arXiv:2601.18451v1 [cs.CV] 26 Jan 2026

3DGesPolicy：基于动作控制的音素感知整体语音伴随手势生成

Xuanmeng Sha 大阪大学 日本
shaxuanmeng@gmail.com

Liyun Zhang 大阪大学 日本
liyun.zhang@lab.ime.cmc.osaka-u.ac.jp

Tomohiro Mashita 大阪电气通信大学 日本
mashita@osakac.ac.jp

Naoya Chiba 大阪大学 日本
chiba@nchiba.net

Yuki Uranishi 大阪大学 日本
yuki.uranishi.cmc@osaka-u.ac.jp

摘要

生成融合全身动作与面部表情的整体伴 speech 手势时，现有的部件分解或帧级回归方法会导致身体动作的协调在语义上不连贯，并产生空间不稳定、毫无意义的动作。我们提出了 3DGesPolicy，这是一种新颖的基于动作的框架，通过机器人学中的扩散策略，将整体手势生成重新定义为连续轨迹控制问题。通过将帧间变化建模为统一的整体动作，我们的方法能够有效学习帧间整体手势运动模式，并确保动作轨迹在空间和语义上连贯，同时符合真实的运动流形。为进一步弥合表现力对齐方面的差距，我们提出了手势-音频-音素（GAP）融合模块，能够深入整合并细化多模态信号，确保语音语义、身体动作与面部表情之间实现结构化的细粒度对齐。在 BEAT2 数据集上进行的大量定量和定性实验表明，与其他最先进方法相比，3DGesPolicy 在生成自然、富有表现力且与语音高度对齐的整体手势方面表现有效。

CCS 概念

• **计算方法 → 计算机视觉；机器人视觉；神经网络。**

关键词

伴 speech 生成、手势生成、扩散策略、轨迹控制、3D 人体动画

1 引言

逼真的 3D 数字人对于虚拟助手、社交机器人以及 AR/VR 远程临场等应用至关重要，在这些应用中

允许免费为个人或课堂使用而复制本作品的全部或部分内容（无论数字版还是纸质版），但前提是不得为营利或商业利益而复制或分发，且所有副本均须保留本声明及首页的完整引文。对于本作品中归他人而非作者所有的组件，必须尊重其版权。注明出处后允许摘录。若要以其他方式复制、重新发布、张贴到服务器或重新分发至列表，须事先获得具体许可和/或支付费用。请通过 permissions@acm.org 申请许可。

, ,© 2026 版权所有归所有者/作者所有。出版权已授权 ACM。ACM ISBN 978-x-xxxx-xxxx-x/YY/MM https://doi.org/10.1145/nnnnnnn.nnnnnnn

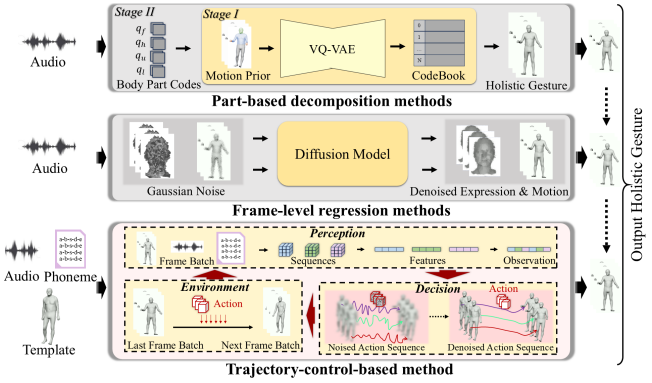

图1：整体语音伴随手势生成范式的比较。不同于采用离散码本或绝对手势回归的基于部件或帧级方法，3DGesPolicy 将此任务重新定义为基于动作的轨迹控制问题。通过在扩散策略训练循环中对音素级统一手势动作进行优化，我们的框架确保生成稳定、具有空间连贯性和语义连贯性的动作。

动作质量会直接影响用户参与度和沟通效果 [6, 7, 9]。实现这一目标的关键在于伴 speech 手势合成。尽管语音驱动的孤立身体手势 [1, 41, 43, 44]和说话人脸 [11, 36, 48] 已取得令人瞩目的成果，但要实现整合二者的整体伴 speech 手势合成仍然十分困难，因为这涉及身体部位协调和语义对齐[4, 9]。

现有方法主要采用两种思路。如图1所示，基于部位的分解方法 [4, 9]利用离散码本 [10] 来预测身体部位运动。尽管这种设计提供了强有力的运动先验，但量化的身体片段会削弱部件间协调和运动细节，导致生成语义不连贯的手势 [3]。相比之下，帧级回归方法[8, 12] 利用扩散模型直接回归手势序列，从而保留运动细节和多样性。然而，这类方法对噪声敏感，并且严重依赖训练数据分布 [12]，缺乏捕捉语音对齐运动的鲁棒性，且容易生成语义无意义的动作。此外，这两种

种范式往往都忽视了三维空间中的空间移动和交互，导致生成的结果不自然、不真实，且动作不稳定、漂浮。

为了解决这些局限，我们提出了 3DGesPolicy，这是一种新颖的轨迹控制框架，利用源自机器人控制的扩散策略机制 [13] 生成整体的、伴随语音的表现性手势。不同于以往预测绝对手势或离散身体编码的工作，我们将全身和面部运动的帧间变化建模为统一的整体动作，并学习一种视觉运动策略，根据当前运动状态、音频和音素级语义生成手势运动轨迹。这种重新表述将任务从手势回归转变为轨迹控制，使模型能够学习三维空间中的语义感知运动模式，而非仅仅拟合绝对手势姿态或分解后的身体部位。这种轨迹控制方法确保动作稳定且有据可依，自然遵循物理约束，从而生成更逼真的结果，同时对各种语音输入保持鲁棒性。

为确保语义对齐并消除无意义的动作，我们设计了一种新颖的手势-音频-音素（GAP）融合模块，用于深度整合多模态信号。GAP 模块为音频和音素特征分别采用独立的编码器，随后通过跨模态注意力机制，将它们与手势表征进行对齐和融合。这一设计确保生成的动作直接受到音素层面语言语义的引导，避免产生过多或无关的动作，同时促进身体运动与面部表情之间与语音对齐的协调，实现精细的音素级同步，生成自然、语义丰富且富有情感表现力的结果。

大量实验表明，与其他最先进方法相比，3DGesPolicy 在生成自然且富有表现力的身体和面部区域手势方面效果显著。我们的主要贡献有三点：

- **我们提出了 3DGesPolicy，这是一种轨迹控制框架，将整体手势生成重新表述为基于动作的扩散策略问题**。 通过将帧间变化建模为统一动作，我们的方法能够学习稳定、自然的运动轨迹，并自然地跟随语音动态。

- **我们提出了手势-音频-音素（GAP）融合模块，实现跨模态的细粒度音素级对齐**。 通过自适应门控和跨模态注意力，该模块深入融合语音声学、语言语义、身体动作和面部表情。

- **大量实验验证了我们的方法在整体手势生成方面的有效性和优越性**。 定量评估和定性评估均表明，与其他最先进的方法相比，我们的方法能够生成自然、稳定的全身运动，实现精准的唇动与语音同步，并产生语义丰富的手势。

2 相关工作

2.1 整体共语手势生成

语音伴随手势生成旨在创建与语音和情感相协调的全身运动。尽管近期研究着重分析对话中的细微情感动态， [38, 39], 生成方法则旨在合成这些线索。早期研究包括基于规则的方法 [37, 40]和数据驱动方法 [1, 41–44]。近年来，整体生成逐渐兴起，旨在协调身体姿态、面部表情和唇部运动 [4, 5, 9, 27]。当前方法主要遵循两种范式：基于部位的分解和帧级回归。

2.1.1 基于部件的分解方法。 这类方法使用身体部位的离散表示来生成手势。TalkSHOW[5] 使用组合式向量量化变分自编码器 [28] 对身体和手部动作进行编码。EMAGE [4] 引入掩码音频-手势建模，以改善重建效果。MambaTalk [9] 集成选择性状态空间模型，以提高效率。然而，离散量化往往会丢失细粒度细节 [3]。此外，将身体部位分开处理会削弱协调性，导致整体动作不自然。

