arXiv:2603.01418v1 [cs.CV] 2 Mar 2026

UniTalking：用于说话人像生成的统一音视频框架

Hebeizi Li [1,2*†], Zihao Liang [1*], Benyuan Sun[1‡], Zihao Yin[1], Xiao Sha[1], Chenliang Wang[1], Yi Yang[1✉1] 中央媒体技术研究院，华为[2] 计算机科学与工程学院，北京航空航天大学
‡ 项目负责人，✉ 通讯作者

摘要

尽管 *Veo3* 和 *Sora2* 等最先进的音视频生成模型展现出卓越的能力，但其闭源性质使其架构和训练范式难以获取。为弥合可访问性与性能之间的差距，我们推出 *UniTalking*，这是一个统一的端到端扩散框架，可生成高保真语音和唇形同步视频。该框架的核心是采用多模态 *Transformer* 块，通过共享的自注意力机制，显式建模音频与视频潜在标记之间细粒度的时间对应关系。借助预训练视频生成模型的强大先验，该框架既能确保达到最先进水平的视觉保真度，又支持高效训练。此外，*UniTalking* 还具备个性化语音克隆能力，可根据简短的音频参考生成具有目标风格的语音。定性和定量结果表明，我们的方法能够生成高度逼真的说话人像，在唇形同步准确度、音频自然度和整体感知质量方面均优于现有开源方法。



图1. 我们用于说话人像生成的统一音视频框架示意图。UniTalking通过多种输入模态促进同步音频和视频的生成。

1. 引言

人工智能生成内容（AIGC）的蓬勃发展，很大程度上得益于扩散模型的进步，使高保真图像、视频和音频合成取得了显著成功。早期研究主要聚焦于单模态生成，而近来研究前沿已转向音频转视频和视频转音频合成等跨模态任务。然而，一个关键且尚未得到充分探索的领域，是对彼此交织的多种模态进行统一生成。在现实应用中，视频和音频并非仅仅按顺序出现，而是同步呈现，共同构成连贯的感知体验。这凸显了现有方法中的一个重大空白：它们往往无法有效地对这种联合分布进行建模。

现有的联合音视频生成方法揭示了大规模专有模型与可及的学术研究之间的显著分歧。近来，尚未发布的闭源系统，如谷歌的Veo3和OpenAI的Sora2，已展现出生成高度连贯视听内容（包括可信的说话人像）的最先进水平能力。然而，这些模型无法获取，其架构、训练数据和方法均未公开，这对学术进展与可复现性构成了关键障碍。因此，更广泛的研究社区仍不得不应对存在根本性局限的现有开源范式。

在公众可访问的领域中，这些方法大致可分为两种范式。第一种采用级联两阶段流程，先生成一种模态（例如音频），再将其作为条

*同等贡献
†工作完成于华为实习期间。

件信号，用于驱动另一种模态（例如视频）的合成。尽管流程直观，但这种管线往往难以解决时间错位和误差累积问题，导致整体连贯性明显不足。第二种范式旨在通过端到端模型同时直接生成两种模态。然而，这类现有工作的一个显著局限是，主要聚焦于将拟音与视频同步，例如生成海浪拍岸的画面及其对应的轰鸣声。这些方法无法满足语音等复杂任务所需的精度。因此，开发一个开放且可复现的框架，以生成具备帧级、音素精准唇部同步的说话人像，仍是一项艰巨而迫切的挑战。这一局限严重制约了人工智能生成内容在自动化影视配音、逼真虚拟化身和交互式数字人等高影响力领域的应用。

为弥合这一关键差距，我们提出了 UniTalking，这是一种用于端到端统一生成说话肖像的新型框架。如图1所示，UniTalking 支持统一音视频说话肖像的生成，并能基于参考图像和音频制作个性化内容。我们框架的核心是多模态 Transformer 块，它能够在潜在空间中实现音频与视觉模态之间的细粒度对齐。通过采用共享自注意力机制，该模块能够显式建模视素（视觉）与音素（听觉）之间的时间对应关系，确保生成的唇部运动与合成语音精准同步。此外，模块内与文本条件进行交叉注意力，确保生成内容严格遵循输入文本的语义和歌词信息，从而实现同步且语义准确的生成。总之，我们的主要贡献如下：

- 我们提出了 UniTalking，这是一种新型端到端框架，采用对称双流架构，将说话人脸视频与相应的语音音频生成统一起来，从而应对连贯音视频合成中的一项重大挑战。
- 我们提出了一种新颖的架构设计：多模态 Transformer 模块中的联合注意力机制作用于拼接后的音视频词元，从而实现精细的时间对齐，并提升唇形同步准确度。
- 我们的框架采用多模态条件方案，能够合成说话人像，在保留图像中特定视觉身份的同时，模仿参考音频片段中的发声风格。
- 我们为开源统一说话人像生成树立了新的最先进水平，在 Lip-Sync、感知质量以及身份/风格保真度方面均展现出卓越性能。

2. 相关工作

我们将本研究置于近期音视频生成进展的背景下。我们将既有研究分为三个主要领域：音频到视频（A2V）合成、视频到音频（V2A）合成，以及新兴的统一音视频生成领域。这些领域中的许多最先进的方法都采用潜在扩散模型（LDM） [25]，通常以扩散Transformer（DiT） [23]，为骨干网络，并使用流匹配 [19]等目标进行训练。

2.1. 音频到视频肖像生成

音频驱动的说话人像生成近年来日益受到关注，这主要得益于基础视频模型 [13, 28, 36]的进步。Hallo3 [6]率先探索了这一方向，通过为 CogVideoX [36]引入专门的注意力机制，实现音视频同步和身份保持。类似地，混元视频头像 [2]基于混元视频 [13]，构建，通过计算音频特征与面部区域之间的交叉注意力，实现可控表情。大量研究利用了强大的 Wan 模型 [28]。例如，OmniHuman [18] 支持对表情强度进行控制，而 MultiTalk [14]则通过设计特定的注意力图来合成多人对话。这些级联式方法虽然有效，但生成的视频均以预先存在的音频为条件。

