# AQ-GT：用于伴随语音手势合成的时间对齐量化 GRU-Transformer

Hendric Voß
h
voss@techfak.uni-bielefeld.de
社会认知系统组 比勒费尔德大学
德国

Stefan Kopp
s
kopp
@techfak.uni-bielefeld.de 社会
认知系统组 比勒费尔德大学 德国

arXiv:2305.01241v2 [cs.HC] 8 May 2023

## 摘要

生成逼真且与上下文相关的伴随语音手势，是创建多模态人工智能体过程中一项颇具挑战、但日益重要的任务。以往方法侧重于学习伴随语音手势表征与生成动作之间的直接对应关系，这虽然能产生看似自然的手势，但在人类评估中往往缺乏说服力。我们提出一种方法，利用带有量化流程的生成对抗网络对部分手势序列进行预训练。由此得到的码本向量在我们的框架中同时作为输入和输出，构成手势生成与重建的基础。通过学习潜在空间表征之间的映射，而非直接将其映射到向量表征，该框架能够生成高度逼真且富有表现力的手势，充分复现人类的运动与行为，同时避免生成过程中的伪影。我们通过将该方法与成熟的伴随语音手势生成方法以及现有的人类行为数据集进行比较，对其进行了评估。我们还开展了消融研究以检验研究结果。结果表明，我们的方法明显优于当前最先进水平，且生成的手势在一定程度上与人类手势难以区分。我们公开提供数据流程和生成框架。

## CCS 概念

• **以人为中心的计算** → **交互系统与工具**；交互设计实证研究； **人机交互理论、概念与模型**； • **计算方法** → **神经网络**；**学习潜在表征**；无监督学习。

## 关键词

机器学习，深度学习，伴随语音手势，手势合成，多模态数据，量化，Transformer，门控循环单元，时间对齐

允许为个人或课堂使用而免费制作本作品全部或部分内容的数字副本或纸质副本，前提是这些副本不得用于营利或商业利益，也不得进行复制或分发，且副本须附上本声明及首页的完整引文。对于本作品中归 ACM 以外其他方所有的组件，必须尊重其版权。注明出处后可以摘录。若要以其他方式复制、转载、发布到服务器或分发至列表，则须事先获得特别许可和/或支付费用。请向 permissions@acm.org 申请许可。*Conference'17*，*2017*年7月，美国华盛顿哥伦比亚特区© 2023 Association for Computing Machinery。ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00 https://doi.org/XXXXXXX.XXXXXXX

**ACM 参考格式：**
Hendric Voß 和 Stefan Kopp。2023。AQ-GT：用于伴随语音手势合成的时间对齐量化 GRU-Transformer。载于 *ACM* 会议论文集（*Conference'17*）。ACM，纽约，NY，美国，10 页。https://doi.org/XXXXXXX.XXXXXXX

## 1 引言

人类之间的有效交流涉及多种模态，例如语音、面部表情和身体手势。理解和生成这些多模态信号的能力，使我们能够进行有意义且细致入微的对话，并且常见于日常互动中 [10, 45]。因此，研究人员投入了大量精力，致力于准确处理、解读和生成这些交流线索，以促进自然、流畅的人机交互 [30]。

对于具身智能体而言，生成肢体语言和手势等非语言线索来配合口语，是一项尤其具有挑战性的任务。近年来，研究人员开发了机器学习方法，用于根据给定的语言输入预测手势。这类语音驱动手势合成框架在生成与给定口语话语或文本相匹配的逼真手势运动方面取得了令人瞩目的成果，但在生成兼具人类手势的交际有效性和丰富性的动作方面仍面临困难 [37]。

本文提出了一种旨在学习手势序列离散潜在空间的框架。所提出的 AQ-GT 方法结合门控循环单元（GRU）与 Transformer，学习向量的中间表示，并通过时间对齐网络架构构建全新的伴随语音手势。为此，我们为手势合成领域作出了若干贡献。首先，我们提出了一种全新的数据处理流程，可对任意单目视频中的完整三维身体关节进行跟踪。该流程能够自动获取精确的手势信息，从而有助于学习自然的手势。其次，我们提出了一种全新的伴随语音手势合成方法。该方法利用手势的离散潜在空间，结合门控循环单元（GRU）与 Transformer 的优势，生成与伴随语音紧密对齐的自然手势。第三，我们对所提出的框架进行了全面评估，同时采用客观和主观指标，结果表明，该框架优于当前最先进的方法，生成的手势在一定程度上与人类手势难以区分。最后，我们开展了消融研究，分离并分析所提出模型中各个组成部分对

整体输出质量的贡献。我们已在线提供一段展示伴随语音手势生成示例的视频[1]。

## 2 相关工作

生成伴随语音手势一直是社会交互式多模态智能体（如具身对话智能体、3D 虚拟形象或社会辅助机器人）的长期目标。早期方法主要依赖人工设计的基于规则的模型，这些模型基于经验或理论见解构建。其中一个例子是行为标记语言 [30]，它允许通过在 XML 格式中指定功能到形式的映射来生成多模态行为。类似地，BEAT 工具包 [11]采用可扩展的规则集，这些规则集源自对人类对话行为的语言学和语境分析。后续方法采用动态学习方法，主要使用基于核的概率分布模型或随机马尔可夫模型，从动作捕捉序列中学习动作模式，或从时间分段中学习高层结构 [9, 20]。由于这类技术具有高度可定制性且计算需求低，因此也被用于 Pepper 和 Nao 等商用机器人 [38]。

近年来，深度机器学习的发展对多模态行为处理和虚拟智能体行为生成领域产生了重大影响 [12, 28]。因此，数据驱动的方法已成为各种人类与智能体、人类与人类互动任务中生成复杂行为的主要方法 [42, 51]。

在深度学习发展的早期阶段，主要目标是训练虚拟智能体展现自然行为，并生成视觉效果吸引人的伴随语音手势，而无需耗时的手工设计技术。研究重点是使用单一输入模态。早期尝试中，Fan 等人进行的一项研究 [17] 采用双向长短期记忆（LSTM）模型，仅根据文本输入和给定的初始手势输入生成新的伴随语音手势序列。类似地，Ferstl 和 McDonnell [19] 采用了具有编码器—解码器架构的循环神经网络，并使用韵律语音特征进行训练，以生成短动作序列；而 Ginosar 等人[22] 则提出了一种结合语音音频数据和初始姿势来生成手势的模型。

较新的研究纳入了多种输入模态，各自具有不同的维度和结构，并推动了更复杂的数据驱动方法的发展。在对伴随语音手势的研究中，Holden 等人 [26] 采用前馈网络，并通过循环函数对其进行训练以生成手势。相比之下，Henter 等人 [25] 采用了不同的深度学习技术，使用带有可逆 1x1 卷积的 Glow 网络来生成伴随语音手势。Ling 等人 [33] 采用了相关方法，将变分自编码器与深度强化学习相结合，以实现伴随语音手势的目标导向控制。

当前研究凸显了生成对抗网络（GAN）在生成高度逼真结果方面的潜力，其能够协调融合多种模态，从而增强对人类交流的表征。Ahuja 等人 [4] 设计了一种用于学习风格嵌入的方法，使模型能够为不同说话者生成独特的手势风格。该方法支持生成丰富多样的手势变化，并可在生成过程中进行动态调整。为了在 GAN 框架中整合音频、文本和说话人身份，Yoon 等人[48]设计了一种技术，能够有效生成与自然人类交流中观察到的手势高度相似的手势。该方法生成的手势能够反映口语词语的语义和节奏，从而在虚拟环境中模拟真实的人类互动。同样，Ao 等人[6] 最近提出了一种兼顾节奏和语义方面的手势合成方法。具体而言，他们将基于节奏的分段流程与语音和动作的神经嵌入相结合，并借鉴了语言学理论中的相关见解。

