



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 12 月 8 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  
December 8<sup>th</sup>, 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

手势是交流中重要的细节补充与情感表达载体。面向虚拟人/数字形象的肢体控制中,传统方案往往依赖穿戴式检测设备或其他专用传感器,以稳定获取人体运动信息,但这类设备带来较高的使用门槛与部署成本。

随着图像识别技术的发展,计算机能够通过相机识别用户动作并实现无穿戴的动捕式交互,在一定程度上降低了设备负担。然而,这类方案仍要求用户在镜头前实际表演手势动作;在远程交流或沉浸式场景中,手势作为情感表达工具往往伴随空间占用与体力消耗,人们需要投入较多精力活动肢体以增强说服力或表达效果,从而使虚拟世界中的手势表达依然显得繁琐。

近年来,手势与语言的关系被进一步分析,使得计算机能够从文本或语音生成与语义一致的手势动画,让数字人物在发言中融入更符合人类交流习惯的非语言表达。若将该技术用于用户的实时数字形象,则有望提供更轻便的手势控制方案:用户无需实际做出手势,只需在相机前说话即可,同时自然地产生语音、面部表情与头部旋转等信息。面部表情可补充情感线索,头部旋转可为手势朝向与叙事空间锚定提供指导,从而改进依赖空间方位与路径方向的手势推理;且这些信息的实时获取技术已相对成熟,具备良好的在线性能。

然而,当前协同语音手势生成技术通常以获取完整文本为前提,通过语义分析生成高精度手势,难以直接适配逐字输入的用户语音实时场景。因此,本论文提出 FaceCapGes 方法,基于语音、面部表情与头部旋转三种实时信息生成在线实时的 3D 手势骨骼动画。该方法可为用户的实时数字形象添加手势动画,无需用户实际做出手势动作;同时模型不依赖未来输入,能够在实时环境中提升虚拟形象的表达能力。

以往研究已对面部表情、语音与手势的关系进行分析,并提出成熟的学习方法。本模型在现有级联架构基础上,增加头部姿态特征分析模块,同时引入滑动窗口机制以实现架构的实时运行。本模型所依赖的框架及额外添加模块性能开销低,与面部捕捉和头部姿态计算任务同时运行时,仍具备良好的实时性能。

主观评价结果表明,在自然性方面,本方法与当前主流方法相当;在响应速度上,具有显著优势。生成的手势与语音高度对齐,且具备良好的实时交互表现。此外,本模型可部署在 iPhone 等轻量设备上,只要输入格式兼容 ARKit 面部捕捉标准,就能广泛应用于各类实时互动场景。

**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

Gestures are a crucial channel for enriching communication with complementary details and emotional expression. In the control of virtual humans or digital avatars, traditional approaches often rely on wearable sensing devices or other dedicated sensors to reliably capture human motion; however, such hardware introduces substantial barriers to use and deployment costs.

With advances in visual recognition, camera-based systems can capture user movements and enable markerless, mocap-style interaction, partially reducing the dependence on wearables. Nevertheless, these approaches still require users to physically perform gestures in front of the camera. In remote communication or immersive scenarios, expressive gesturing often entails non-trivial spatial and physical effort: users must invest energy in moving their limbs to strengthen persuasion or convey emotions, which can make gestural expression in virtual worlds cumbersome.

Meanwhile, growing understanding of the relationship between gestures and language has enabled computers to generate semantically consistent gestures from text or speech, allowing digital characters to incorporate nonverbal behaviors that better match human conversational habits. If applied to a user's real-time avatar, such techniques could provide a more lightweight control paradigm: instead of performing gestures, users only need to speak in front of a camera, naturally producing speech, facial expressions, and head rotations. Facial expressions can supplement affective cues, while head rotations can guide gesture orientation and narrative spatial anchoring, improving the inference of gestures that depend on spatial references and motion directions. Moreover, real-time capture of facial expressions and head pose is relatively mature and offers strong online performance.

However, most existing co-speech gesture generation methods assume access to complete text and rely on semantic analysis to generate high-quality gestures, making them difficult to adapt to real-time scenarios with incremental, word-by-word user speech. To address this, we propose FaceCapGes, a method that generates online, real-time 3D skeletal gesture animations from three streams of real-time signals: speech, facial expressions, and head rotations. FaceCapGes can augment a user's live digital avatar with expressive gestures without requir-

ing the user to physically gesture, and it does not depend on future inputs, thereby enhancing avatar expressiveness in live interactive settings.

Prior work has analyzed the relationships among facial expressions, speech, and gestures, and has established effective learning paradigms. Building on existing cascaded architectures, our model introduces an additional head-pose feature analysis module and incorporates a sliding-window mechanism to enable real-time operation. The underlying framework and added modules incur low computational overhead, and the system maintains robust real-time performance even when running concurrently with facial capture and head-pose estimation.

Subjective evaluations show that our method achieves naturalness comparable to current mainstream approaches, while providing a significant advantage in responsiveness. The generated gestures are tightly aligned with speech and demonstrate strong real-time interactive behavior. Furthermore, our model can be deployed on lightweight devices such as iPhones; as long as the input format is compatible with the ARKit facial capture standard, it can be broadly applied to a wide range of real-time interactive scenarios.

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

# 目 录

<b>第 1 章 引言 .....</b>	<b>1</b>
1.1 研究背景和意义.....	1
1.2 研究内容.....	2
1.3 论文组织架构.....	3
<b>第 2 章 相关工作 .....</b>	<b>5</b>
2.1 手势的定义与身体姿态的参数化表示.....	5
2.1.1 手势的定义与范围 .....	5
2.1.2 身体姿态的参数化表示 .....	6
2.2 面部表情的定义与参数化表示.....	8
2.3 国内外研究现状.....	10
2.3.1 手势生成的研究目标 .....	10
2.3.2 手势生成的演变 .....	12
2.3.3 当代生成研究的策略趋势 .....	13
2.4 实时生成的理论基础与可行性分析.....	14
2.5 本文的工作与创新点.....	15
2.6 本章小结.....	16
<b>第 3 章 方法 .....</b>	<b>17</b>
3.1 研究定位与总体设计思路.....	17
3.2 系统整体框架与模块定位.....	17
3.2.1 信号采集与系统配置层 .....	18
3.2.2 手势生成模型层（FaceCapGes） .....	19
3.2.3 渲染驱动层 .....	19
3.3 问题定义.....	19
3.3.1 任务描述 .....	20
3.3.2 输入与输出模态 .....	20
3.3.3 学习目标与优化形式 .....	21

3.4 级联架构与输入模态设计.....	21
3.4.1 级联架构的原理与理论背景 .....	21
3.4.2 基线模态继承与实时适配 .....	22
3.4.3 头部姿态模态的引入与结构位置 .....	26
3.4.4 小结：从语义驱动到反应调节的信号层级 .....	27
3.5 训练与损失函数设计.....	27
3.6 实现与训练配置.....	28
3.7 本章小结.....	30
<b>第 4 章 评估 .....</b>	<b>31</b>
4.1 评估设置.....	31
4.2 客观评估指标与实现细节.....	31
4.2.1 Fréchet Gesture Distance (FGD) .....	31
4.2.2 Speech-Gesture Rhythm Correlation (SRGR).....	33
4.2.3 Beat Alignment (BA) .....	33
4.2.4 L1-based Diversity (L1DIV).....	34
4.3 模型对比概览.....	35
4.4 用户调研.....	35
4.4.1 评估系统与实验配置 .....	35
4.4.2 平衡拉丁方排序设计 .....	38
4.4.3 结果与分析 .....	38
4.5 定性分析.....	40
4.6 定量分析.....	40
4.7 消融实验分析.....	41
4.8 性能评估.....	41
<b>第 5 章 结论 .....</b>	<b>43</b>
<b>参考文献.....</b>	<b>45</b>
<b>附录 A 平衡拉丁方排序设计 .....</b>	<b>49</b>
<b>致 谢.....</b>	<b>51</b>
<b>学术论文和科研成果目录.....</b>	<b>53</b>

个人简历..... 55



## 插图

- 图 2.1 示例图来自<sup>[21]</sup>，展示 FACS AU45 (blink) 对应的 ARKit 中的两种 BlendShape 基形: eyeBlinkLeft (闭左眼) 与 eyeBlinkRight (闭右眼)。  
.....9
- 图 2.2 BlendShape 线性插值效果示例: eyeBlink 从  $w=0$  (左) 到  $w=1$  (右) 的连续变化, 中图为中间值。BS 权重可直接用于实时渲染驱动。 .....9
- 图 2.3 MediaPipe Face Mesh 关键点结构示意图, 截取自<sup>[1]</sup>。 ..... 10
- 图 3.1 系统整体架构与数据流示意图..... 18
- 图 3.2 双向与单向 LSTM 对比示意图。双向结构 (上) 在每个时间步同时利用历史与未来帧特征进行建模; 单向结构 (下) 仅基于历史帧进行递推, 以保持因果性并支持流式推理。 ..... 24
- 图 3.3 滑动窗口训练机制: 模型通过自回归循环预测  $M$  帧, 每步使用  $N$  帧上下文并预测第  $N + 1$  帧。损失函数累计所有预测帧的误差, 保证因果性与时间平滑性。 ..... 26
- 图 3.4 头部姿态编码器结构示意图。输入为 Rot6d 表示的 6 维旋转向量, 经两层前馈网络与 ReLU 非线性映射, 输出 12 维紧凑潜在表征。该编码结果与语音、面部嵌入拼接后输入解码器, 用于补充动作的方向与节奏信号。 . 27
- 图 3.5 FaceCapGes 模型结构: 音频、面部、头部编码器分别提取模态特征后拼接, 输入至 LSTM 解码器生成躯干与手部动作, 仅保留最后一帧输出作为当前时刻预测, 符合帧级实时推理设定。训练阶段历史姿态序列比目标长度少一帧, 需进行零填充。 ..... 28
- 图 3.6 BEAT 数据集的骨架拓扑结构。蓝色部分为 FaceCapGes 模型的控制区域, 涵盖上肢与三段脊椎关节, 其余节点保持静态。 ..... 29
- 图 4.1 用户研究系统界面示意图。左、中、右位置随机分配给三种模型, 音频与面部表情完全一致, 仅身体动作不同。 ..... 37
- 图 4.2 16 名参与者对三种模型在“真实感”、“同步性”和“多样性”三个维度的主观排名结果。 ..... 38
- 图 4.3 平衡拉丁方实验版本中 8 位 VR 用户的主观排名结果, 评价维度包括“真实感”、“同步性”和“多样性”。 ..... 39

图 4.4 GT、CaMN、DiffSHEG 与 FaceCapGes 的动作效果对比。本模型手势响应自然，方向与头部朝向一致。 .....	40
--	----

\*



表 格

表 2.1 不同旋转表示方式的空间连续性与使用示例.....8

表 2.2 两类手势生成任务在约束与可利用条件上的对比..... 11

表 3.1 输入输出模态符号与维度..... 20

表 4.1 在 BEAT 四位说话人测试集上的定量评估结果。FaceCapGes 在所有指标  
上优于 CaMN，且与 DiffSHEG 表现相当。..... 40

表 4.2 说话人 2 的消融实验结果。头部姿态对提升手势自然性与表达力有显著  
作用。..... 41

表 4.3 在 RTX 4090 上的帧级推理速度评估结果。..... 41

\*



算 法

\*

$\epsilon$  介电常数  
 $\mu$  磁导率  
 $\epsilon$  介电常数  
 $\mu$  磁导率



## 第1章 引言

### 1.1 研究背景和意义

近年来,随着元宇宙、虚拟社交与直播等领域的相关技术日趋成熟,用户已能够使用任意外观的虚拟人作为交互载体,在虚拟空间中与异地用户进行交流。虚拟人3D模型的姿态由其内部骨架关节的旋转参数(如欧拉角、四元数等)定义,最终通过蒙皮渲染技术完成可视化。得益于自动骨骼绑定技术,骨骼动画的生成可消除不同3D模型间的骨架拓扑差异,实现跨模型复用。

在虚拟人交互中,穿戴式动作捕捉设备是实时驱动手势的传统方案,贴合肢体的标记点可将肢体运动实时转换为相同的骨骼动画,提供直观、准确的操控。尽管其精度较高,但对于大多数用户而言,此类设备存在功能用途单一、硬件成本昂贵、便携性差等问题,限制了使用频率。因此,在当前的虚拟人交互应用中,仅少数专业用户会使用此类设备,而多数用户在设备限制下无法简单控制虚拟人肢体,导致了两种用户体验之间的不一致。

针对普通用户对低门槛虚拟人交互的需求,基于相机的动作捕捉技术<sup>[1-2]</sup>成为主流替代方案。该技术无需额外硬件,仅通过手机或电脑的内置相机即可实时捕捉用户动作,用于转化为骨骼动画。然而,该方案仍存在三种局限:一是这种方法要求用户在空中面向相机做出手势,过程中难以同步操作键盘、鼠标,造成操作冲突;二是手势活动范围受相机视野限制,在自然使用距离通过手机或电脑的内置相机拍摄用户,将严重限制手势的捕捉范围;三是持续的手势动作会产生体力消耗,在直播等长时间使用场景中,用户疲劳问题将变得显著。”

因此,我们提出一种新的需求:一种无需用户实际做出手势,仅通过用户的实时语音、面部捕捉与头部姿态,实时生成与语义和情感相匹配的手势骨骼动画。该方法旨在降低使用门槛,并解决操作冲突与体力消耗问题。

然而,现有研究尚未提供成熟的解决方案满足上述需求。首先,当前多数手势生成方法依赖完整的语音或文本输入。由于用户的语音输入是逐字进行的,计算机需要等待用户的未来输入才能解析当前的语义,造成动画生成的延迟。其次,现有研究未对用户的头部姿态模态做深入研究。头部姿态对手势的节奏与朝向具有明确的关系,且可以通过常规相机实时捕捉,但利用该模态来增强生成手势的自然度的相关研究尚不充分。

为此, 本文提出了一种新颖的实时手势生成模型。该模型以帧为单位, 输入语音、面部表情与头部姿态数据, 并逐帧输出对应的骨骼动画。本文首次在实时手势生成中, 将头部姿态作为一种新的模态引入, 利用其与自然手势节奏与朝向的高度相关性, 结合语音和面部信息共同提升生成动作的自然度。我们采用级联多模态架构与自回归训练来融合这些模态, 以学习其联合表征, 从而在严格的实时约束下增强手势的表现力。

本文的主要贡献如下:

1. 提出了 FaceCapGes, 一种帧级实时手势生成模型, 使用户无需动作捕捉设备或实际做出手势, 仅通过语音等常见输入即可驱动虚拟人的手势动画;
2. 将头部姿态作为新模态引入多模态级联架构, 在实时生成约束下有效提升了手势的自然性与表现力;
3. 通过实验结果表明, 该模型在手势自然性、语音-手势对齐度与实时响应方面展示了良好的性能。其框架适用于所有兼容 ARKit 的设备。

## 1.2 研究内容

本文的研究目标是构建一种基于语音、面部捕捉与头部姿态的实时数字人手势生成模型, 实现无需动作捕捉设备即可驱动虚拟角色自然表达的实时动画系统。该研究旨在解决现有手势生成方法对未来上下文的依赖及实时性不足的问题, 从而为虚拟人交互提供更低门槛、更高沉浸度的解决方案。

为实现上述目标, 本文的主要研究内容如下:

其一, 设计一种帧级手势生成架构。为实现帧级实时推理, 本架构在基线模型<sup>[3]</sup>的基础上做两种调整: 一方面, 保留其语音、面部表情等可实时获取的输入模态, 移除非实时模态的输入分支与特征处理模块; 另一方面, 采用滑动窗口式自回归训练方式, 确保模型仅依赖历史与当前帧信息进行推理 (不依赖未来上下文), 同时通过窗口内时序依赖建模保持动作的时间连续性。由此使架构可处理逐帧输入的实时流数据。

其二, 提出一种引入头部姿态的新型模态融合策略。在现有语音、面部表情生成手势的模型基础上, 将头部姿态作为终端模态引入, 设计头部姿态编码器并提取特征, 以增强生成手势的朝向的自然性。