2.2. 视频到音频生成

视频到音频（V2A）生成提供了一种互补方法，即利用静音视频合成相应的音频。该领域分为拟音合成和语音生成两类。

对于拟音，早期方法如 Diff-Foley [22] 利用对比预训练来对齐音视频特征。近期模型，如 FoleyCrafter [39]，MMAudio [4]，和 Kling-Foley [30]，，采用了 MM-DiT 框架，通过对多种模态进行联合注意力计算，实现了最先进水平的效果。对于视频到语音（V2S），Faces2Voices [12] 采用三阶段流水线来建模内容、音色和韵律。DeepAudio [38]采用模块化多阶段方法，而 AudioGen-Omni [31]则扩展了 Kling-Foley [30] 架构以纳入歌词，实现了高质量的语音合成。

2.3. 统一音视频生成

统一生成是在单个模型中同时处理两种模态的新兴且关键的研究方向。JavisDiT [20] 采用双分支 DiT，并通过多模态双向交叉注意力模块实现交互。Universe-1 [29]采用专家拼接（SoE）方法，整合 Wan2.1 [28]和 ACE-Step [9]等预训练单模态模型。其他工作，如 Ovi [21] 和 OmniTalker [32]，也采用双流

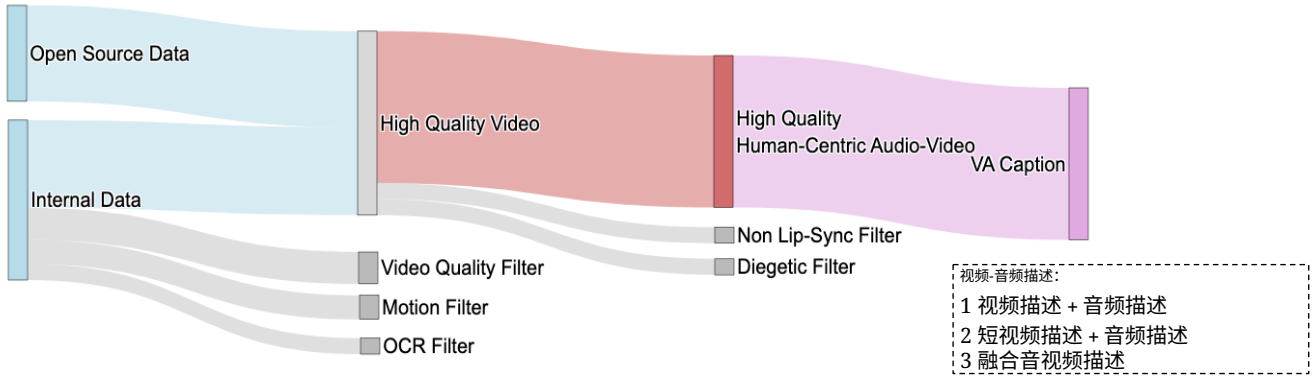


图2。构建以人为中心的音视频数据集的数据处理流程。该流程采用顺序过滤，首先进行单模态（视频和音频）检查，随后进行跨模态过滤。最后，为高质量的音视频对添加多层级多模态描述。

架构，并引入专用融合模块或交叉注意力层来实现模态交互。

UniTalking 与这些方法有所不同。尽管我们也采用双流架构，但我们的对称双分支设计以强大的视频基础模型为初始化，专为实现无缝的模态对齐而设计。此外，不同于以往主要专注于文本驱动生成的工作，我们的框架独特地整合了视觉身份和声学风格的多模态条件输入，使统一的端到端模型能够实现更高程度的个性化。

3. 数据准备

我们的数据集主要来自两个来源：开源的 OpenHumanVid 语料库 [16] 以及一个大规模内部数据集。我们对这一初始语料库采用严格的多阶段过滤流程，以筛选出高质量、以人为中心的语音内容（见图 2）。如前所述，该流程依次应用视频、音频和音视频过滤模块，以去除不理想的内容（例如静态视频、低信噪比音频或较差的唇形同步）。完成筛选后，我们为每对经过过滤的音视频生成多层级文本标注。此标注过程会生成不同粒度的描述。除生成描述外，我们还为每段视频生成参考音频，以便用于训练参考音频到音视频的生成。最终筛选得到的以人为中心的数据集包含 230 万个对齐的音视频样本。

3.1. 数据处理流程

视频-音频过滤。 我们采用一个由三个阶段组成的顺序过滤流程来筛选数据集：

• 视频过滤：此初始流程处理视觉流，旨在过滤掉静态、包含文字叠加层或整体视觉质量较低的视频。

• 音频过滤：第二个流程分析音频流，以移除静音、缺少语音或信噪比（SNR）较低的样本。在此过程中，我们使用 PANNs 和 SentenceASD 模型执行语音事件检测。

• 音视频过滤：最终流程评估跨模态关系。该流程用于过滤仅包含场景内音频的视频（使用 LightASD），以及唇形同步对齐效果不佳的样本（使用 Lip-Sync）。

视频-音频联合字幕生成。我们生成三种不同的文本描述，用于训练联合音视频模型。前两种格式是拼接描述，即分别生成模态专属描述，再将其组合起来。具体包括：（a）详细的视频描述加音频描述；（b）简短精炼的视频描述加音频描述。这些独立描述由一组描述生成模型生成，包括 Qwen3-VL[35], Whisper-V3[24], 和 Qwen3-Omni- Captioner[34]。第三种格式是融合描述（c），我们将视频和音频流直接输入 Qwen3-Omni[34]，生成一段统一的视听描述。训练期间，我们随机抽取三种提示词中的一种作为输入。

