ReactMotion：基于说话者话语的反应式听者动作生成

罗成[1†]、吴碧珠[2, 4, 5†]、李冰[1*]、任建锋[4]，白瑞斌[4]，曲荣[5]，沈琳琳[2, 3*]和 Bernard Ghanem[1]

[1] 阿卜杜拉国王科技大学[2] 深圳大学计算机视觉研究所、人工智能学院[3] 深圳大学广东省智能信息处理重点实验室[4] 宁波诺丁汉大学计算机科学学院[5] 诺丁汉大学计算机科学学院，英国项目主页：https://reactmotion.github.io

arXiv:2603.15083v1 [cs.CV] 16 Mar 2026

摘要。 本文提出一项新任务——基于说话者话语的反应式听者动作生成，旨在生成自然逼真的听者身体动作，以恰当地回应说话者的话语。然而，由于人类反应本质上具有非确定性，听者非语言行为的建模仍未得到充分探索，且颇具挑战。为推动这一任务的发展，我们提出了 ReactMotionNet，这是一个大规模数据集，其中说话者话语与多个候选听者动作配对，并根据恰当程度进行标注。该数据集设计明确体现了倾听者行为的一对多特性，并提供了超越单一真实值动作的监督。基于这一数据集设计，我们制定了专为评估反应恰当性而设计的偏好导向评估协议，而传统的动作指标侧重于输入与动作的对齐，忽略了这一点。我们进一步提出 ReactMotion，这是一个统一的生成框架，可联合建模文本、音频、情绪和动作，并通过基于偏好的目标函数进行训练，以鼓励生成既恰当又多样的听者响应。大量实验表明，ReactMotion 优于检索基线和级联式基于大语言模型的流水线，能够生成更加自然、多样且恰当的听者动作。

关键词： 双人交互 · 交互式人工智能系统

1 引言

对双人类交流进行建模，对于虚拟代理 [33]，数字人类 [50, 104]，和社交机器人 [71]至关重要。尽管以往的研究推动了语音到语音对话 [15]，和基于语言的接口 [1, 28]，的发展，也研究了听者的面部反应 [57, 70]，，但反应式听者身体动作仍在很大程度上遭到忽视

† 贡献相同。 * 通讯作者。

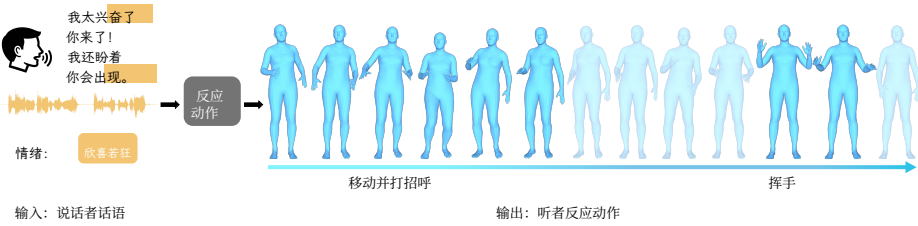


图 1：所提出的新任务示意图：语音话语驱动的反应式听者动作生成。给定说话者的话语，即转录文本和/或音频（可选地辅以情绪信息），生成模型（如我们的 ReactMotion）会为听者生成相应的响应式身体动作序列。

尽管它们是面对面互动的核心组成部分。听者常通过姿势和细微的手势表达投入与理解，而生成这类反馈对于自然的双人交流至关重要。

我们提出了一项新任务——语音话语驱动的反应式听者动作生成，旨在根据说话者话语的音频和/或转录文本，生成能够恰当回应该话语的自然听者身体动作。与主要实现输入内容的文本到动作生成 [21, 62, 75, 76, 96] 或音频驱动的动作生成 [88]不同，我们的设定对对话反应进行建模，其中说话者线索是间接的，输出本质上是一对多的。

这项任务面临三项挑战。(i) 同一话语可能引发多种有效的听者反应 [57, 70]。这种非确定性的听者行为给听者动作响应的建模带来了重大挑战。(ii) 据我们所知，目前尚无公开的大规模数据集，能够为每个话语提供多种听者反应身体动作。(iii) 反应恰当性难以评估。基于单一真实标注或动作多样性的指标，不足以衡量听者反应的恰当性。

为应对这些挑战，我们推出了**ReactMotionNet**，这是一个精心整理的数据集，包含 151,328 组（说话者话语，听者动作）配对。**与通常为每种条件仅提供一个目标的既有动作数据集不同**，我们为每条话语关联多个 候选反应，并将其标注为三个偏好等级：金级、银级和负级。这种分级设计捕捉了一对多的歧义，并支持**偏好式监督与评估** [11, 13, 102]。此外，我们提出了一种可扩展的流程，可将现有动作数据重新用于双人说话者—听者配对的数据集构建，**避免依赖成本高昂的说话者—听者动作捕捉**。

为评估反应恰当性，我们提出一种**等级感知排序协议**。我们训练一个多模态评判网络，在相同说话者输入的条件下为候选反应打分并排序，并报告其相对于金级、银级或负级的胜率。**这种相对评估超越了单一参考相似度**，更能体现同一话语可能存在多种恰当反应。最后，我们提出**ReactMotion**，一个统一的生成框架，联合建模说话人转录文本、情绪和音频

以生成听者动作。我们利用分层标注，通过基于偏好的目标函数，在训练中学习每个话语组内的相对比较。

贡献。 (i) 据我们所知，我们首次提出了在双人互动中根据说话者语音生成听者身体动作的任务。(ii) 我们提出了**ReactMotionNet**，这是一个包含多层级（金级/银级/负级）听者反应动作的新数据集，并设计了分层感知评估协议来评估反应恰当性，从而推动对听者非语言回应行为的研究。(iii) 我们提出了**ReactMotion**，这是一种统一的多模态生成模型，能够处理说话者的多种线索，并生成高质量的听者身体动作来回应说话者。

2 相关工作

人体动作生成。 人体动作生成可以以多种模态为条件，包括文本 [8, 30, 42, 48, 52, 63, 78, 84, 95, 98], 动作类别 [60, 64, 74], 以及音乐 [37, 39, 40, 90] 或语音 [38,45,85]等音频信号。在这些方法中，基于文本和音频驱动的动作生成与我们的设置最为相关。基于文本的方法根据明确的动作描述生成动作 [4, 10, 18, 26, 34, 61, 79, 80, 97, 101, 105], ，而音频驱动方法则合成与时间同步的声学信号相匹配的手势 [7, 53, 99]。具有代表性的建模范式包括基于 Transformer 的潜变量模型（例如， [43, 60, 100]）、结合自回归建模的离散动作分词（例如， [3, 9, 91, 96]），以及基于扩散的框架（例如， [2, 22, 44, 76]）。

除了单人生成之外，近期研究 [24, 41, 53, 55, 73, 81] 将动作合成扩展到了多人场景。这些方法通常通过以联合动作的明确文本描述，或双方的音频流为条件来生成多人动作。相比之下，我们的问题设定有所不同：目标动作并未由明确的动作指令或同步信号直接指定。相反，模型必须根据说话者的话语（包括转录文本、音频和情绪线索）推断隐含的互动意图，并为听者生成符合社交规范的反应动作。这要求模型推理跨说话者动态关系，而非直接进行条件到动作的映射。

人体反应生成。 人体反应生成对人工智能交互系统至关重要。口语建模已从级联式自动语音识别→ 大语言模型 → 文本转语音流程，发展到端到端和全双工语音到语音模型 [15, 66, 77, 94], ； 与此同时，面部反应生成也已从条件生成对抗网络 [27] 发展到具备不确定性感知能力的扩散方法 [49, 51, 57, 70, 103]。视听面对面对话建模也得到了探索 [14, 57, 59, 103]。

在三维人体建模中，大多数方法都根据表演者动作 [12, 17, 46, 47, 87]合成反应者动作。例如，InterFormer [12] 在Transformer中采用时空注意力，而 ReGenNet [87] 和 ReMoS [17]则使用扩散模型生成全身动作。最近，HERO [93] 生成

直接从 RGB视频生成3D反应动作，并结合表演者的面部表情来捕捉情感线索。与之不同的是，我们的方法根据说话者的话语生成3D反应者动作，其中包含转录文本、音频和可选的情绪标注。转录文本是一种轻量、易用的模态，音频提供丰富的语音线索，而情绪标签则明确指示情绪状态，从而有助于更有效地建模交互。

3D人体交互数据集。近期的数据集推动了多人动态和交互感知3D动作的研究。一些工作 [20, 25, 41, 86, 92] 提供成对的人体动作，将交互建模为对称的运动学耦合，即根据一位参与者的动作预测另一位参与者的动作。这种方法虽有助于实现空间协调，却忽略了推动对话的语言和情感信号。

其他数据集 [31, 32, 35, 56, 67, 68, 93] 提供无声的RGB视频以及3D反应动作，提供了更丰富的背景信息，但仍缺少语音语义和情感线索，而这些对于交际意图至关重要。一些数据集 [24,36,55,73] 同时包含人际互动的音频和动作，但其动作主要集中在上半身，例如手臂，而且仅限于一对一的说话者-倾听者配对。

相比之下，我们的数据集提供说话者话语与听者反应动作之间的一对多映射。每个话语都有多个反应，并标记为金级、银级和负级，分别代表恰当、部分恰当和无关的反应，因此更适用于实际应用。此外，动作更加丰富，例如跳跃，从而能够呈现更多样的人体反应。

3 任务定义

本文研究双人互动中的倾听者反应动作生成，互动由一位说话者 和一位听者组成。给定说话者话语 C^s，目标是生成适当的听者反应身体动作，记为 R^l。形式上，目标是学习以下条件分布：

$$p_\theta\left(R^l \mid C^s\right), \qquad C^s \in \left\{A^s,\ T^s,\ (A^s, T^s),\ (A^s, E^s),\ (T^s, E^s),\ (A^s, T^s, E^s)\right\}. \tag{1}$$

其中， A^s 表示说话人音频， T^s 表示对应的文本转录，E^s 表示说话者情绪， θ 表示模型参数。如公式 1 所示， C^s 可以由说话者话语的不同模态或其组合构成。在推理阶段，可以从 $p_\theta(R^l \mid C^s)$ 中采样得到多样的听者反应。

与传统的文本到动作生成不同，说话者话语并未明确指定目标听者动作。因此， C^s 到R^l 的映射本质上是一对多的，这要求模型在保持多样性的同时，生成符合语境的动作。

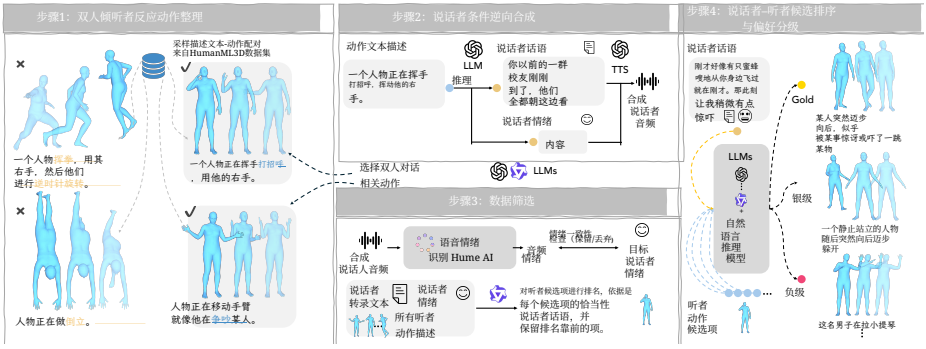


图 2：ReactMotionNet 数据集构建。 我们整理双人对话中的听者动作（步骤 1），通过逆向推断和文本转语音（TTS）合成说话者条件（步骤 2），筛除不可靠样本（步骤 3），并对说话者–听者配对进行排名和重新分级，形成金级/银级/负级偏好（步骤 4）。

4 ReactMotionNet 数据集

为弥合现有三维人体动作交互数据集与现实世界对话动态之间的差距，我们构建了一个名为 ReactMotionNet 的数据集，其中包含一对多说话者话语—听者反应映射，并提供分级适切性标注。为构建该数据集，我们提出了一种新颖的数据构建流程（图 2），利用强大的大语言模型 [58, 89]，将现有的人体动作数据重新整理为说话者—听者动作—响应配对，从而避免了成本高昂的数据收集。

4.1 数据集构建流程

步骤 *1*：双人倾听者反应动作整理。 不同于现有主要关注静止站立时上半身动作的音频驱动三维人体交互数据集，我们从更具动态性且应用广泛的 HumanML3D 数据集 [19]中整理动作。借助动作的文本描述，我们使用多个基于大语言模型的验证器（例如 ChatGPT-o1 [29],ChatGPT-o3 mini [58]）筛除与对话无关的动作（例如，倒立）。此步骤得到一组具有反应语义的动作，作为听者的反应动作。

步骤 *2*：说话者条件逆向合成。对于上一步中的每个听者动作 R^l，我们推断出多个合理的说话者话语，这些话语可能会引发观察到的反应。具体而言，我们将听者动作的描述文本输入 OpenAI o3-mini [1, 58, 69]，以生成潜在的说话人转录文本 T^s 及其对应的情绪标签 E^s。我们在话语生成过程中纳入情绪因素，因为说话者的情绪状态会影响听者的反应。例如，同一句转录文本“随你怎么做”，可能会引发不同的回应：

表 1：数据集统计。#配对数是带标签的说话者—听者配对（即候选反应）的总数。#转录文本、#音频和#情绪分别表示不同转录文本、音频文件和情绪类别的数量。#动作表示不同动作序列的数量。#动作/话语表示每条说话者话语对应的候选动作平均数量。标签计数表示金级/银级/负级候选项的数量（#$\mathcal{G}$/#$\mathcal{S}$/#$\mathcal{N}$）。

划分	#配对	说话者话语			听者反应	#动作/话语。	标签 (y)
		#转写	#音频	#情绪	#动作	(平均值)	(#$\mathcal{G}$/#$\mathcal{S}$/#$\mathcal{N}$)
训练集	137,879	6,631	6,631	46	1,822	20.79	7,527 / 30,862 / 99,490
Val	6,790	841	841	40	195	8.07	903 / 1,682 / 4,205
Test	6,659	826	826	39	197	8.06	877 / 1,652 / 4,130
All	151,328	8,298	8,298	47	2,029	18.24	9,307 / 34,196 / 107,825

支持性的语气可能会让听者开心地原地跳跃，而沮丧的语气则可能让听者感到受伤，转身离开。给定 T^s 和 E^s，我们使用 GPT-4o mini 文本转语音 [28] 合成相应的说话人音频 A^s。这些步骤生成了一组可能的说话者话语（A^s、 T^s、 E^s）。

步骤 *3*：数据筛选。我们执行一系列流程以确保数据集质量。首先，对于每条说话者话语，我们都会核实合成音频 A^s 是否如实体现了预期情绪 E^s。具体而言，我们使用自动语音情感识别器（即，Hume AI [6]）分析生成的音频，并丢弃预测情绪与指定情绪标签不一致的话语。接下来，我们将每条保留的说话者话语与步骤 1 中获得的每个听者反应动作 R^l 的描述文本配对。然后，我们使用 Qwen（Qwen3-235B-A22B-Instruct） [89] ，为每一组说话者话语与听者动作描述文本配对评定其双人对话适切性分数。对于每条说话者话语，我们仅保留得分最高的若干条听者反应动作，从而剔除不恰当的配对。