其三, 搭建实验系统与用户测试环境。本文基于主流渲染引擎构建了虚拟人驱动与手势动画的可视化系统。该系统作为评估平台, 为后续的主观实验提供了统一环

境。

综上所述，本文引入头部姿态的新模态，来辅助手势生成模型从历史与现在信息推理手势的能力，并且提供了一个实验场景验证模型的推理质量与实时性能。

### 1.3 论文组织架构

本文共分为六个章节，内容安排如下：

第一章为绪论，介绍本研究的背景与意义，阐述研究目标、主要内容及核心贡献，并说明论文的整体组织结构。

第二章为相关工作，回顾了国内外在语音驱动手势生成、多模态学习及实时面部捕捉技术等方面的研究现状，分析了现有方法的不足，并明确了本文的研究定位与创新点。

第三章为方法，详细介绍了本文提出的 FaceCapGes 模型的整体架构与算法流程，包括多模态输入编码、姿态解码机制、滑动窗口自回归训练策略及对抗优化过程。

第四章为评估，阐述了模型在 BEAT 数据集上的实验设置与性能评估，涵盖客观指标、用户主观实验及消融研究；同时分析了模型在实时交互场景中的表现与应用潜力。

第五章为结论，总结了本文的主要研究成果与贡献，讨论了当前系统的局限性，并对未来在语义手势生成与跨平台实时驱动方向上的研究进行了展望。

第六章为附录与致谢部分，包含用户实验的平衡拉丁方设计、附加实验结果、参考文献及致谢内容。





## 第2章 相关工作

### 2.1 手势的定义与身体姿态的参数化表示

#### 2.1.1 手势的定义与范围

在人类交流中,手势是常与语音同时出现的身体动作信号,它承担语义补充、情感表达与互动调节等多重功能。这类动作通常呈现出明确的表达意图,使其区别于姿态平衡,移动等纯功能性动作<sup>[4]</sup>。因此,手势与语言并非两个相互独立的系统,而是源于共同的认知的表达机制。

在本研究的广义定义下,手势的运动形式不局限于手部运动,还可扩展至头部运动、躯干姿态等与表达相关的所有上半身动作。

**Kendon 连续体** Kendon 连续体<sup>[4-5]</sup>将与表达相关的手势行为,基于语言化程度做了以下分类:*gesticulation*(随语手势)→*language-like gestures*(语言样手势)→*pantomime*(拟态/哑剧式动作)→*emblems*(约定俗成的象征手势)→*sign language*(手语)。越靠左的类型通常更依赖当前的言语与语境、形式更即兴;越靠右则越接近离散的符号系统,规约化程度高,意义更稳定,可在缺少口语的情况下独立传达。

当前人机交互、虚拟人/数字人驱动等方向的手势生成,多数工作聚焦于 Kendon 连续体最左侧的随语手势,即说话过程中自然出现、与语音节奏与语义强关联的上肢动作。其含义往往依赖当前的口语与语境,脱离它们时通常难以传达清晰含义。

相对而言,连续体右侧的手势更接近文化中约定俗成的符号,可以脱离口语独立传达含义。例如,表达称赞的拍手动作,或表示“请保持安静”的嘴前竖起食指的动作,这些被视为象征手势,通常不在随语手势生成领域的研究对象中。

本文的研究范围据此限定在随语手势的学习与生成。

**随语手势的分类** McNeill<sup>[5]</sup>将随语手势进一步划分为四种基本类型:

1. **Iconic gestures (形象性手势)**: 以具象方式描绘事物的外形、空间路径或动作特征。例如,用手势勾勒一个物体的轮廓,或划出一道线表示移动的轨迹。此类手势与语言内容直接对应,表达具体语义。
2. **Metaphoric gestures (隐喻性手势)**: 表达抽象概念或思维结构的手势,比如用双手做出捧起一个物体放到一边的动作,表示“先把这个话题放一边”。这种手

势并不描绘实体，而是以具象化的方式呈现抽象语义。

3. **Deictic gestures (指示性手势)**: 指向空间中的对象、人物或方向，常用于对话焦点的指明与注意引导。
4. **Beat gestures (节奏型手势)**: 与语音重音、韵律或节奏同步的节奏性动作，通常不承载具体语义，但可用于强调语音节奏，引起听众对说话内容的注意。

这四类手势构成了随语手势在语义与语篇功能上的主要维度，并在自然交流中常以复合形式出现。在手势生成任务中，研究通常将其视为不同的可生成目标：其中节奏型手势由于与语音韵律高度同步、对齐与建模相对容易，长期以来在数据驱动方法中更常被优先刻画；而形象性与隐喻性等语义相关手势则对文本的语义推理有更高要求，因而是更具挑战性的方向。

鉴于本文以低延迟在线驱动为目标，本文优先建模与韵律强耦合的节奏型手势；语义一致的形象性/隐喻性手势留作后续工作。

**头部手势的分类** 除手部动作外，头部动作同样是手势的重要组成部分。头部的点动与摆动在时间结构上常与手势及语音节奏保持同步<sup>[6]</sup>，在语用功能上既能辅助语音韵律的组织，也能表达态度与指向信息。

在不同研究中，头部动作被从多个维度加以分析，其主要功能可归纳为以下几方面：

1. **韵律相关 (prosodic)** 动作反映语音重音与句法节奏的对应关系<sup>[7]</sup>；
2. **语义或态度相关 (semantic/attitudinal)** 动作表达说话者的情绪倾向与交际意图<sup>[4-5]</sup>；
3. **指向相关 (deictic)** 动作通过转头或注视方向建立叙事空间的参照<sup>[5]</sup>。

此外，研究表明头部动作的启动时间往往早于发声<sup>[8]</sup>。具体而言，头部动作存在启动与加速过程，若其峰值需与重读音节的时间对齐，则动作必须提前起势。因此，头部动作可能对即将到来的语音韵律具有前瞻性。

这一特征揭示了头部动作与语音之间的时序关系，说明视觉模态中的运动信号有时可先于声学事件出现。本文研究也因此关注头部动作，并将其纳入输入模态。

### 2.1.2 身体姿态的参数化表示

手势作为身体运动的子集，其生成和识别依赖于身体姿态的连续建模。因此，在进一步讨论手势生成方法之前，在此明确身体姿态的参数化表示方式。

**骨架结构的定义** 在计算机动画与动作捕捉领域，身体姿态通常由骨架结构和关节旋转参数共同定义。骨架结构描述了人体各关节的拓扑关系及层级依赖；而每个姿态帧由一组关节旋转参数所确定，这些参数定义了相对于父节点的旋转变换。在不同的系统与任务中，骨架结构的具体形式往往有所差异，这种差异直接影响姿态数据的表示与学习方式。

在不同的系统与任务中，骨架结构可以遵循各自的标准，因此，不同的数据集、3D 模型或神经网络往往基于自身定义的关节层级与命名体系进行训练与标注。例如，AMASS 数据集<sup>[9]</sup>采用 SMPL 拓扑结构<sup>[10]</sup>，BEAT 数据集<sup>[3]</sup>使用简化上半身骨架。近年来的自动骨骼绑定与骨架归一化方法，通过学习或优化关节对应关系，实现了不同拓扑之间的姿态重定向（pose retargeting）<sup>[11-12]</sup>，从而消除了模型依赖于特定骨架结构的限制。

**旋转参数的选取** 在确定骨架结构之后，具体的关节状态可通过多种旋转参数进行描述。常见的旋转参数表示方法包括：

1. 欧拉角（Euler Angles）：通过三个顺序旋转角表示姿态，直观且参数维度低（3 维），适合存储，但运算存在万向节锁（Gimbal Lock）问题，导致特定姿态下自由度丢失，且插值过程易产生非物理运动；
2. 四元数（Quaternion）：以四维单位向量（ $q_0^2 + q_1^2 + q_2^2 + q_3^2 = 1$ ）表示旋转，避免欧拉角的奇异性的同时，实现平滑插值。然而在神经网络回归中通常需要显式处理单位长度约束，且同一旋转存在双覆盖（ $q$  与  $-q$  等价），优化不稳定。
3. Axis-Angle：以旋转轴（三维单位向量）与旋转角度（标量）的乘积表示旋转，参数紧凑（3 维），是参数化人体模型 SMPL<sup>[10]</sup>的主要姿态存储格式。但角度取值范围存在周期性（ $[0, 2\pi)$ ），导致参数空间不连续，影响神经网络学习；
4. 旋转矩阵：以  $3 \times 3$  的正交矩阵，表示坐标系或物体绕某个轴旋转一定角度后的新姿态。常用于计算机图形学、机器人学和物理仿真中，但正交约束（ $R^T R = I$  且  $\det(R) = 1$ ）提高神经网络的学习难度。
5. Rot6d<sup>[13]</sup>：将  $3 \times 3$  旋转矩阵的前两列展开为六维向量，通过 Gram-Schmidt 正交化过程自动保持列向量的正交性与单位长度，无需额外约束。该表示既继承了旋转矩阵的数值稳定性，又解决了其参数冗余和正交性约束难题，同时具备连续性优势。

在计算机图形与实时渲染中，通常采用四元数或旋转矩阵进行骨骼变换与插值，以保证数值稳定性和计算效率。然而，在深度学习任务中，这些表示方式的约束会对

训练效率造成影响：四元数需满足单位长度约束；旋转矩阵则需满足元素的正交约束。近年来的研究表明<sup>[13]</sup>，Rot6d 既规避了上述两种表示的约束难题，又保持了姿态空间的连续性与物理合理性，在动作生成与姿态预测任务中展现出更优的可学习性。

手势生成方法对旋转参数的选取经历了一个演变过程。早期工作采用的方法各异（如欧拉角、Axis-Angle），多受限于特定框架或历史因素；而近期研究，出于对神经网络训练稳定性的追求，已显著趋向于采用 Rot6d 表示。这一趋势的实例对比见表 2.1。

表 2.1 不同旋转表示方式的空间连续性与使用示例

表示方式	维度	连续性	典型应用场景	训练使用示例
欧拉角	3	存在万向节锁	姿态存储	CaMN <sup>[3]</sup>
四元数	4	连续，但存在单位长度约束	图形学	—
Axis-Angle	3	角度部分不连续	神经网络	DiffSHEG <sup>[14]</sup>
旋转矩阵	9	连续，但存在正交约束	图形学、机器人学	MambaGesture <sup>[15]</sup>
Rot6d	6	连续	神经网络	近期多数 <sup>[16-18]</sup>

因此，本文在姿态生成模型中统一使用 Rot6d 表示每个关节的旋转状态，以确保训练阶段的平滑收敛与推理阶段的稳定性。

## 2.2 面部表情的定义与参数化表示

面部表情是非语言交流的重要组成部分，与语音、手势共同传递情感和态度信息。与身体动作不同，面部表情主要由皮肤形变和局部运动构成，无法通过骨骼旋转直接建模，因此需要专门的参数化表示方式。

目前常见的面部状态描述方法主要有两类：

**BlendShape 权重模型** FACS (Facial Action Coding System)<sup>[19]</sup> 提出复杂表情可分解为若干可组合的基本动作单元 (Action Units, AU)，为表情的参数化表示提供了理论基础。受此启发，BlendShape 模型将面部表情表示为一组可线性叠加的形变基 (morph targets)，每个基形对应一个权重控制的局部表情变化。复杂表情由多个基形的组合生成，其思想与 FACS 的动作单元体系相呼应。

与基于骨骼的形变不同，BlendShape 直接在顶点层面定义几何偏移量，因此对网格拓扑结构高度依赖：每个基形的顶点偏移需与基础网格逐点对应。在具有相同网格结构的角色模型之间，BlendShape 集合可以直接复用；但若拓扑（如顶点数量、索引

或连线关系) 发生变化, 偏移数据将无法一一对应, 从而难以在不同模型间映射或重定向。这种拓扑依赖性限制了其跨模型的通用性。

尽管如此, BlendShape 在表现精细面部表情和软组织形变方面具有显著优势。其标准化权重接口、实时可驱动性与渲染兼容性, 使其成为虚拟人和表情捕捉系统的主流表示形式, 并被广泛应用于如 ARKit<sup>[20]</sup> 等实时动画框架, 以及主流渲染引擎中。

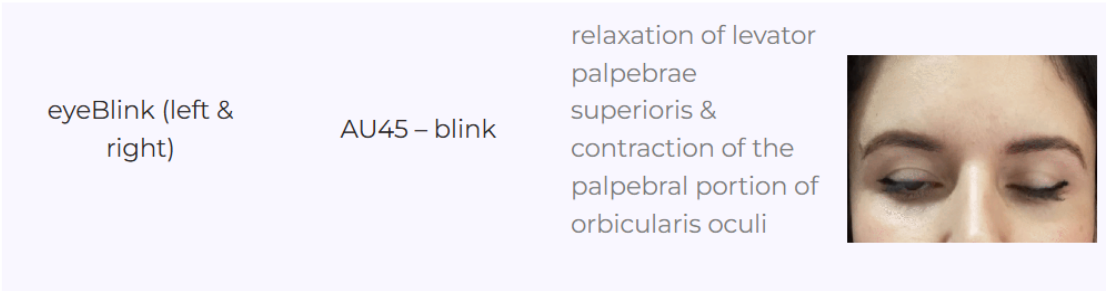


图 2.1 示例图来自<sup>[21]</sup>, 展示 FACS AU45 (blink) 对应的 ARKit 中的两种 BlendShape 基形: eyeBlinkLeft (闭左眼) 与 eyeBlinkRight (闭右眼)。

图 2.2 展示了 eyeBlinkLeft、eyeBlinkRight 的 ARKit BlendShape 基形在权重  $w \in [0, 1]$  下的线性插值效果 (从张眼到闭眼)。



图 2.2 BlendShape 线性插值效果示例: eyeBlink 从  $w=0$  (左) 到  $w=1$  (右) 的连续变化, 中图为中间值。BS 权重可直接用于实时渲染驱动。

**关键点坐标模型 (Landmark-based Representation)** 关键点模型通过检测面部若干语义特征点 (2D 或 3D 坐标) 来描述几何结构变化。典型实现包括 MediaPipe Face Mesh<sup>[1]</sup> 与 OpenFace 系列。

与基于网格形变的 BlendShape 不同，关键点模型并不依附于任何具体网格拓扑，而是在几何空间中以语义一致的特征点集合形式定义面部结构。这种表示方式并不描述模型的形变，而是对“人脸几何”的抽象建模，因此常用于表情识别、头部姿态估计等分析任务，而较少用于驱动渲染。

在多模态学习与特征分析中，BlendShape 表示具有较高的统一性和可量化性，适合以固定维度向量作为模型输入，并易于应用于不同虚拟角色的动画驱动。因此，本文在系统设计中采用与 ARKit 兼容的 BlendShape 参数作为面部模态输入特征。

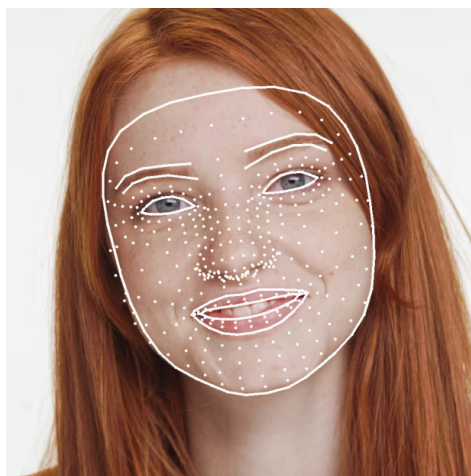


图 2.3 MediaPipe Face Mesh 关键点结构示意图，截取自<sup>[1]</sup>。

## 2.3 国内外研究现状

### 2.3.1 手势生成的研究目标

#### 2.3.1.1 研究目标类型的差异

语音驱动的手势生成研究在总体目标上虽一致——即让虚拟角色的动作与语音内容、节奏协调一致——但在使用场景与系统角色上存在显著差异。

现有研究大体可分为两类：

**为 AI 的虚拟形象生成手势** 这一类研究的目标是让 AI 驱动文本对话系统的虚拟形象具备手势表现力。

模型可以一次性生成下一句语音或文本，因此可以访问完整的未来信息，包括整句音频、文本和语义上下文。

典型方法通过编码完整句子的节奏与语义，预测整段动作轨迹，以最大化动作与语义的一致性和整体流畅性。

这类方法适合 AI 驱动的系统或离线生成的应用场景，如合成视频。

**为用户的虚拟人生成实时手势** 本文所聚焦的目标类型属于第二类。

在用户实时说话的过程中，系统需根据当前语音流（以及可选的面部表情与头部姿态）即时生成同步手势。

此任务具有严格的实时性约束与因果性限制：模型在每一时刻只能使用当前及过去的信息，而不能访问未来语音或文本内容。

因此，常见的整句式规划或滑动预测方法不再适用。

该任务更接近实时交互系统，而非内容生成系统。

2.3.1.2 任务约束与可利用条件的差异

这两类研究目标在可利用的信息条件和评价重心上存在本质区别：

表 2.2 两类手势生成任务在约束与可利用条件上的对比

对比维度	AI 虚拟形象生成	用户虚拟人实时生成
输入信息	完整句级语音或文本（可使用未来信息）	实时语音流，仅使用过去与当前帧
输出目标	整句手势序列（离线生成）	连续流式手势（逐帧生成）
时间约束	不必要实时	帧级实时性（<50ms 推理延迟）
评价重点	整体语义一致性与美学自然度	瞬时同步性、动作平滑与交互稳定性
应用场景	离线动画、内容合成、AI 虚拟直播	实时虚拟人、视频会议、用户虚拟直播
语音模态	作为输入或由文本生成（TTS 输出）	作为实时输入特征（语音流）
手部手势	生成目标（输出）	生成目标（输出）
面部表情	通常为生成目标（输出）	可通过设备实时采集，作为输入辅助推理
头部姿态	通常为生成目标（输出）	可实时采集并作为输入特征，用于同步推理

