arXiv:2301.10047v2 [cs.GR] 2 Feb 2023

DiffMotion：使用去噪扩散模型进行语音驱动手势合成

Fan Zhang[1], [2][0000−0002−9534−1777]、Naye Ji[2](✉)[0000−0002−6986−3766]、Fuxing Gao[2],和 Yongping Li[1,3]

[1] 中国澳门澳门科技大学人文艺术学院[2] 中国浙江传媒学院传媒工程学院 {fanzhang,jinaye,fuxing}@cuz.edu.cn[3] 宁波财经学院数字技术与工程学院 liyongping@nbufe.edu.cn

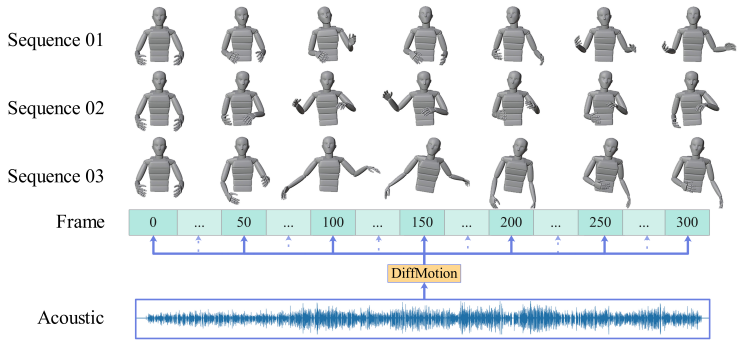

图 1. DiffMotion 的随机样本即使在输入语音音频相同的情况下，也能在序列内部和序列之间生成许多各异但自然的输出手势。

摘要。 语音驱动手势合成是虚拟人创作领域中日益受到关注的方向。然而，一个关键挑战在于语音与手势之间固有的复杂一对多映射。以往研究探索并借助生成模型取得了显著进展。尽管如此，大多数合成手势仍远不够自然。本文提出 DiffMotion，一种基于扩散模型的新型语音驱动手势合成架构。该模型由自回归时序编码器和去噪扩散概率模块组成。编码器提取语音输入和历史手势的时序上下文。扩散模块学习一个参数化马尔可夫链，逐步将简单分布转换为复杂分布，并根据伴随语音生成手势。与基线方法相比，客观和主观评估

证实了我们的方法能够生成自然且多样的手势动作，并展示了基于扩散的模型在语音驱动手势合成方面的优势。项目页面：https://zf223669.github.io/DiffMotionWebsite/。

关键词： 手势生成· 手势合成· 跨模态· 语音驱动· 扩散模型。

1 引言

近年来，随着元宇宙的兴起，3D 虚拟人技术日益流行。为了塑造自然且令人产生共鸣的角色，一项主要任务是实现自然且与语言表达相匹配的非语言（伴随语音）手势，使其能够像人类一样进行交流。迄今为止，尽管动作捕捉系统能够完成这项任务，但它们需要专用硬件、场地和演员，成本可能很高。自动生成是生成手势成本最低的方式，在制作过程中无需人工参与。语音驱动手势生成是一种很有潜力的解决方案。然而，根据输入语音生成相关且时机恰当的手势，面临的主要挑战是语音与手势之间固有的跨模态一对多映射。也就是说，这种关联难以建模：同一话语往往会伴随明显不同的手势，且在不同时间发生，即使说话者相同或不同[16]。

在一对一映射的假设下，以往基于规则的方法或确定性深度学习方法都无法完成这一任务。前者受限于给定的手势单元，导致动作重复；后者通过最小化均方误差进行训练，因此容易趋向于平均姿势。因此，当前研究已转向概率生成模型（如生成对抗网络、变分自编码器和归一化流）。尽管如此，与原始动作捕捉数据集[31]相比，大多数合成手势仍显著缺乏自然感。扩散模型能够生成高质量、多样化的样本，在各种生成任务中展现出令人瞩目的效果。然而，扩散模型在语音驱动手势合成任务中仍鲜少受到关注。

本文提出**DiffMotion**，一种新颖的基于扩散的概率架构，用于语音驱动手势生成。该模型使用大量未经结构化且未经人工标注的手势数据进行训练。此外，如图1所示，我们的方法能够估计自然、多样的手势，甚至能够生成数据集中未出现过的手势。我们的贡献如下：