2.1.2 帧级回归方法。 这些方法使用扩散模型在连续空间中回归手势，具有更高的多样性。DiffSHEG [8] 对面部与身体之间的联合分布进行了建模。ExGes [12] 采用检索增强框架和对比学习 [29, 30] 来识别关键帧。然而，由于这些方法预测的是绝对姿态，因此对噪声较为敏感，鲁棒性不足。这往往会导致产生毫无意义、偏离真实运动流形的动作。

2.2 扩散策略模型

扩散策略 [13] 是一种强大的机器人学框架，可从视觉观测中学习复杂的操作任务。先进的视觉感知和平移技术 [14–17] 可进一步增强此类观测编码的鲁棒性。该框架已应用于强化学习 [18, 19], 、模仿学习 [20, 21], 和运动规划 [22–24]。近期研究在这一框架中加入了多视角条件化和优化 [25, 26] 3 [31] 。值得注意的是，D Diffusion Policy 将感知与决策解耦，从而提升了 3D 任务的表现。受这些机器人控制领域成功经验的启发，我们将扩散策略范式应用于手势生成，以实现稳定且自然的轨迹控制。

3 方法

3.1 问题表述

我们将整体共语手势生成重新表述为一个连续轨迹控制问题。在这一范式下，我们旨在学习一种手势动作策略（ 3DGesPolicy），驱动一个三维人体模型模板执行流畅、自然、富有表现力且与语音同步的身体动作和面部表情。

给定整体手势 $g^{1:N} = (g^1, g^2, \ldots\ldots, g^N) \in \mathbb{R}^{55\times6+100+4+3}$，其中包括采用 Rot6D 表示的 55 个姿态关节、 $\mathbb{R}^{100}$ FLAME参数、 $\mathbb{R}^4$足部接触标签，以及连续 N帧的 $\mathbb{R}^3$ 全局平移，我们将时间微分手势偏差定义为动作 $a_0^{1:N} = (a_0^1, a_0^2, \ldots\ldots, a_0^N) \in \mathbb{R}^{N\times D_a}$，其中 D_a表示动作维度。该

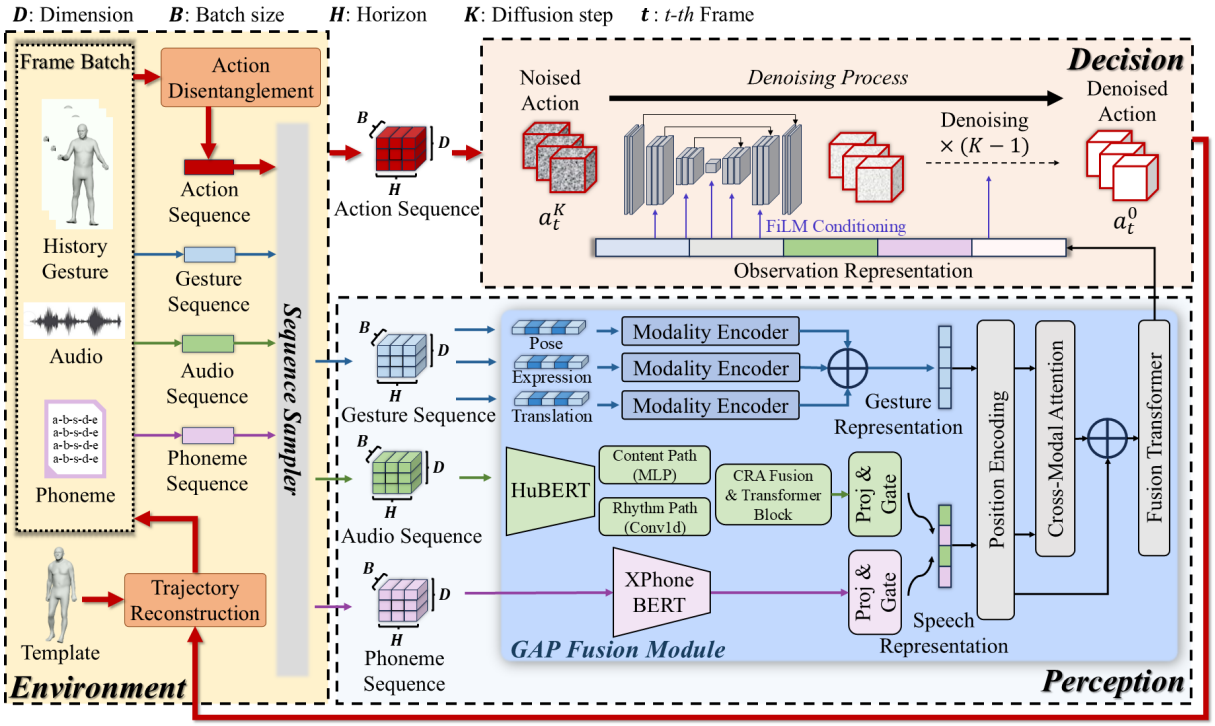

图2：3DGesPolicy架构概览。我们的架构首先将历史手势解耦为动作序列，然后由感知模块将手势、音频和音素序列编码为观测表征，作为带有GAP融合模块的决策模块的条件信息，通过去噪生成动作，并驱动模板输出后续的整体手势。

动作空间能够自然地确保运动平滑性，并缓解绝对位置预测中常见的抖动问题。在以音频上下文 $s^{1:N}$、音素信号 $m^{1:N}$和历史手势状态 g_h为条件的情况下，高斯噪声动作序列 $a_k^{1:N}$会被迭代地细化为干净的动作序列 $a_0^{1:N}$，其中 $k \in \{1, \ldots\ldots, K\}$ 表示扩散步。因此，我们框架的目标是学习一种扩散策略，根据包含音频、音素和当前手势的多模态观测 $O \in \{O_s, O_m, O_g\}$ a_0 预测最优动作序列。该问题可表述为：

$$a_0^{1:N} = \text{3DGesPolicy}(a_K^{1:N}, O, K), \tag{1}$$

其中，K 表示扩散步骤的总数。利用预测得到的动作序列 $a_0^{1:N}$ 和模型手势模板g_{temp}，第 n帧的整体手势g_0^n 表示为：

$$g_0^n = g_{temp} + \sum_{i=1}^{n} a_0^i \cdot \Delta t, \quad n \in \{1, \ldots, N\}, \tag{2}$$

其中，Δt 表示时间步长。通过预测动作，我们的方法能够生成连续、平滑且动态的整体手势轨迹。

3.2 架构

3.2.1 概述。 我们将框架设计为一个用于机器人学的闭环策略学习系统，由三个模块组成：环境、感知和决策。如图 2 所示，环境模块通过解耦原始手势序列来构建动作空间，并将手势、音频和音素等多模态输入采样为固定长度的时间窗口，为策略学习奠定稳定基础。感知模块通过我们提出的 GAP 融合模块，将历史手势序列以及同步音频和音素序列编码为全面的观测表示。决策模块在条件扩散过程中，对于手势的动作策略训练发挥着重要作用；它在感知模块提供的观测表示的引导下，迭代地将含噪动作序列细化为干净的手势轨迹。最后，将动作累积到 3D人体模型模板上，以重建整体手势。这种架构设计使我们的模型能够学习流畅且富有表现力的手势控制策略，自然地遵循语音节奏和语义。

3.2.2 环境。 环境模块构建动作空间，并准备多模态数据，以实现高效的策略训练。通过将原始手势序列解耦为可学习的时间变化，并将可变长度的输入整理为固定时域批次，该模块确保策略在简洁且信息丰富的上下文中进行训练，从而提升运动一致性和训练稳定性。

动作形式化： 我们方法的一个关键组成部分是对动作序列进行形式化，使其能够有效捕捉涵盖身体动作和面部

表情的整体手势动态。我们将动作定义为时间差分表示，用于编码连续帧之间的运动学变化，并采用自适应缩放机制。这种以运动为中心的形式化方式使我们的模型能够学习平滑的手势运动轨迹，同时自然地保持时间连续性。

给定一个整体手势序列 $g^{1:N} = \{g^1, g^2, \ldots\ldots g^N\} \in \mathbb{R}^{N \times D_g}$，我们首先计算基本的时间位移算子：

$$\mathcal{D}^n_{temporal} = g^{n+1} - g^n, n \in [0, N-1] \tag{3}$$

该算子捕获原始的帧间运动向量，为动作构建提供基础。为了提高对动作的敏感度和训练稳定性，我们采用一种自适应缩放机制，根据动作强度调节动作幅度。缩放因子定义如下：

$$\Lambda_{adaptive} = \exp(-\beta \cdot \|\mathcal{D}^n_{temporal}\|^2_F) \tag{4}$$

