A^2-LLM：一种端 到端对话式音频数字人大型语言模型

Xiaolin Hu [*] [1 2] Hang Yuan [*3 2] Xinzhu Sang [1] Binbin Yan [1] Zhou Yu [3] Cong Huang [2] Kai Chen [2]

arXiv:2602.04913v1 [cs.LG] 4 Feb 2026

摘要

开发富有表现力且响应迅速的对话式数字人，是下一代人机交互的基石。尽管大语言模型（LLM）显著增强了对话能力，当前大多数系统仍依赖连接独立模块的级联架构。这些流水线常常受累积误差、高延迟和较差实时性能的困扰。由于无法获取底层对话语境，这些流水线本质上会优先保证僵硬的口型同步，而牺牲情感深度。为应对这些挑战，我们提出A^2-LLM，一种端到端对话式音频数字人大型语言模型，可在统一框架内联合推理语言、音频韵律与三维面部运动。为便于训练，我们引入FLAME-QA，这是一个高质量多模态数据集，旨在以问答格式将语义意图与富有表现力的面部动态对齐。借助深层语义理解，A^2-LLM能够生成超越简单口型同步的情感丰富的面部运动。实验结果表明，我们的系统在保持实时效率（500毫秒延迟，0.7实时因子）的同时，实现了更优的情感表现力。代码与数据集将发布于项目仓库。

1. 引言

大语言模型（LLM）已迅速从纯文本系统发展为能够处理和生成语音的多模态智能体，使更自然的人机交互（HCI）成为可能。尤其是，近期的大型音频语言模型（音频语言模型）在语音理解方面展现出强大性能，并且能够生成语音，超越传统的级联方法，并实现直接、实时的语音交互。

在现有系统中，GPT-4o（OpenAI，2024）是首个端到端音频交互式大型模型，可直接将音频输入转换为音频输出，无需依赖中间文本表示，因此能够保留副语言线索和全局对话连贯性。继 GPT-4o 之后，涌现出一系列开源大型音频语言模型（Fang 等人，2024；Zeng 等人，2024；Fang 等人，2025；Ding 等人，2025；Wu 等人，2025；Huang 等人，2025；Li 等人，2025b），其中 Step-Audio2 在情感语音理解与生成方面表现尤为出色。尽管取得了这些进展，当前的音频语言模型仍局限于纯音频交互，无法满足许多 HCI 场景的需求（Cassell，2001）。在自然的面对面交流中，语音与面部表情紧密关联，人们会同时依靠听觉和视觉线索来推断意图和情绪（McGurk & MacDonald，1976；Yehia 等人，1998）。

为通过视觉反馈补充音频交互，近期的视频生成模型（Xu 等人，2024b；a）已实现令人瞩目的照片级真实感。然而，这些模型依赖二维像素生成，缺乏底层的三维几何，因此无法部署于沉浸式虚拟现实/扩展现实和裸眼三维显示器中，因为这些场景要求具备立体的 6DoF 一致性。

因此，原生三维面部动画仍是沉浸式人机交互的标准。当前主流范式是将语音信号直接映射为 FLAME（Li 等人，2017）的参数。FLAME 是一种广泛采用的三维可变形模型（3DMM），能够将面部几何高效地解耦为形状、表情和姿态属性。尤为重要的是，将这些参数化控制与先进的神经渲染流水线（Grassal 等人，2022；Kerbl 等人，2023；Qian 等人，2024）相结合，能够兼顾高保真合成与几何一致性，满足此类沉浸式环境的要求。

为实现面对面交互，系统通常采用级联架构（自动语音识别 → 大语言模型 → 文本转语音 → 动画）（Huang 等人，2024；Zhang 等人，2023）。尽管这种流程采用模块化设计，却难免存在延迟高、累积误差以及根本性的语义—情感鸿沟。将大语言模型推理压缩为音频波形，会丢失对话语境，往往导致表情不匹配

[1]信息光子学与光通信国家重点实验室，北京邮电大学，中国北京 [2]中关村人工智能研究院，中国北京 [3]华东师范大学金融与统计学院，中国上海。通信作者：Xinzhu Sang <xzsang@bupt.edu.cn>，Kai Chen <kaichen@zgci.ac.cn>。

预印本。*2026*年*2*月*6*日。

——例如，说出“哈哈”时，嘴唇机械地同步运动，而上面部却僵硬、毫无表情。近期的研究尝试通过注入显式情绪条件（例如 DiffPoseTalk（Sun 等人，2024）、ARTalk（Chu 等人，2025））或将提取出的语义特征补充到解耦的骨干网络中（例如 Wav2Sem（Li 等人，2025a））来缓解这一问题。然而，这两种策略都受限于其解耦特性：它们主要关注唇形重建，无法实现真正的对话级情绪理解。

在本工作中，我们通过提出**A^2-LLM**来应对这些挑战，即一种端到端对话式音频数字人大型语言模型。有别于解耦范式，我们的框架在统一的模态空间内联合推理语言、音频韵律与三维面部运动。具体而言，我们借助残差向量量化变分自编码器（Zeghidour 等人，2021）将面部动态离散为层次化词元，并使用专门的运动连接器将其桥接到语言骨干网络。通过联合解码同步的音频与面部词元，A^2-LLM在源头消除了语义—情感鸿沟。这种整体方法不仅将级联误差与系统延迟降至最低，还确保面部行为——从唇形同步到上面部表情——深深植根于对话语境，而非仅仅从声学表面推断而来。

为便于训练，我们引入**FLAME-QA**。与仅限于非结构化音视频（AV）对的传统数据集不同（Nagrani 等人，2017；Afouras 等人，2018；Zhang 等人，2021），我们的数据集被组织为多模态问答三元组。具体而言，每个样本将一个输入问题与一个多模态回答配对，该回答同时包含同步的语音与面部动态。纳入这一前序问题施加了先前工作所缺乏的语义约束：它促使模型生成的面部行为不仅与声学信号同步，而且在逻辑上以对话上下文为条件。

总而言之，我们的贡献如下：

FLAME-QA 数据集： 我们提出了一个以问答三元组形式构建的多模态数据集。与非结构化的音视频对不同，这种格式提供了严格对齐的语义监督，确保面部动态在逻辑上受对话上下文的制约。

A^2-大语言模型框架： 我们提出了一个统一框架，可联合解码音频和面部标记。这种端到端方法消除了语义—情感鸿沟，从而实现响应迅速（~500毫秒）且情感根基深厚的对话交互。

生成式面部能力： 我们系统地验证了大规模语义理解与富有表现力的 3D 运动之间的潜在对齐。通过利用残差向量量化变分自编码器和运动连接器，我们证明了细粒度、语境感知的面部动态能够有效地植根于模型原生的语义处理之中。

2. 相关工作

端到端音频语言模型的演进。 口语对话智能体的范式正从高延迟的模块化级联架构转向统一的大型音频语言模型（LALMs）（Huck Yang & Heck，2025）。JoyVoice（Yu et al.，2025）、EMOVA（Chen etal.，2025）和 Step-Audio（Wuet al.，2025）等近期创新将语音处理直接集成到大语言模型骨干网络中，实现了丰富的情感韵律，并能处理复杂的多说话人场景。然而，尽管 GPT-4o（OpenAI，2024）等模型具有出色的听觉表现力，它们仍然是“无具身”的。它们擅长生成声学信号，却缺乏原生的能力来控制三维形态化具身形象，因而在视觉面对面交互中留下了显著的空白——在这种交互中，听觉与视觉线索天然耦合。

