/TemplateVersion (2025.1)

HM-Talker: 用于高保真说话人头像合成的混合运动建模

Shiyu Liu[1], Kui Jiang[1], Xianming Liu[1], Hongxun Yao[1], Xiaocheng Feng[1]

[1]Harbin Institute of Technology University
liushiyu_aiia@stu.hit.edu.cn, { jiangkui, csxm, h.yao } @hit.edu.cn, xcfeng@ir.hit.edu.cn

Abstract

音频驱动的说话人视频生成在人机交互中增强了用户参与度。然而，目前的方法经常产生运动模糊和嘴唇抖动的视频，主要是因为它们依赖于对音频与面部运动相关性进行隐式建模——这种方法缺乏明确的发音先验（即与语音相关的面部运动的解剖性指导）。为了解决这一限制，我们提出了 HM-Talker，一种生成高保真、时间一致的说话人视频的新框架。HM-Talker 利用混合运动表示，结合隐式和显式运动线索。显式线索使用行动单元（AUs），解剖定义的面部肌肉运动，同时结合隐式特征以减少音素与视觉音位的不匹配。具体而言，我们的跨模态解缠模块（CMDM）从音频输入中直接提取 AUs，与视觉线索对齐，同时提取互补的隐式/显式运动特征。为减少显式特征中身份依赖偏差并增强跨个体泛化，我们引入了混合运动建模模块（HMMM）。该模块动态合并随机配对的隐式/显式特征，强制身份无关学习。这些组件结合起来，能够在不同身份间实现稳健的嘴唇同步，推动个性化说话人视频合成。广泛的实验表明，HM-Talker 在视觉质量和嘴唇同步准确性方面优于当前最先进的方法。

介绍

音频驱动的动画头像合成已成为多媒体技术的一个重要前沿，展示出在提高交互应用中的用户参与度方面的巨大潜力。其核心目标是在合成的过程中，使面部表情和嘴唇运动与输入音频信号精确同步，同时保持主题身份的完整性。

虽然基于 2D 的方法 (Prajwal et al. 2020; Zhong et al. 2023; Zhang et al. 2023b) 使用生成模型从单个图像生成会说话的肖像，但由于 3D 建模不足，它们经常产生机械伪影和解剖上不合理的动作。当代 3D 方法使用神经辐射场（NeRF） (Mildenhall et al. 2021) 或 3D 高斯喷射 (3DGS) (Kerbl et al. 2023) 显著提高了视觉质量和时间一致性。早期的方法采用动态 NeRF 与音频投影特征来解耦头部-躯干运动 (Guo et al. 2021)，而其他方法通过多分辨率哈希编码 (Tang et al. 2022) 增强空间-声学表示，或通过三平面编码 (Li et al. 2023) 减少哈希碰撞。然而，这些方法主要依赖于隐式表示来将音频映射到面部运动（Figure 1 (a)），在稳定结构和细粒度控制方面遇到困难。

最近的研究结合了显式先验（例如，3DMM）以实现可推广的动画合成 (Gong et al. 2025; Chu and Harada 2025)，但在应用于特定身份的谈话头生成时，由于可推广的中性统计先验（Figure 1 (b)），在重建精度上有所妥协。一些努力强调了精确的音频唇部对齐，使用

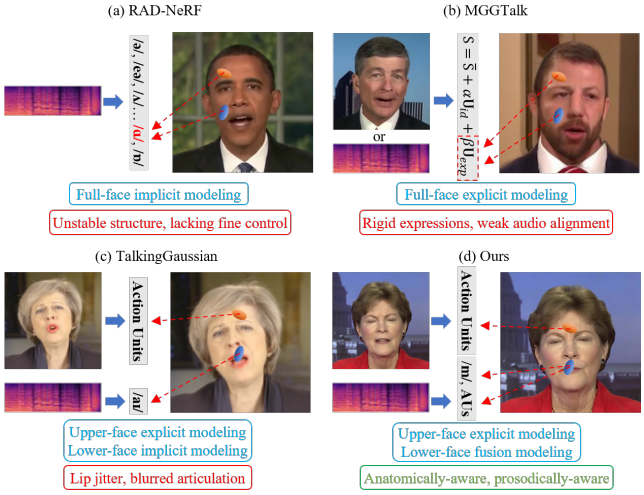

Figure 1: 说话人头像合成。目前的方法主要通过纯隐式 (a) 或纯显式 (b) 方案来建模下脸部运动，因此存在表情僵硬、音频对齐不良或运动模糊和嘴唇抖动等问题。我们的方法采用混合显式-隐式的公式进行下脸部运动建模，实现了对解剖结构感知和韵律感知的面部动画合成。

混合形状 (Peng et al. 2024) 和动作单位（AUs） (Li et al. 2025) 实现了面部上下运动的同步优化和分解。对于较低面部动态，技术进一步将口腔运动与表情隔离，使用分离的运动重建 (Li et al. 2025) 或音频驱动的点云 (Xie et al. 2025) 来提高发音精确度。尽管有这些进展，一个关键限制依然存在：对音频导出的隐式特征的依赖常常导致低面部运动的音素-视素错位，表现为唇部抖动和发音模糊（Figure 1 (c)）。