前者可以在生成阶段规划动作节奏与语义对应，而后者需在无未来信息的条件下维持自然与同步，并保证输出连续、平滑且无突跳，从而对模型结构、输入模态与延迟控制提出更高要求。

### 2.3.1.3 评估方法

语音驱动手势生成研究的评价体系通常涵盖生成质量、时序匹配与表达多样性等多个维度。研究者既关注生成动作的自然性与视觉流畅度，也重视其与语音信号在节奏和语义层面的对应关系。近年来，随着实时生成和多模态扩展任务的发展，相关评估方法也逐渐体系化，可概括为以下几类：

1. **自然性 (Naturalness)** —— 衡量生成动作在运动平滑性、速度变化及能量分布上的合理性，常采用 FGD (Fréchet Gesture Distance)<sup>[22]</sup>、运动速度统计、或主观“自然度”评分等指标。
2. **同步性 (Synchronization)** —— 评估手势在时间上与语音重音或韵律事件的对齐程度，常用 BA (Beat Alignment)<sup>[23]</sup>、DTW (Dynamic Time Warping) 等方法，以及基于重读检测的主观同步性评价。
3. **多样性 (Diversity)** —— 衡量模型在不同语音输入下生成的动作变化程度，通常以轨迹分布的方差、速度曲线差异或 L1DIV 等指标度量，以防止模型陷入单一模式或过度平滑。
4. **语义相关性 (Semantic Relevance)** —— 反映生成动作与语义关键词或情绪类别的一致性，可通过 SRGR (Semantic Relevance to Gesture Ratio)<sup>[3]</sup> 等指标或人工标注语义标签对齐评估。
5. **实时性与稳定性 (Latency & Robustness)** —— 在面向交互系统的研究中，还需评估帧级推理延迟与输出平滑性，以确保动作流连续且系统响应及时。

总体而言，现有评价体系既包含客观的运动学与统计指标，也结合主观感知评分，在不同任务目标下可形成“自然性—同步性—多样性—相关性”的综合评估框架。

### 2.3.2 手势生成的演变

近年来，语音驱动手势生成经历了从规则设计到数据驱动模型、再到多模态扩展与实时生成的持续演变。这一过程不仅体现了算法架构的更新，也反映了研究目标与应用场景的变化：从基于语言规则的行为映射，到学习语音—动作关系的深度生成模型，再到面向交互的多模态实时系统。

#### 2.3.2.1 规则驱动阶段

早期的手势生成系统主要依赖语言学规则与专家知识构建<sup>[6,24-26]</sup>。这类方法通过语义分类或韵律规则将语音片段映射为预定义的手势模板（如指示、肯定、节奏性动



作),并以有限的动作库组合出手势序列。典型代表如 BEAT toolkit 与 Robot Behavior Toolkit,它们可在虚拟代理或机器人中实现基于语音的同步动作。然而,手势词典与语法规则的人工设计成本较高,难以覆盖自然语音中的多样变化,导致生成结果缺乏自然性与个体差异。

### 2.3.2.2 数据驱动阶段

随着大规模语音与动作配对数据的出现,研究者开始采用统计学习和深度神经网络模型学习语音—手势映射关系。在此阶段,语音通常作为唯一输入模态,模型通过 LSTM、GRU 或 MLP 等结构预测连续手势序列。典型代表如 CaMN 模型<sup>[3]</sup>,其基于 BEAT 数据集<sup>[3]</sup>训练级联网络,将 LSTM、全连接网络与 GAN 结构相结合,实现从语音到动作的端到端预测。

然而,该类模型多使用欧拉角或离散旋转参数作为手势表示,生成结果容易出现抖动与不连续。后续工作引入更平滑的表示方式,如 Rot6d<sup>[13,16-18]</sup>或 Axis-Angle<sup>[14]</sup>,显著提升了动作流畅性。与此同时,为解决语音与手势间的多对多映射问题,研究者引入了 VQ-VAE<sup>[16,27]</sup>与扩散模型<sup>[14,28-31]</sup>,在保持自然性的同时提升了生成多样性与表现力。

尽管这些方法在客观指标与视觉效果上均优于传统模型,但它们普遍假设可访问完整语音或文本上下文,属于“整句式(non-streaming)”生成,推理延迟较高,不适用于实时应用。即便是推理效率较高的模型(如 CaMN、DiffSHEG),也因上下文缓冲机制而引入明显延迟。

### 2.3.2.3 多模态扩展阶段

为进一步提升动作表现力与语音理解能力,部分研究引入视觉模态或语言语义特征。例如,CaMN<sup>[3]</sup>在语音输入的基础上融合面部捕捉信息以增强表现;EMAGE<sup>[16]</sup>与 DiffSHEG<sup>[14]</sup>同时生成手势与面部动作;DiffTED<sup>[30]</sup>实现了端到端的视频合成。

这些多模态生成方法在提升虚拟智能体的自然感与沉浸感方面表现优异,但其任务假设仍基于整句输入,因此主要用于 AI 虚拟形象生成或离线内容创作场景,而非实时用户交互。

## 2.3.3 当代生成研究的策略趋势

近年来的研究主要呈现以下两类技术趋势:

1. **扩散模型与高保真生成(Diffusion-based generation):**近年来扩散模型在手势生成任务中表现出卓越的动作自然度与多样性<sup>[14,28-30,32]</sup>。这类模型通常以完整

语音或文本片段为条件，在多阶段去噪过程中逐步生成高保真手势序列。然而，其生成过程依赖未来信息与整句上下文，因此主要用于离线生成或虚拟形象合成。本文的研究目标与之互补，聚焦于在实时约束下实现低延迟、因果一致的动作生成。

2. 语义增强方向 (Semantic-aware generation): 部分研究尝试通过语义或文本特征扩展生成范围，以覆盖 *iconic* 或 *metaphoric* 手势。例如，Yoon 等 (2020)<sup>[33]</sup> 将语音与文本嵌入结合，Alexanderson 等 (2023)<sup>[32]</sup> 引入上下文风格控制，实现了语义相关的动作变化。这一方向旨在增强生成结果的语义一致性与表现力，与本文关注的实时性问题互为补充：前者通过丰富输入语义扩展表达范围，而本文则探索在严格因果条件下通过多模态信号增强空间与时间表达。

## 2.4 实时生成的理论基础与可行性分析

从生成可行性的角度，现有研究普遍认为 *beat* 手势可在无语义理解的条件下由语音韵律直接驱动生成。大量语音驱动手势研究证实，仅凭语音的能量、时长与音高变化即可合成自然的节奏性上肢动作<sup>[22,34-35]</sup>。这些研究所生成的动作在时间结构上与语音重音同步，体现了语音与手势共享的时间规划机制。

相比之下，*iconic*（形象性）、*metaphoric*（隐喻性）与 *deictic*（指向性）手势均依赖语义或指向关系，需要从上下文分析语义与语境，难以在严格实时的因果条件下生成。Kucherenko 等<sup>[23]</sup> 的可预测性分析进一步验证了这一点：他们发现手势的语义类别和空间指向性在语音特征中几乎不可预测，即便结合文本特征，预测性能也相当有限，而节奏阶段（*phase*）相关特征在音频中则具有显著更高的可预测性。这表明，在缺乏未来语义与全局上下文的实时场景中，仅凭语音模态，模型只能稳定生成节奏层面的动作。

为突破这一限制，本文引入头部姿态模态作为补充输入信号。头部姿态能在实时因果条件下提供部分空间与时间线索：其转头与注视方向反映互动焦点，点头与抬头与语音重读共现，能够在不依赖未来语义信息的前提下，为手势生成提供弱先验约束。这种模态扩展为实时系统提供了理论上的可行性基础，使模型能够在语音之外获得关于节奏、方向与视角的附加信息。

**头部姿态对手势预测的贡献** 头部动作在自然语音中常呈现出一定的时间前瞻性<sup>[8]</sup>：其启动往往早于对应韵律词的发声，这意味着视觉模态可能比声学信号更早反映语

音节奏的变化趋势。这种时序特性为实时生成任务提供了潜在的预测窗口，使系统能够在语音节奏变化尚未显现前，就提前捕获相关的动态线索。因此，头部姿态在实时生成中不仅提供同步参考，也可能在时间上形成前驱信号，为手势节奏的自然启动提供时序优势。

**头部姿态对空间锚定与视角一致性的贡献** 头部姿态模态为实时语音驱动的手势生成提供了关键的空间参照信号。其与语音韵律在时间组织上高度耦合。即使在无未来语义信息的条件下，头部的转向与注视变化仍能反映说话者的注意焦点与叙述方向，从而帮助模型在动作生成中保持空间的连贯性与方向一致性。这一机制使系统能够在时间与空间两个维度上同步对齐语音与动作，让生成的手势在视觉上更具互动感与表达意图。

在 McNeill 的四类手势体系中，头部姿态的引入主要强化了两类动作的生成：(1) 对 *beat* 手势而言，它为语音重读和节奏段落提供显式的时间协同信号，使手部与头部动作在韵律层面更加一致；(2) 对 *iconic* 手势而言，它在具有路径与方向特征的动作中提供空间参考，使模型能够在叙事空间中更稳定地确定动作的方位与轨迹方向。通过这两方面的强化，系统在保持实时性的同时获得了更自然的节奏衔接与空间表达。

与此同时，本文亦明确头部姿态模态的作用边界：其核心优势在于捕捉方向、焦点与时序节奏，而非手型语义或复杂形态描摹等细粒度语义特征。换言之，它主要改善手势的位置、方向与视角依附，而非手势的形状描绘或语义内容。对于依赖抽象语义或外指参照的 *metaphoric* 与 *deictic* 手势，仍需语言或上下文模态的补充。

总体而言，头部姿态为实时生成提供了介于韵律与语义之间的关键中层约束。其时间上的前瞻性与空间上的指向性共同帮助模型在低延迟条件下保持自然、连贯且空间协调的动作表现，从而在因果生成框架内有效拓展了语音驱动手势的可表达范围，并为节奏主导型 (*beat-like*) 动作的实时生成提供了结构的支持。

## 2.5 本文的工作与创新点

本文研究的目的是设计一种能够在实时条件下运行的语音驱动手势生成模型，使用户无需动作捕捉设备或特定硬件，仅通过语音输入即可驱动虚拟人的上肢与头部动作。与以往主要面向离线生成或 AI 虚拟形象生成的研究不同，本文关注的任务场景是“用户实时交互”，因此系统必须在严格因果的条件下运行，即只能利用当前与过去的输入帧信息，无法依赖未来语音或文本内容进行整体规划。

现有的高精度手势生成模型在离线场景中表现优异，特别是基于扩散模型或 VQ-VAE 的方法<sup>[14,16,28,30]</sup>，能够生成自然、连贯且语义相关的手势序列。然而，这些模型通常需要整句语音或文本作为输入，推理过程依赖未来信息以分析语义结构与节奏特征，因此难以直接应用于实时交互系统。在无未来信息约束下，模型必须在信息不完整的情况下进行预测，这会显著影响动作生成的自然性与语义一致性。

为缓解上述问题，本文提出利用用户的面部表情与头部姿态作为辅助输入模态，为手势生成过程提供额外的非语言信号。面部表情能反映说话者的情绪与语气变化，在语义模糊的情况下有助于生成更具情感表达的动作；头部姿态则能反映注意方向与交互焦点，在语音节奏变化时为手势生成提供时序上的参考。通过将这两种模态与语音信号联合输入模型，系统能够在实时条件下获得更多上下文线索，从而在保持低延迟的同时提升手势生成的自然度与一致性。

本文以 CaMN 模型<sup>[3]</sup>为基础进行扩展。CaMN 原为离线级联结构，输入包括语音与面部捕捉特征，输出包含手部与头部姿态。本文首先将其输入机制改写为逐帧输入流形式，并引入新的头部特征分析模块，将头部姿态信号作为独立通道输入至级联网络的末端层，以强化模型的时序响应能力。在这一结构下，系统可在实时语音流输入条件下逐帧生成动作输出，实现语音、面部与头部信号的联合驱动。

实验结果表明，本文提出的模型在实时性与动作自然性之间取得良好平衡。在典型的桌面端环境下，单帧推理时间约为 1 毫秒之内，能够满足实时交互需求；同时在 FGD、BA 与主观评测中表现出与离线模型相近的手势自然度与同步性。由此，本文的方法证明了在严格因果条件下，通过多模态输入融合可以有效提升实时手势生成的表现力与稳定性。

## 2.6 本章小结

本章综述了语音驱动手势生成领域的相关研究现状与发展脉络。首先，对手势的概念与在计算机中的参数化表示进行了阐述，说明了手势与面部表情、头部姿态在虚拟人交互中的角色与差异。随后，从研究目标的角度分析了不同任务设定之间的区别，指出现有大多数工作聚焦于为 AI 虚拟形象生成整句级动作，而缺乏面向用户实时交互的研究。在此基础上，回顾了手势生成方法从规则驱动到数据驱动、从单模态到多模态的演变过程，总结了现有模型虽在生成质量上取得显著进展，但在实时性与因果性方面仍存在局限。

最后，结合本文的研究目标，提出了面向实时交互的语音驱动手势生成方案。

## 第3章 方法

### 3.1 研究定位与总体设计思路

手势可根据语义依赖性与时间结构复杂度区分为可语义生成与可韵律生成两大类。本文的研究聚焦于严格实时的语音驱动任务，在此条件下模型无法访问未来语音或完整语义，因此重点生成与语音韵律同步的 *beat-like* 手势及头部动作，并通过面部与头部模态的联合输入进一步增强表达性与自然度。这种选择既符合认知语言学的节奏—动作一致性原则，也符合实时系统的因果约束。

基于上述定位，本文提出了一种基于音频、面部 BlendShape 权重和头部姿态输入的帧级多模态级联手势生成模型 *FaceCapGes*。模型旨在在实时条件下实现自然、同步且具有一定指向性的上半身动作生成，在不依赖语义理解或未来上下文的前提下，通过多模态输入弥补语音模态预测能力的不足。

为避免与系统级说明混淆，本文模型在第 3.2 节首先给出端到端系统定位；随后在第 3.3 节形式化定义任务与符号体系；在第 3.4 节详细阐述模型的级联结构与输入模态设计；最后在第 3.5 节说明训练过程与优化方法；并在第 3.6 节介绍实现细节与训练配置。

### 3.2 系统整体框架与模块定位

本节介绍整个系统的端到端驱动流程及模块职责划分。如图 3.1 所示，系统整体架构由五个层级组成：用户配置层、设备层、中间件层、手势生成模型层以及渲染与驱动层。各层之间通过多模态信号接口进行连接，实现从信号采集到虚拟人动作生成的端到端实时处理。

*FaceCapGes* 模型位于中间层，承担多模态输入到上半身姿态输出的核心推理任务，而输入采集与渲染模块分别负责信号获取与结果展示。

为实现基于语音、面部捕捉与头部姿态的实时数字人驱动系统，本文构建了完整的信号采集、动作生成与渲染展示的处理管线。*FaceCapGes* 模型作为该系统的核心计算模块，负责在实时约束下从多模态输入推理出当前帧的上半身骨骼姿态。