生成式虚拟形象与情感对齐。为了可视化语音，相关研究已发展为两条主要路线：二维视频合成和三维面部动画。尽管二维扩散模型（Low 和 Wang，2025）实现了照片级真实感，却缺乏沉浸式虚拟现实/扩展现实所需的几何一致性。相反，原生三维方法通常依赖音频驱动回归。正如人机交互研究（Zhang 等人，2025；Llanes-Jurado 等人，2024）所指出的，这类解耦方法往往存在语义—情感鸿沟——生成的唇部运动千篇一律，无法体现对话的情感语境（例如，语气悲伤，面部却毫无表情），从而显著增加用户的认知负荷。不同于纯粹的音频响应代理（Park 等人，2024），A^2-LLM 利用离散运动标记化，将面部动态视为一种语言。通过在我们提出的 FLAME-QA 数据集上训练，我们确保生成的表情不仅与音频同步，还能以对话的语义意图为逻辑依据。

3. 预备知识

3.1. 端到端音频语言模型

为绕过级联系统固有的延迟和错误传播问题，我们基于 Step- Audio-2-mini（Wu 等人，2025）的架构构建了一个端到端的 LALM。该模型以 Qwen2.5-7B（Xu 等人，2025）为初始化模型，并采用 Qwen2-Audio 的编码器，旨在无需中间文本瓶颈即可直接感知和生成语音。该架构集成了音频编码器、适配器、大语言模型解码器和流匹配去标记器。关键在于，解码器以自回归方式生成离散文本与音频标记交错组成的统一序列。这种同步机制确保解码器的隐藏状态能够内在地捕捉语义推理和

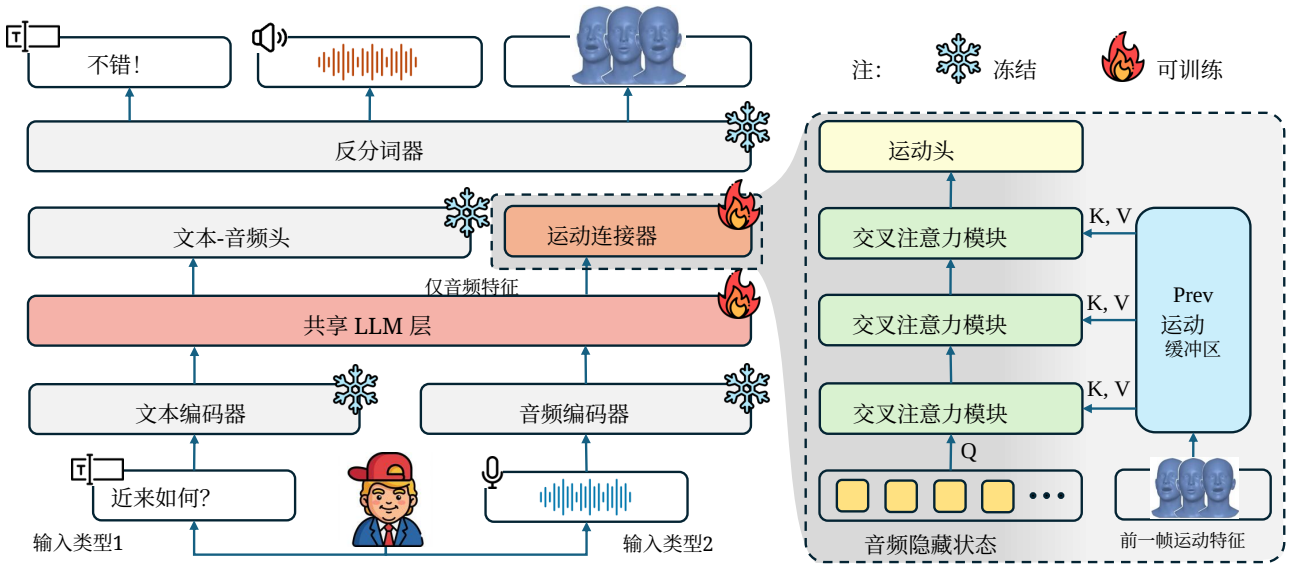


图 *1.* **A 示意图[2]-大语言模型。左侧：**统一框架中，共享 LLM 层与运动连接器经过联合优化，可生成同步的语音和面部动态。**右侧：**运动连接器通过运动头预测分层令牌。它采用分段式自回归设计，利用音频隐藏状态（作为查询）关注缓冲的运动历史（键和值），以保持时间连续性。

声学韵律，从而形成丰富且在时间上对齐的特征空间，这对于驱动富有表现力的面部运动至关重要。

3.2. FLAME 3D 面部表示

为使对话智能体拥有几何一致的化身，我们采用 FLAME 3D 可变形人脸模型（Li 等人，2017）。FLAME 以数学方式将面部几何解耦为低维参数，用于控制身份、姿态和表情。形式上，三维人脸网格 $\mathbf{V} \in \mathbb{R}^{5023\times 3}$ 通过可微函数 $M(\beta,\ \psi,\ \theta)$重建：

$$\mathbf{V} = M(\beta, \psi, \theta) \tag{1}$$

其中，$\beta \in \mathbb{R}^{|\beta|}$ 表示形状参数（身份），$\psi \in \mathbb{R}^{|\psi|}$ 表示表情参数，$\theta \in \mathbb{R}^{|\theta|}$ 表示姿态参数（颈部、下颌和全局旋转）。

在本研究中，我们专注于合成动态对话行为。我们固定身份形状 β，并对每一帧 t的时变状态向量$\mathbf{x}_t \in \mathbb{R}^D$ 进行建模。具体而言，我们的配置采用 50 个表情参数、3 个下颌姿态参数、3 个全局姿态参数和 2 个眼睑参数，从而得到维度为 $D = 58$的紧凑运动空间。

3.3. 问题表述

我们将该任务表述为联合多模态生成。给定对话上下文$\mathbf{C}$ 和用户音频$\mathbf{A}^{in}$，目标是合成一个由三种同步模态组成的响应 $\mathbf{R}$：

文本$\mathbf{Y}^t$、语音音频$\mathbf{Y}^a$以及三维面部运动$\mathbf{Y}^m$。具体而言，我们将面部运动模态表示为 FLAME 参数的连续序列$\mathbf{M} = (\mathbf{x}_1, \ldots, \mathbf{x}_T) \in \mathbb{R}^{T\times D}$，其中$\mathbf{x}_t$表示第 3.2 节中定义的逐帧状态向量。3.2。

生成过程分解为两个相互耦合的阶段。首先，LALM 骨干网络自回归地生成交错的文本和音频词元。为确保紧密的视听同步，同时避免增加上下文长度，我们采用音频锚定 运动策略。具体而言，我们提取大语言模型在音频生成过程中产生的隐藏状态序列$\mathbf{H}$，并通过可学习连接器 $\mathcal{F}_\theta$将其映射为面部动作词元：

$$\mathbf{Y}^m_t = \mathcal{F}_\theta\left(\mathbf{H}_t, \mathbf{Y}^m_{<t}\right) \tag{2}$$

这种表述确保$\mathbf{Y}^m$ 既以$\mathbf{Y}^t$ 中的语义推理为基础，也以$\mathbf{Y}^a$中捕获并包含于$\mathbf{H}$的声学韵律为基础。

为支持这一目标，我们需要将数据格式化为多模态问答对。形式上，我们将一个训练样本定义为元组 $(\mathbf{Q},\mathbf{R})$，其中响应 $\mathbf{R}$ 被分解为对齐的模态：$(\mathbf{Q},\mathbf{R}_{text}, \mathbf{R}_{audio}, \mathbf{R}_{visual})$。由于缺乏此类遵循指令的音视频数据，我们构建了 **FLAME-QA** 数据集（第4节）。