指数函数提供了基于自然动作的加权方式，其中 β 控制敏感度的衰减率。这确保细微动作获得较高的缩放权重，同时防止大幅动作主导学习过程。$\|\cdot\|_F$ 表示弗罗贝尼乌斯范数。最终的动作形式化表示为：

$$a^n = \varepsilon_{scaling} \cdot \Lambda_{adaptive} \cdot \mathcal{D}^n_{temporal} \tag{5}$$

其中$\varepsilon_{scaling}$ 为基础缩放参数。这种微分表示在轨迹重建过程中通过时间累积自然地保持动作连续性。

计算动作后，同步序列会被划分为称为时间视界的定时长片段 H，以便序列采样器获得局部一致的时间上下文。这些小片段为感知模块准备多模态数据，以便其编码历史上下文，并为决策模块细化手势轨迹。

3.2.3 感知。感知模块将观测视界内的多模态观测转换为统一的表示 O，作为动作生成的条件引导。通过整合手势、音频和音素模态的特征，该模块确保语音内容与全身动作之间具有强烈的音素级同步性。

*GAP*融合模块：感知模块的核心是我们提出的手势-音频-音素（GAP）融合机制，它通过层次编码和跨模态注意力实现细粒度的多模态对齐。该模块通过专用编码器处理每种模态：手势组件 O_g（姿态、表情、平移）由各模态专用的 Transformer 编码；音频特征 O_s 由预训练的 HuBERT 模型 [46] 提取，并通过采用受 [4],启发的内容-韵律注意力（CRA）的双路径编码器进行处理；音素特征 O_m 则由预训练的 XPhoneBERT 模型[47]提取。随后，通过自适应门控机制对齐音频和音素特征，再通过跨模态注意力将融合后的语音表示与手势特征整合起来。这一设计使模型能够捕捉语义和韵律线索，并在声学语言信号与运动动态之间建立明确的对应关系。最终的多模态表示$O = \{O_g, O_s, O_m\}$提供丰富的条件信号，用以引导决策模块生成具有精确音素级表现力和时间连贯性的手势动作。详细的架构说明见补充材料。

3.2.4 决策。条件去噪扩散模型是学习动作策略的核心，其依据为 [31]。网络从长度为 H的高斯噪声序列 a_K开始，经过 K 个扩散步骤，逐步对潜在动作序列进行去噪，从而恢复干净的动作 a_0。该反向扩散过程在数学上表示为：

$$a_{k-1} = \alpha_k(a_k - \gamma_k \epsilon_\theta(a_k, k, O)) + \sigma_k \mathcal{N}(0, \mathrm{I}), \tag{6}$$

其中，ϵ_θ表示噪声预测网络，而 α_k、 γ_k和 σ_k是依赖噪声调度的系数，用于控制噪声去除速率。$\mathrm{N}(0, \mathrm{I})$表示生成过程中添加的随机噪声。

经过 K 次迭代后，我们得到预测动作轨迹 $a_0^{1:N}$。为了恢复最终的整体手势，必须将这些微分动作积分到手势状态空间中。第 n 帧的整体状态 g_0^n重建如下：

$$g_0^n = g_{temp} + \sum_{i=1}^{n} \mathcal{G}(a_0^i), \tag{7}$$

其中，g_{temp} 表示模板手势模型。$\mathrm{G}(\cdot)$ 是逆变换函数，其表达式为：

$$\mathcal{G}(a^i) = a^i / (\varepsilon_{scaling} \cdot \Lambda_{adaptive}) \tag{8}$$

3.2.5 训练目标。为确保合成的整体手势既具有概率多样性，又满足物理一致性，我们采用复合目标函数，联合优化策略训练稳定性、手势重建准确性和动态速度保真度。

扩散损失：在每个训练步骤 k中，向真实动作序列 a_0添加随机噪声 $\epsilon^k \sim \mathrm{N}(0, \mathrm{I})$，以获得含噪输入 a_k。去噪网络 ϵ_θ以观测表示 O为条件进行训练，以预测添加的噪声。扩散损失定义为：

$$\mathcal{L}_{\text{diff}} = \mathrm{MSE}(\epsilon^k, \epsilon_\theta(\overline{\alpha_k} a_0 + \overline{\beta_k} \epsilon^k, k, O)), \tag{9}$$

其中，$\overline{\alpha_k}$ 和 $\overline{\beta_k}$为噪声调度。

重建损失：为防止长序列中的累积漂移，我们通过计算预测手势 $g^{1:N}$ 与真实序列 $\hat{}g^{1:N}$之间的欧几里得距离，显式监督手势轨迹：

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_n[\frac{1}{N} \sum_{n=1}^{N} \|g_0^n - \hat{g}_0^n\|^2], \tag{10}$$

其中， g_0^n 和 $\hat{}g_0^n$ 分别表示第 n帧的预测手势和真实手势。

速度损失：为增强运动平滑性和动态保真度，我们采用速度损失来捕捉时间运动模式：

$$\mathcal{L}_{\text{vel}} = \mathbb{E}_n[\frac{1}{N} \sum_{n=1}^{N} \|(g_0^{n-1} - g_0^n) - (\hat{g}_0^{n-1} - \hat{g}_0^n)\|^2]. \tag{11}$$

表1：3DGesPolicy 与 BEAT2 数据集上最先进方法的定量比较。3DGesPolicy 在所有指标上均表现出色。FGD、DIV、MSE 和 LVD 的单位分别为 10^{-1}、1、10^{-8} 和 10^{-5}。

方法	模态 表情	模态 手势	指标 FGD ↓	指标 DIV ↑	指标 MSE ↓	指标 LVD ↓
FaceFormer [48]	✓	✗	-	-	7.787	7.593
CodeTalker [36]	✓	✗	-	-	8.026	7.766
DisCo [41]	✗	✓	9.417	9.912	-	-
CaMN [1]	✗	✓	6.644	10.86	-	-
DSG [49]	✗	✓	8.811	11.49	-	-
Habible 等人 [50]	✓	✓	9.040	8.213	8.614	8.043
TalkSHOW [5]	✓	✓	6.209	13.47	7.791	7.771
DiffSHEG [8]	✓	✓	8.986	11.91	7.665	8.673
EMAGE [4]	✓	✓	5.512	13.06	7.680	7.556
MambaTalk [9]	✓	✓	5.366	13.05	**6.289**	6.897
3DGesPolicy（我们的方法）	✓	✓	**3.190**	**14.03**	7.020	**2.200**

总目标： 最终训练目标是三个组成部分的加权和：

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{diff}} + \lambda_2 \mathcal{L}_{\text{rec}} + \lambda_3 \mathcal{L}_{\text{vel}}, \tag{12}$$

其中， λ_1、 λ_2 和 λ_3 用于平衡各损失项的相对重要性。

3.2.6 实现细节。 我们使用 DDIM 调度器实现 3DGesPolicy，训练步数为 100 步，推理步数为 10 步。模型使用 AdamW 优化器进行训练，共训练 200 个 epoch，批量大小为 32，初始学习率为 1×10^{-5}。各损失项采用经验确定的权重进行平衡：$\lambda_1 = 10$、$\lambda_2 = 1$ 和 $\lambda_3 = 1$。所有实验均在单张 NVIDIA RTX 6000 Ada Generation GPU 上进行，训练耗时 28 小时 1 分 20 秒。

4 实验

我们开展了全面的实验，以验证 3DGesPolicy 框架的有效性。我们在 BEAT2 数据集上进行定量和定性评估，将我们的方法与身体手势生成方法 [1, 41, 49],、面部生成方法 [36, 48] 以及最先进的整体生成方法 [4, 5, 8, 9, 50] 进行比较，并采用多项指标进行评估。为进一步根据视听用户体验评估我们的方法，我们设计了一项用户研究，通过成对比较评估自然度、同步性和情感表现力。此外，我们还进行了消融研究，以分析框架中各关键组件的贡献。更多结果见**补充材料**。

4.1 实验设置

4.1.1 基线。我们将 3DGesPolicy 与身体手势生成方法 [1, 41, 49] 和面部生成方法 [36,48], 进行比较，随后主要与最先进的整体生成方法 [4, 5, 8, 9, 50] 进行比较。所有方法均在相同的实验设置下进行评估，以确保比较公平有效。