3.2. 参考数据生成

除了标准的文本到视频音频（T2VA）和图像到视频音频（TI2VA）生成外，我们的模型还支持以人声音频参考为条件进行合成，从而控制输出音色。由于源数据集缺少相应的参考音频，我们合成这些样本以创建必要的训练对。

我们在此任务中采用 IndexTTS2 [42]模型。对于每个视频，我们通过向 IndexTTS2 提供两个输入来生成新的音频片段：(1) 视频中的原始音频，作为目标音色参考，以及

(2) 随机生成的文本提示。这些提示由随机采样的词语和数字序列构成。我们设置了时长限制，确保每个合成的参考音频片段时长在 3 到 5 秒之间。为了扩充数据，我们重复这一过程，为每个视频生成三个不同的参考音频样本。

4. 方法

本节首先介绍我们工作所基于的生成建模基本原理。随后，我们将介绍 UniTalking 的整体架构，详细说明其多模态处理的核心组件，并概述旨在实现高效且稳定收敛的渐进式训练策略。

4.1. 预备知识

流匹配。 我们的框架基于流匹配 [19], , 这是一种无需仿真即可训练连续归一化流（CNF）的方法。CNF 学习一个依赖时间的向量场 $v_\theta(x_t, t)$，它定义了一个概率流常微分方程，$\mathrm{d}x_t/\mathrm{d}t = v_\theta(x_t, t)$，将简单的先验样本$x_0 \sim \mathcal{N}(0, \mathrm{I})$在 $t = 0$变换为数据样本 x_1在 $t = 1$。模型直接使用条件流匹配（CFM）目标进行训练，该目标沿着简单的线性路径 $x_t = (1 - t)x_0 + tx_1$回归至目标向量$x_1 - x_0$：

$$\mathcal{L}_{CFM} = \mathbb{E}_{t,q(x_1),p_0(x_0)} \left[\| v_\theta(x_t - t) - (x_1 - x_0) \|^2 \right]. \tag{1}$$

通过采样 x_0 并使用数值求解器求解学习到的 ODE 来完成生成。在所提出的框架中，独立音频生成任务的目标函数定义为 $L_{total} = L^a_{CFM}$，而其他联合生成任务的目标函数则定义为 $L_{total} = L^a_{CFM} + L^v_{CFM}$。

无分类器引导。 我们采用无分类器引导（CFG） [11]来控制合成。向量场$v_\theta(x_t, t, c)$在条件 c下进行训练，该条件会被随机替换为空标记 $\emptyset$。在推理时，引导后的向量场 $\hat{}v_\theta$ 通过外推得到：

$$\hat{v}_\theta(x_t, t, c) = v_\theta(x_t, t, \emptyset) + \omega \cdot (v_\theta(x_t, t, c) - v_\theta(x_t, t, \emptyset)), \tag{2}$$

其中 $\omega > 1$为引导尺度。在 UniTalking 中，条件 c 可由文本提示（c_{text}）、身份图像（c_{image}）和参考音频（c_{audio}）组合构成，从而支持多种任务，包括文本到音视频（T2AV）、文本图像到音视频（TI2AV）以及文本参考到音视频（TR2AV）。

4.2. 整体架构

UniTalking的架构如图3所示，基于多模态扩散变换器（MM-DiT） [8]。整个框架采用流匹配目标进行训练，作为连续归一化流，并通过无分类器引导进行引导，详情见第4.1节。模型的核心是一个变换器主干网络，由 N 个MM-DiT模块组成，这些模块处理由潜在视频标记和音频标记拼接而成的序列。每个模块都采用对称的双流设计来处理这些标记，以促进模型学习共享的时间动态。变换器的输出是去噪后的潜在表示，随后将其解码为连贯的视频帧和音频波形。

为充分利用大规模预训练的优势，我们模型的视频流继承了 Wan2.2-5B 的架构和权重 [28], , 为视觉合成提供了强大的先验。至关重要的是，为弥合不同模态之间的表示差距，音频流被设计为与视频流完全相同的孪生分支。这种架构对称性有助于实现无缝的潜空间融合。音频分支的参数采用随机初始化，并经过训练以匹配其预训练视觉对应分支的能力。

4.3. 潜在表示

为确保计算效率，并使学习聚焦于核心生成任务，所有输入模态均由专用的预训练编码器投影到紧凑的潜在空间中，这些编码器在整个训练过程中保持冻结。

视频与音频潜变量。 视频帧由来自 Wan2.2 的 3D因果变分自编码器 [33] 编码，实现了高效的时空压缩 $16 \times 16 \times 4$。对于音频，我们采用来自 MMAudio 的 1D变分自编码器 [4]。原始波形经短时傅里叶变换（STFT） [26] 转换为梅尔频谱图，然后进行编码。推理过程中，去噪后的音频潜变量被解码回梅尔频谱图，并通过 BigVGAN声码器 [15]合成为 44.1kHz 波形。

条件潜变量。 模型由多模态条件引导。文本提示使用冻结的 UMT5 模型进行嵌入 [5], , 而参考音频片段则由同一个冻结的 MMAudio VAE 编码器编码。为确保训练稳定性，文本和参考音频潜变量的序列长度分别固定为 512 和 257。较长的序列会被截断，较短的序列则用零填充。这种冻结编码器设置为多模态 DiT 模块提供了稳定且语义丰富的潜在空间。