4. FLAME-QA 数据集

现有数据集，如 TFHP（Sun 等人，2024）和 HDTF（Zhang 等人，2021），仅提供原始音视频对，缺乏我们的问题表述所需的指令遵循结构（Wei 等人，2021；Wang 等人，2023；Liu 等人，2023；Ouyang 等人，2022）。为此，我们构建了 FLAME-QA，这是一个大规模多模态数据集，包含 ∼100经过专门筛选和格式化的高质量样本，用于端到端虚拟形象训练。

数据整理流程。我们从 VoxCeleb（Nagrani 等人，2019）的 80 万个原始片段中筛选数据，并实施严格的双流处理流程。在视觉流中，我们使用 SMIRK（Retsinas 等人，2024）从每一帧中提取高保真 FLAME 参数（β，ψ，θ），作为真实运动目标。同时，声学语义流使用 Whisper（Radford 等人，2022）进行自动语音识别转录，并通过辅助模块标注情感状态。为确保数据质量，我们利用大语言模型评估语义连贯性，过滤掉逻辑不连贯或语音不清晰的片段，最终得到包含 10 万个样本的精炼子集。

多模态三元组构建。为了满足第（$\mathbf{Q}$，$\mathbf{R}$）节 3.3 中定义的结构要求，我们必须合成缺失的前置指令。对于每条选定的转录（作为真实回复文本 $\mathbf{R}_{text}$），我们会提示大语言模型生成一个上下文感知问题$\mathbf{Q}_{text}$，然后使用 IndexTTS2（**将其合成为语音**（Zhou 等人，2025）。这一过程生成完全对齐的三元组，其中面部运动 $\mathbf{R}_{visual}$ 在语义上锚定于由$\mathbf{Q}$建立的对话上下文。

任务格式化与表情增强。为防止灾难性遗忘并保持通用音频能力，我们将数据组织为四个在统一目标下训练的子任务：音频问答（音频问答）、复述（语音重合成）、文本转语音（文本转语音）和自动语音识别（自动语音识别）。此外，为突破新闻类数据集有限的情感范围，我们精心构建了一个包含 ∼1k 个样本的高动态子集。这些样本由 InfiniteTalk（Yang 等人，2025b）生成，包含极端表情和丰富的情感内容（如大笑、惊讶），为生动的情感生成提供了有力监督。

5. 方法

在本节中，我们介绍 A²-LLM，这是一种端到端的多模态语言模型，旨在实现富有表现力且结合语境的感知驱动的面部运动生成。如图 1所示，我们的系统摒弃了级联管线，在统一框架内联合建模语言、音频韵律和面部动态。我们首先定义联合多模态生成任务以及音频锚定建模

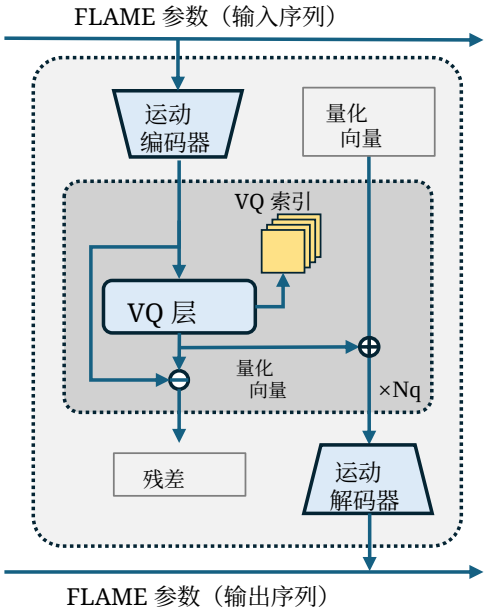


图 *2.* **运动标记化模块示意图。** 该模块通过残差向量量化（RVQ），将连续 FLAME 参数序列编码为紧凑的分层离散标记，从而实现高效的自回归建模。

方法（第3.3节）。接下来，我们介绍残差运动标记化（第5.1节），该方法通过时序 RVQ-VAE 将连续的面部动态离散化为分层标记。随后，我们介绍运动连接器（第 5.2节），这是一个利用历史条件注意力将这些标记与大语言模型骨干网络对齐的模块。接着，我们详细说明模型架构，以及用于平衡指令遵循与运动表现力的三阶段课程策略（第5.3节）。最后，我们定义统一训练目标（第 5.4节），以优化系统，实现文本、音频和运动的同步生成。

5.1. 残差运动标记化

为对高频面部动态进行建模，同时保持时间一致性，我们采用了配备时间压缩机制的残差向量量化 VAE（残差向量量化变分自编码器）（Zeghidour 等人，2021）。我们的残差向量量化变分自编码器架构如图2所示。形式上，给定一个原始运动序列 $\mathbf{M} \in \mathbb{R}^{T\times D}$（其中 T 对应于 25 Hz 的音频帧率），编码器会将其在时间维度上压缩 G 倍（例如，$G = 5$），从而得到一个紧凑的潜在序列 $\mathbf{z} \in \mathbb{R}^{(T/G)\times d_z}$。

为同时捕捉语义结构和细粒度细节，我们使用包含 N_q 个码本的残差向量量化（RVQ）对 $\mathbf{z}$ 进行离散化。量化过程将潜向量分解为离散码之和：

$$\mathbf{q} = \sum_{j=1}^{N_q} \mathbf{e}_{k_j}, \quad \text{where } \mathbf{e}_{k_j} = \mathcal{C}_j(\mathbf{r}_{j-1}) \tag{3}$$

其中，$\mathcal{C}_j$ 表示第 j 个码本，残差递归定义为 $\mathbf{r}_j = \mathbf{r}j-1 - \mathbf{e}k_j$。
整体训练目标结合了重建保真度与码本承诺：

$$\mathcal{L}_{\text{VQVAE}} = \mathcal{L}_{\text{rec}} + \| \text{sg}[\mathbf{z}] - \mathbf{q}\|_2^2 + \gamma\|\mathbf{z} - \text{sg}[\mathbf{q}]\|_2^2 \quad (4)$$

其中 $\text{sg}[\cdot]$ 表示停止梯度算子。为确保动画在几何上准确且在时间上稳定，我们不单纯依赖参数回归。遵循 ARTalk（Chu 等人，2025）的方法，我们通过明确引入顶点级监督（针对嘴唇和面部区域）以及时间导数约束（速度和加速度）来构建 $\mathcal{L}_{\text{rec}}$，从而确保精准的口型同步和平滑的运动过渡。

5.2. 具有层次化上下文的运动连接器

运动连接器弥合了音频中心大语言模型与离散运动空间之间的模态鸿沟。根据骨干网络架构，大语言模型解码器以高频率（25 Hz）生成音频标记。因此，对于当前处理窗口，我们获得一系列与音频对齐的隐藏状态 $\mathbf{H} \in \mathbb{R}^{T\times d}$。

特征下采样与对齐。由于目标 RVQ 标记处于压缩的时间空间中（如第5.1节所述），直接进行逐帧映射并不合理。因此，我们先通过一维卷积层进行特征级下采样，再进行线性投影。此操作将输入$\mathbf{H}$ 映射至压缩后的运动分辨率$\mathbf{H}' \in \mathbb{R}^{(T/G)\times d}$，从而严格确保语义特征与目标运动标记速率对齐。

历史条件注意力。为确保片段之间平滑过渡并保持时间连续性，我们通过交叉注意力 Transformer 采用分段自回归机制。我们不孤立地生成运动，而是明确地根据前一时间片段的运动轨迹来预测当前运动。具体而言，前一片段的离散运动标记经投影后形成稠密嵌入，作为上下文记忆。在 Transformer 层中，当前下采样后的音频特征 $\mathbf{H}'$ 充当查询，而历史运动嵌入则充当键和值。该设计使模型能够检索历史运动状态，在合成由当前音频韵律驱动的新动态时保持连贯性。