这激发了我们的核心研究问题：能否探索混合运动特征（来自视觉和音频输入的显性和隐性线索），以最大限度地减少身份特定的偏见，同时确保精确和可推广的唇同步？

为了解决这个问题，我们提出了 HM-Talker，一种创新的混合运动建模框架，用于准确的音频驱动下的面部动画。与先前方法仅依赖于图像衍生的显式特征或音频衍生的隐式特征来进行下脸部运动不同，HM-Talker 结合了两者的优点，实现基于解剖定位的重建并促进节奏上的一致性。具体而言，我们提出了跨模态解缠结模块（CMDM），通过提取显式/隐式运动特征并在模态间对齐运动表示，来推动混合建模的面部动画。CMDM 具

有两个主要功能：1）从面部视频中提取上脸部（$c^e_{v,u}$）和下脸部（$c^e_{v,l}$）的显式表示；2）学习音频衍生的隐式运动特征（$c^i_{a,l}$），并将其投影到视觉发音空间（$c^e_{a,l}$），以便与基于图像的面部动作单元对齐。这使得即使在音频驱动的条件下也能进行基于定位的重建。为了进一步减轻显式运动中的身份依赖偏差并增强跨个体的泛化能力，我们提出了混合运动建模模块（HMMM），它通过门控注意力动态聚合音频衍生的隐式特征与基于面部动作单元的显式特征（音频预测或图像提取）。同时，在训练期间随机选择三对显式-隐式特征，强制进行身份无关的适应，并在音频驱动的场景下提高鲁棒泛化。通过解缠特定模态和特定身份的信息，同时加强跨模态一致性，我们的框架能够稳健地整合隐式和显式运动线索，以实现高保真、身份无关的面部动画。

一般来说，贡献如下所示。

- 我们提出了一种新颖的说话人面部合成框架，该框架将混合隐式音频特征与显式动作单元（AU）提示相结合，以提高唇部发音的准确性和时间对齐。通过将韵律语音动态与基于解剖学的视觉先验相结合，我们的方法实现了可解释和可推广的唇部运动控制。
- 我们设计了一个跨模态解耦模块（CMDM），该模块引入了音频-视觉发音空间投影，将音频衍生的隐式特征转换为视觉兼容的显式特征。这促进了身份不变的音频衍生 AU 表示的学习，在推理过程中充当显式运动提示，并减少与主体相关的过拟合。
- 我们开发了一种混合运动建模模块（HMMM），该模块配备了门控注意机制用于随机特征选择，动态融合不同显式-隐式运动特征组合。这种强制随机化增强了音频驱动的泛化能力，同时保持了解剖学上的合理性。

我们的研究扩展了 TalkingGaussian（Li et al. 2025），这是一个基于三维高斯点云（3DGS）的最先进（SOTA）的音频驱动会说话的人头生成框架。它引入了两个关键创新以实现语音同步渲染：脸部和嘴巴分解以及用于人头建模的可变形高斯场。具体而言，该方法采用了两个不同的分支——面部分支和口中分支——每个分支都使用一对持久性高斯场和基于网格的运动场来合成表现力丰富且时间上连贯的人头。

持续高斯场存储了规范参数 $\theta = \{\mu, s, q, \alpha, f\}$，这些参数通过从语音视频帧的静态三维高斯重建初始化。这些参数分别保留了面部和口腔内部区域的特定身份几何。基于网格的运动场通过一个混合编码器-解码器网络估计每个基元的变形 $\delta = \{\Delta\mu, \Delta s, \Delta q\}$：

$$\delta = \text{MLP}(\mathcal{H}(\mu) \oplus \boldsymbol{C}), \tag{1}$$

其中 $\oplus$ 表示连接，$\mathcal{H}(\mu)$ 是用于空间编码的三平面哈希编码器，$\boldsymbol{C} = \{a, e\}$ 包括音频嵌入 a 和表达参数 e。此公式明确地将几何形变与外观特征解耦。

可微分高斯渲染通过视角依赖的高斯的透明合成来合成像素 p 的颜色 $\mathcal{C}$：

$$\mathcal{C}(p) = \sum_{i \in N} c_i \hat{\alpha}_i \prod_{j=1}^{i-1} (1 - \hat{\alpha}_j), \tag{2}$$

其中 c_i 是通过球谐函数预测的颜色，$\hat{\alpha}_i$ 是第 i 个高斯的二维投影不透明度。总像素不透明度 $\mathcal{A}$ 的表达式为：

$$\mathcal{A}(p) = \sum_{i \in N} \hat{\alpha}_i \prod_{j=1}^{i-1} (1 - \hat{\alpha}_j), \tag{3}$$

在这项工作中，我们主要关注于增强人脸分支。关于口内分支的详细信息可以在附录中找到。

概述

如 Figure 2 所示，我们的 HM-Talker 从单眼视频片段中合成音频驱动的说话头，并采用混合运动表示。给定一个由肖像帧 $\mathcal{I}_{1:T}$ 和音频组成的输入视频，我们的目标是学习保持身份的 3D 高斯表示，并通过 CMDM 和 HMMM 在模拟语音条件发音动态的同时保留特定主题属性的先导运动网络。

根据 TalkingGaussian（Li et al. 2025）的三阶段优化模式，我们的框架通过静态初始化、运动学习和微调进行训练。在所有阶段中，我们采用统一的损失函数，该函数结合了像素级保真度、感知细节和跨模态对齐特性。具体来说，我们使用 $\mathcal{L}_1$ 损失和 D-SSIM 项目来监督低级图像重建，使用 LPIPS 损失（Zhang et al. 2018）来增强感知真实感，并使用对齐损失来确保音频预测特征和图像衍生特征的一致性。总体训练目标被表述为：

$$\mathcal{L}_{DF} = \mathcal{L}_1 + \lambda_1 \mathcal{L}_{D\text{-}SSIM} + \lambda_2 \mathcal{L}_{LPIPS} + \lambda_3 \mathcal{L}_{align}, \tag{4}$$

