arXiv:2601.04236v1 [cs.SD] 4 Jan 2026

# SmoothSync：基于量化音频的抗抖动节拍同步手势生成双流扩散变换器

蒋玉娇[a]，*，廖清敏[a]，卢宗庆[a]

[a]清华大学深圳国际研究生院，清华大学，中国深圳，*518055*

**摘要**

伴语手势生成是一个重要的研究领域，旨在合成与语音同步的类人手势。现有方法常面临节奏不一致、动作抖动、脚部滑动以及多次采样多样性有限等问题。本文提出一种新颖的框架 SmoothSync，它利用双流扩散变换器（DiT）架构中的量化音频令牌来合成整体手势并增强采样变化。具体而言，我们（1）通过互补的变换器流融合音频-运动特征，实现更出色的同步；（2）引入抖动抑制损失，以提升时间平滑性；（3）采用概率音频量化，从相同输入生成不同的手势序列。为可靠地评估抖动情况下的节拍同步，我们提出 Smooth-BC，这是一种对动作噪声不太敏感的节拍一致性指标鲁棒变体。在 BEAT2 和 SHOW 数据集上进行的大量实验表明，SmoothSync 的表现优于现有最先进方法：在 BEAT2 上，FGD 提升 30.6%，Smooth-BC 提升 10.3%，多样性提升 8.4%；同时分别将抖动和脚部滑动减少 62.9% 和 17.1%。代码将于未来发布，以促进后续研究。

关键词：伴语手势生成、扩散变换器、双流架构

*通讯作者电子邮箱：jiangyj20@mails.tsinghua.edu.cn（蒋玉娇）

## 1. 引言

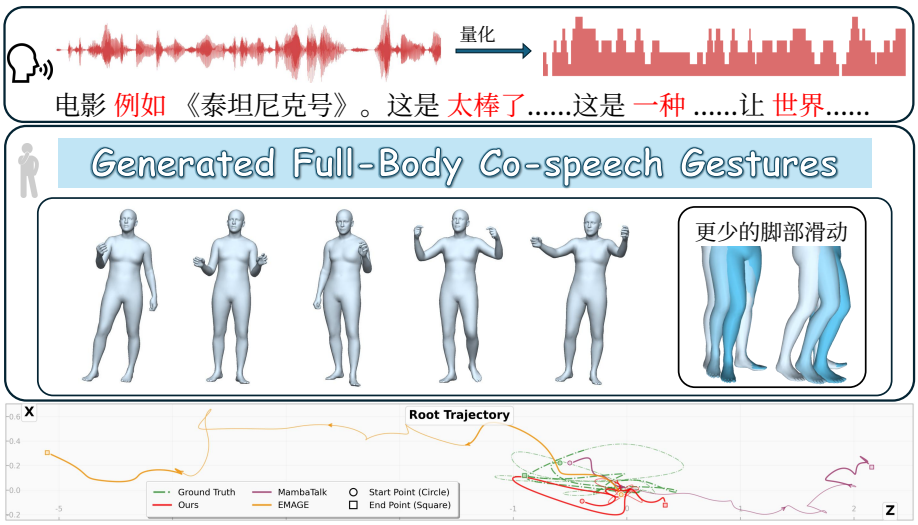


图 1：**SmoothSync** 以量化音频梅尔能量作为输入，并通过双流网络高效融合音频和动作特征，以进行去噪生成。我们的方法生成的动作更加平滑，脚部滑动更少，整体运动范围也与真实数据更加吻合。相比之下，以往方法生成的动作存在严重漂移和过度抖动。**放大查看**以获得更清晰的视图。

伴语手势生成是计算机视觉、计算机图形学和人机交互领域的一项基础挑战，应用范围涵盖从虚拟化身到具身人工智能系统。其目标是合成自然且富有表现力的身体动作，使其与语音音频在时间上保持同步，同时与口语内容保持语义连贯性。然而，由于音频与动作之间并非简单的一一映射，该领域仍难以生成兼具**精准节奏**、**视觉流畅**和**风格多样**的动作。

近期的研究探索了多种深度学习方法，包括生成对抗网络（GAN） [1, 2]，基于向量量化变分自编码器（VQ-VAE）的自回归模型 [3, 4, 5]，扩散模型 [6, 7]，以及状态空间模型 [5, 8]。然而，当前方法面临三项关键局限，妨碍了其实际部署：**动作质量问题：**现有方法存在动作伪影，包括抖动、脚部滑动和时间不一致性。许多基于 VQ-VAE 的——

方法 [4, 5, 9] 依赖离散多部件动作表示，这类表示会丢失细粒度的动作细节，也难以维持平滑的时间过渡。**多样性有限：**大多数方法采用大规模预训练音频编码器 [10, 11]，其输出具有确定性，在多次采样轮次中几乎没有变化；这与自然的人类行为相悖，因为相同的语音可以伴随不同的手势。**身体覆盖不完整：**许多方法 [6, 12] 仅关注上半身手势，忽略了包括全局平移在内的全身动力学，因而限制了它们在沉浸式环境中的适用性。

为应对这些挑战，我们提出了**SmoothSync**，这是一种新颖的双流扩散 Transformer 框架，革新了音频和动作特征的处理与融合方式，从而实现高质量的手势合成。我们的核心洞见是，有效的音频—动作融合需要先进行**模态专属处理**，再进行**跨模态整合**。与现有方法将各模态独立处理或简单拼接特征不同，我们的双流架构为音频令牌和运动令牌保留了独立的处理路径，使每种模态都能依据归纳偏置进行处理，并通过联合注意力机制实现跨模态交互。

我们的方法包含三项关键创新：（1）**双流架构**，通过并行流处理音频和动作，并采用模态专属的归一化和注意力机制，随后进行联合跨模态注意力；（2）**抖动抑制损失**，在保留自然表现力的同时，明确惩罚高频动作伪影；（3）**量化梅尔频谱图特征**，在保持强同步性的同时，使相同输入能够生成多样化的手势。

为在存在动作伪影的情况下实现可靠的节奏对齐评估，我们引入了**Smooth-BC 指标**，该指标能够滤除由抖动引起的虚假节拍检测，同时准确测量真实的节奏同步。我们的综合评估还包括用于运动质量评估的新指标：抖动与脚滑指标 [13]，用于量化传统测量方法经常忽略的动作伪影；以及样本间多样性指标，用于衡量针对同一输入生成的多个样本之间的变化。

在 BEAT2 和 SHOW数据集上的实验 [4, 3]表明，Smooth-Sync 达到了最先进性能，能够生成高质量、多样化且具有节奏同步性的、

全身体势，同时显著减少动作伪影，如图 1 所示。

我们的主要贡献如下:

• 我们提出了**SmoothSync**，一种统一的双流扩散框架，可根据音频输入生成高质量、多样化的全身手势。在BEAT2和SHOW数据集上，该方法的抖动显著降低，节奏对齐效果更优，表现大幅超越当前最先进方法。

• 我们提出了一种**双流Transformer架构**，采用模态专属处理和联合跨模态注意力，并结合**量化梅尔频谱图特征提取**流程，在保持强同步性的同时实现多样化的手势生成。

• 我们提出了**Smooth-BC**，这是一种稳健的节拍一致性指标，可通过滤除抖动伪影实现可靠的节奏评估，弥补传统评估协议的局限。

## 2. 相关工作

**伴语手势生成**。 早期方法依赖基于规则的系统 [14,15, 16] 和统计模型 [17, 18],, 但这些方法在自然交互方面缺乏灵活性。

深度学习通过不同范式革新了这一领域。**VQ-VAE自回归模型**（如 TalkSHOW [3], 、EMAGE [4], 和 MambaTalk [5] ）将动作离散化为词元, 以进行自回归生成，而 SemTalk [19] 和 Echo-Mask [9] 则采用残差向量量化（RVQ-VAE）。然而，离散化误差会导致动作抖动和脚部滑动伪影。这些方法将身体划分为不同部分，分别对每个部分进行重建，从而产生不自然且在时间上不一致的手势。**扩散模型**已成为一种替代方案。DiffGesture [6], 、LivelySpeaker [12], 以及近期的 MDT-A2G [20], 、DiffSHEG [21], 和 MMoFusion [22] 展现了稳定的训练过程和高质量的生成能力。扩散模型能够处理多模态分布，因此非常适合于

手势生成。**状态空间模型**（如 MambaTalk [5] 和 MambaGesture [8]）试图利用其在长序列建模方面的高效率，但难以捕捉身体各部分之间的动态变化，因误差累积以及对高频运动细节的约束不足而产生抖动伪影。关于运动预测的相关工作 [23, 24] 已探索运动学建模方法。

一个关键局限是仅关注上半身手势。许多方法 [25,6, 3, 12, 26]忽略了全身动力学和全局平移。SMPLX [27]模型取得了重大进展，提供了涵盖身体、面部和手部的统一表示，使 EMAGE [4] 和 MambaTalk [5]等近期工作能够实现全身合成。