多尺度层次化预测。为了捕捉面部动态的完整谱系——从语义层面的下颌运动到细粒度的微表情——我们的连接器旨在预测完整的分层 RVQ 索引集合。我们采用 N_q 个特定层的预测头，每个预测头负责对第 k 个量化层的概率分布进行建模。系统经过优化，通过多任务学习目标进行优化，并汇总所有层次的交叉熵损失：

$$\mathcal{L}_{motion} = \sum_{k=1}^{N_q} \mathcal{L}_{CE}(\mathbf{q}^{(k)}, \hat{\mathbf{q}}^{(k)}) \quad (5)$$

其中 $\mathbf{q}^{(k)}$ 表示真实值标记索引。该公式鼓励模型在统一的上下文中联合学习粗粒度语义结构与细微的高频细节。

5.3. 多模态大语言模型与训练策略

A^2-大语言模型基于 Step-Audio-2-mini（Wu etal., 2025）骨干网络构建。为赋予模型丰富的面部运动能力，同时保留其语言智能，我们采用低秩适应（LoRA）（Huet al., 2022），并提出三阶段课程方案，以应对多模态联合训练中的稳定性与收敛性权衡。

阶段 1：运动连接器预训练。我们的首要目标是使运动连接器收敛。实验表明，统一学习率会导致梯度坍塌，这是因为初始化后的连接器具有较高方差，并与预训练大语言模型相互作用。因此，我们采用差分学习率策略（例如，LoRA：$1e^{-4}$，连接器：$1e^{-5}$）。然而，这会造成收敛速度差异：快速学习的 LoRA 模块迅速过拟合并遭受灾难性遗忘，而学习缓慢的连接器则需要更多步数才能捕捉细粒度动态。这形成了两难局面：过早停止会导致运动抖动，而延长训练则会损害语言能力。

阶段 2：通过 LoRA 重置进行联合对齐。为解决这一问题，我们采用 LoRA 重置策略。阶段 1 中连接器收敛后，我们丢弃过拟合的 LoRA 权重，同时保留已收敛的运动连接器。随后，我们重新初始化 LoRA，并使用相同的差分学习率进行第二轮联合微调。在这一阶段，预训练连接器充当稳定的语义锚点。全新的 LoRA 模块能够以少得多的步数有效地与运动模态对齐，从而避免因训练时间过长而导致的灾难性遗忘。

阶段 3：情感指令微调。最后，为了赋予虚拟形象生动的情感特征，我们继续使用精心筛选的 ∼1k 个高自由度情感问答样本训练模型。此阶段起到风格注入的作用，使系统能够生成细腻且富有共情力的面部表情，充分体现对话的情感语境。

5.4. 训练目标

我们使用标准的交叉熵损失来优化 A^2-LLM 框架。与骨干网络的架构保持一致，我们将文本和音频词元视为统一的交错序列，而非彼此独立的模态。因此，总训练损失的公式为：

$$\mathcal{L} = \mathcal{L}_{\text{seq}} + \lambda \mathcal{L}_{\text{motion}} \tag{6}$$

其中，$\mathcal{L}_{\text{seq}}$ 表示文本与音频联合序列的自回归下一标记预测损失，而 $\mathcal{L}_{\text{motion}}$ 表示预测运动标记的分层交叉熵损失（如公式5所定义）。超参数 λ 用于平衡主生成骨干网络与运动连接器之间的梯度幅值，确保语义连贯性、声学韵律和面部动态得到同步优化。

6. 实验

本节系统评估了模型的性能，包括其实时能力、语言能力和面部表情合成能力。最后，我们通过分阶段训练开展消融实验，以验证该训练策略的有效性。

6.1. 实时行为

为验证 A^2-LLM 作为响应式数字人的能力，我们使用三个关键指标评估其推理延迟和处理效率。**首个标记时间（TTFT）**用于衡量从用户输入到生成首个音频标记的延迟。**首个动作时间（TTFA）**表示生成首个面部运动前的总延迟。**实时因子（RTF）**定义为生成时间与合成内容时长之比，用于量化吞吐效率。**性能分析**。

表1。 实时性能比较。A^2-大语言模型（优化版）利用图编译来最大限度地减少内核启动开销。级联系统的结果是初步估算值，根据不同响应长度取平均得出。

方法	首个标记时间（毫秒）	首个动作时间（毫秒）	RTF
级联系统	-	13,730	> 1.0
级联（流式）	-	3,232	> 1.0
A^2-大语言模型（基础）	47.24	850.15	1.005x
A^2-大语言模型（优化）	50.71	**535.53**	**0.703x**

如表1所示，利用静态图编译可显著降低开销。优化后的 A^2-LLM 首个动作时间为 535 毫秒，实时因子为 0.7 倍。这证实，我们的系统生成面部动态的速度快于实时播放，从而确保交互流畅且不会阻塞。

与基线方法的比较。我们以标准级联系统作为基准。批处理会产生难以承受的延迟（>13 秒），即使经过优化的流式处理管线也会受到累积模块延迟的影响（~3.2秒），而 A^2-大语言模型的统一架构消除了这些瓶颈。通过联合解码音频和运动，我们的模型将延迟降低了一个数量级（首个动作时间为 ~0.5 秒），有效弥合了处理与感知之间的差距。

6.2. 语言能力

为严格评估整合面部运动 token 是否会损害模型的语言智能，我们使用 *Open-VoiceBench*开展了全面评估。该基准由 Baichuan-Audio 提出（Li 等人，2025b），专门用于评估原生音频模型在各类领域中的能力，包括逻辑推理、知识检索和指令遵循。

基线方法。我们将由 8B 大语言模型骨干网络、0.16B LoRA 适配器和 0.12B 运动连接器组成的 A^2-LLM，与领先的端到端音频模型进行基准测试：GLM-4-Voice（9B）（Zeng 等人，2024）、Baichuan-Audio（7B）（Li 等人，2025b）、StepAudio-Chat（132B）（Huang 等人，2025）以及 Qwen2.5-Omni（7B）（Xu 等人，2025）。这些基线方法的结果直接取自其官方开源报告。此外，我们还报告了纯文本模型 Qwen3-4B 和 Qwen3-8B 的结果（Yang 等人，2025a），将其作为纯语言理解能力的理论上限。

结果与分析。表2中的定量结果表明了两个关键发现。首先，A^2-LLM 在音频原生系统中表现最佳，在 *Alpaca Eval*（74.20）和*Trivia QA*（79.90）上均取得最高分。值得注意的是，在*Reasoning QA*方面，我们的英语微调数据与中文基准测试之间存在语言不匹配，导致模型偶尔不遵循指令。其次，更重要的是，我们的模型保持了与纯文本基线方法相当的性能。这证实，我们的训练策略（第5.3节）能够有效对齐视觉模态，同时避免灾难性遗忘。面部动态的生成并未与模型容量争夺资源，而是与语言骨干网络在语义上保持一致，从而确保富有表现力的输出始终以高阶推理为基础。

6.3. 面部动作质量与动态

评估端到端对话模型的面部运动质量面临着一项独特挑战：生成差距。与从固定音频重建运动的传统音频驱动任务不同，A^2-LLM 会生成具有独特韵律和节奏的语音，因此自然会与预先录制的真实标注（GT）数据集有所不同。为解决这一问题，我们采用多维评估协议，同时评估空间表现力和时间动态。

表2. 在 OpenVoiceBench 上的评估。纯文本 Qwen3模型作为语义上界。最佳原生音频结果以**粗体**显示。缩写：Alp.：AlpacaEval，LQ：Llama Questions，RQA：Reasoning QA，TQA：Trivia QA，WQ：Web Questions。