## 3 数据

目前已有多个用于训练语音伴随手势生成算法的数据集。然而，这些数据集通常存在一定的局限性 [37]。例如，许多数据集仅包含少量手势序列，而其他数据集则仅提供身体关节的二维追踪信息，或者未纳入手部追踪信息，例如手指动作。据我们所知，只有两个数据集提供三维身体关节信息，且数据规模足以满足我们的研究需求：BEAT 数据集 [34] 和 Talking with Hands 数据集 [31, 50]。遗憾的是，这两个数据集涵盖的说话者数量都相当有限，这可能会限制模型在不同说话者风格之间的泛化能力。因此，我们决定自行收集并构建一个数据集，称为 BiGe 数据集。

### 3.1 数据采集

为了成功训练伴随语音手势生成算法，所提供的数据必须满足某些要求。尤其是，数据中的大量噪声或显著的可变性会对生成的手势产生不利影响。由于我们的目标是捕捉演讲者的全身手势，尤其是手势，因此需要确保人物完整地出现在画面中。此外，人物必须面向摄像机，以避免手势被遮挡。考虑到消除摄像机运动颇具挑战，我们还要求视频来源使用固定摄像机。为满足这些标准，我们选择使用与 Yoon 等人相同的数据来源 [49],，即 YouTube 上的官方 TED 频道 [1]。为了通过多语言演讲丰富我们的数据集，我们还选用了 TEDx Talks Youtube 频道，其中收录了不同语言的国际演讲 [2]。我们总共收集了 6945 个视频，采用尽可能高的视频和音频格式，并包含自动及人工标注的字幕。接下来，我们筛除了没有可用字幕、分辨率低于 1280x720 像素，或缺少逐词时间信息字幕的视频。经过筛选后，数据集中剩下 4327 个视频，总时长为 1021 小时。

### 3.2 数据处理

为了从收集的视频中提取人体关节关键点信息，创建了一个经过修改的数据处理流程，其基础是

[1]https://vimeo.com/823756031

**表 1：** BiGe 数据集与感兴趣镜头（SoT）概览

| | |
|---|---|
| 视频数量 | 4327 |
| 视频总时长 | 1021 小时 |
| 带有 SoT 的视频 | 2756 |
| SoT 的总长度 | 260.6 小时 |
| SoT 的平均长度 | 17 s |
| SoT 数量 | 54,360（平均每段视频 20 个） |

Yoon 等人 [49] 的工作。为了跟踪视频中每个人的 59 个 3D 全身关节，我们结合使用了 AlphaPose 库 [18], 、VideoPose3D 模型 [39], 和 MediaPipe Hands 模型 [52]。该流程首先通过 AlphaPose 库中的 YOLOv3 模型 [41], 检测视频每一帧中的所有人物。然后，通过 FastPose 模型 [53] 的实现估计每个关节的 2D 位置。随后，使用 VideoPose3D 模型 [39] 将跟踪到的、不包括手部的关节转换为 3D 坐标。最后，使用 MediaPipe Hands 模型，通过将此前跟踪到的 2D 手部关节提升为 3D 坐标，确定手部的 3D 位置。完成跟踪后，我们根据检测到的镜头切换将视频剪辑成多个片段。随后，我们为整个数据集确定感兴趣镜头（SoT），即满足 Yoon 等人 [49] 所确立的移除条件的所有片段。由于跟踪流程为每个关节提供了置信度分数，我们使用该分数进一步细化 SoT 的选择，移除其中中位数低于阈值 $\tau$ 的 SoT；我们将该阈值设为 0.05。表 1 概述了所得 BiGe 数据集的特征：该数据集包含 2756 个视频，从中提取了 54.360 个 SoT，平均长度为 17 秒。为创建数据集，我们将这些数据划分为 2156 个训练视频、300 个验证集视频和 300 个测试视频。

## 4 AQ-GT 架构

所提出模型架构的整体结构如图 1 所示。该模型由两部分组成：第一部分（以黄色标示）是一个量化变分自编码器，用于为序列数据生成结构化嵌入空间表示，涵盖原始音频和手势数据。第二部分（以绿色标示）是一个根据音频、文本和说话人 ID 输入，以自回归方式生成手势帧的模型。该部分包括输入数据的预处理与整合、一个 GRU-Transformer、一个用于说话人预测的网络，以及一个时间对齐器。我们在模型名称“A Q-GT”中强调了时间对齐器的使用，以及 GRU-Transformer 所采用的量化潜在空间。下面将介绍各个组件及其训练过程。

### 4.1 离散潜在空间中的量化

受 Esser 等人工作的启发， [15], 我们采用向量量化变分自编码器（VQ-VAE） [43] 与生成对抗网络（GAN） [23, 35]相结合的方法。我们将 VQ-VAE-2 模型 [40] 与带散度惩罚的 Wasserstein 生成对抗网络结合使用（WGAN-div） [47] 来学习简洁手势序列的离散潜在空间。原始 VQ-VAE 模型 [43] 由两个主要组件构成：编码器和解码器，两者共用同一个“码本”。编码器将观测映射为一系列离散潜变量，而解码器则利用这些离散变量重建原始观测。为此，编码器使用从输入空间 $x$到向量 $E$（$x$）的非线性映射，然后根据该向量与共享码本中原型向量 $c_i$， $i \in 1...C$之间的距离对其进行量化。Razavi 等人 [40] 提出的 VQ-VAE-2 模型可视为 VQ-VAE 的扩展，它引入了层次结构。在该模型中，每个层次级别都会学习一个码本，分别记为 $C_{top}$、 $C_{bottom}$，两个码本的组合记为 $C$。用于训练 VQ-VAE-2 的损失函数为：

$$\mathcal{L}_{vq}(\mathrm{x}, D(\mathrm{z})) = ||\mathrm{x} - D(\mathrm{z})||_2^2 + ||sg[E(\mathrm{x})] - \mathrm{z}||_2^2 + \alpha||sg[\mathrm{z}] - E(\mathrm{x})||_2^2 \tag{1}$$

其中， $E$ 表示编码器， $D$ 表示解码器， $sg$ 指停止梯度， $\alpha$ 是控制正则化项影响的超参数。公式 1 中的第一项衡量输入与解码器输出之间的重构误差，而第二项和第三项分别促使模型使用码本中的编码，并最小化编码器输出与所选编码之间的距离。

量化模型中一个值得注意的问题是码本坍塌，即码本中的很大一部分被弃用，只有有限的子集得到利用。尽管 VQ-VAE-2 模型通过引入层次化学习增强了码本的稳定性，但这一问题仍可能出现。Esser 等人 [15] 近期的研究表明，将 VQ-VAE 与 GAN 相结合，能够得到感知上更多样且训练更稳定的码本。鉴于这些发现，我们在 VQ-VAE-2 模型的对抗训练中加入了一个判别器，该判别器处理 VQ-VAE-2 的解码输出，并输出一个无界标量值。借鉴 Wu 等人 [47], 研究的见解，我们在训练中采用 Wasserstein 散度目标；与带梯度惩罚的 Wasserstein 损失 [24]相比，这一目标能够提升输出的保真度和稳定性。我们将 VQ-VAE-2 与 WGAN-div 的这一组合称为 VQ-VAE-2wdiv 模型。