。

其中 $\mathcal{L}_{align} = \mathcal{L}_1(\boldsymbol{c}^e_{a,l}, \boldsymbol{c}^e_{v,l})$ 在补偿网络中监督跨模态投影。该统一公式确保了跨音频和视觉模态的结构一致性、感知质量和运动连贯性。

基于之前的工作（Peng et al. 2024），我们采用音频视觉编码器作为基础音频编码器，以从原始语音输入中提取可泛化的音频特征。此外，结合流模型（Teed and Deng 2020）的三维可变模型（3DMM）（Paysan et al. 2009）被用于估算头部姿态，从而进一步推断出相机姿态。此外，语义分割模型（Yu et al. 2018; Kvanchiani et al. 2023）被用于执行面部区域分割。

跨模态解耦模块

模型面部运动涉及调和隐含的音频线索与结构上有根据但对身份敏感的显性特征。为了解决这个问题，我们引入了 CMDM，它共同建模三种运动表示：图像衍生的 AU、音频衍生的隐含特征，以及通过补偿网络预测的音频 AU。该设计支持仅音频设置下的 AU 监督，并通过跨模态 AU 对齐促进身份不变的表示。

显式运动建模。给定输入肖像序列 $\mathcal{I}_{1:T}$，我们使用 OpenFace（Baltrusaitis et al. 2018）提取 17 动作单元 $\mathcal{A}$，并将其分为上半脸 $\mathcal{A}_u = \{AU_{01}, AU_{02}, AU_{04}, AU_{05}, AU_{06}, AU_{07}, AU_{45}\}$ 和下半脸 $\mathcal{A}_l = \{AU_{09}, AU_{10}, AU_{12}, AU_{14}, AU_{15}, AU_{17}, AU_{20}, AU_{23}, AU_{25}, AU_{26}\}$ 子集。上半脸 AU 特征被直接串联成一个运动特征 $\boldsymbol{c}^e_{v,u} = \bigoplus_{i \in \mathcal{A}_u} AU_i \in \mathbb{R}^7$。

对于下脸 AUs，直接连接引入了两个主要挑战：（1）有限的表示能力导致对共现模式的过拟合，以及（2）弱的归纳偏向无法推断到未见过的发声。为了解决这些问题，我们引入了一个残差增强的 MLP，用于在保持原始 AU 语义的同时编码非线性的 AU 交互：

$$\boldsymbol{c}^e_{v,l} = \text{MLP}(\mathcal{A}_l) \oplus \mathcal{A}_l \in \mathbb{R}^{32}. \tag{5}$$

