中图法分类号:TP37;TP18　　**文献标识码**: A　　**文章编号**: 1006-8961(2026)11-3991-14

论文引用格式:Xuan E Y, Li Y, Li Z W, Yao M M and Guo R Z. 2026. Adaptive diffusion model for co-speech holistic motion generation. Journal of Image and Graphics, 31(11):3991-4004(宣恩允, 李游, 李梓维, 姚萌萌, 郭仁忠. 2026. 用于语音驱动整体人体运动的自适应扩散模型. 中国图象图形学报, 31(11):3991-4004)[DOI:10.11834/jig.250531]

用于语音驱动整体人体运动的自适应扩散模型

宣恩允[1,2],李游[1,3*],李梓维[1,2],姚萌萌[1],郭仁忠[1,3]

1. 人工智能与数字经济广东省实验室(深圳),深圳　518083; 2. 深圳大学计算机与软件学院,深圳　518060;
3. 深圳大学建筑与城市规划学院/智慧城市研究院,深圳　518060

摘　要: **目的**　语音驱动的整体运动生成旨在同时实现富有表现力的手势和与语音精确同步的面部表情。这两个任务具有不同本质:手势生成是非确定性的,同一段语音可对应多种自然动作,需要高多样性;而面部表情生成是确定性的,需要与音素精确对应,要求高准确性。现有方法面临3个关键局限:1)采用固定架构设计强制施加任务间关系,阻碍模型捕捉手势与表情之间的真实动态联系;2)使用人工设计的静态损失权重,无法适应训练过程中任务重要性的动态变化;3)过度依赖最小化与真实数据的差异,导致手势过拟合而抑制多样性。本文旨在开发一个统一的自适应框架,在无需人工干预的情况下同时满足上述的双重目标。**方法**　提出一个基于扩散模型的新框架,通过基于任务不确定性的多任务学习,自适应地平衡确定性的面部表情生成与非确定性的手势生成。该方法引入可学习的不确定性损失权重,能够在训练期间动态调整损失权重,使手势和表情任务自主挖掘并优化它们之间的关系,达到最优的效果,并且该方法减轻了调整参数的负担。**结果**　在BEAT(body-expression-audio-text)数据集上的实验表明,本文方法在面部表情的FD(Fréchet distance)指标上达到9.18(最优),在手势多样性上达到52.5(最高)。用户研究进一步验证了该方法在手势多样性、面部同步性和整体运动质量等方面的优越性。**结论**　本文提出的自适应扩散框架通过自适应任务平衡机制,成功解决了整体运动生成中面部同步性与手势多样性之间的权衡问题,实现了两个基本标准的同时满足,为语音驱动的虚拟形象动画提供了一种有效的解决方案。代码地址:https://doi.org/10.57760/sciencedb.j00240.00175。

关键词: 协同语音运动生成;语音驱动手势生成;多任务学习(MTL);扩散模型;人工智能生成内容

Adaptive diffusion model for co-speech holistic motion generation

Xuan Enyun[1,2], Li You[1,3*], Li Ziwei[1,2], Yao Mengmeng[1], Guo Renzhong[1,3]

1. Guangdong Laboratory of Artificial Intelligence and Digital Economy(Shenzhen), Shenzhen　518083, China; 2. College of Computer Science and Software Engineering, Shenzhen University, Shenzhen　518060, China; 3. School of Architecture and Urban Planning / Research Institute for Smart Cities, Shenzhen University, Shenzhen　518060, China

Abstract: **Objective** Co-speech holistic motion generation aims to simultaneously produce expressive gestures and accurately synchronized facial expressions. However, the two tasks have fundamentally different characteristics. Gesture generation is non-deterministic, representing a one-to-many mapping in which the same speech can correspond to various natural motions, thus requiring high diversity. Conversely, facial expression generation, especially lip movements, is deterministic and represents a one-to-one mapping that requires precise correspondence with phonemes, thus demanding high accuracy. Existing methods suffer from three critical limitations. First, numerous methods employ fixed architectural

收稿日期:2025-11-04;**修回日期**:2026-04-09;**预印本日期**:2026-04-16

*通信作者:李游　liyou@gml.ac.cn

基金项目:人工智能与数字经济广东省实验室(深圳)项目(GML-26420009)

Supported by:Guangdong Laboratory of Artificial Intelligence and Digital Economy (Shenzhen)(GML-26420009)

designs, such as unidirectional conditional flows, imposing rigid task relationships and limiting the model's capability to capture the true dynamic connections between gestures and expressions. Second, the use of manually designed static loss weights prevents the framework from adapting to dynamic changes in task importance during training. Third, over-relying on minimizing differences with ground-truth data often leads to gesture overfitting, thereby suppresses diversity. These limitations force existing methods into making unavoidable tradeoffs between facial synchronization and gesture diversity. Aiming to address this issue, this research introduces a unified adaptive framework that autonomously models and dynamically balances the relationship between the two tasks through learnable uncertainty mechanisms. The proposed framework aims to simultaneously achieve gesture diversity and expression accuracy without requiring manual intervention. **Method** A novel diffusion-based framework is introduced, leveraging uncertainty-based multitask learning for adaptive task balancing in holistic motion generation. This approach represents the first application of uncertainty-based loss weighting to speech-driven holistic motion synthesis. The core idea is to treat gesture and facial expression generation as distinct tasks within a unified framework, allowing their relationship to naturally emerge during training. The framework employs a denoising diffusion probabilistic model that operates on concatenated gesture and facial expression representations. The architecture incorporates shared features, including WavLM audio representations, word embeddings, speaker identity, and timestep encoding, alongside task-specific features such as Gaussian noise vectors and seed motion sequences, to capture commonalities and distinct requirements of each task. Cross-local attention mechanisms are adopted to capture long-range dependencies across timesteps and modalities, while self-attention layers refine task-specific patterns. A key innovation of this framework lies in the introduction of learnable parameters that represent task-dependent homoscedastic uncertainty for gestures and expressions. The overall training objective integrates the losses of both tasks and dynamically weights them according to these uncertainty parameters. This formulation automatically balances task contributions: larger uncertainty values reduce penalties to encourage diversity, whereas smaller values increase penalties to enforce precision. The uncertainty parameters are jointly optimized with the model parameters, allowing the framework to automatically determine the optimal task weighting without manual intervention. **Result** Comprehensive evaluations performed on the 76-h BEAT dataset, featuring 30 speakers and a 98/16/16 data split, demonstrate notable improvements. The proposed method achieved the highest gesture diversity score (52.5), outperforming MambaTalk (51.6), diffusion-based approach for speech-driven holistic 3D expression and gesture generation (DiffSHEG) (47.4), DiffuseStyleGesture (DSG) (48.5), and cascaded motion network (CaMN) (43.2). Additionally, the method achieved the best semantic relevance score (SRGR: 0.324). For facial expressions, the proposed framework achieved the lowest Fréchet distance, outperforming MambaTalk, DiffSHEG, and SAiD. Ablation studies confirm the critical role of uncertainty-based weighting, as removing it decreased gesture diversity from 52.5 to 47.2 and increased facial FD from 9.18 to 10.5. The learned uncertainty parameters converged to weights of 0.506 and 0.494 for gestures and expressions, respectively, demonstrating autonomous task balancing. When applied to DiffSHEG and MambaTalk, the proposed mechanism further improved gesture diversity, validating its strong generalizability. Qualitative analysis revealed that the generated gestures exhibit substantially greater diversity than those produced by baselines, closely imitating ground truth. Additionally, user studies involving 17 participants evaluating nine groups of videos exhibited a clear preference for the proposed method in terms of gesture diversity, facial synchronization, and overall quality dimensions. **Conclusion** This research presents a novel adaptive diffusion framework that successfully addresses the fundamental challenge of simultaneously achieving precise facial synchronization and diverse gesture generation. By introducing uncertainty-based learnable parameters within a multitask learning paradigm, the proposed method automatically optimizes task relationships without requiring manual parameter tuning, thereby achieving superior performance in deterministic expression synthesis and non-deterministic gesture generation. Experimental results demonstrate substantial improvements in facial accuracy (FD: 9.18), gesture diversity (52.5), and semantic relevance (SRGR: 0.324), with user studies confirming enhanced realism. These findings indicate that the proposed framework offers an effective solution for creating lifelike virtual agents and introduces new research directions for holistic motion generation through adaptive multitask learning. The codebase of the paper: https://doi.org/10.57760/sciencedb.j00240.00175.