4.1.2 数据集。我们在 [4] 介绍的 BEAT2 数据集上训练并评估模型。BEAT2 在原始 BEAT 数据集 [1] 的基础上扩展而来，包含来自 30 位说话者的 76 小时高质量动作捕捉数据，并将其标准化为网格表示，同时配有同步的音频、文本和帧级语义标注 [32, 33]。遵循 [9], 确立的标准协议，我们采用 BEAT2 标准子集，并按 85%/7.5%/7.5% 的比例划分训练集、验证集和测试集，以确保与先前工作的比较公平 [4, 9]。

4.1.3 评估指标。为全面评估生成的面部动画质量，我们采用四项指标：Fréchet 手势距离（FGD） [34] 用于衡量生成手势与真实值手势之间的分布相似性，以评估其真实感。多样性（DIV） [35] 通过计算不同手势片段之间的平均 L1 距离来衡量手势变化程度。均方误差（MSE） [36]通过计算预测顶点位置与真实值顶点位置之间的平均平方距离，量化面部顶点的位置精度。唇部顶点距离（LVD） [5]用于衡量生成的面部顶点与真实值面部顶点之间的 L1 差异，为面部运动质量提供补充评估，且对唇部运动精度尤为敏感。这些指标的计算公式列于补充材料中。

4.2 定量评估

我们在表1中将3DGesPolicy与最先进方法进行了比较，包括仅表情方法 [36, 48],仅手势方法 [1, 41, 49], 和整体生成方法 [4, 5, 8, 9, 50] 。与基线方法相比，我们的方法取得了更优的性能。在FGD和LVD上的显著优势验证了3DGesPolicy在保留稳定且语义上有意义的动作，以及通过轨迹控制建立语音信号与唇部运动之间精确的音素级同步方面的有效性。

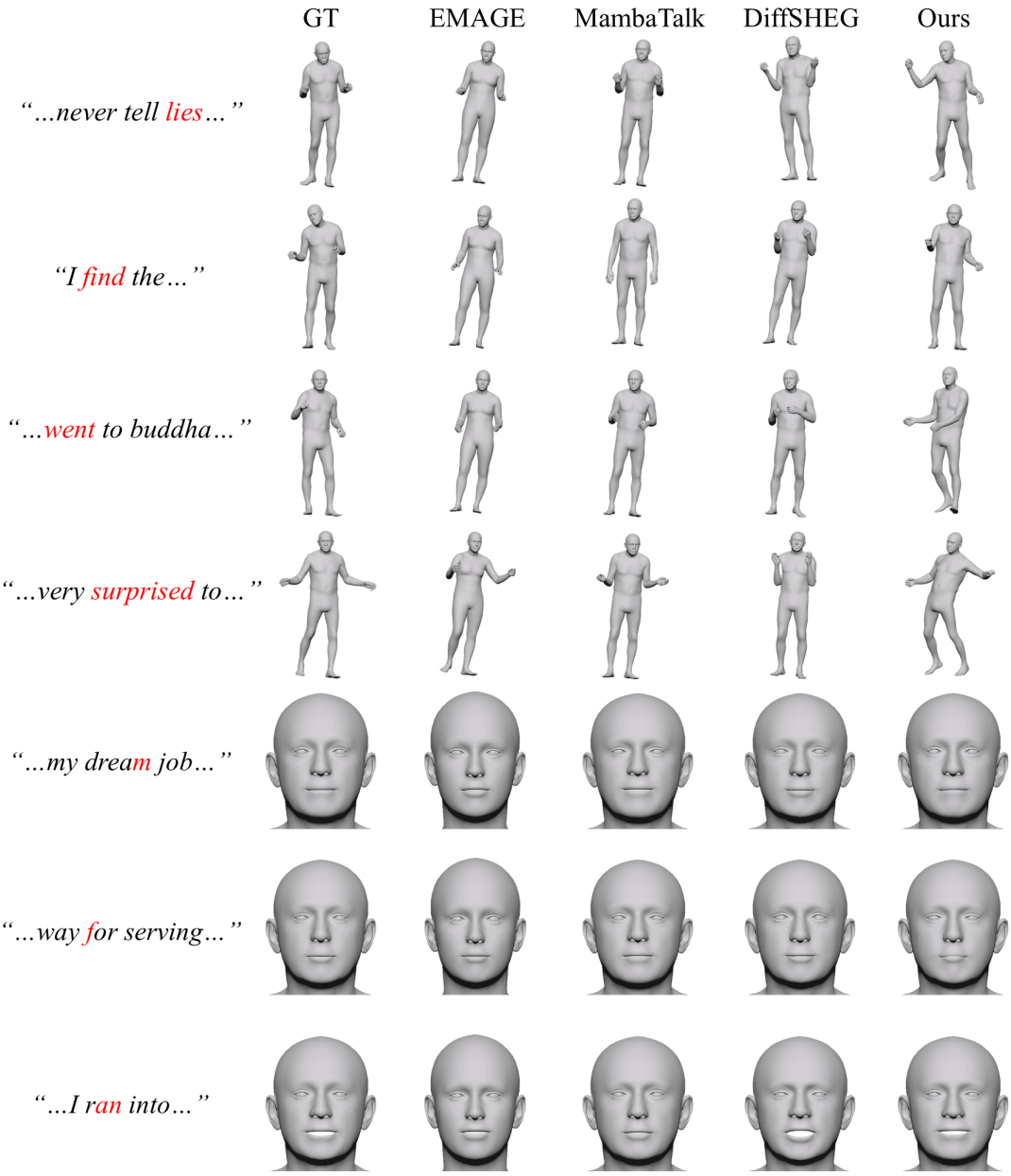

图3：BEAT2 数据集上的整体手势生成定性比较。与基线方法相比，我们的方法生成的身体动作表现力更强，也更贴合语音，同时避免了静态或语义无关的手势；此外，它实现了更出色的唇动与语音同步，并能准确捕捉细微的语音细节。

4.3 定性评估

我们在BEAT2模板上对整体手势生成进行了定性比较。图3的上半部分显示，我们的3DGesPolicy能够生成自然且多样的手势，并与语音语义高度一致。通过将任务重新定义为轨迹控制，我们的方法生成了稳定且扎根于三维空间的动作，显著减少了不自然的滑动与现有基线方法相比，伪影更少。此外，我们的策略能够通过有意义的手势动作有效捕捉语音内容，同时避免过度或无意义的动作。例如，对于“went”和“find”等词语，我们的方法会生成恰当的指向手势。对于“.....非常惊讶.....”这样的短语，我们的方法还会通过动态姿势表现情绪强度，而基线方法则相对静止。图3下半部分展示了我们出色的唇动与语音同步表现。我们的方法

表2：用户研究结果。自然度、同步性和情感的偏好百分比。更高的自然度以及较高的情感和同步性评分表明，我们的方法在实现精准语义同步和富有表现力的动作的同时，确保了更强的真实感。

方法	自然度（%）↑		同步性（%）↑		情感（%）↑	
	Ours	对比方法	Ours	对比方法	Ours	对比方法
我们的方法 对比 DiffSHEG [8]	**91.67**	8.33	**58.33**	41.67	**66.67**	33.33
我们的方法 对比 EMAGE [4]	**83.33**	16.67	**83.33**	16.67	**58.33**	41.67
我们的方法 对比 MambaTalk [9]	**91.67**	8.33	50.00	50.00	50.00	50.00
我们的方法 对比 真实值	25.00	**75.00**	25.00	**75.00**	0.00	**100.00**

表 3：针对动作、GAP融合模块和音素组件的消融研究。动作的构造对于自然动作至关重要，而 GAP 模块和音素特征对于准确的面部动画和唇部同步至关重要。

模型配置	FGD ↓	MSE ↓	LVD ↓
无动作	23.563 ± 0.481	82.530 ± 0.269	18.030 ± 0.405
Sim-GAP 融合	3.713 ± 0.138	12.847 ± 0.177	2.500 ± 0.098
无音素	4.009 ± 0.125	16.698 ± 0.242	2.400 ± 0.076
完整模型	**3.190** ± 0.119	**7.020** ± 0.164	**2.200** ± 0.088