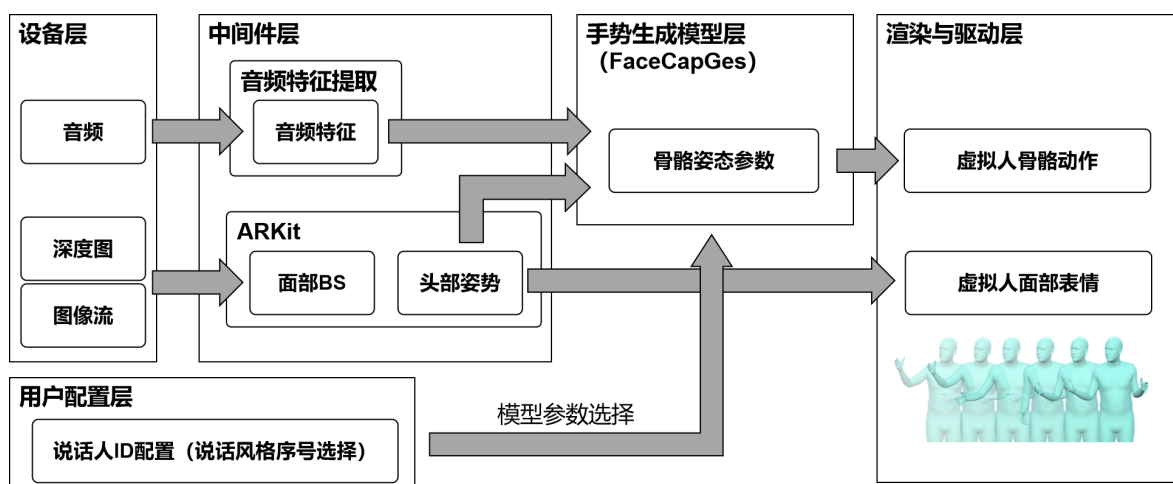


图 3.1 系统整体架构与数据流示意图

### 3.2.1 信号采集与系统配置层

该层位于系统整体架构的输入端，用于从用户端设备实时获取多模态信号，并在系统初始化阶段完成运行参数的配置。整体结构可划分为设备层、中间件层与用户配置层三个部分，如图 3.1 所示。

**设备层** 设备层负责采集语音与视觉模态信号。语音信号由麦克风实时录制，采样率与帧移可根据运行设备性能调整；视觉信号由前置 TrueDepth 摄像头获取面部深度图与视频流，并作为 ARKit 面部追踪模块的输入。

**中间件层** 中间件层通过 Apple 提供的 ARKit 框架，将设备层的原始图像流与深度图转化为结构化特征。ARKit 输出两类主要数据：(1) **面部表情特征** ARKit 提供 52 维 BlendShape 系数向量，用于描述关键肌肉群的局部形变状态。该特征能够反映用户的表情、口型与情感变化，并以帧级形式同步输出。(2) **头部姿态特征** ARKit 在 ARFaceAnchor 对象中输出一个  $4 \times 4$  齐次变换矩阵  $\mathbf{T} = [\mathbf{R}|\mathbf{t}]$ ，其中左上角的  $\mathbf{R} \in \mathbb{R}^{3 \times 3}$  为旋转矩阵，右上角的  $\mathbf{t} \in \mathbb{R}^{3 \times 1}$  为平移向量。本研究从矩阵中提取旋转部分，并将其转换为 Rot6D<sup>[13]</sup> 表示形式，以提升旋转空间的连续性与模型训练的稳定性。

同时，音频流在中间件层中被传入特征提取模块以生成时间序列特征。模型训练阶段使用 Librosa 库离线提取 Mel 频谱、短时能量 (RMS) 与基频  $F_0$  等声学特征，以保证特征精度与一致性。系统运行阶段可由等价的实时特征提取模块（如 torchaudio 或 TensorFlow Audio）逐帧生成对应特征，以实现端到端的低延迟运行。

**用户配置层** 用户配置层负责系统初始化阶段的模型与参数设定。用户可在应用中选择说话风格，对应加载不同说话人 ID 配置下的模型权重。该配置仅在系统启动时生效，不参与实时推理过程。

本层提供的多模态信号经中间件处理后，以统一的数据接口传递至手势生成模型，实现语音、表情与头部姿态的实时融合输入。

### 3.2.2 手势生成模型层 (FaceCapGes)

FaceCapGes 模块位于系统的中间层，是本文提出的核心计算单元。该模块接收来自信号采集与系统配置层的三类输入特征：语音特征、面部 BlendShape 系数以及头部姿态参数，并在不依赖未来帧的条件下，逐帧预测用户当前时刻的上半身骨骼姿态。

生成的骨骼姿态采用 Rot6D 连续旋转表示形式，覆盖上半身 47 个关节的旋转参数。模型内部通过级联多模态编码结构提取时序相关特征，并利用单向 LSTM 解码器完成时间依赖建模，从而在保持实时性的同时，生成与语音节奏、表情变化及头部朝向高度一致的自然手势。

FaceCapGes 输出的姿态数据通过统一接口传递至渲染与驱动模块，与实时面部捕捉信号共同驱动虚拟角色的整体动作。由于模型仅依赖当前与历史帧输入，可与输入层以固定帧率并行运行，实现端到端的低延迟推理。

### 3.2.3 渲染驱动层

该模块位于系统输出端，负责将手势生成模型与面部捕捉结果共同转化为虚拟人的实时动作表现。系统将 FaceCapGes 模型输出的上半身骨骼姿态与 ARKit 实时生成的 52 维面部 BlendShape 系数传递至渲染引擎，由引擎内编写的脚本模块解析并映射至目标虚拟人的骨骼与表情控制接口，从而实现多模态动作驱动。

渲染模块采用基于 GPU 的蒙皮计算与实时光照模型，以确保动画的平滑性和视觉一致性。最终，系统能够在实时流式输入条件下稳定运行，同步呈现语音、表情与身体动作，以自然流畅的数字人形象实现从多模态信号输入到可视化输出的完整驱动流程。

## 3.3 问题定义

在整体系统中，FaceCapGes 模块承担着从多模态输入信号到上半身骨骼姿态预测的核心任务。为了明确模型的输入输出结构与学习目标，本节对该问题进行形式化

定义。

### 3.3.1 任务描述

目标是在实时条件下，根据用户当前时刻的语音、面部表情与头部姿态信息，预测其对应的上半身骨骼姿态。模型需能够逐帧生成与语音节奏、面部动态和头部转动方向相协调的自然手势动作，而不依赖未来的输入帧或整句语音信息。

形式上，可以将该任务定义为一个多模态时序映射函数：

$$\hat{\mathbf{v}}_t^B = f_\theta(\mathbf{v}_{t-N:t}^A, \mathbf{v}_{t-N:t}^F, \mathbf{v}_{t-N:t}^H), \quad (3.1)$$

其中  $f_\theta$  表示由参数  $\theta$  控制的预测模型， $N$  为历史窗口长度。各模态输入定义如下： $\mathbf{v}_t^A$  表示语音模态在时刻  $t$  的特征向量，由麦克风信号经特征提取模块得到； $\mathbf{v}_t^F$  表示面部模态的输入，为 ARKit 输出的 52 维标准化 BlendShape 系数； $\mathbf{v}_t^H$  表示头部模态的输入，为 ARKit 得出的头部旋转矩阵经 Rot6D 表示；而  $\hat{\mathbf{v}}_t^B$  为模型在当前时刻预测的上半身骨骼姿态向量。模型仅利用当前及过去  $N$  帧的输入信息估计  $\hat{\mathbf{v}}_t^B$ ，从而满足严格的实时推理约束。

### 3.3.2 输入与输出模态

FaceCapGes 模型的输入由三种可同时实时获取的模态组成：语音特征、面部 BlendShape 权重及头部姿态参数；输出为当前帧的上半身骨骼旋转状态。各模态的符号与维度如表 3.1 所示。

表 3.1 输入输出模态符号与维度

模态	符号	维度	描述
语音特征	$\mathbf{v}_t^A$	$\mathbb{R}^{1067}$	由音频信号提取的时序特征（Mel 频谱、能量、基频等）
面部 BlendShape	$\mathbf{v}_t^F$	$\mathbb{R}^{52}$	ARKit 输出的标准化表情权重向量
头部姿态	$\mathbf{v}_t^H$	$\mathbb{R}^6$	采用 Rot6d 表示的头部旋转参数
骨骼姿态（输出）	$\hat{\mathbf{v}}_t^B$	$\mathbb{R}^{6 \times 47}$	上半身 47 个关节的旋转状态

输入序列  $(\mathbf{v}_{t-N:t}^A, \mathbf{v}_{t-N:t}^F, \mathbf{v}_{t-N:t}^H)$  描述了用户在过去  $N$  帧内的语音与表情动态信息。模型通过学习其时序变化规律，逐帧生成对应的骨骼姿态输出  $\hat{\mathbf{v}}_t^B$ 。在推理阶段，模型仅访问至时刻  $t$  的输入序列，无法访问任何未来帧信息，保证了生成过程的因果性与实时性。



### 3.3.3 学习目标与优化形式

在训练阶段，给定来自多模态语音动作数据集（如 BEAT）的配对样本：

$$(\mathbf{v}_t^A, \mathbf{v}_t^F, \mathbf{v}_t^H, \mathbf{v}_t^B) \quad (3.2)$$

模型的学习目标是 minimized 预测姿态  $\hat{\mathbf{v}}_t^B$  与真实姿态  $\mathbf{v}_t^B$  之间的差异。综合考虑空间重构误差与时序平滑性约束，总体优化目标可表示为：

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \lambda_v \mathcal{L}_{vel} + \lambda_a \mathcal{L}_{acc} + \lambda_{adv} \mathcal{L}_{adv} \quad (3.3)$$

其中：

- $\mathcal{L}_{rec}$  为姿态重构损失，衡量单帧旋转角度误差；
- $\mathcal{L}_{vel}$ 、 $\mathcal{L}_{acc}$  分别约束预测序列的速度与加速度连续性；
- $\mathcal{L}_{adv}$  为对抗损失，用于提升生成手势的自然性；
- $\lambda_v, \lambda_a, \lambda_{adv}$  为对应的权重系数。

通过最小化上述综合损失，模型能够在不依赖未来帧的条件下生成自然、流畅且与语音节奏相匹配的上半身动作序列。

## 3.4 级联架构与输入模态设计

### 3.4.1 级联架构的原理与理论背景

现有语音驱动手势生成模型多采用多模态融合结构，其中以 CaMN<sup>[3]</sup> 为代表的级联架构在设计理念上具有代表性。其核心思想是将语音、面部表情与身体动作视为语义表达的不同层级：语音模态承担语义与节奏驱动作用，面部模态反映情感与意图，身体动作则是语言与情绪的外化呈现。CaMN 采用自上而下的处理顺序，即依次对语音、面部和动作模态进行建模，从而以层次化结构保持模态间的语义依存关系。

这种设计符合人类交流中“语言、表情、动作”一体化的认知规律<sup>[4-5]</sup>。语音先规划语义与节奏，面部表情作为情绪强化信号随后产生，最终通过身体动作完成完整的非语言表达。模型中，语音编码器输出的时间嵌入被输入至面部编码器，再与面部特征融合后驱动动作解码器，从而保持语义一致性并增强表现力。

然而，CaMN 的原始设计面向离线整句生成任务，需要访问未来上下文以维持全局连贯性。在实时场景下，这种依赖将引入显著延迟并破坏因果性。FaceCapGes 在继承其层次思想的同时，对输入模式、训练方式与模态选择进行了系统性重构，以满足帧级实时约束。

### 3.4.2 基线模态继承与实时适配

FaceCapGes 保留了 CaMN 的语音与面部模态结构，但针对实时生成任务进行了适配性改写。

#### 3.4.2.1 说话人 ID 分支移除

如图 3.1 所示，用户配置层会设置说话人 ID 配置用于模型切换，但该模态在本模型中不属于网络输入。在基线模型 CaMN 中，输入模态包含显式的说话人 ID 向量，用于在同一模型内区分不同演讲者的风格差异。然而在实时交互场景下，该分支并非必要：用户身份通常固定，且说话风格的变化频率远低于帧级推理速度。因此，FaceCapGes 移除了 ID 输入分支，并采用“单说话人训练”策略，即针对每个说话人独立训练模型参数。实验表明，该方式能在保持收敛稳定的同时显著提升动作的自然性与节奏一致性。从系统使用角度看，不同模型可视为“说话风格配置文件”，用户仅在需要时切换对应参数，该操作发生频率低，不会影响实时推理性能。

#### 3.4.2.2 输入模态继承

语音特征通过时间卷积网络（Temporal Convolutional Network, TCN）和多层感知机（MLP）编码，以捕捉短时节奏模式；面部模态采用相似结构，并在中间层融合语音嵌入，从而增强语音与表情之间的语义关联。语音编码器  $E_A$  与面部编码器  $E_F$  的输出定义为：

$$\mathbf{z}_t^A = E_A(\mathbf{v}_{t-N:t}^A), \quad \mathbf{z}_t^F = E_F(\mathbf{v}_{t-N:t}^F; \mathbf{z}_t^A) \quad (3.4)$$

其中  $\mathbf{z}_t^A \in \mathbb{R}^{128}$ ,  $\mathbf{z}_t^F \in \mathbb{R}^{32}$ 。这两个编码器负责提取低层次语音节奏与表情动态信息，为后续模态融合提供稳定上下文表征。

此外，系统在此基础上引入头部姿态模态  $\mathbf{v}_t^H$ ，用于补充空间方向与节奏信号。其编码器  $E_H$  将 Rot6D 表示的头部旋转向量映射为紧凑潜在表征：

$$\mathbf{z}_t^H = E_H(\mathbf{v}_{t-N:t}^H), \quad (3.5)$$

编码器结构将在第 3.4.3 节详细说明。

#### 3.4.2.3 输出模态继承（身体姿态解码）

在输入模态经过编码与融合后，模型需将多模态特征映射至对应的身体姿态空间。为实现层次化的动作生成与结构协调，本文将上半身的输出区域划分为两个互补分支：躯干 (Torso, T) 与上肢 (Upper limbs, U)。躯干部分包含脊椎的三个主要控制关节，用于确定身体的姿态基准与运动节奏；上肢部分包含双臂及手部关节，负责生成

与语音节奏及情绪表达相呼应的细节动作。最终的上半身姿态表示为两者的组合：

$$\mathbf{v}^B = \mathbf{v}^T \otimes \mathbf{v}^U, \quad (3.6)$$

其中  $\otimes$  表示通道维度拼接操作。

该分层设计继承了 CaMN 的层次预测思路：模型首先生成相对稳定的躯干姿态以确定整体方向，再以此为条件预测上肢动作，从而在实时生成中保持整体协调性与自然度。

具体而言，来自语音、面部与头部编码器的特征  $\mathbf{z}_i^A$ 、 $\mathbf{z}_i^F$ 、 $\mathbf{z}_i^H$  会与历史姿态序列  $(\mathbf{v}_{i-N}^B, \dots, \mathbf{v}_i^B)$  拼接，组成多模态隐向量：

$$\mathbf{z}_i^M = \mathbf{z}_i^A \otimes \mathbf{z}_i^F \otimes \mathbf{z}_i^H \otimes (\mathbf{v}_{i-N}^B, \dots, \mathbf{v}_i^B), \quad (3.7)$$

其中  $\otimes$  表示通道维度拼接操作，时间末帧采用零填充以对齐维度。

随后， $\{\mathbf{z}_0^M, \dots, \mathbf{z}_N^M\}$  经两个单向 LSTM 解码器，分别生成躯干与上肢的潜在特征：

$$\mathbf{z}^T = \text{LSTM}_T(\mathbf{z}_0^M, \dots, \mathbf{z}_N^M), \quad \mathbf{z}^U = \text{LSTM}_U(\mathbf{z}_0^M, \dots, \mathbf{z}_N^M), \quad (3.8)$$

并通过独立的 MLP 模块还原为旋转参数：

$$\hat{\mathbf{v}}^T = \text{MLP}_T(\mathbf{z}^T), \quad \hat{\mathbf{v}}^U = \text{MLP}_U(\mathbf{z}^U). \quad (3.9)$$