Key words: co-speech motion generation; speech-driven gesture generation; multi-task learning (MTL); diffusion model; artificial intelligence generated content

0 引 言

随着机器人技术和深度学习的飞速发展，虚拟人和社交机器人等数字化身在游戏、动画、电影制作以及人机交互系统中发挥着越来越重要的作用。这些虚拟形象的表现力在很大程度上依赖于生成与口语内容相对应的同步面部表情和自然手势的能力（Badler，1997；Ekman 和 Friesen，1978；McNeill 和 Arnheim，1994）。高质量的虚拟形象不仅需要嘴巴随特定词语准确张开，还需要手势与对话内容的转折相辅相成。因此，有效的整体运动生成需要满足两个基本标准：实现面部表情的高度同步性（确定性任务）以及确保手势的多样性和语义相关性（非确定性任务）（Yi 等，2023）。下面将详细介绍相关工作研究进展。

首先是音频驱动表情生成，在面部动画合成领域，早期的 Voice Puppetry 系统（Brand，1999）首次实现了从音频直接生成面部动画。随后，Massaro 等人（2012）以及 Taylor 等人（2012）的研究提供了更灵活的控制方法，但仍需要人工干预。深度学习算法的出现极大地推动了该领域的发展。Eskimez 等人（2018）、Chen 等人（2019）和 Greenwood 等人（2018）的工作专注于合成面部标志点。Transformer（Vaswani 等，2017）架构的引入带来了突破性进展，Faceformer（Fan 等，2022）、MakeItTalk（Zhou 等，2020）、SAiD（speech-driven blendshape facial animation with diffusion）（Park 和 Cho，2023）和 Imitator（Thambiraja 等，2023）等方法显著提高了表情合成的准确性，能够预测捕捉核心变形的系数，实现参数化表示（Edwards 等，2016；Jeni 等，2015；Cudeiro 等，2019；Xing 等，2023）。

其次是音频驱动手势生成领域，手势生成因其非确定性特征而成为更具挑战性的任务。正如赵宝全等人（2024）的三维数字人运动生成综述的总结，早期方法主要基于生成对抗网络（generative adversarial network，GAN）（Goodfellow 等，2014）和循环神经网络（recurrent neural network，RNN），如长短期记忆网络（long short-term memory，LSTM）（Graves 等，2013）和门控循环单元（gated recurrent unit，GRU）架构（Chung 等，2014）。Liu 等人（2022）提出的 CaMN（cascaded motion network）模型和 Yoon 等人（2020）的工作采用了这些架构进行手势生成。近年来，基于扩散的模型展现出生成更逼真手势的能力，Yang 等人（2023）提出的 DiffuseStyleGesture、Wang 等人（2026）提出的 MMoFusion 和 MotionStreamer（Xiao 等，2025）模型通过扩散机制显著提升了动作的真实性。还有些工作专注于数字人风格化、多模态驱动等（潘烨 等，2025）。然而，这些方法分别对面部和手势任务建模，无法捕捉人类交流的全部范围。

上述两种方法只能从音频生成单一模态的运动参数，为弥合单一模态生成的不足，近年来出现了一些联合建模手势和面部表情的整体运动生成方法。Habibie 等人（2021）的早期工作使用卷积神经网络（convolutional neural network，CNN）和 GAN 进行全身合成，将其视为多任务问题并为身体不同部分设置独立解码器。TalkSHOW（Yi 等，2023）采用 Transformer 和码本训练模型，使用 Wav2vec（Baevski 等，2020）提取音频特征合成面部表情，并利用 VQ-VAE（vector-quantized variational autoencoders）（Kingma 和 Welling，2014）生成身体动作。EMAGE（expressive masked audio gesture）（Liu 等，2024）分别使用 MG2G（masked gesture-to-gesture）和 A2G（audio-to-gesture）生成身体动作和面部表情，通过在训练中引入掩码手势先验和交叉注意力机制（Huang 等，2019）提升性能。DiffSHEG（diffusion-based approach for speech-driven holistic 3D expression and gesture generation）（Chen 等，2024）基于扩散模型（Ho 等，2020）实现全身动作生成，设计了从表情到手势的单向条件流以学习联合分布。MambaTalk（Xu 等，2024）和其他最新方法（Sun 等，2025）也尝试对手势和面部表情进行联合建模。然而，这些方法在满足前述标准时仍面临两个关键局限：1）手势与面部表情之间微妙且依赖上下文的关系难以直接建模。强制施加固定关系（如 DiffSHEG 的单向数据流）可能阻碍模型捕捉真实动态关系，从而影响性能。2）人工设计的损失权重无法适应训练过程中任务重要性的动态变化。这些缺陷迫使现有方法在面部同步性和手势多样性之间做出妥协：优先考虑面部准确性时，过度约束的建模会抑制手势多样性；允许手势自由度时，任务平衡不足又会导致面部同步性下降。

针对上述问题，本文方法参考了多任务学习的思想，得到了合适的解决方案。近年来，多任务学习（multi-task-learning，MTL）在计算机视觉领域取得了

显著成功，能够通过单一模型同时学习多个任务，如图像分类与语义分割（Ilyas等，2022）、目标检测与分割（He等，2022）。MTL的发展主要集中在两个方向：1）研究神经网络框架和优化过程，Ma等人（2019）和Tang等人（2020）探讨了底层特征如何传递到多个上层任务；2）研究如何平衡多任务训练的损失权重，Kendall等人（2018）和Gong等人（2019）提出基于任务不确定性动态估计损失权重的方法，其中不确定性参数与网络参数共同学习。Yu等人（2020）通过梯度投影技术缓解了多任务学习中的梯度冲突问题。然而，虽然Habibie等人（2021）将全身动作合成视为多任务学习问题，但未考虑任务特定权重的优化，这为本研究提供了切入点。

针对上述问题，本文方法借鉴多任务学习领域的思想，提出了一个基于自适应多任务学习的扩散框架，这是首个将基于不确定性的多任务学习应用于整体运动生成的方法。如图1所示，本文方法使手势和表情任务能够在训练期间自主发现并优化它们的关系，而非依赖预定义的建模假设。通过引入可学习的不确定性参数来动态调整任务权重，模型能够在无需手动干预的情况下自适应地实现高面部同步性，同时保持手势多样性。本文的主要贡献包括：1）提出一个新颖的基于扩散的自适应框架，通过自适应任务平衡统一手势和面部表情的生成，在单一模型中平衡了面部表情精度和手势多样性的不同要求；2）引入基于不确定性的可学习参数用于自适应多任务平衡，这是首次将基于不确定性的损失加权应用于整体运动生成，优化了面部表情与音频的同步性，同时促进了多样化和富有表现力的手势生成；3）通过全面的定量和定性实验以及用户研究，验证了本文方法的有效性，在面部表情精度、手势多样性和整体运动真实感方面取得了显著改进。

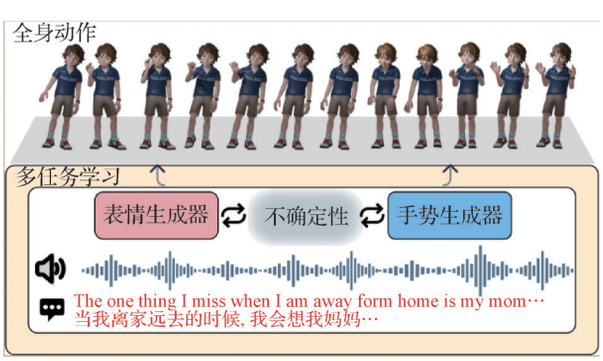


图1　音频驱动全身运动框架图

Fig. 1　A simple framework of co-speech holistic motion generation

1　本文方法

本节阐述提出的自适应扩散框架，首先介绍模型的整体架构（1.1节），随后详细说明特征编码（1.3节）、解码（1.4节）过程，并重点解析作为本文核心的自适应不确定性损失函数（1.5节）。本文提出的框架整体流程如图2所示。本文的整体思路是使用扩散模型从完全的噪声预测出一个联合向量，然后将这个向量按照维度拆分为手势和面部表情。该过程主要包含去噪（denoising）和扩散（diffusion）两个阶段。