能够准确再现细微的语音特征，例如发出“m”音素时完全闭唇，以及发出“f”音素时的唇齿音发音。EMAGE 和 MambaTalk 等竞争方法往往无法捕捉这些细节，导致发音不完整。这些视觉结果表明，GAP融合模块实现了精确的音素级对齐，这与我们在 LVD 指标上的定量提升一致。

4.4 用户研究

我们通过 A/B 测试开展了一项用户研究，将我们的 3DGesPolicy 与 DiffSHEG [8], EMAGE [4], 和 MambaTalk [9]进行比较。我们请参与者根据三个维度评估视频对：自然度、语音同步性和情感表现力。我们从 30 位参与者那里收集了 360 份有效评估。如表2所示，我们的方法在几乎所有指标上都优于其他方法，在自然度方面的得分尤其高。这些结果证实，与以往方法相比，我们基于动作的轨迹控制能够生成更接近三维空间中真实人体动态的运动模式。此外，同步性和情感方面的出色表现也验证了我们的方法能够实现更优的视听对齐和情感表达。

4.5 消融研究

我们在 BEAT2 数据集上开展消融研究，以验证三个核心组件：动作形式化、GAP融合模块和音素整合。我们使用 FGD、MSE 和 LVD 进行性能评估，其单位分别为 10^{-1}、10^{-8} 和 10^{-5}。结果见表3。

4.5.1 动作构式。我们将基于动作的轨迹控制与一种采用绝对姿态进行直接手势回归的变体进行比较。结果显示，所有指标均大幅下降，尤其是 FGD 和 MSE。这验证了我们的核心贡献：基于动作的构式不仅通过时间累积确保动作连续性，还使模型能够通过轨迹控制学习遵循语音动态的运动模式。

4.5.2 GAP 融合模块。 为评估我们的融合架构，我们测试了一个简化版本（Sim-GAP），该版本用简单拼接取代自适应门控和跨模态注意力。我们观察到，面部指标 MSE 和 LVD 显著下降，这表明简单拼接无法实现细粒度对齐。这验证了我们的 GAP 模块能够提供精确音素级同步所需的结构化表示。

4.5.3 音素整合。 最后，我们通过完全移除音素特征，仅以手势和音频观测为条件，评估音素级语义信息的作用。结果显示，所有指标均持续下降。这验证了音素能够提供关键的语言时序线索，对于实现准确的唇形同步以及语音与手势之间的语义对齐至关重要。

5 结论

我们提出了 3DGesPolicy，这是一种新颖的轨迹控制框架，将整体共语手势生成重新定义为基于动作的扩散策略问题。通过将帧间变化建模为统一动作，并引入用于音素级多模态对齐的 GAP 融合模块，我们的方法能够生成自然的手势，并实现精准的唇动与语音同步。在 BEAT2 数据集上的定量和定性结果表明，与最先进的方法相比，我们的方法在动作质量和表现力方面均有显著提升。对三维人体手势关键点进行轨迹控制，或可通过精细手部细节和手指动作进一步改进我们的方法。未来的工作将聚焦于这一机制

以提升手部动作的还原度，同时保持整体协调性。

6 补充材料

6.1 概述

在本补充材料中，我们首先在第 6.2 节提供 BEAT2 的更多实验结果，并将我们的方法与其他最先进方法进行比较。在第 6.3 节中，我们介绍各项指标的计算公式。在第 6.4 节中，为确保模型可复现性，我们详细介绍序列采样器的设置以及 GAP 融合模块的网络架构。最后，我们在第 6.5 节中介绍用户研究的详细信息。

6.2 更多实验结果

由于实验部分篇幅有限，我们在图4中展示了 BEAT2 数据集的更多定性比较结果。这些补充结果进一步展现了我们的方法在生成自然、生动且语义丰富的手势动作方面的优越性能。

上半部分展示了全身手势合成的更多对比结果。我们的方法始终能够生成语义丰富、情感表现力强的身体动作，并与语音内容紧密对齐。在需要强调的语境中，例如单词“it”和“anger”，我们的方法能够生成清晰的手部指向手势，而基线方法则无法捕捉这些有意义的表情。同样，在表达负面情绪或列举事物时，例如“house”和“parents”，其他方法生成的动作表现力不如我们的结果。这些观察验证了我们的轨迹控制框架能够有效理解语音语义，并合成自然体现说话者交际意图的运动模式，从而生成语义连贯且情感恰当的身体手势。

下半部分展示了面部表情合成的更多结果，体现了我们的方法在音素级唇部同步方面的优势。在快速的多音节发音过程中，例如“fantastic”中的“s”音素，我们的方法能够准确再现与真实值一致的唇形，而其他方法生成的嘴部动作过于平滑，无法捕捉各个音节的发音。同样，对于“watching”中的“w”音素和“kind”中的“in”音，我们的方法生成的唇部构型与真实值高度吻合，发音准确度也明显高于基线方法。这些结果证实了我们在GAP融合模块中采用的音素级架构设计是有效的，能够实现更精确的面部运动合成和细粒度视听对齐。

6.3 评估指标

具体的指标计算方法如下：

i) Fréchet 手势距离 *(FGD) [34]*：该指标用于评估真实值与生成的身体手势之间的贡献距离：

$$\mathrm{FGD}(\mathbf{g},\hat{\mathbf{g}}) = \|\mu_r - \mu_g\|^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right), \tag{13}$$

其中，μ_r 和 Σ_r 分别表示从真实人类手势 $\mathbf{g}$中提取的潜在特征分布 $\mathbf{z}_r$ 的均值和协方差矩阵，而 μ_g 和 Σ_g 则表示合成手势 $\hat{\mathbf{g}}$ 的潜在特征分布 $\mathbf{z}_g$ 的相应统计量。遵循先前的工作 [4, 9], , 我们使用一个预训练自编码器来提取这些潜在特征，该自编码器由基于骨架卷积神经网络（SKCNN）的编码器和基于全卷积神经网络的解码器组成。FGD 值越低，表示生成的手势越接近真实人类动作的分布。

*ii) L1*多样性（*DIV*）*[35]*：该指标通过计算多个手势片段之间的平均 L1 距离，量化生成的手势序列的变化程度。多样性分数越高，表明生成动作的变化越大、自然度越高。对 N 个动作片段的计算公式如下：

$$\mathrm{DIV} = \frac{1}{2N(N-1)}\sum_{i=1}^{N}\sum_{j=1}^{N}\left\|p_t^i - \hat{p}_t^j\right\|_1, \tag{14}$$

其中，p_t^i表示第 t 帧的关节位置，i表示第几个动作片段。我们在整个测试数据集上评估多样性。请注意，计算关节位置时，我们将平移设为零，以确保 L1多样性仅关注局部运动模式。

iii) 均方误差（*MSE*）*[36]*：该指标通过计算预测顶点位置与真实顶点位置之间平方欧氏距离的均值，衡量生成面部标志点的位置准确性：

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(f_i - \hat{f}_i)^2, \tag{15}$$

其中，n 表示面部顶点总数，f_i 表示第 i 个顶点的ˆ真实值位置，f_i 表示其预测位置。对所有顶点求和，以获得平均位置误差。MSE 值越低，面部运动重建的准确性越高。

iv) 唇部顶点距离（*LVD*）*[5]*：该指标通过测量唇部区域顶点的速度差，评估语音与面部表情之间的同步性。它计算生成的面部顶点与真实值面部顶点速度之间的平均 L1 距离：

$$\mathrm{LVD} = \frac{1}{n}\sum_{i=1}^{n}\left\|f_i' - \hat{f}_i'\right\|_1, \tag{16}$$

其中，n 表示唇部区域中的顶点数，f_{ri} 表示 i号顶点的真实值速度，$f_{\hat{r}i}$ 表示生成的面部表情中对应顶点的速度。对所有唇部顶点求和，以计算平均绝对速度差。LVD 值越低，表示视听同步效果越好，唇部运动也越自然。

6.4 模型可复现性

6.4.1 序列采样器详情。为在可控的计算预算内实现高效的策略学习，序列采样器会将长序列划分为称为观测时域的固定时长窗口 H。经过大量实验评估后，我们设置 $H = 16$ 以获得最佳性能。每个采样批次都包含一个长度为 $N_{obs} = 8$ 的观测窗口，用于编码