最终拼接得到当前帧的完整上半身姿态：

$$\hat{\mathbf{v}}^B = \hat{\mathbf{v}}^T \otimes \hat{\mathbf{v}}^U. \quad (3.10)$$

在推理阶段，解码器隐状态在时间步之间保持连续，与前述输入模态特征配合，使模型在保持因果性的同时具备自然的时间平滑性。由于该部分结构沿用自基线模型，本文不再赘述。接下来，将介绍时间建模结构的改动及其对实时性的适配。

#### 3.4.2.4 时间建模结构改动与因果性约束

基线模型 CaMN 使用双向 LSTM 生成完整序列的骨骼姿态，输入与输出片段长度一致。由于双向结构在每个时间步都依赖未来帧隐状态，虽然能增强整体平滑性，但不满足实时生成场景的因果约束。为实现严格的实时性，本文将时间建模模块改为单向 LSTM，使模型在每一时间步仅依赖过去  $N$  帧的输入并预测当前帧的骨骼姿态。虽然单向 LSTM 结构上仍会输出与输入片段等长的时间序列，但训练时仅计算其最后一帧的预测误差：

$$\hat{\mathbf{v}}_t = \text{LSTM}(\mathbf{z}_{t-N:t}^M)_N, \quad \mathcal{L}_{\text{causal}} = \|\hat{\mathbf{v}}_t - \mathbf{v}_t\|_2^2, \quad (3.11)$$

其中，下标  $N$  表示取 LSTM 输出序列的最后一帧作为当前时刻的预测结果。

该策略通过在前  $N$  帧内累积隐状态，于第  $N + 1$  帧完成当前姿态预测，从而建立严格的因果时序映射。

在推理阶段，FaceCapGes 采用长度为  $N$  的显式输入窗口，并在时间步之间保留 LSTM 的隐状态。虽然单向 LSTM 理论上能够通过递推隐状态存储历史信息，但由于隐状态为压缩形式，难以完全保留短时节奏与相位特征。因此，显式窗口输入与隐状态记忆在模型中形成互补：前者提供局部的高分辨率上下文，后者维持全局的时序连贯性。这种设计在保证因果性的前提下提高了生成的稳定性与自然性，也是实现实时语音驱动动作生成的关键因素之一。

图 3.2 展示了双向与单向结构的差异：单向结构仅依赖历史帧输入，更适合在流式序列中逐帧输出预测结果。

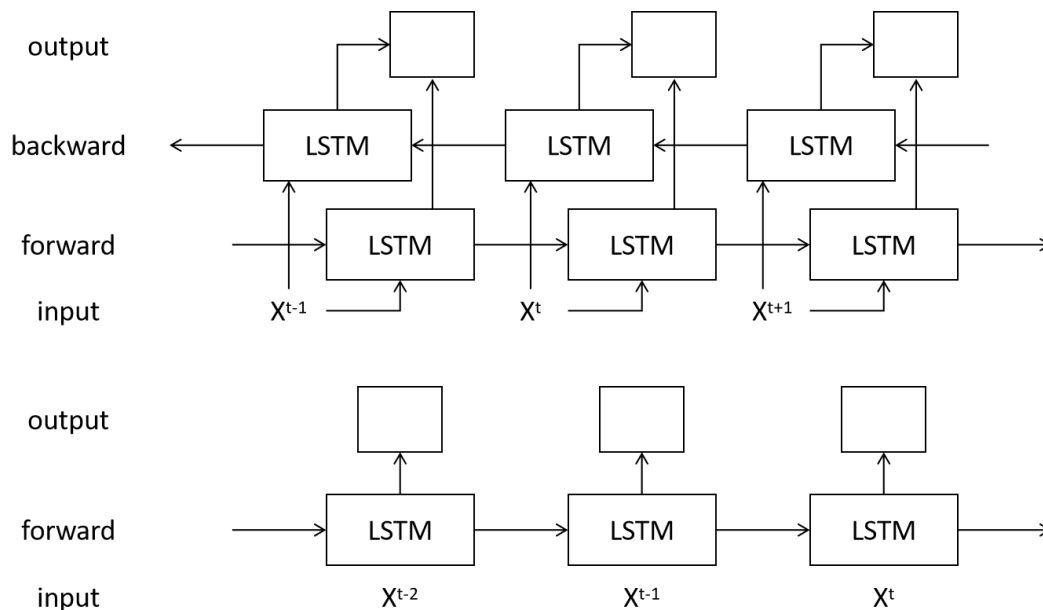


图 3.2 双向与单向 LSTM 对比示意图。双向结构（上）在每个时间步同时利用历史与未来帧特征进行建模；单向结构（下）仅基于历史帧进行递推，以保持因果性并支持流式推理。

### 3.4.2.5 滑动窗口式自回归训练

训练时，每段输入为  $N + M$  帧，前  $N$  帧作为输入上下文，后  $M$  帧逐步预测（见图 3.3）。其中，前  $N$  帧的历史姿态可视为模型的因果历史窗口，其设计思路与基线模型中附加片段首部的“历史缓冲”类似，但在意义上有所不同。CaMN 在双向时间

建模下使用该缓冲以补足片段外部上下文，而 FaceCapGes 则将其重新定义为实时预测所需的前序姿态帧数量，即当前帧推理所依赖的显式时间上下文。

在训练阶段，模型采用纯自回归（pure autoregressive）方式展开，即在每一步预测后，将自身生成的历史帧作为下一步输入，而非采用教师强制（teacher forcing）。这种方式使模型在优化过程中暴露于自身预测的分布，保持训练与推理过程的一致性，避免了教师强制常见的暴露偏差（exposure bias），即推理阶段模型面对自身生成数据时性能下降的问题。在每个窗口内连续预测  $M$  步后，累计所有预测帧的误差并计算平均损失：

$$\mathcal{L}_{\text{total}} = \frac{1}{M} \sum_{i=1}^M \|\hat{\mathbf{g}}_i - \mathbf{g}_i\|_2^2. \quad (3.12)$$

虽然该纯自回归方式在训练早期的收敛速度略慢，但生成稳定性更高，可有效抑制长期序列中的误差积累。此外，这种机制特别契合实时虚拟人等持续交互场景：模型在此类应用中并非针对短语段离线生成，而是与语音流持续同步、长时间运行。在这种场景下，纯自回归训练使模型在遇到自身预测误差时能够动态修正节奏，从而在长时间交互中保持自然的动作连续性与节奏稳定性。

每一预测步中定义一个长度为  $N+1$  的滑动窗口，其中前  $N$  帧为输入，第  $N+1$  帧为预测目标。形式化定义如下：

$$\mathbf{g}_i^H = (\mathbf{g}_{i-N}, \dots, \mathbf{g}_{\min(N, i-1)}) \otimes (\hat{\mathbf{g}}_{\max(N+1, i-N)}, \dots, \hat{\mathbf{g}}_{i-1}), \quad (3.13)$$

$$\hat{\mathbf{g}}_i = \text{FaceCapGes}(\mathbf{v}_{i-N}, \dots, \mathbf{v}_i; \mathbf{g}_i^H), \quad (3.14)$$

其中  $\otimes$  表示时间拼接操作。该机制在每步仅依赖过去信息，从而保持因果性约束；同时通过窗口内的滚动更新，在不引入未来帧的前提下实现平滑过渡。推理阶段模型以单帧为输入流，输出当前时刻的上半身姿态，实现端到端低延迟生成。

需要指出的是，由于滑动窗口机制依赖前  $N$  帧的上下文信息，模型在序列开端无法立即生成动作，即存在一个短暂的“冷启动”阶段。然而在本文的目标应用场景——实时语音驱动虚拟人系统——中，模型作为常驻进程持续运行，而非针对离散语句反复初始化。因此该延迟仅在首次启动时出现约  $N$  帧（0.3–0.5 秒）的等待，对用户体验影响可忽略。这一延迟被视为实时生成框架下的合理权衡。

通过以上适配，FaceCapGes 在保持 CaMN 级联优势的同时显著降低系统延迟，实现实时稳定的语音与面部驱动手势生成。

然而，仅依赖这两种模态仍存在动作方向与节奏响应不足的问题。为此，下一节将在级联结构末端引入头部姿态模态，以补充时序反馈与空间方向信号。

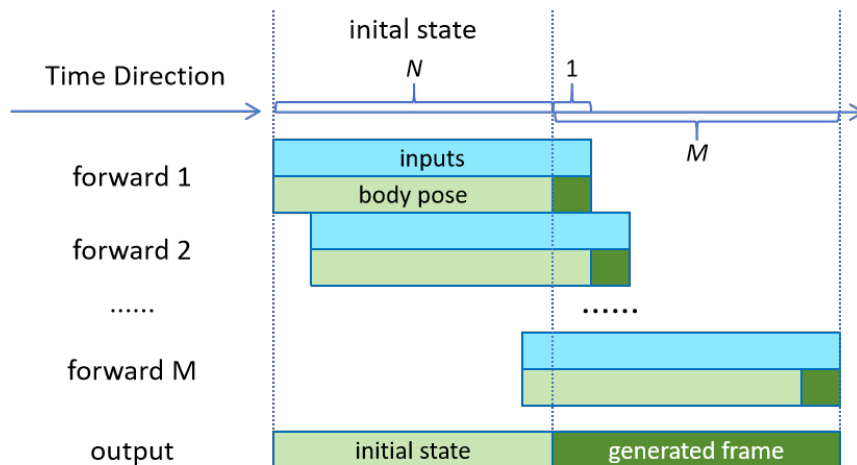


图 3.3 滑动窗口训练机制：模型通过自回归循环预测  $M$  帧，每步使用  $N$  帧上下文并预测第  $N+1$  帧。损失函数累计所有预测帧的误差，保证因果性与时间平滑性。

### 3.4.3 头部姿态模态的引入与结构位置

在模型结构设计中，我们考察了头部姿态特征与其他模态的多种组合方式。具体而言，分别尝试了：(1) 将头部姿态特征在编码阶段与语音或面部特征进行早期融合；(2) 在解码阶段以前两者的嵌入结果为条件，预测头部姿态特征作为辅助信号。实验结果显示，这两种交互方式均未带来显著性能提升，部分设置甚至出现训练收敛速度下降或动作节奏轻微错位的情况。

这一现象与认知层面的规律相符。头部动作虽然与语音韵律在时间上存在同步性，但在认知层面并非由语音或表情直接驱动，也难以反向推导这些模态的动态变化。换言之，三者更可能属于并行协同关系，共享节奏与注意机制，但不构成单向的预测链。

基于此观察，本文在最终架构中采用弱耦合的后级输入设计：头部姿态特征在语音与面部特征编码完成后，以独立通道的形式拼接至多模态隐向量  $\mathbf{z}_t^M$ ，而非在编码阶段进行显式交互。该处理方式在保持整体结构简洁性的同时，仍保留头部姿态在方向、节奏及注意焦点方面的补充作用。

实验表明，在此配置下模型的整体自然度与时序稳定性得到改善，说明头部姿态虽非语音或表情的从属模态，但作为空间与节奏的辅助信号仍具有积极贡献。

**输入特征与表示** 本文仅使用头部旋转信息，不引入平移位移特征。BEAT 数据集中演讲者多为站姿，录制中存在身体移动，若直接使用位移作为输入，噪声易混入；此

外，目标应用场景中的坐姿用户分布不同，直接建模位移会削弱泛化性。因此仅采用旋转特征，并使用连续且可微的 Rot6d 表示，以避免欧拉角与四元数的奇异性问题。

**编码器结构** 图 3.4 所示为头部姿态编码器结构。该编码器由两层前馈网络组成，输入为 Rot6d 表示的 6 维向量：

$$\mathbf{z}_t^H = E_H(\mathbf{H}_{t-N:t}; \mathbf{z}_t^A, \mathbf{z}_t^F), \quad (3.15)$$

其中  $E_H$  的具体形式为：

$$\mathbf{h}_1 = \text{ReLU}(W_1 \mathbf{H}_t + b_1), \quad (3.16)$$

$$\mathbf{z}_t^H = W_2 \mathbf{h}_1 + b_2, \quad (3.17)$$

网络维度设置为：输入 6，中间层 36，输出 12。在特征层面，其输出与语音、面部嵌入拼接后输入解码器，形成从语义到反应的多层信号流。



图 3.4 头部姿态编码器结构示意图。输入为 Rot6d 表示的 6 维旋转向量，经两层前馈网络与 ReLU 非线性映射，输出 12 维紧凑潜在表征。该编码结果与语音、面部嵌入拼接后输入解码器，用于补充动作的方向与节奏信号。

**架构位置与实验依据** 在早期实验中，我们尝试将头部姿态与语音、面部特征早期融合，但该方式导致手势方向轻微抖动、语音与动作节奏错位，收敛速度减慢。分析认为，原因在于头部姿态的反应性与非因果性：若过早参与特征交互，会破坏语音—手势因果映射。因此本文采用后级融合策略，在语音与面部特征编码完成后再输入头部特征。该设计使模型先形成语义骨架，再由头部姿态进行方向修正与节奏调节。

### 3.4.4 小结：从语义驱动到反应调节的信号层级

FaceCapGes 在 CaMN 的语音—表情级联架构基础上，通过实时适配与头部模态引入实现了信号层级的扩展。前两级模态承担语义与情感驱动，而新增头部姿态层作为反应性调节模块，在实时条件下为手势提供动态节奏与空间反馈。

## 3.5 训练与损失函数设计

在前述滑动窗口式自回归训练框架下，每段输入序列包含  $N$  帧上下文与  $M$  帧预测结果。模型在每个窗口中输出连续的动作序列  $\hat{\mathbf{g}} \in \mathbb{R}^{M \times 6 \times 47}$ ，并以重构与对抗两类损失共同约束生成质量。

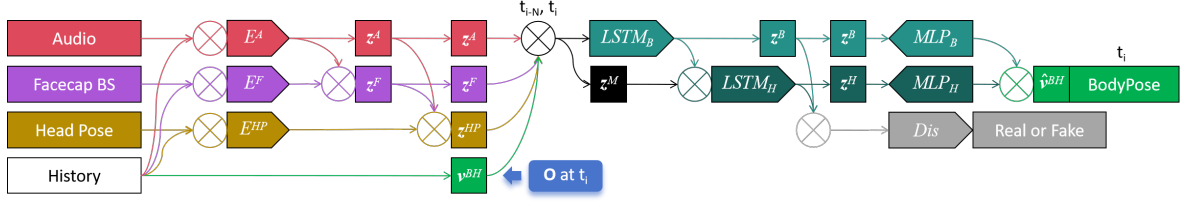


图 3.5 FaceCapGes 模型结构：音频、面部、头部编码器分别提取模态特征后拼接，输入至 LSTM 解码器生成躯干与手部动作，仅保留最后一帧输出作为当前时刻预测，符合帧级实时推理设定。训练阶段历史姿态序列比目标长度少一帧，需进行零填充。

**手势重构损失** 重构项  $\mathcal{L}_{GestureRec}$  由位置、速度与加速度误差组成，用以平衡空间准确性与时间平滑性：

$$\mathbf{g}' = \mathbf{g}_t - \mathbf{g}_{t-1}, \quad \mathbf{g}'' = \mathbf{g}'_t - \mathbf{g}'_{t-1}, \quad (3.18)$$

$$\mathcal{L}_{GestureRec} = \mathcal{L}_{rec}(\mathbf{g}, \hat{\mathbf{g}}) + \mathcal{L}_{vel}(\mathbf{g}', \hat{\mathbf{g}}') + \mathcal{L}_{acc}(\mathbf{g}'', \hat{\mathbf{g}}''), \quad (3.19)$$

其中  $\mathcal{L}_{rec}$  保证空间姿态重构精度， $\mathcal{L}_{vel}$  和  $\mathcal{L}_{acc}$  强调动态平滑性与时序一致性。该组合设计在自回归预测中能有效缓解抖动与速度漂移问题。

**对抗损失** 为进一步提升动作的自然度，引入对抗项  $\mathcal{L}_{Adv}$ ：

$$\mathcal{L}_{Adv} = -\mathbb{E}[\log(Dis(\hat{\mathbf{g}}))], \quad (3.20)$$

其中判别器  $Dis$  以完整的动作序列为输入，判别其是否来自真实数据分布。该项损失约束生成序列的整体动力学分布，促进生成动作在节奏、加速度和能量变化上与真实表演者一致。训练中通过交替优化生成器与判别器的参数，保持两者的平衡。