1.1　整体架构

如图2所示，在去噪阶段，本文方法结合了共享特征和任务特定特征来生成统一的运动。共享特征包括时间步 t、说话人ID（identity）、音频输入和相应的文本转录等基本信息，这些特征为手势和面部表情生成任务奠定了共同基础。这些共享特征建立了一个共同的基础，确保生成运动中手势和面部表情之间的连贯性和同步性。

为了满足每个任务的独特要求，本文方法将任务特定特征——即种子手势（或表情）和带噪声的手势（或表情）——与共享特征集成。这个组合输入通过交叉局部注意力层进行处理，在捕获交互和上下文依赖关系的同时解码两个任务。随后，输出经过自注意力层以增强特征细化，然后每个任务计算各自的损失。

遵循Kendall等人（2018）的方法，本文方法引入了两个可学习参数 σ_1 和 σ_2 动态调整手势和面部表情任务的损失权重。这种基于不确定性的加权机制使模型能够自适应地平衡两个任务，使每个任务都能达到最优性能，而无需手动调整损失权重。通过让模型自主调整其关注点，本文方法的框架确保了与音频的高质量面部同步和丰富多样的手势表达，实现了整体和同步的运动生成。

1.2　通过扩散模型生成运动

扩散模型通过两个阶段运行：扩散（前向）和去噪（反向）过程（Ho等，2020）。在扩散阶段，噪声逐渐添加到原始数据 $\boldsymbol{x}_0$ 中，而在去噪阶段，模型反转此过程以恢复 $\boldsymbol{x}_0$。在本文的框架中，手势和面部表

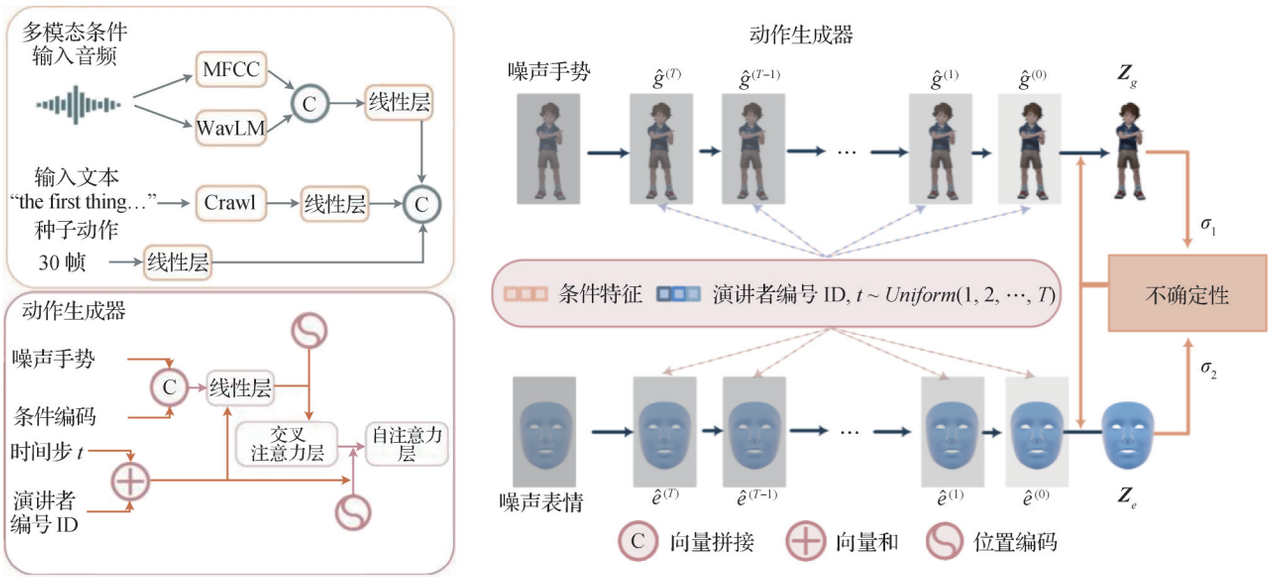


图 2　使用多任务学习联合音频驱动手势生成任务与音频驱动表情生成任务框架图

Fig. 2　A framework diagram of using multi-task learning to combine co-speech gesture generation task and audio-driven expression generation task

情被连接成一个称为运动的统一表示，作为模型的 $\boldsymbol{x}_0$ 输入。扩散模型以音频、说话人 ID 和文本转录等外部特征为条件。

1.2.1　扩散过程

前向过程生成一系列带噪变量 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T$，其中 $\boldsymbol{x}_T$ 是纯噪声。在每个时间步 t，根据高斯分布添加噪声，具体为

$$q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}\right) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{1-\beta_t}\,\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}\right) \tag{1}$$

式中，β_t 是噪声调度，$\mathcal{N}()$ 为正态分布，$\boldsymbol{I}$ 为单位矩阵。

或者，也可以直接采样，即

$$q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_0\right) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\overline{\alpha}_t}\,\boldsymbol{x}_0, \left(1-\overline{\alpha}_t\right)\boldsymbol{I}\right) \tag{2}$$

式中，$\overline{\alpha}_t = \prod_{i=1}^{t}\left(1-\beta_i\right)$ 表示 $\left(1-\beta_i\right)$ 的累积乘积。

1.2.2　去噪过程

反向过程旨在逐步去除每一步的噪声，从 $\boldsymbol{x}_t$ 预测 $\boldsymbol{x}_{t-1}$，具体为

$$p_\theta\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta\left(\boldsymbol{x}_t, t\right), \boldsymbol{\Sigma}_\theta\left(\boldsymbol{x}_t, t\right)\right) \tag{3}$$

模型被训练来预测前向过程中添加的噪声，从 $\boldsymbol{x}_T$ 开始，迭代去噪直到生成 $\boldsymbol{x}_0$，该 $\boldsymbol{x}_0$ 近似原始运动数据，$\boldsymbol{\mu}_\theta$ 为分布均值，$\boldsymbol{\Sigma}_\theta$ 为均方差矩阵。

本文方法的目标是在给定某些条件 c 的情况下生成长度为 N 的人体运动 $\boldsymbol{x}_{1:N}$。类似于 Tevet 等人(2023)和 Yang 等人(2023)，本文方法直接预测运动而不是像 Ho 等人(2020)那样预测 $\boldsymbol{\Sigma}_\theta\left(x_t, t\right)$。去噪模块从输入噪声 $\boldsymbol{x}_t$、时间步 t 和条件 c 重建原始运动 x_0，具体为

$$\hat{\boldsymbol{x}}_0 = Denoise\left(\boldsymbol{x}_t, t, c\right) \tag{4}$$

式中，*Denoise* 为去噪函数。

1.3　特征编码

本文模型使用共享和任务特定的特征来统一生成手势和面部表情，从而在满足各个任务需求的同时实现连贯的运动合成。

共享特征包括：1)音频和文本特征。WavLM (waveform language model)(Chen 等，2022)提取 1 133 维的音频特征 $\boldsymbol{Z}_a$。来自 Crawl-300d-2M 嵌入(Mikolov 等，2018)的文本特征 $\boldsymbol{Z}_w$(300 维)与音频特征拼接后，投影到 96 维的 $\boldsymbol{Z}_f$，编码了语义和声学信息。2)说话人 ID。独热编码后投影到 384 维的嵌入 $\boldsymbol{Z}_{\mathrm{id}}$，捕捉说话人特定的特征。3)时间步 t。通过多层感知机(multilayer perceptron，MLP)编码为 384 维向量 $\boldsymbol{Z}_t$，与 $\boldsymbol{Z}_{\mathrm{id}}$ 结合形成 $\boldsymbol{Z}_s$，用于在整个扩散过程中进行时序建模。