模型	Alp.	LQ	RQA	TQA	WQ
纯文本（上界）					
Qwen3-4B	76.08	75.70	39.75	82.66	75.64
Qwen3-8B	78.19	82.16	39.29	86.39	83.93
原生音频模型					
GLM-4-Voice	48.90	71.00	26.50	46.60	51.50
Baichuan-Audio	69.60	74.50	**37.90**	54.20	60.30
StepAudio-Chat	56.53	72.33	30.00	56.80	**73.00**
Qwen2.5-Omni	72.76	75.33	31.88	57.06	62.80
A²-大语言模型	**74.20**	**75.60**	34.60	**79.90**	60.60

评估协议：伪预言机策略。我们采用参考模型比较策略，使用**Diff-PoseTalk** (Sun 等人，2024) 作为伪预言机。DiffPoseTalk 是一种优先考虑发音动作精度而非实时效率的高保真扩散模型，是唇形同步的理想“上限”。通过基于 A相同的音频合成结果生成参考运动²-LLM，我们建立了统一的比较基准，以便与领先的基线方法进行比较（ARTalk（Chu 等人，2025）、FaceFormer（Fan 等人，2022）、CodeTalker（Xing 等人，2023））。所有评估均在 FLAME-QA 留出测试集上进行。

空间分析：同步与表现力。我们首先使用**嘴部张开差异 (MOD)** 评估帧级空间质量，以衡量唇部同步精度；并使用**Up- perFace 动态 (UFD)**评估情感表现力。

表 3. 空间质量比较。A²-大语言模型在唇形同步（嘴部张开差异）方面表现出竞争力，同时相比音频驱动基线，在上半脸表现力方面显著更优。

模型	嘴部张开差异 (↓)	UFD (↑)
ARTalk	**4.60 ± 0.80**	9.40 ± 1.29
CodeTalker	5.29 ± 0.84	2.38 ± 0.17
FaceFormer	5.75 ± 0.90	3.14 ± 0.35
A²-大语言模型（我们的方法）	5.08 ± 0.88	**11.13 ± 1.48**

如表 3所示，A²-大语言模型取得了颇具竞争力的嘴部张开差异（MOD）分数（5.08），与表现出色的基线方法 ARTalk 相当，并优于其他方法。这证实了我们的端到端架构能够成功学习细粒度的构音控制。更重要的是，A²-大语言模型在表现力方面占据优势（UFD：11.13）。以音频为中心的基线方法往往由于对视听相关性较弱而产生“冻结”的上面部，而我们的模型则利用语义理解注入上下文感知表情（例如，惊讶引发的扬眉），实现真正的具身对话

体验。

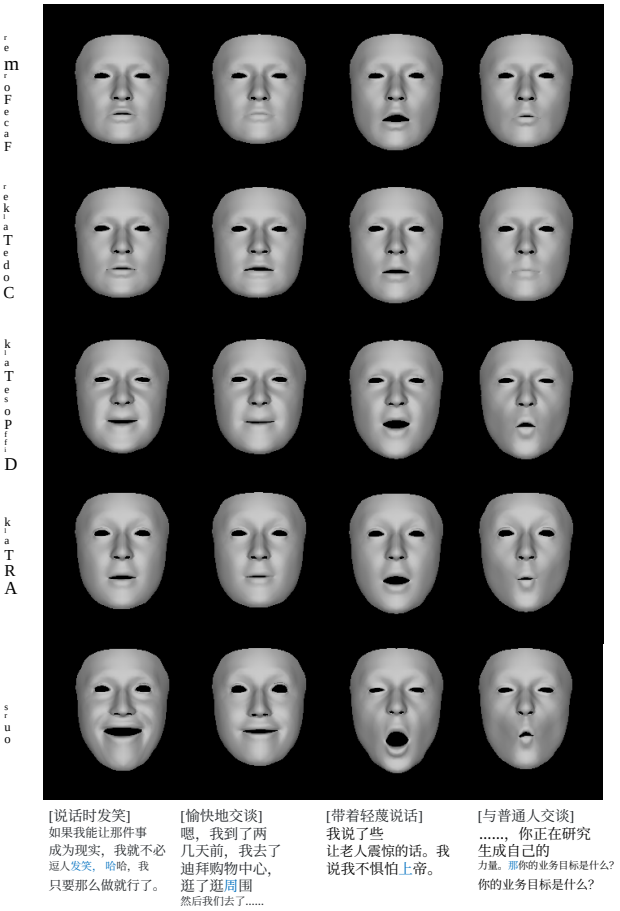


图3. 面部表现力的视觉对比。A²-大语言模型能够自发地生成与回复内容更为匹配的面部运动，而无需显式的情感条件。

时序分析：频率与动态。除了空间对齐，真实感动画还需要具备物理合理性和时间一致性。我们在频域中分析运动动态，并将模型与最强基线 ARTalk 进行比较。

表 4中的结果表明，A²-LLM 在五个关键维度上具有更高的动态保真度：**(1)全局节奏：** 时间相关性，用于衡量整体运动趋势与参考数据的同步程度，结果显示显著提升（+112%），表明我们的模型能更好地捕捉全局对话节奏。**(2)动态趋势：** 速度相关性，用于表示运动方向（一阶导数）的一致性；我们的模型得分为正（0.111），而 ARTalk 得分为负，凸显了我们在方向准确性方面的优势。**(3)嘴部形变：** 唇宽相关性，用于量化对微笑和元音发音至关重要的水平嘴部拉伸，提升了 26%，验证了发音控制更加精细。**(4) 运动能量：** 生动度比率，即预测运动方差与参考运动方差之比，达到 1.09（更接近目

表4. 时间动态分析。与 ARTalk 相比，A²-大语言模型表现出更高的节奏一致性（相关性）和更自然的运动幅度（生动度）。

指标	ARTalk	A²-大语言模型（我们的方法）
时间相关性（↑）	0.218	**0.464**
速度相关性（↑）	-0.309	**0.111**
唇宽相关性（↑）	0.477	**0.604**
生动度比率（目标 ≈ 1.0）	0.804	**1.087**
峰值对齐（毫秒）（↓）	116.6	**114.3**

标值 1.0），证实我们避免了回归模型中常见的“过度平滑”效应。**(5)延迟：** 峰值对齐，用于衡量最大张口的时间滞后，降低至 114 毫秒，确保音视频同步更加紧密。

定性分析。为验证较高的 UFD 分数是否能转化为有意义的表现力，我们在不同语义语境下展示了图中的视觉对比。3 。基线方法往往保持中性表情，主要关注唇部发音，而 A²-LLM 展现出更优异的语义-视觉对齐能力。

例如，在笑声片段（“哈哈”）中，我们的模型会自然地生成脸颊上扬和微笑嘴型；在表达轻蔑陈述（“我并不害怕上帝”）时，则会呈现细微的皱眉和不屑的表情。同样，模型能够区分轻松叙述（“四处漫步”）与常规商务询问，并据此调整面部紧张度。这些结果证实，A²-LLM 成功利用骨干网络的语义理解能力，生成语境感知且情感一致的面部行为，超越了简单的音频与唇部同步。

表5。 **用户偏好研究结果**。 胜率表示在成对比较中，A²-大语言模型的秩高于（表现力更强）竞争基线的案例百分比（$N = 60$）。

我们的方法 vs. 基线	胜率（%）	平局（%）	负率（%）
我们的方法 vs. DiffPoseTalk（虚拟形象 1）	**71.7%**	10.0%	18.3%
我们的方法 vs. ARTalk（虚拟形象 2）	**75.0%**	5.0%	20.0%