步骤 *4*：说话者–听者候选排序与偏好分级。给定一对由说话者话语及其在步骤 3 中对应的一个听者反应动作组成的数据，我们使用多个智能体（即 ChatGPT-o1 [29],ChatGPT-o3 mini [58], 和 Qwen3-235B-A22B-Instruct [89]）对该数据对进行评估。他们依据以下两项指标进行评分：（1）语义适切性（反应是否符合该话语），以及（2）对话合理性（其是否听起来像自然的双人对话回应）。我们进一步使用自然语言推理（NLI）模型[7]，以验证听者动作描述文本是否能从说话者话语中逻辑合理地推断出来。随后，我们对各智能体的分数进行加权求和，得到最终分数，并依据预定义的阈值将该数据对标记为金级、银级或负级。

[6] https://www.hume.ai/expression-measurement[7] https://huggingface.co/MoritzLaurer/deberta- v3- large- zeroshot- v1.1- all-33

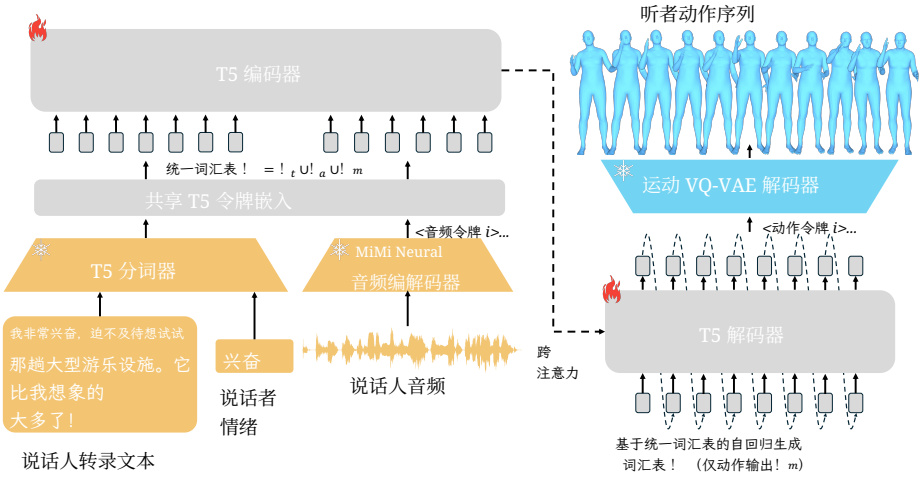


图 3：ReactMotion框架概览。我们使用特定于模态的分词器将原始数据（即说话者的话语（包括转录文本、音频和情绪）以及听者的反应动作）转换为离散特殊令牌。利用这些分词器，我们采用序列到序列模型整合不同模态的信息，并学习根据说话者的话语生成听者的反应动作。

4.2 数据集统计

总体而言，我们的数据集包含 151,328 对带标签的（说话者话语、听者反应动作）数据，涵盖 8,298 条独特的说话者话语和 2,029 个独特的听者反应动作。平均而言，每条说话者话语对应 18.24 个候选反应动作，凸显了听者反应一对多 的特性。总体来看，分别有 9,307、34,196 和 107,825 对数据被标记为金级、银级和负级，体现了候选反应的分级恰当性 。我们按照说话者话语 以 8:1:1 的比例划分训练集/验证集/测试集，从而确保不同划分之间的说话者话语互不重叠（即同一话语不会出现在多个划分中）。表 1 列出了详细统计数据。我们的自动化构建流程还支持轻松扩展到更大规模的数据集。

5 方法

我们提出 ReactMotion，这是一个用于基于说话者话语的反应式听者动作生成的统一框架。如图 3 所示，我们首先引入模态专属分词器，将原始输入（即说话者话语（包括转录文本、音频和情绪）以及听者的反应动作）转换为离散特殊令牌。借助这些分词器，我们使用序列到序列模型统一不同模态的信息，并学习该任务的条件分布（公式 1）。为捕捉双人互动的一对多特性，我们进一步使用基于组级偏好的学习目标训练模型，

该目标明确允许针对同一说话者话语生成多个恰当的反应。

5.1 模态特定分词

我们采用模态专属分词器，将不同模态的原始数据转换为离散标记。

音频分词。 我们使用 Moshi [15]（其神经音频编解码器 MiMi），将说话者话语 A^s 中的音频波形转换为离散码。具体而言，我们采用其音频编码器 $\mathcal{E}_{\text{aud}}(\cdot)$，从 A^s中提取音频特征，然后使用基础码本 $\mathcal{C}_{\text{aud}}$对其进行量化。

$$h_a^s = \mathcal{E}_{\text{aud}}(A^s), \qquad x_a^s = \mathcal{Q}_{\text{aud}}(h_a^s), \tag{2}$$

其中，量化器 $\mathcal{Q}_{\text{aud}}(\cdot)$ 会将特征映射到码本 $\mathcal{C}_{\text{aud}}$中距离最近的条目，并输出相应的码本索引 x_a^s。这些索引随后被视为离散音频标记，使统一模型能够纳入音频信息，同时保留有助于反应行为的韵律和副语言线索。

动作分词。 我们将听者的响应动作 R^l 表示为离散标记，其方式 [96],与音频分词过程类似：

$$h_m^l = \mathcal{E}_{\text{mot}}(R^l), \qquad x_m^l = \mathcal{Q}_{\text{mot}}(h_m^l). \tag{3}$$

其中， $\mathcal{E}_{\text{mot}}$ 和 $\mathcal{Q}_{\text{mot}}$ 分别是动作编码器和量化器，x_m^l 是动作码本 $\mathcal{C}_{\text{mot}}$中的离散索引。

此外，统一模型以离散令牌形式预测的听者反应动作，可以通过以下方式映射回原始动作数据：

$$h_m^l = \mathcal{Q}_{\text{mot}}^{-1}(x_m^l), \qquad R^l = \mathcal{D}_{\text{mot}}(h_m^l), \tag{4}$$

其中， $\mathcal{Q}_{\text{mot}}^{-1}(\cdot)$将离散令牌索引映射为码本中的向量，而 VQ-VAE 动作解码器 [82, 96] $\mathcal{D}_{\text{mot}}(\cdot)$则将这些向量解码回原始动作数据。

5.2 统一的序列到序列建模

借助上述模态专属分词器，我们现在可以将不同模态的信息表示在统一空间中，从而使序列到序列模型能够根据说话者话语生成听者反应动作。

具体而言，我们采用 T5-base [65] 作为序列到序列骨干网络，并扩展其原始文本词汇表 V_t，纳入音频和动作词汇表：

$$V = V_t \cup V_m \cup V_a \cup V_s, \tag{5}$$

其中，V_m 是动作码本 $\mathcal{C}_{\text{mot}}$的编码索引，表示为{<动作令牌 i>}$_{i=0}^{|V_{\mathcal{C}_{\text{mot}}}|-1}$；$V_a$ 是音频码本$\mathcal{C}_{\text{aud}}$的编码索引，分别表示为 {<音频令牌 i>}$_{i=0}^{|V_{\mathcal{C}_{\text{aud}}}|-1}$。$V_s$ 包含特殊标记，例如 <动作令牌>、 </动作令牌>、 <音频标记>、</音频标记>、<情绪> 和 </情绪>，用于包裹动作、音频和情绪标记序列。

这一统一词汇表使我们能够以通用格式对不同模态或其组合 C^s 条件下的听者动作响应生成进行建模，并通过单一模型实现。具体而言，我们首先将说话者话语 C^s 和听者反应动作 R^l 的离散编码纳入固定的提示模板。由于篇幅限制，此处展示了一个仅以说话人音频为条件的粗略任务模板示例；详细示例以及其他条件下使用的模板见附录 A.2。

输入：你正在对说话者-听者二元互动进行建模。给定说话者_音频： [音频令牌占位符], , 仅返回听者反应动作标记序列。**输出**： [动作令牌占位符]

现在，生成听者反应动作的建模过程可表示为自回归过程，其中每个动作令牌以概率 $p_\theta\left(x_t^{\text{out}} \mid x^{\text{in}}(C^s)\ x_{<t}^{\text{out}}\right)$ 生成。这里，$x^{\text{in}}(C^s)$ 是嵌入输入说话者话语 C^s 的任务模板的输入标记序列，而 x^{out} 是输出标记序列，即听者反应动作 x_m^l。

5.3 组级偏好学习

单个说话者话语 C^s 可能对应多个合理的听者反应动作 R^l。直接使用这类一对多配对进行微调，可能会导致模型退化为平均且安全的行为，例如静止不动。为缓解这一问题，我们使用组级偏好学习来训练模型。

对于每个说话者话语 C^s，我们从每个标签中随机采样与之对应的听者动作，构建一个组 $\{\mathcal{G},\mathcal{S},\mathcal{N}\}$，其中 $\mathcal{G}$、 $\mathcal{S}$ 和$\mathcal{N}$ 分别表示标记为金级、银级和负级的动作集合。集合中的每个动作 R^l 均表示为一个动作令牌序列 x_m^l。我们使用长度归一化的条件对数似然 [5, 54, 83]计算每个动作的预测分数：

$$\ell(x_m^l \mid C^s) = \frac{1}{|x_m^l|} \sum_{t=1}^{|x_m^l|} \log p_\theta\left(x_{m,t}^l \mid x^{\text{in}}(C^s), x_{m,<t}^l\right). \tag{6}$$

然后，我们使用平滑对数均值指数算子对标签相同的动作的预测分数进行汇总：

$$\ell_{\mathcal{A}}(C^s) = \log\left(\frac{1}{|\mathcal{A}(C^s)|} \sum_{x^l 2_m \in \mathcal{A}(C^s)} \exp\left(\ell(x_m^l \mid C^s)\right)\right), \quad \mathcal{A} \in \{\mathcal{G},\mathcal{S},\mathcal{N}\}. \tag{7}$$

由此得到 C^s 的三个预测分数，即 $\ell_{\mathcal{G}}$、$\ell_{\mathcal{S}}$ 和 $\ell_{\mathcal{N}}$，分别对应金级、银级和负级集合。

由于金级动作优于银级动作，银级动作又优于负级动作，因此模型会被引导生成 $\ell_{\mathcal{G}} > \ell_{\mathcal{S}} > \ell_{\mathcal{N}}$。我们通过软间隔排序损失来强制执行这一顺序：

$$\begin{aligned}\mathcal{L}_{\text{rank}} = &\log\big(1+\exp\big(m-(\ell_{\mathcal{G}}-\ell_{\mathcal{S}})\big)\big) \;+\; \log\big(1+\exp\big(m-(\ell_{\mathcal{S}}-\ell_{\mathcal{N}})\big)\big) \\ &+\; \lambda_{gn}\,\log\big(1+\exp\big(m-(\ell_{\mathcal{G}}-\ell_{\mathcal{N}})\big)\big),\end{aligned} \tag{8}$$

其中，m 指定不同标签之间的间隔，而 λ_{gn} 控制金级≻负级约束的强度。

采用频率重加权的训练目标。为减轻频繁出现的动作序列所占的主导地位，我们根据动作序列 ID 应用逆频率加权。令 i为一个组的索引（对应一条说话者话语），并令 r_{ij} 表示组 i中第 j 个候选项的动作序列 ID。我们将 $\text{freq}(r)$ 计算为动作 ID r 在训练集 中出现的次数，并赋予条目权重 $\tilde{w}_{ij} = \frac{1}{\sqrt{\text{freq}(r_{ij})}}$。接着，我们定义组权重为组内条目权重的平均值，$w_i = \frac{1}{|\mathcal{C}_i|}\sum_j \tilde{w}_{ij}$，其中$\mathcal{C}_i$ 表示组 i的候选集。最后，我们在应用排序损失的同时最大化聚合金级分数：

$$\mathcal{L} = \frac{\sum_i w_i\Big(-\ell_{\mathcal{G}}^{(i)} + \lambda_{\text{rank}}\mathcal{L}_{\text{rank}}^{(i)}\Big)}{\sum_i w_i}. \tag{9}$$

6 实验

6.1 实现细节

我们使用默认的 AdamW 优化器和余弦学习率调度训练 ReactMotion，共进行 100,000 次迭代。学习率设置为 2×10^{-5}，并进行 1,000 个预热步数。我们在单个 NVIDIA A100 GPU 上训练，每设备批量大小为 8，并进行 2 步梯度累积。我们使用六种条件变体（T, A, $T{+}A$, $T{+}E$, $A{+}E$, $T{+}A{+}E$），并应用模态丢弃（$p{=}0.3$）以提高鲁棒性（有关实现的更多细节，请参见附录 A）。

6.2 评估协议

评估指标。(i) 反应恰当性，即生成的反应式人体动作对说话者输入的响应程度，是我们任务的核心目标。受基于偏好的评估范式 [6,11,13,16,72,102]启发，我们使用组级胜率**Win(g>G)**、**Win(g>S)**和**Win(g>N)**来评估反应恰当性。具体而言，我们将最佳生成样本 g 与标注为金级（G）、银级（S）和负级（N）的听者动作进行比较，并计算其相对于各参考等级的胜率。相对于某一参考等级获胜，是指

表 2：严格模态缺失（Strict-L2）条件下的多模态评判网络可靠性。我们在**测试集**上评估六种输入模式（文本 T、音频 A、情绪 E及其融合），报告成对胜率（Win(G>N)、Win(G>S)、Win(S>N)）和排序指标（平均倒数排名(G)、归一化折损累计增益@K），并采用分级相关性 G>S>N。

Mode	Win(G>N) ↑	Win(G>S) ↑	Win(S>N) ↑	平均倒数排名(G) ↑	归一化折损累计增益@3 ↑	归一化折损累计增益@5 ↑	归一化折损累计增益@10 ↑
T	0.992	0.873	**0.983**	0.829	0.864	**0.878**	0.932
A	0.992	0.872	**0.983**	**0.832**	**0.866**	**0.878**	**0.933**
T+E	**0.993**	0.876	0.982	0.826	0.857	0.876	0.929
A+E	0.992	0.874	**0.983**	0.831	0.865	**0.878**	**0.933**
T+A	**0.993**	**0.879**	0.982	0.820	0.855	0.875	0.928
T+A+E	**0.993**	0.878	0.982	0.828	0.859	**0.878**	0.930

较高参考等级（例如，银级），表明生成动作的排名高于质量更高的标注响应，反映出更强的反应恰当性。为实现这一评估，我们训练了一个**多模态评判网络**，以便在相同说话者输入条件下对生成的反应式身体动作进行排名。有关评判网络的详细信息见附录。我们还报告**Gen@3**，即在同一组内 $\{\mathcal{G}, \mathcal{S}, \mathcal{N}\}$ 加上生成候选项中，某个生成候选项排名位于前3的组所占比例。(ii)动作质量通过在动作特征空间中计算的弗雷歇 Inception 距离（FID） [23] 进行衡量；(iii) 多样性遵循人体动作生成 [82, 96]，以生成样本之间的平均成对嵌入距离进行衡量。（有关评估指标的更多细节，请参见附录 B.4。）

多模态评判网络验证。由于评判网络是衡量反应恰当性的核心，我们在具有分级恰当性标注（G/S/N）的样本上对其进行验证。具体而言，我们计算分级一致性胜率 **Win(G>S)**、**Win(G>N)**和**Win(S>N)**，以检验评判器是否会为更恰当的反应分配更高分数。数值越高，表明评判器越可靠。我们进一步报告**平均倒数排名(G)**，用于衡量金级反应的排名位置，以及**归一化折损累计增益@3/归一化折损累计增益@5/归一化折损累计增益@10**，以评估前K 个候选项中的分级排序质量。