(1) 我们提出了 DiffMotion，这是首个基于扩散的生成模型，用于解决跨模态语音驱动手势合成问题。(2) 我们创新性地结合了自回归时序编码器和去噪扩散概率模块，该模块能够学习手势与伴随语音之间的复杂规律，并生成符合语音节奏的逼真、多样的动作。(3) 实验表明，无论从客观还是主观角度来看，DiffMotion 的表现均优于最先进的基线方法。

2 相关工作

由于当前研究正从确定性转向生成式模型，我们仅简要讨论语音驱动手势生成的新颖方法。

Ylva 等人[29]提出使用多个判别器的生成对抗网络[10] ，将语音转换为三维手势动作。尽管这种方法在标准回归损失训练方面取得了显著改进，但数据集需要手工标注。此外，他们承认，由于将其视为手势阶段分类任务，生成的动作缺乏真实感且存在不连续现象。值得注意的是，Alexanderson 等人 [24] 利用概率生成模型——即能够处理精确对数似然和潜变量推断的归一化流 [5][6], ——构建了一个基于 Glow[7,18] 的网络，该网络源自归一化流和循环神经网络，成功地对给定语音输入时手势的条件概率分布进行建模，并能根据相同的语音信号生成多种动作。此后，Taylor 等人 [23] 将变分自编码器与归一化流结合进行扩展，证明该方法只需使用一小部分可训练参数，就能生成富有表现力且接近真实值的身体动作。尽管归一化流足以捕捉高维复杂性且仍可训练，但这些方法需要对变换施加拓扑约束[5,6,32]。

扩散模型（ [27]中的一项综述）采用更灵活的架构，通过参数化马尔可夫链逐步将简单分布转换为复杂的数据分布，并可通过优化变分下界高效训练。在图像合成[4,9], 领域超越 GAN 后，扩散模型在计算机视觉[15], 和自然语言处理[1]等多种生成任务中展现出卓越的效果。尤其是，扩散模型在多模态建模 [2,33],和时间序列预测 [22],方面取得了令人瞩目的成果，在跨模态序列到序列建模与生成领域具有巨大潜力。然而，它在语音驱动手势合成任务中尚未受到足够关注。受上述讨论启发，我们设计了一种基于扩散模型的架构，以探索扩散模型在语音驱动手势生成中的能力。

3 我们的方法

我们提出了用于跨模态语音驱动手势合成的 **DiffMotion**。该模型旨在根据语音特征和历史手势生成手势动作。本节首先在第 3.1 节对问题进行形式化定义，然后在第 3.2 节详细介绍 DiffMotion 架构。最后，在第 3.3 节和第 3.4 节介绍训练和推理过程。

3.1 问题 formul化

我们将手势特征和声学信号分别记为 $x_t^0 \in [x_1^0, ..., x_t^0, ..., x_T^0]$ 和 $c_t \in [c_1, ..., c_t, ..., c_T]$, 其中 $x_t^0 = \mathbb{R}^D$ 表示帧

t 处的三维骨架关节角度， D 表示骨架关节的通道数。 c_t 是当前的声学子序列信号，由从语音声学特征序列 $[a_1, ..., a_T]$ 中截取的子序列构成， T 是序列长度。令 $p_\theta(\cdot)$ 表示概率密度函数（PDF），其旨在近似真实的手势数据分布 $p(\cdot)$，并便于采样。给定当前帧 t 之前的历史帧 t'，以及过去的窗口帧 τ 和初始动作 $x^0_{1:\tau}$，我们的任务是根据姿态 $x^0_t \sim p_\theta(\cdot)$ 在给定历史姿态 $x^0_{t'-\tau:\ t'-1}$ 和作为协变量的声学信号 c_t 时的条件概率分布，逐帧生成该姿态：

$$\begin{aligned} x_t^0 \sim p_\theta\left(x_t^0 | x^0_{t-\tau:t-1}, c_t\right) \approx p(\cdot) &:= p\left(x_t^0 | x^0_{t-\tau:t-1}, c_t\right) \\ &:= p\left(x^0_{1:\tau}\right) \cdot \prod\nolimits_{t'=\tau+1}^{t} p\left(x^0_{t'} | x^0_{t'-\tau:t'-1}, c_{t'}\right) \end{aligned} \tag{1}$$