EMAGE [4] 是一个重要的里程碑，它采用掩码音频手势建模，并针对不同身体部位使用四个组合式 VQ-VAE。尽管取得了令人瞩目的成果，但它仍受限于离散表示瓶颈，需要复杂的多阶段流水线，且时序建模能力有限。我们的方法通过扩散直接建模连续SMPLX参数，从而解决这些问题。

**音频特征设计**。 许多近期方法 [21, 19, 28] 采用大规模预训练编码器，例如 Wav2vec2 [11], HuBERT [29], 或 Whisper [10]。这些编码器虽能提供丰富的语义表示，却会导致确定性输出，且样本间多样性有限。相比之下，我们的方法使用量化梅尔频谱图特征 [30]，以实现多样化的手势生成。

**运动质量评估**。 常用的 BC（节拍一致性）指标 [31,32, 33] 对噪声极为敏感，纯噪声动作也能获得高分，因此并不可靠。为解决这一局限，我们提出了 Smooth-BC，这是一种能够滤除噪声干扰、可靠评估节奏对齐的稳健指标。

## 3. 方法

### *3.1.* 总体架构

我们提出了一种新颖的双流扩散变换器（DiT）架构，能够有效捕捉语音音频与人体运动之间的复杂交互，

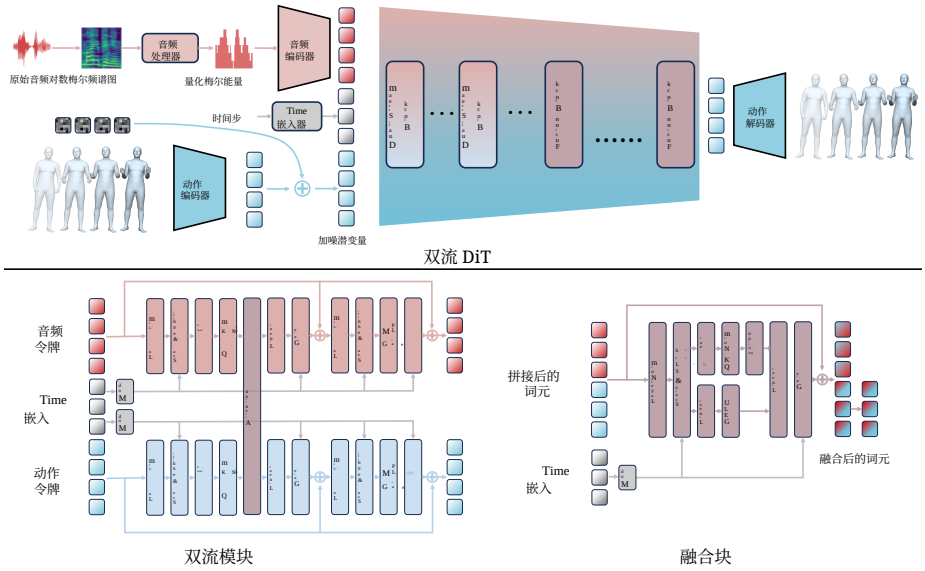


图 2：我们的双流 DiT架构概览。该框架通过独立路径分别处理音频和动作，然后利用专门的 Transformer 模块将二者融合，从而实现高质量的语音驱动动作生成。

其设计灵感源自可扩展扩散 Transformer 的最新进展 [34]。如图 2 所示，该架构由三个关键组件构成：（1）音频编码器，用于从语音信号中提取稳健的多模态特征；（2）动作编码器，用于归一化人体动作表示，以及运动解码器，用于将处理后的特征转换回原始动作空间；（3）采用专门融合机制的双流 Transformer 网络。

### *3.2.* 音频编码

给定输入音频后，我们提取梅尔频谱图特征，并应用量化以离散化音频表示：

$$\mathbf{F}_a = \text{Embed}(\text{Quantize}(\text{MelSpec}(\text{audio}))). \tag{1}$$

**音频特征提取：**我们使用 25 毫秒窗口和 10 毫秒步长从音频中提取对数梅尔频谱图特征，并应用指数变换并计算幅度，以获得梅尔能量。

**梅尔能量增强：**训练期间，我们通过低频噪声增强和归一化进行数据增强。我们采样随机参数并生成正弦噪声：

$$A \sim \mathcal{U}(0,4), \quad \phi \sim \mathcal{U}(0,2\pi), \quad f_r \sim \mathcal{U}(150,1200), \tag{2}$$

$$\mathbf{n}(t) = A \sin\left(\frac{2\pi t}{f_r} + \phi\right), \quad t \in \{0,1,\ldots,T-1\}. \tag{3}$$

增强后的特征计算如下：

$$\mathbf{x}'_{t,d} = \begin{cases} \max(x_{\text{thresh}}, (\mathbf{x}+\mathbf{n})_{t,d}) & \text{if } \mathbf{x}_{t,d} > x_{\text{thresh}}, \\ (\mathbf{x}+\mathbf{n})_{t,d} & \text{otherwise}, \end{cases} \tag{4}$$

其中 $x_{\text{thresh}} = \min(\mathbf{x} + 0.1(\max(\mathbf{x}) - \min(\mathbf{x}))$.

最后，我们将特征归一化为 $[0,1]$:

$$\mathbf{x}_{aug} = \text{clamp}\left(\frac{\mathbf{x}' - \min(\mathbf{x})}{\max(\mathbf{x}) - \min(\mathbf{x})}, 0, 1\right). \tag{5}$$

**时间量化：** 我们对音频特征应用三阶段时间量化过程，以改善时间对齐并获得更离散的表示。

给定长度为 $\mathbf{v} \in [0,1]^{T\times D}$、维度为 $T$、窗口大小为 $D$ 的输入特征序列，以及窗口大小 $w$，我们执行以下操作：

**步骤 1：采用自适应偏移进行时间下采样** with adaptiveoffset:

$$s = \begin{cases} \text{randint}(0, w-1) & \text{training} \\ \lfloor w/2 \rfloor & \text{inference} \end{cases}, \quad \mathbf{v}_{\text{down}} = \mathbf{v}[s :: w]. \tag{6}$$

**步骤 2：时间上采样** 通过重复每帧 $w$ 次并调整长度：

$$\mathbf{v}_{\text{up}} = \text{repeat}(\mathbf{v}_{\text{down}}, w)[:T]. \tag{7}$$

**步骤 3：数值量化**采用自适应分箱：

$$\mathbf{q}_i = \begin{cases} \mathbb{I}[\mathbf{v}_{\text{up},i} > 0.05] & \text{if } n_{\text{bins}} = 2 \\ \text{clamp}(\lfloor \mathbf{v}_{\text{up},i} \cdot n_{\text{bins}} \rfloor, 0, n_{\text{bins}} - 1) & \text{otherwise} \end{cases}. \tag{8}$$

随后通过学习得到的嵌入器，将量化音频特征嵌入到与动作特征维度相匹配的空间中。

### *3.3.* 双流扩散变换器网络

**动作表示：** 我们使用 SMPLX 的 [27] rot6d 参数化来表示人体运动。完整的动作向量$\mathbf{m}_t \in \mathbb{R}^{333}$ 包含姿态旋转$\mathbf{R}_t \in \mathbb{R}^{55\times6}$ 和全局平移 $\mathbf{T}_t \in \mathbb{R}^3$。这种表示方式在保持计算效率的同时，捕捉了人体动作的完整复杂性，并避免了其他旋转表示固有的奇异性。

**双流架构：** 我们的核心创新在于双流处理范式：在进行策略性融合之前，为音频和运动令牌分别保留独立的处理路径。该设计使每种模态都能依据其模态特定归纳偏置进行处理，同时支持跨模态交互。

动作编码器使用预先计算的均值和标准差统计量对动作序列进行归一化：

$$\mathbf{m}_0^{norm} = \frac{\mathbf{m}_0 - \boldsymbol{\mu}_m}{\boldsymbol{\sigma}_m}. \tag{9}$$

在训练期间，随后会对归一化后的动作施加噪声扰动：

$$\mathbf{m}_t = \sqrt{\tilde{\alpha}_t} \cdot \mathbf{m}_0^{norm} + \sqrt{1-\tilde{\alpha}_t} \cdot \boldsymbol{\epsilon}, \tag{10}$$

其中，$\mathbf{m}_0^{norm}$表示归一化后的动作，$\boldsymbol{\epsilon}$ 为高斯噪声，$\tilde{\alpha}_t$ 遵循预定义的噪声调度。

音频和动作特征均被线性投影到变换器隐藏维度：

$$\mathbf{X}_a = \text{Linear}_a(\mathbf{F}_a), \quad \mathbf{X}_m = \text{Linear}_m(\mathbf{m}_t). \tag{11}$$

**双流模块：** 每个模块通过并行特征提取处理音频和运动令牌，随后采用联合注意力机制。该模块由以下部分组成：

- **调制：** 遵循 DiT 设计原则，我们针对每种模态，分别应用由时间步嵌入条件化的自适应层归一化：

$$\alpha_{1,i}, \beta_{1,i}, \gamma_{1,i}, \alpha_{2,i}, \beta_{2,i}, \gamma_{2,i} = \text{Mod}_i(\mathbf{t}), \tag{12}$$

$$\hat{\mathbf{X}}_i = (1+\alpha_{1,i}) \cdot \text{LayerNorm}(\mathbf{X}_i) + \beta_{1,i}, \tag{13}$$