表 2 显示，评判器在全部六种模式和两个测试集上，都能始终保持预期的偏好顺序，且**几乎完美地区分**不同类别。金级几乎总是优于负级（Win(G>N) ≈ 0.99），银级也明显优于负级（Win(S>N) $\approx$ 0.98），这表明评判器能够可靠地区分较差的动作与合理的动作。同时，金级以明显的间隔胜过银级（Win(G>S) ≈ 0.87–0.88），体现了其对细粒度质量差异的敏感度，而不只是简单地排斥负级。评判器还取得了较高的排名质量（MRR(G) ≈ 0.82–0.84；nDCG@5 ≈ 0.87–0.88；nDCG@10 $\approx$ 0.93），展现出稳定且有意义的前K 名排序。

尽管我们的多模态评判网络是在多种输入模态上训练的，即，文本（T）、音频（A）和情绪（E），但它通过 Strict-L2 支持缺失模态。被禁用的模态会被无信息输入替代（全填充文本、全填充音频码或未知情绪标记）。因此，评判网络能够处理任意模态子集；即使只有一种

模态，它在评估中也表现良好。（有关评判网络的更多详情，请参见附录 B.1 和 B.2）。

6.3 定量结果

由于倾听者反应动作生成仍处于探索初期，我们评估了一组具有代表性的基线方法。**（a）随机选择**从 HumanML3D [19]中均匀采样一个动作序列。**（b）检索**采用先前 HumanML3D 文本到动作工作 [82,96] 中的文本-动作匹配网络，计算文本-动作相似度，并根据说话人转录文本从训练集中检索最近邻的听者动作序列。我们还考虑了更强的**级联式大语言模型→文本到动作**基线：给定一段说话者话语（以及情绪），大语言模型 [89]首先生成听者动作描述，随后将其输入文本到动作生成器，以合成最终动作。我们分别采用 Qwen3-30B-A3B（305 亿参数）和在我们的训练集（说话者话语、听者动作描述）配对数据上进行微调的 Qwen3-4B-Thinking（40 亿参数）作为大语言模型。生成的描述文本被输入两个具有代表性的文本到动作生成器：T2M-GPT [96] 和 MG-MotionLLM [82]。有关基线方法的更多细节请参见附录 B.3。

表 3 显示，ReactMotion 在响应恰当性方面优于所有基线方法。在级联式大语言模型→文本到动作流程中，大语言模型→MG-MotionLLM $*$表现最强，优于随机选择和检索。然而，尽管采用了强大的动作生成器，与银标准参考进行严格比较时，其表现仍然不佳（胜率(g>S)），这表明先生成描述文本、再生成动作的两阶段流程难以生成高度恰当的听者反应。

相比之下，ReactMotion 在各种输入模式下均实现了近乎完美的胜率(g>N)，并显著提升了胜率(g>S)和 Gen@3。我们的完整模型（$T+A+E$）取得了最佳总体胜率，同时保持较低的 FID 和有竞争力的多样性。尽管检索通过构造方式实现了最高的多样性，但其恰当性远低于我们的方法，真实感也更差。更多实验结果见附录 D。

6.4 定性结果

我们在图 4 中展示了具有代表性的示例，对比了我们的 ReactMotion（我们的方法）、经过交叉熵训练的变体（交叉熵）以及大语言模型→MG-MotionLLM $*$，后者采用在训练集上微调的 Qwen [89]（Qwen3-4B-Thinking），并在相同说话者条件下与金级和银级参考反应进行比较。总体而言，**ReactMotion** 生成的反应动作既与说话者内容在语义上保持一致，又能表现出丰富的强度变化。例如，对于带有兴奋情绪的 utterance“The energy in here feels electric right now”，我们的模型会生成幅度更大、更具动感的上身和手臂动作，更好地体现“electric”所传达的高能量感，并符合金级反应中呈现的交流风格。

相比之下，**银级**反应表现为快速的手部挥动，但整体能量较低，因此与兴奋状态不太相符。

表 3：测试集上的定量结果。主要评估指标是 Win(g>N)、Win(g>S)、Win(g>G) 和用于衡量反应恰当性的 Gen@3。我们还评估了动作质量（FID）和多样性。 * 表示大语言模型使用测试集说话者话语和听者动作描述对进行微调。

方法	输入模态	Win(g>N)↑	Win(g>S)↑	Win(g>G)↑	Gen@3↑	FID↓	多样性↑
GT	-	-	-	-	-	0.278	6.187
随机选择	-	0.265	0.122	0.006	0.099	42.363	9.880
检索	T	0.392	0.252	0.130	0.206	7.429	**8.207**
大语言模型→T2M-GPT	$T+E$	0.138	0.038	0.016	0.199	49.920	4.946
大语言模型→T2M-GPT *	$T+E$	0.171	0.027	0.017	0.350	42.589	6.102
大语言模型→MG-MotionLLM	$T+E$	0.775	0.245	0.044	0.345	23.629	5.082
大语言模型→MG-MotionLLM *	$T+E$	0.883	0.274	0.047	0.380	25.723	4.546
ReactMotion（我们的方法）	T	0.993	0.774	0.258	0.916	**4.706**	4.789
ReactMotion（我们的方法）	A	0.992	0.614	0.164	0.864	6.221	4.009
ReactMotion（我们的方法）	$T+E$	0.990	0.696	0.206	0.930	5.422	4.475
ReactMotion（我们的方法）	$A+E$	0.993	0.736	**0.323**	**0.981**	6.485	4.162
ReactMotion（我们的方法）	$T+A$	0.993	0.651	0.215	0.931	6.560	4.145
ReactMotion（我们的方法）	$T+A+E$	**1.000**	**0.797**	0.266	0.960	4.760	4.804

交叉熵 变体往往会退化为通用且条件约束较弱的响应（例如，交叉双臂等静态姿势），表明其利用偏好结构和对反应行为一对多特性的建模能力有限。最后，**大语言模型→文本到动作** 基线通常会生成重复性动作（例如，几乎持续挥手），时间变化有限，因此似乎不太适合双人交流，因为反应通常会随时间演变（例如，双手抬起和放下、姿势变化以及细微转身）。此外，由于双人反应难以用自然语言描述，大语言模型生成的域外描述可能存在噪声，导致 MG-MotionLLM 生成效果下降，包括动作序列过短。

6.5 用户研究

我们招募了59名志愿者，并开展用户研究，以评估由**ReactMotion （我们的方法）**生成的听者动作的反应恰当性，并与两个基线（**交叉熵** 变体和**大语言模型→MG-MotionLLM** *）以及组内最佳 **银级** 参考进行比较。在每种情况下，参与者观看基于相同说话者话语（显示音频及其转录文本）生成的两个动作视频（A/B），并选择更恰当的听者反应。每位参与者完成36个案例，涵盖六种说话者条件（每种条件进行六次两两比较）。

如图5所示，**我们的方法**优于生成式基线模型，对**交叉熵**的胜率为 67.8%，对**大语言模型→MG-MotionLLM** *的胜率为 72.0%。**我们的方法**与**银级**参考方案相比也颇具竞争力，在**银级**与**我们的方法**的比较中获得 44.1% 的投票，显著高于**交叉熵**（31.9%）和**大语言模型→MG-MotionLLM** （31.4%）。

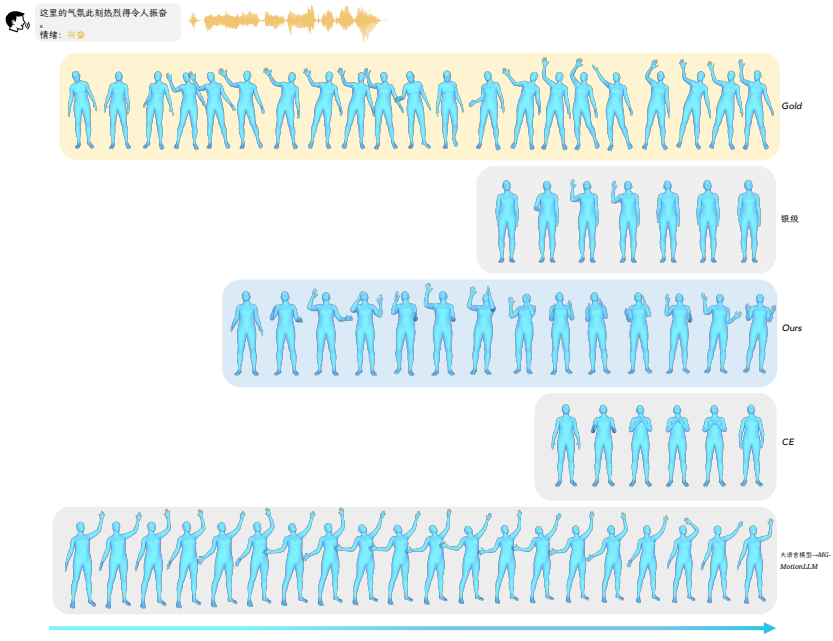


图 4：定性结果。我们比较金级和银级听者反应、由我们的 ReactMotion（我们的方法）生成的动作、交叉熵训练变体（交叉熵），以及级联式大语言模型→文本到动作基线；这些结果均以相同的说话者话语为条件。我们将生成的三维动作序列可视化。

6.6 消融研究

模态研究。我们在表 3 中研究了输入模态的影响。总体而言，多模态融合在各种设置下表现最佳。文本是最强的单一线索，具有较高的对齐度和最低的单模态 FID（例如，T: Win(g>N)=0.993, Win(g>S)=0.774, FID=4.706）。单独使用音频在细粒度适切性方面表现较弱，但加入情绪后性能显著提升（最佳 Win(g>G)=0.323 和 Gen@3=0.981）。完全融合（$T+A+E$）表现最为均衡，取得最佳 Win(g>N)=1.000、较高的 Win(g>S)=0.797，以及较低的 FID=4.760。

组级偏好学习的消融研究。 表4对我们的组级偏好学习目标中的关键组件进行了消融。与仅使用交叉熵训练相比，我们的**完整模型** 显著提升了反应的恰当性和动作质量（例如，Win(g>S)：0.741→0.797；Gen@3：0.938→0.960；FID：6.555→4.760）。移除**逆频率重加权** 会导致恰当性下降幅度最大，尤其是在与最高层级对比时（Win(g>G)：0.266→0.220），这凸显了抑制高频、通用动作主导地位的重要性。移除**排序损失** 会降低保真度（FID：4.760→5.950），同时提高多样性（4.804→5.453），这表明排名约束有助于确保各层级之间正确的相对顺序。最后，重新

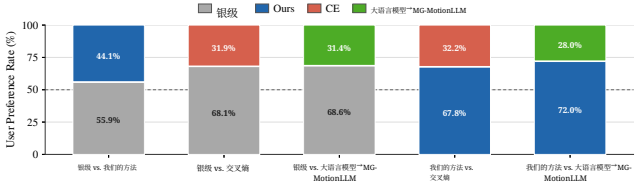


图5：用户研究的反应恰当性。

表4：消融研究 在测试划分上的结果（除非另有说明，均使用 $A{+}T{+}E$ ）。w/o 表示训练时未使用相应组件。**CE 基线** 通过将每个说话者输入与单个金级反应配对作为监督，仅使用交叉熵损失训练相同的模型。

方法	胜率(g>N)↑	胜率(g>S)↑	胜率(g>G)↑	生成@3↑	FID↓	多样性↑
CE 基线	0.990	0.741	0.262	0.938	6.555	5.448
我们的方法（完整）	**1.000**	**0.797**	**0.266**	**0.960**	**4.760**	4.804
不使用 逆频率重加权	0.979	0.704	0.220	0.946	5.177	4.929
不使用 $\mathcal{L}_{\text{rank}}$	0.996	0.781	0.260	**0.960**	5.950	**5.453**
不使用 $\ell_{\mathcal{G}}$	0.996	0.712	0.215	0.943	6.376	4.493

移动 $\ell_{\mathcal{G}}$ 会持续损害恰当性和质量，这表明对黄金反应进行似然监督仍然是必要的。

7 结论

我们提出基于说话者话语的反应式听者动作生成，这是一项用于建模双人互动中听者动作响应的新任务。为支持这项任务，我们推出**ReactMotionNet**，这是一个多模态数据集，能够明确捕捉人类行为固有的非确定性：针对每条说话者话语，我们提供多个候选听者动作及偏好标注，从而实现超越单一“真实”响应的监督。基于这一数据集设计，我们制定了专为反应式动作生成量身打造的偏好导向评估协议。最后，我们提出**ReactMotion**，这是一个处理多模态说话者线索的统一框架，在动作质量和反应恰当性方面显著优于强基线。我们相信，这项工作为未来建模双人互动的研究奠定了基础。

补充材料概要

补充材料安排如下：

- **A 节** 介绍实现细节，包括模型配置、词汇表构建、优化设置和训练超参数。

- **A.1 节**介绍 ReactMotion 的模型大小。
- **A.2 节**：针对不同说话人条件设置的提示模板；
- **B 节**进一步提供更多评估详情，包括：
  - **B.1 节**：多模态评判网络的构建；
  - **B.3 节**：基线方法的详细介绍。
- **B.4 节**介绍了评估指标，涵盖反应恰当性、动作质量和多样性。

- **C 节**提供了有关 ReactMotionNet 数据集的补充统计信息和分析。

- **D.1 节**展示了超参数敏感性分析，包括全扫描结果、代表性配置和热图可视化。

- **D.2 节**评估了所提方法的推理效率。
- **D.3节**介绍了用户研究的流程和结果。
- **D.4节**展示了具有代表性的失败案例。
- **E节**讨论了当前框架的局限性。

A 实现细节

表5总结了我们实验中采用的关键实现细节和训练超参数。具体而言，ReactMotion采用T5-base序列到序列骨干网络，包含2.229亿个骨干网络参数；扩展词汇表后，可训练参数为2.359亿个。根据方法部分的描述，原始文本词汇表（$|V_t| = 32,100$）扩充了动作令牌（$|V_m| = 512$）、MiMi 音频词元（每个码本包含$|V_a| = 2,048$个；共8个码本）以及用于标记不同模态边界的模态专用特殊令牌，最终统一词汇表的大小为63，338。值得注意的是，为了完整起见，词汇表包含全部8个MiMi码本中的词元；但在实际训练中，我们只使用基础码本中的词元，以加快训练过程。模型以经过词元化的说话者话语为输入，并以自回归方式预测听者反应动作令牌；源序列和目标序列的最大长度分别设为512和256。我们使用AdamW训练模型，学习率为 2.0×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$，权重衰减为0.0，预热步数为1,000，每设备批量大小为8，梯度累积步数为2，总优化步数为100,000。为捕捉说话者话语到合理听者反应的一对多映射，训练采用所提出的组级偏好目标，并使用 $\lambda_{\text{rank}} = 0.25$, $\lambda_{\text{gn}} = 0.25$和间隔$m = 0.5$。此外，我们以0.3的概率应用模态丢弃，以提高模型对缺失模态的鲁棒性；偏好优化期间则使用长度归一化LogSumExp聚合，以获得稳定的集合级分数。

表5： 训练中使用的实现细节和超参数 .