我们的判别器的损失函数如下：

$$\mathcal{L}_{D_{wdiv}}(\mathrm{x}, D(\mathrm{z})) = Dis(\mathrm{x}) - Dis(D(\mathrm{z})) + \delta\|\nabla_{\hat{\mathrm{x}}} Dis(\hat{\mathrm{x}})\|^p \tag{2}$$

其中， $Dis$ 是判别器， $\delta$ 是用于确定散度惩罚强度的超参数。损失函数的第一部分 $Dis(\mathrm{x}) - Dis(D(\mathrm{z}))$用于评估真实样本 $\mathrm{x}$ 与我们的 VQ-VAE-2 $D(\mathrm{z})$ 输出之间的差异。第二项 $\delta\|\nabla_{\hat{\mathrm{x}}} Dis(\hat{\mathrm{x}})\|^p$是散度惩罚，它促使生成样本 $D(\mathrm{z})$ 接近真实数据分布。

生成器损失函数可简单表示为

$$\mathcal{L}_{G_{wdiv}} = Dis(D(\mathrm{z})) \tag{3}$$

该方法通过最小化判别器输出的结果来实现。因此，VQ-VAE-2 模型生成器的完整损失函数由 $\mathrm{L}^{vq}$ 和 $\mathrm{L}^{G_{wdiv}}$ 两个损失函数组合而成：

$$\mathcal{L} = \beta\mathcal{L}_{vq} + \gamma\mathcal{L}_{G_{wdiv}} \tag{4}$$

我们还引入了超参数 $\beta$ 和 $\gamma$，用于调整两个损失函数各自的贡献。

## 4.2 输入模态

在我们提出的框架中，我们结合使用三种输入模态（初始或先前的手势位置、文本和音频），以及说话人 ID。每种模态都经过特定的预处理。为了使框架更加灵活，我们允许初始信息和生成的手势序列具有任意长度。为此，我们引入两个参数：$N$表示初始信息中的帧数，我们将其设为 4；$M$ 表示手势序列生成的帧数，我们将其设为 30。为确保预处理流程的一致性，我们以每秒 15 帧的速率对手势进行采样，从而生成一段 2 秒长的新手势序列。

对于先前手势帧 $g$的预处理，我们在整个手势训练集上训练 VQ-VAE-2wdiv 模型，学习率为 $2e-5$，批次大小为 256。下文将此模型称为 $VQ_G$。对于引入的超参数，我们采用贝叶斯搜索进行超参数搜索，并根据结果将 $\alpha$、$\beta$ 和 $\gamma$ 分别设为 0.25、1 和 1。经过 200 个训练轮次后，我们恢复验证损失最低时的权重，并冻结模型权重。我们没有使用三维空间中的绝对位置，而是将父骨骼的相对位置作为输入和输出向量。最后，我们为 512 个可能的码本向量分别构建嵌入空间，并根据码本索引返回选定的嵌入。

文本预处理使用预训练的 BERT 模型 [14],，将给定文本 $tex$转换为词元，并提取其高层表示。对于我们的框架，我们返回一个固定长度的表示向量，最多包含 300 个词元。

音频处理流程采用三种不同的模型，以有效处理原始音频数据 $a$并从中提取有意义的特征。这些模型包括 Wav2Vec2 框架 [7],、我们定制的预训练 VQ-VAE-2wdiv 模型，以及 Onset Encoder 模型 [32]。Wav2Vec2 框架是一个基于语音输入训练的半监督模型，我们使用它处理原始音频数据并获取模型最后一层的隐藏状态。然而，由于 Wav2Vec2 框架主要侧重于根据给定的音频数据重建语音，因此训练过程中可能会丢失重要的韵律特征。为此，我们还使用降采样至 16 千赫的原始音频数据片段（每段 0.25 秒）训练自己的 VQ-VAE-2wdiv 模型。我们从随机选取的 YouTube 视频中收集了 1500 小时的音频数据，并将其划分为 1200 小时的训练数据和 150 小时的验证和测试数据。我们使用与 $VQ_G$ 相同的学习率、批次大小和超参数配置。由于该数据集远大于我们的手势数据集，我们训练模型 100 轮，随后冻结验证损失最低时的权重。我们将此模型称为 VQA 模型。在整个框架的训练过程中，我们使用该模型处理音频并返回向量 $E_x$。

近期研究凸显了声音起始在伴随语音手势生成中的重要性 [6, 32]。通过利用这一信息，模型能够使生成的手势与音频输入更有效地同步，从而更贴合语音的节奏。为此，我们采用了 Liang 等人提出的 Onset Encoder 模型 [32]。该模型首先使用滤波器检测音频的起始，然后将语音输入解耦为语义相关和语义无关的线索。与另外两个模型不同，Onset Encoder 模型的权重并未冻结，而是在框架训练期间进行更新。

为适应不同的说话者风格，我们使用说话人身份 $sp$，并为数据集中每位可能的说话人创建一个嵌入空间。将嵌入输入多层感知机（MLP）后，我们采用 Kingma 和 Welling 提出的“重参数化技巧” [29]，在潜在空间中建立多元正态分布。正如先前研究所示 [4, 21]，这项技术可用于生成新的、未见过的手势风格。除了嵌入空间，我们还定义了一个由 MLP 网络组成的“说话人决策网络”，该网络根据输入向量$W$预测当前说话人身份 $sp^*$，并据此返回相应的嵌入层。我们使用第 5 节所述的 Softmax 对比损失训练该网络。

## 4.3 GRU-Transformer

基于数据生成伴随语音手势，需要对序列手势数据中的时间依赖关系进行建模。门控循环单元（GRU） [13] 因其高度灵活且易于使用 [6, 32, 36, 48]，是常见的选择。GRU 尤其擅长捕捉非线性依赖关系，这对于对序列手势数据进行建模尤为重要，因为这类数据通常呈现复杂的时间动态。然而，GRU 的性能常受诸多挑战制约，例如收敛速度慢、学习效率有限以及梯度消失问题 [46]。与此同时，Transformer 架构 [44] 已被用于学习结构化输入数据中的长期依赖关系。尽管 Transformer 潜力可观，但由于计算需求高，且处理复杂非线性手势数据的能力有限，因此并未广泛用于伴随语音手势生成 [3, 8, 54]。

我们提出融合这两种方法，以全面且稳健地建模伴随语音手势复杂的时间（序列）结构。具体而言，我们使用 Transformer 捕捉输入序列中的广泛全局依赖关系，同时利用 GRU 学习非线性、局部依赖关系。如图 2 所示，我们提出了一种 GRU-Transformer 模型，使用 GRU 和 Transformer 流水线同时处理输入数据。Transformer 模块由一个带有时间信息预处理的多头自注意力模块组成，随后接入一个多层感知机（MLP）以降低输出维度。由于注意力头的最大数量必须能够被输入维度整除，我们将头数设为 $n_{heads} = max \quad n \in 1\ 2\ 3\ \ldots\ 12 \quad n \qquad n_i \ = \ 0 \qquad n_i$

({,,,,: 模数 })，其中等于输入维度。GRU1 模块由一个 GRU 层和其后的层归一化组成。GRU2 模块则使用四个堆叠的 GRU 层。通过跳跃连接，所有模块的中间输出都会加到输出向量上。