自回归时序编码器提取条件信息，而 $p_\theta(\cdot)$ 旨在逼近由去噪扩散模块训练的 $p(\cdot)$。我们将在第 3.2 节详细讨论这两个模块。

3.2 DiffMotion 架构

DiffMotion 架构由两个模块组成：自回归时序编码器（AT-Encoder）和去噪扩散概率模块（DDPM）。整体架构如图 2 所示。

自回归时序编码器（AT-Encoder）。AT-Encoder 采用多层长短期记忆网络，对语音声学特征的时序上下文以及截至帧 $t-1$ 的过往姿势进行编码，得到更新后的隐藏状态 h_{t-1}：

$$h_{t-1} = g\left(x^0_{t-\tau:t-1}, c_t, h_{t-2}\right) = \mathrm{LSTM}_\theta\left(\mathrm{concatenate}\left(\mathrm{x^0_{t-\tau:t-1}}, \mathrm{c_t}\right), \mathrm{h_{t-2}}\right), \tag{2}$$

其中， h_t 表示长短期记忆网络隐藏状态的演变。长短期记忆网络 $_\theta$ 通过共享权重 θ 和 $h_0 = 0$ 进行参数化。接着，我们使用一系列中性（平均）姿势作为初始动作 $x_{1:\tau}$。因此，我们可以通过动作扩散模型近似公式（1），其负对数似然（NLL）为：

$$p_\theta\left(x^0_{1:\tau}\right) \cdot \prod\nolimits_{t'=\tau+1}^{t} p_\theta\left(x^0_{t'} | h_{t'-1}\right), \quad NLL := \Sigma^t_{t'=\tau+1} - \log p_\theta\left(x^0_{t'} | h_{t'-1}\right) \tag{3}$$

去噪扩散概率模块(DDPM)。 DDPM 是一种形式为 [25] 的潜变量模型 p_θ：$= \int p_\theta\left(x^{0:N}\right) dx^{1:N}$，其中 $x^1, ..., x^N$ 是与第 n 个扩散时间阶段的数据 x^n 具有相同维度的潜变量。该模块包含两个过程，即扩散过程 和 生成过程。在训练时间，扩散过程通过优化数据似然的变分界，逐渐将原始数据（x^0）转换为白噪声（x^N）。在推理时，生成过程通过使用朗之万采样的马尔可夫链[19]逆转这一加噪过程，从而恢复数据。

手势可以通过在每一帧从条件数据分布中采样生成，然后反馈给 AT-Encoder 以生成下一帧。扩散过程和生成过程中的马尔可夫链如下：

$$
\begin{aligned}
&p\left(x^n|x^0\right) = \mathcal{N}\left(x^n; \sqrt{\overline{\alpha}^n}x^0, \left(1-\overline{\alpha}^n\right)I\right) \quad and \\
&p_\theta\left(x^{n-1}|x^n, x^0\right) = \mathcal{N}\left(x^{n-1}; \tilde{\mu}^n\left(x^n, x^0\right), \tilde{\beta}^n I\right),
\end{aligned} \tag{4}
$$

其中 $\alpha^n := 1-\beta^n$ 且 $\alpha^n := \prod_{i=1}^{n}\alpha^i$。如 [9]所示，$\beta^n$ 是一个递增的方差调度 $\beta^1, \ldots\ldots \beta^N$，其中 $\beta^n \in (0, 1)$，且 $\tilde{\beta}^n := \frac{1-\alpha^{n-1}}{1-\alpha^n}\beta^n$。