设置	数值
序列到序列骨干模型	T5-base [65]
文本分词器	T5-base分词器 [65]
音频分词器	MiMi 神经音频编解码器 [15]
动作分词器	来自 T2M-GPT 的 VQ-VAE [96]
每设备批量大小	8
梯度累积步数	2
训练步数	100,000
预热步数	1,000
优化器	AdamW
Adam β_1	0.9
Adam β_2	0.999
权重衰减	0.0
学习率	2.0×10^{-5}
最大源序列长度	512
最大目标长度	256
文本词汇表大小 $\lvert V_t \rvert$	32,100
音频码本大小 $\lvert V_a \rvert$	2,048
MiMi 音频码本数量	8
运动 VQ-VAE 码本大小 $\lvert V_m \rvert$	512
总词汇表大小 $\lvert V \rvert$	49,002
骨干网络参数	222.9M
词汇表扩展后的总可训练参数	235.9M
排序损失权重 λ_{rank}	0.25
金负样本损失权重 λ_{gn}	0.25
排序间隔 m	0.5
模态丢弃率	0.3
LogSumExp 归一化	已启用

A.1 模型大小

表 6： ReactMotion 的模型配置和参数。

指标	值
骨干网络参数	222.9M
可训练参数总数 235.9M	
统一词汇表大小	49,002

表6总结了ReactMotion的模型大小。该模型基于T5-base骨干网络构建，包含2.229亿个参数；扩展词汇表以纳入多模态令牌后，可训练参数为2.359亿个。

A.2 提示模板

为支持不同说话人条件设置下的统一生成，我们将可用的说话者线索转换为固定的自然语言提示模板。

给定由转录文本、音频和可选情绪标注组成的说话者话语，我们通过有选择地启用相应字段来构建输入提示。模型被要求以严格的格式输出仅听者动作标记序列，不附加任何自然语言内容。

形式上，对于说话者话语 C^s，提示构造如下：

输入：
你正在对说话者-听者二元互动进行建模。输入：- SPEAKER_转录文本： [说话者转录文本] - SPEAKER_音频： [说话人音频] - SPEAKER_情绪： <情绪> [说话者情绪] </情绪>输出：仅返回严格符合以下格式的一系列听者动作令牌：<动作令牌> <动作令牌 i> ... </动作令牌>请勿输出任何其他文字。

在实际应用中，提示中的字段会根据所选条件模式启用或禁用。例如，当使用转录文本但不使用音频时，SPEAKER_音频字段留空；禁用情绪时，则完全省略情绪行。此设计使我们能够在单一统一框架中处理纯文本、纯音频、文本+音频、文本+情绪、音频+情绪以及文本+音频+情绪设置。

下面展示几个具体示例。

纯文本条件（T）。

输入：
你正在对说话者-听者二元互动进行建模。输入：- 说话者_转录文本： [说话人转录文本] - 说话人音频： ¯输出：仅返回听者动作令牌序列，格式必须严格如下：<动作令牌> <动作令牌 i> ... </动作令牌>

不要输出任何其他文字。

文本+情绪条件（$T+E$）。

输入：
你正在对说话者-听者二元互动进行建模。输入：- 说话者_转录文本： [说话者转录文本] - 说话人音频： ¯ - 说话者_情绪： <情绪> [说话者情绪] </情绪>输出：仅返回听者动作令牌序列，格式必须严格如下：<动作令牌> <动作令牌 i> ... </动作令牌>

请勿输出任何其他文字。

纯音频条件（A）。

输入：
你正在对说话者-听者二元互动进行建模。输入：- 说话者转录文本： ¯ - 说话者_音频： [说话人音频]输出：仅返回听者动作令牌序列，格式必须严格如下：

<动作令牌> <动作令牌 i> ... </动作令牌>请勿输出任何其他文字。

音频+情绪条件（$A+E$）。

输入:
你正在对说话者-听者二元互动进行建模。输入：- 说话者转录文本：¯ - 说话者_音频： [说话人音频] - 说话者_情绪： <情绪> [说话者情绪] </情绪>输出：仅返回听者动作令牌序列，并严格遵循以下格式：<动作令牌> <动作令牌 i> ... </动作令牌>

不要输出任何其他文字。

文本+音频条件（$T+A$）。

输入:
你正在对说话者-听者二元互动进行建模。输入：- 说话者_转录文本： [说话者转录文本] - 说话者_音频： [说话人音频]输出：仅返回一串听者动作标记，格式必须严格如下：<动作令牌> <动作令牌 i> ... </动作令牌>

不要输出任何其他文字。

文本+音频+情绪条件（$T+A+E$）。

输入:
你正在对说话者-听者二元互动进行建模。输入：- 说话者_转录文本： [说话者转录文本] - 说话者_音频： [说话人音频] - 说话者_情绪： <情绪> [说话者情绪] </情绪>输出：仅返回一串听者动作标记，格式必须严格如下：<动作令牌> <动作令牌 i> ... </动作令牌>不要输出任何其他文字。

给定构造的提示 $x^{\text{in}}(C^s)$，模型自回归地预测听者动作标记序列 x^{out}，如下所示

$$p_\theta\left(x_t^{\text{out}} \mid x^{\text{in}}(C^s), x_{<t}^{\text{out}}\right).$$

其中，$x^{\text{in}}(C^s)$ 表示根据说话者话语 C^s 实例化的提示序列，x^{out} 表示输出听者动作词元序列。

B 附加评估细节

B.1 多模态评判器网络

为了评估生成的听者动作的反应恰当性，并支持最佳 K 选择，我们训练了一个多模态评判网络，如

图 6 所示。给定说话者话语 C^s 和候选听者动作令牌序列 x_m^l，评判网络 s_ψ 会输出一个标量兼容性分数

$$s_\psi(C^s, x_m^l) \in \mathbb{R}, \tag{10}$$

其中，较大的值表示候选听者动作对于给定说话者话语更为合适。

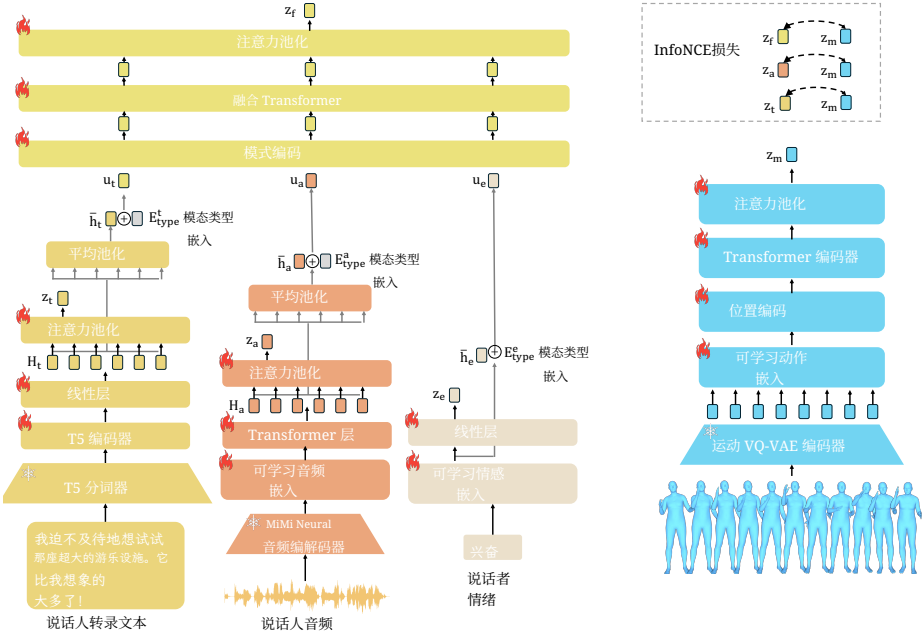


图 6：多模态评判网络的架构。给定一段说话者话语和一个候选听者动作，评判器通过三个模态专属分支对转录文本、MiMi 音频词元和离散情绪标签进行编码，生成模态嵌入 z_t、z_a 和 z_e，以及用于构成融合词元 u_t、u_a 和 u_e 的隐藏状态摘要。向这些词元中添加模态类型嵌入和模式嵌入，随后通过融合 Transformer 和注意力池化进行处理，以获得统一条件嵌入 z_f。与此同时，由 VQ-VAE 动作词元表示的候选听者动作通过动作 Transformer 编码，并池化为动作嵌入 z_m。评判器在共享归一化评分空间中计算条件嵌入与动作嵌入之间的兼容性。训练期间，对融合嵌入和辅助模态专属嵌入应用组级 InfoNCE 目标，从而能够在完整和部分说话者话语条件下实现可靠评分。雪花和火焰图标分别表示冻结模块和可训练模块。

架构。 它包含三个分支，用于编码说话者话语 C^s中的不同模态：文本转录、音频和情绪；

一个融合分支，用于整合 C^s 中的可用信息，同时允许缺失模态；以及一个动作分支，用于编码响应动作。所有分支都会将特征从维度 d 投影到维度为 d_o 的共享评分空间。默认情况下，所有分数空间嵌入都会进行 ℓ_2归一化。

文本分支。令 x_t^s 表示经过分词的说话人转录文本，并令 $M_t \in\{0,1\}^{T_t}$ 表示其填充掩码，其中 $M_t(j) = 1$ 表示第 j 个词元有效，$M_t(j) = 0$ 表示填充。我们使用 T5编码器 $\mathcal{E}_{\mathrm{T5}}(\cdot)$对转录文本进行编码，并将得到的隐藏状态投影到共享隐藏空间：

$$H_t = W_t\,\mathcal{E}_{\mathrm{T5}}(x_t^s) + b_t, \tag{11}$$

其中 $H_t \in \mathbb{R}^{T_t\times d}$、$W_t \in \mathbb{R}^{d\times d_{\mathrm{T5}}}$ 和 $b_t \in \mathbb{R}^d$ 是可学习参数。接着，我们将词元级特征聚合为评分空间中的文本嵌入：

$$\tilde{z}_t = \mathrm{AttnPool}_t(H_t; M_t), \qquad z_t = \mathrm{L2Norm}(\tilde{z}_t), \tag{12}$$

其中 $\tilde{z}_t,\ z_t \in \mathbb{R}^{d_o}$，$\mathrm{AttnPool}_t(\cdot)$表示一种掩码注意力池化算子，它根据 M_t忽略填充位置；L2Norm 表示 ℓ_2 归一化。

音频分支。令 x_a^s 表示通过 MiMi神经编解码器分词器 [15]获得的说话人音频词元序列。由于 MiMi 音频由多个码本表示，我们首先将离散词元映射为嵌入，添加可学习的码本级嵌入和位置嵌入，然后使用 Transformer 编码器处理所得序列：

$$H_a = \mathcal{E}_a\big(\mathrm{Emb}_a(x_a^s) + E^a_{\mathrm{lvl}} + E^a_{\mathrm{pos}}\big), \tag{13}$$

其中，$H_a \in \mathbb{R}^{T_a\times d}$，$\mathrm{Emb}_a(\cdot)$表示可学习音频令牌嵌入层，$E^a_{\mathrm{lvl}}$ 表示可学习码本级嵌入，E^a_{pos} 表示可学习位置嵌入，$\mathcal{E}_a$ 表示音频Transformer编码器。令 $M_a \in \{0,1\}^{T_a}$ 表示音频填充掩码，其中 $M_a(j) = 1$ 表示有效音频令牌，而 $M_a(j) = 0$ 表示填充。令牌级音频特征经过池化后得到音频嵌入：

$$\tilde{z}_a = \mathrm{AttnPool}_a(H_a; M_a), \qquad z_a = \mathrm{L2Norm}(\tilde{z}_a), \tag{14}$$

其中 $\tilde{z}_a,\ z_a \in \mathbb{R}^{d_o}$。

情绪分支。令e^s 表示离散的说话者情绪标签。我们将其映射为可学习嵌入，并将其投影到共享评分空间中：

$$h_e = \mathrm{LayerNorm}\big(\mathrm{Emb}_e(e^s)\big), \qquad \tilde{z}_e = W_e h_e + b_e, \qquad z_e = \mathrm{L2Norm}(\tilde{z}_e), \tag{15}$$

其中，$\mathrm{Emb}_e(\cdot)$是可学习情绪嵌入表，$h_e \in \mathbb{R}^d$、$W_e \in \mathbb{R}^{d_o\times d}$、$b_e \in \mathbb{R}^{d_o}$和 $\tilde{z}_e$，$z_e \in \mathbb{R}^{d_o}$。

融合分支。为统一所有可用信息，我们在 d 维隐藏空间中为每种模态构建一个融合令牌。令 $o \subseteq \{t, a, e\}$

表示活跃模态集，并令 $\delta_k(o) \in \{0,1\}$ 表示在模式 o 下模态$k \in \{t,a,e\}$ 是否可用。

对于文本和音频，我们通过在有效位置上进行掩码平均池化来汇总隐藏状态:

$$\bar{h}_t = \begin{cases} \dfrac{\sum_{j=1}^{T_t} M_t(j)\, H_t(j)}{\sum_{j=1}^{T_t} M_t(j)}, & \delta_t(o)=1, \\ 0, & \delta_t(o)=0, \end{cases} \quad \bar{h}_a = \begin{cases} \dfrac{\sum_{j=1}^{T_a} M_a(j)\, H_a(j)}{\sum_{j=1}^{T_a} M_a(j)}, & \delta_a(o)=1, \\ 0, & \delta_a(o)=0, \end{cases} \tag{16}$$

其中，$H_t(j)$ $H_a(j)$ $\in \mathbb{R}^d$ 表示第 j 个隐藏状态。对于已由单个隐藏向量表示的情绪，我们定义

$$\bar{h}_e = \begin{cases} h_e, & \delta_e(o)=1, \\ 0, & \delta_e(o)=0. \end{cases} \tag{17}$$

随后，我们构建模态专属融合令牌

$$u_t = \bar{h}_t + E^t_{\text{type}}, \qquad u_a = \bar{h}_a + E^a_{\text{type}}, \qquad u_e = \bar{h}_e + E^e_{\text{type}}, \tag{18}$$

其中，u_t， u_a， $u_e \in \mathbb{R}^d$以及 E^t_{type}、 E^a_{type}和 E^e_{type}是可学习的类型嵌入。

为显式编码哪些模态处于激活状态，我们进一步引入可学习的模式嵌入 $E_{\text{mode}}(o)$ $\in \mathbb{R}^d$。初始融合令牌序列为

$$X_f^{(0)} = \begin{bmatrix} u_t + E_{\text{mode}}(o) \\ u_a + E_{\text{mode}}(o) \\ u_e + E_{\text{mode}}(o) \end{bmatrix} \in \mathbb{R}^{3\times d}. \tag{19}$$

由于某些模态可能缺失，我们定义一个模态存在掩码

$$M_f = \big[\delta_t(o),\, \delta_a(o),\, \delta_e(o)\big] \in \{0,1\}^3. \tag{20}$$

融合序列通过带掩码的 Transformer 编码器进行处理:

$$H_f = \mathcal{E}_f\big(X_f^{(0)}; M_f\big), \qquad \tilde{z}_f = \text{AttnPool}_f(H_f; M_f), \qquad z_f = \text{L2Norm}(\tilde{z}_f), \tag{21}$$

其中 $H_f \in \mathbb{R}^{3\times d}$, $\mathcal{E}_f$表示多模态融合 Transformer， $\tilde{z}_f$， z_f $\in \mathbb{R}^{d_o}$。

动作分支。 每个候选听者动作都表示为动作令牌序列 x^l_m，该序列由动作 VQ-VAE 分词器生成。我们将动作令牌映射为嵌入，添加位置嵌入，并使用动作 Transformer 对其进行编码:

$$H_m = \mathcal{E}_m\big(\text{Emb}_m(x^l_m) + E^m_{\text{pos}}\big),\ \tilde{z}_m = \text{AttnPool}_m(H_m; M_m),\ z_m = \text{L2Norm}(\tilde{z}_m), \tag{22}$$

其中， $H_m \in \mathbb{R}^{T_m\times d}$, $M_m \in \{0,1\}^{T_m}$ 为动作填充掩码，$\text{Emb}_m(\cdot)$为动作令牌嵌入层， E^m_{pos} 为动作位置嵌入， $\mathcal{E}_m$为动作 Transformer 编码器，以及 $\tilde{z}_m$， $z_m \in \mathbb{R}^{d_o}$。