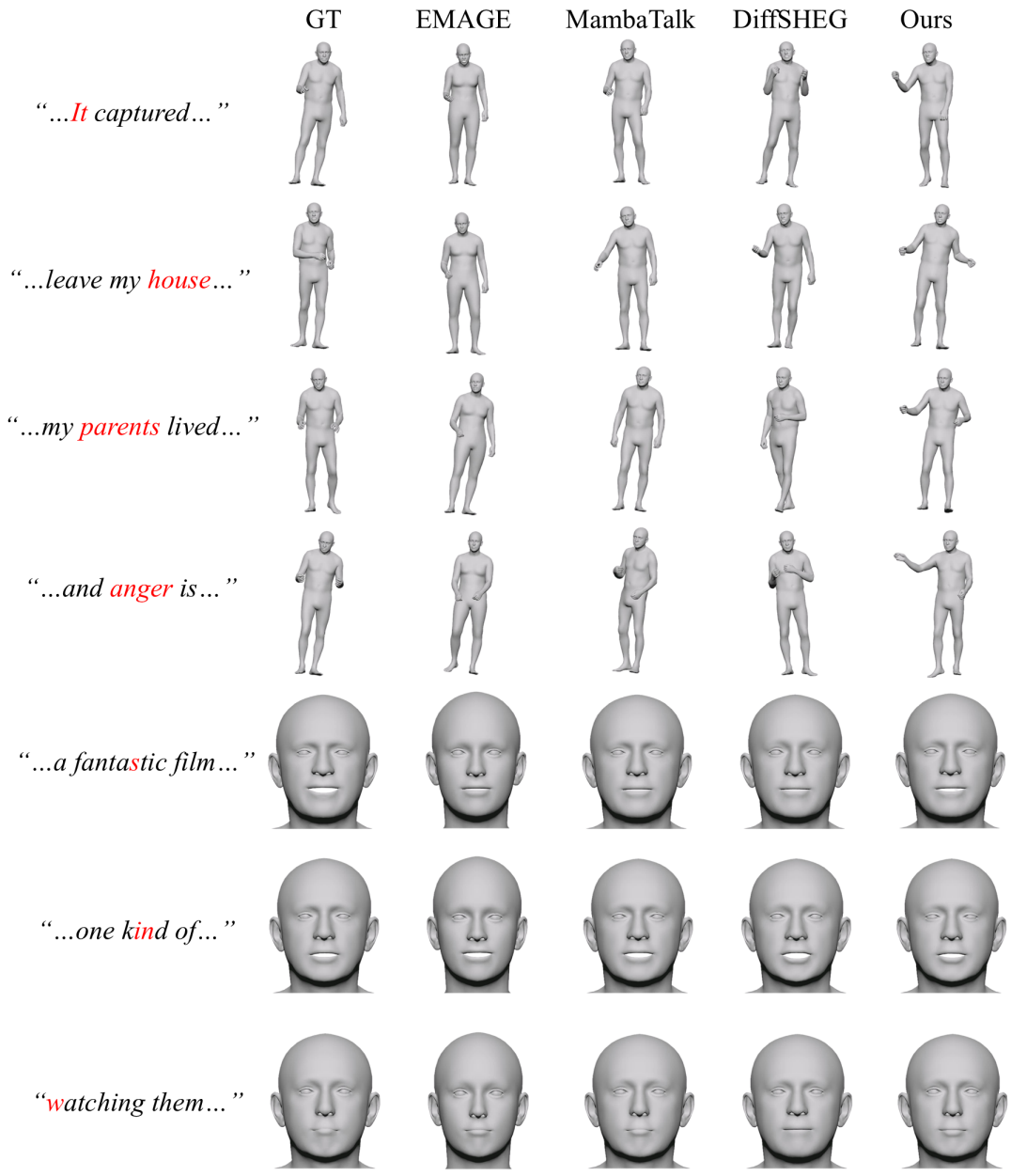

图4：　BEAT2数据集上手势生成和面部生成的补充实验结果

历史上下文，以及一个长度为 $N_{act} = 8$ 的动作预测窗口，用于生成未来手势。这种划分确保策略能够从局部一致的时间上下文中学习，同时保留充足的历史信息以维持动作连续性。采样器会同时处理手势序列 $g^{1:N}$、音频序列 $s^{1:N}$、音素序列$m^{1:N}$以及计算得到的动作序列 $a^{1:N}$，以保持时间同步性。该机制使模型能够处理可变长度的输入，并将其整理为适用于 GPU 加速的标准化批次结构。

*6.4.2 GAP*融合模块架构细节。 GAP融合模块采用分层编码策略，以实现手势、音频和音素模态之间的细粒度对齐。本节提供完整的架构规范，以确保结果可复现。

i) 模态专用编码器：每种模态在融合前，都会先经过专用编码器处理。对于手势组件，我们使用三个独立的模态编码器模块，分别处理姿态（55 × 3维）、表情（100 维）和平移（3 维）。每个模态编码器都由一个两层多层感知机组成

投影层，后接一个用于时间建模的 Transformer 块。音频编码器采用受 [4] 启发的双路径架构。从预训练的 HuBERT 模型 [46], 中提取 1024 维特征后，我们通过并行的内容路径和节奏路径对其进行处理。内容路径使用两层多层感知机，而节奏路径采用核大小为 5 和 3 的一维卷积来捕捉时间模式。随后，这两条路径通过内容-韵律注意力（CRA）进行融合，该机制计算注意力权重，以自适应地结合内容和节奏信息。在音素编码方面，我们利用预训练的 XPhoneBERT 模型 [47] 获取 768 维的上下文化音素表征。这些特征经过音素编码器处理，该编码器由时间投影层和核大小为 3 和 5 的一维卷积组成，用于捕捉局部音素模式，随后再通过一个用于序列建模的 Transformer 块。所有编码器都会将各自的输入投影到统一的 256 维潜在空间。

ii) 音频-音素对齐：为建立音频与音素模态之间的时间连贯性，我们采用一个带有自适应门控机制的音频-音素对齐模块。该模块通过独立的线性层投影音频特征 O_s 和音素特征 O_m，随后将两者拼接，并将结果输入带有 Sigmoid 激活的两层多层感知机，以计算门控权重 $\alpha \in [0,1]$。对齐后的语音特征表示为 $O_{speech} = \alpha \cdot O_s + (1-\alpha) \cdot O_m$，其中门控根据局部时间上下文自适应地平衡声学与语言信息。

iii) 跨模态融合：编码后，我们通过逐元素相加，将姿态、表情和平移中的手势特征组合起来，形成统一的运动特征 $O_{motion} = O_{pose} + O_{exp} + O_{trans}$。

运动特征和语音特征均通过添加正弦位置编码来保留时间顺序信息。随后，我们应用跨模态注意力，其中语音特征作为查询，运动特征作为键和值。该注意力机制在声学-语言信号与运动动态之间建立明确的对应关系。输出通过残差连接与语音特征融合：$O_{fused} = \text{CrossAttention}(O_{speech}, O_{motion}) + O_{speech}$。最后，两个堆叠的 TransformerBlock 对融合后的特征进行细化，并通过一个三层多层感知机将其投影为维度为 1024 的最终观测表示。

iv) 实现细节：所有 Transformer 块均使用 8 个注意力头，Dropout 率为 0.1。多层感知机的隐藏维度设为 512（潜在维度的两倍）。除使用 Sigmoid 激活的门控机制外，所有中间层均采用 ReLU 激活。按照预归一化架构，在每个注意力和前馈操作之前应用层归一化。

6.5 用户研究

用户研究界面如图5所示。参与者需要并排比较，并根据个人偏好选择更好的动画。每组视频对均从示例中随机选取，并随机切换我们模型与竞争模型生成的视频所在的位置。每位参与者的完成时间约为 20 分钟，共需观看 12 组视频对，每组回答 3 个问题。参与者需要回答的问题是

呈现给参与者的问题如下：(1) 关于自然度和流畅度，哪段动画中的手势看起来更像真人，且没有明显的扭曲、突兀的动作变化或抖动？(2) 关于同步性，哪段动画中的手势和面部表情与语音节奏及音频更协调，整体表现更和谐？(3) 关于表现力，哪段动画中的手势和面部表情情感更丰富，并能通过动作和面部特征有效传达情绪？