用户偏好研究 为评估感知到的情感质量，我们开展了一项主观用户研究（$N = 60$ 个样本）。参与者根据表现力对随机排列的三元视频组进行排序，将我们的方法与 DiffPoseTalk 和 ARTalk 进行比较。如表 5 所示，A²-大语言模型相对于两种基线方法的偏好率均超过 71%。这些结果表明，与当前的音频驱动方法相比，人类评估者往往认为由我们的统一框架生成的面部动态在语境上更具吸引力。

6.4. 联合训练策略消融

为验证我们的训练策略对同步性和运动质量的影响，我们将最终模型与**仅适配器** 基线（冻结大语言模型）进行了比较。

表6. 唇部动态消融。仅适配器 基线存在严重延迟（515 毫秒）且相关性较差，而联合训练可确保相位精确对齐。

指标	仅适配器	联合训练
时间相关性（↑）	0.028	**0.464**
速度相关性（↑）	0.019	**0.111**
唇宽相关性（↑）	0.057	**0.604**
生动度比率（≈ 1.0）	0.850	**1.087**
峰值对齐（ms）（↓）	515.05	**114.30**

如表6所示，仅适配器基线表现出严重的同步滞后（峰值对齐 ∼515 ms），且运动相关性可忽略不计（0.028），表明冻结的大语言模型无法预判精确的发音时序。相比之下，联合训练将延迟大幅降低至114 ms，并恢复了正向速度相关性（0.111）。这证实了微调大语言模型对于使面部标记的相位和节奏与语音信号对齐至关重要。

7. 结论

在本项工作中，我们提出了**A²-LLM**，这是一个开创性的端到端框架，将语言处理、音频生成和面部动画统一于单一的多模态大模型中。通过突破传统的级联管线，我们有效弥合了语义—情感鸿沟，使数字化身能够生成不仅实现口型同步，而且深深植根于对话语境和情感意图的面部表情。

我们的贡献主要有三方面：提出一种统一架构，实现∼500毫秒的交互延迟；构建**FLAME-QA**，首个面向语义感知三维面部动画的大规模指令微调数据集；以及提出一种分层运动建模策略，同时确保稳定性和表现力。大量实验表明，在动态多样性和用户偏好方面，A²-LLM 显著优于现有的音频驱动基线。

局限性与未来工作。 尽管取得了这些进展，当前的局限性也指明了令人振奋的未来方向。首先，我们的模型目前针对英语进行了优化；拓展多语言支持是优先事项。其次，尽管我们专注于面部动态，将这一端到端范式拓展到全身姿势生成仍是一项尚待解决的挑战。我们相信，A²-LLM 为打造下一代富有共情力且真正具备交互能力的数字人奠定了坚实基础。

参考文献

Afouras, T., Chung, J. S., and Zisserman, A. LRS3-TED：一个用于视觉语音识别的大规模数据集。*arXiv* 预印本 *arXiv:1809.00496*，2018。

Cassell, J. 具身对话代理：用户界面中的表征与智能。《*AI Magazine*》, 22(4): 67–67, 2001.

Chen, K., Gou, Y., Huang, R., Liu, Z., Tan, D., Xu, J., Wang, C., Zhu, Y., Zeng, Y., Yang, K.等。EMOVA：赋能语言模型以鲜明情感看、听与说。载于计算机视觉与模式识别会议论文集，第5455–5466页，2025。

Chu, X., Goswami, N., Cui, Z., Wang, H.与Harada, T. ARTalk：基于自回归模型的语音驱动3D头部动画。载于*SIGGRAPH Asia 2025*会议论文集，第1–9页，2025。

Ding, D., Ju, Z., Leng, Y., Liu, S., Liu, T., Shang, Z., Shen, K., Song, W., Tan, X., Tang, H.等。Kimi-audio技术报告。*arXiv*预印本*arXiv:2504.18425*，2025。

Fan, Y.、Lin, Z.、Saito, J.、Wang, W. 和 Komura, T. Face-former：基于语音驱动的使用 Transformer 的三维面部动画。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第18770–18780页，2022年。

Fang, Q.、Guo, S.、Zhou, Y.、Ma, Z.、Zhang, S. 和 Feng, Y. Llama-omni：借助大语言模型实现无缝语音交互。*arXiv*预印本 *arXiv:2409.06666*，2024年。

Fang, Q.、Zhou, Y.、Guo, S.、Zhang, S. 和 Feng, Y. Llama-omni2：基于大语言模型的实时语音聊天机器人与自回归流式语音合成。*arXiv*预印本 *arXiv:2505.02625*，2025年。

Grassal, P.-W.、Prinzler, M.、Leistner, T.、Rother, C.、Nießner, M. 和 Thies, J. 从单目RGB视频生成神经头部头像。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第18653–18664页，2022年。

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 等。Lora：大语言模型的低秩适应。*ICLR*，1(2):3，2022。

Huang, A., Wu, B., Wang, B., Yan, C., Hu, C., Feng, C., Tian, F., Shen, F., Li, J., Chen, M., 等。Step-audio：智能语音交互中的统一理解与生成。*arXivpreprint arXiv:2502.11946*，2025。

Huang, R., Li, M., Yang, D., Shi, J., Chang, X., Ye, Z., Wu, Y., Hong, Z., Huang, J., Liu, J. 等。Audiogpt：理解并生成语音、音乐、声音和说话头部。载于*AAAI*人工智能大会论文集，第 38 卷，第 23802–23804 页，2024 年。Huck Yang, A. S. 和 Heck, L. P. 基于大语言模型的口语对话代理。*arXiv* 电子印本, 第 arXiv–2512 页，2025 年。Kerbl, B., Kopanas, G., Leimk¨uhler, T. 和 Drettakis, G. 用于实时辐射场渲染的 3D高斯泼溅。*ACM*图形学汇刊，42(4), 2023 年 7 月。Li, H., Dai, J., Zhao, X., Zhou, F., Pan, J. 和 Li, L. Wav2Sem：用于 3D语音驱动面部动画的即插即用音频语义解耦。载于计算机视觉与模式识别会议论文集，第 183–192 页，2025a。

Li, T., Bolkart, T., Black, M. J., Li, H., and Romero, J. 从4D扫描中学习面部形状和表情模型。*ACM Trans.Graph.*, 36(6):194–1，2017。

Li,T.,Liu, J., Zhang, T., Fang, Y., Pan,D.,Wang,M., Liang, Z., Li, Z., Lin, M., Dong, G. 等。Baichuan-Audio：面向端到端语音交互的统一框架。*arXiv*预印本 *arXiv:2502.17239*，2025b。

Liu, H., Li, C., Wu, Q. 和 Lee, Y.J. 视觉指令微调。神经信息处理系统进展，36:34892–34916，2023。

Llanes-Jurado, J., G´omez-Zaragoz´a, L., Minissi, M. E., Alca˜niz, M. 和 Mar´ın-Morales, J. 基于大语言模型开发用于社会情绪诱发的对话式虚拟人。《专家系统与应用》，246:123261，2024。

Low, C. 和 Wang, W. Talkingmachines：通过自回归扩散模型实现实时音频驱动的 FaceTime风格视频。*arXiv*预印本 *arXiv:2506.03099*，2025。

McGurk, H. 和 MacDonald, J. 聆听嘴唇，观看声音。《自然》，264(5588):746–748，1976。

Nagrani, A.、Chung, J. S. 和 Zisserman, A. Voxceleb：大规模说话人识别数据集。*arXiv*预印本 *arXiv:1706.08612*，2017。

Nagrani, A.、Chung, J. S.、Xie, W. 和 Zisserman, A. Voxceleb：真实环境下的大规模说话人验证。计算机科学与语言，2019。

OpenAI。GPT-4o 系统卡，2024。

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal,S., Slama, K.,Ray, A. 等. 利用人类反馈训练语言模型遵循指令. 神经信息处理系统进展, 35:27730–27744, 2022.