任务特定特征包括：1)噪声向量。高斯噪声转换为 384 维向量 $\boldsymbol{Z}_{\mathrm{g_noise}}$(手势)或 $\boldsymbol{Z}_{\mathrm{e_noise}}$(表情)，为多样的运动探索提供随机性。2)种子序列。真实(ground truth，GT)序列的前 30 帧通过线性层处理成 96 维向量 $\boldsymbol{Z}_g$(手势)和 $\boldsymbol{Z}_e$(表情)，以保持时间连贯性并引导自然的运动轨迹。

1.4 特征解码

特征通过注意力层进行处理，以理解特征间的关系并生成高质量的运动。1）交叉局部注意力。拼接后的特征$\left[\boldsymbol{Z}_s, \boldsymbol{Z}_f, \boldsymbol{Z}_g, \boldsymbol{Z}_{\text{noise}}\right]$与位置编码（Kitaev等，2020）结合，捕捉跨时间步的长程依赖关系，使模型能够理解上下文关系。2）自注意力。通过允许特征关注每种模态内部的相关信息（Vaswani等，2017）提炼任务特定的模式，通过选择性地关注重要特征分量，增强手势和表情生成的精度。任务特定的解码器独立处理提炼后的特征，确保每个运动生成任务都能达到最佳性能。

1.5 不确定性调整损失函数

现有的整体运动生成方法存在两类主要局限：1）采用固定的架构设计强制施加任务间关系（如DiffSHEG的单向条件流将面部表情特征注入手势生成，限制了手势的自由探索空间）；2）使用手动调整的静态损失权重无法适应训练过程中任务重要性的动态变化。这些方法难以适应手势（非确定性）和面部表情（确定性）这两个任务的本质差异。

为解决这一问题，本文方法基于一个关键洞察设计了自适应损失加权机制：手势生成是非确定性的（一对多映射），而面部表情生成是确定性的（一对一映射）。与通过架构设计强加任务关系不同，本文方法让模型通过可学习的不确定性参数自主发现最优的任务平衡策略。

每个任务使用Huber损失（Huber，1992）计算预测误差，具体为

$$L_{\text{task}} = E_{x_0 \sim q(x_0|c), t \sim [1,T]}\left[L_{\text{Huber}}\left(x_0 - \widehat{x_0}\right)\right] \tag{5}$$

遵循Kendall等人（2018）的方法，本文方法引入可学习的参数σ_1和σ_2表示任务不确定性，构建动态加权的总损失。具体为

$$L_{\text{total}} = \frac{1}{2\sigma_1^2} L_{\text{gesture}} + \frac{1}{2\sigma_2^2} L_{\text{expression}} + \log\left(\sigma_1\right) + \log\left(\sigma_2\right) \tag{6}$$

式中，σ_1和σ_2分别表示手势和表情任务的同方差不确定性。对数项起正则化作用，防止不确定性参数趋向无穷大。这种设计的核心优势在于：较大的σ值会减小对应任务的损失权重$1/(2\sigma^2)$，允许更大的预测方差以鼓励多样性；较小的σ值则增大损失权重，强制模型更严格地优化该任务。

不确定性参数与模型参数联合优化，使模型能够在训练过程中自主发现最优的任务平衡策略。手势和表情任务通过共享的音频特征保持连贯性，但各自的解码器独立优化，不存在强制的单向数据流。σ_1和σ_2作为“软约束”，让模型自主平衡两个任务，而非通过架构设计强加固定关系。实验表明，模型学习到的权重配置（σ_1对应0.506，σ_2对应0.494）成功适应了两个任务的本质差异，在无需人工调参的情况下同时满足了面部同步性和手势多样性的要求。

2 实验结果和讨论

2.1 实验设置

本文在BEAT（body-expression-audio-text）（Liu等，2022）数据集上进行实验，该数据集是一个全面的多模态数据集，专为协同语音手势生成而设计。它包含来自30位说话人的76 h高质量动作捕捉数据，涵盖4种语言，具有愤怒、快乐和悲伤等各种情绪表达。BEAT整合了身体动作、面部表情、音频和文本，非常适合生成由语音驱动的富有表现力的同步手势和面部动作。

本文遵循CaMN模型代码的训练集/验证集/测试集划分规则，从2、4、6和8位说话人中选择片段，每位说话人选择98/16/16个片段分别作为训练/验证/测试集。每个片段大约持续1 min，训练以30 帧/s的帧率进行。视频数据被划分为150帧的片段进行训练。本文分别评估了面部表情和手势生成，将本文结果与SAiD（speech-driven blendshape facial animation with diffusion）（Park和Cho，2023）、CaMN（cascaded motion network）（Liu等，2022）、DSG（DiffuseStyleGesture）（Yang等，2023）、DiffSHEG（Chen等，2024）和MambaTalk（Xu等，2024）进行比较。此外，本文设计了用户研究来评估模型生成内容的整体有效性。本文方法使用单个RTX 4090 GPU进行训练，迭代180 000次，批大小为200。对于手势生成，本文方法使用轴角旋转表示而不是欧拉角进行重新训练。模型的输入是音频及其对应的文本。

2.2 评估指标

2.2.1 手势生成

1）FGD（Fréchet gesture distance）（Yoon等，2020）。使用从预训练模型中提取的特征衡量真实手势和生成手势之间的分布相似性。本文在训练集上重新训

练了VAE以适应使用轴角旋转表示的维度。

2)多样性(diversity)(Li等,2021a)。通过序列之间的平均成对距离量化生成手势的变化。

3)节拍对齐(beat align)(Li等,2021b)。使用Chamfer距离评估手势与节拍的同步性。

4)语义相关性手势复现率(semantic relevance gesture recall,SRGR)(Liu等,2022)。评估手势的语义相关性和多样性。

2.2.2 面部动画生成

本文方法采用SAiD(Park和Cho,2023)指标评估表情准确性和音频同步性。

FD(Fréchet distance)(Dowson和Landau,1982)比较预测和真实blendshapes值在潜在空间中的分布。

WInD(Wasserstein inception distance)(Dimitrakopoulos等,2020)衡量真实和生成表情分布之间的Wasserstein距离。

2.3 定量分析

本文分别评估了手势合成和表情生成。对于手势生成,本文方法与CaMN(Liu等,2022)、DSG(Yang等,2023)、DiffSHEG(Chen等,2024)和MambaTalk(Xu等,2024)进行比较;对于面部表情生成,本文方法与DiffSHEG、SAiD(Park和Cho,2023)和MambaTalk进行比较。

对于手势动作生成,本文实验不仅需要评估其生成的准确性,还要评估其多样性,这是因为一段音频所适配的动作并非单一的,对于同一段音频,人们会做出不同的动作。如表1所示,MambaTalk的FGD值最低,这说明它与真值的关联非常强,这会要求数据的质量非常高。如果数据中出现抖动,其预测也会产生抖动,这是实际场景中不希望的。然而本文模型达到了最高的多样性(52.5)和SRGR(0.324),说明本文模型生成更多样的动作并且这些动作是符合语义的。FGD值的升高也符合音频与手势一对多的关系,说明本文模型具有更好的泛化性,不会对真值的质量有太高的依赖。

对于面部表情生成,可以认为它是一个确定性的任务,因为一段音频的发音规则对应的口型基本上是确定的,因此需要评估表情生成的准确性与同步性。如表2所示,本文模型达到最低的FD值(9.18),这说明本文方法生成的面部表情的准确度最高,最接近真值。

表1 在BEATv0上的手势生成定量评估

Table 1 Quantitative evaluation of gesture generation on BEATv0 dataset

方法	多样性↑	SRGR↑	FGD↓	节拍对齐↑
CaMN	43.2	0.318	4.45	0.850
DSG	48.5	0.314	4.32	0.932
DiffSHEG	47.4	0.319	3.75	0.953
MambaTalk	51.6	0.320	**3.43**	**0.961**
本文	**52.5**	**0.324**	5.32	0.920