其中 $i \in \{a, m\}$。

- **并行 QKV 生成：**每种模态独立生成查询、键和值向量：

$$\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i = \text{Linear}_i(\hat{\mathbf{X}}_i), \quad i \in \{a, m\}. \tag{14}$$

- **QK 归一化：** 我们对查询向量和键向量进行归一化，以稳定训练：

$$\mathbf{Q}_i, \mathbf{K}_i = \text{RMSNorm}(\mathbf{Q}_i, \mathbf{K}_i), \quad i \in \{a, m\}. \tag{15}$$

- **联合跨模态注意力：** 音频和运动令牌通过拼接注意力进行融合：

$$\mathbf{Q}_j, \mathbf{K}_j, \mathbf{V}_j = \text{Concat}([\mathbf{Q}_a, \mathbf{Q}_m], [\mathbf{K}_a, \mathbf{K}_m], [\mathbf{V}_a, \mathbf{V}_m]) \tag{16}$$

$$\mathbf{A}_a, \mathbf{A}_m = \text{Split}(\text{Attention}(\mathbf{Q}_j, \mathbf{K}_j, \mathbf{V}_j)). \tag{17}$$

**残差连接和多层感知机：**每种模态都会在注意力层和多层感知机层中应用门控残差连接：

$$\mathbf{X}_i = \mathbf{X}_i + \gamma_{1,i} \cdot \text{Linear}_i(\mathbf{A}_i), \tag{18}$$

$$\mathbf{X}_i = \mathbf{X}_i + \gamma_{2,i} \cdot \text{MLP}_i((1 + \alpha_{2,i}) \cdot \text{LayerNorm}(\mathbf{X}_i) + \beta_{2,i}), \tag{19}$$

其中 $i \in \{a,m\}$。

这种集成的双流设计使模型能够通过并行 QKV 生成捕捉各模态特有的特征，并通过联合注意力捕捉跨模态依赖关系，这对于高质量的语音驱动运动合成至关重要。

**融合块：** 除了双流模块外，我们还采用融合块，通过统一注意力和多层感知机计算来处理拼接后的音频-运动特征。每个融合块都会执行：

- **调制处理：**拼接后的特征经过时间步

  条件归一化：

$$\alpha_f, \beta_f, \gamma_f = \text{Mod}_f(\mathbf{t}), \tag{20}$$

$$\mathbf{X}_{mod} = (1 + \alpha_f) \cdot \text{LayerNorm}(\mathbf{X}_{cat}) + \beta_f. \tag{21}$$

• **联合 QKV 和多层感知机生成：** 单个线性变换同时生成注意力和多层感知机特征：

$$[\mathbf{QKV}, \mathbf{M}] = \text{Split}(\text{Linear}_1(\mathbf{X}_{mod})), \tag{22}$$

其中，**QKV** 包含查询、键和值向量，**M** 包含多层感知机输入特征。

• **并行处理：** 注意力和多层感知机同时计算：

$$\mathbf{A}_f = \text{Attention}(\text{RMSNorm}(\text{Split}(\mathbf{QKV}))), \tag{23}$$

$$\mathbf{M}_{act} = \text{Activation}(\mathbf{M}). \tag{24}$$

• **输出整合：** 结果通过残差连接进行合并：

$$\mathbf{X}_{out} = \mathbf{X}_{cat} + \gamma_f \cdot \text{Linear}_2(\text{Concat}([\mathbf{A}_f, \mathbf{M}_{act}])) \tag{25}$$

这种融合块设计在单一的统一计算中高效结合了跨模态注意力与前馈处理，从而实现音频和运动模态之间的有效信息整合。

**运动解码器：** 运动解码器是一个简单的反归一化层，可将处理后的潜在特征转换回原始动作空间。它使用预先计算的统计量进行逆归一化：

$$\mathbf{m}_{final} = \mathbf{X}_m \cdot \boldsymbol{\sigma}_m + \boldsymbol{\mu}_m, \tag{26}$$

其中，$\mu_m$ 和 $\boldsymbol{\sigma}_m$ 是根据训练数据计算得出的均值和标准差统计量，$\mathbf{X}_m$ 表示来自双流变换器的处理后动作特征。

### *3.4.* 长运动生成

对于较长的动作序列，我们采用 MimicMotion [35] 中的基于片段的生成策略。我们将序列划分为彼此重叠的时间片段，并分别进行处理。通过在重叠区域进行渐进式融合，并利用前一片段的运动上下文，确保各片段之间的时间连贯性。这种方法使我们能够生成任意长度的动作序列，同时保持局部运动质量和全局时间一致性。

### *3.5.* 训练目标

我们的训练目标由三个主要部分组成：rot6d重建损失、平移重建损失和抖动损失。这些损失共同确保生成的动作准确、平滑且符合物理规律。

**重建损失：**

$$\mathcal{L}_{\text{rot6d}} = \frac{1}{\sum_{t=1}^{T} M_t} \sum_{t=1}^{T} M_t \cdot \frac{1}{D_{\text{rot6d}}} \left\| \mathbf{R}_t^{\text{pred}} - \mathbf{R}_t^{\text{gt}} \right\|_2^2, \tag{27}$$

$$\mathcal{L}_{\text{trans}} = \frac{1}{\sum_{t=1}^{T} M_t} \sum_{t=1}^{T} M_t \cdot \frac{1}{D_{\text{trans}}} \left\| \mathbf{T}_t^{\text{pred}} - \mathbf{T}_t^{\text{gt}} \right\|_2^2, \tag{28}$$

其中，$M_t$ 是帧$t$的有效性掩码，$D_{\text{rot6d}}$ 是 rot6d 表示的维度，$D_{\text{trans}}$ 是平移向量的维度。

**抖动损失：**为抑制高频抖动并确保时间平滑性，我们引入了一种基于三维关节位置三阶有限差分（加加速度）的抖动损失：

$$\mathcal{L}_{\text{jitter}} = \frac{1}{\sum_{t=1}^{T-3} M_t^{\text{jitter}}} \sum_{t=1}^{T-3} M_t^{\text{jitter}} \cdot \frac{1}{D_{\text{joints}}} \left\| \mathbf{J}_t^{\text{pred}} - \mathbf{J}_t^{\text{gt}} \right\|_2^2, \tag{29}$$

其中，$D^{\text{joints}}$ 是关节数量，三阶差分定义为：

$$\mathbf{J}_t = \left\| (\mathbf{P}_{t+3} - 3\mathbf{P}_{t+2} + 3\mathbf{P}_{t+1} - \mathbf{P}_t) \cdot \text{fps}^3 \right\|_2, \tag{30}$$

其中，$\mathbf{P}_t$ 表示帧$t$处的三维关节位置，帧率是采样频率。**总损失：**最终训练目标是上述损失的加权和：

$$\mathcal{L}_{\text{total}} = \lambda_{\text{rot6d}} \mathcal{L}_{\text{rot6d}} + \lambda_{\text{trans}} \mathcal{L}_{\text{trans}} + \lambda_{\text{jitter}} \mathcal{L}_{\text{jitter}}, \tag{31}$$

其中我们设置了 $\lambda_{\text{rot6d}} = 1$、 $\lambda_{\text{trans}} = 1$ 和 $\lambda_{\text{jitter}} = 1 \times 10^{-9}$。

## 4. 实验

### *4.1.* 实验设置

**数据集**。 我们在两个数据集上评估我们的方法：BEAT2 和 SHOW。**BEAT 2** [4]是 BEAT 的扩展版本， [36], 包含 76 小时的多说话者录制内容（30 位说话者），并配有同步音频、转录文本和语义标注。

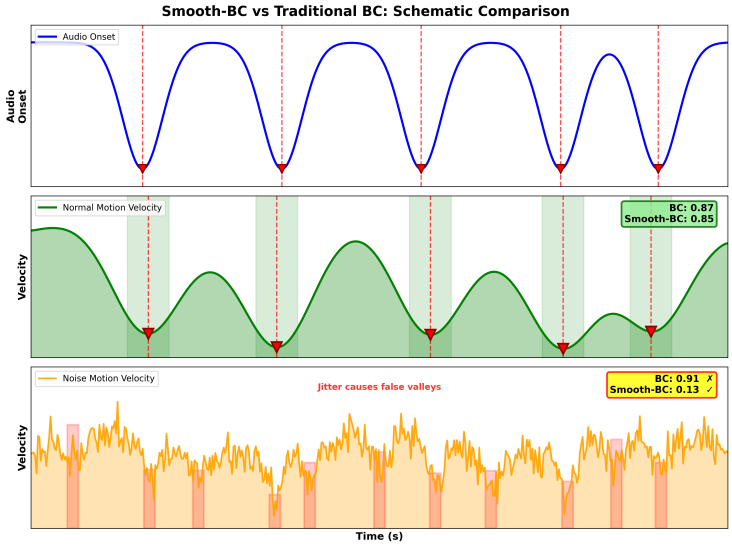


图 3：提出 **Smooth-BC** 指标，通过对速度极值点附近的斜率施加严格约束，排除因抖动伪影而被错误检测到的运动节拍。

运动数据以统一网格参数和帧级标签表示。遵循 EMAGE [4], ，我们采用 BEAT2 标准子集，并按 85%/7.5%/7.5% 划分训练/验证/测试集。**SHOW** [3] 提供了来自四位说话者的 26.9 小时真实世界脱口秀视频，其中包含同步的三维人体网格重建和音频流。动作采用 SMPLX [27] 表示，以 30 帧率捕获；音频则以 22 kHz 采样率录制。我们遵循 TalkSHOW [3], 中的方案，筛除时长超过三秒的片段，并按 80%/10%/10% 划分训练、验证和测试集。