图像派生的显式运动特征 $\boldsymbol{c}^e_{v,l}$ 使模型能够通过非线性映射学习高阶共激活模式，同时通过跳跃连接保留原始 AU 输入。

隐式运动建模。显式建模提供解剖学上的控制，但自然的唇部运动也取决于音频驱动的动态，而专注于 ASR

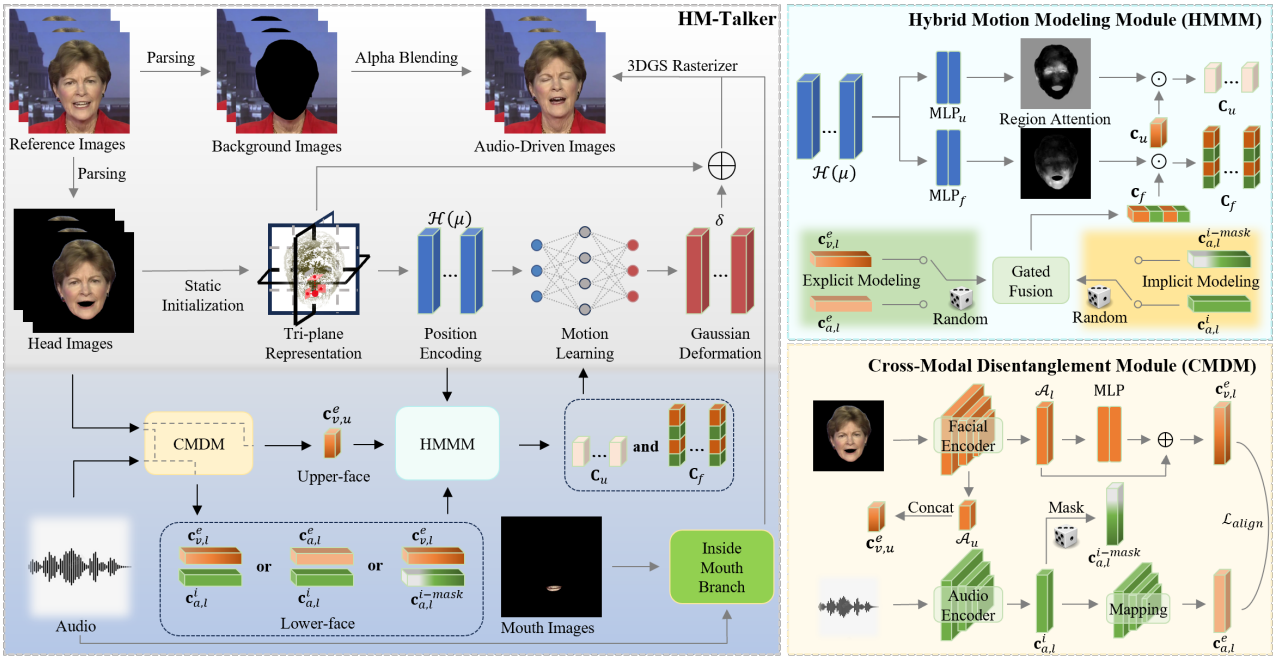

Figure 2: HM-Talker 概述。参考图像在语义上被分解为头部图像、嘴部图像和背景图像。头部图像初始化一个静态高斯场。然后，三平面编码器从这个场中提取位置编码，记为 $\mathcal{H}(\mu)$ 。同时，音频输入和头部图像通过跨模态解耦模块（CMDM）进行处理。该模块输出显式运动特征（$\boldsymbol{c}^e_{a,l}$ ，$\boldsymbol{c}^e_{v,l}$ ）和隐式运动特征（$\boldsymbol{c}^i_{a,l}$ ，$\boldsymbol{c}^{i-mask}_{a,l}$ ）。这些特征与上脸特征（$\boldsymbol{c}^e_{v,u}$ ）组合后输入到混合运动建模模块（HMMM）。HMMM 使用 $\mathcal{H}(\mu)$ 计算区域特定的注意力。然后它融合随机选择的运动特征对，以生成下脸控制向量 $\boldsymbol{C}_f$ 。这个向量与上脸控制向量 $\boldsymbol{C}_u$ 一起，预测应用于静态高斯场的变形 δ 。最后，3DGS 光栅化器渲染动态面部图像。这个结果与内部嘴部分的输出和背景图像进行阿尔法混合，生成音频驱动的输出。

的音频编码器通常会忽略这一点。为了解决这个问题，我们利用了 SyncTalk 的预训练音频视觉编码器（Peng et al. 2024），它通过跨模态对齐捕捉了韵律感知的语位动态 $a \in \mathbb{R}^{512}$ ，从而消除了对手工制作的运动学规则的需求。

我们将这些音频特征分割成重叠的 8 帧窗口，然后通过一个层次网络 AudioNet 来进行降维，以提炼出紧凑的时间嵌入。随后，AudioAttNet 对这些嵌入应用一维卷积和基于注意力的时间聚合，生成音频导出的隐式运动特征 $\boldsymbol{c}^i_{a,l} \in \mathbb{R}^{32}$ 。这个流程强调感知上显著的音频区域，同时抑制噪声，从而产生用于唇部运动生成的稳健韵律表现。（AudioNet 和 AudioAttNet 的架构细节在附录中提供）。

从隐式到显式的投影。为了减轻从图像中导出的显式运动特征中的身份纠缠并实现无图像推理，我们引入了轻量级的音频到动作单元映射器（A2AM）。这个具有 ReLU 激活的多层感知器（MLP）将隐式运动特征 $\boldsymbol{c}^i_{a,l} \in \mathbb{R}^{32}$ 映射到显式的基于动作单元的特征 $\boldsymbol{c}^e_{a,l} \in \mathbb{R}^{32}$ ，并由对齐损失 $\mathcal{L}_{align}$ 进行监督：

$$\boldsymbol{c}^e_{a,l} = \sigma_2(\boldsymbol{W}_2(\sigma_1(\boldsymbol{W}_1\boldsymbol{c}^i_{a,l} + \boldsymbol{b}_1)) + \boldsymbol{b}_2), \tag{6}$$

其中 $\boldsymbol{W}_{1,2}$ 和 $\boldsymbol{b}_{1,2}$ 是可学习参数，而 $\sigma_{1,2}$ 是激活函数。这个简单但有效的设计利用了需要最小转换的高级嵌入。$\mathcal{L}_{align}$ 监督确保了对异常值的鲁棒性，使得音频驱动推理能够重建视觉表达，并保证模态对齐。

混合运动建模模块

HMMM 通过门控注意力动态融合隐式音频特征与显式 AU 基表示来整合异构运动线索。与依赖于特定身份视觉先验的图像驱动方法不同，音频驱动的建模由于其解耦和与说话者无关的特性，自然支持跨身份泛化。为了在保留显式线索的优势的同时充分利用这一点，我们采用了一种随机特征选择策略，从三个变体中随机抽样特征对进行融合，以鼓励泛化，同时确保运动在时间上是一致的并且在解剖学上是连贯的。

随机门控特征融合。我们的可学习门控注意机制协同结合显式和隐式运动特征，以平衡解剖控制与音频驱动的变异性。具体来说，它通过以下门控操作融合显式的下脸特征 $\boldsymbol{c}^e_{:,l}$ 和音频导出的隐式特征 $\boldsymbol{c}^{i*}_{a,l}$ ：

$$\boldsymbol{c}_f = \mathcal{G}(\boldsymbol{c}^{i*}_{a,l}, \boldsymbol{c}^e_{:,l}) = \alpha \odot \boldsymbol{c}^e_{:,l} + (1-\alpha) \odot \boldsymbol{c}^{i*}_{a,l}, \tag{7}$$