注:加粗字体表示各列最优结果。“↑”表示值越高越好;“↓”表示值越低越好。

表2 表情生成定量评估FD和WInD

Table 2 Quantitative evaluation of expression generation FD and WInD

方法	FD/$\times10^{-5}$↓	WInD/$\times10^{-4}$↓
SAiD	69.3	8.47±0.119
DiffSHEG	12.4	**2.19**±0.142
MambaTalk	9.32	2.56±0.132
本文	**9.18**	2.67±0.327

注:加粗字体表示各列最优结果。“↑”表示值越高越好;“↓”表示值越低越好。

这些结果验证了基于不确定性的学习显著优于固定权重方法,能够实现最优的任务特定目标:手势的高多样性和面部表情的精确同步。

2.4 定性分析

除了上述的定量指标对比,本文实验还进行了定性分析以直观地比较生成效果。图3展示了本文模型与GT、CaMN、DSG、DiffSHEG和MambaTalk的对比。本文模型展现出最丰富的手势动作,其余的模型都会受到GT的影响而刻意去模仿GT的动作,这在一对多任务中是不理想的。其中,CaMN模型与音频的节奏同步性不高,手势表现为迟钝、不灵动;DSG模型有时会产生动作抖动和人物位置移动的现象;DiffSHEG模型与GT的关联性太高,在有些GT动作较少或者有抖动时,其生成的动作也会较少活动或出现抖动。本文模型在符合音频语义、音频节奏的基础上,不会太模仿GT的动作,这更接近一对多任务的理想目标。

图4展示了本文模型与GT、SAiD、DiffSHEG和MambaTalk的对比。在面部表情上,SAiD的运动幅

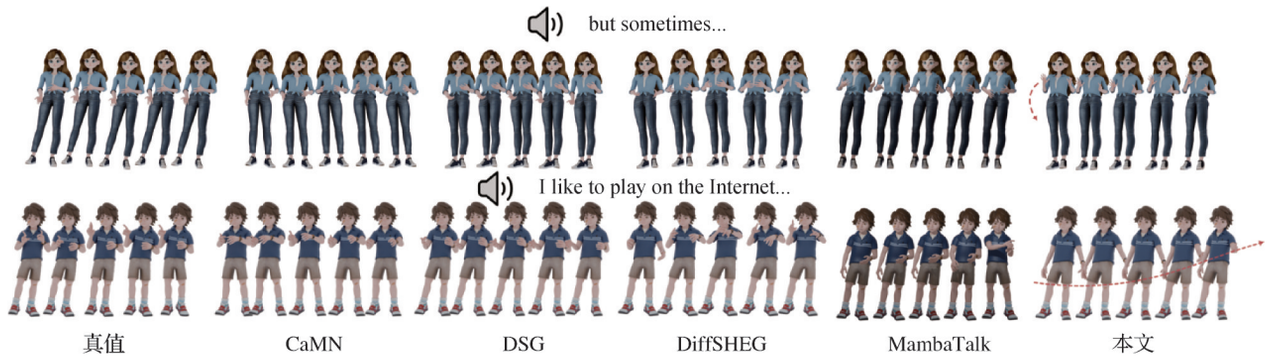


图3　本文方法的手势生成结果与其他工作的对比结果

Fig. 3　Comparison of gesture generation results of our method with other works

度较小,导致效果不好;DiffSHEG和MambaTalk的运动频率会出现滞后于音频的情况。而本文模型体现出了最高的音频同步率,同时在说到一些语速较快的单词时,能较好地跟上GT嘴唇的运动节奏,本文模型的嘴唇动作也是最明显的。另外,本文实验绘制了JawOpen Blendshape随时间变化的折线图,如图5所示,可以更清晰地观测到本文模型曲线与GT的曲线同步率最高,与GT的均方误差(mean square error,MSE)也是最低的,其他曲线都会有明显的时间滞后或者错误,说明本文模型在口型同步与准确度上效果最优。

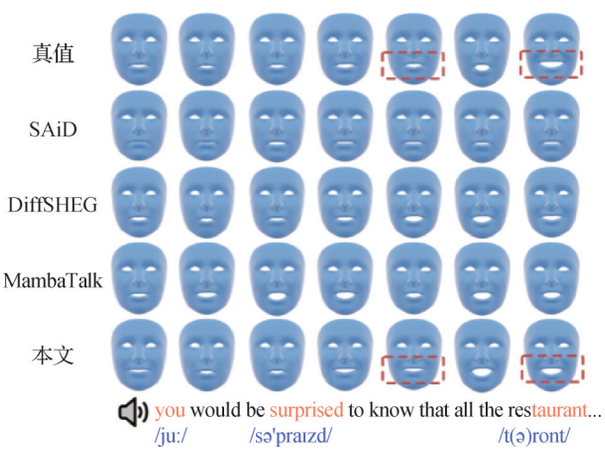


图4　本文方法与其他工作的表情生成结果对比

Fig. 4　Comparison of expression generation results of our method with other works

2.5　用户研究

本文实验使用Streamlit制作了一个网页用于用户测试。该用户实验在测试集中选取视频,制作了9组视频进行测试,每个视频大约30 s左右。

在实验中招募了17位参加者,他们具有较好的英语水平、不同的教育背景和不同的研究领域。实验中对每组视频设计了3个问题,以评估生成动作的多样性、生成表情的同步性以及综合评估全身动作的整体生成效果:

1)您认为哪个模型的动作更加多样化和富有变化(考虑动作幅度和重复性等因素)?

2)您认为哪个模型的面部表情更好(嘴唇动作是否与音频发音对齐良好?)?

3)综合考虑身体动作和面部表情,您认为哪个

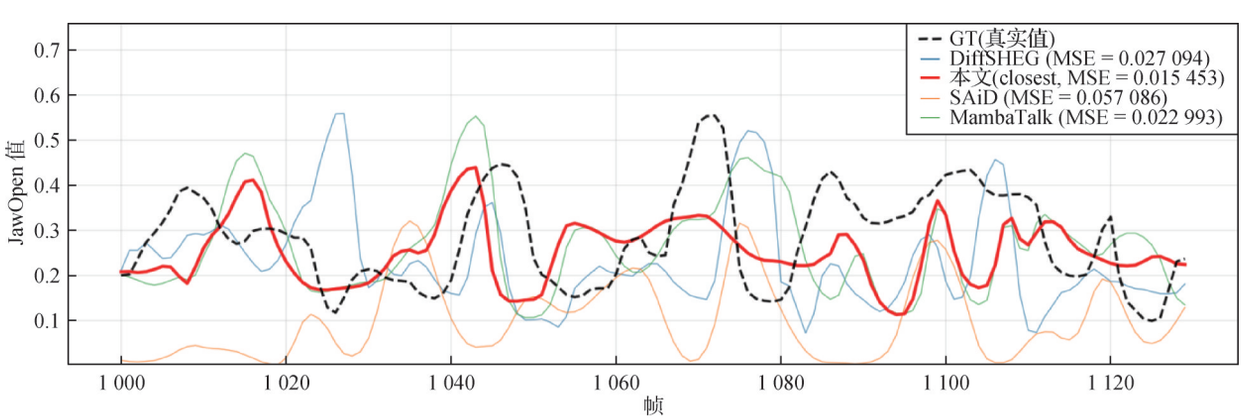


图5　JawOpen Blendshape数值对比图

Fig. 5　Comparison of JawOpen Blendshape

模型的整体表现更好?