兼容性评分。令 $\phi(\cdot,\ \cdot)$ 表示嵌入空间中的兼容性函数。给定条件嵌入 $z \in \mathbb{R}^{d_o}$ 和动作嵌入$z_m \in \mathbb{R}^{d_o}$，我们定义

$$\phi(z, z_m) = \alpha\, z^\top z_m, \qquad \alpha = \exp(\tau), \tag{23}$$

其中，τ 是可学习温度参数，$\alpha > 0$ 是相应的缩放因子。由于所有分数空间嵌入都经过 ℓ_2归一化，式（23）是缩放后的余弦相似度。

融合兼容性分数定义为

$$s_\psi(C^s, x_m^l) = \phi(z_f, z_m). \tag{24}$$

此　外，我们还计算模态特定兼容性分数　　s

$$s_\psi^{(k)}(C^s, x_m^l) = \phi(z_k, z_m), \qquad k \in \{t, a, e\}, \tag{25}$$

使评判器能够根据部分说话者话语为候选动作打分。

组级对比训练。对于每个说话者话语 C_i^s，我们构建一个候选集

$$\mathcal{U}_i = \mathcal{G}(C_i^s) \cup \mathcal{S}(C_i^s) \cup \mathcal{N}(C_i^s), \tag{26}$$

其中，$\mathcal{G}(C_i^s)$、$\mathcal{S}(C_i^s)$ 和 $\mathcal{N}(C_i^s)$ 分别表示金级、银级和负级听者动作集。训练期间，我们从每个层级中随机抽取少量候选项，并对它们进行联合编码。

为增强模型对不完整条件的鲁棒性，我们在训练期间随机改变活跃模态集 o 。这使评判器在不同条件模式下都能保持可靠，包括纯文本和纯音频等单模态设置。

令 $\mathcal{P}_i \subseteq \mathcal{U}_i$ 表示与 C_i^s 关联的正样本集；在我们的默认设置中，$\mathcal{P}_i = \mathcal{G}(C_i^s)$。给定条件嵌入 z_i（可以是融合嵌入 z_f，也可以是活跃的特定模态嵌入 z_t，z_a，z_e），我们优化以下组级 InfoNCE 目标：

$$\mathcal{L}_{\mathrm{con}}(z) = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log \frac{\sum\limits_{x \in \mathcal{P}_i} \exp\big(\phi(z_i, z_m(x))\big)}{\sum\limits_{x \in \mathcal{U}_i} \exp\big(\phi(z_i, z_m(x))\big) + \sum\limits_{b \in \mathcal{B}_{\mathrm{bank}}} \exp\big(\beta\, \phi(z_i, z_m(b))\big)}, \tag{27}$$

其中，$\mathcal{B}$ 是小批次，$\mathcal{B}_{\mathrm{bank}}$ 是一个辅助动作库，用于提供额外的通用负样本；$z_m(x)$ 表示候选项 x 的动作嵌入，$z_m(b)$ 表示从动作库中采样的动作的嵌入，β 控制动作库负样本的贡献。动作库可以防止评判器对通用或模板式动作赋予过高的兼容性分数。

我们始终将公式（27）应用于融合嵌入 z_f。对于各模态的辅助损失，我们仅将其应用于以下模态：

当前模式 o:

$$\mathcal{L}_{\text{judge}} = \lambda_{\text{f}}\,\mathcal{L}_{\text{con}}(z_f) + \sum_{k\in o} \lambda_k\,\mathcal{L}_{\text{con}}(z_k), \qquad k \in \{t, a, e\}, \tag{28}$$

其中 λ_{f}、λ_t、λ_a、λ_e 为用于平衡不同损失项的损失权重。

多模态评判网络验证。由于评判网络是我们评估协议的核心，我们进一步验证其排名是否符合标注的等级排序 $\mathcal{G} \succ \mathcal{S} \succ \mathcal{N}$。对于任意等级 $\mathcal{A} \in \{\mathcal{G}, \mathcal{S}, \mathcal{N}\}$，我们定义其在条件 C^s 下的平均评判分数为

$$\bar{s}_{\mathcal{A}}(C^s) = \frac{1}{|\mathcal{A}(C^s)|} \sum_{x\in\mathcal{A}(C^s)} s_\psi(C^s, x). \tag{29}$$

随后，我们报告 **Win(G>S)**、**Win(G>N)** 和 **Win(S>N)**，其定义为

$$\text{Win}(\mathcal{A} > \mathcal{B}) = \frac{1}{|\mathcal{D}|} \sum_{C^s\in\mathcal{D}} \kappa\big(\bar{s}_{\mathcal{A}}(C^s), \bar{s}_{\mathcal{B}}(C^s)\big), \tag{30}$$

其中 $(\mathcal{A}, \mathcal{B}) \in \{(\mathcal{G}, \mathcal{S}), (\mathcal{G}, \mathcal{N}), (\mathcal{S}, \mathcal{N})\}$，$\mathcal{D}$ 表示说话者话语的评估集，并且

$$\kappa(u, v) = \begin{cases} 1, & u > v, \\ 0.5, & u = v, \\ 0, & u < v. \end{cases} \tag{31}$$

我们进一步报告 **MRR(G)**，其定义为

$$\text{MRR}(\mathcal{G}) = \frac{1}{|\mathcal{D}|} \sum_{C^s\in\mathcal{D}} \frac{1}{\min_{x\in\mathcal{G}(C^s)} \text{rank}_{C^s}(x)}, \tag{32}$$

其中，$\mathcal{U}(C^s)$ 中的所有候选项均按 $s_\psi(C^s,\ x)$ 降序排列，$\text{rank}_{C^s}(x)$ 表示候选项 x 的最终排名（从 1 开始）。

最后，我们报告 **nDCG@3**、**nDCG@5** 和 **nDCG@10**，使用分级相关性标签 2、1 和 0，分别对应金级、银级和负类候选项。这些指标用于验证学习得到的评判器所生成的排名是否与标注的恰当性结构一致。

Strict-L2 缺失模态注入。对于部分条件评估，我们采用*Strict-L2* 缺失模态注入协议。给定一个活跃模态集$o \subseteq \{t, a, e\}$，每个不可用模态在由其编码器分支处理之前 都会被替换为空输入。这不同于较弱的掩蔽策略：该策略仅在融合阶段移除某个模态，但仍允许其编码器读取原始输入。

形式上，令 $\delta_t(o)$、$\delta_a(o)$ 和 $\delta_e(o)$ 分别表示在模式 o 下文本、音频和情绪是否处于激活状态。对于文本，如果 $\delta_t(o) = 0$，我们会将转录文本替换为全填充序列，并将其填充掩码设为零:

$$x_t^s \leftarrow \texttt{PAD}, \qquad M_t(j) = 0,\ \forall j. \tag{33}$$

对于音频，如果 $\delta_a(o) = 0$，我们会将所有编解码器标记替换为音频填充索引，并将所有时间步标记为已填充：

$$x_a^s \leftarrow \texttt{PAD}_a, \qquad M_a(j) = 0,\ \forall j. \tag{34}$$

对于情绪，如果 $\delta_e(o) = 0$，我们会用专门的未知符号替换原始标签：

$$e^s \leftarrow \texttt{<unk>}. \tag{35}$$

在融合阶段，还会通过 M_f 屏蔽相应的模态标记。

因此，不可用的模态不会向最终条件表征提供任何语义信息。该协议严格检验评判器能否仅凭实际可用的说话者信号，可靠地评估听者动作。除非另有说明，所有部分条件可靠性实验均在此 Strict-L2 协议下进行。

B.2 评判网络的实现细节

表7： 多模态评判网络的超参数。

参数	值
骨干编码器	T5-base
隐藏维度 d	768
嵌入维度	512
Transformer 注意力头	12
Transformer 层	6
前馈维度	3072
随机失活	0.1
温度	0.07
记忆库大小	4096
优化器	AdamW
学习率	5×10^{-5}
权重衰减	0.01
批量大小	16
训练轮次	50
λ_{f}	1.0
λ_{t}	0.5
λ_{a}	0.5
λ_{e}	0.2

多模态评判网络采用基于Transformer的架构实现，用于评估说话者话语与候选听者动作之间的兼容性。文本模态由预训练的T5-base编码器进行编码，而音频标记、情绪标签和动作令牌则

表8： 我们在验证集和测试集上，针对六种输入模式（文本 T、音频 A、情绪 E及其融合）评估了多模态匹配评判器。我们报告基于平均分比较的成对胜率（胜率(G>N)、胜率(G>S)、胜率(S>N)）和排序指标（平均倒数排名(G)、归一化折损累计增益@K，采用分级相关性 G>S>N），其中 G=$\mathcal{G}$ （金级）、S=$\mathcal{S}$ （银级）和 N=$\mathcal{N}$ （负级）。

Mode	划分	胜率(G>N) ↑	胜率(G>S) ↑	胜率(S>N) ↑	平均倒数排名(G) ↑	归一化折损累计增益@3 ↑	归一化折损累计增益@5 ↑	归一化折损累计增益@10 ↑
Val								
T	Val	0.990	0.873	0.985	0.839	0.878	0.891	0.939
A	Val	0.990	0.873	0.985	**0.842**	**0.881**	**0.893**	**0.940**
T+A	Val	0.993	**0.883**	**0.988**	0.840	0.875	0.890	0.937
T+E	Val	**0.994**	0.881	**0.988**	0.841	0.875	0.891	0.938
A+E	Val	0.990	0.875	0.985	0.840	0.878	0.892	0.939
T+A+E	验证集	0.993	0.882	**0.988**	0.840	0.876	0.890	0.937
Test								
T	Test	0.992	0.873	**0.983**	0.829	0.864	**0.878**	0.932
A	Test	0.992	0.872	**0.983**	**0.832**	**0.866**	**0.878**	**0.933**
T+A	Test	**0.993**	**0.879**	0.982	0.820	0.855	0.875	0.928
T+E	Test	**0.993**	0.876	0.982	0.826	0.857	0.876	0.929
A+E	Test	0.992	0.874	**0.983**	0.831	0.865	**0.878**	**0.933**
T+A+E	测试集	**0.993**	0.878	0.982	0.828	0.859	**0.878**	0.930

经嵌入处理后，通过 Transformer 编码器进行处理，以获得各模态的表示。这些表示被投影到共享嵌入空间中，并在其中计算最终的兼容性分数。

表7总结了训练评判网络时使用的关键超参数。模型采用 768 的隐藏维度，并将表示投影到 512 维的嵌入空间中。Transformer 编码器使用 12 个注意力头和 6 层，前馈维度为 3072。训练采用 AdamW 优化器，学习率为 5×10^{-5}，权重衰减为 0.01，批量大小为 16。使用大小为 4096 的记忆库，为对比训练提供额外的负样本。

B.3 基线方法

真实值。 我们使用测试集中的真实听者动作序列作为上界参考。

随机选择。 我们从 HumanML3D [19] 中随机采样一个动作序列，作为朴素基线。

检索。 遵循标准的文本—动作匹配协议 [19, 82], ，我们将说话者转录文本与候选动作进行匹配，并从训练集中返回最近邻中的 top-1，以此检索听者动作。具体而言，我们使用 [19], 中的预训练文本编码器和动作编码器；这些编码器采用对比目标进行训练，使匹配的文本—动作对在共享嵌入空间中彼此接近，而不匹配的配对则以一定间隔分开。文本编码器将输入转录文本映射为语义特征向量，而动作编码器先将姿态序列转换为动作片段代码，再将其映射为动作特征向量。实际使用时，文本编码器遵循 [19], 中的架构，动作编码器则采用隐藏层大小为 1,024 的双向GRU实现。

级联大语言模型→文本到动作。我们构建级联基线：首先提示大语言模型，根据说话者转录文本和情绪生成听者反应动作的描述文本。随后，我们将生成的描述文本输入文本到动作（T2M）模型，以合成最终动作。这里，我们选用两个大语言模型——Qwen3-30B-A3B 和微调后的 Qwen3-4B-Thinking，以及两个有代表性的文本到动作生成器：T2M-GPT 和 MG-MotionLLM。

因此，**大语言模型→文本到动作-GPT** 表示使用 Qwen3-30B-A3B 和 T2M-GPT 的级联方案，而 **大语言模型→文本到动作-GPT*** 则表示将微调后的 Qwen3-4B-Thinking 与 T2M-GPT 配合使用。同理，**大语言模型→MG-MotionLLM**表示使用 Qwen3-30B-A3B 和 MG-MotionLLM 的级联方案，而**大语言模型→MG-MotionLLM*** 则表示将微调后的 Qwen3-4B-Thinking 与 MG-MotionLLM 配合使用。

为使主表简洁，我们在$T+E$ 设置下报告级联基线。

B.4 评估指标

我们从三个互补的角度评估模型性能：(i) 反应恰当性、(ii)动作质量和 (iii)多样性。

反应恰当性。反应恰当性衡量生成的听者动作对说话者话语的响应效果。对于每条说话者话语 C^s，标注的听者动作被划分为三个相关性等级：金级 $\mathcal{G}(C^s)$、银级 $\mathcal{S}(C^s)$和负级 $\mathcal{N}(C^s)$。令

$$\widehat{\mathcal{R}}^l(C^s) = \{\hat{x}^l_{m,1}, \ldots, \hat{x}^l_{m,M}\} \tag{36}$$

表示同一条件下生成的听者动作序列集合。为评估相对恰当性，我们使用 B.1 节介绍的多模态评判网络，该网络会为候选听者动作分配兼容性分数

$$s_\psi(C^s, x^l_m) \tag{37}$$

to a 以说话者输入 C^s 为条件的候选听者动作 x^l_m。

对于任意候选集 $\mathcal{A}$（C^s），我们将其平均评判分数定义为

$$\bar{s}_{\mathcal{A}}(C^s) = \frac{1}{|\mathcal{A}(C^s)|} \sum_{x^l_m \in \mathcal{A}(C^s)} s_\psi(C^s, x^l_m). \tag{38}$$

为简洁起见，我们将生成集合和三个标注层级的平均分数记为

$$g(C^s) = \bar{s}_{\widehat{\mathcal{R}}^l}(C^s),\ G(C^s) = \bar{s}_{\mathcal{G}}(C^s),\ S(C^s) = \bar{s}_{\mathcal{S}}(C^s),\ N(C^s) = \bar{s}_{\mathcal{N}}(C^s). \tag{39}$$

随后，我们报告 **胜率(g>G)**、**胜率(g>S)** 和 **胜率(g>N)**，其定义如下：

$$\mathrm{Win}(g > \mathcal{A}) = \frac{1}{|\mathcal{D}|} \sum_{C^s \in \mathcal{D}} \kappa\big(g(C^s), \bar{s}_{\mathcal{A}}(C^s)\big), \qquad \mathcal{A} \in \{\mathcal{G}, \mathcal{S}, \mathcal{N}\}, \tag{40}$$

其中， $\mathcal{D}$ 表示评估集，并且

$$\kappa(u,v)=\begin{cases}1, & u>v,\\ 0.5, & u=v,\\ 0, & u<v.\end{cases} \tag{41}$$

直观而言，**Win(g>N)**衡量生成的动作是否优于明显不恰当的反应，**Win(g>S)** 是针对中等恰当反应的更严格标准，而**Win(g>G)** 则是最具挑战性的标准，用于与高度恰当的标注反应进行比较。值越高，表示反应恰当性越强。