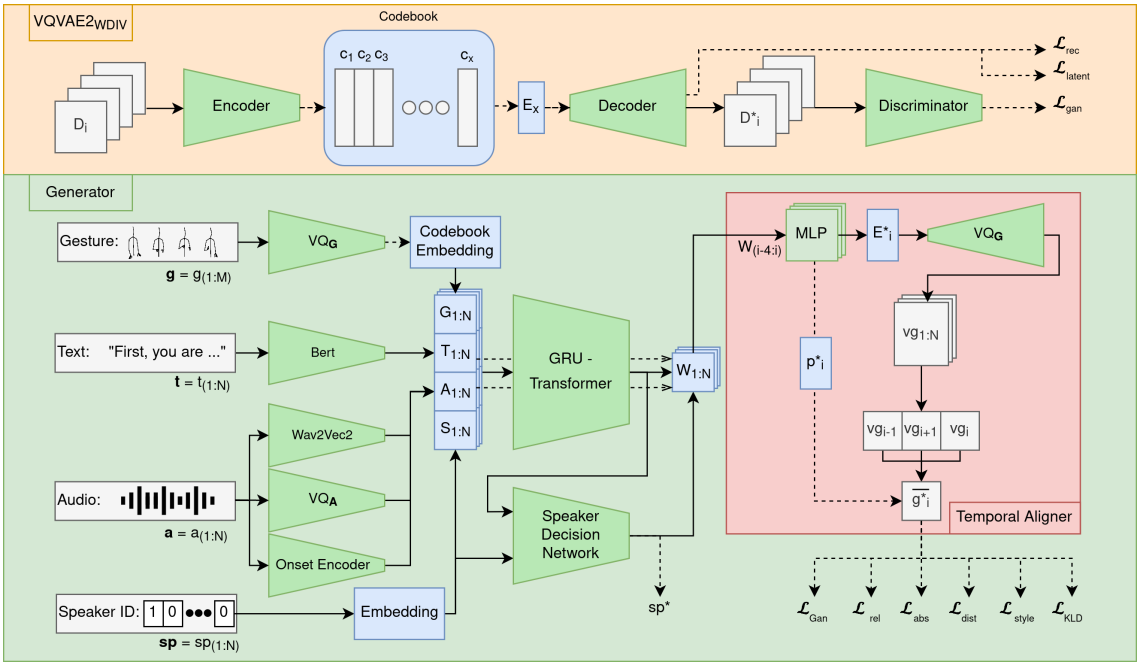


**图 1：AQ-GT模型概览。上方：添加了 GAN 判别器的 VQVAE2 网络。下方：包含输入模态、预处理、GRU-Transformer 和时间对齐网络（红色）的生成器网络。**

## 4.4 时间对齐器

生成伴随语音手势的一项挑战，是在相关的言语事件与手势事件之间建立时间同步性。尽管 Transformer 和 GRU 都能够从输入中学习时间信息，但在实际应用中，这些层仍难以生成令人信服且没有突兀、不必要手势伪影的手势 [16, 36]。因此，我们添加了一个时间对齐网络，用于学习先前生成帧的时间依赖关系，并对重建序列进行对齐，以消除生成手势中的突发噪声。我们的时间对齐器结构如图 1（红色方框）所示。为构建该对齐器，我们复用了第 4.2 节中介绍的 VQG。

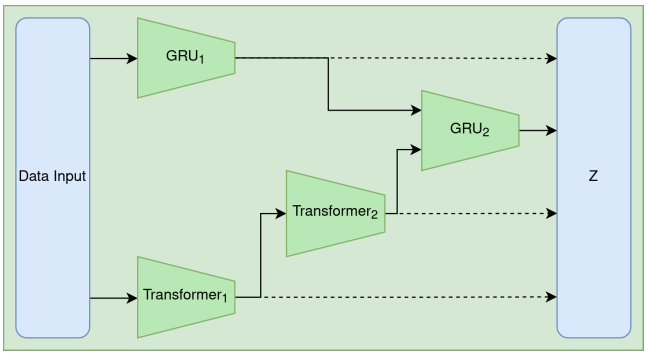


**图 2：GRU-Transformer 组件的架构。输入和输出数据以蓝色表示；模型层以绿色表示；虚线表示跳跃连接。**

接收到组合向量$W$后，我们应用滑动窗口遍历前面三帧，并将它们与当前帧以及上一次迭代的结果拼接起来。然后，我们将组合向量输入多层感知机，并保存结果以供下一次迭代使用。形式上，我们计算时间向量

$$f_c(\mathrm{W}_t) = MLP(f_c(\mathrm{W}_{t-1}) + \Big\|_{i=0}^{-3} \mathrm{W}_{t-i}) \tag{5}$$

由此，我们为每个时间步 $t$ 重构网络 $VQ_G$ 的向量 $E_x$，记作 $E_t^*$。利用该向量 $E_t^*$，我们使用 $VQ_G$ 网络重构下一帧手势 $vg_t$。由于 $VQ_G$ 是在由四帧组成的序列上训练的，我们将 $vg_{t-1}$ 的第三个元素、$vg_t$ 的第二个元素和 $vg_{t+1}$ 的第一个元素组合起来，构成手势向量 $g_t^*$。形式上，我们执行以下操作：

$$g_t^* = \frac{1}{3}\sum_{i=0}^{2} vg_{t-1+i,2-i} \tag{6}$$

构建向量 $g_t^*$ 后，我们将其与向量 $p_t^*$取平均；后者是通过一个多层感知机，根据时间向量 $f_c$（$\mathrm{W}_t$）的输出计算得到的。这样可以确保网络能够校正向量 $g_t^*$重建中残留的噪声。

## 5 生成器模型训练

生成器模型的训练过程如图 1 所示。在该模型中，经过独立预处理的输入数据被组合成一个输入向量，作为 GRU-Transformer 的输入。利用其输出，我们通过说话人决策网络预测当前说话人的身份。我们将 GRU-Transformer 的输出、说话人决策

网络，并将输入文本和音频编码为一个向量$W$。利用该向量，我们通过时间对齐器合成手势 $G^*$。

我们采用第 4.1 节所述的相同 WGAN-div 方法训练该模型。为了训练该框架，我们定义了多个损失函数，以指导训练的不同部分。在计算两组向量之间的距离时，我们使用 Huber损失 [27], , 并将其记作 $HL$。

首先，我们定义绝对重构损失和相对重构损失。为此，我们比较生成器模型的输出与真实值之间的 Huber距离，以及通过重构姿态数据的骨架结构得到的绝对 Huber距离；这两种比较均针对生成的手势和真实值进行。形式上，我们将重构损失定义为生成姿态的骨架与真实骨架之间的平均 Huber距离：

$$\mathcal{L}_{abs} = \frac{1}{N}\sum_{i=1}^{N} LH(Rc(g_i), Rc(g_i^*))\mathcal{L}_{rel} \quad = \frac{1}{N}\sum_{i=1}^{N} LH(g_i, g_i^*) \quad (7)$$

其中， $Rc$ 是绝对关节位置的重构函数， $N$ 是样本量， $g_i^*$ 是生成器模型输出的手势向量。

为确保生成涉及双手的正确手势，我们定义了一种损失，用于衡量左右手臂所有骨骼之间的成对距离，并将其与真实值进行比较。形式上，我们将距离损失定义为：

$$\mathcal{L}dist = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} HL(\mathrm{g_i}_{l,j} - \mathrm{g_i}_{r,j}, \mathrm{g^*_i}_{l,j} + \mathrm{g^*_i}_{r,j}) \quad (8)$$