为了量化评估生成结果的主观质量，本文对17位受试者在9组测试视频上的投票记录进行了深度的统计与显著性分析，具体结果如图6所示。与常规的单纯频数统计不同，为验证受试者对不同模型的偏好差异是否具有统计学意义，本文引入了卡方检验(Chi-square test)。在假设受试者对3种模型的主观选择概率相等的前提下，统计结果呈现出极高的显著性。具体而言，在面部表情同步性方面，本文模型获得了压倒性的主观偏好($\chi^2 = 144.04, p < 0.001$)。这表明自适应不确定性机制在处理确定性的口型生成任务时，成功引导模型施加了严格的对齐约束，显著优于基线模型。在手势动作多样性($\chi^2 = 77.80, p < 0.001$)与全身整体合成效果($\chi^2 = 98.16, p < 0.001$)方面，本文模型同样以极高的统计学显著性胜出了DiffSHEG与CaMN等基线模型。结合受试者的定性反馈，本文模型之所以在非确定性的手势生成中脱颖而出，是因为其生成了更广泛、更具表现力的手势范围，有效捕捉了真实人类运动的内在变化性。传统的基线模型往往倾向于过度拟合真值(GT)动作，导致在面对一对多的映射时表现出动作的僵硬或幅度受限；而本文提出的自适应框架在保持语义连贯性的同时，赋予了手势更大的探索自由度。这种兼顾确定性约束与非确定性自由度的能力，证明了本文模型能够更准确、更拟真地反映人类对音频提示的多元化自然反应。

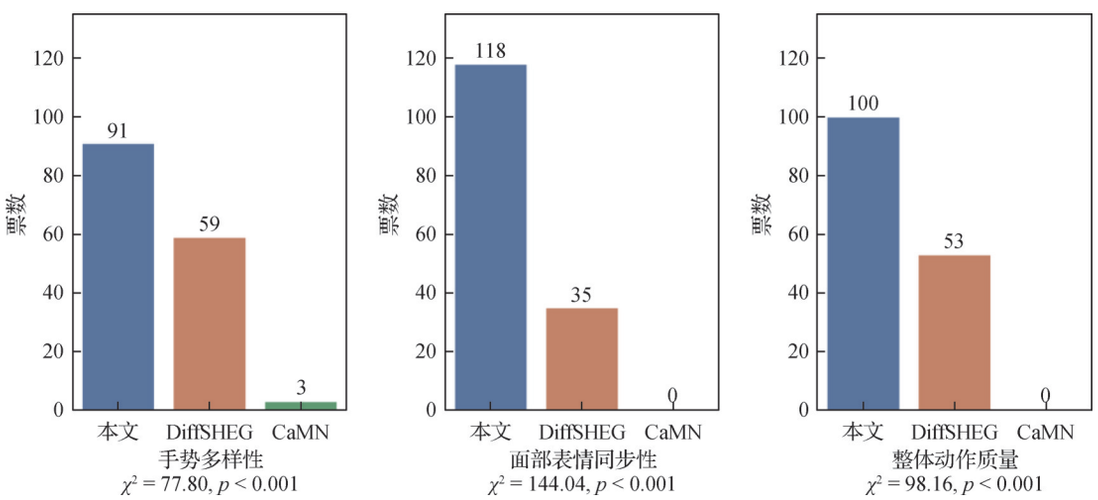


图6　基于卡方显著性检验的用户研究结果

Fig. 6　User study results based on Chi-square significance test

2.6　模型消融

消融实验对两个关键组件进行消融研究。基于不确定性的加权参数(uncertainty-based weighting parameter，UWP)和多任务学习框架。表3和表4显示了详细的消融实验结果。

在分析这些结果之前，首先介绍实验中涉及的模型配置。表3中的模型配置包括：完整模型是本文提出的包含UWP机制的完整框架；消除UWP指移除不确定性加权参数，采用固定的1∶1任务权重；单独训A2E表示仅训练音频到面部表情生成任务(audio-to-expression)，不包含手势生成；单独训A2G表示仅训练音频到手势生成任务(audio-to-gesture)，不包含面部表情生成；DiffSHEG加UWP是在DiffSHEG基础上应用本文提出的UWP机制；DiffSHEG是原始的DiffSHEG基线方法；MambaTalk的处理方式与DiffSHEG相同。表4中的模型配置包括：单独训A2E对应仅训练音频到表情任务；单独训A2G对应仅训练音频到手势任务(同上)；消除UWP是采用固定0.5∶0.5权重的多任务模型；完整模型包含可学习的不确定性参数。

2.6.1　UWP的有效性

为了验证该可学习参数的有效性，进行了两组实验，并对手势合成和表情合成两个任务分别进行评估。1)两个任务的权重参数从1∶1开始学习；2)两个任务的权重参数固定为1∶1。实验结果如表3和表4所示。带有UWP的模型在面部表情生成上达到了更低的FD值(9.18)，说明准确度提升了；同时提高了手势合成的多样性(52.5)和SRGR

表3　在BEATv0上的消融实验

Table 3　Ablation experiment on BEATv0 dataset

方法	FD/×10^{-5}↓	WInD/×10^{-4}↓	多样性↑	SRGR↑	FGD↓	节拍对齐↑
完整模型	<u>9.18</u>	2.67±0.327	<u>52.5</u>	<u>0.324</u>	5.32	0.920
消除UWP	10.5	2.49±0.211	47.2	0.314	3.98	0.923
单独训A2E	10.0	2.35±0.207	–	–	–	–
单独训A2G	–	–	48.4	0.316	4.31	0.930
DiffSHEG加UWP	10.6	**2.17**±0.182	51.5	0.322	5.62	0.950
DiffSHEG	12.4	<u>2.19</u>±0.142	47.4	0.319	<u>3.75</u>	0.953
MambaTalk加UWP	**8.87**	2.43±0.237	**54.3**	**0.334**	5.02	<u>0.958</u>
MambaTalk	9.32	2.56±0.132	51.6	0.320	**3.43**	**0.961**

注：加粗、下划线字体表示各列最优、次优结果。“–”表示无相关实验结果。“↑”表示值越高越好；“↓”表示值越低越好。

表4　多任务学习的定量分析

Table 4　Quantitative experiments on multi-task learning

方法	任务权重(手势/表情)	开始时loss/×10^{-3}	结束时loss/10^{-3}	损失差值/10^{-3}
单独训A2E	0 / 1	0 / 251	0 / 6.62	0 / 244
单独训A2G	1 / 0	254 / 0	6.28 / 0	**248** / 0
消除UWP	0.5 / 0.5	251 / 260	7.82 / 8.16	244 / 252
完整模型	0.506 3 / 0.493 7	246 / 292	7.67 / 8.22	239 / **284**

注：加粗字体表示最后一列最优结果。

(0.324)，FGD升高到5.32，说明提升了动作合成的多样性，增强了模型的泛化性，减少了对真值的依赖程度。

实验中也对DiffSHEG、MambaTalk应用了UWP。实验结果表明，两者都可以提升手势多样性，验证了该方法在不同架构上的普适性。值得注意的是，即使DiffSHEG采用单向条件流(面部表情→手势)的刚性架构，应用UWP后仍能显著改善性能，说明自适应损失加权机制能够部分缓解刚性建模带来的问题。此外，应用UWP后面部表情的损失下降更快(如表4所示)，这也说明其能增强表情的学习能力。

2.6.2　多任务学习框架的有效性

为了验证多任务学习框架的有效性，本文分别对手势合成和表情合成任务进行单独训练，实验结果如表3和表4所示。当两个任务联合训练时，若没有UWP，则可以降低FGD的值(3.98)，同时面部表情生成也可以达到相当的水平(FD：10.5)；若有UWP，则可以达到更理想的效果，提高了多样性(52.5)，使生成的动作更为多样，FD降至9.18，这是实际应用场景中希望看到的结果。虽然手势和表情的关联性没有这么强，但联合训练可以让模型捕捉到两者之间的关系，从而提升整体性能。

2.6.3　多任务学习的训练动态分析

表4详细展示了不同模型配置在训练过程中的损失变化情况，为理解UWP机制的工作原理提供了重要洞察。单独训练A2E模型时，表情任务的损失从初始的251降至6.62，下降幅度达244；单独训练A2G模型时，手势任务的损失从254降至6.28，下降幅度达248。这两个单任务基线的表现说明各任务单独训练时均能有效收敛。