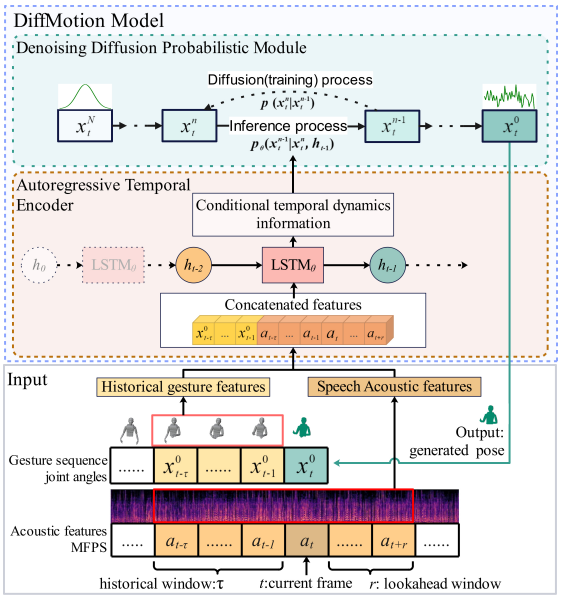

图2。 DiffMotion 示意图。该模型由一个自回归时间编码器和一个去噪扩散概率模块组成。

3.3 训练

训练目标是优化参数 θ，使其通过真实噪声 $\epsilon \sim \mathcal{N}(0, I)$ 与预测噪声 ϵ_θ 之间的均方误差（MSE）损失来最小化负对数似然：

$$
\mathbb{E}_{x_t^0,\epsilon,n}[||\epsilon - \epsilon_\theta\left(\sqrt{\overline{\alpha}^n x_t^0} + \sqrt{1-\overline{\alpha}^n}\epsilon, h_{t-1}, n\right)||^2], \tag{5}
$$

这里的 ϵ_θ 是一个神经网络，它使用输入 x_t^0、 h_{t-1} 和 n 来预测ϵ，并采用了与 [21] 相似的架构。完整的训练流程如算法 1 所示。

算法 1： 逐帧训练 $t \in [\tau+1, T]$

输入： 数据 $x_t^0 \sim p\left(x_t^0 | x_{t-\tau:t-1}^0,\ c_t\right)$ 和 LSTM 状态 h_{t-1}

重复

初始化 $n \sim$均匀分布(1，.....， N) 和 $\epsilon \sim \mathcal{N}(0,\ I)$

对其执行梯度更新

$$\nabla_\theta ||\epsilon - \epsilon_\theta \left(\sqrt{\overline{\alpha}_n} x_t^0 + \sqrt{1-\overline{\alpha}_n}\epsilon, h_{t-1}, n\right)||^2$$

直到 收敛;

3.4 推断

训练后，我们希望使用变分推断生成与原始数据分布匹配的新手势（$x_t^0 \sim p_\theta\left(x_t^0 | x_{t-\tau:t-1}^0,\ c_t\right)$）。首先，我们将 AT-Encoder 应用于初始动作 $x_{1:\tau}$ 的中性姿态序列，通过公式 2 得到隐藏状态 h_{t-1}。然后，我们按照算法 2 中的采样过程，获取当前帧的一个样本 x_t^0。 σ_θ 是 $p_\theta\left(x^{n-1}|x^n\right)$的标准差。我们选择 $\sigma_\theta := \tilde{\beta}^n$。

算法 2： 通过退火朗之万动力学采样 x_t^0

输入： 噪声 $x_t^N \sim \mathcal{N}(0, I)$ 和状态 h_{t-1}

对于 $n = N$ **到** 1，**执行**

如果 $n > 1$，**则**

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

else

$z = 0$

结束如果

$x_t^{n-1} = \frac{1}{\sqrt{\alpha^n}}\left(x_t^n - \frac{\beta^n}{\sqrt{1-\alpha^n}}\epsilon_\theta\left(x_t^n, h_{t-1}, n\right)\right) + \sqrt{\sigma_\theta}z$

结束循环

返回： $x^0{}_t$

推理时，将过去的姿态 $[x_{t-\tau-1}^0, ..., x_{t-1}^0]$ 和声学特征$[a_{t-\tau}, ...a_{t+r}]$ 拼接后输入 AT-Encoder 提取上下文，再将其作为条件信息传递给扩散模块，以生成当前手势（x_t^0）。DiffMotion 每帧输出一个手势，并将其反馈至已生成的手势序列中。同时，过去姿态窗口从 $[t-\tau, t-1]$ 滑动至 $[t-\tau+1, t]$，声学窗口也向前移动一帧，以生成下一个手势（x_{t+1}^0），如图 2 所示。