其中 $K$ 是手臂关节的数量。

我们还加入了说话人对比损失，该损失结合了 Yoon 等人提出的说话人风格损失 [48] 和我们提出的说话人判别 Softmax 损失。我们的训练方法基于输入 $T$ 和原始说话人身份 $\boldsymbol{sp}id(1)$生成手势，同时最大化具有相同输入 $T$的随机说话人身份 $\boldsymbol{sp}id(2)$之间的距离。形式上，我们定义：

$$f_{st} = \mathrm{CE}(\mathrm{G}(\boldsymbol{T}, \boldsymbol{sp}_{id(1)}), \mathrm{G}(\boldsymbol{T}, \boldsymbol{sp}_{id(2)})) \quad (9)$$

$$f_{sp} = \mathrm{HL}(\mathrm{SPD}(\boldsymbol{T}, \boldsymbol{sp}_{id(1)}), \mathrm{SPD}(\boldsymbol{T}, \boldsymbol{sp}_{id(2)})) \quad (10)$$

$$\mathcal{L}_{\mathrm{style}} = -\mathbb{E}\left[\min\left(\frac{f_{st} + f_{sp}}{\|\boldsymbol{sp}_{id(1)} - \boldsymbol{sp}_{id(2)}\|_1}, \epsilon\right)\right] \quad (11)$$

其中 $SPD(T,\ sp)$是说话人决策网络的 Softmax 输出， $CE$ 是交叉熵损失。为便于采样不同风格，我们使用说话人身份特征空间与$N(0,\ I)$之间的 Kullback-Leibler（KL）散度损失（$\mathrm{L}_{KLD}$），以假设风格嵌入服从高斯分布。

最后，我们需要考虑自然手势中关节向量的位置依赖性和时间依赖性。为此，我们将关节位置向量关于时间的一阶至六阶导数纳入损失函数。形式化地，我们定义：

$$f_{dop}(g) = \frac{1}{6}\sum_{i=1}^{6}\frac{d^i g}{dt^i} \quad (12)$$

$$\mathcal{L}_{DoP} = \frac{1}{N}\sum_{i=1}^{N} LH(f_{dop}(g_i), f_{dop}(g_i^*)) \quad (13)$$

**表 2：所提出的模型与 BiGe 数据集上的三种最先进模型的比较。FGD 和 MAJE 分数越低越好；多样性距离越高越好。**

| 方法 | FGD ↓ | 多样性 ↑ | MAJE ↓ |
|---|---|---|---|
| 真实值 | 0.00 | 42.128 | 0.00 |
| 三模态 [48] | 2.023 | 38.980 | 0.0119 |
| SEEG [32] | 2.018 | 38.660 | 0.0121 |
| HA2G [36] | 1.346 | 41.332 | 0.0096 |
| **我们的方法 (AQ-GT)** | **0.3977** | **42.885** | **0.0078** |

总体损失函数由以下各项组成：

$$\mathcal{L}total = \pi_1\mathcal{L}_{gan} + \pi_2(\mathcal{L}_{rel} + \mathcal{L}_{abs}) \quad (14)$$

$$+ \pi_3\mathcal{L}dist + \pi_4\mathcal{L}_{\mathrm{style}} + \pi_5\mathcal{L}_{KLD} + \pi_6\mathcal{L}_{DoP} \quad (15)$$

其中 $\pi_1$、 $\pi_2$、 $\pi_3$、 $\pi_4$、 $\pi_5$、 $\pi_6$ 为用于调整损失权重的超参数。我们对框架中的所有参数进行了超参数搜索。根据搜索结果，我们使用批次大小 190、 $\pi_1$ 为 2、 $\pi_2$、 $\pi_3$、 $\pi_4$、以及 $\pi_6$ 为 20、 $\pi_5$ 为 0.004。我们采用不设上限的训练轮次，并结合停止条件来结束训练；如果连续 50 轮后验证损失没有下降，则恢复表现最佳的权重。随后，我们训练模型 368 轮，并恢复第 318 轮的权重。我们采用逐步递减的学习率调度策略来优化模型性能。具体而言，我们将学习率初始化为 $1e{-}4$。每 20 轮将学习率乘以 0.75，最低降至 $1e-5$。为降低过拟合的可能性，我们在每个层之间以及网络输入处使用丢弃法，每 25 轮逐步增加 0.05，最高增至 0.3。

## 6 评估

为了评估所提出的框架，我们将生成的手势与其他最先进水平的框架所生成的手势进行比较，既使用现有的客观指标，也通过用户研究考察人们对生成结果的主观感知差异。

### 6.1 客观评估

为确保比较公平，我们在大型 BiGe 数据集上训练了三个现有框架，并比较结果。具体而言，我们将 Trimodal 框架 [48], 、SEEG 框架 [32], 和 HA2G 框架 [36] 与我们提出的模型进行评估比较。遗憾的是，由于无法获取 Ao 等人提出模型的实现 [6] ，我们无法将自己的模型与之进行比较。

我们遵循每个框架的原始配置，仅调整输出维度，以适应数据集中更高的关节维度。我们使用 Fréchet 手势距离（FGD） [48],多样性距离 [36], 以及生成的关节位置与真实值之间绝对误差的均值（MAJE）来衡量各模型的性能。由于多样性距离的方差较大，我们报告了 1000 次计算的平均值。结果见表 2。

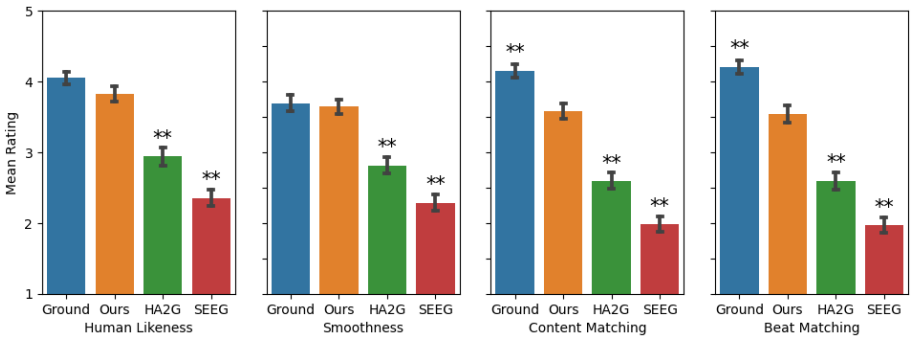


**图 3：主观评估研究的结果。真实值以蓝色标示。我们的方法框架以黄色标示。HA2G 框架以绿色标示。SEEG 框架以红色标示。星号表示显著效应（*：p< 0.05，**：p < 0.005）。**

实验结果表明，我们的模型在所有评估指标上均优于现有框架。值得注意的是，AQ-GT 框架的 FGD 比 HA2G 低三倍。就多样性距离而言，我们的框架优于所有其他框架，甚至超过了人类真实值。需要注意的是，多样性衡量的是基于潜在表征的随机手势样本对之间的绝对距离。因此，尽管评估结果令人鼓舞，但这也可能意味着手势的离散表示会产生离群点，并将其映射到潜在空间中的不同区域，从而增大多样性距离。

为了与更广泛的已发表框架进行比较，我们还在 TED 手势数据集 [49] 上训练了模型。我们的模型架构和配置保持不变，仅调整了输入和输出维度，以适应不同的身体关节维度。为避免造成不公平的优势，我们还在 TED 数据集上预训练了一个新的 $VQ_G$ 模型，并在框架的训练过程中将其纳入其中。结果如表 3 所示，再次表明，与所有其他方法相比，我们的方法在 FGD 指标上取得了显著提升。