4.4. 多模态 Transformer 块

为有效整合多模态条件并强化时间连贯性，我们对标准 DiT 模块进行了三项改进，重点针对其注意力机制和位置编码。

首先，标准自注意力被替换为联合-

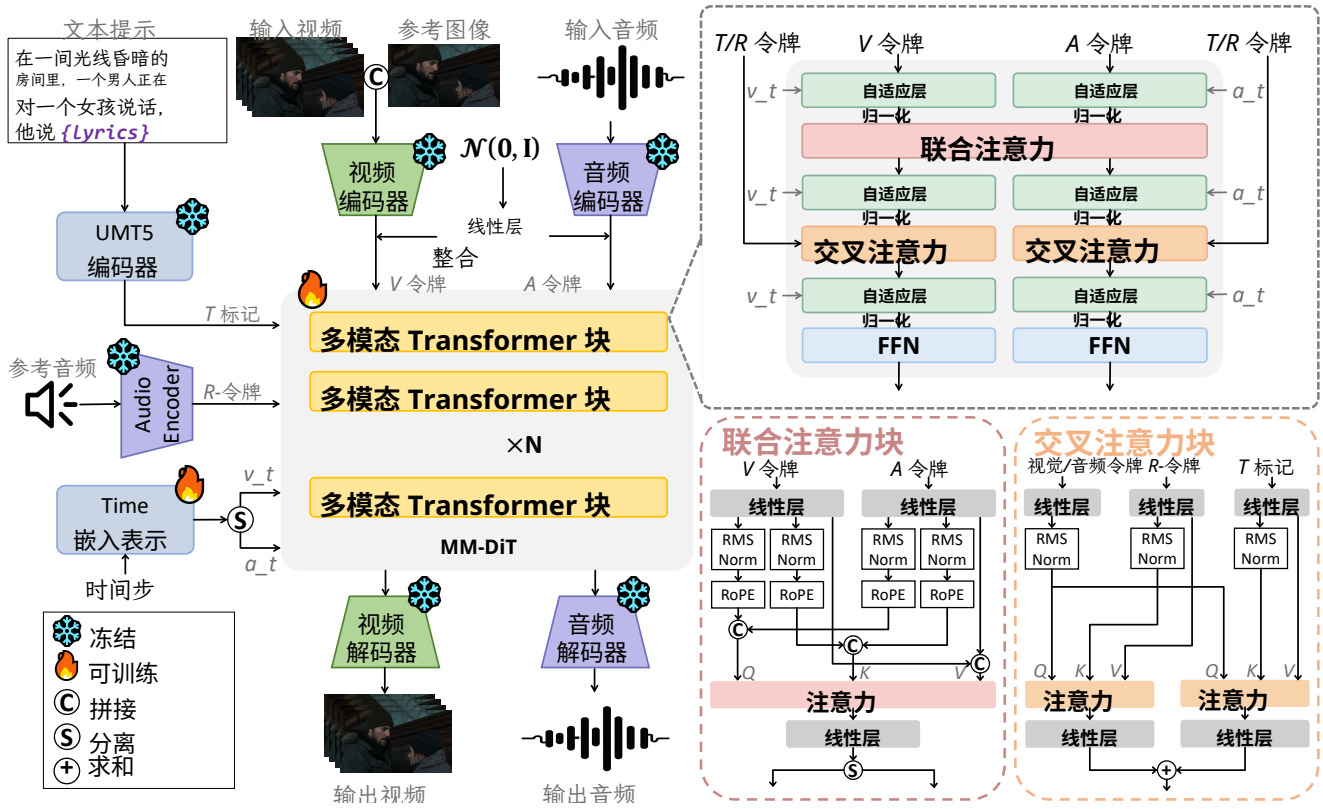


图3. UniTalking 的架构。我们的框架使用双流多模态 DiT（MM-DiT）骨干网络，为说话人像联合生成同步的音频和视频；该网络通过流匹配训练为连续归一化流。模型的核心是多模态 Transformer 块，详见右侧。该模块对拼接后的音视频令牌应用联合注意力，以确保精确的时间对齐；并应用交叉注意力来融入文本和声学风格等多模态条件。冻结的组件以雪花标记，可训练的部分则以火焰标记。

注意力机制。来自视频与音频流的潜在标记被拼接在一起，迫使模型在单次注意力操作中学习模态内与模态间依赖关系，以促进整体的音视频理解。

其次，为了联合以文本（c_{text}）和参考音频风格（c_{audio}）为条件，我们扩展了交叉注意力模块。我们为参考音频条件增加了一组辅助键值投影层。主流中的潜在标记分别关注各个条件，并通过逐元素求和融合其输出，从而使生成结果既忠实于文本，又与参考音频的风格保持一致。

第三，我们采用一种专门设计的旋转位置编码（RoPE）[27] 策略。遵循 OVI [21], 的做法，我们沿时间（t）轴应用标准 RoPE。然而，对于音频标记的空间（h， w）维度，我们使用基于单一固定位置的 RoPE。这种各向异性设计促使模型优先关注时间动态，从而增强音频流与视频流之间的对齐。更多细节见附录。

4.5. 训练策略

由于预训练视频分支与随机初始化的音频分支之间存在初始化不平衡，我们提出了一种渐进式两阶段训练策略，以实现稳定收敛。

音频分支预训练。 我们的训练流程首先在文本转语音（TTS）任务上单独训练音频分支。我们在复合文本转语音任务上进行训练，同时优化有音色参考和无音色参考的合成。为保留其他模态的表征，我们采用参数高效方法：冻结与文本分支和视频分支相关的所有参数（包括其前馈神经网络和注意力投影层）。仅微调音频输入投影层，以及音频分支的前馈神经网络和注意力投影层。我们发现，仅微调这些参数就足以生成优质语音。这一阶段促使音频分支学习从语义到声学的有效映射，使其表征能力与文本分支对齐。我们的实验发现，这一阶段对于最终音视频生成的生成音频质量至关重要，

至关重要。

多任务联合音视频训练。 在具备强大音频分支的情况下，我们对整个 UniTalking 框架进行端到端训练。此处的目标是学习音频与视频之间复杂的时间相关性。我们采用多任务学习目标，在文本到音视频（T2AV）、文本视频到音频（TV2A）、文本图像到音视频（TI2AV）和文本参考到音视频（TR2AV）等任务之间交替，以迫使模型学习各模态之间的双向关系，从而实现精确的唇部同步。我们认为，这种交替会形成强大的协同学习动态，迫使模型学习视听语音的整体且解耦的表示。