4 实验

我们从客观和主观两个方面将 DiffMotion（**扩散模型**）与先前的基线方法进行比较。为确保公平且便于比较，我们选择基线方法时遵循以下标准：1）采用GENEA研讨会[30] 推荐的 Trinity手势数据集[13]；2）骨架结构与扩散模型一致；3）关节角度以指数映射[8]表示；4）提供开源代码。基于流的方法 StyleGestures（**SG**）[24],符合上述要求。此外，由于 Audio2Getsture 方法（**AG**）[11]的输出格式与**扩散模型** 和**SG**略有不同，因此仅对其进行主观评估。所有实验均包含真实值（**真实值**）。我们重点研究三维上半身节拍手势，这类手势占所有伴随语音的手势的 50% 以上，并且在节奏上与相应语音[17]相联系。

4.1 训练数据处理

我们使用的 Trinity手势数据集包含 23 段录制，总计 244 分钟的动作捕捉数据和音频，内容是一位以英语为母语的男性在不同话题下进行即兴演讲。演员的动作由 20 台 Vicon 摄像机以 59.94 帧每秒（fps）的速度捕捉，骨架包含 69 个关节。

我们选取了从第一个脊柱关节到双手和头部的上半身骨架，并排除了手指动作，仅保留 15 个上半身关节。我们遵循 [24] 提出的数据处理方法，获得了 20,665$\times$2 个样本。每个样本以 80×27 语音特征（80 帧（4 秒）的 27 通道 Mel 频率功率谱，MFPS）作为输入，以 80×45 关节角度特征作为输出。帧率从 60 fps 下采样至 20 fps。每个关节角度均用指数映射表示，以避免不连续性。

4.2 模型设置

AT-Encoder 中的长短期记忆网络包含 2 层，隐藏状态为 $h_t \in \mathbb{R}^{512}$。我们采用了 [22], 提出的类似网络结构，用于 ϵ_θ。四元方差调度从 $\beta_1 = 1 \times 10^{-4}$ 开始，至 $\beta^N = 0.1$结束。我们在推理阶段将分位数设为=0.5，以获取中位数。

实际上，手势必须提前准备好 [3,12]。我们让时刻 t 的控制输入 c_t 不仅包含当前的声学特征 a_t，还包含周围 $c_t = a_{t-\tau:t+r}$的窗口，其中前瞻 $r = 20$ 的设置确保能够考虑足够多的未来信息，如图 2 所示。为避免动画抖动，我们使用 Savitzky-Golay平滑滤波器[20] 沿帧方向对序列的所有关节通道进行滤波，并将窗口长度设为 31、多项式阶数设为 4。

该模型基于 TorchLightning 框架构建，批量大小为 80，使用 Adam优化器，学习率为 1.5×10^{-3}。所有实验均在 Intel i9 处理器和单块 NVIDIA GTX 3090 GPU 上运行。

4.3 客观评估

我们从真实感、时间一致性和多样性三个方面，对 DiffMotion **扩散模型** 、**SG** 和**真实值**进行定量比较：1）**真实感**：我们采用 L_1 关节位置距离[11] 和三维关键点正确率（PCK）[28] 来评估生成动作的真实感。2）**时间一致性**：引入节拍一致性分数（BCS）指标[11] ，用于衡量动作与语音的节拍相关性（时间一致性）。3）**多样性**：多样性指标用于评估生成的手势之间以及同一序列中的变化。对于相同的语音输入，我们合成多个动作序列，每个序列采样 N 次，然后将每个序列切分为长度相等且互不重叠的片段。最后，计算所有片段的平均 L_1 距离。

结果列于表1。定量结果表明，我们的方法在真实感和时间一致性方面优于**SG** 。在多样性指标上，我们的模型得分更高（平均为 6.25），优于**SG**（3.3）和**真实值**（2.6）。这表明**扩散模型** 和**SG** 都能够生成**真实值**中未出现过的新颖手势；由于扩散模型能够采样到范围更广的样本，因此我们的方法取得了比**SG**更好的结果[22]。

表1. 定量结果。 ↑ 表示数值越高越好， ↓ 表示数值越低越好。我们进行了 20 次测试，并报告其平均分和最高分（括号内）。

方法	L_1↓	PCK↑	BCS↑	多样性↑
扩散模型（我们的方法）	**11.6(10.26)**	**0.61(0.35)**	0.79(0.80)	**6.25(6.31)**
SG	16.35(15.22)	0.41(0.50)	0.68(0.69)	3.3(3.6)
GT	0	0	**0.93**	2.6

