22，第 784–789 页，2022年。[Chang 等人，2022b] Ziyi Chang、Edmund JC Findlay、Haozheng Zhang 和 Hubert PH Shum。《使用去噪扩散概率模型统一人体动作合成与风格迁移》。arXiv 预印本，arXiv:2212.08526，2022年。[Chen 等人，2022] Sanyuan Chen、Chengyi Wang、Zhengyang Chen、Yu Wu、Shujie Liu、Zhuo Chen、Jinyu Li、Naoyuki Kanda、Takuya Yoshioka、Xiong Xiao、Jian Wu、Long Zhou、Shuo Ren、Yan min Qian、Yao Qian、Jian Wu、Michael Zeng、Xiangzhan Yu 和 Furu Wei。《WavLM：面向全栈语音处理的大规模自监督预训练》。IEEE《精选信号处理主题期刊》，16(6)：1505–1518，2022年。[Dabral 等人，2022] Rishabh Dabral、Muhammad Hamza Mughal、Vladislav Golyanik 和 Christian Theobalt。《Mo-fusion：一种基于去噪扩散的动作合成框架》。CoRR，abs/2212.04495，2022年。[Ghorbani 等人，2022] Saeed Ghorbani、Ylva Ferstl、Daniel Holden、Nikolaus F Troje 和 Marc-Andr´e Carbonneau。《Zeroeggs：基于示例的语音零样本手势生成》。arXiv 预印本 arXiv:2209.07556，2022年。[Ginosar 等人，2019] Shiry Ginosar、Amir Bar、Gefen Ko- havi、Caroline Chan、Andrew Owens 和 Jitendra Malik。《学习对话手势的个体风格》。发表于 IEEE 计算机视觉与模式识别会议（CVPR），第3497–3506页，2019年。[Habibie 等人，2021] Ikhsanul Habibie、Weipeng Xu、Dushyant Mehta、Lingjie Liu、Hans-Peter Seidel、Gerard Pons-Moll、Mohamed Elgharib 和 Christian Theobalt。《从视频中学习语音驱动的三维对话手势》。发表于 IVA ’21：ACM智能虚拟代理国际会议，第101–108页，2021年。[Habibie 等人，2022] Ikhsanul Habibie、Mohamed Elgharib、Kripasindhu Sarkar、Ahsan Abdullah、Simbarashe Nyat- sanga、Michael Neff 和 Christian Theobalt。《一种基于运动匹配的框架：从语音合成可控手势》。发表于 SIGGRAPH ’22：计算机图形学与交互技术特别兴趣小组会议，第46:1–46:9页，2022年。[Ho 和 Salimans，2022] Jonathan Ho 和 Tim Salimans。《无分类器扩散引导》。CoRR，abs/2207.12598，2022年。[Ho 等人，2020] Jonathan Ho、Ajay Jain 和 Pieter Abbeel。《去噪扩散概率模型》。发表于《神经信息处理系统进展》，33:6840–6851，2020年。[Huber，1992] Peter J Huber。《位置参数的稳健估计》。发表于《统计学突破》，第492–518页，1992年。[Kim 等人，2022] Jihoon Kim、Jiseob Kim 和 Sungjoon Choi。《FLAME：基于自由形式语言的动作合成与编辑》。arXiv 预印本 arXiv:2209.00349，2022年。[Kitaev 等人，2020] Nikita Kitaev、Lukasz Kaiser 和 Anselm Levskaya。《Reformer：高效的Transformer》。发表于第8届国际学习表征会议（ICLR），2020年4月26日至30日。[Kucherenko 等人，2021] Taras Kucherenko、Patrik Jonell、Youngwoo Yoon、Pieter Wolfert 和 Gustav Eje Henter。《基于通用数据对手势生成系统进行的大规模众包评估：GENEA挑战赛2020》。发表于 IUI ’21：第26届智能用户界面国际会议，4月13日至17日，第11–21页，2021年。[Li 等人，2021a] Jing Li、Di Kang、Wenjie Pei、Xuefei Zhe、Ying Zhang、Zhenyu He 和 Linchao Bao。《Audio2gestures：利用条件变分自编码器从语音音频生成多样化手势》。发表于2021年 IEEE/CVF计算机视觉国际会议（ICCV），第11273–11282页，2021年。[Li 等人，2021b] Jing Li、Di Kang、Wenjie Pei、Xuefei Zhe、Ying Zhang、Zhenyu He 和 Linchao Bao。《Audio2gestures：利用条件变分自编码器从语音音频生成多样化手势》。发表于2021年