我们还报告**Gen@3**，该指标衡量在相同说话者话语下，至少有一个生成动作是否在所有候选项中排名前 3。对于每个C^s，我们构建候选池

$$\mathcal{C}(C^s)=\mathcal{G}(C^s)\cup\mathcal{S}(C^s)\cup\mathcal{N}(C^s)\cup\widehat{\mathcal{R}}^l(C^s), \tag{42}$$

按 $s_\psi(C^s,\cdot)$ 降序排列 $\mathcal{C}(C^s)$ 中的所有候选项，并将候选项 x_m^l 的排名记为 $\mathrm{rank}_{C^s}(x_m^l)$。随后我们计算

$$\mathrm{Gen@3}=\frac{1}{|\mathcal{D}|}\sum_{C^s\in\mathcal{D}}\mathbb{I}\left[\min_{\hat{x}_m^l\in\widehat{\mathcal{R}}^l(C^s)}\mathrm{rank}_{C^s}(\hat{x}_m^l)\le 3\right]. \tag{43}$$

该指标尤其适用于我们的任务，因为反应式倾听者行为本质上是一对多的：同一说话者话语可能对应多个合理的听者反应，而**Gen@3** 评估模型能否在有限的候选预算内生成至少一个极具竞争力的反应。

动作质量。我们在动作特征空间中使用弗雷歇 Inception 距离 (FID) [23] 评估动作质量。设 $f_{\mathrm{eval}}(x_m^l)$ 表示由预训练动作评估网络提取的动作序列特征表示。我们计算测试集中生成的动作和真实动作的特征统计量，然后测量两个高斯分布之间的弗雷歇距离：

$$\mathrm{FID}=\|\mu_r-\mu_g\|_2^2+\mathrm{Tr}\Big(\Sigma_r+\Sigma_g-2(\Sigma_r\Sigma_g)^{1/2}\Big), \tag{44}$$

其中，$(\mu_r,\ \Sigma_r)$ 和 $(\mu_g,\ \Sigma_g)$ 分别是真实动作特征和生成动作特征的均值与协方差。FID 越低，说明生成的动作越接近真实听者动作的分布，因此反映出整体动作质量越好。

多样性。 由于单个说话者话语可能对应多种合理的听者反应，因此评估生成的动作的多样性也很重要。遵循先前的人体动作生成研究 [82, 96], ，我们在相同的动作特征空间中衡量多样性。给定所有生成的动作的集合，我们

随机抽取两个大小相等的子集 S_d，分别记作 $\{\hat{x}^{l}_{m,\,1},\ \ldots\ldots,\ \hat{x}^{l}_{m,\,S_d}\}$ 和 $\{\hat{x}^{l\prime}_{m,\,1},\ \ldots\ldots,\ \hat{x}^{l\prime}_{m,\,S_d}\}$，并将多样性定义为

$$\text{Diversity} = \frac{1}{S_d}\sum_{i=1}^{S_d}\left\|f_{\text{eval}}(\hat{x}^{l}_{m,i}) - f_{\text{eval}}(\hat{x}^{l\prime}_{m,i})\right\|_2 . \tag{45}$$

多样性越高，表示生成的动作变化越丰富，也越不容易坍缩为少数重复性动作模式。

表 9： 组级偏好训练的完整超参数扫描结果。我们改变排序间隔 m、排序损失权重 λ_{rank} 和正例与负例权重 λ_{gn}。我们报告成对偏好指标（胜率 (g> 负级)、胜率 (g>S)、胜率 (g>金级)），以及 Gen@3、FID 和多样性。

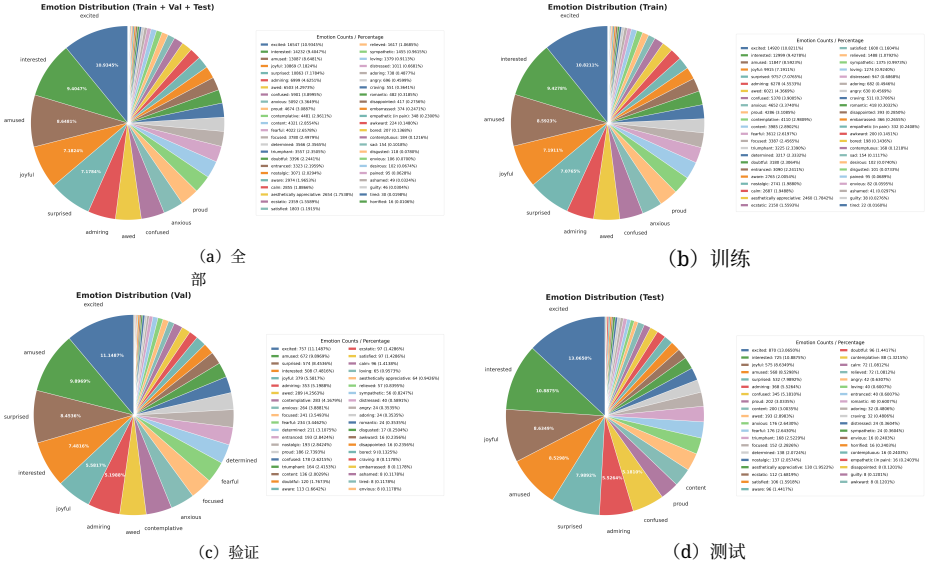
m	λ_{rank}	λ_{gn}	胜率(g>负级) ↑	胜率(g>S) ↑	胜率(g>金级) ↑	Gen@3 ↑	FID ↓	多样性 ↑
0.00	**0.00**	**0.00**	0.9976	0.7809	0.2585	0.9600	5.2638	5.3005
0.00	**0.00**	**0.25**	0.9976	0.7809	0.2633	0.9600	5.2638	5.3005
0.00	0.00	0.50	0.9976	0.7809	0.2615	0.9600	5.2638	5.3005
0.00	0.00	1.00	0.9964	0.7809	0.2615	0.9613	5.2638	5.3005
0.00	**0.25**	**0.00**	0.9988	0.7809	0.2331	0.9467	5.9644	4.6993
0.00	0.25	0.25	0.9952	0.7482	0.2240	0.9467	5.2102	4.8197
0.00	0.25	0.50	0.9988	0.7288	0.2137	0.9455	5.3426	4.9865
0.00	0.25	1.00	0.9939	0.7815	0.2458	0.9528	5.3948	4.7384
0.00	**0.50**	**0.00**	0.9927	0.7760	0.2548	0.9443	4.6552	4.7315
0.00	**0.50**	**0.25**	0.9952	0.7730	0.2482	0.9600	5.4479	4.4127
0.00	0.50	0.50	0.9952	0.7476	0.2379	0.9600	5.9814	4.3124
0.00	0.50	1.00	0.9939	0.7694	0.2512	0.9576	5.3426	4.5137
0.00	1.00	0.00	0.9964	0.7548	0.2312	0.9443	6.5379	3.9613
0.00	1.00	0.25	0.9891	0.7306	0.2391	0.9479	7.0065	3.9543
0.00	1.00	0.50	0.9964	0.7391	0.2125	0.9540	5.5322	4.4312
0.00	1.00	1.00	0.9855	0.6731	0.1925	0.9407	6.8036	3.9632
0.50	**0.00**	**0.00**	0.9964	0.7809	0.2597	0.9600	5.2638	5.3005
0.50	0.00	0.25	0.9976	0.7809	0.2639	0.9613	5.2638	5.3005
0.50	0.00	0.50	0.9976	0.7809	0.2615	0.9600	5.2638	5.3005
0.50	0.00	1.00	0.9952	0.7809	0.2615	0.9588	5.2638	5.3005
0.50	**0.25**	**0.00**	0.9939	0.7494	0.2349	0.9407	5.0807	4.8318
0.50	**0.25**	**0.25**	1.0000	0.7966	0.2663	0.9600	4.7596	4.8039
0.50	**0.25**	**0.50**	0.9903	0.7337	0.2343	0.9407	4.8888	4.6845
0.50	0.25	1.00	0.9952	0.8184	0.2778	0.9552	5.1955	4.8183
0.50	**0.50**	**0.00**	0.9964	0.8287	0.3057	0.9625	5.8396	4.1884
0.50	0.50	0.25	0.9952	0.7579	0.2318	0.9310	5.3855	4.3443
0.50	0.50	0.50	0.9952	0.7736	0.2385	0.9625	6.2371	4.3488
0.50	0.50	1.00	0.9952	0.6762	0.1913	0.9467	6.1306	4.3766
0.50	1.00	0.00	0.9915	0.7337	0.2403	0.9492	6.7096	3.9289
0.50	1.00	0.25	0.9915	0.7082	0.2149	0.9443	5.4811	4.1878
0.50	1.00	0.50	0.9673	0.6132	0.1901	0.9334	6.9334	3.9102
0.50	1.00	1.00	0.9891	0.6168	0.1834	0.9237	6.5986	3.9541
1.00	**0.00**	**0.00**	0.9976	0.7809	0.2597	0.9600	5.2638	5.3005

m	λ_{rank}	λ_{gn}	胜率(g>N) ↑	胜率(g>S) ↑	胜率(g>G) ↑	生成@3 ↑	FID ↓	多样性 ↑
1.00	0.00	0.25	0.9976	0.7809	0.2609	0.9588	5.2638	5.3005
1.00	0.00	0.50	0.9964	0.7809	0.2609	0.9600	5.2638	5.3005
1.00	0.00	1.00	0.9976	0.7809	0.2627	0.9588	5.2638	5.3005
1.00	0.25	0.00	0.9964	0.8008	0.2851	0.9516	6.0285	4.2946
1.00	0.25	0.25	0.9939	0.7676	0.2464	0.9552	5.1537	4.6242
1.00	0.25	0.50	0.9939	0.7821	0.2682	0.9516	5.3639	4.5391
1.00	0.25	1.00	0.9988	0.8117	0.2706	0.9625	5.1943	4.6935
1.00	0.50	0.00	0.9927	0.7524	0.2288	0.9528	5.3754	4.3702
1.00	0.50	0.25	0.9952	0.7361	0.2288	0.9455	5.6698	4.2394
1.00	0.50	0.50	0.9903	0.7113	0.2010	0.9516	5.8942	4.3384
1.00	0.50	1.00	0.9915	0.6501	0.1816	0.9310	5.6888	4.2328
1.00	1.00	0.00	0.9952	0.6562	0.1973	0.9310	7.0648	3.9867
1.00	1.00	0.25	0.9849	0.5938	0.1774	0.9262	7.4283	3.8852
1.00	1.00	0.50	0.9921	0.5914	0.1798	0.9104	8.6083	3.6349
1.00	1.00	1.00	0.9831	0.5847	0.1731	0.9237	6.2941	3.9609
2.00	0.00	0.00	0.9976	0.7809	0.2567	0.9600	5.2638	5.3005
2.00	0.00	0.25	0.9976	0.7809	0.2585	0.9600	5.2638	5.3005
2.00	0.00	0.50	0.9964	0.7809	0.2579	0.9600	5.2638	5.3005
2.00	0.00	1.00	0.9976	0.7809	0.2627	0.9613	5.2638	5.3005
2.00	0.25	0.00	0.9952	0.7639	0.2512	0.9540	5.6781	4.4907
2.00	0.25	0.25	0.9891	0.7433	0.2452	0.9588	5.1178	4.7459
2.00	0.25	0.50	0.9964	0.7815	0.2603	0.9588	5.6664	4.3494
2.00	0.25	1.00	0.9939	0.7748	0.2785	0.9697	5.7083	4.1561
2.00	0.50	0.00	0.9939	0.7264	0.2228	0.9516	6.1482	4.1211
2.00	0.50	0.25	0.9964	0.6477	0.1828	0.9249	6.7075	3.8914
2.00	0.50	0.50	0.9964	0.6326	0.1901	0.9249	5.4215	4.1601
2.00	0.50	1.00	0.9909	0.6610	0.1907	0.9370	6.8355	3.7096
2.00	1.00	0.00	0.9927	0.6423	0.1998	0.9298	7.1093	3.8085
2.00	1.00	0.25	0.9715	0.6483	0.2046	0.9407	6.8560	3.7436
2.00	1.00	0.50	0.9752	0.6362	0.1907	0.9298	6.1279	3.8659
2.00	1.00	1.00	0.9655	0.5648	0.1544	0.9140	6.1125	4.0394

C ReactMotionNet 数据集的更多详情

ReactMotionNet 具备三项适用于研究倾听者反应动作生成的理想特性。首先，它提供了大规模 监督，包含超过 151K 个带标签的说话者–听者配对。其次，它通过将每段说话者话语与多个候选反应式动作关联起来，明确体现了听者行为一对多 的特性。第三，它通过金级、银级和负级标签提供分级监督 ，支持生成建模和偏好感知评估。此外，数据集按互不重叠的说话者话语进行划分，从而能够更清晰地评估模型对未见对话情境的泛化能力。

ReactMotionNet 共包含 151,328 个带标签的说话者–听者配对，涵盖 8,298 条不同的说话者话语和 2,029 种不同的听者响应动-作。平均而言，每条说话者话语对应 18.24 个候选反应



（a）全部　（b）训练

（c）验证　（d）测试

图 7： 完整数据集以及训练/验证/测试划分中的情绪分布。

动作，进一步凸显了听者反应行为固有的一对多 特性。在所有配对中，分别有 9,307、34,196 和 107,825 个被标注为金级、银级和负级，体现了候选反应恰当性的等级差异 。我们按说话者话语 划分数据集，按照 8:1:1 的比例分为训练集、验证集和测试集，确保各划分中的话语互不重叠 即，每条话语只出现在一个划分中。

该数据集涵盖 *47* 种情绪类别，包括钦佩、爱慕、审美欣赏、愉悦、愤怒、焦虑、羞愧、觉察、敬畏、尴尬、无聊、平静、困惑、沉思、轻蔑、满足、渴望、欲求、坚定、失望、厌恶、忧虑、怀疑、欣喜若狂、窘迫、感同身受（痛苦）、着迷、嫉妒、兴奋、恐惧、专注、内疚、惊骇、感兴趣、喜悦、充满爱意、怀旧、痛苦、自豪、宽慰、浪漫、悲伤、满意、惊讶、同情、疲倦和胜利。如图 *7* 所示，这些情绪标签在整个数据集及各个划分中呈现出广泛但不均衡的分布，使 *ReactMotionNet* 成为一个用于建模多样化情感对话回应的真实基准。

D 更多实验结果

D.1 超参数敏感性分析

我们研究分组偏好训练对排名边际 m、排名损失权重 λ_{rank} 以及正例与负例权重 λ_{gn} 的敏感性。我们

表10： 从完整搜索中选出的代表性超参数配置。我们重点展示 Gen@3、胜率(g>S) 和胜率(g>G)，以及 FID 和多样性。

配置	m	λ_{rank}	λ_{gn}	胜率(g>N)↑	胜率(g>S)↑	胜率(g>G)↑	Gen@3↑	FID↓	多样性↑
C1	2.00	0.25	1.00	0.9939	0.7748	0.2785	0.9697	5.7083	4.1561
C2	0.50	0.50	0.00	0.9964	0.8287	0.3057	0.9625	5.8396	4.1884
C3	0.00	0.50	0.00	0.9927	0.7760	0.2548	0.9443	4.6552	4.7315
C4	1.00	0.00	0.25	0.9976	0.7809	0.2609	0.9588	5.2638	5.3005

主要考察 **Gen@3**，该指标衡量生成的动作能否在相同候选预算下被排入最合理的候选项前列。此外，我们还报告 **Win(g>S)** 和 **Win(g>G)**，分别评估相对于中等质量和高质量参考候选项的相对偏好质量。我们还纳入 FID 和多样性，用于表征动作真实感和输出多样性。