表2。参数数量、训练时间和每帧平均合成时间及其95%置信区间。

方法	参数数量↓	训练时间↓	合成时间↓	
扩散模型（我们的方法）	**10.4M**	**8.02分钟**	1.60±0.06秒	
SG	109.34M	6.2H	**0.08±0.06秒**	

我们还在表2中报告了参数数量、训练时间和每帧平均合成时间。结果表明，**扩散模型**(10.4M)的参数数量远少于**SG**(109.34M)。我们的方法训练时间更短（8.02分钟），而**SG**需要6.2小时。然而，由于扩散模型固有的特性，[27], **扩散模型**(1.60±0.06)的生成阶段耗时比SG(0.08±0.06)更长。

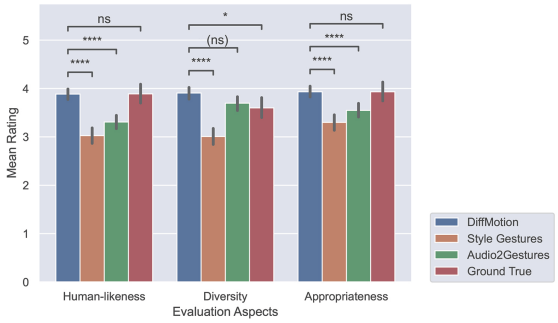

图 3. 平均评分及其 95%置信区间。星号表示显著效应（*: $p < 0.05$，****: $p < 1.00e-04$，ns：差异不显著）。

表 3. 平均感知评分。

模型	拟人度↑	多样性↑	适当性↑
DM	**3.89±0.95**	**3.91±0.96**	**3.93±0.93**
SG	3.03±1.34	3.01±1.43	3.30±1.36
AG	3.61±1.12	**3.89±0.92**	3.55±1.12
GT	**3.89±0.93**	3.60±1.02	**3.93±0.92**

4.4 主观评估

语音驱动手势生成的最终目标是产生自然、令人信服的动作。考虑到手势合成的客观评估通常较为棘手，而且并不总能转化为人类观察者所感知到的更高主观质量 [24,26], ，我们进行了主观人类感知评估。我们设计了一套问题，包含三个评估维度，并采用五点李克特量表对基线方法（**SG**、**AG**）、**扩散模型**和**真实值**进行主观评估。三个维度分别是拟人度、多样性和适当性：1）**拟人度**：生成的手势是否自然，且看起来像真实人类的动作，不考虑语音；2）**多样性**：哪种动作包含更多手势模式；3）**适当性**：时间一致性，即生成的手势是否与语音的节奏相匹配。

首先，我们训练了每个模型，并使用相同的语音音频作为输入生成了 20 个片段。每个片段时长为 18 秒。接着，我们随机选取每个模型生成的 3 个片段进行评估。然后，我们使用 GENEA_可视化工具[14]构建视频，供用户研究使用。

我们招募了 30 名志愿参与者，其中男性 16 名、女性 14 名，年龄为 19 至 23 岁。他们（20 名来自中国，10 名为来自美国、英国等地的国际学生）都擅长英语。我们请他们按照评估维度进行评分，分数范围为 1 至 5 分，分别代表最差和最好。

首先，我们向所有参与者介绍了该方法，并向他们展示了一些不属于估值集的示例片段。参与者完全理解流程后，我们开始了正式实验。所有参与者都

被要求佩戴耳机并坐在电脑屏幕前。实验环境安静且没有干扰。参与者不知道每段视频分别对应哪种方法。视频播放顺序随机，但确保每段视频都出现三次，由参与者观看并评分。

我们采用单因素方差分析来确定各模型在三个评估方面的得分是否存在差异。结果如图 3 和表 3 所示。我们提出的**扩散模型** 的平均评分与另外两个基线模型之间存在统计学上的显著差异，而与**真实值**之间则不存在统计学上的显著差异。**扩散模型** 和**真实值** 的拟人度得分分别为 3.89 ± 0.95 和 3.89 ± 0.93，适当性得分分别为 3.93 ± 0.93 和 3.93 ± 0.92。有趣的是，在多样性评估方面，**扩散模型** (3.91 ± 0.96) 与**AG**(3.89 ± 0.92)之间不存在显著差异，且二者的得分均高于**真实值**(3.60 ± 1.02)。这些结果表明，**扩散模型** 生成更丰富、真实值中未出现的手势的能力不逊于**AG** 。结果表明，所提出的方法优于以往最先进的方法，并证明基于扩散的方法有益于语音驱动手势生成任务。