**表 3：所提出模型与 TED 数据集上最先进系统的比较。FGD 越低越好；多样性距离越高越好（SEEG 和 Rythmic Gesticulator 未报告多样性距离）。**

| 方法 | FGD ↓ | 多样性 ↑ |
|---|---|---|
| 真实值 | 0.00 | 110.821 |
| Joint Embedding [5] | 22.083 | 91.223 |
| Speech2Gesture [22] | 19.254 | 98.095 |
| Attention Seq2Seq [49] | 18.154 | 92.176 |
| SEEG [32] | 3.751 | - |
| 三模态 [48] | 3.729 | 102.539 |
| HA2G [36] | 3.072 | **108.086** |
| Rythmic Gesticulator [6] | 2.04 | - |
| **我们的方法（AQ-GT）** | **1.612** | 104.76 |

此外，我们的框架取得了第二高的多样性距离分数，仅次于 HA2G 框架。

## 6.2 人工评分者研究

由于评估伴随语音手势的质量具有很强的主观性，手势生成框架通常会通过人工评分者进行评估 [37]。因此，我们开展了一项人工评分者研究，将我们提出的框架与 Liu 等人 [36]和 Liang 等人 [32]提出的框架进行比较评估。为此，我们从测试集中随机选取了十段时长为 20 秒的音频/文本序列，并使用这三个模型生成伴随语音手势。本研究以在线评估的形式开展，共有 70 名英语参与者（女性和男性各占 50%）参加。参与者需要按随机顺序观看视频，并使用 5 点李克特量表，针对拟人度、动作流畅度，以及手势在内容和时间（节拍）上与伴随语音的匹配程度进行评分。在正式研究之前，我们对 10 名参与者开展了预调查，以估算每段序列所需的时间。根据预调查结果，我们在分析中排除了在调查页面停留时间不足 30 秒的参与者，以及用时偏离平均值超过两个标准差的参与者。应用这些条件后，最终保留了 39 名参与者。由于结果未通过正态性和方差齐性检验，我们使用广义线性混合效应模型（GLMM）进行重复测量分析。随后，为比较不同模型之间的两两显著性效应，我们采用了 Dunn 事后检验。结果如图 3 所示，表明本文提出的 AQ-GT 方法在所有评分类别中均显著优于 HA2G 和 SEEG 框架。值得注意的是，我们的框架在拟人度和动作流畅度方面的评分与人类真实值相当（差异不显著），而在内容匹配和节拍匹配方面的评分仍显著低于真实值。这些发现表明，尽管我们的方法能够生成恰当、流畅且自然的手势，但在生成能够准确反映语音内容或传达大量额外含义的可信手势方面仍有不足。

**表 4：所提出的原始模型与四种不同版本的比较，这些版本省略了单个组件（FGD 和 MAJE 越低越好；多样性越高越好）。**

| 方法 | FGD ↓ | 多样性 ↑ | MAJE ↓ |
|---|---|---|---|
| 真实值 | 0.00 | 42.128 | 0.00 |
| **不含门控循环单元** | 57.75 | 16.235 | 0.0330 |
| **不使用时间对齐器** | 1.653 | 39.054 | 0.0157 |
| **不使用 VQG** | 1.342 | 41.698 | 0.0102 |
| **无 Transformer** | 0.777 | 43.592 | 0.0086 |
| **原文** | **0.3977** | **42.885** | **0.0078** |

## 7 消融研究

为了评估所提出框架中各个组件的作用，我们采用四种不同的架构配置开展了消融研究。我们通过有选择地移除所有门控循环单元层（w/oGRU）或所有 Transformer 层（w/oTransformer），评估门控循环单元-Transformer 的影响。对于时间对齐器，我们通过移除使用 $VQ_G$ 对手势进行的最终重建（w/o VQG）或移除整个时间对齐器（w/o Temporal Aligner）来进行评估。每种修改后的架构均重新训练了 50 个训练轮次，并恢复了验证损失最佳时对应的权重。测试集上的表现报告于表 4。如表所示，移除门控循环单元层对所有报告的指标影响最大。对生成结果进行目视检查也发现，关节出现严重变形，且看不到任何手势。另一方面，移除 Transformer 层会导致 FGD 和 MAJE 略有下降，但多样性距离值有所升高。如上文所述，这可能表明多样性距离存在问题，因为该指标极易受到手势中噪声和离群点的影响。尽管如此，这也表明 Transformer 层确实发挥了预期作用：学习帧之间的长期依赖关系，并减少生成手势中的噪声。

移除 $VQ_G$ 重建会导致所有指标显著下降，但仍取得了与 HA2G 框架相近的结果（见表 2）。移除整个时间对齐器会使性能进一步大幅下降，结果更接近三模态和 SEEG 框架。这清楚地表明，学习生成的中间序列中的时间依赖关系有助于生成更加连贯、自然的手势，因此通常应将其纳入未来的伴随语音手势合成框架中。

## 8 结论

本文提出了一种新颖的伴随语音手势合成框架，称为 AQ-GT。该框架结合并运用了多种方法和技术来生成伴随语音手势运动。最值得注意的是，我们结合门控循环单元和 Transformer 层，对多模态行为中的非线性时间依赖关系进行建模。此外，我们使用向量量化自编码器，为两种模态（语音和手势）中的序列数据提供离散潜在空间表征，并采用专门的时间对齐网络来重建自然手势。为了训练这一复杂模型，我们收集了大量高精度的多模态数据，并使用扩展的数据处理流程，从单目视频中追踪完整的三维身体关节。评估结果表明，该方法能够合成超越当前最先进水平的伴随语音手势，无论在客观指标还是主观指标方面均表现出色；在拟人性和流畅度方面，部分合成结果甚至与人类手势难以区分。通过消融研究，我们证明了每个组件都能对客观指标产生积极影响。

总体而言，所提出的 AQ-GT框架为该领域作出了贡献，并推动了伴随语音手势合成当前最先进水平的发展。尽管我们的框架能够生成令人信服、自然的手势，但仍面临一些挑战。首先，多样性距离表明，由离散潜在表示生成的手势中可能存在伪影。此外，主观研究显示，当前框架（与所有其他基于学习的方法一样）仍无法涵盖人类交流中的意义维度，因此主要生成支持说话行为的对话“运动手势”，却很少提供额外信息。基于学习的伴随语音手势合成还存在另一个固有缺点，即无法引导或控制手势的生成。在未来的工作中，我们将研究离散表示的缺点及其改进方法，并探索能否通过传达额外含义来引导手势生成，使之更贴合对话内容。在目前已另行发表的工作中，我们已开始扩展 AQ-GT模型，使其能够支持手势合成中的形式与意义目标特征，这些特征是从规模较小但标注丰富的数据集中学习得到的。这表明，所提出的 AQ-GT模型可作为合适的基础，进一步探索如何生成的不仅是伴随语音的动作，更是作为多模态交际行为不可或缺组成部分的、真正有意义的手势。

