上海交通大学硕士学位论文

基于面部捕捉、语音、头部运动的在线实时数字人驱动

姓　　名：花泉　润

导　　师：杨旭波教授

学　　号：122037990002

申 请 学 位：工学硕士

学 科/专 业：专业

院　　系：计算机学院

2025 年 7 月 15 日

A Dissertation Submitted to

Shanghai Jiao Tong University for the Degree of Master

FACECAPGES: REAL-TIME FRAME-BY-FRAME GESTURE GENERATION FROM AUDIO, FACIAL CAPTURE, AND HEAD POSE

Author: Jun Hanaizumi

Supervisor: Prof. Xubo Yang

Depart of XXX

Shanghai Jiao Tong University

Shanghai, P.R. China

July 15th, 2025

上海交通大学

学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全知晓本声明的法律后果由本人承担。

学位论文作者签名：

日期：　　年　　月　　日

上海交通大学

学位论文使用授权书

本人同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。

本学位论文属于：

□ **公开论文**

□ **内部论文**，保密 □ 1 年 / □ 2 年 / □ 3 年，过保密期后适用本授权书。

□ **秘密论文**，保密 ____ 年（不超过 10 年），过保密期后适用本授权书。

□ **机密论文**，保密 ____ 年（不超过 20 年），过保密期后适用本授权书。

（请在以上方框内选择打“✓”）

学位论文作者签名：　　指导教师签名：

日期：　　年　　月　　日　　日期：　　年　　月　　日

摘 要

近年来，基于人工智能的手势生成方法主要依赖语音模态为虚拟角色合成动作，从而减少对手工动画的依赖。然而，此类模型通常需要完整的语音或文本片段作为输入，难以满足如直播、元宇宙等对延迟敏感的实时应用需求。

面向用户的实时手势生成面临多重挑战。模型在无法访问未来信息的情况下，需仅依靠过去输入推断合适的手势，并确保与语音的时间对齐及自然表达效果。尽管已有研究尝试引入面部线索提升生成质量，但“头部姿态”作为一种易于获取、自然伴随语音出现的模态信息，仍未被充分利用。

为此，本文提出了 FaceCapGes，一种多模态级联网络，结合语音、面部捕捉和头部姿态三类信息，实现实时手势生成。本模型不依赖未来上下文，通过级联架构引入头部姿态信息，从而提升动作自然性。该设计特别适用于坐姿或受限场景，使用户仅凭面部与头部动作，即可驱动虚拟角色产生富有表现力的手势。

主观评价结果显示，FaceCapGes 在自然性方面与当前主流方法相当，同时在响应速度上具显著优势。所生成的手势与语音高度对齐，并具备良好的实时交互表现。此外，本模型可部署于如 iPhone 等轻量设备，只要输入格式兼容 ARKit 面部捕捉标准，即可广泛应用于多种实时互动场景中。

关键词：协同语音手势生成，数字人驱动，面部捕捉，多模态学习

Abstract

Recent AI-based gesture generation methods primarily rely on speech modalities to synthesize gestural motions for virtual avatars, reducing the need for manually crafted animations. However, these models typically require full speech or text segments as input, making them unsuitable for latency-sensitive applications such as live streaming or metaverse interactions.

Real-time gesture generation for user-facing applications poses significant challenges. Without access to future information, models must infer appropriate gestures solely from past inputs, while maintaining temporal alignment with speech and ensuring natural expressiveness. While prior work has explored facial cues to enhance generation quality, head pose —an easily captured modality that naturally accompanies speech —remains underutilized.

To address this, we propose FaceCapGes, a multimodal cascaded network that integrates speech, facial capture, and head pose for real-time gesture generation. Without relying on future context, our model incorporates head pose through a cascaded architecture to improve naturalness. This design is especially suited for seated or constrained settings, where users can drive expressive gestures for virtual avatars using only facial and head movements.

Subjective evaluations show that our method achieves naturalness on par with existing state-of-the-art models. The generated gestures demonstrate good alignment with speech and exhibit a significant advantage in real-time responsiveness. The model can run on lightweight devices such as an iPhone, provided the input is compatible with ARKit-based facial capture formats, enabling a wide range of real-time interactive scenarios.