其中 $\alpha = \text{MLP}_g(\boldsymbol{c}^{i*}_{a,l} \oplus \boldsymbol{c}^e_{:,l})$ 是一个可学习的权重。此公式允许模型根据上下文需求自适应地关注基于生理的线索或韵律动态。显式特征 $\boldsymbol{c}^e_{:,l}$ 和隐式特征 $\boldsymbol{c}^{i*}_{a,l}$ 分别从三个预定义对之一采样，对应于三个融合路径：1) $\mathcal{P}_{\text{audio}}$ ：使用由 A2AM 生成的 $\boldsymbol{c}^e_{a,l}$ ，然后进行门控融合 $\mathcal{G}(\boldsymbol{c}^i_{a,l}, \boldsymbol{c}^e_{a,l}) \to \boldsymbol{c}_f$ ；2) $\mathcal{P}_{\text{masked}}$ ：对音频输入 $\boldsymbol{c}^{i-mask}_{a,l} = \mathcal{M}_a \cdot \boldsymbol{c}^i_{a,l}$ 应用元素级掩码，然后通过 $\mathcal{G}(\boldsymbol{c}^{i-mask}_{a,l}, \boldsymbol{c}^e_{v,l}) \to \boldsymbol{c}_f$ 与 $\boldsymbol{c}^e_{v,l}$ 融合；3) $\mathcal{P}_{\text{vanilla}}$ ：直接通过 $\mathcal{G}(\boldsymbol{c}^i_{a,l}, \boldsymbol{c}^e_{v,l}) \to \boldsymbol{c}_f$ 将 $\boldsymbol{c}^i_{a,l}$ 与 $\boldsymbol{c}^e_{v,l}$ 融合。该策略模拟推理时视觉线索的缺失，增强鲁棒性并鼓励模型在超越身份特定外观特征

方面的泛化。简单的选择操作计算效率高且完全可并行化，使得在没有额外开销的情况下实现有效的正则化。

融合运动特征 $\boldsymbol{c}_f$ 和上脸运动特征 $\boldsymbol{c}^e_{v,u}$ ，然后与特定区域的注意力图 (Guo et al. 2022) 相结合，以生成用于嘴唇动画和面部表情的空间调制变形场：

$$\boldsymbol{C}_f = \boldsymbol{c}_f \odot \mathrm{MLP_f}(\mathcal{H}(\mu)), \boldsymbol{C}_u = \boldsymbol{c}^e_{v,u} \odot \mathrm{MLP_u}(\mathcal{H}(\mu)). \tag{8}$$

这种以注意力为驱动的机制确保了视觉上可解释的肌肉激活和与语音相关的发音提示能够在空间上连贯且时间上同步地被利用。最后，我们将两个特征向量 $\boldsymbol{C}_u, \boldsymbol{C}_f$ 同时作为控制条件输入到预测网络中：

$$\delta_{face} = \mathrm{MLP}(\mathcal{H}(\mu) \oplus \boldsymbol{C}_u \oplus \boldsymbol{C}_f). \tag{9}$$

高斯变形 δ_{face} 和规范参数 θ_{face} 首先使用 TalkingGaussian 中的可微高斯渲染模块被渲染为面部图像 C_{face} 及其对应的 alpha 图 A_{face} 。同时，内部口腔分支生成一个口腔内部图像 C_{mouth} 及其 alpha 图 A_{mouth} 。最终的音频驱动输出 $\hat{I}_{head}$ 通过对这两个图像进行 alpha 混合获得：

$$\hat{I}_{\mathrm{head}} = C_{\mathrm{face}} \times A_{\mathrm{face}} + C_{\mathrm{mouth}} \times (1 - A_{\mathrm{mouth}}). \tag{10}$$

实验

实验设置

数据集。根据该领域中已建立的研究协议 (Ye et al. 2023; Li et al. 2023; Guo et al. 2021; Tang et al. 2022)，我们的评估框架采用五个公共可访问的视频序列，以保持跨方法的公平比较。该数据集包括三位男性被试（“Lieu”、“Jae-in”和“Obama”）和两位女性被试（“May”和“Shaheen”），平均时长为 7,637 帧，以 25 帧每秒的速度捕获。所有录制内容均保持肖像为中心的构图，主要分辨率为 512×512，除了“Obama”和“Jae-in”中观察到的 450×450 分辨率。

比较基准。我们的比较分析包括三个不同类别的当代方法：2D 生成架构（IP-LAP (Zhong et al. 2023)，TalkLip (Wang et al. 2023)，DINet (Zhang et al. 2023b)），神经辐射场实现（AD-NeRF (Guo et al. 2021)，RAD-NeRF (Tang et al. 2022)，ER-NeRF (Li et al. 2023)，SyncTalk (Peng et al. 2024)），和 3D 高斯散射技术（GaussianTalker (Cho et al. 2024)，TalkingGaussian (Li et al. 2025)）。这一选择确保了对不同技术范式的前沿解决方案的全面覆盖。

静态图像质量评估。我们采用峰值信噪比（PSNR）进行像素级准确性评估，使用结构相似性指数（SSIM）(Wang et al. 2004) 进行结构完整性评估，以及使用学习感知图像局部相似性（LPIPS）(Zhang et al. 2018) 度量来量化感知质量差异，特别是关于细节保留的方面。

动态运动评估。我们使用 SyncNet（Chung and Zisserman 2017a,b) 进行唇同步准确性分析，辅以生成面部表情与参考面部表情之间的地标距离（LMD）(Chen et al. 2018) 测量。同步保真度通过置信分数（Sync-C）和误差距离（Sync-D）指标进一步量化。面部肌肉活动通过 OpenFace (Baltrusaitis et al. 2018) 提取的动作单元（AUs）(Prince et al. 2015) 进行分析，特别关注上面部区域差异（AUE-U）与口腔发声器错误（AUE-L）。