Park, S. J., Kim, C. W., Rha, H., Kim, M., Hong, J., Yeo, J. H. 和 Ro, Y. M. Let's go real talk：面向面对面对话的口语对话模型. *arXiv*预印本 *arXiv:2406.07867*, 2024.

Qian, S., Kirschstein, T., Schoneveld, L., Davoli, D., Gieben- hain, S., and Nießner, M. Gaussianavatars：采用绑定骨骼的三维高斯构建的照片级真实感头部数字人。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第 20299–20309 页，2024 年。

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. 通过大规模弱监督实现鲁棒语音识别，2022 年。

Retsinas, G., Filntisis, P. P., Danecek, R., Abrevaya, V. F., Roussos, A., Bolkart, T., and Maragos, P. 通过神经合成分析实现三维面部表情。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 2490–2501 页，2024 年。Sun, Z., Lv, T., Ye, S., Lin, M., Sheng, J., Wen, Y.-H., Yu, M., and Liu, Y.-j. DiffPoseTalk：通过扩散模型实现语音驱动的风格化三维面部动画和头部姿态生成。*ACM*图形学汇刊 *(TOG)*，43(4)：1–9，2024 年。Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. Self-instruct：通过自生成指令对语言模型进行对齐。载于计算语言学协会第 *61* 届年会论文集（第 *1* 卷：长篇论文），第 13484–13508 页，2023 年。Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. 微调后的语言模型是零样本学习器。*arXiv*预印本 *arXiv:2109.01652*，2021 年。Wu, B., Yan, C., Hu, C., Yi, C., Feng, C., Tian, F., Shen, F., Yu, G., Zhang, H., Li, J., 等。Step-audio 2 技术报告。*arXiv*预印本 *arXiv:2507.16632*，2025 年。Xing, J., Xia, M., Zhang, Y., Cun, X., Wang, J., and Wong, T.-T. CodeTalker：基于离散运动先验的语音驱动三维面部动画。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 12780–12790 页，2023 年。Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., 等。Qwen2. 5-omni 技术报告。*arXiv*预印本 *arXiv:2503.20215*，2025 年。Xu, M., Li, H., Su, Q., Shang, H., Zhang, L., Liu, C., Wang, J., Yao, Y., and Zhu, S. Hallo：用于肖像图像动画的分层音频驱动视觉合成。*arXiv*预印本 *arXiv:2406.08801*，2024a。

Xu, S., Chen, G., Guo, Y.-X., Yang, J., Li, C., Zang, Z., Zhang, Y., Tong, X., 和 Guo, B. Vasa-1：实时生成逼真的音频驱动说话人脸。神经信息处理系统进展，37:660–684，2024b。Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., 等。Qwen3 技术报告。*arXiv* 预印本 *arXiv:2505.09388*，2025a。Yang, S., Kong, Z., Gao, F., Cheng, M., Liu, X., Zhang, Y., Kang, Z., Luo, W., Cai, X., He, R., 等。InfiniteTalk：用于稀疏帧视频配音的音频驱动视频生成。*arXiv* 预印本 *arXiv:2508.14033*，2025b。Yehia, H., Rubin, P., 和 Vatikiotis-Bateson, E. 声道与面部行为的定量关联。*SpeechCommunication*，26(1-2):23–43，1998。Yu, F., Wang, T., Wu, Y., Zhu, L., Deng, W., Han, W., Wang, W., Hu, L., Liang, X., He, X., 等。JoyVoice：用于拟人化多说话人对话合成的长上下文条件控制。*arXiv* 预印本 *arXiv:2512.19090*，2025。Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., 和 Tagliasacchi, M. Soundstream：端到端神经音频编解码器。*IEEE/ACM*音频、语音与语言处理汇刊，30:495–507，2021。Zeng, A., Du, Z., Liu, M., Wang, K., Jiang, S., Zhao, L., Dong, Y., 和 Tang, J. GLM-4-Voice：迈向智能、类人且端到端的语音聊天机器人。*arXiv* 预印本 *arXiv:2412.02612*，2024。Zhang, D., Li, S., Zhang, X., Zhan, J., Wang, P., Zhou, Y., 和 Qiu, X. Speechgpt：赋予大语言模型内在的跨模态对话能力。*arXiv* 预印本 *arXiv:2305.11000*，2023。Zhang, Y., Lucas, M., Bem-haja, P., 和 Pedro, L. AI 与人类生成的语音和虚拟形象：重新思考用户参与度与认知负荷。教育与信息技术，第 1–20 页，2025。Zhang, Z., Li, L., Ding, Y., 和 Fan, C. 基于光流引导的一次性说话人脸生成及高分辨率视听数据集。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 3661–3670 页，2021。Zhou, S., Zhou, Y., He, Y., Zhou, X., Wang, J., Deng, W., 和 Shu, J. IndexTTS2：情感表达与时长控制自回归零样本文本转语音领域的突破。*arXiv* 预印本 *arXiv:2506.21619*，2025。

A. 数据处理

FLAME-QA 构建流程

我们从约 80 万段 VoxCeleb 视频片段开始，并应用以下自动化流程：

- 转录。 我们使用 Whisper v3 Large 获取每个片段的文本转录稿。
- 情感估计。 语音情感识别模块会预测话语级情感标签，之后将其用作辅助监督。
- 清理与问答生成。 我们提示 GPT-5.1 清理转录稿、移除嘈杂或不完整的片段，并生成相应的问题，从而得到问答对。只有通过质量筛选条件（例如长度和可懂度）的片段才会被保留，最终得到约 10 万个问答对。
- 问题音频合成。 我们使用 IndexTTS2 文本转语音模型，根据生成的文本合成语音问题，获得配对的问题音频、答案音频（原始音频）和文本。

富含情感的合成问答。 对于富含情感的子集，我们首先使用大语言模型对问答文本进行抽样和编辑，以突出多样化的情感内容和个性化风格。问答文本通过文本转语音转换为音频，然后输入 InfiniteTalk，生成表情丰富的说话头像视频，共约 0.8K 个片段。我们对所有视频应用 SMIRK，提取帧级 FLAME 表情与姿态参数，作为真实运动监督。

我们使用 GPT-5.1 作为主要的数据清理引擎。具体而言，我们将原始自动语音识别转录文本与精心设计的提示词（见图 4）一并输入 GPT-5.1，以纠正识别错误、删除语流不流畅现象和无关内容，并构建高质量问答对。

转录稿清理与筛选提示词

系统：
你是一名乐于助人的助手，专门负责评估文本质量并为问答训练数据集生成问题。

用户：
评估以下文本是否适合作为问答训练数据集中的答案（完整的英文句子、语法正确、内容有意义）。文本：{答案文本}如果适合，请生成一个自然的问题。如果不适合，请将问题设为 null。请以 JSON 格式回复：{“is suitable”：true 或 false，“question”：“你的问题”或 null}仅回复 JSON 对象。

图 4. GPT-5.1 用于清理自动语音识别转录文本、评估其是否适合作为问答答案，并以 JSON 格式生成相应问题的提示词，用于构建 FLAME-QA 数据集。

B. 实现细节

硬件与环境。 所有实验，包括残差运动分词器的训练和 A^2-LLM 的指令微调，均在配备 $8\times$ 英伟达 A100（80GB）图形处理器的计算节点上进行。该框架使用 PyTorch 实现。

残差向量量化变分自编码器架构与训练。残差运动离散化模块将连续的 FLAME 参数序列（25 fps）压缩为 5 Hz 分辨率的离散码（时间分组大小为 $G = 5$）。该架构由编码器和解码器组成，两者的基础通道宽度均为 128，且各有 3 个残差块。潜在嵌入维度为 256，由 $N_q = 6$ 个层次量化器进行离散化，每个量化器的码本大小为 256。