T2AV 是核心的联合生成任务，模型仅根据文本提示同时合成音频和视频。它建立了粗粒度对齐，因为生成内容必须与共享的底层脚本相对应。TV2A 为时间对齐提供了严格的单向监督信号。在该任务中，我们引入了从音频标记到视频令牌的注意力掩码，从而防止音频影响视频分支，并使模型能够基于真实视频的潜在特征进行音频预测。该任务迫使音频分支学习细微的逐帧视觉动作所产生的精确声学结果，直接教会模型从视素到音素的细粒度映射。而 TI2AV 和 TR2AV 则赋予模型生成个性化内容的能力。

通过交替执行这四项任务，我们构建了一组多维约束，防止模型找到过于简单的解。这一全面的策略迫使模型学习同步、可控且个性化的视听语音所遵循的真实底层流形，从而得到稳健且通用的最终模型。

5. 实验

5.1. 实现细节

按照多阶段训练策略，音频分支预训练的初始阶段使用批量大小为 256、学习率为 1×10^{-5}，在内部文本转语音数据上共训练 100,000 步。第二阶段的视频音频联合训练使用批量大小为 64、学习率为 1×10^{-5}，训练 100,000 步。训练过程采用 AdamW优化器进行优化，其中 $\beta_1 = 0.9$，$\beta_2 = 0.999$，$\epsilon = 10^{-8}$。训练调度基于 UniPC [40] 实现。为实现高效训练，所有训练均须使用`bf16`精度，并采用全分片数据并行（FSDP） [41]。

由于该模型基于 Wan2.2，因此两种结构的框架细节高度一致。例如，MM-DiT模块的数量为 $N = 30$，模型维度配置为 dim =3072，注意力采用 24 个头。此外，整个模型的参数总数为 $10B$。

5.2. 结果可视化

音视频同步。 统一音视生成领域的一项基本挑战，是这两种模态之间的时间对齐。为展示模型生成结果的对齐质量，图4展示了随机选取样本的视频帧和音频梅尔频谱图。通过比较视频帧与梅尔频谱图的变化可以发现，生成视频中的说话人像呈现出准确的唇音同步。此外，将 UniTalking 生成音频的梅尔频谱图与真实值进行比较可见，两者呈现出高度相似的模式，表明生成音频非常准确。文本提示”没有任何背景音乐”生成的视频没有背景噪声，因此梅尔频谱图能够更清晰地突出人声成分。文本提示”一声短促的笑声” 也与样本 1 的最后一帧视频相对应，这表明模型在生成结果时能够遵循输入文本。**交叉注意力可视化**。 为进一步展示 UniTalking 中视频与音频模态之间的相关性，图 5展示了不同层中联合注意力模块对视频标记与音频标记之间的注意力图。注意力图揭示了视频与音频模态之间存在明显的依赖关系，语义相关区域获得了更高的权重。音频到视频注意力图表明，面部和身体区域受益于音频标记。相反，视频到音频注意力图显示，音频标记仅受益于嘴唇区域。不同训练迭代次数的模型呈现出一致的相关模式。此外，注意力图中一些不匹配的高亮区域会随着训练迭代次数的变化而随机改变。据推测，这是训练策略或数据噪声导致的模型缺陷所致。

5.3. 实验设置

T2AV联合生成。 对于音频-视频联合生成，我们将其与近期最先进的方法 Universe-1 [29]和 OVI [21]进行比较。我们针对每组模型比较开展了一项盲式两两偏好研究，共有 20 名参与者。测试集包含 50 条用于说话肖像生成的测试提示词。

TR2AV联合生成。 对于TR2AV任务，最关键的评估指标是生成内容与参考内容之间的一致性。在本研究中，实验

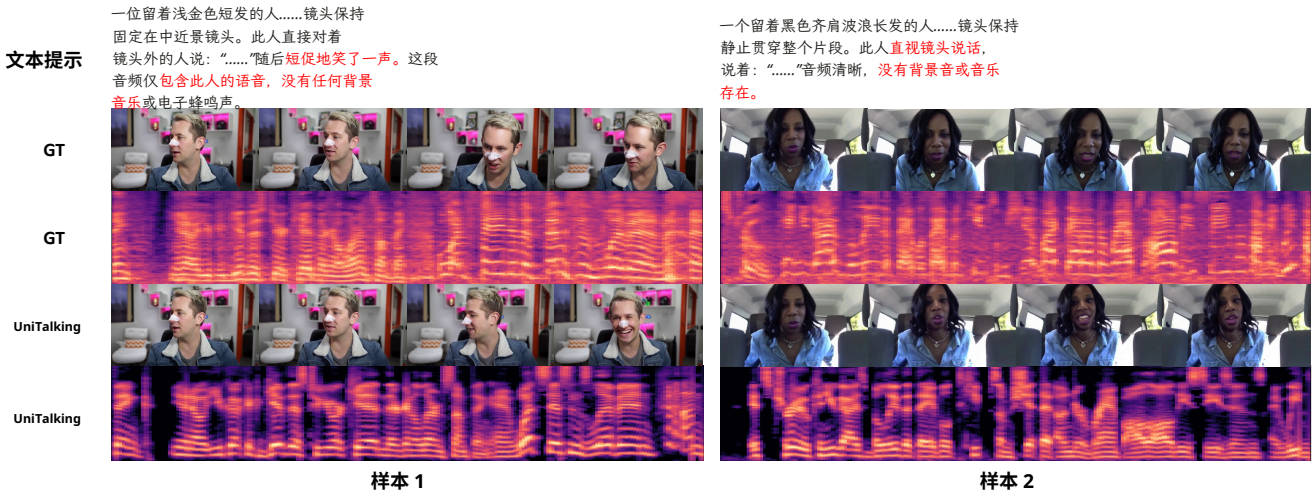


图4. 生成的音视频说话人肖像和真实音视频数据的视频帧与梅尔频谱图可视化。顶部是对应的文本提示，其中某些关键词以 **粗体**和**红色**突出显示。生成的视频和音频在语义上与真实值一致。