4.5 消融实验

我们发现，扩散步数 N 是影响手势生成质量和效果的关键超参数。尽管较大的 N能够使 x^N 近似服从高斯分布[25], ，但会增加推理过程所需的时间。为平衡生成效率与效果，我们调整了提前停止的轮数（50 轮），并在其他超参数均保持不变的情况下，评估了 $N = 1$，50，100，200，500，.....1000，.....2500。结果见表 4。较快的生成速度会导致抖动，出现在 $N < 100$时。
$N \geq 100$时可以实现连续运动（尤其是 $N = 100$时，我们取得了最佳结果）。然而，随着 N增加，耗时也会大幅增加。尽管如此， $N > 1000$时，模型偶尔会生成异常姿势，因为扩散步数会破坏原始信息[32]的细节。

Ta表 4. 扩散步数的评估　　骤 N。

指标	1	100	300	500	600	800	1000	1500	2000	2500
合成时间 每帧（秒）	0.004 ±0.001	0.34 ±0.02	0.97 ±0.04	1.60 ±0.06	1.91 ±0.06	2.55 ±0.08	3.23 ±0.10	4.81 ±0.14	6.43 ±0.19	8.02 ±0.23
训练时间 （分钟）	8.63 ±0.06	8.63 ±0.07	8.62 ±0.10	8.51 ±0.08	8.55 ±0.09	8.52 ±0.08	8.93 ±0.08	8.84 ±0.08	8.74 ±0.07	8.67 ±0.08
训练损失 验证损失	0.99 0.99	0.32 0.34	0.20 0.22	0.17 0.19	0.16 0.18	0.15 0.16	0.13 0.15	0.12 0.13	0.10 0.11	0.09 0.10

5 结论

本文提出了一种用于自动合成伴语手势的新型框架 **DiffMotion**。该框架由自回归时序编码器和去噪扩散概率模块组成。该架构能够学习手势与其伴随语音之间复杂的一对多映射，并生成自然、多样且时机恰当的三维伴语手势。实验表明，我们的系统在定量和定性方面均优于以往的基线方法。不过，该系统仍存在一些局限性。例如，模型倾向于生成缺乏舒展感的冗余动作，而且由于 DDPM 固有的特性，推理速度较慢。未来，我们计划尝试以下方向：1）考虑为动作舒展留出呼吸空间；2）引入语义和情感表达等多模态特征；3）研究能够让用户与虚拟人交互界面进行交互的实时系统。

致谢。本研究得到中国浙江省重点项目和发展项目（编号：2021C03137）、中国浙江省公益性技术应用研究项目（编号：LGF22F020008）以及浙江省影视媒体技术重点实验室（编号：2020E10015）的资助。

参考文献

1. Austin, J., Johnson, D.D., Ho, J., Tarlow, D., van den Berg, R.：离散状态空间中的结构化去噪扩散模型。神经信息处理系统进展 **34**，17981–17993（2021） 2. Avrahami, O., Lischinski, D., Fried, O.：用于自然图像文本驱动编辑的混合扩散。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第18208–18218页（2022） 3. David, M.：手势与思维。芝加哥大学出版社（2008） 4. Dhariwal, P., Nichol, A.：扩散模型在图像合成方面胜过生成对抗网络。神经信息处理系统进展 **34**，8780–8794（2021） 5. Dinh, L., Krueger, D., Bengio, Y.：NICE：非线性独立成分估计。arXiv预印本 arXiv:1410.8516（2014） 6. Dinh, L., Sohl-Dickstein, J., Bengio, S.：使用Real NVP进行密度估计。arXiv预印本 arXiv:1605.08803（2016） 7. Eje, H.G., Simon, A., Jonas, B.：MoGlow：使用归一化流进行概率式、可控的动作合成。ACM图形学汇刊 **39**(6)，1–14（2020） 8. Grassia F. Sebastian：使用指数映射对旋转进行实用参数化。Journal of Graphics Tools **3**(3)，29–48（1998） 9. Ho, J., Jain, A., Abbeel, P.：去噪扩散概率模型。神经信息处理系统进展 **33**，6840–6851（2020） 10. Ian, G., Jean, P.A., Mehdi, M., Bing, X., David, W.F., Sherjil, O., Aaron, C., Yoshua, B.：生成对抗网络。神经信息处理系统进展 **27**（2014）