实施细节。对于每个肖像视频，我们首先并行训练面部分支和嘴内部分支，总共进行 50,000 次迭代。在这个阶段，面部分支由隐式和显式运动特征的混合驱动，随机通过 HMMM 选择。之后，两个分支共同进行额外的 15,000 次迭代的微调。训练期间使用 Adam 和 AdamW 优化器。损失权重设置如下：$\lambda_1 = 0.2$ 、$\lambda_2 = 0.5$ 和 $\lambda_3 = 1\mathrm{e}{-3}$ 。所有模块的学习率设置为 5e−4 。所有实验均在 RTX 3090 GPU 上进行。尽管我们的模型在训练时使用了音频和图像输入以增强运动表现，但在评估期间，我们仅对面部下半部分的运动进行音频输入的推理。

与 SOTA 的比较

自重构。为了全面评估重构性能，我们在所有数据集上采用 10:1 的训练-验证划分。如 Table 1 和 Table 2 所示，我们的方法在渲染质量、动作准确性和效率方面优于现有方法。具体而言，其 PSNR 为 35.15 dB，超过基于 NeRF 的 SyncTalk（34.51 dB）和基于 3DGS 的 TalkingGaussian（32.48 dB）。在动作逼真度方面，我们的方法将 LMD 减少到 2.514，将 AUE-L 减少到 0.53，突出了基于 AU 的显式特征编码在面部动作传递中的有效性。值得注意的是，我们的 Sync-C 得分为 7.807，甚至超过了专门的 2D 唇同步基线，验证了我们混合隐式-显式模型在捕捉唇部发音方面的优势。基于 3DGS，我们的框架在实现实时渲染速度（110 FPS）和训练效率（0.51 小时）方面，达到了最先进的视觉质量，并与 TalkingGaussian 匹敌。

口型同步。为了评估泛化能力，我们使用来自未见过的说话者（验证集“Lieu”和“Shaheen”）的域外音频来驱动仅在“May”数据集上训练的模型，包括具有挑战性的性别不匹配情况。正如在 Table 3 中报道的那样，我们的模型始终获得最高的口型同步分数，展示了即使在不同性别间转换发音时，模型也具有强大的语音泛化能力。这验证了我们的框架在保持说话者不变的运动模式方面的有效性，尽管包含了身份条件先验。此外，我们对不同音频输入下的中间表示进行 t-SNE 可视化。如 Figure 4 所示，我们的模型成功保持了隐式和显式运动特征间的区分，并通过门控融合有效地融合它们，证实了稳健的模态解耦。

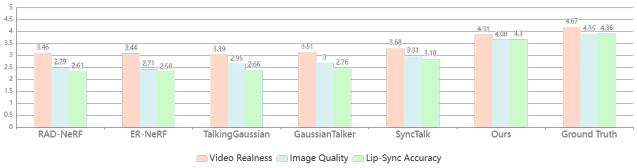

Figure 3: 用户研究。评分范围为 1 到 5，数字越高表示性能越好。

图像质量比较。我们在帧级别针对最先进的时间建模方法进行定性比较：SyncTalk、TalkingGaussian，以及我们提出的方法。选择与目标音素相对应的关键视频帧以突显音素-语素对齐。如 Figure 5 所示，我们的方法在各个音素类别中生成的视觉输出最为一致。例如，在发音时，例如宽口音素如 /ʌ/ 或微妙音素如 /ə/ ，我们的模型保持了精确的视觉对齐，而其他方法则显示出明显的不匹配（在红色框中突出显示）。在如 /eɪ/ 的发音中，尽管竞争方法生成了类似的嘴形，我们的方法更好地保留了口腔内部细节（在黄色框中突出显示），实现了更高的感知真实感。这表明，即便没有显式优化内部口腔结构，在联合训练期间面部运动建模的改进自然延伸到内部发音器上，从而实现连贯和准确的说话头合成。

	Methods	Rendering Quality			Motion Quality			Efficiency	
		PSNR ↑	LPIPS ↓	SSIM ↑	LMD ↓	AUE-(L/U) ↓	Sync-C ↑	Time	FPS
NeRF	AD-NeRF (Guo et al. 2021)	30.07	0.1042	0.9689	2.998	1.01/0.97	6.053	18.7h	0.11
NeRF	RAD-NeRF (Tang et al. 2022)	31.95	0.0620	0.9660	2.847	0.74/0.76	5.742	5.3h	28.7
NeRF	ER-NeRF (Li et al. 2023)	32.47	0.0395	0.9658	2.639	0.62/0.54	6.531	2.1h	31.2
NeRF	SyncTalk (Peng et al. 2024)	34.51	0.0221	0.9959	2.607	0.55/0.29	7.502	2.0h	52
3DGS	GaussianTalker (Cho et al. 2024)	32.69	0.0442	0.9952	2.726	0.67/0.59	6.234	3.2h	95
3DGS	TalkingGaussian (Li et al. 2025)	32.48	0.0309	0.9950	2.616	0.60/0.28	6.246	0.5h	108
3DGS	HM-Talker (Ours)	35.15	0.0207	0.9971	2.514	0.53 / 0.22	7.807	0.51h	110