（a）样本 1



（b）样本 2

图5. 联合注意力中视频令牌与音频标记之间注意力图的可视化。子图(a)和(b)展示了两个不同的随机样本。子图中的第一行表示音频到视频注意力，第二行表示视频到音频注意力。不同列对应于从不同推理采样步骤和不同 Transformer 块中得到的权重。

设置采用Qwen3-Omni [34]，并利用MiniMax 多语言测试集 [37] 来评估生成音频与参考音频之间的说话人相似度。

文本转语音生成。对于文本转语音生成，我们将模型与单模态文本转语音模型和 OVI [21] 进行比较。我们遵循 OVI 中的文本转语音评估协议，并在模型上使用 Seed-TTS test-en 数据集 [1]进行测试。词错误率（WER）用作评估指标。

5.4. 结果

T2AV 上的盲测偏好研究。我们对 T2AV 生成进行主观评估，使用 OVI [21] asthe 基线模型，并计算所有模型在视频质量、音频质量和音视频同

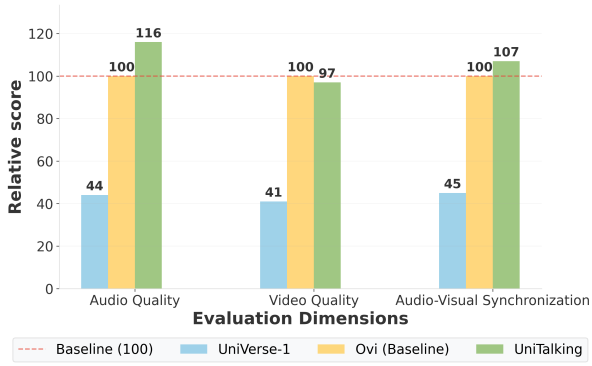


图6. 三项指标的盲测偏好研究结果：视频质量、音频质量和音视频同步。OVI 的结果以基线 100 进行归一化，其他方法则报告相对得分。

步方面的胜负率。胜负率的计算方式为$rate = \frac{win+tie}{lose+tie}$。我们构建了一个包含 50 条说话测试提示词的测试集。如图6所示，我们的方法在音频质量和音视频同步方面表现最佳。与 OVI 相比，所提出的方法在音频质量和音视频同步方面分别达到 116% 和 107%。在视频质量方面，我们的模型与 OVI 不相上下。由于 OVI 和我们的方法都使用了预训练 Wan 2.2 模型，这一表现相对合理。

我们还使用 Sync-C 和 Sync-D 唇形同步指标进行了客观比较。如表1所示，我们的方法在 Sync-C 上比 Universe-1 高出 3.02 分，性能与 Sora2 模型不相上下。在 Sync-D 上，我们的方法分别以 3.92 分和 0.55 分的优势超过 Universe-1 和 OVI。在这项分析中，我们观察到 OVI 模型获得了显著较高的 Sync-C 得分。我们推测这可能是指标存在潜在偏差所致，因为 Sync-C 似乎更偏好嘴部动作夸张的生成对象，而非发音动作更自然的对象。

表1. T2AV 上的 Lip-Sync 评估结果

方法	Sync-C ↑	Sync-D ↓
Universe-1 [29]	1.85	11.97
OVI [21]	6.56	8.6
Sora2	5.35	7.78
UNITALKING（我们的方法）	4.87	8.05

与参考音频的相似度评估。表2 展示了音色相似度对比实验的结果。与 Qwen3-Omni 一致 [34]，，我们选择了 Eleven- Labs、MiniMax [37], 和 Qwen3-Omni [34] 作为对比方法。可以看出，我们的方法具备一定程度的语音克隆能力。尽管难以超越最先进大型音频模型的性能，UniTalking 仍实现了较高的语音相似度

表2. TR2AV 上说话人相似度的结果。UniTalking 的文本提示补充了视频描述。

方法	说话人相似度 ↑ 英语	中文
ElevenLabs	0.613	0.677
MiniMax [37]	0.756	0.780
Qwen3-Omni [34]	0.773	0.772
UNITALKING（我们的方法）	0.703	0.662

可与 ElevenLabs 相媲美。

文本转语音生成评估。遵循 OVI 的评估设置[21]，，我们评估了第一阶段模型的文本转语音（TTS）生成能力。如表3所示，我们的模型达到了与最先进（SOTA）方法相当的性能，证明了音频训练流程在生成高保真语音方面的有效性。此外，消融研究表明，省略第一阶段训练会导致最终音视频模型生成的音频质量显著下降。我们推测，这是由于模态收敛速率不平衡所致：视频分支受益于在大规模数据上的广泛预训练，而音频分支的训练则相对不足。

表 3. 文本转语音评估结果。我们遵循 OVI [21] 的评估协议。

Type	方法	词错误率 ↓
TTS	Fish Speech [17]	**0.008**
	F5-TTS [3]	0.018
	CosyVoice [7]	0.034
	FireRedTTS [10]	0.038
统一	Ovi-Aud [21]	0.035
	UNITALKING（我们的方法）	0.038

6. 结论

在这项工作中，我们提出了 UniTalking，这是一种支持 T2VA、TI2VA 和 TR2VA 的多任务音视频生成统一框架。我们采用统一的注意力机制，在共享的潜在空间中直接对齐文本、视频和音频模态。这种方法不同于依赖交叉注意力的传统双主干网络。实验结果证明了我们方法的有效性。尽管该框架是在会说话的人像生成任务上验证的，但我们认为 UniTalking 框架广泛适用于通用音视频合成，包括音效和音乐。