# 参考文献


[1] [无日期]。 TED — youtube.com。https://www.youtube.com/c/TED/videos。 [访问于 2 月 16 日-2023]。[2] [无日期]。TEDx 演讲 — youtube.com。https://www.youtube.com/channel/UCsT0YIqwnpJCM-mx7-gSA4Q。 [访问于 2 月 16 日-2023]。[3] Chaitanya Ahuja、Dong Won Lee、Ryo Ishii 和 Louis-Philippe Morency。2020。没有被遗漏的手势：学习口语与自由形式手势之间的关系。载于《计算语言学协会论文集：EMNLP 2020》精选论文。1884–1895。[4] Chaitanya Ahuja、Dong Won Lee、Yukiko I. Nakano 和 Louis-Philippe Morency。2020。《伴随语音的手势动画风格迁移：一种多说话人条件混合方法》。http://arxiv.org/abs/2007.12553 arXiv:2007.12553 [cs]。[5] Chaitanya Ahuja 和 Louis-Philippe Morency。2019。Language2Pose：基于自然语言的姿势预测。http://arxiv.org/abs/1907.01108 arXiv:1907.01108 [cs]。[6] Tenglong Ao、Qingzhe Gao、Yuke Lou、Baoquan Chen 和 Libin Liu。2022。节奏手势生成器：基于分层神经嵌入的节奏感知伴随语音手势合成。《ACM 图形学汇刊》41，6（2022 年 12 月），1–19。https://doi.org/10.1145/3550454.3555435 arXiv:2210.01448 [cs, eess]。[7] Alexei Baevski、Henry Zhou、Abdelrahman Mohamed 和 Michael Auli。2020。wav2vec 2.0：用于语音表征自监督学习的框架。https://doi.org/10.48550/arXiv.2006.11477 arXiv:2006.11477 [cs, eess]。[8] Uttaran Bhattacharya、Nicholas Rewkowski、Abhishek Banerjee、Pooja Guhan、Aniket Bera 和 Dinesh Manocha。2021。Text2gestures：一种为虚拟智能体生成情感性身体手势的基于 Transformer 的网络。载于《2021 IEEE 虚拟现实与三维用户界面会议（VR）》。IEEE，1–10。[9] Matthew Brand 和 Aaron Hertzmann。2000。风格机器。载于《第 27 届计算机图形学与交互技术年度会议论文集（SIGGRAPH '00）》。ACM Press/Addison-Wesley Publishing Co.，美国，183–192。https://doi.org/10 1145 344779 344865 。/。[10] Justine Cassell、David McNeill 和 Karl-Erik McCullough。1999。言语-手势不匹配：关于语言与非语言信息共享同一底层表征的证据。《语用学与认知》7，1（1999），1–34。[11] Justine Cassell、Hannes Högni Vilhjálmsson 和 Timothy Bickmore。2004。BEAT：行为表达动画工具包。载于《逼真角色：工具、情感功能与应用》，Helmut Prendinger 和 Mitsuru Ishizuka（编）。Springer，柏林，海德堡，163–185。https://doi.org/10.1007/978-3-662-08373- 4 8[12] ¯ Changchun Liu、P. Rani 和 N. Sarkar。2005。人机交互中情感识别机器学习技术的实证研究。载于《2005 IEEE/RSJ 智能机器人与系统国际会议》。IEEE，加拿大阿尔伯塔省埃德蒙顿，2662–2667。https://doi.org/10.1109/IROS.2005.1545344[13] Kyunghyun Cho、Bart Van Merriënboer、Caglar Gulcehre、Dzmitry Bahdanau、Fethi Bougares、Holger Schwenk 和 Yoshua Bengio。2014。使用 RNN 编码器-解码器学习用于统计机器翻译的短语表征。arXiv 预印本 arXiv:1406.1078（2014）。[14] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019。BERT：用于语言理解的深度双向 Transformer 预训练。https://doi.org/10.48550/arXiv.1810.04805 arXiv:1810.04805 [cs]。[15] Patrick Esser、Robin Rombach 和 Björn Ommer。2021。驾驭 Transformer 进行高分辨率图像合成。http://arxiv.org/abs/2012.09841 arXiv:2012.09841 [cs]。[16] Angela Fan、Thibaut Lavril、Edouard Grave、Armand Joulin 和 Sainbayar Sukhbaatar。2020。通过反馈记忆解决 Transformer 的一些局限。arXiv 预印本 arXiv:2002.09402（2020）。[17] Yuchen Fan、Yao Qian、Feng-Long Xie 和 Frank K. Soong。2014。基于双向 LSTM 循环神经网络的文本转语音合成。载于《Interspeech 2014》。ISCA，1964 1968 10 21437 2014 443 –。https://doi.org/。/[18] Interspeech。- Hao-Shu Fang、Jiefeng Li、Hongyang Tang、Chao Xu、Haoyi Zhu、Yuliang Xiu、Yong-Lu Li 和 Cewu Lu。2022。AlphaPose：实时全身区域多人姿势估计与跟踪。《IEEE 模式分析与机器智能汇刊》（2022）。[19] Ylva Ferstl 和 Rachel McDonnell。2018。探究循环运动建模在言语手势生成中的应用。载于《第 18 届智能虚拟体国际会议论文集》。ACM，澳大利亚新南威尔士州悉尼，93–98。https://doi.org/ 10 1145 3267851 3267898 。/。[20] Aphrodite Galata、Neil Johnson 和 David Hogg。2001。学习行为的可变长度马尔可夫模型。《计算机视觉与图像理解》81，3（3 月 2001 398 413 10 1006 2000 0894），–。[21] https://doi.org/。/cviu。。Saeed Ghorbani、Ylva Ferstl 和 Marc-André Carbonneau。。基于语音的范例式风格化手势生成：GENEA 挑战赛 2022 参赛作品。载于《2022 年多模态交互国际会议论文集》。778–783。[22] Shiry Ginosar、Amir Bar、Gefen Kohavi、Caroline Chan、Andrew Owens 和 Jitendra Malik。2019。学习对话手势的个体风格。http://arxiv.org/abs/1906.04160 arXiv:1906.04160 [cs, eess]。[23] Ian Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron Courville 和 Yoshua Bengio。2020。生成对抗网络。《Commun. ACM》63，11（2020），139–144。[24] Ishaan Gulrajani、Faruk Ahmed、Martin Arjovsky、Vincent Dumoulin 和 Aaron C Courville。2017。改进 Wasserstein GAN 的训练。《神经信息处理系统进展》30（2017）。[25] Gustav Eje Henter、Simon Alexanderson 和 Jonas Beskow。2020。MoGlow：利用归一化流进行概率式、可控的动作合成。《ACM 图形学汇刊》39，6（2020 年 12 月），1–14。https://doi.org/10.1145/3414685. 3417836 arXiv:1905.06598 [cs, eess, stat]。[26] Daniel Holden、Taku Komura 和 Jun Saito。2017。用于角色控制的相位函数神经网络。《ACM 图形学汇刊》36，4（2017 年 8 月），1 13 10 1145 3072959 3073663 – 。https://doi.org/ 。/ 。[27] Peter J. Huber。1964。位置参数的稳健估计。《数理统计年刊》35，1（1964 年 3 月），73–101。https://doi.org/10.1214/ aoms/1177703732 出版方：数理统计学会。[28] Kyung-Min Kim、Chang-Jun Nan、Jung-Woo Ha、Yu-Jung Heo 和 Byoung-Tak Zhang。2015。Pororobot：一款能玩视频问答游戏的深度学习机器人。（2015）。[29] Diederik P Kingma 和 Max Welling。2013。自动编码变分贝叶斯。arXiv 预印本 arXiv:1312.6114（2013）。[30] Stefan Kopp、Brigitte Krenn、Stacy Marsella、Andrew N. Marshall、Catherine Pelachaud、Hannes Pirker、Kristinn R. Thórisson 和 Hannes Vilhjálmsson。2006。迈向多模态生成的通用框架：行为标记语言。载于《智能虚拟代理》（《计算机科学讲义》），Jonathan Gratch、Michael Young、Ruth Aylett、Daniel Ballin 和 Patrick Olivier（编）。Springer，柏林，海德堡，205–217。https://doi.org/10.1007/11821830_17[31] Gilwoo Lee、Zhiwei Deng、Shugao Ma、Takaaki Shiratori、Siddhartha S. Srinivasa 和 Yaser Sheikh。2019。用双手交谈 16.2 m：用于会话动作分析与合成的大规模同步身体—手指动作和音频数据集。载于《IEEE/CVF 国际计算机视觉会议论文集》。763–772。[32] Yuanzhi Liang、Qianyu Feng、Linchao Zhu、Li Hu、Pan Pan 和 Yi Yang。2022。SEEG：语义增强的伴随语音手势生成。载于《IEEE/CVF 计算机视觉与模式识别会议论文集》。10473–10482。[33] Hung Yu Ling、Fabio Zinno、George Cheng 和 Michiel van de Panne。2020。使用动作变分自编码器的角色控制器。《ACM 图形学汇刊》39，4（2020 年 8 月）。https://doi.org/10.1145/3386569.3392422 arXiv:2103.14274 [cs]。[34] Haiyang Liu、Zihao Zhu、Naoya Iwamoto、Yichen Peng、Zhengqing Li、You Zhou、Elif Bozkurt 和 Bo Zheng。2022。BEAT：用于会话手势合成的大规模语义和情感多模态数据集。载于《计算机视觉——ECCV 2022》（《计算机科学讲义》），Shai Avidan、Gabriel Brostow、Moustapha Cissé、Giovanni Maria Farinella 和 Tal Hassner（编）。Springer Nature Switzerland，Cham，612–630。https://doi.org/10.1007/978-3-031-20071- 7 36[35] ¯ Jinlin Liu、Yuan Yao 和 Jianqiang Ren。2019。高分辨率图像合成的加速框架。arXiv 预印本 arXiv:1909.03611（2019）。[36] Xian Liu、Qianyi Wu、Hang Zhou、Yinghao Xu、Rui Qian、Xinyi Lin、Xiaowei Zhou、Wayne Wu、Bo Dai 和 Bolei Zhou。2022。学习用于伴随语音手势生成的分层跨模态关联。载于《2022 IEEE/CVF 计算机视觉与模式识别会议（CVPR）》。IEEE，美国路易斯安那州新奥尔良，10452–10462。https://doi.org/10.1109/CVPR52688.2022.01021[37] Simbarashe Nyatsanga、Taras Kucherenko、Chaitanya Ahuja、Gustav Eje Henter 和 Michael Neff。2023。数据驱动的伴随语音手势生成综合综述。https://doi.org/10.1111/cgf.14776 arXiv:2301.05339 [cs]。[38] Amit Kumar Pandey 和 Rodolphe Gelin。2018。量产型社交人形机器人 Pepper：同类中的首款机器。IEEE《机器人与自动化杂志》25，3（2018 年 9 月），40–48。https://doi.org/10.1109/MRA.2018.2833157 会议名称：IEEE《机器人与自动化杂志》。[39] Dario Pavllo、Christoph Feichtenhofer、David Grangier 和 Michael Auli。2019。利用时间卷积和半监督训练进行视频中的三维人体姿态估计。载于《IEEE/CVF 计算机视觉与模式识别会议论文集》。7753–7762。[40] Ali Razavi、Aaron Van den Oord 和 Oriol Vinyals。2019。利用 VQ-VAE-2 生成多样的高保真图像。《神经信息处理系统进展》32（2019）。[41] Joseph Redmon 和 Ali Farhadi。2018。YOLOv3：一项渐进式改进。arXiv 预印本 arXiv:1804.02767（2018）。[42] Adriana Tapus、Antonio Bandera、Ricardo Vazquez-Martin 和 Luis V. Calderita。2019。感知人及其与他人的互动以用于社交机器人学——综述。《模式识别快报》118（2019 年 2 月），3–13。https: //doi.org/10 1016 2018 03 006 . /j.patrec. . . Aaron Van Den Oord、Oriol Vinyals 等。2017。神经离散表示学习。《神经信息处理系统进展》30（2017）。[44] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017。注意力机制就是你所需要的一切。《神经信息处理系统进展》30（2017）。