Table 1: 自我重建的比较。与基于 NeRF 或 3DGS 的方法相比，我们在大多数指标上达到了领先的性能。最佳和次佳的结果分别用粗体和下划线表示。

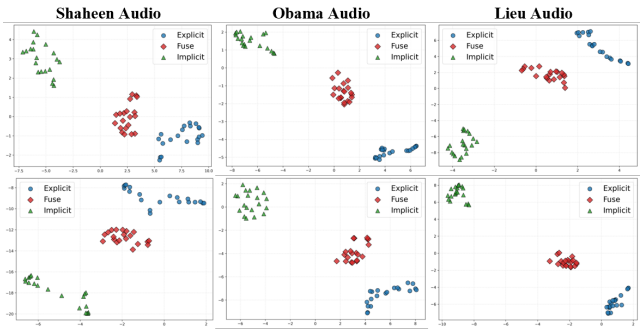

Figure 4: 基于三个 20 帧剪辑的运动特征的 t-SNE 可视化。每行对应一个剪辑；每列代表不同的音频输入。这里，“显式”表示音频预测的显式特征。

用户研究。我们进行了一个用户研究，邀请了 30 位非专业参与者评价 35 段视频（5 个身份 × 7 种方法，每段 10 秒）。参与者在一个 5 分制的评分表上对视频真实感、图像质量和口型同步准确性进行评分。在所有方面，我们的方法始终优于先前的基线方法（Figure 3 ）。尤其是在视频真实感上获得了 4.31 分，在图像质量上获得了 4.08 分，分别比第二好的方法高出 17 % 和 23 %。此外，在口型同步准确性上，我们的方法获得了 4.10 分，大大缩小了与真实情况（4.36 分）的差距。

消融研究

我们在自重建设置下进行消融研究，以评估我们 HM-Talker 框架中关键组件的贡献。由于 CMDM 提供了 HMMM 使用的解耦运动特征，这两个模块在评估中并不是完全可以分离的。因此，我们通过在它们相互连接的约束下修改其内部设置来评估每个模块的影响。更多的定性比较和额外的消融研究，请读者参阅附录。

	Methods	Motion Quality		
		LMD ↓	AUE-(L/U) ↓	Sync-C ↑
GAN	IP-LAP (Zhong et al. 2023)	3.161	1.00/-	7.040
GAN	DINet (Zhang et al. 2023b)	3.230	1.09/-	7.455
GAN	TalkLip (Wang et al. 2023)	3.285	0.82/-	6.657
	HM-Talker (Ours)	2.636	0.61 / 0.26	7.756

Table 2: 自重构的比较结果。

Methods	“Shaheen” Audio		“Lieu” Audio	
	Sync-D ↓	Sync-C ↑	Sync-D ↓	Sync-C ↑
Ground Truth	6.239	10.015	6.840	8.372
DINet (Zhang et al. 2023b)	8.201	7.295	8.226	6.470
IP-LAP (Zhong et al. 2023)	9.819	5.316	9.392	5.077
TalkLip (Wang et al. 2023)	9.553	5.488	11.679	3.151
RAD-NeRF (Tang et al. 2022)	12.012	3.054	12.044	2.449
ER-NeRF (Li et al. 2023)	9.775	5.529	10.017	4.782
SyncTalk (Peng et al. 2024)	8.903	6.350	7.508	7.780
GaussianTalker (Cho et al. 2024)	8.926	6.576	10.943	4.198
TalkingGaussian (Li et al. 2025)	11.450	3.179	9.849	5.039
HM-Talker (Ours)	7.590	7.972	7.292	7.994

Table 3: 唇同步比较

Setting	PSNR ↑	AUE-(L/U) ↓	Sync-C ↑	LMD ↓
Gate Fusion	35.155	0.53/ 0.22	7.807	2.514
Concat Fusion	35.100	0.52 /0.24	7.818	2.515
Fix Alpha	35.053	0.54/0.31	7.770	2.527
Purely Implicit	35.063	0.61/0.27	7.747	2.531
Purely Explicit	34.944	0.53/0.32	7.382	2.588
ActionUnit	35.155	0.53/ 0.22	7.807	2.514
3DMM	35.117	0.52 /0.26	7.679	2.534
BlendShape	35.106	0.53/0.25	7.731	2.520
AVE	35.155	0.53/0.22	7.807	2.514
DeepSpeech	34.847	0.72/0.22	6.230	2.718
Hubert	35.018	0.47 / 0.20	6.890	2.531
CMDM (AU)	35.155	0.53 / 0.22	7.807	2.514
ExpNet (3DMM)	34.257	0.85/0.25	6.404	3.181

Table 4: 关于不同设置的消融研究。

跨模态解耦模块。为了评估 CMDM 的必要性，我们用 SadTalker 的跨模态编码器 ExpNet (Zhang et al. 2023a) (Table 4 ，第 12-13 行) 替换它。虽然 ExpNet 在音频到 3DMM 映射中具有良好的泛化能力，但在我们的设定中，它未能产生精确且具身份感知的面部动态，从而导致唇同步效果下降。我们还比较了不同显式运动表示，使用 3DMM 和 BlendShape 替换动作单元 (Table 4 ，第 6-8 行)。结果保持一致，表明我们的框架具有鲁棒性，虽然 AU 由于其细粒度设计提供了更好的发音控制。最后，我们通过用 DeepSpeech (Amodei et al. 2016) 和 HuBERT (Hsu et al. 2021) (Table 4 ，第 9-11 行) 替代 AVE 来评估音频编码器的影响。使用 DeepSpeech 时，性能显著下降，其专注于音素的训练缺乏 viseme 级别的细节。在这种情况下，融合权重 α 收敛到 1，表明模型抑制了嘈杂的音频特征——展示了我们门控融合的适应性。

混合运动建模模块。我们通过以下比较验证我们的门控注意力：(1) 基于 MLP 的融合，(2) 固定融合 (α=0.5)，(3) 纯显式 (α=1)，和 (4) 纯隐式 (α=0) 建模 (Table 4 ，第 1-5 行)。具有自适应门控的混合建模实现了最佳的整体性能。静态融合比率和单模设置表现不佳，而 MLP 融合性能相似但缺乏可解释性。相反，我们的门控机制动态平衡特征的可靠性，兼顾了性能和透明度。