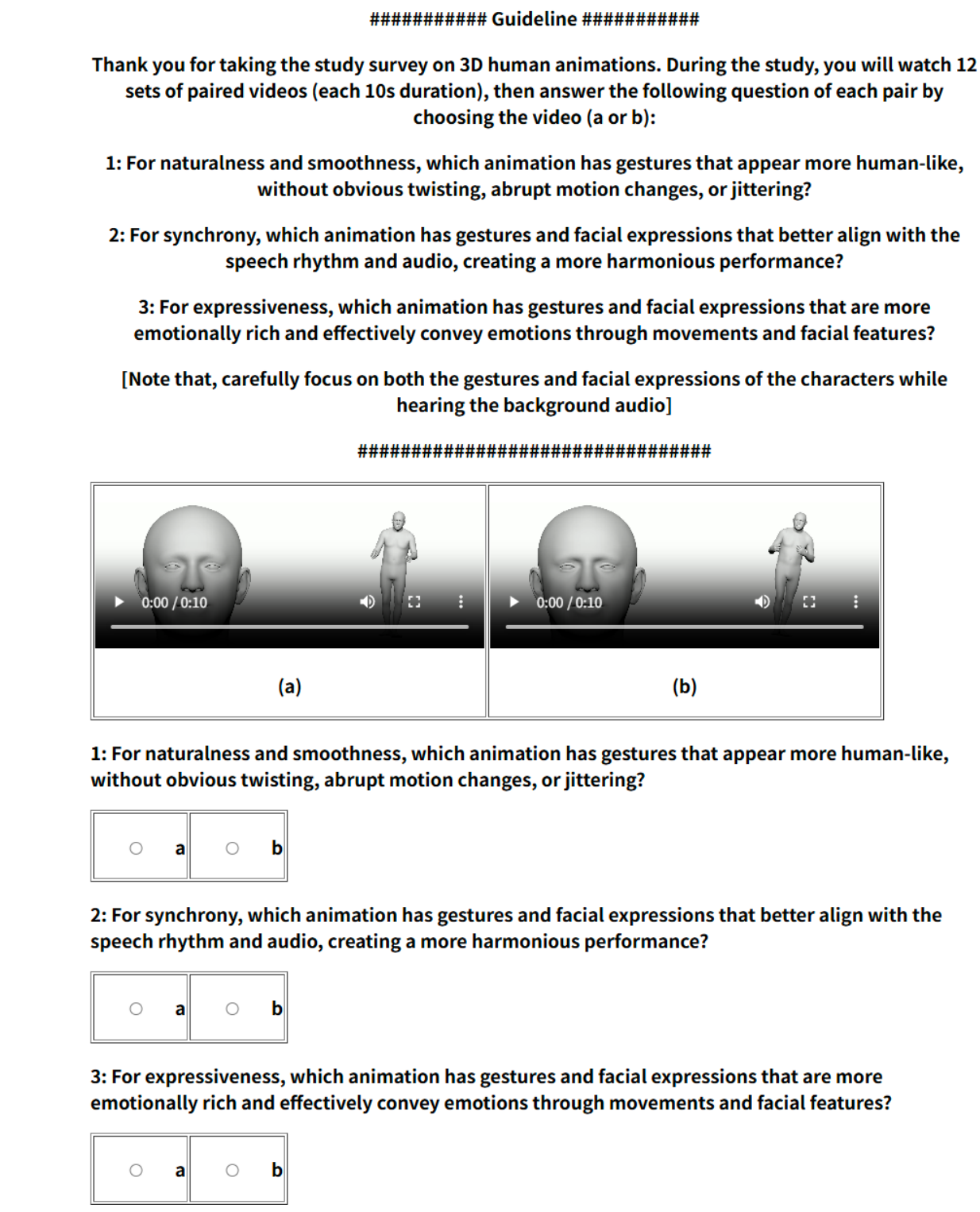

图5：设计的用户研究界面。每位参与者需要回答12组视频对；受篇幅限制，此处仅展示一组视频对。

参考文献

[1] H. Liu、Z. Zhu、N. Iwamoto、Y. Peng、Z. Li、Y. Zhou、E. Bozkurt 和 B. Zheng，《Beat：用于对话手势合成的大规模语义与情感多模态数据集》，载于《欧洲计算机视觉会议》。Springer，2022年，第612–630页。[2] X. Liu、Q. Wu、H. Zhou、Y. Du、W. Wu、D. Lin 和 Z. Liu，《音频驱动的伴随语音手势视频生成》，《神经信息处理系统进展》，第35卷，第21 386–21399页，2022年。[3] D. P. Kingma 和 M. Welling，《自编码变分贝叶斯》，arXiv预印本 arXiv:1312.6114，2013年。[4] H. Liu、Z. Zhu、G. Becherini、Y. Peng、M. Su、Y. Zhou、X. Zhe、N. Iwamoto、B. Zheng 和 M. J. Black，《EMAGE：通过表达性掩码音频手势建模实现统一的整体共语手势生成》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2024年，第1144–1154页。[5] H. Yi、H. Liang、Y. Liu、Q. Cao、Y. Wen、T. Bolkart、D. Tao 和 M. J. Black，《从语音生成整体式三维人体运动》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2023年，第469–480页。[6] S. Alexanderson、R. Nagy、J. Beskow 和 G. E. Henter，《聆听、去噪、行动！基于扩散模型的音频驱动运动合成》，《ACM图形学汇刊》（TOG），第42卷，第4期，第1–20页，2023年。[7] Y. Zhi、X. Cun、X. Chen、X. Shen、W. Guo、S. Huang 和 S. Gao，《LivelySpeaker：迈向语义感知的伴随语音手势生成》，载于《IEEE/CVF国际计算机视觉会议论文集》，2023年，第20 807–20817页。[8] J. Chen、Y. Liu、J. Wang、A. Zeng、Y. Li 和 Q. Chen，《DiffSHEG：一种基于扩散的方法，用于实时语音驱动的整体式三维表情与手势生成》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2024年，第7352–7361页。[9] Z. Xu、Y. Lin、H. Han、S. Yang、R. Li、Y. Zhang 和 X. Li，《MambaTalk：利用选择性状态空间模型实现高效的整体手势合成》，《神经信息处理系统进展》，第37卷，第20 055–20080页，2024年。[10] Z. Geng、C. Wang、Y. Wei、Z. Liu、H. Li 和 H. Hu，《作为组合式词元的人体姿态》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2023年，第660–671页。[11] X. Sha、L. Zhang、T. Mashita、N. Chiba 和 Y. Uranishi，《3DFacePolicy：基于动作控制的音频驱动三维面部动画》，2025年。 [在线]。可访问：https://arxiv.org/abs/ 2409 10848 。X. Zhou、F. Li、M. Chen、Y. Zhou、P. Wan[12]、D. Zhang、Y. Jin、Z. Fan、H. Liu 和 J. He，《ExGes：面向音频驱动手势合成的表达性人体运动检索与调制》，arXiv预印本 arXiv:2503.06499，2025年。[13] C. Chi、Z. Xu、S. Feng、E. Cousineau、Y. Du、B. Burchfiel、R. Tedrake 和 S. Song，《扩散策略：通过动作扩散进行视觉运动策略学习》，《国际机器人研究杂志》，第44卷，第10–11期，第1684–1704页，2025年。[14] L. Zhang、P. Ratsamee、B. Wang、Z. Luo、Y. Uranishi、M. Higashida 和 H. Takemura，《全景感知图像到图像转换》，载于《IEEE/CVF冬季计算机视觉应用会议论文集》，2023年，第259–268页。[15] L. Zhang、P. Ratsamee、Z. Luo、Y. Uranishi、M. Higashida 和 H. Takemura，《用于增强目标识别与视觉里程计的全景级图像到图像转换》，《IEEE视频技术电路与系统汇刊》，第34卷，第2期，第938–954页，2023年。[16] L. Zhang、P. Ratsamee、Y. Uranishi、M. Higashida 和 H. Takemura，《将热图像转换为彩色图像以增强热视觉的视觉里程计》，载于《2022年IEEE安全、防护与救援机器人国际研讨会（SSRR）》。IEEE，2022年，第33–40页。[17] L. Zhang、N. Liu、Y. Hou 和 X. Liu，《基于多阈值S-F的不均匀光照图像分割》，《光电工程》，第41卷，87，2014年7，81第。页–。[18] J. Lee、J. Hwangbo、L. Wellhausen、V. Koltun 和 M. Hutter，《在复杂地形上学习四足运动》，《科学机器人》，第5卷，第47期，文章编号 eabc5986，2020年。[19] S. Levine、C. Finn、T. Darrell 和 P. Abbeel，《端到端训练深度视觉运动策略》，《机器学习研究杂志》，第17卷，40，2016年39，1第。页–。[20] A. Rajeswaran、V. Kumar、A. Gupta、G. Vezzani、J. Schulman、E. Todorov 和 S. Levine，《利用深度强化学习和示范数据学习复杂的灵巧操作》，arXiv预印本 arXiv:1709.10087，2017年。[21] A. Zeng、P. Florence、J. Tompson、S. Welker、J. Chien、M. Attarian、T. Armstrong、I. Krasin、D. Duong、V. Sindhwani 等，《Transporter网络：重构视觉世界以实现机器人操作》，载于《机器人学习会议》。PMLR，2021年，第726–747页。[22] P. Florence、C. Lynch、A. Zeng、O. A. Ramirez、A. Wahid、L. Downs、A. Wong、J. Lee、I. Mordatch 和 J. Tompson，《隐式行为克隆》，载于《机器人学习会议》。PMLR，2022年，第158–168页。[23] B. Ichter、J. Harrison 和 M. Pavone，《为机器人运动规划学习采样分布》，载于《2018年IEEE机器人与自动化国际会议（ICRA）》。IEEE，2018年，第7087–7094页。[24] X. Sha、Z. Jia、W. Sun、Y. Hao、X. Xiao 和 H. Hu，《基于HoloLens的混合现实机器人控制系统开发》，载于《智能机器人与应用》，H. Yu、J. Liu、L. Liu、Z. Ju、Y. Liu 和 D. Zhou 编。