超参数扫描揭示了若干一致的规律。首先，引入较小的正排序间隔有益，而且比不设间隔更可靠。在 $\lambda_{\text{rank}} = 0.25$ 和 $\lambda_{\text{gn}} = 0.25$ 条件下，将 m 从 0 增加到 0.5，可使 Win(g>S) 从 0.7482 提高至 0.7966，Win(g>G) 从 0.2240 提高至 0.2663，Gen@3 从 0.9467 提高至 0.9600，同时将 FID 从 5.2102 降至 4.7596。尽管在某些情况下，较大的间隔可以进一步提高 Gen@3，但这种提升并不总是伴随着偏好对齐或动作质量的改善，这表明过大的间隔可能导致目标过度专门化。

其次，λ_{rank} 是扫描中最敏感的超参数。适度的排序监督有益，而取值过大往往会损害对齐和生成质量。例如，在 $m = 0.5$ 和 $\lambda_{\text{gn}} = 0.25$ 条件下，将 λ_{rank} 从 0.25 增加到 0.5 和 0.1，会使 Win(g>S) 从 0.7966 降至 0.7579 和 0.7082，使 Win(g>G) 从 0.2663 降至 0.2318 和 0.2149，并使 FID 从 4.7596 恶化至 5.3855 和 5.4811。这表明，过大的排序压力会使优化偏向相对排序，从而牺牲生成保真度。

第三，λ_{gn} 具有次要但不可忽视的影响，取适中 值时能够实现最优权衡。在 $m = 0.5$ 和 $\lambda_{\text{rank}} = 0.25$条件下，将$\lambda_{\text{gn}} = 0.25$ 设为该值，相较于 $\lambda_{\text{gn}} = 0$，可提升 Win(g>S)、Win(g>G) 和 Gen@3，同时降低 FID。相比之下，进一步将 λ_{gn} 增至 1.0 虽然会略微提高成对偏好得分，但会降低 Gen@3 并使 FID 恶化，这表明更强的金级与负级分离并不一定能提升整体生成质量。

因此，我们在所有主要实验中均采用 $m = 0.5$、$\lambda_{\text{rank}} = 0.25$ 和 $\lambda_{\text{gn}} = 0.25$，因为这一设置处于参数扫描的稳定区域，并且在面向偏好和生成的评估标准上取得了最均衡的整体表现。

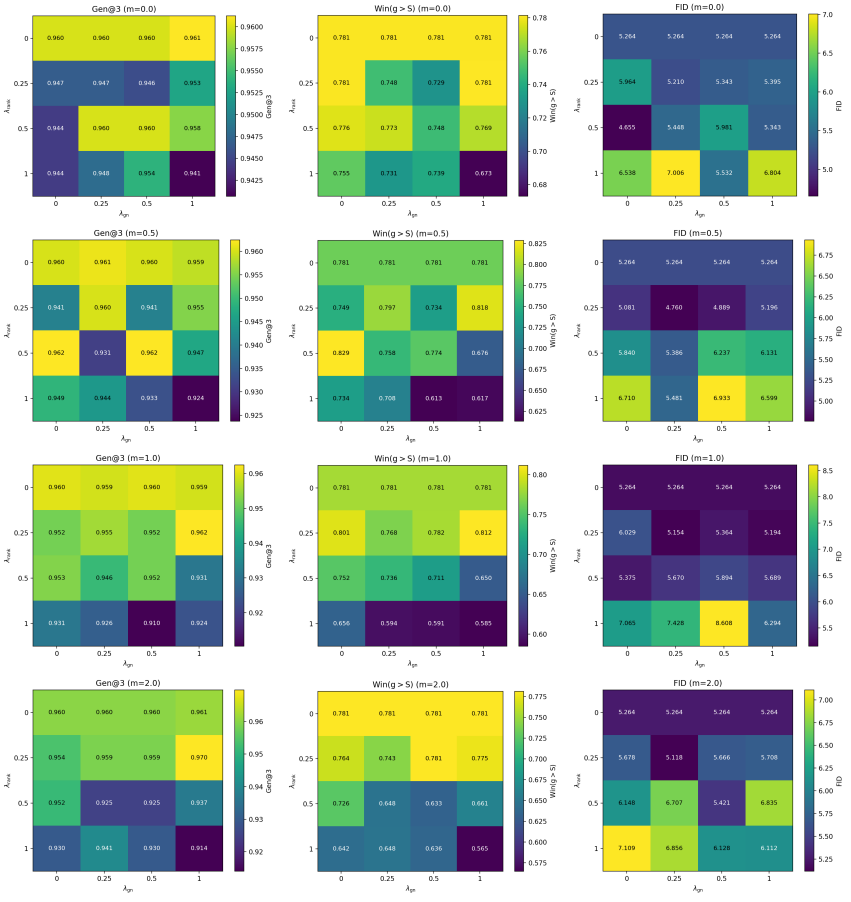


图 8： 不同排序间隔下的超参数敏感性热力图。我们展示了 Gen@3、Win(g>S) 和 FID 随 λ_{rank} 和 λ_{gn}的变化。

D.2 推理效率

表 11 列出了所提出的 ReactMotion 的推理效率。在推理过程中，ReactMotion 在单张 NVIDIA A100 80GB GPU 上运行，并以说话者的多模态输入为条件，自回归地生成听者动作令牌。在评估中，模型生成了 50 个听者反应动作，分别对应 50 段说话者话语。总计生成 1,830 个动作令牌，用时 28.8 秒，生成吞吐量达到每秒 63.6 个令牌和每秒 1.74 个动作序列，对应每个听者动作序列约 0.60 秒的平均延迟。

随后，使用 VQ-VAE 解码器将生成的动作令牌解码为关节序列。该解码器每秒可处理 39.1 个动作序列，计算开销极小。因此，完整流程

表 11：单张 NVIDIA A100（80GB）上的推理效率。

指标	数值
词元生成速度	63.6 词元/秒
动作生成速度	1.74 轮/秒
端到端生成速度	1.66 轮/秒
每个样本的平均延迟	∼0.60 秒
VQ-VAE 解码速度	39.12 转/秒

实现了每秒 1.66 个动作序列的端到端吞吐量。这些结果表明，ReactMotion 在模型容量与推理效率之间保持了良好的平衡，能够在对话场景中实现近实时的响应动作生成。

D.3 用户研究的更多细节

我们在腾讯问卷平台上开展了一项用户研究，用于评估由 **ReactMotion （我们的方法**）生成的听者动作，并将其与两个生成式基线进行比较，即 **交叉熵**变体和 **大语言模型→MG-MotionLLM** *，以及组内最佳**银级**参考结果。共有59名志愿者（16名女性和43名男性）参与了本次研究，他们都具有机器学习或深度学习相关背景，并通过在线问卷完成研究。在每次试验中，我们向参与者呈现一对基于同一说话者话语生成的倾听者动作视频（A/B），同时显示说话者的转录文本并播放相应的音频。参与者需要选择其中展现出更恰当听者反应动作的视频。为避免位置偏差，两种待比较的动作会被随机分配到A/B位置。每位参与者完成36次试验，涵盖六条说话者话语，每种条件下进行六次两两比较。对于**银级**条件，我们根据动作文本描述，在每个说话者—条件组中选取最佳候选，并渲染出动作片段。

图5中的结果揭示了三个值得注意的发现。首先，**我们的方法**始终优于两种生成式基线：相较于**交叉熵**，偏好率达到67.8%；相较于**大语言模型→MG-MotionLLM**，偏好率达到72.0%。这证明了我们的统一多模态序列到序列形式优于标准交叉熵训练和级联生成流程。其次，尽管**银级** 参考整体上仍然更出色，**我们的方法**与**银级** 的差距明显小于任一基线：在与**银级**的比较中，**我们的方法** 获得了44.1%的投票，而**交叉熵** 和**大语言模型→MG-MotionLLM**仅分别获得31.9%和31.4%。这表明，**我们的方法** 生成的动作在感知上更接近高质量的组内参考。第三，这些结果凸显了所提出的组级偏好学习目标的有效性。该目标显式建模了金级、银级和负级反应之间的排序，并在人工评估中带来了更恰当的倾听者行为。同时，**我们的方法** 与**银级**之间仍存在差距，这表明听者反应动作

生成仍然具有挑战性，在动作自然度、语境准确性和多样性方面仍有进一步提升的空间。

D.4 失败案例

虽然该模型在许多场景中都能有效生成符合语境的听者动作，但在复杂对话中捕捉更深层的对话意图仍然颇具挑战。在含义模糊或较为少见的情境中，恰当的听者行为需要更深入地理解意图，而当前模型在这些情况下的鲁棒性可能仍然有限。这表明，进一步提升双人交互中的意图觉察交互建模能力，是未来一项颇具前景的研究方向。

E 局限性

由于我们是首个探索这一任务的研究团队，因此设计了一种相对简单但有效的模型架构，以保持训练稳定性和计算效率。这种设计使我们能够在不过度增加架构复杂度的情况下验证方法的核心思想。所提出的方法已取得令人期待的成果，证明了其可行性和有效性。尽管如此，仍有很大的改进空间。未来的工作可以探索更先进的网络架构和更精细的训练技术，以进一步提升性能。

参考文献

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., 等：GPT-4 技术报告（2023） 2. Alexanderson, S., Nagy, R., Beskow, J., Henter, G.E.：聆听、去噪、行动！使用扩散模型进行音频驱动的动作合成。ACM Transactions on Graphics (TOG)**42**(4), 1–20（2023） 3. Ao, T., Gao, Q., Lou, Y., Chen, B., Liu, L.：Rhythmic Gesticulator：利用层次化神经嵌入实现节奏觉察的伴语手势合成。ACM Transac- tions on Graphics (TOG)**41**(6), 1–19（2022） 4. Barquero, G., Escalera, S., Palmero, C.：利用混合位置编码实现无缝人体动作组合。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集（2024） 5. Bishop, C.M., Nasrabadi, N.M.：模式识别与机器学习，第4卷。Springer（2006） 6. Bradley, R.A., Terry, M.E.：不完全区组设计的秩分析：I. 成对比较法。Biometrika **39**(3/4), 324–345（1952） 7. Chen, B., Li, Y., Ding, Y.X., Shao, T., Zhou, K.：在基于提示的伴随语音动作生成中实现协同全身控制。载于：ACM国际多媒体会议（ACM MM）论文集。第6774–6783页（2024）

8. Chen, C., Zhang, J., Lakshmikanth, S.K., Fang, Y., Shao, R., Wetzstein, G., Fei-Fei, L., Adeli, E.：动作的语言：统一三维人体动作的言语与非言语语言。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集。第6200–6211页（2025年6月）
9. Chen, C., Zhang, J., Lakshmikanth, S.K., Fang, Y., Shao, R., Wetzstein, G., Fei-Fei, L., Adeli, E.：动作的语言：统一三维人体动作的言语与非言语语言。载于：计算机视觉与模式识别会议论文集。第6200–6211页（2025年） 10. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.：通过潜在空间中的动作扩散执行指令。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集。第18000–18010页（2023年） 11. Chiang, W.L., Zheng, L., Sheng, Y., Angelopoulos, A.N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M., Gonzalez, J.E., 等：Chatbot Arena：一个通过人类偏好评估大语言模型的开放平台。载于：国际机器学习会议（ICML）（2024年） 12. Chopin, B., Tang, H., Otberdout, N., Daoudi, M., Sebe, N.：用于人体反应生成的交互变换器。IEEE多媒体汇刊 (TMM)**25**, 8842–8854（2023年） 13. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.：基于人类偏好的深度强化学习。神经信息处理系统进展（NeurIPS）**30**（2017年） 14. Chu, X., Liu, R., Huang, Y., Liu, Y., Peng, Y., Zheng, B.：Unils：用于统一聆听与说话的端到端音频驱动虚拟形象。arXiv预印本 arXiv:2512.09327（2025年） 15. Défossez, A., 等：Moshi：用于实时对话的语音文本基础模型。arXiv预印本 arXiv:2410.00037（2024年） 16. Dubois, Y., Galambosi, B., Liang, P., Hashimoto, T.B.：长度控制版AlpacaEval：一种消除自动评估器偏差的简单方法。arXiv预印本 arXiv:2404.04475（2024年） 17. Ghosh, A., Dabral, R., Golyanik, V., Theobalt, C., Slusallek, P.：ReMoS：用于双人交互的3D动作条件反应合成。载于：欧洲计算机视觉会议（ECCV）。第418–437页（2024年） 18. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.：Momask：3D人体动作的生成式掩码建模。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集。第1900–1910页（2024年） 19. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.：从文本生成多样且自然的三维人体动作。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集。第5152–5161页（2022年） 20. Guo, W., Bie, X., Alameda-Pineda, X., Moreno-Noguer, F.：多人极端动作预测。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集。第13053–13064页（2022年） 21. Han, B., Peng, H., Dong, M., Ren, Y., Shen, Y., Xu, C.：AMD：自回归动作扩散。载于：Wooldridge, M.J., Dy, J.G., Natarajan, S.（编），人工智能促进协会（AAAI）。第2022–2030页（2024年） 22. He, X., Huang, Q., Zhang, Z., Lin, Z., Wu, Z., Yang, S., Li, M., Chen, Z., Xu, S., Wu, X.：通过动作解耦扩散模型生成伴语手势视频。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集。第2263–2273页（2024年）

23. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.：采用双时间尺度更新规则训练的生成对抗网络会收敛到局部纳什均衡。《神经信息处理系统进展》（NeurIPS）**30**（2017年）24. Ho, L., Huang, Y., Qin, D., Shi, M., Tse, W., Liu, W., Yamagishi, J., Komura, T.：Interact：日常场景中两人动态、富有表现力且具有交互性的活动大型数据集。《ACM计算机图形与交互技术论文集》（PACMCGIT）**8**（4），1–27（2025年）25. Hu, T., Zhu, X., Guo, W., Su, K.：通过正向动作表示实现高效交互识别。《工程中的数学问题》**2013年**（1），795360（2013年）26. Huang, Y., Wan, W., Yang, Y., Callison-Burch, C., Yatskar, M., Liu, L.：Como：通过语言引导的姿势代码编辑实现可控动作生成。载于：欧洲计算机视觉会议（ECCV），第180–196页（2024年）27. Huang, Y., Khan, S.M.：Dyadgan：生成双人交互中的面部表情。载于：IEEE计算机视觉与模式识别会议研讨会（CVPRW）（2017年）28. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., 等：GPT-4o系统卡（2024年）29. Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., 等：OpenAI o1系统卡（2024年）30. Jeong, M., Hwang, Y., Lee, J., Jung, S., Kim, W.H.：Hgm^3：采用困难词元挖掘的分层生成式掩码动作建模。载于：国际学习表征会议（ICLR）（2025年）31. Khirodkar, R., Bansal, A., Ma, L., Newcombe, R., Vo, M., Kitani, K.：Ego-humans：以自我为中心的三维多人基准。载于：IEEE/CVF国际计算机视觉会议（ICCV），第19807–19819页（2023年）32. Khirodkar, R., Song, J.T., Cao, J., Luo, Z., Kitani, K.：Harmony4d：真实场景近距离人类交互视频数据集。《神经信息处理系统进展》（NeurIPS）**37**，107270–107285（2024年）33. Kim, D.Y., Lee, H.K., Chung, K.：元宇宙中的虚拟形象中介体验：虚拟形象真实感对用户-虚拟形象关系的影响。《零售与消费者服务杂志》**73**，103382（2023年）34. Kim, J., Kim, J., Choi, S.：FLAME：自由形式的基于语言的动作合成与编辑。载于：人工智能促进协会（AAAI），第37卷，第8255–8263页（2023年）35. Ko, W.R., Jang, M., Lee, J., Kim, J.：Air-act2act：用于教授机器人非言语社会行为的人际交互数据集。《国际机器人研究杂志》**40**（4-5），691–697（2021年）36. Lee, G., Deng, Z., Ma, S., Shiratori, T., Srinivasa, S.S., Sheikh, Y.：用手交谈16.2 m：用于对话动作分析与合成的同步身体-手指动作和音频大型数据集。载于：IEEE/CVF国际计算机视觉会议（ICCV），第763–772页（2019年）37. Li, B., Zhao, Y., Zhelun, S., Sheng, L.：Danceformer：基于音乐条件的三维舞蹈生成与参数化动作Transformer。载于：人工智能促进协会（AAAI），第36卷，第1272–1279页（2022年）38. Li, J., Kang, D., Pei, W., Zhe, X., Zhang, Y., Bao, L., He, Z.：Audio2gestures：从音频生成多样化手势。《IEEE可视化与计算机图形学汇刊》（TVCG）**30**（8），4752–4766（2023年）