未来工作我们的模型性能受到可用训练资源和数据规模的限制，尤

其是与闭源模型相比时。此外，该框架尚不支持多人参考生成，例如 Sora2 中的“Cameo”，它利用多个音视频片段作为输入。我们将这些挑战留待未来工作。

参考文献

[1] Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao 等。Seed-tts：一系列高质量、多用途的语音生成模型。arXiv 预印本 arXiv:2406.02430，2024年。7[2] Yi Chen, Sen Liang, Zixiang Zhou, Ziyao Huang, Yifeng Ma, Junshu Tang, Qin Lin, Yuan Zhou 和 Qinglin Lu。混元视频头像：面向多个角色的高保真音频驱动人物动画。arXiv 预印本 arXiv:2505.20156，2025年。2[3] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chun-hui Wang, JianZhao JianZhao, Kai Yu 和 Xie Chen。F5-tts：借助流匹配生成流畅且忠实语音的讲述者。在第63届计算语言学协会年会论文集（第1卷：长篇论文）中，第6255–6271页，2025年。8[4] Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing 和 Yuki Mitsufuji。MMAudio：驾驭多模态联合训练，实现高质量视频到音频合成。在计算机视觉与模式识别会议上，第28901–28911页，2025年。2、4[5] Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang 和 Orhan Firat。Unimax：面向大规模多语言预训练的更公平、更有效的语言采样。arXiv 预印本 arXiv:2304 09151，2023年 4 。 。[6] Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang 和 Siyu Zhu。Hallo3：借助视频扩散 Transformer实现高度动态且逼真的肖像图像动画。在计算机视觉与模式识别会议上，第21086–21095页，2025年。2[7] Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma 等。CosyVoice：基于监督语义标记的可扩展多语言零样本文本转语音合成器。arXiv 预印本 arXiv:2407.05407，2024年。8[8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas M¨uller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel 等。扩展用于高分辨率图像合成的整流流Transformer。在国际机器学习会议上，2024年。4[9] Junmin Gong, Sean Zhao, Sen Wang, Shengyuan Xu 和 Joe Guo。ACE-Step：迈向音乐生成基础模型。arXiv 预印本 arXiv:2506.00045，2025年。2[10] Hao-Han Guo, Yao Hu, Kun Liu, Fei-Yu Shen, Xu Tang, Yi-Chen Wu, Feng-Long Xie, Kun Xie 和 Kai-Tuo Xu。FireRedTTS：面向行业级生成式语音应用的基础文本转语音框架。arXiv 预印本 arXiv:2409.03283，2024年。8[11] Jonathan Ho 和 Tim Salimans。无分类器扩散引导。arXiv 预印本 arXiv:2207.12598，2022年。4[12] Ji-Hoon Kim、Jeongsoo Choi、Jaehun Kim、Chaeyoung Jung 和 Joon Son Chung。从面孔到声音：学习用于高质量视频到语音的层次化表征。载于计算机视觉与模式识别会议，第 15874–15884 页，2025年。2[13] Weijie Kong、Qi Tian、Zijian Zhang、Rox Min、Zuozhuo Dai、Jin Zhou、Jiangfeng Xiong、Xin Li、Bo Wu、Jianwei Zhang 等。混元视频：大型视频生成模型的系统化框架。arXiv 预印本 arXiv:2412.03603，2024年。2[14] Zhe Kong、Feng Gao、Yong Zhang、Zhuoliang Kang、Xi-aoming Wei、Xunliang Cai、Guanying Chen 和 Wenhan Luo。让他们说话：音频驱动的多人对话视频生成。arXiv 预印本 arXiv:2505 22647，2025 2 . . [15] Sang-gil Lee、Wei Ping、Boris Ginsburg、Bryan Catanzaro 和 Sungroh Yoon。BigVGAN：一种经过大规模训练的通用神经声码器。arXiv 预印本 arXiv:2206 04658，2022 4 . . [16] Hui Li、Mingwang Xu、Yun Zhan、Shan Mu、Jiaye Li、Kaihui Cheng、Yuxuan Chen、Tan Chen、Mao Ye、Jingdong Wang 等。OpenHumanVid：用于提升以人为中心的视频生成的大规模高质量数据集。载于计算机视觉与模式识别会议，第 7752–7762 页，2025年。3[17] Shijia Liao、Yuxuan Wang、Tianyu Li、Yifan Cheng、Ruoyi Zhang、Rongzhi Zhou 和 Yijin Xing。Fish-Speech：利用大语言模型实现先进的多语言文本转语音合成。arXiv 预印本 arXiv:2411.01156，2024年。8[18] Gaojie Lin、Jianwen Jiang、Jiaqi Yang、Zerong Zheng、Chao Liang、Yuan Zhang 和 Jingtuo Liu。OmniHuman-1：重新思考单阶段条件人类动画模型的规模化。载于计算机视觉与模式识别会议，第 13847–13858 页，2025年。2[19] Yaron Lipman、Ricky TQ Chen、Heli Ben-Hamu、Maximil-ian Nickel 和 Matt Le。生成建模的流匹配。arXiv 预印本 arXiv:2210.02747，2022年。2, 4[20] Kai Liu、Wei Li、Lai Chen、Shengqiong Wu、Yanhao Zheng、Jiayi Ji、Fan Zhou、Rongxin Jiang、Jie bo Luo、Hao Fei 等。JAVISDiT：具有层次化时空先验同步的联合音视频扩散变换器。arXiv 预印本 arXiv:2503.23377，2025年。2[21] Chetwin Low、Weimin Wang 和 Calder Katyal。OVI：用于音视频生成的双骨干跨模态融合。arXiv 预印本 arXiv:2510.01284，2025年。2, 5, 6, 7, 8[22] Simian Luo、Chuanhao Yan、Chenxu Hu 和 Hang Zhao。Diff-Foley：基于潜在扩散模型的同步视频转音频合成。NeurIPS，36：48855–48876，2023年。2[23] William Peebles 和 Saining Xie。基于变换器的可扩展扩散模型。载于国际计算机视觉大会，第 4195–4205 页，2023年。2[24] Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever。通过大规模弱监督实现鲁棒语音识别。载于国际机器学习大会，第 28492–28518 页。机器学习研究论文集，2023年。3