**评估指标**。 我们使用弗雷歇手势距离（FGD） [37] 来衡量生成手势的真实感，使用节拍一致性（BC） [32] 来评估音频-动作同步，使用内部多样性 [38] 来衡量单个生成的动作序列中的多样性，并使用交互多样性 [38]来衡量不同采样轮次之间的多样性。为了评估动作抖动和脚部滑动等动作伪影，我们引入了抖动与脚滑指标 [13]。此外，我们观察到 BC 指标对动作抖动高度敏感，纯噪声序列会表现出异常高的 BC值（BC=7.868，如表4中的噪声动作变体所示）。为解决这一

关键局限，我们提出了 **Smooth-BC**指标。该指标在检测运动节拍时，对速度极值附近的斜率施加严格约束，从而有效滤除由抖动伪影引起的虚假节拍检测。如图3所示，传统 BC 会将所有速度极小值视为有效的运动节拍，使高频抖动人为抬高分数；而 Smooth-BC 则应用斜率约束，将真实的节奏性运动与噪声引起的波动区分开来。噪声运动实验验证了其鲁棒性：虽然纯噪声的 BC=7.868，但其 Smooth-BC分数几乎为零（0.004965），这证实 Smooth-BC 能够可靠地衡量真实的节奏同步，而不会受到动作伪影的误导。有关所有评估指标的完整细节，请参阅补充材料。

**实现细节**。 我们使用 8 块 V100 GPU，在 BEAT2数据集上进行训练，批量大小为 16。我们采用 AdamW优化器，学习率设为 1e-4，并训练 50,000 次迭代。每个训练序列由 320 帧组成。对于扩散过程，我们在训练期间采用包含 1,000 步的 DDPM调度，并在推理时切换到包含 50 步的 DDIM调度，以加快生成速度。我们的方法在单块 RTX 4090 GPU 上实现了高效推理，速度达到 160 FPS（对于 30fps 动作，相当于 5.3× 倍实时速度）；处理 320 帧序列时，模型包含 154.47M 个参数，内存占用为 2.28 GB。详细的计算分析见补充材料。

### *4.2.* 定量结果

SmoothSync 在全面的评估指标上实现了最先进性能，在动作质量、节奏对齐和生成多样性方面均取得了显著提升。

**动作真实性与多样性**。 表1展示了我们在BEAT2数据集上与基线方法的定量比较。SmoothSync取得了**最佳FGD分数**，表明其动作真实性更优。更值得注意的是，我们的方法取得了**最高的Intra-Diversity分数**，证明SmoothSync能够在单个序列中生成变化更丰富的动作。在保持具有竞争力的Beat Consistency性能（排名第二）的同时，我们的方法成功平衡了节奏对齐

表1：**在BEAT2数据集上的定量比较**。 SmoothSync取得了最佳的FGD（动作真实性）和Intra-Diversity分数，同时保持了有竞争力的BC表现。结果表明，与现有方法相比，该方法的动作质量和多样性更优。我们报告FGD $\times 10^{-1}$、BC $\times 10^{-1}$。DIV表示Intra-Diversity。

| 方法 | FGD↓ | BC↑ | DIV↑ |
|---|---|---|---|
| S2G [1] | 28.15 | 4.683 | 5.971 |
| Trimodal [37] | 12.41 | 5.933 | 7.724 |
| HA2G [25] | 12.32 | 6.779 | 8.626 |
| DisCo [39] | 9.417 | 6.439 | 9.912 |
| CaMN [36] | 6.644 | 6.769 | 10.86 |
| DSG [40] | 8.811 | 7.241 | 11.49 |
| Habibie 等人 [2] | 9.040 | 7.716 | 8.213 |
| TalkSHOW [3] | 6.209 | 6.947 | <u>13.47</u> |
| EMAGE [4] | 5.512 | 7.724 | 13.06 |
| DiffSHEG [21] | 8.986 | 7.142 | 11.91 |
| MambaTalk [5] | 5.366 | **7.812** | 13.05 |
| EchoMask [9] | 4.623 | 7.738 | 13.37 |
| SemTalk [19] | <u>4.278</u> | 7.770 | 12.91 |
| **SmoothSync** | **4.151** | <u>7.781</u> | **14.00** |

与运动多样性——这是以往方法难以实现的一项挑战性权衡。

**动作质量与伪影减少**。表2提供了扩展评估，重点关注对实际应用至关重要的动作质量指标。SmoothSync 在所有质量指标上都展现出卓越性能：(1) 与 MambaTalk 相比，FGD 提升了 30.6%，表明其能够生成明显更逼真的动作；(2) 相较于 EMAGE，Smooth-BC 提升了 10.3%，证明在适当滤除动作伪影后，其节奏对齐能力更优；(3)**与 EMAGE 相比，抖动减少了 62.9%**，表明其动作生成过程明显更加平滑；(4) 与 MambaTalk 相比，足部滑动减少了 17.1%，表明其足部接触一致性更佳。值得注意的是，SmoothSync 是**唯一实现非零交互多样性的方法**，而 EMAGE 和 MambaTalk 的得分均为

表2：**动作质量指标的扩展比较**。SmoothSync 在所有指标上均显著优于基线方法，在动作真实性、节拍一致性、内部和交互多样性方面表现最佳，同时大幅减少了动作伪影。我们使用不同的随机种子生成4个样本，并计算它们之间的交互多样性。我们报告 Smooth-BC $\times 10^{-1}$、抖动 $\times 10^{2}$、足部滑动 $\times 10^{-2}$、交互多样性 $\times 10^{-2}$。

| 方法 | FGD↓ | Smooth-BC↑ | 内部多样性↑ | 抖动↓ | 足部滑动↓ | 交互多样性↑ |
|---|---|---|---|---|---|---|
| EMAGE [4] | 5.512 | 4.064 | 13.06 | 1.351 | 0.9276 | 0 |
| MambaTalk [5] | 5.366 | 3.427 | 13.05 | 1.447 | 0.7693 | 0 |
| **SmoothSync** | **3.723** | **4.484** | **14.16** | **0.5008** | **0.6376** | **8.090** |

表3：**在SHOW数据集上的定量比较**。SmoothSync在所有评估指标上均优于TalkSHOW，展现出更高的动作真实性（FGD）、更稳健的节拍一致性（Smooth-BC）、更强的多样性（内部多样性和交互多样性），以及显著减少的抖动伪影。结果验证了我们的方法对真实场景脱口秀录音的泛化能力。

| 方法 | FGD↓ | Smooth-BC↑ | 内部多样性↑ | 抖动↓ | 交互多样性↑ |
|---|---|---|---|---|---|
| TalkSHOW [3] | 5.072 | 2.396 | 3.291 | 0.6795 | 2.295 |
| **SmoothSync** | **4.996** | **2.749** | **3.959** | **0.2987** | **2.902** |

0，凸显了我们的方法针对相同音频输入生成多样手势变体的独特能力——这是实现自然类人行为的关键要求，而现有方法完全无法解决这一问题。

**跨数据集泛化**。 为进一步验证我们方法的泛化能力，我们在SHOW数据集上开展了额外实验 [3], 。该数据集包含真实场景脱口秀录音，其特征与BEAT2不同。表3展示了我们的方法与TalkSHOW在SHOW数据集上的定量比较。SmoothSync在**所有指标上**均表现出色：FGD提升1.5%，Smooth-BC提升14.7%，内部多样性提高20.3%，抖动降低56.0%。尤为值得注意的是，SmoothSync的跨样本多样性高出26.4%，展现了我们的方法从相同音频输入生成多样化手势序列的独特能力。这些结果验证了我们的框架能够有效泛化至录制条件、说话风格和运动特征各异的多种数据集。