Key words: co-speech gesture generation, virtual avatar driving, face-capture, multimodal learning

目 录

插 图

*

表 格

*

算 法

*

ϵ	介电常数
μ	磁导率
ϵ	介电常数
μ	磁导率
ϵ	介电常数
μ	磁导率
ϵ	介电常数
μ	磁导率
ϵ	介电常数
μ	磁导率
ϵ	介电常数
μ	磁导率
ϵ	介电常数
μ	磁导率
ϵ	介电常数
μ	磁导率
ϵ	介电常数
μ	磁导率
ϵ	介电常数
μ	磁导率
ϵ	介电常数
μ	磁导率
ϵ	介电常数
μ	磁导率
ϵ	介电常数
μ	磁导率
ϵ	介电常数
μ	磁导率
ϵ	介电常数
μ	磁导率
ϵ	介电常数
μ	磁导率
ϵ	介电常数
μ	磁导率
ϵ	介电常数
μ	磁导率
ϵ	介电常数

μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率
ϵ 介电常数
μ 磁导率

第 1 章 引言

在面对面的交流中，手势在增强语言表达的清晰性和表现力方面起着至关重要的作用。在虚拟人场景中，手势生成模型能够基于语音、面部线索或文本等模态自动合成身体动作。所生成的手势通常以三维骨架运动的形式表示，并可通过实时渲染引擎（如 Unity 或 Unreal Engine）重定向至虚拟角色绑定结构中。然而，现有方法多依赖完整语音或文本输入，不适用于用户输入持续变化、不可预知的交互式实时场景。

这一局限在元宇宙、直播和远程社交平台等虚拟人应用中尤为突出。尽管动作捕捉设备可以实现直观的手势控制，但其价格昂贵、使用复杂，许多用户难以获取。因此，对于使用台式机、智能手机或 VR 设备的用户而言，往往只能依赖预设动画，限制了表达的自然性与个性化。为了弥合这一差距，从语音与摄像头输入实时生成手势成为提升虚拟人交互能力的关键。

近年来，研究者已提出多种手势生成方法，包括多层感知机（MLP）[gesticulator2020,beatcamn]、循环神经网络（RNN）[beatcamn]、图神经网络[gesturemaster2022]、扩散模型[diffsheg,diffgesture,DiffTED2024]、VQ-VAE[emage]以及自监督学习方法[diffusion-self-supervised2023]。这些模型在生成质量上表现优异，但大多数依赖离线处理或未来上下文信息，难以满足实时应用需求。部分在线模型已被提出[towards_realtime_co_speech_gesture_generation,realtime_gesture_animation_generation]，但往往仍需短时未来窗口，并忽略了常见设备中可用的重要模态信息。

为解决上述问题，本文提出 FaceCapGes，一种无需动作捕捉设备即可实现用户驱动的虚拟人控制的帧级实时手势生成模型。本模型支持在无未来上下文的情况下进行在线推理，能从实时输入中持续预测手势。我们首次引入“头部姿态”作为新模态，与语音和面部捕捉结合，利用其与自然手势节奏的高度相关性，提升生成动作的自然度。各模态通过级联多模态架构进行处理，以学习其联合关系，从而在实时约束下提升手势表达力。

本方法可部署于轻量设备，具体实现基于配备 Apple ARKit 面部捕捉系统的 iPhone。尽管模型本身计算效率高，具备实时推理能力，但目前仍依赖 BEAT 数据集提供的 ARKit 专用面部数据格式，从中提取面部表情与头部动作。因此，跨平台部署能力仍受限，对通用摄像头、桌面环境或 VR 头显的支持仍是一个开放问题。

通过在 BEAT 数据集上的实验表明，引入头部姿态显著提升了手势自然性，同时保持了低延迟的实时性能。实验结果说明，即使在无动作捕捉设备与未来输入的

情况下，用户也可仅通过语音与头部动作驱动富有表现力的虚拟角色，只要输入符合 ARKit 格式即可。

综上，本文的主要贡献包括：

1. 提出 FaceCapGes，一种帧级实时手势生成模型，使用户无需动作捕捉设备，仅通过语音即可驱动虚拟人手势动画；
2. 将头部姿态作为新模态引入多模态级联架构，在实时约束下提升手势的自然性与表现力；
3. 实验结果在手势自然性、语音对齐度以及实时响应方面表现优异，适用于兼容 ARKit 的设备。