[45] Petra Wagner、Zofia Malisz 和 Stefan Kopp。2014。互动中的手势与语音：概述。209–232 页。[46] Xin Wang、Jiabing Xu、Wei Shi 和 Jiarui Liu。2019。OGRU：一种优化的门控循环单元神经网络。载于物理学杂志：会议系列，第 1325 卷。IOP Publishing，012089。[47] Jiqing Wu、Zhiwu Huang、Janine Thoma、Dinesh Acharya 和 Luc Van Gool。2018。GAN 的 Wasserstein 散度。载于欧洲计算机视觉会议论文集（*ECCV*）。653–668。[48] Youngwoo Yoon、Bok Cha、Joo-Haeng Lee、Minsu Jang、Jaeyeon Lee、Jaehong Kim 和 Geehyuk Lee。2020。基于文本、音频和说话人身份三模态语境的言语手势生成。*ACM* 图形学汇刊 39，6（2020 年 12 月），1–16。https://doi.org/10.1145/3414685.3417838 arXiv:2009.02119 [cs] 存档：https://web.archive.org/web/20230405100117/https://arxiv.org/abs/2009.02119

[49] Youngwoo Yoon、Woo-Ri Ko、Minsu Jang、Jaeyeon Lee、Jaehong Kim 和 Geehyuk Lee。2019。机器人学习社交技能：面向人形机器人的端到端伴随语音手势生成学习。载于*2019*年国际机器人与自动化会议（*ICRA*）。IEEE，加拿大魁北克省蒙特利尔，4303–4309。https://doi. org/10.1109/ICRA.2019.8793720

[50] Youngwoo Yoon、Pieter Wolfert、Taras Kucherenko、Carla Viegas、Teodor Nikolov、Mihail Tsakov 和 Gustav Eje Henter。2022。《GENEA 挑战赛 2022：对数据驱动的伴随语音手势生成的大规模评估》。载于*2022* 年多模态交互国际会议论文集。736–747。[51] Chuang Yu 和 Adriana Tapus。2019。《用于多模态情绪识别的交互式机器人学习》。载于社交机器人学（计算机科学讲义），Miguel A. Salichs、Shuzhi Sam Ge、Emilia Ivanova Barakova、John-John Cabibihan、Alan R. Wagner、Álvaro Castro-González 和 Hongsheng He（编）。Springer International Publishing，Cham，633–642。https://doi.org/10.1007/978-3-030-35888-4 59_[52] Fan Zhang、Valentin Bazarevsky、Andrey Vakunov、Andrei Tkachenka、George Sung、Chuo-Ling Chang 和 Matthias Grundmann。2020。MediaPipe Hands：设备端实时手部追踪。*arXiv* 预印本 *arXiv:2006.10214*（2020）。[53] Feng Zhang、Xiatian Zhu 和 Mao Ye。2019。快速人体姿态估计。载于 *IEEE/CVF* 计算机视觉与模式识别会议论文集。3517–3526。[54] Wenlin Zhuang、Jinwei Qi、Peng Zhang、Bang Zhang 和 Ping Tan。2022。文本/语音驱动的全身动画。*arXiv* 预印本 *arXiv:2205.15573*（2022）。