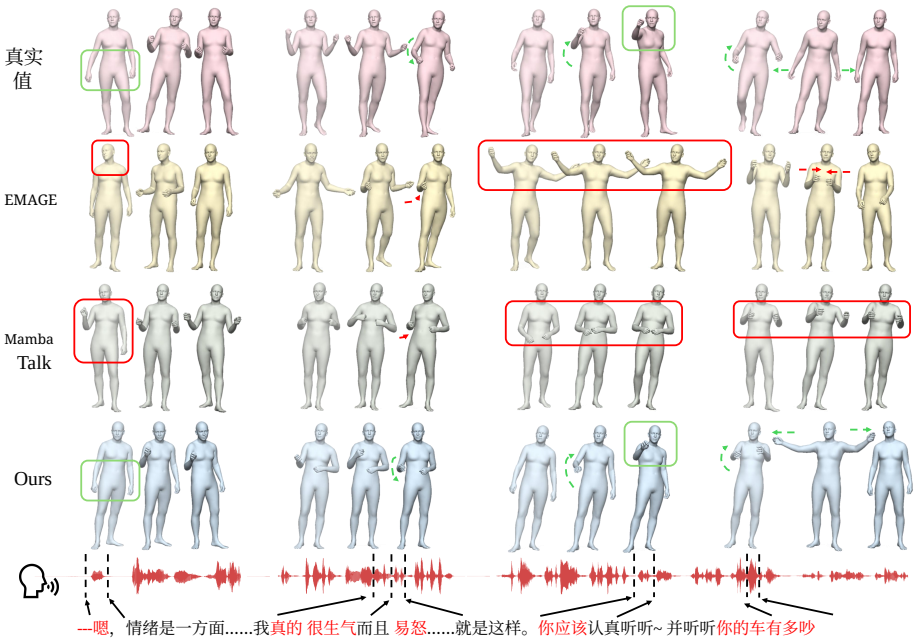


图4：**BEAT2数据集上的比较**。与基线方法相比，我们的方法生成的动作会在语音停顿期间恰当地停顿，能依照语音节奏在强调处做出正确的上下手部动作，并生成更接近真实值的手势；在“你应该”等强语义线索处，手部动作在语义上也更加恰当。在“有多吵”处，SmoothSync生成了幅度较大的向外伸展动作，其垂直方向与真实值并非完全一致，这展现了SmoothSync的多样性能力。

### *4.3.* 定性结果

图4提供了全面的视觉对比，展现了我们的方法在多个具有挑战性的场景中明显的定性优势。

**节奏精度与时间对齐**。 对比结果表明，SmoothSync 在捕捉语音节奏和时间动态方面表现更出色。在语音停顿期间，我们的方法会生成**恰当的静止状态**，与真实行为高度吻合；而基线方法则会持续产生不必要的动作，破坏自然的语音—手势同步。在语音重音标记的强调点上，SmoothSync 会生成**精准的上下手部动作**，使其与语音节奏保持一致，体现了我们的双流架构在捕捉音频 - 动作相关性方面的有效性。

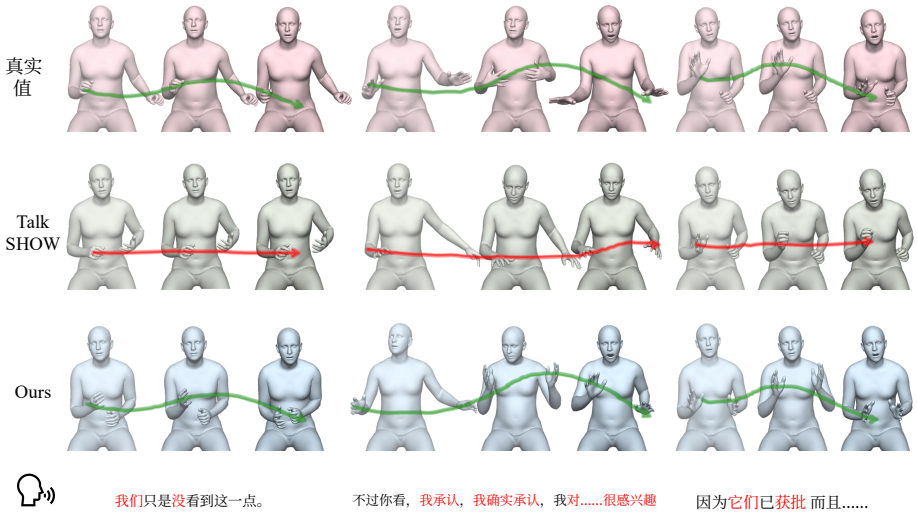


图5：**在SHOW数据集上的比较**。与TalkSHOW有限的动作范围相比，我们的方法生成的手势幅度更大、多样性更高，呈现出的节奏模式也与真实值更加吻合。SmoothSync还展现出更优的语义匹配能力：在“只是”和“它们”等重读词处，我们的方法生成的手势与语音的语义内容和韵律重音更加契合。

**语义连贯性与手势适切性**。 在“You should”等语义突出的位置，SmoothSync 生成的手势**明显更接近真实值**，且手部动作在语义上恰如其分，表明其对语音—手势语义关系的理解更出色。相比之下，EMAGE 生成的手势完全错误，而 MambaTalk 生成的动作则毫无语义。

**多样性与自然变化**。 “音量多大”片段展现了 SmoothSync 独特的多样性能力。我们的方法生成**大幅度向外伸展的动作**，在保持相似手势语义的同时，呈现出执行细节上的自然变化，展示了自然的人类式变异：相同的语音可以伴随不同但同样恰当的手势。

**SHOW数据集上的定性比较**。图5中的定性比较揭示了我们的方法在 SHOW 数据集上的显著优势。TalkSHOW生成的手势运动多样性有限，幅度也较小，而 Smooth-

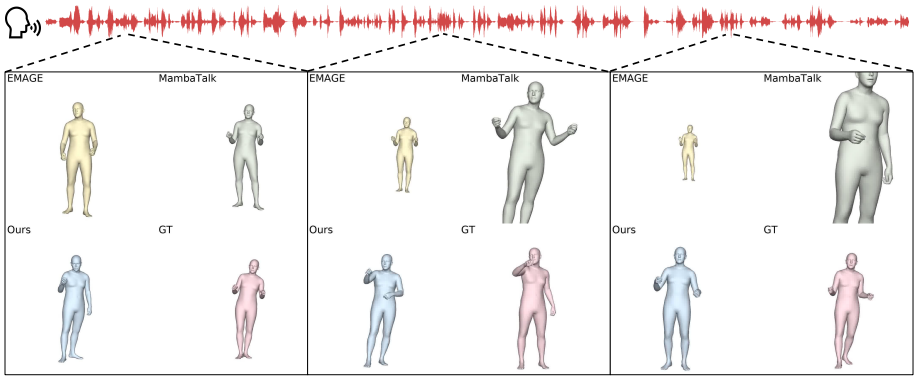


图 6：**长序列全局平移比较**。 与 EMAGE 和 MambaTalk 的运动漂移随时间逐渐累积，导致角色出现不自然的位移不同，我们的方法在扩展序列中展现出更优异的时间一致性和全局平移稳定性。SmoothSync 在整个序列中保持的平移范围与真实行为高度吻合。本比较凸显了我们的方法在生成自然手势的同时维持长期空间连贯性的有效性。

Sync生成的动作幅度更大、变化更丰富，更贴合真实值中观察到的节奏模式。此外，我们的方法在语义韵律对齐方面表现更优——在“just”和“they”等强调词处，SmoothSync生成的手势能恰当地体现语义内容和韵律重音，而TalkSHOW则无法捕捉这些细微的语音-手势关系。这些定性观察与表3中的定量结果相结合，证实了我们的框架在不同数据集和录制条件下具有强大的泛化能力。

**长序列一致性与运动漂移**。 全身手势生成的一项关键挑战，是在扩展序列中保持时间一致性并防止不自然的运动漂移。如图6所示，我们比较了生成的动作在长音频序列中的全局平移轨迹。EMAGE 和 MambaTalk 都受到**严重运动漂移**的影响，累积位移误差导致角色在场景中不自然地滑动。随着时间推移，这种漂移会逐渐加剧，使这些方法不适用于长篇内容生成。与之形成鲜明对比的是，SmoothSync 保持了**出色的长序列一致性**，其

表4：**消融研究**展示了我们框架中各组件的有效性。噪声动作变体使用纯噪声作为运动输入，表明与 BC 指标相比，Smooth-BC 指标对噪声表现出更强的鲁棒性。

| 方法 | FGD↓ | BC↑ | Smooth-BC↑ | 内部多样性↑ | 抖动↓ | 足部滑动↓ | 交互多样性↑ |
|---|---|---|---|---|---|---|---|
| SmoothSync | **3.723** | 7.155 | **4.484** | **14.16** | **0.5008** | **0.6376** | 8.090 |
| **音频特征输入（将量化梅尔能量更改为：）** | | | | | | | |
| 梅尔能量 | 4.798 | 7.236 | 3.753 | 13.10 | 0.9182 | 0.7582 | 7.840 |
| Whisper [10] | 6.011 | **7.918** | 0.3294 | 11.61 | 14.20 | 2.471 | 7.235e-4 |
| Wav2vec2 [11] | 8.914 | 7.858 | 0.1935 | 9.420 | 9.000 | 1.886 | 5.235e-3 |
| **DiT架构（从双流改为：）** | | | | | | | |
| 仅解码器 | 5.964 | 7.497 | 2.988 | 11.92 | 1.418 | 0.7838 | 6.650 |
| **– 抖动损失** | 4.644 | 7.796 | 3.244 | 13.85 | 1.683 | 1.094 | **9.069** |
| **噪声动作** | 66.74 | 7.868 | 0.004965 | 13.57 | 78.66 | 17.29 | 1.564 |