为实现实时推理，模型采用滑动窗口的自回归训练策略，并在不依赖未来信息的情况下处理各输入模态。此外，我们使用六维连续旋转表示（Rot6d）[rot6d] 替代传统的欧拉角表示，从而提升动作多样性。后续章节将详细介绍模型架构、训练策略及在 BEAT 数据集上的评估结果，并在主观质量、时序对齐与推理延迟方面与现有手势生成模型进行对比（见图 **??**，表 **??**，表 **??**）。

第 2 章 数学与引用文献的标注

2.1 数学

2.1.1 数字和单位

宏包 siunitx 提供了更好的数字和单位支持：

- 12 345.678 90
- $1 \pm 2\mathrm{i}$
- 0.3×10^{45}
- $1.654 \times 2.34 \times 3.430$
- $\mathrm{kg\,m\,s^{-1}}$
- μm μm
- Ω Ω
- 10 和 20
- 10, 20 和 30
- 0.13 mm, 0.67 mm 和 0.80 mm
- 10 ~ 20
- 10 °C ~ 20 °C

2.1.2 数学符号和公式

按照国标 GB/T 3102.11—1993《物理科学和技术中使用的数学符号》，微分符号 d 应使用直立体。除此之外，数学常数也应使用直立体：

- 微分符号 d：`\dd`
- 圆周率 π：`\uppi`
- 自然对数的底 e：`\ee`
- 虚数单位 i, j：`\ii \jj`

公式应另起一行居中排版。公式后应注明编号，按章顺序编排，编号右端对齐。

$$\mathrm{e}^{\mathrm{i}\pi} + 1 = 0, \tag{2.1}$$

$$\frac{\mathrm{d}^2 u}{\mathrm{d}t^2} = \int f(x)\,\mathrm{d}x. \tag{2.2}$$

公式末尾是需要添加标点符号的，至于用逗号还是句号，取决于公式下面一句是接着公式说的，还是另起一句。

$$\frac{2h}{\pi}\int_0^{\infty}\frac{\sin(\omega\delta)}{\omega}\cos(\omega x)\,\mathrm{d}\omega=\begin{cases}h, & |x|<\delta,\\ \frac{h}{2}, & x=\pm\delta,\\ 0, & |x|>\delta.\end{cases}\tag{2.3}$$

公式较长时最好在等号“=”处转行。

$$\begin{aligned}&I(X_3;X_4)-I(X_3;X_4\mid X_1)-I(X_3;X_4\mid X_2)\\ =&[I(X_3;X_4)-I(X_3;X_4\mid X_1)]-I(X_3;X_4\mid \tilde{X}_2)\qquad(2.4)\\ =&I(X_1;X_3;X_4)-I(X_3;X_4\mid \tilde{X}_2).\qquad(2.5)\end{aligned}$$

如果在等号处转行难以实现，也可在 +、−、×、÷ 运算符号处转行，转行时运算符号仅书写于转行式前，不重复书写。

$$\begin{aligned}\frac{1}{2}\Delta(f_{ij}f^{ij})=2\Bigg(&\sum_{i<j}\chi_{ij}(\sigma_i-\sigma_j)^2+f^{ij}\nabla_j\nabla_i(\Delta f)\\ &+\nabla_k f_{ij}\nabla^k f^{ij}+f^{ij}f^k\left[2\nabla_i R_{jk}-\nabla_k R_{ij}\right]\Bigg).\end{aligned}\tag{2.6}$$

2.1.3　定理环境

示例文件中使用 ntheorem 宏包配置了定理、引理和证明等环境。用户也可以使用 amsthm 宏包。

这里举一个“定理”和“证明”的例子。

定理 2.1 (留数定理)　假设 U 是复平面上的一个单连通开子集，$a_1,\ldots,a_n$ 是复平面上有限个点，f 是定义在 $U\backslash\{a_1,\ldots,a_n\}$ 上的全纯函数，如果 γ 是一条把 $a_1,\ldots,a_n$ 包围起来的可求长曲线，但不经过任何一个 a_k，并且其起点与终点重合，那么：

$$\oint_{\gamma}f(z)\,\mathrm{d}z=2\pi\mathrm{i}\sum_{k=1}^{n}\mathrm{I}(\gamma,a_k)\operatorname{Res}(f,a_k).\tag{2.7}$$