对于消除UWP的多任务模型，采用固定的0.5∶0.5权重配置，手势任务损失从251降至7.82(下降244)，表情任务损失从260降至8.16(下降252)。虽然两个任务都能收敛，但由于采用均衡权重，模型难以针对任务特性进行自适应优化。相比之下，本文的完整模型通过可学习的不确定性参数，自主学习到0.506 3∶0.493 7的任务权重配置。尽管手势任务初始损失略低(246)，但最终收敛至7.67(下降239)；表情任务初始损失较高(292)，但最终收敛至8.22(下降284)。关键发现在于，UWP机制使表情

任务的损失下降幅度显著增大(284 vs 252)，说明模型为表情任务分配了更强的优化能力。这与面部表情生成作为确定性任务需要更精确优化的特点相吻合，验证了UWP机制能够根据任务本质特性自适应调整优化策略，在无需人工干预的情况下实现任务间的最优平衡。

3 结 论

针对语音驱动整体运动生成中面部同步性与手势多样性难以兼顾的问题，本文提出了一种基于不确定性的自适应多任务学习框架。该框架通过引入可学习的不确定性参数动态调整任务权重，使模型能够自主平衡确定性的面部表情生成与非确定性的手势生成。

本文的主要贡献包括：1)构建了统一的自适应框架，将整体运动生成抽象为多任务学习问题，避免了现有方法中刚性架构的局限性；2)首次将不确定性建模应用于整体运动生成领域，通过可学习参数实现任务间的动态平衡，无需人工调参；3)在BEAT数据集上验证了方法的有效性，在手势多样性(52.5)、语义相关性(SRGR：0.324)和面部表情精度(FD：9.18)上均达到最优，且该机制应用于DiffSHEG、MambaTalk后手势多样性都得到提升，验证了方法的可迁移性。

实验结果表明，不确定性参数能够自主学习到合理的任务权重配置(手势0.506、面部表情0.494)，有效缓解了多任务学习中的权重调参难题。消融实验证实，带有不确定性加权的多任务学习框架能够显著提升手势多样性和表情同步性。用户研究进一步验证了生成动作的自然性和表现力。

本研究也存在一定局限性：模型对训练数据质量仍有依赖，虽然相比现有方法已有改善，但在数据噪声较大时仍可能影响生成质量；当前框架主要关注手势和面部表情两个模态，未来可扩展至身体姿态、头部运动等更多模态；不确定性参数的收敛速度受初始化和学习率影响，需要研究更稳定的优化策略。

未来研究可从以下方向展开：1)将基于不确定性的动态权重机制推广到更多模态的联合生成；2)提升模型在少样本或跨数据集场景下的泛化能力；3)结合情感识别等高层语义信息，增强生成动作的表现力和个性化特征；4)优化模型推理效率，使其更适用于实时交互场景。

本文提出的自适应多任务学习框架为语音驱动整体运动生成提供了新的解决思路，在虚拟数字人、人机交互等领域具有广阔的应用前景。

参考文献(References)

Badler N. 1997. Virtual humans for animation, ergonomics, and simulation//Proceedings of 1997 IEEE Nonrigid and Articulated Motion Workshop. San Juan, USA: IEEE: 28-36 [DOI: 10.1109/NAMW.1997.609848]

Baevski A, Zhou H, Mohamed A and Auli M. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 12449-12460

Brand M. 1999. Voice puppetry//Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques. Los Angeles, USA: ACM Press/Addison-Wesley Publishing Co.: 21-28 [DOI: 10.1145/311535.311537]

Chen J M, Liu Y F, Wang J N, Zeng A L, Li Y and Chen Q F. 2024. DiffSHEG: a diffusion-based approach for real-time speech-driven holistic 3D expression and gesture generation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 7352-7361 [DOI: 10.1109/cvpr52733.2024.00702]

Chen L L, Maddox R K, Duan Z Y and Xu C L. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 7824-7833 [DOI: 10.1109/CVPR.2019.00802]

Chen S Y, Wang C Y, Chen Z Y, Wu Y, Liu S J, Chen Z, et al. 2022. WavLM: large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing, 16(6): 1505-1518 [DOI: 10.1109/jstsp.2022.3188113]

Chung J, Gulcehre C, Cho K H and Bengio Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling [EB/OL]. [2025-10-15]. https://arxiv.org/pdf/1412.3555.pdf

Cudeiro D, Bolkart T, Laidlaw C, Ranjan A and Black M J. 2019. Capture, learning, and synthesis of 3D speaking styles//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 10093-10103 [DOI: 10.1109/cvpr.2019.01034]

Dimitrakopoulos P, Sfikas G and Nikou C. 2020. WInD: Wasserstein inception distance for evaluating generative adversarial network performance//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona,

Spain: IEEE: 3182-3186 [DOI: 10.1109/ICASSP40776.2020.9053325]

Dowson D C and Landau B V. 1982. The Fréchet distance between multivariate normal distributions. Journal of Multivariate Analysis, 12(3): 450-455 [DOI: 10.1016/0047-259x(82)90077-x]

Edwards P, Landreth C, Fiume E and Singh K. 2016. JALI: an animator-centric viseme model for expressive lip synchronization. ACM Transactions on Graphics, 35(4): #127 [DOI: 10.1145/2897824.2925984]

Ekman P and Friesen W V. 1978. Facial action coding system: A technique for the measurement of facial movement. Palo Alto, CA: Consulting Psychologists Press

Eskimez S E, Maddox R K, Xu C L and Duan Z Y. 2018. Generating talking face landmarks from speech//Proceedings of the 14th International Conference on Latent Variable Analysis and Signal Separation. Guildford, UK: Springer: 372-381 [DOI: 10.1007/978-3-319-93764-9_35]

Fan Y R, Lin Z J, Saito J, Wang W P and Komura T. 2022. FaceFormer: speech-driven 3D facial animation with transformers//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 18749-18758 [DOI: 10.1109/CVPR52688.2022.01821]

Gong T, Lee T, Stephenson C, Renduchintala V, Padhy S, Ndirango A, et al. 2019. A comparison of loss weighting strategies for multi task learning in deep neural networks. IEEE Access, 7: 141627-141632 [DOI: 10.1109/ACCESS.2019.2943604]

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. 2014. Generative adversarial nets//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 2672-2680

Graves A, Jaitly N and Mohamed A R. 2013. Hybrid speech recognition with deep bidirectional LSTM//2013 IEEE Workshop on Automatic Speech Recognition and Understanding. Olomouc, Czech Republic: IEEE: 273-278 [DOI: 10.1109/ASRU.2013.6707742]

Greenwood D, Matthews I A and Laycock S D. 2018. Joint learning of facial expression and head pose from speech//Proceedings of the 19th Annual Conference of the International Speech Communication Association. Hyderabad, India: ISCA: 2484-2488 [DOI: 10.21437/interspeech.2018-2587]

Habibie I, Xu W P, Mehta D, Liu L J, Seidel H P, Pons-Moll G, et al. 2021. Learning speech-driven 3D conversational gestures from video//Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents. [s.l.]: ACM: 101-108 [DOI: 10.1145/3472306.3478335]

He H J, Xu H Z, Zhang Y, Gao K, Li H X, Ma L F, et al. 2022. Mask R-CNN based automated identification and extraction of oil well sites. International Journal of Applied Earth Observation and Geoinformation, 112: #102875 [DOI: 10.1016/j.jag.2022.102875]

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #574

Huang Z L, Wang X G, Huang L C, Huang C, Wei Y C and Liu W Y. 2019. CCNet: criss-cross attention for semantic segmentation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 603-612 [DOI: 10.1109/ICCV.2019.00069]

Huber P J. 1992. Robust estimation of a location parameter//Kotz S and Johnson N L, eds. Breakthroughs in Statistics: Methodology and Distribution. New York, USA: Springer: 492-518 [DOI: 10.1007/978-1-4612-4380-9_35]

Ilyas T, Mannan Z I, Khan A, Azam S, Kim H and De Boer F. 2022. TSFD-Net: tissue specific feature distillation network for nuclei segmentation and classification. Neural Networks, 151: 1-15 [DOI: 10.1016/j.neunet.2022.02.020]

Jeni L A, Cohn J F and Kanade T. 2015. Dense 3D face alignment from 2D videos in real-time//Proceedings of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). Ljubljana, Slovenia: IEEE: 1-8 [DOI: 10.1109/FG.2015.7163142]

Kendall A, Gal Y and Cipolla R. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: IEEE: 7482-7491 [DOI: 10.1109/CVPR.2018.00781]

Kingma D P and Welling M. 2014. Auto-encoding variational bayes//Proceedings of the 2nd International Conference on Learning Representations. Banff, Canada: [s.n.]