全局平移范围在整个序列中都与真实行为高度吻合。我们的方法有效地将空间运动限制在合理范围内，同时保留自然的手势表现力，展现了双流架构和抖动抑制训练目标在长期时序建模中的稳健性。

### *4.4.* 消融研究

为系统地验证框架中各个组件的有效性，我们在BEAT2数据集上进行了全面的消融实验。我们分析了不同音频特征表示、网络架构和训练目标的影响。

**音**频特征输入。为了验证我们提出的量化梅尔能量的有效性，我们比较了以梅尔能量、Whisper [10], 和 Wav2- vec2 [11] 特征作为输入的实验。如表4所示，与量化梅尔能量相比，使用梅尔能量的实验在 Smooth-BC、内部多样性和交互多样性指标上的表现有所下降，这表明量化增强操作有效提升了模型泛化能力和多样性。使用 Whisper 和 Wav2vec2 特征的实验虽然取得了较高的 BC 分数（7.918、7.858），但它们

表5：**抖动损失权重敏感性**。尽管 $\lambda_{\text{jitter}} = 10^{-9}$ 看起来较小，但它作用于加加速度幅值，并按 fps $^3$ = 27,000 进行缩放。 $\lambda_{\text{jitter}} = 10^{-9}$ 在真实感、多样性和抖动降低之间实现了最佳平衡。

| $\lambda_{\text{jitter}}$ | FGD↓ | Smooth-BC↑ | 多样性↑ | 抖动↓ |
|---|---|---|---|---|
| 0（无损失） | 4.644 | 3.244 | 13.85 | 1.683 |
| $10^{-11}$ | 4.136 | 3.421 | 13.41 | 1.474 |
| $10^{-10}$ | <u>4.113</u> | 4.087 | <u>14.00</u> | 1.076 |
| **$10^{-9}$** | **3.723** | <u>4.484</u> | **14.16** | <u>0.5008</u> |
| $10^{-8}$ | 5.096 | **4.591** | 13.81 | **0.3699** |

抖动指标极高，Smooth-BC 分数则显著偏低（0.3294、0.1935），这表明其较高的 BC 值是由抖动伪影人为抬高的，而非源于真正的节奏对齐。相比之下，我们的量化梅尔能量方法兼具较高的 BC 和稳健的 Smooth-BC，展现出真正的节奏同步，且没有动作伪影。此外，大规模预训练模型的交互多样性分数尤其低，这表明其信息内容过多，导致在 BEAT2 数据集上严重过拟合，并丧失泛化能力。

**双流DiT架构的效**果。为验证我们提出的双流 DiT 架构的有效性，我们对仅解码器架构和双流架构进行了实验比较。在仅解码器架构中，音频条件通过交叉注意力注入网络；而我们的双流架构则在融合之前分别对音频和运动模态进行特征学习。如表4所示，仅解码器架构在所有指标上的表现均逊于双流架构，这表明双流架构能够更好地整合音频和动作信息，生成节奏感更强、更加自然的动作。

**抖**动**损失的影响**。 我们分析了抖动损失及其权重 $\lambda_{\text{jitter}}$ 对生成质量的影响。如表4所示，移除抖动损失后，BC 指标显著提升，而 Smooth-BC 下降，抖动与脚滑指标则大幅上升。这表明，抖动损失能够有效抑制动作抖动，同时保持节奏一致性。

我们进一步对 $\lambda_{\text{jitter}}$ 进行了敏感性分析，如表5所示。我们选定的值 $\lambda_{\text{jitter}} = 10^{-9}$ 取得了最佳的 FGD 和多样性，与不使用抖动损失相比，抖动减少了 70.2%。更大的权重（$10^{-8}$）会使动作过度平滑，降低真实感（FGD ↑37%）和多样性；而较小的权重则不足以有效抑制抖动。

**Smooth-BC 指标的稳健性验证**。 为验证我们提出的 Smooth-BC 指标对动作伪影的稳健性，我们开展了一项诊断实验：在保持原始音频不变的同时，使用纯噪声作为运动输入。如表4（噪声运动行）所示，该噪声序列取得了异常高的 BC分数 7.868，与许多有效的手势生成方法相当，甚至更高，这表明传统 BC 指标受到高频噪声的严重干扰。相比之下，同一噪声序列的 Smooth-BC 分数仅为 0.004965（几乎为零），证实我们提出的指标能够成功过滤由抖动造成的虚假节拍检测。噪声输入下 BC 与 Smooth-BC 分数的巨大差异（7.868 对 0.004965，相差 1,585×）验证了即使存在严重动作伪影，Smooth-BC 仍能提供可靠的节奏评估，因此是评估手势生成质量时更值得信赖的指标。

## 5. 结论

本文介绍了 SmoothSync，这是一种新颖的双流扩散变换器框架，用于生成高质量、与节拍同步的伴语手势。我们的方法通过引入双流架构，有效捕捉音频与动作之间的跨模态关系；采用抖动抑制损失，在保持自然节奏的同时显著减少动作伪影；并使用量化梅尔频谱特征，使相同输入能够生成多样化的手势。此外，我们还提出了 Smooth-BC 指标，通过滤除噪声干扰，更可靠地评估节奏对齐情况，从而弥补传统评估指标的局限性。在 BEAT2 和 SHOW 数据集上开展的全面实验表明，Smooth-

Sync 达到了最先进性能，能够生成高质量、节奏同步、无抖动且多样化的伴语手势。

## 附录A. 评估指标

**弗雷歇手势距离 (FGD)**。 FGD 指标 [37] 量化了真实手势序列与合成手势序列之间的分布相似性。该指标按照生成模型感知评估中确立的方法，计算通过预训练编码器网络提取的特征分布之间的瓦瑟斯坦-2距离：

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

其中，$\mu_r, \Sigma_r$ 表示从真实手势序列$z_r$ 中得到的潜在表示的均值和协方差$\mathbf{g}$，而$\mu_g$，$\Sigma_g$ 表示从生成的动作$\hat{\mathbf{g}}$中得到的统计量。我们使用 EMAGE [4] 在 BEAT2数据集上提供的预训练网络提取特征。

**内部多样性**。 多样性指标 [38] 通过成对动作分析评估手势变化程度，得分越高表示动作越丰富。该指标计算所有手势配对之间的平均 L1 范数：

$$\text{Intra-Diversity} = \frac{1}{2N(N-1)} \sum_{l=1}^{N} \sum_{j=1}^{N} \left\| p_l^i - p_j^i \right\|_1, \tag{A.2}$$

其中，$p_l$ 表示时间帧$l$上的关节坐标。多样性评估涵盖整个评估数据集。值得注意的是，计算关节位置时会排除全局平移分量，以确保该指标仅反映局部关节动作变化。

**节拍一致性（BC）**。BC 指标 [32]量化了手势节律与声学韵律模式之间的时间同步。BC值越高，表示音频-动作对齐效果越好。音频节拍对应语音起始检测，而运动节拍则根据上半身发音动作中速度极小值确定（不包括手指关节）。一致性计算如下：

$$\mathrm{BC} = \frac{1}{g} \sum_{b_g \in g} \exp\left(-\frac{\min_{b_a \in a} \|b_g - b_a\|^2}{2\sigma^2}\right), \tag{A.3}$$

其中，$g$和$a$ 分别表示检测到的手势和声学节拍时间戳集合。

**Smooth-BC.**我们的分析表明，传统的 BC 评估对抖动十分敏感，高频噪声会人为抬高同步分数。Smooth-BC 指标通过结合带有斜率约束的峰值检测，解决了这一缺陷，从而有效区分真实的节奏模式与抖动引起的虚假波动。

Smooth-BC 算法首先计算上半身速度：

$$v_t = \|\mathbf{J}_{t+1} - \mathbf{J}_t\|_2, \tag{A.4}$$

其中，$\mathbf{J}_t$ 表示时间帧$t$上的三维关节坐标。

随后，算法使用带有斜率约束的稳健峰值检测算法，将速度极小值识别为运动节拍。对于速度序列 $\mathbf{v} = [v_1, v_2, \ldots, v_T]$，我们首先将速度取反，以便将极小值作为峰值检测：$\tilde{\mathbf{v}} = -\mathbf{v}$。峰值集合（对应速度极小值）定义如下：

$$\mathcal{P} = \{i \mid w \le i \le T - w, \text{ satisfying conditions } C_1, C_2, C_3\}, \tag{A.5}$$