**总体损失** 综合两项目标，模型的最终训练目标为：

$$\mathcal{L}_{total} = \mathcal{L}_{GestureRec} + \lambda_{adv} \mathcal{L}_{Adv}, \quad (3.21)$$

其中  $\lambda_{adv}$  为对抗损失的权重（实验中设为 0.1）。该设计在保持运动学准确性的同时，提升了时序的自然性与节奏感。

### 3.6 实现与训练配置

在上述训练目标下，FaceCapGes 基于 PyTorch 实现，所有实验在单张 NVIDIA RTX 4090 GPU 上进行。



本文基于 BEAT 数据集<sup>[3]</sup> 进行训练与评估。该数据集包含多模态同步的语音、面部 blendshape 与全身动作信息，以 15 fps 记录多位专业表演者的演讲片段，覆盖多种语义与情绪场景。其标准骨架结构如图 3.6 所示，共包含 47 个关节节点。FaceCapGes 仅预测其中的上半身部分，包括上肢及躯干的三个主要控制点（蓝色区域所示），以聚焦语音驱动手势中的表达性动作。下肢关节保持静态以保证骨架一致性。

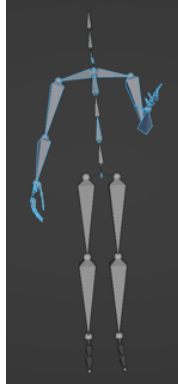


图 3.6 BEAT 数据集的骨架拓扑结构。蓝色部分为 FaceCapGes 模型的控制区域，涵盖上肢与三段脊椎关节，其余节点保持静态。

本文选取表演者 ID 2、4、6、8 的数据进行训练与测试，其中 2、4 为男性，6、8 为女性，确保在性别与说话风格上的分布均衡。训练集与测试集均包含相同的表演者，但使用不同的演讲片段，在预处理阶段已进行严格划分以避免片段交叉。

**训练配置** 训练时输入窗口的前序帧数设为  $N = 16$ ，预测长度为  $M = 34$ ，训练片段的切割步长为 10 帧。相邻片段因此存在部分重叠，从而在保证充分上下文信息的同时提升数据覆盖率与时间连续性。优化器采用 Adam<sup>[36]</sup>，学习率设为  $2 \times 10^{-4}$ ， $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ 。批大小设为 256。为防止早期训练阶段的不稳定，对抗项在第 10 epoch 后引入，整体训练共 374 个 epoch。在损失计算中，前  $N$  帧的历史窗口仅作为因果上下文输入，不参与重构与对抗项的误差回传。损失函数采用第 ?? 节所述的复合重构与对抗目标。

**姿态表示** 所有身体动作均转换为连续可微的 Rot6d 表示，使用 EMAGE<sup>[16]</sup> 中的实现方法，以避免欧拉角奇异性与四元数的符号不确定性。

**运行性能** 在实时推理阶段，FaceCapGes 能以 15 FPS 的速度驱动虚拟角色，满足实时语音交互应用的延迟要求。

### 3.7 本章小结

本章系统介绍了 FaceCapGes 模型的总体设计与关键技术细节。首先，从系统整体出发，阐述了实时语音驱动虚拟人生成管线的总体框架，说明了模型在输入采集、手势生成与渲染驱动中的定位与功能。针对基线模型 CaMN 的结构特征，本文在保持多模态级联优势的基础上，对时间建模与实时适配机制进行了系统性改进：移除了说话人 ID 输入分支，采用独立模型对应不同说话风格；引入单向 LSTM 与滑动窗口式自回归训练以保证因果性与流式生成能力。

在模型的模态设计上，本文重点分析了语音与面部模态的继承机制，并在级联架构末端加入头部姿态编码器，以补充方向性与节奏反馈信号。同时，结合对抗优化与多级重构损失，构建了兼顾空间准确性与时间自然性的训练目标。实验部分将进一步验证这些结构设计对实时性、平滑性与自然度的提升效果。

## 第4章 评估

### 4.1 评估设置

我们使用四种广泛采用的指标对模型进行评估:Fréchet Gesture Distance (FGD)<sup>[37]</sup>、L1 动作多样性 (L1DIV)<sup>[3]</sup>、节奏对齐度 (Beat Alignment, BA)<sup>[3]</sup>，以及语义相关动作召回率 (SRGR)<sup>[3]</sup>。所有评估均在 BEAT 数据集上进行，选用说话人编号为 2、4、6 和 8。生成的身体姿态与真实标签均采用相同骨架拓扑的 BVH 格式。FGD 特征由 EMAGE 模型<sup>[16]</sup>提取，L1DIV、BA 和 SRGR 的实现使用 BEAT 官方提供的代码。所有指标在四位说话人上取平均，以减少个体差异带来的偏差。

### 4.2 客观评估指标与实现细节

为全面评价模型在动作自然性、节奏同步性与多样性等方面的表现，本文在客观指标层面采用四项度量：Fréchet Gesture Distance (FGD)、Speech-Gesture Rhythm Correlation (SRGR)、Beat Alignment (BA) 以及 L1-based Diversity (L1DIV)。这些指标分别对应生成动作在分布一致性、语音同步性与变化丰富性等不同维度，共同构成对模型质量的综合评估体系。

其中，FGD 作为生成分布的核心统计指标，需要训练额外的动作自编码器 (AutoEncoder, AE) 作为特征提取器；其余指标则直接基于生成序列与语音信号的时间对应关系进行计算。本节首先介绍 FGD 的计算原理与评估模型结构，随后依次阐述其它三项指标的定义与计算方法。

#### 4.2.1 Fréchet Gesture Distance (FGD)

FGD<sup>[37]</sup>用于衡量生成手势分布与真实手势分布之间的统计距离，灵感源自图像生成领域的 Fréchet Inception Distance (FID)。不同于图像任务直接利用 Inception 网络特征，在动作生成领域，特征空间需由单独训练的动作自编码器定义。该自编码器通过重构任务学习手势的潜在表示，使潜在空间具备对运动模式的压缩与区分能力。在该潜在空间中，假设真实分布与生成分布的高维嵌入向量分别为  $\mathcal{N}(\mu_r, \Sigma_r)$  与  $\mathcal{N}(\mu_g, \Sigma_g)$ ，则 FGD 定义为：

$$\text{FGD} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}). \quad (4.1)$$

较小的 FGD 值表示生成动作的统计分布更接近真实数据，可反映动作的整体自然度与风格一致性。

**评估模型结构与训练配置.** 本文在每位说话人的训练集上分别训练一组评估用自编码器，以避免跨说话人分布差异对指标的干扰。自编码器输入为以 Rot6d 表示的上半身骨架序列，训练目标为最小化位置、速度与加速度的多尺度重构误差：

$$\mathcal{L}_{AE} = \|\hat{\mathbf{g}} - \mathbf{g}\|_2^2 + \lambda_v \|\hat{\mathbf{g}}' - \mathbf{g}'\|_2^2 + \lambda_a \|\hat{\mathbf{g}}'' - \mathbf{g}''\|_2^2, \quad (4.2)$$

其中  $\lambda_v = 0.1, \lambda_a = 0.1$ ,  $\mathbf{g}'$ 、 $\mathbf{g}''$  分别为速度与加速度序列。

训练配置如下：输入片段长度为 32 帧，批大小 256，隐藏层维度 128，学习率  $1.2 \times 10^{-4}$ ，优化器为 Adam；共训练 400 轮。训练片段步长设为 10，以增加样本数量并保持时间连续性。

**骨架拓扑敏感的自编码器.** 在基线模型 CaMN 的 FGD 评估中，使用了基于时间卷积的平铺向量编码器（Embedding-based AutoEncoder）。该结构将整帧姿态作为高维向量输入，对旋转参数的数值尺度高度敏感，当使用 Rot6d 表示时，各关节分量的方差差异会在潜在空间中被放大，导致潜在分布协方差矩阵奇异，从而引起 FGD 数值爆炸。

为避免此问题，本文采用基于骨架拓扑卷积的自编码器（Skeleton-aware AutoEncoder）作为 FGD 特征提取器。该模型在编码层中引入骨架邻接矩阵  $\mathbf{A}$ ，通过局部卷积核在相邻关节之间共享权重：

$$\mathbf{h}_i^{(l+1)} = \sigma \left( \sum_{j \in \mathcal{N}(i)} A_{ij} W^{(l)} \mathbf{h}_j^{(l)} + b^{(l)} \right), \quad (4.3)$$

从而在空间上实现局部归一与结构平滑。这一设计使特征提取对旋转表示形式不敏感，可在欧拉角、Axis-Angle 与 Rot6d 等不同表示下保持稳定的潜在分布。在本文的实验中，该结构显著提高了 FGD 的鲁棒性，避免了 EmbeddingNet 评估器在 Rot6d 表示下协方差爆炸的现象。

**指标计算流程.** 在获得评估模型后，分别将生成序列与真实序列输入自编码器的编码器部分，提取潜在特征  $\mathbf{z}_g$  与  $\mathbf{z}_r$ ，再计算两者的均值与协方差以求得 FGD。评估时按说话人独立计算，再取平均值作为总体指标。

### 4.2.2 Speech-Gesture Rhythm Correlation (SRGR)

SRGR 指标<sup>[3]</sup>用于衡量生成手势与语音韵律在节奏层面的同步性，其核心思想是比较语音能量包络与手势运动速度包络的时间相关程度。与整句级语义匹配不同，SRGR 反映的是语音与手势在局部时间尺度上的节奏耦合强度。

**定义与原理.** 设语音能量序列为  $\mathbf{e} = \{e_t\}$ ，通过对语音短时能量（Short-Time Energy, STE）进行平滑得到；手势运动速度定义为各关节旋转向量的一阶差分模长的平均值：

$$v_t = \frac{1}{J} \sum_{j=1}^J \|\mathbf{r}_{t,j} - \mathbf{r}_{t-1,j}\|_2, \quad (4.4)$$

其中  $J$  为关节数量， $\mathbf{r}_{t,j}$  为第  $j$  个关节在帧  $t$  的旋转向量。

对两条时间序列  $\mathbf{e}$  与  $\mathbf{v}$  进行标准化后，定义其皮尔逊相关系数（Pearson Correlation）为：

$$\text{SRGR} = \frac{\text{Cov}(\mathbf{e}, \mathbf{v})}{\sigma_e \sigma_v}. \quad (4.5)$$

SRGR 值越高，表示手势运动节奏与语音节奏的同步程度越强，通常取值范围在  $[-1, 1]$ 。在实际实验中，为减少局部异常的影响，本文在 1.5 秒滑动窗口内计算局部相关系数并取平均作为最终结果。

**计算流程.** 1. 对语音信号计算短时能量，采用窗口长度 50 ms、步长 10 ms；2. 对生成与真实手势分别计算全身平均运动速度曲线；3. 将两条曲线统一至相同时间分辨率并进行归一化；4. 滑动计算局部皮尔逊相关系数，最后取平均值作为 SRGR。

该指标能客观反映模型在语音驱动节奏一致性方面的表现，在主观实验中也与“同步性”评分显著相关。

### 4.2.3 Beat Alignment (BA)

BA（Beat Alignment）指标<sup>[3]</sup>用于衡量手势关键动作与语音重读节拍的时间对齐程度，反映模型在**时序同步性（temporal synchronization）**方面的性能。与 SRGR 的全局相关性不同，BA 更注重**事件级的对齐精度**。

**定义与原理.** 设语音节拍集合为  $\mathcal{B}_s = \{t_k^s\}$ ，通过检测语音短时能量或梅尔倒谱系数（MFCC）的局部峰值得到；手势峰集合为  $\mathcal{B}_g = \{t_m^g\}$ ，定义为手势速度  $v_t$  的局部极大

点集合。对于每个语音节拍  $t_k^s$ ，计算其最近手势峰的时间差  $\Delta t_k = \min_m |t_k^s - t_m^g|$ 。则 BA 指标定义为：

$$BA = 1 - \frac{1}{K} \sum_{k=1}^K \frac{\Delta t_k}{\tau}, \quad (4.6)$$

其中  $K$  为节拍数， $\tau$  为允许最大偏差阈值（本文取  $\tau = 0.5 \text{ s}$ ）。当  $\Delta t_k < \tau$  时，记为一次成功对齐。因此 BA 越接近 1，表明手势动作越能准确响应语音重读节拍。

**计算流程.** 1. 通过能量包络峰值检测获取语音节拍点；2. 通过运动速度极值检测获取手势关键帧；3. 计算两者的最短时间偏差并归一化；4. 平均所有节拍对齐率，得到整体 BA 指标。

本文使用该指标评估模型在节奏突变与强调语气时的响应能力。在主观观察中，BA 与视觉上“语音同步自然度”呈正相关。

#### 4.2.4 L1-based Diversity (L1DIV)

L1DIV 指标<sup>[3]</sup>衡量生成手势序列的多样性（diversity），即不同语音输入下模型生成结果的动作差异程度。该指标反映模型在保持自然性的同时，能否避免生成收敛到平均动作模式（mode collapse）的倾向。

**定义与原理.** 设模型对同一语音片段的  $N$  次生成结果为  $\{\mathbf{g}^{(1)}, \mathbf{g}^{(2)}, \dots, \mathbf{g}^{(N)}\}$ ，则 L1 多样性定义为任意两序列间平均 L1 距离：

$$L1DIV = \frac{2}{N(N-1)} \sum_{i < j} \frac{1}{T} \sum_{t=1}^T \|\mathbf{g}_t^{(i)} - \mathbf{g}_t^{(j)}\|_1. \quad (4.7)$$

当  $N = 2$  时，该指标即为两次独立生成的平均动作差异。

**计算流程.** 1. 对同一语音输入重复生成多次（不同随机噪声或 dropout 路径）；2. 对每帧姿态计算关节旋转的 L1 差；3. 对时间与样本求平均得到整体 L1DIV 值。

较高的 L1DIV 表明模型生成具有较强的多样性，但过高可能意味着动作不稳定或噪声放大。因此，L1DIV 通常与 FGD 联合分析：FGD 反映真实度，L1DIV 反映丰富度，两者共同平衡模型在自然性—多样性维度上的表现。

### 4.3 模型对比概览

表 ?? 总结了各模型的输入输出模态特征。值得注意的是，我们的模型是唯一同时支持“头部姿态”输入且完全不依赖未来信息的方案。

模型	输入模态					未来信息	输出
	音频	面部捕捉	头部姿态	说话人 ID	情绪		
CaMN	✓	✓	×	✓	✓	✓	身体
DiffSHEG	✓	×	×	✓	×	✓	身体 + 面部
本方法	✓	✓	✓	*	×	×	身体

### 4.4 用户调研

为进一步验证模型在真实交互环境中的表现，本文进行了用户主观评估实验，比较 FaceCapGes、DiffSHEG<sup>[14]</sup> 与 CaMN<sup>[3]</sup> 三个模型在动作自然性、同步性与多样性方面的主观质量。本节首先介绍用户研究系统与实验配置，随后报告主观评价结果与分析。

#### 4.4.1 评估系统与实验配置

**实验材料与呈现方式。** 用户评估所使用的手势动画均基于测试集语音片段生成，并以 BVH（Biovision Hierarchy）文件形式保存。BVH 是一种通用的动作捕捉数据格式，通过层级化定义骨骼结构与帧级旋转参数，可直接导入 3D 动画与虚拟人系统。本文使用的 BVH 文件采用欧拉角旋转表示，以保证与 Unity 的骨骼系统兼容。

由于三种对比模型的输出旋转参数形式不同——CaMN 模型直接输出欧拉角，DiffSHEG 输出 Axis-Angle，而 FaceCapGes 输出 Rot6d 表示——因此在导出 BVH 之前，需将后两者统一转换为欧拉角表示，以便在同一骨架拓扑下进行动画比较。

所有对比模型（FaceCapGes、CaMN、DiffSHEG）均使用相同的训练集与骨骼定义，其中 CaMN 与 DiffSHEG 采用各自论文公开的原版参数。虚拟人角色选自 BEAT 数据集提供的公开演讲者模型集合，并经过 Unity 的 Mecanim 自动骨骼绑定系统匹配。该系统会自动配对 BVH 文件中定义的骨骼层级与虚拟人模型的骨骼节点，从而在不依赖手动权重绘制的情况下完成动作重定向。

本系统当前支持的虚拟人模型需同时具备骨骼绑定与 ARKit 兼容的 BlendShape 参数。基于此约束，实验选用了 CaMN 论文公开的男女两名演讲者模型，二者均满