施普林格：施普林格国际出版，2019年，第571–581页。[25] T.-W. Ke、N. Gkanatsios 和 K. Fragkiadaki，“3D扩散演员：基于3D场景表征的策略扩散”，arXiv预印本 arXiv:2402.10885，2024年。[26] A. Z. Ren、J. Lidard、L. L. Ankile、A. Simeonov、P. Agrawal、A. Majumdar、B. Burch- fiel、H. Dai 和 M. Simchowitz，“扩散策略策略优化”，arXiv预印本 arXiv:2409.00588，2024年。[27] E. Ng、J. Romero、T. Bagautdinov、S. Bai、T. Darrell、A. Kanazawa 和 A. Richard，“从音频到照片级真实化身：合成人类对话”，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2024年，第1001–1010页。[28] A. Van Den Oord、O. Vinyals 等，“神经离散表征学习”，《神经信息处理系统进展》，第30卷，2017年。[29] T. Chen、S. Kornblith、M. Norouzi 和 G. Hinton，“视觉表征对比学习的简易框架”，载于《国际机器学习会议》。PMLR，2020年，第1597–1607页。[30] X. Sha 和 T. Mashita，“3D局部部件上的动量对比学习评估”，载于《2024年国际网格与云计算研讨会（ISGC）论文集——PoS（ISGC2024）》，第458卷，2024年，第004页。[31] Y. Ze、G. Zhang、K. Zhang、C. Hu、M. Wang 和 H. Xu，“3D扩散策略：通过简易3D表征实现可泛化的视觉运动策略学习”，arXiv预印本 arXiv:2403.03954，2024年。[32] L. Zhang、Z. Lian、H. Liu、T. Takebe 和 Y. Nakashima，“Simlabel：面向缺失标签多标注者学习的相似度加权半监督方法”，arXiv预印本 arXiv:2504.09525，2025年。[33] L. Zhang、J. Ke、S. Fan、X. S ha 和 Z. Lian，“多标注者倾向学习的统一评估框架”，2025年。 [在线]。网址：https://arxiv.org/abs/2508.10393[34] Y. Yoon、B. Cha、J.-H. Lee、M. Jang、J. Lee、J. Kim 和 G. Lee，“基于文本、音频和说话人身份三模态上下文的语音手势生成”，《ACM图形学汇刊（TOG）》，第39卷，第6期，第1–16页，2020年。[35] J. Li、D. Kang、W. Pei、X. Zhe、Y. Zhang、Z. He 和 L. Bao，“Audio2gestures：利用条件变分自编码器从语音音频生成多样化手势”，载于《IEEE/CVF国际计算机视觉会议论文集》，2021年，第11 293–11302页。[36] J. Xing、M. Xia、Y. Zhang、X. Cun、J. Wang 和 T.-T. Wong，“CodeTalker：基于离散运动先验的语音驱动3D面部动画”，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2023年，第12 780–12 790页。[37] S. Kopp 和 I. Wachsmuth，“为对话代理合成多模态话语”，《计算机动画与虚拟世界》，第15卷，52，2004 1，39第。– 页。[38] L. Zhang，“Microemo：视频对话中基于微表情动态的时间敏感型多模态情绪识别”，arXiv预印本 arXiv:2407 16552，2024年。[39] L. Zhang、Z. Luo、S. Wu 和 Y. Nakashima，“Microemo：视频对话中基于细微线索动态的时间敏感型多模态情绪识别”，载于《第二届多模态与负责任情感计算国际研讨会论文集》，2024年，第110–115页。[40] P. Wagner、Z. Malisz 和 S. Kopp，“交互中的手势与语音：综述”，第209–232页，2014年。[41] H. Liu、N. Iwamoto、Z. Zhu、Z. Li、Y. Zhou、E. Bozkurt 和 B. Zheng，“DisCo：面向多样化伴随语音手势合成的解耦隐式内容与节奏学习”，载于《第30届ACM国际多媒体会议论文集》，2022年，第3764–3773页。[42] Y. Yoon、W.-R. Ko、M. Jang、J. Lee、J. Kim 和 G. Lee，“机器人学习社交技能：面向人形机器人的协同语音手势生成端到端学习”，载于《2019年国际机器人与自动化会议（ICRA）》。IEEE，2019年，第4303–4309页。[43] U. Bhattacharya、N. Rewkowski、A. Banerjee、P. Guhan、A. Bera 和 D. Manocha，“Text2gestures：一种用于为虚拟代理生成情绪化身体手势的基于Transformer的网络”，载于《2021年IEEE虚拟现实与3D用户界面会议（VR）》。IEEE，2021年，第1–10页。[44] X. Qi、C. Liu、L. Li、J. Hou、H. Xin 和 X. Yu，“Emotiongesture：音频驱动的多样化情绪化伴随语音3D手势生成”，《IEEE多媒体汇刊》，第26卷，第10 420–10430页，2024年。[45] X. Zhang、J. Li、J. Zhang、Z. Dang、J. Ren、L. Bo 和 Z. Tu，“Semtalk：帧级语义强调的整体协同语音动作生成”，载于《IEEE/CVF国际计算机视觉会议论文集》，2025年，第13 761–13 771页。[46] W.-N. Hsu、B. Bolte、Y.-H. H. Tsai、K. Lakhotia、R. Salakhutdinov 和 A. Mohamed，“HuBERT：通过隐藏单元掩码预测进行自监督语音表征学习”，《IEEE/ACM音频、语音与语言处理汇刊》，第29卷，第3451–3460页，2021年。[47] L. T. Nguyen、T. Pham 和 D. Q. Nguyen，“XPhoneBERT：一种用于文本转语音音素表征的预训练多语言模型”，arXiv预印本 arXiv:2305.19709，2023年。[48] Y. Fan、Z. Lin、J. Saito、W. Wang 和 T. Komura，“FaceFormer：基于Transformer的语音驱动3D面部动画”，载于《IEEE/CVF计算机视觉与

计算机视觉与模式识别，2022年，第18 770–18 780页。[49] S. Yang、Z. Wu、M. Li、Z. Zhang、L. Hao、W. Bao、M. Cheng 和 L. Xiao，“Diffus- estylegesture：基于扩散模型的风格化音频驱动协同语音手势生成”，*arXiv*预印本 *arXiv:2305.04919*，2023年。

[50] I. Habibie、W. Xu、D. Mehta、L. Liu、H.-P. Seidel、G. Pons-Moll、M. Elgharib 和 C. Theobalt，“从视频中学习语音驱动的三维对话手势”，载于第*21*届 *ACM* 智能虚拟代理国际会议论文集，2021年，第101–108页。