我们使用 AdamW（lr $= 1 \times 10^{-4}$）优化模型，训练 6 个训练轮次，全局批次大小为 512（每个图形处理器 64）。正则化包括 0.05 的丢弃率和权重为 $\gamma = 0.25$ 的码本承诺损失。向量量化损失的权重设为 $\lambda_{\text{vq}} = 1.0$。

重建目标。为确保整体运动保真度，我们优化了一个复合目标。总重建损失由参数空间回归和顶点空间约束组成：

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{param}} + 10^5(\mathcal{L}_{\text{lips}} + \mathcal{L}_{\text{face}}) + 10^2(\mathcal{L}_{\text{vel}} + \mathcal{L}_{\text{acc}}) \tag{7}$$

各组成部分及其具体配置定义如下：

1) 参数回归：我们在58维 FLAME 参数空间中，以单位权重（$\lambda_{\text{param}} = 1.0$）确保基线一致性：

$$\mathcal{L}_{\text{param}} = \|\mathbf{M} - \hat{\mathbf{M}}\|_2^2 \tag{8}$$

2) 空间几何一致性：为捕捉精细的形态，我们将参数映射到规范零姿态空间中的三维顶点。为补偿公制尺度顶点坐标幅值较小的问题，我们对唇部和面部区域的 L_2 距离赋予较高权重（$\lambda_{\text{lips}} = \lambda_{\text{face}} = 1.0 \times 10^5$）：

$$\mathcal{L}_{\text{geo}} = \|\mathbf{V}_{lips} - \hat{\mathbf{V}}_{lips}\|_2^2 + \|\mathbf{V}_{face} - \hat{\mathbf{V}}_{face}\|_2^2 \tag{9}$$

3) 时间动态一致性：为确保平滑性并抑制抖动，我们对完整顶点序列的速度（$\Delta\mathbf{V}$）和加速度（$\Delta^2\mathbf{V}$）施加约束。这些项的权重为 $\lambda_{\text{vel}} = \lambda_{\text{acc}} = 1.0 \times 10^2$：

$$\mathcal{L}_{\text{dyn}} = \|\Delta\mathbf{V} - \Delta\hat{\mathbf{V}}\|_2^2 + \|\Delta^2\mathbf{V} - \Delta^2\hat{\mathbf{V}}\|_2^2 \tag{10}$$

运动连接器配置。运动连接器充当大语言模型骨干网络与离散运动空间之间的桥梁。它被实现为一个由 6 层组成的轻量级 Transformer 解码器。我们将隐藏维度设为 768，并在每层使用 12 个注意力头，以捕捉复杂的音频-运动依赖关系。为防止联合对齐阶段出现过拟合，我们在整个网络中采用 0。1 的丢弃率。

LoRA配置。为在不产生过高计算成本的情况下，将大语言模型骨干网络高效适配于多模态生成，我们采用低秩适应（LoRA）。我们将秩配置为 $r = 64$，缩放因子设为 α =32，丢弃率设为 0。05。适配器被注入 Transformer 块的所有线性层，具体针对注意力投影（q proj、k proj、v proj、o proj）和前馈网络模块（gate proj、up proj、down proj）。

C. 评估指标

为全面评估生成的三维面部动画质量，我们采用一组涵盖几何精度、表情动态和时间同步的多样化指标。所有指标均在规范零姿态空间中计算（即移除全局旋转和头部平移），以便纯粹关注面部变形。

C.1. 空间与表现力指标

嘴部张开距离（嘴部张开差异）。嘴部张开差异用于衡量唇形同步的几何精度。它计算预测序列与真实序列之间垂直嘴部张开度的平均绝对误差（MAE）。设 $h(\mathbf{V})$ 表示给定顶点帧 $\mathbf{V}$ 中的垂直嘴部张开度（上下唇之间的距离）。该指标定义如下：

$$\text{MOD} = \frac{1}{T}\sum_{t=1}^{T} |h(\hat{\mathbf{V}}_t) - h(\mathbf{V}_t)| \times 1000 \tag{11}$$

ˆ其中，$\mathbf{V}$ 和$\mathbf{V}$ 分别表示预测顶点和真实顶点。结果以毫米（mm）为单位报告。值越低，表示唇部闭合和张口度越准确。

上脸动态（UFD）。UFD 是一种无参考指标，用于量化上脸部运动（例如眨眼、扬眉）的强度。它衡量相对于中性

模板而言，顶点位移的帧间变化率。首先，我们计算上脸区域中每个顶点 i 相对于中性面部 $\mathbf{V}_{neu}$:$d_{t,\ i} = \|\mathbf{v}_{t,\ i} - \mathbf{v}_{neu,\ i}\|_2$的位移范数。随后，将 UFD 计算为该位移随时间变化的平均值:

$$\text{UFD} = \frac{1}{T-1}\sum_{t=1}^{T-1}\left(\frac{1}{|S_{up}|}\sum_{i\in S_{up}}|d_{t,i} - d_{t-1,i}|\right) \times 10^5 \tag{12}$$

其中 S_{up} 表示 t 上脸索引集合。值越高，表示面部表情越丰富、越具动态性 s.

C.2. 时间动态与同步

为评估生成动作的物理合理性和细粒度同步性，我们在频域和时域中分析嘴部开合信号 $O_t = h(\mathbf{V}_t)$ 以及嘴部宽度信号 W_t （嘴唇两角之间的水平距离）。

1) 时间相关性： 该指标用于衡量全局节奏对齐程度。它计算预测嘴部开合序列与真实值嘴部开合序列之间的皮尔逊相关系数（PCC）:

$$\rho_{\text{temp}} = \frac{\text{Cov}(O_{pred}, O_{gt})}{\sigma(O_{pred})\sigma(O_{gt})} \tag{13}$$

值越高，表示生成动作越能遵循正确的开合节奏。

2) 速度相关性： 为评估运动的方向一致性，我们计算速度曲线（一阶导数）的相关性。令 $v_t = O_t - O_{t-1}$。该指标定义如下:

$$\rho_{\text{vel}} = \text{PCC}(v_{pred}, v_{gt}) \tag{14}$$

正值表示虚拟形象的嘴部张合与真实标注严格保持相位同步。

3) 唇宽相关性： 尽管垂直张合在语音中占主导地位，水平伸展对于情绪表达（例如微笑）也至关重要。我们计算嘴宽序列的皮尔逊相关性:

$$\rho_{\text{width}} = \text{PCC}(W_{pred}, W_{gt}) \tag{15}$$

4) 生动度比率： 基于回归的模型常会出现“过度平滑”，导致运动显得呆板。生动度比率比较预测值与真实标注之间速度（运动能量）的标准差:

$$R_{\text{live}} = \frac{\sigma(v_{pred})}{\sigma(v_{gt}) + \epsilon} \tag{16}$$

接近 1.0 的值表示生成动作具有自然的动态幅度，既不过度平滑，也不会抖动。

5) 峰值对齐延迟（峰值对齐）： 该指标用于量化同步延迟。我们分别检测两组嘴部张开曲线中的局部最大值（峰值）。对于真实标注 $p_{gt}^{(i)}$ 中的每个峰值，我们在预测值 $p_{pred}^{(j)}$ 中找到最近的峰值，并计算绝对时间差中位数:

$$\text{Lag} = \text{Median}\left(\min_j |t(p_{gt}^{(i)}) - t(p_{pred}^{(j)})|\right) \tag{17}$$

结果被转换 为毫秒（ms）。值越低，表示与语音信号的同步越紧密。