39. Li, R., Dai, Y., Zhang, Y., Li, J., Yang, J., Guo, J., Li, X.：探索音乐驱动舞蹈生成中的多模态控制。载于：IEEE国际声学、语音与信号处理会议（ICASSP），第8281–8285页（2024年）40. Li, R., Zhang, H., Zhang, Y., Zhang, Y., Zhang, Y., Guo, J., Zhang, Y., Li, X., Liu, Y.：Lodge++：凭借稳健的编舞模式实现高质量、长时舞蹈生成。《IEEE模式分析与机器智能汇刊》（TPAMI），第1–15页（2025年）41. Liang, H., Zhang, W., Li, W., Yu, J., Xu, L.：Intergen：复杂交互下基于扩散的多人动作生成。arXiv预印本，arXiv:2304.05684（2023年）42. Liao, T.H., Zhou, Y., Shen, Y., Huang, C.H.P., Mitra, S., Huang, J.B., Bhattacharya, U.：塑造我的动作：文本驱动的形状感知人体动作合成。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR），第1917–1928页（2025年）43. Liu, H., Zhu, Z., Becherini, G., Peng, Y., Su, M., Zhou, Y., Zhe, X., Iwamoto, N., Zheng, B., Black, M.J.：Emage：通过表现力掩码音频手势建模实现统一的整体式语音伴随手势生成。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR），第1144–1154页（2024年）44. Liu, P., Song, L., Huang, J., Liu, H., Xu, C.：Gesturelsm：基于潜在捷径并采用时空建模的语音伴随手势生成。载于：IEEE/CVF国际计算机视觉会议（ICCV），第10929–10939页（2025年）45. Liu, Y., Cao, Q., Wen, Y., Jiang, H., Ding, C.：迈向可变且协调的整体式语音伴随动作生成。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR），第1566–1576页（2024年）46. Liu, Y., Chen, C., Ding, C., Yi, L.：Physreaction：通过前向动力学引导的四维模仿实现物理合理的实时类人反应合成。载于：ACM国际多媒体会议（ACM MM），第3771–3780页（2024年）47. Liu, Y., Chen, C., Yi, L.：交互式类人机器人：通过社会可供性规范化与预测实现在线全身动作反应合成（2023年）48. Lu, S., Wang, J., Lu, Z., Chen, L.H., Dai, W., Dong, J., Dou, Z., Dai, B., Zhang, R.：Scamo：探索自回归动作生成模型中的缩放定律。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR），第27872–27882页（2025年）49. Luo, C., Song, S., Yan, S., Yu, Z., Ge, Z.：Reactdiff：基础型多样且适当的面部反应扩散模型。载于：ACM国际多媒体会议（ACM MM），第5607–5616页（2025年）50. Luo, C., Wang, J., Li, B., Song, S., Ghanem, B.：Omniresponse：双人交互中的在线多模态对话响应生成。载于：《神经信息处理系统进展》（NeurIPS）（2025年）51. Luo, C., 等：Reactface：双人交互中的在线多样且适当的面部反应生成。arXiv预印本（2024年），arXiv:2305.15748 52. Meng, Z., Xie, Y., Peng, X., Han, Z., Jiang, H.：重新思考文本驱动人体动作生成中的扩散：冗余表示、评估与掩码自回归。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR），第27859–27871页（2025年）

53. Mughal, M.H., Dabral, R., Habibie, I., Donatelli, L., Habermann, M., T heobalt, C.：Convofusion：用于合成伴随语音手势的多模态对话扩散。在：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集。第1388–1398页（2024年）54. Murray, K., Chiang, D.：纠正神经机器翻译中的长度偏差。在：机器翻译会议（WMT）论文集。第212–223页（2018年）55. Ng, E., Romero, J., Bagautdinov, T., Bai, S., Darrell, T., Kanazawa, A., Richard, A.：从音频到照片级真实感化身：合成对话中的人物。在：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集。第1001–1010页（2024年）56. Ng, E., Xiang, D., Joo, H., Grauman, K.：You2me：通过第一人称和第二人称交互，从自我中心视频中推断身体姿态。在：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集。第9890–9900页（2020年）57. Ng, E., 等：学习倾听：对非确定性双人面部运动建模。在：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集（2022年）58. OpenAI：推出OpenAI o3和o4-mini。https://openai.com/index/ openai-o3-mini/（2025年）59. Park, S., Kim, C., Rha, H., Kim, M., Hong, J., Yeo, J., Ro, Y.：来聊点真实的：面向面对面对话的口语对话模型。在：计算语言学协会年会（ACL）论文集。第16334–16348页（2024年）60. Petrovich, M., Black, M.J., Varol, G.：利用Transformer变分自编码器进行动作条件化的3D人体动作合成。在：IEEE/CVF国际计算机视觉会议（ICCV）论文集（2021年）61. Petrovich, M., Black, M.J., Varol, G.：TEMOS：根据文本描述生成多样化人体动作。在：欧洲计算机视觉会议（ECCV）。第480–497页（2022年）62. Petrovich, M., Litany, O., Iqbal, U., Black, M.J., Varol, G., Bin Peng, X., Rempe, D.：用于文本驱动三维人体动作生成的多轨时间线控制。在：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集。第1911–1921页（2024年）63. Pinyoanuntapong, E., Saleem, M.U., Karunratanakul, K., Wang, P., Xue, H., Chen, C., Guo, C., Cao, J., Ren, J., Tulyakov, S.：ControlMM：可控掩码运动生成（2024年）64. Raab, S., Leibovitch, I., Li, P., Aberman, K., Sorkine-Hornung, O., Cohen-Or, D.：MODI：从多样化数据中进行无条件运动合成。在：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集。第13873–13883页（2023年）65. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.：探索利用统一文本到文本Transformer进行迁移学习的极限。《机器学习研究期刊》（JMLR）**21**(140)，1–67（2020年）66. Rubenstein, P.K., 等：AudioPaLM：一种能够说话和倾听的大语言模型。arXiv预印本 arXiv:2306.12925（2023年）67. Ryoo, M.S., Fuchs, T.J., Xia, L., Aggarwal, J.K., Matthies, L.：基于第一人称视频的机器人中心活动预测：它们会对我做什么？在：第十届ACM/IEEE人机交互国际会议（HRI）论文集。第295–302页（2015年）

68. Ryoo, M.S., Matthies, L.：第一人称活动识别：他们在做什么
对我做什么？在：IEEE/CVF国际计算机视觉会议（ICCV）论文集。第2730–2737页（2013年）69. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., 等：OpenAI GPT-5系统卡（2025年）70. Song, S., 等：REACT 2024：第二届多种恰当面部反应生成挑战赛。arXiv预印本（2024 05166 71 2401），arXiv: . . Spaccatini, F., Corlito, G., Sacchi, S.：新的双人关系？社交机器人的拟人化对人类共情的影响。《人类行为中的计算机》**146**，107821（2023年）72. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., Christiano, P.F.：利用人类反馈进行摘要学习。《神经信息处理系统进展》（NeurIPS）**33**，3008–3021（2020年）73. Sun, M., Xu, C., Jiang, X., Liu, Y., Sun, B., Huang, R.：超越对话——生成用于交流的整体三维人类双人动作。《国际计算机视觉期刊》**133**(5)，2910–2926（2025年）74. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.：MotionCLIP：将人体动作生成引入CLIP空间。在：欧洲计算机视觉会议（ECCV）。第358–374页（2022年）75. Tevet, G., Raab, S., Cohan, S., Reda, D., Luo, Z., Peng, X.B., Bermano, A.H., van de Panne, M.：CLOSD：通过多任务角色控制闭合仿真与扩散之间的循环。在：国际学习表征会议（ICLR）（2025年）76. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.：人体动作扩散模型。arXiv预印本 arXiv:2209.14916（2022年），亦称MDM；广泛用作扩散基线。77. Veluri, B., Peloquin, B.N., Yu, B., Gong, H., Gollakota, S.：超越轮流式接口：作为全双工对话智能体的同步大语言模型。在：自然语言处理实证方法会议（EMNLP）论文集。第21390–21402页（2024年）78. Wang, T., Wu, Z., He, Q., Chu, J., Qian, L., Cheng, Y., Xing, J., Zhao, J., Jin, L.：StickMotion：通过绘制简笔小人生成三维人体动作。在：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集。第12370–12379页（2025年）79. Wang, Y., Leng, Z., Li, F.W., Wu, S.C., Liang, X.：FG-T2M：通过扩散模型进行细粒度文本驱动的人体动作生成。在：IEEE/CVF国际计算机视觉会议（ICCV）论文集。第22035–22044页（2023年）80. Wang, Y., Li, M., Liu, J., Leng, Z., Li, F.W., Zhang, Z., Liang, X.：FG-T2M++：由大语言模型增强的细粒度文本驱动人体动作生成。《国际计算机视觉期刊》（IJCV）**133**(7)，4277–4293（2025年）81. Wang, Z., Wang, J., Li, Y., Lin, D., Dai, B.：InterControl：通过控制每个关节实现零样本人体交互生成。在：神经信息处理系统进展（NeurIPS）（2024年）82. Wu, B., Xie, J., Shen, K., Kong, Z., Ren, J., Bai, R., Qu, R., Shen, L.：MG-MotionLLM：一个用于跨多种粒度进行动作理解与生成的统一框架。在：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集。第27849–27858页（2025年）83. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., 等：Google神经机器翻译系统：弥合人工翻译与机器翻译之间的差距（2016年）

84. Xiao, L., Lu, S., Pi, H., Fan, K., Pan, L., Zhou, Y., Feng, Z., Zhou, X., Peng, S., Wang, J.: Motionstreamer：在因果潜在空间中通过基于扩散的自回归模型实现流式动作生成。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集。第 10086–10096 页（2025年） 85. Xu, C., Sun, M., Cheng, Z.Q., Wang, F., Liu, Y., Sun, B., Huang, R., Haupt- mann, A.: Combo：在和谐中实现语音伴随的整体三维人体动作生成与高效可定制适配。IEEE模式分析与机器智能汇刊（TPAMI）第 1–18 页（2025年） 86. Xu, L., Lv, X., Yan, Y., Jin, X., Wu, S., Xu, C., Liu, Y., Zhou, Y., Rao, F., Sheng, X., et al.: Inter-x：迈向通用的人际交互分析。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集。第 22260–22271 页（2024年） 87. Xu, L., Zhou, Y., Yan, Y., Jin, X., Zhu, W., Rao, F., Yang, X., Zeng, W.: Regen- net：迈向人体动作-反应合成。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集。第 1759–1769 页（2024年） 88. Xu, S., Dou, Z., Shi, M., Pan, L., Ho, L., Wang, J., Liu, Y., Lin, C., Ma, Y., Wang, W., et al.: Mospa：由空间音频驱动的人体动作生成（2025年） 89. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 技术报告（2025年） 90. Yang, Y., Huang, Z., Xu, C., He, S.: 用于长期动作生成的拉格朗日运动场。IEEE模式分析与机器智能汇刊（TPAMI）48(2), 1171–1184（2026年） 91. Yi, H., Liang, H., Liu, Y., Cao, Q., Wen, Y., Bolkart, T., Tao, D., Black, M.J.: 从语音生成整体三维人体动作。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集。第 469–480 页（2023年） 92. Yin, Y., Guo, C., Kaufmann, M., Zarate, J.J., Song, J., Hilliges, O.: Hi4d：近距离人体交互的4D实例分割。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集。第 17016– 17027 页（2023年） 93. Yu, C., Zhai, W., Yang, Y., Cao, Y., Zha, Z.J.: Hero：从视频生成人体反应。载于：IEEE/CVF国际计算机视觉会议（ICCV）论文集。第 10262–10274 页（2025年） 94. Zhang, D., Li, S., Zhang, X., Zhan, J., Wang, P., Zhou, Y., Qiu, X.: Speechgpt：赋予大语言模型内在的跨模态对话能力。载于：计算语言学协会论文集。第 15757– 15773 页（2023年） 95. Zhang, J., Fan, H., Yang, Y.: Energymogen：在潜在空间中通过基于能量的扩散模型进行组合式人体动作生成。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集。第 17592–17602 页（2025年）

96. Zhang, J., Zhang, Y., Cun, X., Zhang, Y., Zhao, H., Lu, H., Shen, X., Shan, Y.: 使用离散表示从文本描述生成人体动作。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集。第 14730–14740 页（2023年） 97. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse：使用扩散模型进行文本驱动的人体动作生成。IEEE模式分析与机器智能汇刊（TPAMI）**46**(6), 4115–4128（2024年） 98. Zhang, P., Liu, P., Garrido, P., Kim, H., Chaudhuri, B.: Kinmo：运动学感知的人体运动理解与生成。载于：IEEE/CVF国际计算机视觉会议（ICCV）论文集。第 11187–11197 页（2025年） 99. Zhang, X., Li, J., Zhang, J., Dang, Z., Ren, J., Bo, L., Tu, Z.: Semtalk：通过帧级语义强调实现整体语音伴随动作生成。载于：IEEE/CVF国际计算机视觉会议（ICCV）论文集。第 13761–13771 页（2025年） 100. Zhang, X., Li, J., Zhang, J., Ren, J., Bo, L., Tu, Z.: Echomask：用于整体语音伴随动作生成的语音查询注意力掩码建模。载于：ACM国际多媒体会议（ACM MM）论文集。第 10827–10836 页（2025年） 101. Zhang, Y., Huang, D., Liu, B., Tang, S., Lu, Y., Chen, L., Bai, L., Chu, Q., Yu, N., Ouyang, W.: Motiongpt：微调后的大语言模型是通用运动生成器。载于：人工智能促进协会（AAAI）。第 38 卷，第 7368–7376 页（2024年） 102. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: 使用 MT-Bench 和 Chatbot Arena 评判“以大语言模型为评判器”的方法。神经信息处理系统进展（NeurIPS）**36**, 46595– 46623（2023年） 103. Zhou, M., Bai, Y., Zhang, W., Yao, T., Zhao, T., Mei, T.: 响应式聆听头部生成：基准数据集与基线方法。载于：欧洲计算机视觉会议（ECCV）（2022年） 104. Zhu, Y., Zhang, L., Rong, Z., Hu, T., Liang, S., Ge, Z.: INFP：双人对话中的音频驱动交互式头部生成。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集（2025年） 105. Zou, Q., Yuan, S., Du, S., Wang, Y., Liu, C., Xu, Y., Chen, J., Ji, X.: Parco：部件协调的文本到动作合成。载于：Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G.（编）欧洲计算机视觉会议（ECCV）。第 15114 卷，第 126–143 页（2024年）