11. Jing, L., Di, K., Wenjie, P., Xuefei, Z., Ying, Z., Zhenyu, H., Linchao, B.：Audio2gestures：使用条件变分自编码器从语音音频生成多样化手势。载于：IEEE/CVF国际计算机视觉会议论文集。第11293–11302页（2021） 12. Kendon, A.：手势动作与语音：话语过程的两个方面。《言语与非言语交际的关系》**25**（1980），207–227（1980） 13. Kucherenko, T., Jonell, P., Yoon, Y., Wolfert, P., Henter, G.E.：GENEA挑战赛2020：基于通用数据对手势生成系统进行基准测试。载于：GENEA研讨会2020（2020） 14. Kucherenko, T., Jonell, P., Yoon, Y., Wolfert, P., Henter, G.E.：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。载于：第26届国际智能用户界面会议。第11–21页（2021） 15. Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., Chen, Y.：SRDiff：使用扩散概率模型进行单图像超分辨率。Neurocomputing**479**，47–59（2022） 16. Matthew, B.：语音操控木偶。载于：第26届计算机图形学与交互技术年会论文集。第21–28页（1999） 17. McNeill, D.：《手与心智：手势揭示的思维》。视觉符号学进展，第351页（1992） 18. P., K.D., Prafulla, D.：Glow：采用可逆1×1卷积的生成流。arXiv预印本 arXiv:1807.03039（2018） 19. Paul, L.：论布朗运动理论。《法国科学院院报》**65**(11)，146,530–533（1908），出版方：美国物理教师协会 20. Press, W.H., Teukolsky, S.A.：Savitzky-Golay平滑滤波器。Computers in Physics **4**(6)，669–672（1990） 21. Rasul, K., Seward, C., Schuster, I., Vollgraf, R.：用于多变量概率时间序列预测的自回归去噪扩散模型。载于：国际机器学习会议。第8857–8868页（2021） 22. Rasul, K., Sheikh, A.S., Schuster, I., Bergmann, U., Vollgraf, R.：通过条件归一化流进行多变量概率时间序列预测。arXiv预印本 arXiv:2002.06103（2020） 23. Sarah, T., J onathan, W., David, G., Iain, M.：《使用条件流变分自编码器的语音驱动对话代理》。载于：欧洲视觉媒体制作会议。第1–9页（2021） 24. Simon, A., Eje, H.G., Taras, K., Jonas, B.：使用归一化流进行风格可控的语音驱动手势合成。载于：Computer Graphics Forum，第39卷，第487–496页。Wiley在线图书馆（2020），第2期 25. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.：使用非平衡热力学进行深度无监督学习。载于：国际机器学习会议。第2256–2265页。PMLR（2015） 26. Wolfert, P., Robinson, N., Belpaeme, T.：具身对话代理中手势生成评估实践综述。IEEE Transactions on Human-Machine Systems（2022） 27. Yang, L., Zhang, Z., Hong, S., Zhang, W., Cui, B.：扩散模型：方法与应用综述（2022年9月） 28. Yi, Y., Deva, R.：使用灵活部件混合模型进行人体姿态检测。IEEE模式分析与机器智能汇刊 **35**(12)，2878–2890（2012）

29. Ylva, F., Michael, N., Rachel, M.：多目标对抗式手势生成。载于：运动、交互与游戏。第 1–10 页。ACM，英国泰恩河畔纽卡斯尔（2019） 30. Ylva, F., Rachel, M.：探究循环运动建模在语音手势生成中的应用。载于：第 18 届智能虚拟代理国际会议论文集。第 93–98 页（2018） 31. Yoon, Y., Wolfert, P., Kucherenko, T., Viegas, C., Nikolov, T., Tsakov, M., Henter, G.E.：GENEA挑战赛2022：对数据驱动的伴随语音手势生成进行大规模评估（2022） 32. Zhang, Q., Chen, Y.：扩散归一化流。神经信息处理系统进展 **34** （2021） 33. Zhu, Y., Wu, Y., Olszewski, K., Ren, J., Tulyakov, S., Yan, Y.：用于跨模态和条件生成的离散对比扩散（2022）