IEEE/CVF计算机视觉国际会议（ICCV），第11273–11282页，2021年。[Li 等人， 2022] Siyao Li、Weijiang Yu、Tianpei Gu、Chunze Lin、Quan Wang、Chen Qian、Chen Change Loy 和 Ziwei Liu。《Bailando：利用带编舞记忆的演员-评论家GPT进行3D舞蹈生成》。发表于IEEE/CVF计算机视觉与模式识别会议（CVPR），6月18日至24日，第11040–11049页，2022年。[Liang 等人， 2022] Yuanzhi Liang、Qianyu Feng、Linchao Zhu、Li Hu、Pan Pan 和 Yi Yang。《SEEG：语义增强的伴随语音手势生成》。发表于 IEEE/CVF计算机视觉与模式识别会议（CVPR），第10463–10472页，2022年。[Lin 等人， 2022] Zhenghao Lin、Yeyun Gong、Yelong Shen、Tong Wu、Zhihao Fan、Chen Lin、Weizhu Chen 和 Nan Duan。《GENIE：用于文本生成的大规模扩散模型预训练》。CoRR，abs/2212.11685，2022年。[Liu 等人， 2022] Xian Liu、Qianyi Wu、Hang Zhou、Ying- hao Xu、Rui Qian、Xinyi Lin、Xiaowei Zhou、Wayne Wu、Bo Dai 和 Bolei Zhou。《学习用于伴随语音手势生成的分层跨模态关联》。发表于 IEEE/CVF计算机视觉与模式识别会议（CVPR），第10452–10462页，2022年。[Luvizon 等人， 2023] Diogo Luvizon、Marc Habermann、Vladislav Golyanik、Adam Kortylewski 和 Christian Theobalt。《单摄像机下的场景感知3D多人动作捕捉》。CoRR，abs/2301.05175，2023年。[McGill 等人， 1978] Robert McGill、John W Tukey 和 Wayne A Larsen。《箱线图的变体》。《美国统计学家》，32(1):12–16，1978年。[Mei 和 Patel， 2022] Kangfu Mei 和 Vishal M. Patel。《VIDM：视频隐式扩散模型》。CoRR，abs/2212.00235，2022年。[Nyatsanga 等人， 2023] Simbarashe Nyatsanga、Taras Kucherenko、Chaitanya Ahuja、Gustav Eje Henter 和 Michael Neff。《数据驱动的伴随语音手势生成综合综述》。arXiv 预印本 arXiv:2301.05339，2023年。[Rae 和 Razavi， 2020] Jack W. Rae 和 Ali Razavi。《Transformer需要深层长程记忆吗？》发表于第58届计算语言学协会年会论文集（ACL 2020），线上，2020年7月5日至10日，第7524–7529页，2020年。[Ramesh 等人， 2022] Aditya Ramesh、Prafulla Dhariwal、Alex Nichol、Casey Chu 和 Mark Chen。《利用CLIP潜变量进行分层文本条件图像生成》。arXiv 预印本 arXiv:2204.06125，2022年。[Ren 等人， 2022] Zhiyuan Ren、Zhihong Pan、Xin Zhou 和 Le Kang。《Diffusion Motion：通过扩散模型生成文本引导的3D人体动作》。arXiv 预印本 arXiv:2210.12315，2022年。[Roy 等人， 2021] Aurko Roy、Mohammad Saffar、Ashish Vaswani 和 David Grangier。《利用路由Transformer实现高效的基于内容的稀疏注意力》。计算语言学协会学报，9:53–68，2021年。