足兼容要求，可实现身体与面部的联合驱动。

**播放系统实现。** 我们基于 Unity 自行编写播放脚本，将各模型生成的 BVH 动画用于驱动虚拟人身体骨骼，同时以面部捕捉序列驱动 BlendShape 表情参数，并同步播放原始语音音频。系统支持同时呈现三种模型生成的动画结果：用户可在同一画面中（左、中、右）并行观察三种手势表现，所有语音与面部表情完全一致，唯一变量为身体动作。该设计使参与者能够直接比较不同模型在动作风格、节奏响应与语音同步性方面的差异。

为确保主观评价的公正性与可重复性，系统在每次实验开始前会随机分配三种模型的位置（左、中、右），界面上不会显示模型名称，从而避免潜在偏向。各测试片段的播放顺序在实验前统一设定，以保证不同参与者之间的样本顺序均衡。实验员在播放系统后台记录当前序列与模型对应关系，以便后续结果统计。

**实验界面与设备。** 评估系统提供桌面端与 VR 端两种版本，功能完全一致。VR 版本基于 PICO 设备实现；桌面版支持多窗口并行播放，方便用户同时对比。如图 A.2 所示，播放界面在两种设备上保持统一布局，播放完成后参与者需通过交互界面对三个模型进行排序打分。VR 用户在沉浸式环境中逐一观看三段动画；桌面端用户则可在单屏上同时观察全部模型。因此前者注重细节感知与临场性，后者更有利于整体风格与节奏的一致性对比。

**实验流程与指导。** 实验正式开始前，研究人员向参与者说明了三项主观评价标准的含义，确保所有被试对评分维度理解一致：

- **真实感 (Realism)：** 整体动作是否自然流畅，是否存在明显的违和感，如朝向异常或突然抖动；
- **同步性 (Synchronization)：** 手势动作与语调、语音节奏是否协调一致；
- **多样性 (Diversity)：** 手势是否丰富多变，避免长时间静止或重复单一动作。

在实验过程中，VR 版本于线下环境进行，桌面版通过线上远程环境执行。两种形式均保持实时交流通道，研究人员可在参与者提问时即时解释操作或澄清评分标准。在正式评估阶段，参与者可多次重播当前片段，但不能返回查看先前内容，以减少记忆偏差。所有播放条件（Unity 场景内的相机角度、光照参数、音量与分辨率设置）在全部被试中保持一致，以确保渲染输出的可比性。



需要说明的是，对于 **VR 实验**，所有测试均在相同的线下实验室环境中进行，使用同一套 **PICO** 设备与照明条件；而**桌面端实验**通过远程方式执行，参与者在各自电脑上运行实验程序。研究人员可通过实时屏幕共享观察其操作流程并保持语音沟通，但无法严格控制其所在房间的光照或环境噪声条件。因此，桌面端实验在“观看环境”上存在一定差异，但由于任务内容与播放系统完全相同，且实验员在测试中持续指导，可认为该差异对结果的总体影响有限。

**实验材料与任务设计.** 评估样本来自 BEAT 数据集中四位演讲者（ID 2、4、6、8），其中 2、4 为男性，6、8 为女性。每位演讲者各选取两段平均长度约 1 分钟的语音片段，演讲话题互不重复，共组成 **8 段固定视频样本**。所有实验均使用相同的 8 段样本，但其呈现顺序在不同被试间经过随机化或平衡化处理，以避免顺序效应（order effect）。每段视频均包含三种模型生成的动作版本（FaceCapGes、CaMN、DiffSHEG），并在播放时随机分配其在屏幕的左右位置。参与者在观看每个片段后，根据三项主观标准（真实感、同步性、多样性）对三个模型的表现进行排序评估。



图 4.1 用户研究系统界面示意图。左、中、右位置随机分配给三种模型，音频与面部表情完全一致，仅身体动作不同。

**实验参与者.** 共邀请 16 名参与者（12 名使用 VR 设备，4 名使用桌面端），涵盖不同性别与学术背景。所有参与者在实验前均接受了操作说明与校准，并在系统指导

下完成评分练习。为避免呈现顺序对主观印象造成偏差，另设计了采用平衡拉丁方（Balanced Latin Square）顺序的实验版本，使不同参与者观看样本的顺序均衡分布。

#### 4.4.2 平衡拉丁方排序设计

该版本实验共招募 8 位 VR 用户（4 男 4 女），排序顺序由 HCI 用户研究工具包<sup>[38]</sup>自动生成，确保模型与演讲者组合的呈现顺序在全体被试间均匀分布。所有条件保持一致，唯一变量为视频播放顺序。

#### 4.4.3 结果与分析

本节综合分析两轮用户评估的统计结果与参与者反馈。所有结果均基于相同的 8 段测试样本，其中模型位置与播放顺序在不同被试间随机化或经平衡拉丁方控制，以保证主观评价的公正性。

**电脑用户测评结果。** 如图 4.2 所示，在 16 名参与者的总体评价中，FaceCapGes 在三个维度（真实感、同步性、多样性）上均优于基线模型 CaMN，并在“真实感”方面与离线扩散模型 DiffSHEG 持平。这一结果表明，FaceCapGes 虽在严格的实时因果约束下运行，但仍能保持与非实时生成模型相近的动作自然度与流畅性。

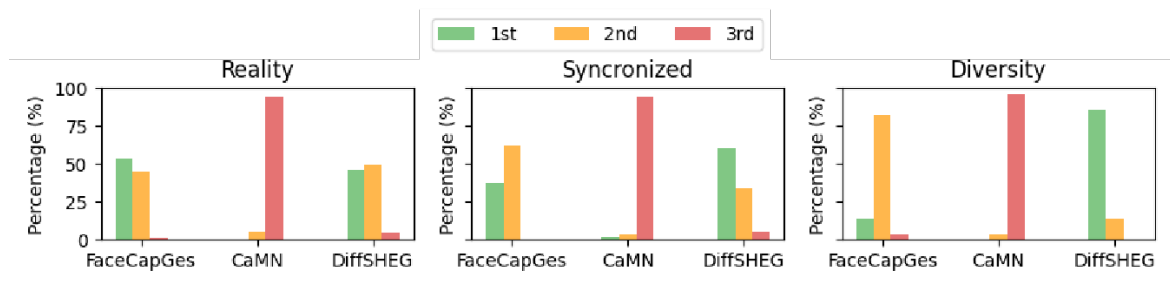


图 4.2 16 名参与者对三种模型在“真实感”、“同步性”和“多样性”三个维度的主观排名结果。

**平衡拉丁方实验结果。** 图 A.1 展示了平衡拉丁方实验版本中 8 位 VR 用户的独立结果，该版本严格控制了模型与演讲者组合的呈现顺序。结果与总体趋势一致，FaceCapGes 在“同步性”与“真实感”上显著优于 CaMN，并在“多样性”指标上与 DiffSHEG 接近。这表明实验结果在不同顺序条件下保持稳定，进一步验证了模型在多维度主观评价中的一致优势。

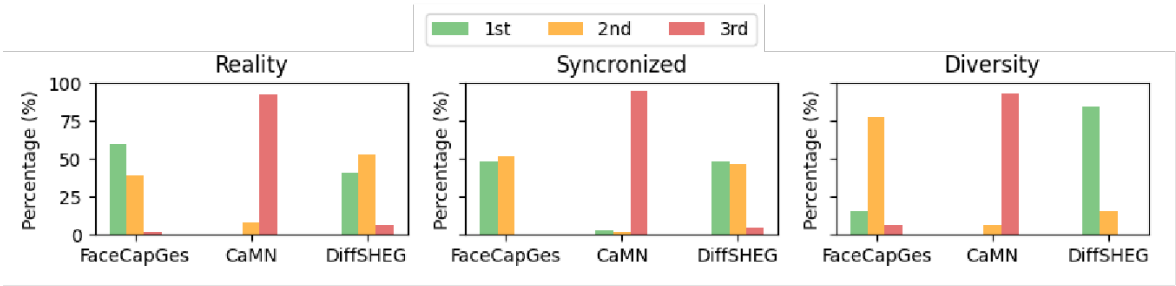


图 4.3 平衡拉丁方实验版本中 8 位 VR 用户的主观排名结果，评价维度包括“真实感”、“同步性”和“多样性”。

**用户反馈分析。** 根据实验后访谈与自由评论汇总，参与者普遍认为 FaceCapGes 的动作过渡自然、节奏感强，手势响应与语音重音、语调变化更加一致。部分 VR 用户指出，在沉浸式视角下能够明显感受到手势的连贯性与时序协调，而 CaMN 在头部与上身动作衔接处偶有“僵硬”或“转向延迟”现象。约三分之一的参与者提到 DiffSHEG 的动作表现力最强，但在某些片段中出现手部摆动幅度过大或抖动情况，导致同步性评分略低。

桌面端用户普遍认为三种模型的差异在节奏与流畅性上最明显；VR 用户则更容易察觉动作的空间一致性与临场协调。两组被试的整体排序趋势一致，说明模型差异具有跨设备一致性。

**结果讨论。** CaMN 得分较低的主要原因在于其采用欧拉角表示，导致部分姿态序列在旋转空间中出现不连续，尤其在头部动作上易产生跳变；而 DiffSHEG 的高方差输出虽提升了动作多样性，但表现出一定的突然颤抖，可能来源于 Axis-Angle 的不连续性。相比之下，FaceCapGes 的因果式时间建模和头部姿态融合策略有效提升了局部动作的平滑性与节奏协调，使其在实时条件下同时兼顾自然度与稳定性。此外，实验的平衡拉丁方版本进一步证明结果在不同呈现顺序下的一致性，排除了顺序偏差对主观评价的显著影响。

综上，用户研究表明 FaceCapGes 在实时生成条件下仍能维持与离线模型相当的主观表现，在动作真实感、节奏同步性和长时间交互稳定性方面均显著优于传统因果结构基线，验证了本文提出的多模态融合与时间建模策略的有效性。

## 4.5 定性分析

我们强烈建议观看电子附录中的演示视频，内容包含 GT、CaMN、DiffSHEG 与本模型的并排展示，能够直观体现时间对齐性、手势响应性以及头-身协调性方面的差异。

如图 4.4 所示，FaceCapGes 能平滑且富有表现力地捕捉说话人动态。与 CaMN 相比，本模型避免了欧拉角带来的不连续问题，生成的身体动作可顺应头部运动趋势。与 DiffSHEG 相比，我们的模型在动作速度上略逊一筹，但能有效避免抖动或夸张姿态。头部姿态的引入也明显提升了手势方向与语义的一致性。

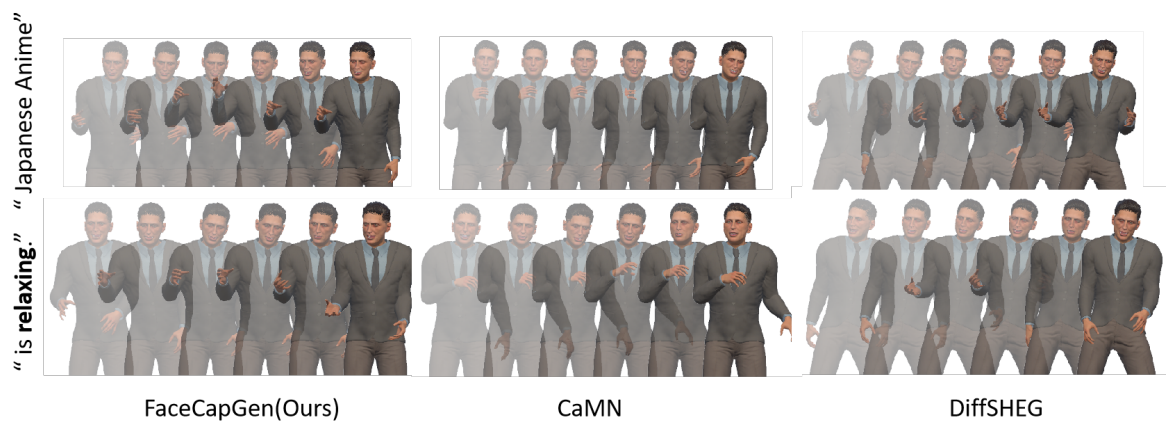


图 4.4 GT、CaMN、DiffSHEG 与 FaceCapGes 的动作效果对比。本模型手势响应自然，方向与头部朝向一致。

## 4.6 定量分析

区域	方法	FGD↓	SRGR↑	BA↑	L1DIV↑
全身	CaMN	47.732	0.098	0.845	7.591
	DiffSHEG	<u>26.846</u>	<b>0.109</b>	<u>0.883</u>	<u>11.028</u>
	本方法	<b>23.385</b>	<u>0.107</u>	<b>0.913</b>	<b>13.284</b>
不含头部	CaMN	49.437	0.103	0.845	7.327
	DiffSHEG	<u>26.847</u>	<b>0.110</b>	<u>0.883</u>	<u>10.990</u>
	本方法	<b>23.384</b>	<u>0.107</u>	<b>0.913</b>	<b>13.330</b>

表 4.1 在 BEAT 四位说话人测试集上的定量评估结果。FaceCapGes 在所有指标上优于 CaMN，且与 DiffSHEG 表现相当。

表 4.1 显示，FaceCapGes 在所有指标上均优于 CaMN。与 DiffSHEG 相比，本模型 FGD 更低，SRGR 相近，且在 L1DIV 上表现最优，表明其生成动作具备良好多样

性。但这一结论与用户主观评分存在一定出入：DiffSHEG 在“多样性”上主观排名更高。

这一差异揭示了 L1DIV 的局限性：它主要衡量空间偏离程度，并不能体现动作频率或视觉表现力。虽然我们的模型在结构上更丰富，但用户普遍认为 DiffSHEG 更“活跃”，吸引注意。这说明 L1DIV 无法完全捕捉感知上的丰富性，进一步强调定量与主观评估的互补作用。

## 4.7 消融实验分析

区域	变体	FGD↓	SRGR↑	BA↑	L1DIV↑
全身	CaMN	32.870	0.111	0.858	7.214
	移除头部姿态	19.591	0.123	0.916	10.642
	移除帧级生成	21.592	<b>0.125</b>	0.892	10.456
	FaceCapGes（本方法）	<b>19.290</b>	0.123	<b>0.918</b>	<b>10.871</b>

表 4.2 说话人 2 的消融实验结果。头部姿态对提升手势自然性与表达力有显著作用。

表 4.2 展示了各模块对模型性能的影响。移除头部姿态会导致所有指标下降，证明其对表达性手势的重要性。

值得注意的是，“非帧级”版本使用未来上下文与双向 LSTM，但其表现反而不及帧级版本。这是因为它采用一次性解码，对每段输入独立处理，无法保证时间连续性，与原 CaMN 的设计一致。而我们的方法通过滑动窗口训练和自回归推理，逐帧依赖历史信息，使动作更连贯、过渡更自然。即使不使用未来帧，也能获得更低的 FGD 和更自然的手势对齐。

## 4.8 性能评估

指标	时间
测试帧数	93015 (f)
推理总时长	504 (s)
平均单帧时间	6.07E-03 (s/f)

表 4.3 在 RTX 4090 上的帧级推理速度评估结果。

为评估实时性能，我们在 BEAT 测试集上使用 batch size 为 1 的配置进行推理测量。如表 4.3 所示，模型平均每帧处理时间小于 7 毫秒，满足实时响应需求。

尽管模型推理本身较快，但系统总延迟还受输入特征提取影响。在我们的原型中，音频特征通过 Librosa 离线提取，93,015 帧总计耗时约 61 秒，单帧不足 1 毫秒。实际部署中可替换为如 PyAudio 或 SoundDevice 的实时提取方法，尽管本文未测量其具体延迟。

ARKit 面部追踪的延迟由开发文档估算约为 60FPS<sup>[2]</sup>。结合模型推理时间，整体系统延迟预计不超过 25ms，足以满足实时虚拟人交互的需求。

## 第 5 章 结论

本文提出了 FaceCapGes，一种基于 CaMN 框架的帧级实时手势生成模型。与现有离线或依赖未来信息的方法不同，FaceCapGes 可使用户仅通过实时音频、面部 blendshape 权重与头部姿态，即可驱动虚拟角色生成自然手势，无需动作捕捉设备。