如果 γ 是若尔当曲线，那么 $\mathrm{I}(\gamma,a_k)=1$，因此：

$$\oint_{\gamma}f(z)\,\mathrm{d}z=2\pi\mathrm{i}\sum_{k=1}^{n}\operatorname{Res}(f,a_k).\tag{2.8}$$

在这里，$\mathrm{Res}(f, a_k)$ 表示 f 在点 a_k 的留数，$\mathrm{I}(\gamma, a_k)$ 表示 γ 关于点 a_k 的卷绕数。卷绕数是一个整数，它描述了曲线 γ 绕过点 a_k 的次数。如果 γ 依逆时针方向绕着 a_k 移动，卷绕数就是一个正数，如果 γ 根本不绕过 a_k，卷绕数就是零。

定理 2.1 的证明。

证明 首先，由……

其次，……

所以…… ■

2.2 引用文献的标注

按照教务处的要求，参考文献外观应符合国标 GB/T 7714 的要求。模版使用 BibLaTeX 配合 biblatex-gb7714-2015 样式包[①]控制参考文献的输出样式，后端采用 biber 管理文献。

请注意 biblatex-gb7714-2015 宏包 2016 年 9 月才加入 CTAN，如果你使用的 TeX 系统版本较旧，可能没有包含 biblatex-gb7714-2015 宏包，需要手动安装。BibLaTeX 与 biblatex-gb7714-2015 目前在活跃地更新，为避免一些兼容性问题，推荐使用较新的版本。

正文中引用参考文献时，使用 `\cite{key1,key2,key3...}` 可以产生“上标引用的参考文献”，如[Yu2001,Cheng1999,LSC1957]。使用 `\parencite{key1,key2,key3...}` 则可以产生水平引用的参考文献，例如 **Li1999**,**Jiang1989**,**Hopkinson1999**。请看下面的例子，将会穿插使用水平的和上标的参考文献：普通图书**Yu2001**,**Jiang1998**，论文集、会议录[CSTAM1990]，科技报告**WHO1970**，学位论文[Zhang1998]，专利文献**Jiang1989**,**HBLZ2001**，专著中析出的文献[Cheng1999,GBT2659]，期刊中析出的文献**Li1999**,**Li2000**，报纸中析出的文献[Ding2000]，电子文献 **Jiang1999**,**Christine1998**,**Xiao2001**。

可以使用 `\nocite{key1,key2,key3...}` 将参考文献条目加入到文献表中但不在正文中引用。使用 `\nocite{*}` 可以将参考文献数据库中的所有条目加入到文献表中。

① https://www.ctan.org/pkg/biblatex-gb7714-2015

第 3 章 浮动体

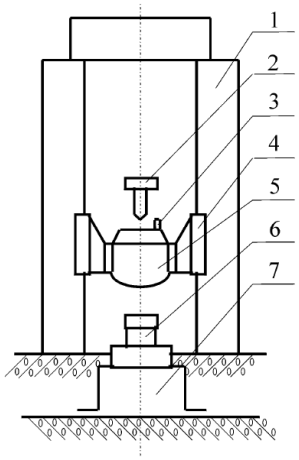
3.1 插图

插图功能是利用 TEX 的特定编译程序提供的机制实现的，不同的编译程序支持不同的图形方式。有的同学可能听说“LATEX 只支持 EPS”，事实上这种说法是不准确的。XETEX 可以很方便地插入 EPS、PDF、PNG、JPEG 格式的图片。

一般图形都是处在浮动环境中。之所以称为浮动是指最终排版效果图形的位置不一定与源文件中的位置对应，这也是刚使用 LATEX 同学可能遇到的问题。如果要强制固定浮动图形的位置，请使用 float 宏包，它提供了 `[H]` 参数。

3.1.1 单个图形

图要有图题，研究生图题采用中英文对照，并置于图的编号之后，图的编号和图题应置于图下方的居中位置。引用图应在图题右上角标出文献来源。当插图中组成部件由数字或字母等编号表示时，可在插图下方添加图注进行说明，如图 3.1 所示。



1. 立柱 2. 提升释放机构 3. 标准冲击加速度计
4. 导轨 5. 重锤 6. 被校力传感器 7. 底座

图 3.1 单个图形示例[He1999]。如果表格的标题很长，那么在表格索引中就会很不美观。可以在前面用中括号写一个简短的标题，这个标题会出现在索引中。

Figure 3.1 Stay hungry, stay foolish.