[Tevet 等人， 2022] Guy Tevet、Sigal Raab、Brian Gordon、Yonatan Shafir、Daniel Cohen-Or 和 Amit H Bermano。人体运动扩散模型。arXiv 预印本 arXiv:2209.14916，2022年。[Vaswani 等人， 2017] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。Attention is all you need。神经信息处理系统进展，第30卷，2017年。[Xie 等人， 2022] Pan Xie、Qipeng Zhang、Zexian Li、Hao Tang、Yao Du 和 Xiaohui Hu。采用 CodeUNet 的向量量化扩散模型，用于生成文本到手语姿态序列。arXiv 预印本 arXiv:2208.09141，2022年。[Yang 等人， 2022] Sicheng Yang、Zhiyong Wu、Minglei Li、Mengchen Zhao、Jiuxin Lin、Liyang Chen 和 Weihong Bao。ReprGesture 在 2022年 GENEA挑战赛中的参赛作品。载于国际多模态交互会议（ICMI），11月7日至11日，第758–763页，2022年。[Yi 等人， 2022] Hongwei Yi、Hualin Liang、Yifei Liu、Qiong Cao、Yandong Wen、Timo Bolkart、Dacheng Tao 和 Michael J Black。根据语音生成整体三维人体运动。arXiv 预印本 arXiv:2212.04420，2022年。[Yoon 等人， 2019] Youngwoo Yoon、Woo-Ri Ko、Minsu Jang、Jaeyeon Lee、Jaehong Kim 和 Geehyuk Lee。机器人学习社交技能：面向人形机器人的伴随语音手势生成端到端学习。载于国际机器人与自动化会议（ICRA），第4303–4309页，2019年。[Yoon 等人， 2020] Youngwoo Yoon、Bok Cha、Joo-Haeng Lee、Minsu Jang、Jaeyeon Lee、Jaehong Kim 和 Geehyuk Lee。基于文本、音频和说话人身份三模态上下文的语音手势生成。《ACM图形学汇刊》（TOG），39(6)：1–16，2020年。[Yoon 等人， 2022] Youngwoo Yoon、Pieter Wolfert、Taras Kucherenko、Carla Viegas、Teodor Nikolov、Mihail Tsakov 和 Gustav Eje Henter。2022年 GENEA挑战赛：对数据驱动的伴语手势生成进行大规模评估。载于国际多模态交互会议（ICMI），11月7日至11日，第736–747页，2022年。[Yuan 等人， 2022] Ye Yuan、Jiaming Song、Umar Iqbal、Arash Vahdat 和 Jan Kautz。PhysDiff：基于物理引导的人体运动扩散模型。arXiv 预印本 arXiv:2212.02500，2022年。[Zhang 等人， 2022] Mingyuan Zhang、Zhongang Cai、Liang Pan、Fangzhou Hong、Xinying Guo、Lei Yang 和 Ziwei Liu。MotionDiffuse：基于扩散模型的文本驱动人体运动生成。arXiv 预印本 arXiv:2208.15001，2022年。[Zhou 和 Wang， 2022] Zixiang Zhou 和 Baoyuan Wang。UDE：用于人体运动生成的统一驱动引擎。arXiv 预印本 arXiv:2211.16016，2022年。[Zhou 等人， 2022] Chi Zhou、Tengyue Bian 和 Kang Chen。GestureMaster：基于图的语音驱动手势生成。载于国际多模态交互会议（ICMI），第764–770页，2022年。

A 概述

在这份补充文件中，我们将进一步分析实验结果。

- 我们详细介绍真实标注手势片段的特征处理流程。
- 我们介绍训练细节，包括数据集描述和实现细节。
- 我们展示所开展用户研究的详细情况。

B 真实标注手势片段

参考手势片段中的每一帧都由一个特征向量 g 表示。参考手势片段中的每一帧都由一个特征向量 $\mathbf{g} = [\mathbf{r}_p, \mathbf{r}_r, \dot{\mathbf{r}}_p, \dot{\mathbf{r}}_r, \rho_p, \rho_r, \dot{\rho}_p, \dot{\rho}_r, g_d]$ 表示，其中 $\mathbf{r}_p \in \mathbb{R}^3$ 和 $\mathbf{r}_r \in \mathbb{R}^4$ 是根节点的位置和旋转，$\dot{\mathbf{r}}_p \in \mathbb{R}^3$ 和 $\dot{\mathbf{r}}_r \in \mathbb{R}^3$ 是根节点的位置速度和旋转速度，$\rho_p \in \mathbb{R}^{3j}$ 和 $\rho_r \in \mathbb{R}^{6j}$ 是局部 3D 关节位置和局部 6D 关节旋转，$\dot{\rho}_p \in \mathbb{R}^{3j}$ 和 $\dot{\rho}_r \in \mathbb{R}^{3j}$ 是局部 3D 关节位置速度和局部 3D 关节旋转速度，j 是关节数量，$g_d \in \mathbb{R}^3$ 是注视方向。此处使用的关节数量为 75，因此 $\mathbf{g} \in \mathbb{R}^{1141}$。去噪过程中使用的手势片段 $\mathbf{g} \in \mathbb{R}^{(8+N)\times 1141}$ 为固定长度片段。

C 训练详情

数据集。 我们在 Ghorbani 等人提出的最新 Zeroeggs 数据集上进行训练和评估。该数据集包含一位以英语讲话的女性演员表演的 67 段独白序列，涵盖 19 种不同的动作风格，总时长为 135 分钟。我们选取了六种较典型且持续时间最长的风格（快乐、悲伤、中性、年老、放松、愤怒）用于训练和验证。随后按照 8:1:1 的比例将数据划分为训练集、验证集和测试集。

实现细节。 在本研究中，手势数据被裁剪为长度为 $N = 80$ （4 秒，20 fps）的片段。我们对所有关节特征维度进行标准归一化（零均值、单位方差），并对手势数据采用镜像增强和长度比例增强。交叉局部注意力网络使用 8 个头、32 个注意力通道和 256 个通道，窗口大小为 11（约 0.5 秒）；每个窗口关注其前一个窗口，并使用 0.1 的 dropout。自注意力网络由 8 层组成，使用 8 个头、32 个注意力通道和 256 个通道，dropout 为 0.1。使用 AdamW [Loshchilov 和 Hutter2019] 优化器（学习率为 3×10^{-5}），批大小为 384，共训练 300000 个样本。我们的模型采用 $T = 1000$ 个加噪步骤和余弦噪声调度进行训练。整个框架在一块 NVIDIA V100 GPU 上约需 3 天完成训练。