本模型融合 LSTM、MLP 以及对抗学习机制，采用级联架构与滑动窗口式自回归训练策略，支持低延迟的在线推理。我们首次将头部姿态作为终端模态引入，有效提升了手势的自然度与协调性。

综合评估表明，FaceCapGes 在生成质量上显著优于 CaMN，并在保持实时性的同时，在主观和客观指标上与主流离线方法表现相当。其模块化设计使其可部署于兼容 ARKit 的轻量设备上，具备良好的实用性与扩展性。

尽管当前系统尚未显式建模语义驱动的手势，未来工作将探索引入语言理解与意图识别机制，以丰富实时交互中的手势语义表达。此外，我们还计划扩展对通用摄像头与非 iOS 平台的支持，通过开发跨平台的面部与姿态捕捉接口，缓解当前对 ARKit 的依赖。





## 参考文献

- [1] KARTYNNIK Y, ABLAVATSKI A, GRISHCHENKO I, et al. Real-time facial surface geometry from monocular video on mobile gpus[J]. arXiv:1907.06724, 2019.
- [2] NYISZTOR K. Introduction to Augmented Reality with ARKit[EB/OL]. 2019. <https://www.pluralsight.com/resources/blog/guides/introduction-to-augmented-reality-with-arkit>.
- [3] LIU H, ZHU Z, IWAMOTO N, et al. BEAT: A Large-Scale Semantic and Emotional Multi-Modal Dataset for Conversational Gestures Synthesis[J]. arXiv preprint arXiv:2203.05297, 2022.
- [4] KENDON A. Gesture: Visible Action as Utterance[M]. Cambridge, UK: Cambridge University Press, 2004.
- [5] MCNEILL D. Hand and Mind: What Gestures Reveal about Thought[M]. Chicago, IL: University of Chicago Press, 1992.
- [6] WAGNER P, MALISZ Z, KOPP S. Gesture and speech in interaction: An overview [J/OL]. Speech Communication, 2014, 57: 209-232. DOI: 10.1016/j.specom.2013.09.008.
- [7] HADAR U, BUTTERWORTH B. Iconic gestures, imagery, and word retrieval in speech[J]. Semiotica, 1989, 75(1/2): 63-83.
- [8] ESTEVE-GIBERT N, PRIETO P, PONS X, et al. The timing of head movements: The role of prosodic heads and edges[J/OL]. The Journal of the Acoustical Society of America, 2017, 141(6): 4727-4739. DOI: 10.1121/1.4986649.
- [9] MAHMOOD N, GHORBANI N, TROJE N F, et al. AMASS: Archive of Motion Capture as Surface Shapes[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 5442-5451.
- [10] LOPER M, MAHMOOD N, ROMERO J, et al. SMPL: A Skinned Multi-Person Linear Model[J]. ACM Transactions on Graphics (TOG), 2015, 34(6): 248:1-248:16.
- [11] GLEICHER M. Retargeting Motion to New Characters[C]//Proceedings of the 25th Annual Conference on Computer Graphics & Interactive Techniques (SIGGRAPH '98). ACM, 1998: 33-42.

- [12] MARTINELLI G, GARAU N, BISAGNO N, et al. Skeleton-Aware Motion Retargeting Using Masked Pose Modeling[C]//European Conference on Computer Vision (ECCV) Workshops, LNCS 15624. Springer, 2024: 287-303.
- [13] ZHOU Y, BARNES C, LU J, et al. On the Continuity of Rotation Representations in Neural Networks[J]. arXiv preprint arXiv:1812.07035v4, 2020.
- [14] CHEN J, LIU Y, WANG J, et al. DiffSHEG: A Diffusion-Based Approach for Real-Time Speech-driven Holistic 3D Expression and Gesture Generation[C]//CVPR. 2024.
- [15] FU C, WANG Y, ZHANG J, et al. MambaGesture: Enhancing Co-Speech Gesture Generation with Mamba and Disentangled Multi-Modality Fusion[EB/OL]. 2024. <https://arxiv.org/abs/2407.19976>. arXiv: 2407.19976 [cs.LG].
- [16] LIU H, ZHU Z, BECHERINI G, et al. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling[J]. arXiv:2401.00374, 2024.
- [17] AO T, ZHANG Z, LIU L. GestureDiffuCLIP: Gesture Diffusion Model with CLIP Latents[J/OL]. ACM Trans. Graph., DOI: 10.1145/3592097.
- [18] CHHATRE K, DANĚČEK R, ATHANASIOU N, et al. AMUSE: Emotional Speech-driven 3D Body Animation via Disentangled Latent Diffusion[C/OL]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024: 1942-1953. <https://amuse.is.tue.mpg.de>.
- [19] EKMAN P, FRIESEN W V. Facial Action Coding System: A Technique for the Measurement of Facial Movement[M]. Palo Alto, California: Consulting Psychologists Press, 1978.
- [20] ARKit in iOS - Tracking and Visualizing Faces. 2024. [https://developer.apple.com/documentation/arkit/arkit\\_in\\_ios/content\\_anchors/tracking\\_and\\_visualizing\\_faces](https://developer.apple.com/documentation/arkit/arkit_in_ios/content_anchors/tracking_and_visualizing_faces).
- [21] OZEL M. ARKit to FACS Cheat Sheet[EB/OL]. 2022. <https://melindaazel.com/arkit-to-facs-cheat-sheet/>.
- [22] GINOSAR S, BAR A, KOHAVI G, et al. Learning Individual Styles of Conversational Gesture[C/OL]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 3497-3506. [https://openaccess.thecvf.com/content\\_CVPR\\_2019/papers/Ginosar\\_Learning\\_Individual\\_Styles\\_of\\_Conversational\\_Ge](https://openaccess.thecvf.com/content_CVPR_2019/papers/Ginosar_Learning_Individual_Styles_of_Conversational_Ge)

- sture\_CVPR\_2019\_paper.pdf.
- [23] KUCHERENKO T, HABERNAL I, BESKOW J, et al. Multimodal Analysis of the Predictability of Hand-Gesture Properties[J/OL]. *Frontiers in Computer Science*, 2021, 3: 10. [https://web.cs.ucdavis.edu/~neff/papers/kucherenko2021hand\\_propert\\_predictability\\_final.pdf](https://web.cs.ucdavis.edu/~neff/papers/kucherenko2021hand_propert_predictability_final.pdf).
- [24] CASSELL J, VILHJÁLMSSON H H, BICKMORE T. BEAT: the Behavior Expression Animation Toolkit[C/OL]//SIGGRAPH '01: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. 2001: 477-486. <https://doi.org/10.1145/383259.383315>. DOI: 10.1145/383259.383315.
- [25] HUANG C M, MUTLU B. Robot behavior toolkit: generating effective social behaviors for robots[C/OL]//HRI '12: Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction. 2012: 25-32. <https://doi.org/10.1145/2157689.2157694>. DOI: 10.1145/2157689.2157694.
- [26] KIPP M. Gesture generation by imitation: from human behavior to computer character animation[C/OL]//. 2005. <https://api.semanticscholar.org/CorpusID:26271318>.
- [27] ZHANG Z, AO T, ZHANG Y, et al. Semantic Gesticulator: Semantics-Aware Co-Speech Gesture Synthesis[J]. *ACM Transactions on Graphics (TOG)*, 2024, 43(4): 1-17.
- [28] ZHU L, LIU X, LIU X, et al. Taming Diffusion Models for Audio-Driven Co-Speech Gesture Generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 10544-10553.
- [29] YANG S, WU Z, LI M, et al. DiffuseStyleGesture: Stylized Audio-Driven Co-Speech Gesture Generation with Diffusion Models[C/OL]//Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23. 2023: 5860-5868. <https://doi.org/10.24963/ijcai.2023/650>. DOI: 10.24963/ijcai.2023/650.
- [30] HOGUE S, ZHANG C, DARUGER H, et al. DiffTED: One-shot Audio-driven TED Talk Video Generation with Diffusion-based Co-speech Gestures[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. 2024: 1922-1931.
- [31] DEICHLER A, MEHTA S, ALEXANDERSON S, et al. Diffusion-Based Co-Speech Gesture Generation Using Joint Text and Audio Representation[C/OL]//ICMI '23:

- Proceedings of the 25th International Conference on Multimodal Interaction. 2023: 755-762. <https://doi.org/10.1145/3577190.3616117>. DOI: 10.1145/3577190.3616117.
- [32] ALEXANDERSON S, KUCHERENKO T, HENTER G E, et al. DiffGesture: Diffusion-based Co-Speech Gesture Generation[C/OL]//Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents (IVA 2023). 2023: 1-8. <https://dl.acm.org/doi/10.1145/3570945.3607282>. DOI: 10.1145/3570945.3607282.
- [33] YOON Y, KIM S, LEE J, et al. Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity[C/OL]//Proceedings of the 2020 ACM International Conference on Multimodal Interaction (ICMI). 2020: 22-30. <https://dl.acm.org/doi/10.1145/3382507.3418838>. DOI: 10.1145/3382507.3418838.
- [34] ALEXANDERSON S, HENTER G E, KUCHERENKO T, et al. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows[C/OL]//Computer Graphics Forum (Proc. Eurographics). 2020: 487-496. <https://people.kth.se/~ghe/pubs/pdf/alexanderson2020style.pdf>.
- [35] KUCHERENKO T, BESKOW J, KJELLSTRÖM H, et al. Moving Fast and Slow: Analysis of Representations and Post-processing in Speech-Driven Gesture Generation[J/OL]. arXiv preprint arXiv:2107.12305, 2021. <https://people.kth.se/~ghe/pubs/pdf/kucherenko2021moving.pdf>.
- [36] KINGMA D P, BA J. Adam: A Method for Stochastic Optimization[EB/OL]. 2017. <https://arxiv.org/abs/1412.6980>. arXiv: 1412.6980 [cs.LG].
- [37] YOON Y, CHA B, LEE J H, et al. Speech gesture generation from the TRIMODAL context of text, audio, and speaker identity[J]. arXiv:2009.02119, 2020.
- [38] SCHWIND V, RESCH S, SEHRT J. The HCI User Studies Toolkit: Supporting Study Designing and Planning for Undergraduates and Novice Researchers in Human-Computer Interaction[C/OL]//Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23). ACM, 2023: 7. DOI: 10.1145/3544549.3585890.

## 附录 A 平衡拉丁方排序设计

为进一步控制呈现顺序带来的偏差，我们设计并实施了用户研究的另一版本，采用 HCI 用户研究工具包<sup>[38]</sup> 生成的平衡拉丁方（Balanced Latin Square）排序。本实验版本共招募 8 位参与者（4 名男性，4 名女性），全部使用 VR 头显设备，以构建性别平衡的样本组。如图 A.1 所示，展示了该版本用户在“真实感”、“同步性”和“多样性”三个维度上的评分结果。

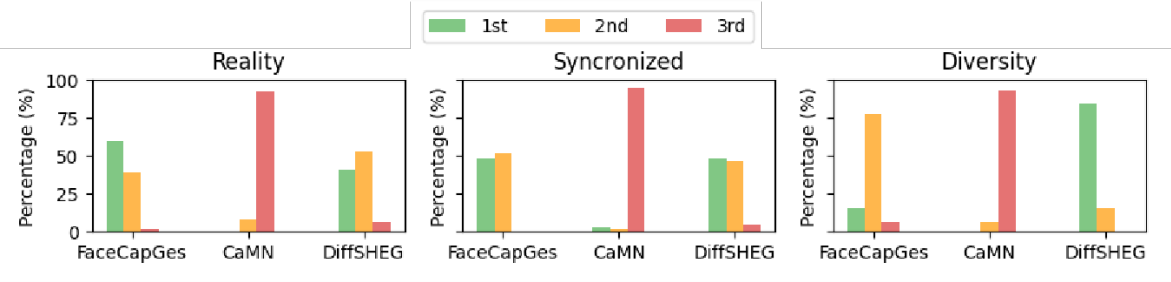


图 A.1 平衡拉丁方实验版本中的用户主观评分结果，共 8 位参与者，评价维度包括“真实感”、“同步性”和“多样性”。

本研究中所用播放界面如图 A.2 所示。该界面在 VR 与桌面端中保持一致，使参与者能够以随机顺序观看三种模型（FaceCapGes（本方法）、DiffSHEG<sup>[14]</sup>、CaMN<sup>[3]</sup>）生成的手势动画。用户可多次重播当前片段，但不能返回浏览先前内容。在每段视频播放结束后，参与者通过拖动界面下方的评分面板，对三个模型进行排序打分。该界面同样适用于平衡拉丁方实验版本，确保所有参与者使用统一的评估流程。



图 A.2 用户研究中使用的手势播放界面示意图。

## 致 谢

感谢那位最先制作出博士学位论文  $\text{\LaTeX}$  模板的物理系同学！

感谢 William Wang 同学对模板移植做出的贡献！

感谢 @weijianwen 学长开创性的工作！

感谢 @sjtug 对 0.10 及之后版本的开发和维护工作！

感谢所有为模板贡献过代码的同学们，以及所有测试和使用模板的各位同学！

感谢  $\text{\LaTeX}$  和 SJTUThesis，帮我节省了不少时间。





## 学术论文和科研成果目录

### 学术论文

- [1] Chen H, Chan C T. Acoustic cloaking in three dimensions using acoustic metamaterials[J]. Applied Physics Letters, 2007, 91:183518.
- [2] Chen H, Wu B I, Zhang B, et al. Electromagnetic Wave Interactions with a Metamaterial Cloak[J]. Physical Review Letters, 2007, 99(6):63903.

### 专利

- [1] 第一发明人, “永动机”, 专利申请号 202510149890.0.



## 个人简历

### 基本情况

某某，yyyy 年 mm 月生于 xxxx。

### 教育背景

- yyyy 年 mm 月至今，上海交通大学，博士研究生，xx 专业
- yyyy 年 mm 月至 yyyy 年 mm 月，上海交通大学，硕士研究生，xx 专业
- yyyy 年 mm 月至 yyyy 年 mm 月，上海交通大学，本科，xx 专业

### 研究兴趣

L<sup>A</sup>T<sub>E</sub>X 排版

### 联系方式

- 地址：上海市闵行区东川路 800 号，200240
- E-mail: john\_smith@sjtu.edu.cn



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

An imperial edict issued in 1896 by Emperor Guangxu, established Nanyang Public School in Shanghai. The normal school, school of foreign studies, middle school and a high school were established. Sheng Xuanhuai, the person responsible for proposing the idea to the emperor, became the first president and is regarded as the founder of the university.

During the 1930s, the university gained a reputation of nurturing top engineers. After the foundation of People's Republic, some faculties were transferred to other universities. A significant amount of its faculty were sent in 1956, by the national government, to Xi'an to help build up Xi'an Jiao Tong University in western China. Afterwards, the school was officially renamed Shanghai Jiao Tong University.

Since the reform and opening up policy in China, SJTU has taken the lead in management reform of institutions for higher education, regaining its vigor and vitality with an unprecedented momentum of growth. SJTU includes five beautiful campuses, Xuhui, Minhang, Luwan Qibao, and Fahu, taking up an area of about 3 225 833 m<sup>2</sup>. A number of disciplines have been advancing towards the top echelon internationally, and a batch of burgeoning branches of learning have taken an important position domestically.

Today SJTU has 31 schools (departments), 63 undergraduate programs, 250 masters-degree programs, 203 Ph.D. programs, 28 post-doctorate programs, and 11 state key laboratories and national engineering research centers.

SJTU boasts a large number of famous scientists and professors, including 35 academics of the Academy of Sciences and Academy of Engineering, 95 accredited professors and chair professors of the "Cheung Kong Scholars Program" and more than 2000 professors and associate professors.

Its total enrollment of students amounts to 35 929, of which 1564 are international students. There are 16 802 undergraduates, and 17 563 masters and Ph.D. candidates. After more than a century of operation, Jiao Tong University has inherited the old tradition of "high starting points, solid foundation, strict requirements and extensive practice." Students

from SJTU have won top prizes in various competitions, including ACM International Collegiate Programming Contest, International Mathematical Contest in Modeling and Electronics Design Contests. Famous alumni include Jiang Zemin, Lu Dingyi, Ding Guangen, Wang Daohan, Qian Xuesen, Wu Wenjun, Zou Taofen, Mao Yisheng, Cai Er, Huang Yanpei, Shao Lizi, Wang An and many more. More than 200 of the academics of the Chinese Academy of Sciences and Chinese Academy of Engineering are alumni of Jiao Tong University.