其中，$w$ 是局部窗口半径。三个条件如下：

$$C_1: \quad \max_{j=i-w}^{i+w-1} \tilde{v}_j \ge h_{\min} \quad \text{(minimum height)}, \tag{A.6}$$

$$C_2: \quad \tilde{v}_i = \max_{j=i-w}^{i+w-1} \tilde{v}_j \quad \text{(local maximum in inverted velocity)}, \tag{A.7}$$

$$C_3: \quad \tilde{v}_{i-w} \le \cdots \le \tilde{v}_{i-1} \le \tilde{v}_i \ge \tilde{v}_{i+1} \ge \cdots \ge \tilde{v}_{i+w-1}$$

$$\text{(slope constraint)}, \tag{A.8}$$

其中，$h_{\min}$ 是最小峰值高度阈值，$w$ 是局部窗口半径，$i$ 是候选峰时间索引。这些约束确保只有在时间上连贯的速度极小值（即动作暂停处）会参与节拍检测，从而滤除原本会被误判为虚假节拍的高频抖动。

**抖动**。 为量化运动抖动伪影，我们计算三维关节位置的三阶有限差分

（加加速度），具体如下： [13]

$$\mathbf{J}^{\text{jerk}}_{t,j} = \mathbf{J}_{t+1,j} - 3\mathbf{J}_{t,j} + 3\mathbf{J}_{t-1,j} - \mathbf{J}_{t-2,j}, \tag{A.9}$$

$$\text{Jitter} = \frac{1}{T-3}\sum_{t=3}^{T}\frac{1}{J}\sum_{j=1}^{J}\left\|\mathbf{J}^{\text{jerk}}_{t,j}\right\|_2 \times \text{fps}^3, \tag{A.10}$$

其中$\mathbf{J}_{t,j}$ 表示第$t$帧关节$j$ 的三维位置，$T$ 表示总帧数，$J$ 表示关节数量，fps 表示帧率。该公式通过计算三阶差分来衡量动作平滑度，抖动值越高，表示时间不连续性越明显。

**足部滑动**。 在地面接触阶段，若双脚看起来在地面上滑动，就会产生脚部滑动伪影。按照 [13], ，我们通过计算地面接触期间预测足部顶点的移动误差来衡量脚部滑动：

$$\text{Foot-Sliding} = \frac{1}{|C|}\sum_{t\in C}\left\|\mathbf{F}^{\text{pred}}_{t+1} - \mathbf{F}^{\text{pred}}_{t}\right\|_2, \tag{A.11}$$

其中， $C = \{t | \|\mathbf{F}^{\text{gt}}{}_{t+1} - \mathbf{F}^{\text{gt}}{}_{t}\|_2 < \text{thr}\}$ 是足部接触帧的集合，$\mathbf{F}^{\text{gt}}{}_{t}$ 和$\mathbf{F}^{\text{pred}}{}_{t}$ 分别表示第$t$帧的真实和预测足部顶点位置。接触阈值 thr 设为 $10^{-2}$。该指标仅在预测的双脚应保持静止时计算其移动误差。

**交互多样性**。 为了衡量由同一输入生成的多个样本之间的多样性，我们计算多模态分数，以评估使用不同随机种子生成的手势序列之间的成对差异：

$$\text{Inter-Diversity} = \frac{1}{N \times \lceil \frac{N}{2} \rceil}\sum_{a=1}^{N}\sum_{b=a+1}^{N}\left\|\hat{\mathbf{M}}_a - \hat{\mathbf{M}}_b\right\|_1, \tag{A.12}$$

其中，$N$ 表示使用不同随机种子，从相同音频输入生成的动作样本数量；$\mathbf{M}_a$ 和 $\mathbf{M}_b$ 表示在相同音频条件下生成的不同动作序列； $\|\cdot\|_1$ 表示 L1 范数，即所有元素绝对差值的均值。交互多样性分数越高，表明模型越能根据相同输入生成多样的动作，反映出模型避免模式崩溃的能力越强。

我们报告 FGD $\times 10^{-1}$、BC $\times 10^{-1}$、Smooth-BC $\times 10^{-1}$、抖动 $\times 10^{2}$、足部滑动$\times 10^{-2}$、交互多样性 $\times 10^{-2}$。

## 附录 B. 计算效率分析

我们提供详细的计算指标，以展示我们方法的实际效率：

**推理性能**。 我们的方法具有高效的推理性能，适用于实时应用。在单张 RTX 4090 GPU 上使用 DDIM 进行 50 步采样时，SmoothSync 的动作生成速度为 160 FPS（每秒帧数），能够以 5.3× 倍实时速度实时生成 30fps 的动作序列。这一性能使我们的方法适用于交互式应用和现场表演场景。

**模型复杂度**。 完整的 SmoothSync 模型包含 154.47M 个参数，其中大部分参数来自双流 Transformer 网络。该模型的规模与其他基于 Transformer 的手势生成方法相当，同时性能更优。

**内存需求**。 处理 320 帧的序列（以 30fps 计算，约为 10.67 秒）时，我们的方法在推理期间需要 2.28 GB GPU 内存。较低的内存占用使其能够部署在 GPU 规格适中的消费级硬件上，从而便于在实际应用中使用。**可扩展性**。 我们高效的架构支持交互式应用，并可部署在消费级硬件上。该方法可通过基于片段的生成来生成较长的动作序列，且不会显著增加内存开销，因为每个片段都是独立处理的。

## 附录 C. 演示视频与定性比较

我们在补充材料中提供了 EMAGE、MambaT-alk 和 SmoothSync 之间全面的定性比较，其中包括演示视频，展示各方法的生成质量，文件夹路径为 Comparison_Videos_with_Baselines/。

对比视频清楚地展示了 EMAGE 和 MambaTalk 生成的动作存在明显的抖动伪影和过度的全局平移漂移，生成的角色会在场景中出现不自然的滑动动作。相比之下，SmoothSync 生成的动作非常稳定，其全局平移范围

与真实行为高度一致。如图6（第4.3节）所示，对扩展序列的全局平移比较表明，EMAGE 和 MambaTalk 的漂移会随时间逐渐累积，而我们的方法能够保持与真实行为高度一致的平移范围，确保整个序列中的角色定位真实自然。

此外，我们的方法生成的动作在保持时间平滑性的同时，与语音音频展现出更强的节奏对齐。生成的手势能够有效传达丰富的语义含义，体现出相较于基线方法对语音-手势关系的更深入理解。

此外，我们在 Inter-Diversity Videos/ 文件夹中提供了演示视频，展示了由相同音频输入生成的多个不同手势样本，凸显了我们的方法在交互多样性方面的优势。相比之下，EMAGE 和 MambaTalk 只能生成确定性输出，完全缺乏人类说话者所展现的自然变化。

**附录D. 数据集详情和许可证**

**BEAT2数据集**。 BEAT2数据集包含来自25位说话者（12位女性、13位男性）的60小时同步音频-动作数据。该数据集分为用于训练的 BEAT2-standard（27小时）和用于增强鲁棒性的 BEAT2-additional（33小时）。

遵循 EMAGE [4], ，我们在 BEAT2-standard 上采用 85%/7.5%/7.5% 的训练/验证/测试划分。为与现有方法进行公平比较，我们的实验聚焦于说话者2数据。

该数据集包含 1762 个序列，每个序列的平均时长为 65.66 秒。每段录音都记录了参与者回应日常问题时自然产生的对话手势。

许可证：BEAT2 根据 Apache-2.0许可证 公开发布。我们遵循数据集作者规定的所有数据使用条款和条件。

**SHOW数据集**。 SHOW数据集 [3]包含来自四位说话者的 26.9 小时脱口秀录音，这些录音是在自然、非受控环境下采集的。动作序列为

动作序列标注了 SMPLX 全身参数，帧率为 30 fps，并与采样率为 22 kHz 的对应音频轨道同步。

遵循 TalkSHOW [3], 中的实验方案，我们剔除时长短于三秒的序列，并按 80%/10%/10% 的比例划分训练集、验证集和测试集。该数据集记录了真实脱口秀场景中的自发对话手势，涵盖多种说话风格。这些特点使 SHOW 成为评估模型鲁棒性及其对真实世界场景泛化能力的宝贵基准。

许可：SHOW 数据集仅供非商业用途。我们遵守所有许可条款，且仅将该数据集用于学术研究。

**附录 E. 基线复现详情**

**EMAGE [4]**。 我们使用作者提供的官方实现和预训练模型进行推理。该模型采用 VQ-VAE 方法进行掩码音频手势建模，并针对不同身体部位（身体、手部、面部和全局平移）分别使用四个 VQ-VAE。

**MambaTalk [5]**。 我们使用作者提供的官方实现和预训练模型进行推理，并遵循论文的方法，采用基于 Mamba 的架构进行整体手势合成。

**TalkSHOW [3]**。 对于SHOW数据集上的实验，我们使用作者提供的官方实现和预训练模型进行推理。TalkSHOW采用基于VQ-VAE的自回归模型，根据音频输入生成全身手势。

**评估协议**。我们使用不同的随机种子为每个测试序列生成 4 个不同的样本，以计算交互多样性指标。为确保公平比较，所有方法均使用相同的测试集和评估脚本。我们使用完全相同的关节集合和坐标系实现所有指标，以确保评估一致。

**参考文献**

[1] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A.Owens, J. Malik, 《学习对话手势的个体风格》，载于：IEEE/CVF计算机视觉与模式识别会议论文集，2019年，第3497–3506页。