[25]Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Bj¨orn Ommer。基于潜在扩散模型的高分辨率图像合成。载于计算机视觉与模式识别会议，第 10684–10695 页，2022年。2[26] Stanley Smith Stevens、John Volkmann 和 Edwin Broomell Newman。用于测量心理音高尺度的量表。《美国声学学会杂志》，8：185–190，1937年。4[27] Jianlin Su、Murtadha Ahmed、Yu Lu、Shengfeng Pan、Wen Bo 和 Yunfeng Liu。RoFormer：采用旋转位置嵌入的增强型 Transformer。《神经计算》，568 127063，2024 5 : . [28] Team Wan、Ang Wang、Baole Ai、Bin Wen、Chaojie Mao、Chen-Wei Xie、Di Chen、Feiwu Yu、Haiming Zhao、Jianxiao Yang 等。Wan：开放且先进的大规模视频生成模型。arXiv 预印本 arXiv:2503.20314，2025年。2, 4[29] Duomin Wang、Wei Zuo、Aojie Li、Ling-Hao Chen、Xinyao Liao、Deyu Zhou、Zixin Yin、Xili Dai、Daxin Jiang 和 Gang Yu。Universe-1：通过专家拼接实现统一音视频生成。arXiv 预印本 arXiv:2509.06155，2025年。2, 6, 8[30] Jun Wang、Xijuan Zeng、Chunyu Qiang、Ruilong Chen、Shiyao Wang、Le Wang、Wangjing Zhou、Pengfei Cai、Ji-ahui Zhao、Nan Li 等。Kling-Foley：用于高质量视频到音频生成的多模态扩散 Transformer。arXiv 预印本 arXiv:2506.19774，2025年。2[31] Le Wang、Jun Wang、Chunyu Qiang、Feng Deng、Chen Zhang、Di Zhang 和 Kun Gai。AudioGen-Omni：用于视频同步音频、语音和歌曲生成的统一多模态扩散Transformer。arXiv 预印本 arXiv:2508.00733，2025年。2[32] Zhongjian Wang、Peng Zhang、Jinwei Qi、Guangyuan Wang、Chaonan Ji、Sheng Xu、Bang Zhang 和 Liefeng Bo。Omnitalker：通过多模态风格模仿实现单次、实时、文本驱动的说话音视频生成。arXiv 预印本 arXiv:2504.02433，2025年。2[33] Pingyu Wu、Kai Z hu、Yu Liu、Liming Zhao、Wei Zhai、Yang Cao 和 Zheng-Jun Zha。改进用于潜空间视频扩散模型的视频变分自编码器。载于计算机视觉与模式识别会议，第 18124–18133 页，2025年。4[34] Jin Xu、Zhifang Guo、Hangrui Hu、Yunfei Chu、Xiong Wang、Jinzheng He、Yuxuan Wang、Xian Shi、Ting He、Xinfa Zhu 等。Qwen3-Omni 技术报告。arXiv 预印本 arXiv:2509.17765，2025年。3, 7, 8[35]An Yang、Anfeng Li、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、Chengen Huang、Chenxu Lv 等。Qwen3 技术报告。arXiv 预印本 arXiv:2505.09388，2025年。3[36] Zhuoyi Yang、Jiayan Teng、Wendi Zheng、Ming Ding、Shiyu Huang、Jiazheng Xu、Yuanming Yang、Wenyi Hong、Xiao-han Zhang、Guanyu Feng 等。CogVideoX：采用专家Transformer的文生视频扩散模型。arXiv 预印本 arXiv:2408.06072，2024年。2[37]Bowen Zhang、Congchao Guo、Geng Yang、Hang Yu、Haozhe Zhang、Heidi Lei、Jialong Mai、Junjie Yan、Kaiyue Yang、Mingqi Yang 等。MiniMax-Speech：采用可学习说话人编码器的内在零样本文本转语音。arXiv 预印本 arXiv:2505.07916，2025年。7, 8

[38] Haomin Zhang、Chang Liu、Junjie Zheng、Zihao Chen、Chaofan Ding 和 Xinhan Di。Deepaudio-v1：迈向多模态、多阶段的端到端视频转语音和音频生成。*arXiv* 预印本 *arXiv:2503.22265*，2025年。2[39] Yiming Zhang、Yicheng Gu、Yanhong Zeng、Zhening Xing、Yuancheng Wang、Zhizheng Wu 和 Kai Chen。Foley-crafter：让无声视频焕发生机，呈现逼真且同步的声音。*arXiv* 预印本 *arXiv:2407.01494*，2024年。2[40] Wenliang Zhao、Lujia Bai、Yongming Rao、Jie Zhou 和 Jiwen Lu。UniPC：用于扩散模型快速采样的统一预测-校正框架。*NeurIPS*，36:49842–49869，2023年。6[41] Yanli Zhao、Andrew Gu、Rohan Varma、Liang Luo、Chien-Chin Huang、Min Xu、Less Wright、Hamid Shojanazeri、Myle Ott、Sam Shleifer 等。PyTorch FSDP：扩展全分片数据并行的经验。*arXiv* 预印本 *arXiv:2304.11277*，2023年。6[42] Siyi Zhou、Yiquan Zhou、Yi He、Xun Zhou、Jinchao Wang、Wei Deng 和 Jingchen Shu。IndexTTS2：在富有情感表现力、时长可控的自回归零样本文本转语音方面取得突破。*arXiv* 预印本 *arXiv:2506.21619*，2025年。3