D 用户研究

对于每种方法，我们从测试集中选取了 10 段测试语音及其对应的测试动作，用于评估。我们确保评估中没有以“悬而未决”方式结束的口语短语。实验

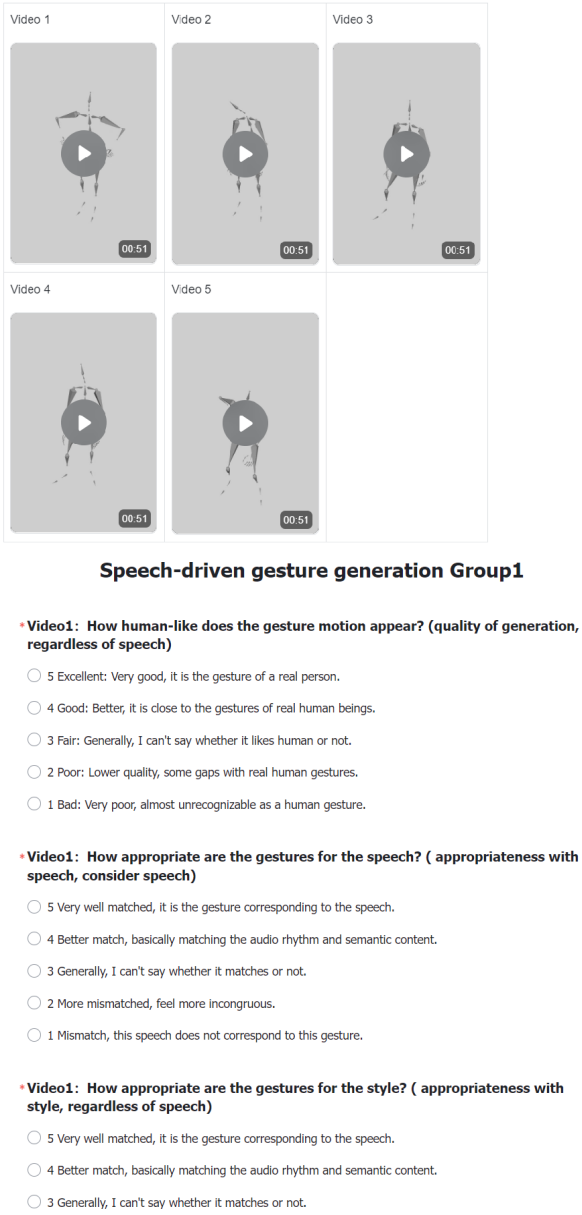


图9：用户研究中用于与现有方法进行比较的并行评分界面截图。

分别由 18 位参与者进行。生成的手势数据通过 Blender [1] 渲染进行可视化。

实验邀请英语熟练度较高的参与者，对拟人度、手势与语音的恰当性以及手势风格恰当性进行评估。在评估过程中，我们提示参与者忽略手指动作。他们需要观看视频，然后对这三个维度进行评分。参与者几乎都是我们实验室的学生 [2]，另有少数来自其他学校，年龄在20至30岁之间。

在拟人度评估中，每个评估页面都会询问

[1] https://www.blender.org/ [2] https://thuhcsi.github.io/

参与者：“手势动作看起来有多像真人？”在手势与语音适切性评估中，每个评估页面都会询问参与者：“手势与语音的配合有多恰当？”至于手势与风格的适切性评估，参与者会被问到：“手势与风格的配合有多恰当？”每个页面都会展示五段模型生成的视频，供参与者按照从 5 到 1 的分数进行评分，分值间隔为 1 分，评分标签从优到劣依次为“优秀”“良好”“一般”“较差”和“差”。

不同维度下不同模型的评估界面示例见图9。参与者表示，我们的框架生成的手势包含许多语义相关和节奏相关的手势。更多可视化结果请参阅我们的补充视频。

E 更多讨论

关于可控性。由于我们使用独热编码来表示风格，因此只需控制方程（8）中的 γ 即可。研究发现，风格 γ 为1的语音具有更明显的对应风格；自然语音或与预期相反的语音（例如由快乐语音生成悲伤手势）可能需要将 γ 设为2或3，才能呈现明显的风格。这是因为语音和风格本身是耦合在一起的。

附录参考文献

[ALoshchilov and Hutter， 2019] Ilya Loschonov and Frank Hutter。解耦权重衰减正则化。载于第七届国际学习表征会议，*ICLR*，*5*月*6*日至*9*日，2019年。