[2] I. Habibie, W. Xu, D. Mehta, L. Liu, H.-P. Seidel, G. Pons-Moll, M. Elgharib, 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。

[3] H. Yi, H. Liang, Y. Liu, Q. Cao, Y. Wen, T. Bolkart, D. Tao, M. J. Black, 根据语音生成整体三维人体运动，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。

[4] H. Liu, Z. Zhu, G. Becherini, Y. Peng, M. Su, Y. Zhou, X. Zhe, N. Iwamoto, B. Zheng, M. J. Black，Emage：通过表达性掩码音频手势建模实现统一的整体性共语手势生成，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] Z. Xu、Y. Lin、H. Han、S. Yang、R. Li、Y. Zhang、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，发表于：第三十八届神经信息处理系统年会，2024年。

[6] L. Zhu、X. Liu、X. Liu、R. Qian、Z. Liu、L. Yu，驯服扩散模型以实现音频驱动的伴语手势生成，发表于：IEEE/CVF计算机视觉与模式识别会议论文集，2023年，第10544–10553页。

[7] S. Alexanderson、R. Nagy、J. Beskow、G. E. Henter，聆听、去噪、行动！利用扩散模型实现音频驱动的动作合成，《ACM图形学汇刊》（TOG）42（4）（2023年）1–20。

[8] C. Fu、Y. Wang、J. Zhang、Z. Jiang、X. Mao、J. Wu、W. Cao、C. Wang、Y. Ge、Y.Liu，《MambaGesture：通过 Mambaand

解耦多模态融合增强伴语手势生成》，载于：第 32 届 ACM 国际多媒体会议论文集，2024年，第 10794–10803 页。

[9] X. Zhang, J. Li, J. Zhang, J. Ren, L. Bo, Z. Tu，EchoMask：基于语音查询的注意力掩码建模，用于整体伴随语音动作生成，载于：第33届ACM国际多媒体会议论文集，2025年，第10827–10836页。

[10] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever，基于大规模弱监督的鲁棒语音识别，载于：国际机器学习会议，PMLR，2023年，第28492–28518页。

[11] A. Baevski, Y. Zhou, A. Mohamed, M. Auli，wav2vec 2.0：一种用于语音表示自监督学习的框架，《神经信息处理系统进展》第33卷（2020年），第12449–12460页。

[12] Y. Zhi、X. Cun、X. Chen、X. Shen、W. Guo、S. Huang、S. Gao，《LivelySpeaker：面向语义感知的伴语手势生成》，载于：IEEE/CVF国际计算机视觉会议论文集，2023年，第20807–20817页。

[13] Z. Shen、H. Pi、Y. Xia、Z. Cen、S. Peng、Z. Hu、H. Bao、R. Hu、X. Zhou，《通过重力视角坐标实现面向世界的人体运动恢复》，载于：SIGGRAPH Asia 2024 Conference Papers，2024年，第1–11页。

[14] J. Cassell、C. Pelachaud、N. Badler、M. Steedman、B. Achorn、T. Becket、B. Douville、S. Prevost、M. Stone，《动画对话：为多个对话代理基于规则生成面部表情、手势和言语语调》，载于：第21届计算机图形学与交互技术年度会议论文集，1994年，第413–420页。

[15] J. Cassell、H. H. Vilhjálmsson、T. Bickmore，《BEAT：行为表达动画工具包》，载于：第 28 届计算机图形学与交互技术年会论文集，2001 年，第 477–486 页。

[16] S. Kopp, I. Wachsmuth，为对话代理合成多模态话语，《计算机动画与虚拟世界》15(1)（2004）39–52。

[17] M. Kipp, M. Neff, K. H. Kipp, I. Albrecht，迈向自然手势合成：以数据驱动方法评估手势合成中的手势单元，载于：智能虚拟代理国际研讨会，施普林格，2007年，第15–28页。[18] S. Levine, P. Krähenbühl, S. Thrun, V. Koltun，手势控制器，《ACM图形学汇刊》29(4)（2010）1–11。[19] X. Zhang, J. Li, J. Zhang, Z. Dang, J. Ren, L. Bo, Z. Tu，SemTalk：通过帧级语义强调实现整体伴语动作生成，载于：IEEE/CVF国际计算机视觉会议论文集，2025年。[20] X. Mao, Z. Jiang, Q. Wang, C. Fu, J. Zhang, J. Wu, Y. Wang, C. Wang, W. Li, M. Chi，MDT-A2G：探索用于伴语手势生成的掩码扩散Transformer，载于：第32届ACM国际多媒体会议论文集，2024年，第3266–3274页。[21] J. Chen, Y. Liu, J. Wang, A. Zeng, Y. Li, Q. Chen, DiffSHEG：一种基于扩散的实时语音驱动整体三维表情与手势生成方法，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。[22] S. Wang, J. Zhang, X. Tan, Z. Xie, C. Wang, L. Ma，MMoFusion：基于扩散模型的多模态伴语动作生成，《模式识别》（2025）111774。[23] C. Gu, C. Zhang, S. Kuriyama，面向动作驱动人体运动预测的方向感知腿部动作学习，《模式识别》150（2024）110317。[24] J. Dai, H. Li, R. Zeng, J. Bai, F. Zhou, J. Pan，Kd-former：用于三维人体运动预测的运动学与动力学耦合Transformer网络，《模式识别》143（2023）109806。[25] X. Liu, Q. Wu, H. Zhou, Y. Xu, R. Qian, X. Lin, X. Zhou, W. Wu, B. Dai, B. Zhou，学习用于伴语手势生成的分层跨模态关联，载于：IEEE/CVF计算机视觉与模式识别会议论文集，2022年，第10462–10472页。

[26] R. Wu, Y. Yu, F. Zhan, J. Zhang, X. Zhang, S. Lu, 《音频驱动的说话人脸生成：兼具多样性与真实感的面部动画》, 《模式识别》144（2023）109865。[27] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, M. J. Black, 《表现力身体捕捉：从单张图像中恢复三维手部、面部和身体》, 载于：IEEE/CVF计算机视觉与模式识别会议论文集, 2019年，第10975–10985页。[28] Y. Cheng, S. Huang, X. Chen, J. Ning, M. Gong, 《Didiffges：用于从语音实时生成手势的解耦半隐式扩散模型》, 载于：AAAI人工智能会议论文集，第39卷，2025年，第2464–2472页。[29] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, 《HuBERT：通过掩码预测隐藏单元进行自监督语音表征学习》, IEEE/ACM音频、语音与语言处理汇刊29（2021）3451–3460。[30] S. S. Stevens, J. Volkmann, E. B. Newman, 《心理音高量值测量尺度》, 美国声学学会杂志8（3）（1937）185–190。[31] D. P. Ellis, 《基于动态规划的节拍跟踪》, 《新音乐研究杂志》36（1）（2007）51–60。[32] R. Li, S. Yang, D. A. Ross, A. Kanazawa, 《AI编舞师：基于音乐条件的三维舞蹈生成》, 载于：IEEE/CVF国际计算机视觉会议论文集，2021年，第13401–13412页。[33] B. Li, Y. Zhao, S. Zhelun, L. Sheng, 《DanceFormer：基于音乐条件的三维舞蹈生成与参数化动作变换器》, 载于：AAAI人工智能会议论文集，第36卷，2022年，第1272–1279页。[34] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel等, 《扩展用于高分辨率图像合成的整流流变换器》,

载于：第四十一届国际机器学习会议，2024年。

[35] Y. Zhang, J. Gu, L.-W. Wang, H. Wang, J. Cheng, Y. Zhu, F. Zou, 《Mimicmotion：基于置信度感知姿态引导的高质量人体动作视频生成》，载于：国际机器学习会议，2025年。

[36] H. Liu, Z. Zhu, N. Iwamoto, Y. Peng, Z. Li, Y. Zhou, E. Bozkurt, 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，施普林格，2022年，第612–630页。

[37] Y. Yoon, B. Cha, J.-H. Lee, M. Jang, J. Lee, J. Kim, 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）39（6）（2020）1–16。

[38] J. Li、D. Kang、W. Pei、X. Zhe、Y. Zhang、Z. He、L.Bao, Audio2gestures：使用条件变分自编码器从语音音频生成多样化手势，载于：IEEE/CVF国际计算机视觉会议论文集，2021年，第11293–11302页。

[39] H.Liu、N.Iwamoto、Z. Zhu、Z.Li、Y. Zhou、E. Bozkurt、B.Zheng, DisCo：用于多样化伴语手势合成的解缠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。

[40] S. Yang、Z. Wu、M. Li、Z. Zhang、L. Hao、W. Bao、M. Cheng、L. Xiao，Diffusestylegesture：使用扩散模型进行风格化音频驱动的伴随语音手势生成，载于：第三十二届国际人工智能联合会议论文集，2023年，第5860–5868页。