Kitaev N, Kaiser L and Levskaya A. 2020. Reformer: the efficient transformer//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: OpenReview.net

Li J, Kang D, Pei W J, Zhe X, Zhang Y, He Z Y, et al. 2021a. Audio2Gestures: generating diverse gestures from speech audio with conditional variational autoencoders//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 11273-11282 [DOI: 10.1109/iccv48922.2021.01110]

Li R L, Yang S, Ross D A and Kanazawa A. 2021b. AI choreographer: music conditioned 3D dance generation with AIST++//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 13381-13392 [DOI: 10.1109/ICCV48922.2021.01315]

Liu H Y, Zhu Z H, Becherini G, Peng Y C, Su M Y, Zhou Y, et al. 2024. EMAGE: towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 1144-1154 [DOI: 10.1109/CVPR52733.2024.00115]

Liu H Y, Zhu Z H, Iwamoto N, Peng Y C, Li Z Q, Zhou Y, et al. 2022. BEAT: a large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 612-630 [DOI: 10.1007/978-3-031-20071-7_36]

Ma J Q, Zhao Z, Chen J L, Li A, Hong L C and Chi E H. 2019. SNR: sub-network routing for flexible parameter sharing in multi-task learning//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI Press: 216-223 [DOI: 10.1609/aaai.v33i01.3301216]

Massaro D W, Cohen M M, Tabain M, Beskow J and Clark R. 2012. Animated speech: research progress and applications//Bailly G, Perrier P and Vatikiotis-Bateson E, eds. Audiovisual Speech Processing. Cambridge, UK: Cambridge University Press: 309-345 [DOI: 10.1017/cbo9780511843891.014]

McNeill D and Arnheim R. 1994. Hand and mind: what gestures reveal about thought. Leonardo, 27(4): #358 [DOI: 10.2307/1576015]

Mikolov T, Grave E, Bojanowski P, Puhrsch C and Joulin A. 2018. Advances in pre-training distributed word representations//Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018. Miyazaki, Japan: European Language Resources Association (ELRA)

Pan Y, Li S X, Tan S, Wei J J, Zhai G T and Yang X K. 2025. Advancements in digital character stylization, multimodal animation, and interaction. Journal of Image and Graphics, 30(2): 334-360 (潘烨, 李韶旭, 谭帅, 韦俊杰, 翟广涛, 杨小康. 2025. 数字人风格化、多模态驱动与交互进展. 中国图象图形学报, 30(2): 334-360) [DOI: 10.11834/jig.230639]

Park I and Cho J. 2023. SAiD: speech-driven blendshape facial animation with diffusion [EB/OL]. [2025-11-01]. https://arxiv.org/pdf/2401.08655.pdf

Sun M Z, Xu C, Jiang X Y, Liu Y, Sun B G and Huang R Q. 2025. Beyond talking-generating holistic 3D human dyadic motion for communication. International Journal of Computer Vision, 133(5): 2910-2926 [DOI: 10.1007/s11263-024-02300-7]

Tang H Y, Liu J N, Zhao M and Gong X D. 2020. Progressive layered extraction (PLE): a novel multi-task learning (MTL) model for personalized recommendations//Proceedings of the 14th ACM Conference on Recommender Systems. [s.l.]: ACM: 269-278 [DOI: 10.1145/3383313.3412236]

Taylor S L, Mahler M, Theobald B J and Matthews I. 2012. Dynamic units of visual speech//2012 ACM SIGGRAPH/Eurographics Symposium on Computer Animation. Lausanne, Switzerland: ACM: 275-284

Tevet G, Raab S, Gordon B, Shafir Y, Cohen-Or D and Bermano A H. 2023. Human motion diffusion model//Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net

Thambiraja B, Habibie I, Aliakbarian S, Cosker D, Theobalt C and Thies J. 2023. Imitator: personalized speech-driven 3D facial animation//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 20564-20574 [DOI: 10.1109/ICCV51070.2023.01885]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Wang S, Zhang J N, Tan X, Xie Z F, Wang C J and Ma L Z. 2026. MMoFusion: multi-modal co-speech motion generation with diffusion model. Pattern Recognition, 169: #111774 [DOI: 10.1016/j.patcog.2025.111774]

Xiao L X, Lu S L, Pi H J, Fan K, Pan L, Zhou Y E, et al. 2025. MotionStreamer: streaming motion generation via diffusion-based autoregressive model in causal latent space//Proceedings of 2025 IEEE/CVF International Conference on Computer Vision. Honolulu, USA: IEEE: 10086-10096 [DOI: 10.1109/iccv51701.2025.00940]

Xing J B, Xia M H, Zhang Y C, Cun X, Wang J and Wong T T. 2023. CodeTalker: speech-driven 3D facial animation with discrete motion prior//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 12780-12790 [DOI: 10.1109/CVPR52729.2023.01229]

Xu Z N, Lin Y K, Han H N, Yang S C, Li R H, Zhang Y C, et al. 2024. MambaTalk: efficient holistic gesture synthesis with selective state space models//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 20055-20080 [DOI: 10.52202/079017-0633]

Yang S C, Wu Z Y, Li M L, Zhang Z S, Hao L, Bao W H, et al. 2023. DiffuseStyleGesture: stylized audio-driven co-speech gesture generation with diffusion models//Proceedings of the 32nd International Joint Conference on Artificial Intelligence. Macao, China: ijcai.org: 5860-5868 [DOI: 10.24963/ijcai.2023/650]

Yi H W, Liang H L, Liu Y F, Cao Q, Wen Y D, Bolkart T, et al. 2023. Generating holistic 3D human motion from speech//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 469-480 [DOI: 10.1109/CVPR52729.2023.00053]

Yoon Y, Cha B, Lee J H, Jang M, Lee J, Kim J, et al. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Transactions on Graphics, 39(6): #222 [DOI: 10.1145/3414685.3417838]

Yu T H, Kumar S, Gupta A, Levine S, Hausman K and Finn C. 2020. Gradient surgery for multi-task learning//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #489 [DOI:

10.48550/arxiv.2001.06782]

Zhao B Q, Fu Y Y, Su Z, Wang R M, Lyu C L and Luo X N. 2024. A survey on multimodal information-guided 3D human motion generation. Journal of Image and Graphics, 29(9): 2541-2565 (赵宝全, 付一愉, 苏卓, 王若梅, 吕辰雷, 罗笑南. 2024. 多模态信息引导的三维数字人运动生成综述. 中国图象图形学报, 29(9): 2541-2565) [DOI: 10.11834/jig.230626]

Zhou Y, Han X T, Shechtman E, Echevarria J, Kalogerakis E and Li D Z Y. 2020. MakeltTalk: speaker-aware talking-head animation. ACM Transactions on Graphics, 39(6): #221 [DOI: 10.1145/3414685.3417774]

作者简介

宣恩允,男,硕士研究生,主要研究方向为三维计算机视觉。E-mail:2300541007@email.szu.edu.cn

李游,通信作者,男,副研究员,硕士生导师,主要研究方向为3D视觉。E-mail:liyou@gml.ac.cn

李梓维,男,硕士研究生,主要研究方向为视觉定位。E-mail:2210273089@email.szu.edu.cn

姚萌萌,男,硕士研究生主要研究方向为点云数据处理。E-mail:yao-mengmeng@gmail.ac.cn

郭仁忠,男,教授,中国工程院院士,主要研究方向为地图制图、地理信息科学和智慧城市。E-mail:guorz@szu.edu.cn