此外，我们采用了三路径训练策略，并固定比例为 $\mathcal{P}_{\text{audio}} : \mathcal{P}_{\text{masked}} : \mathcal{P}_{\text{vanilla}} = 4{:}4{:}2$ 。消融研究揭示了不同路径配置的影响。首先，固定 $\mathcal{P}_{\text{audio}}$ 同时减少 $\mathcal{P}_{\text{masked}}$ 导致所有指标的性能一致下降（Figure 6 ），这强调了音频掩码的重要性。我们将其归因于掩码抑制了隐性和显性特征之间的冗余信息，从而改善了融合。接下来，我们固定 $\mathcal{P}_{\text{vanilla}}$ 并改变 $\mathcal{P}_{\text{audio}}$ 与 $\mathcal{P}_{\text{masked}}$ 之间的比例。当 $\mathcal{P}_{\text{masked}}$ 过高 (7) 或过低 (1) 时，性能下降，证实了我们所选平衡的有效性。我们还考察了 $\mathcal{P}_{\text{masked}}$ 内不同掩码比例。如 Figure 7 所示，性能在掩码率为 $\mathcal{M}_a = 0.2$ 时达到峰值，并随着值的偏离而下降。为了

Figure 5: 图像质量比较的定性结果。与其他方法相比，我们的方法实现了最一致的音素-可见图素对齐性能，其中 TGS 代表 TalkingGaussian。请放大以获得更好的可视化效果。

增强鲁棒性，我们采用随机掩码策略，在训练期间均匀采样 $\mathcal{M}_a$ 从 0.1 到 0.3。这通过更好地平衡正则化和表征学习优于大多数固定设置。最后，我们通过 t-SNE 评估音频预测显性特征的质量，在两个次优设置下：（i）路径比例为 1:7:2 ，以及（ii）较高的掩码率 $\mathcal{M}_a = 0.9$ 。在这两种情况下（Figure 8 ），观察到音频预测与图像导出的特征之间的对齐不良，而我们的默认设置产生更好的重叠，表明融合和解嵌入得到了改善。

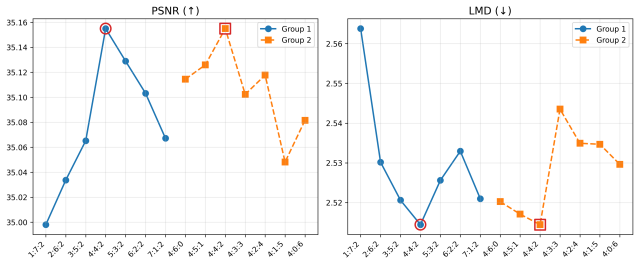

Figure 6: 混合运动建模模块中路径比例的消融研究。

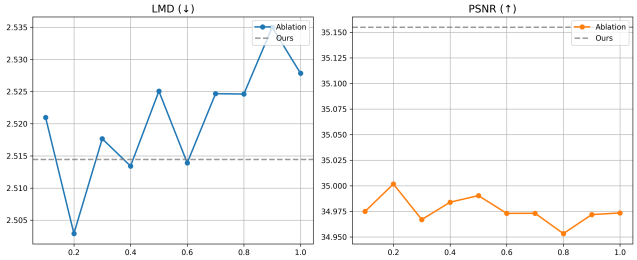

Figure 7: 在混合运动建模模块中的 $\mathcal{M}_a$ 消融研究。

我们提出了 HM-Talker，一种用于高保真、身份无关音频驱动的面部动画的混合运动建模框架。与依赖单一模态线索的先前方法不同，HM-Talker 将解剖学基础

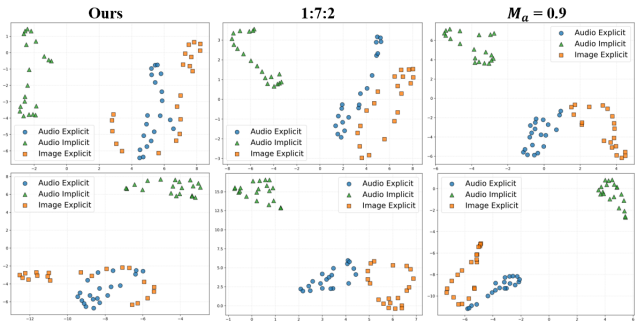

Figure 8: 三段每段包含 20 帧的剪辑的运动特征的 t-SNE 可视化。每行对应一个剪辑；每列代表一个不同的训练设置。

的显式特征与节奏敏感的隐式特征相结合。为实现这一目标，我们引入了两个关键模块：跨模态解耦模块（CMDM），该模块对齐音频和视觉表示，以在音频驱动的设置下启用基于 AU 的监督；以及混合运动建模模块（HMMM），该模块通过门控注意融合多模态运动特征，并采用随机特征配对以增强跨主体的泛化能力。实验证实，HM-Talker 能够在不同身份中产生时间一致且视觉逼真的结果，推进说话头像合成的最新技术。

References

Amodei, D.; Ananthanarayanan, S.; Anubhai, R.; Bai, J.; Battenberg, E.; Case, C.; Casper, J.; Catanzaro, B.; Cheng, Q.; Chen, G.; et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In International Conference on Machine Learning, 173–182.

Baltrusaitis, T.; Zadeh, A.; Lim, Y. C.; and Morency, L.-P. 2018. OpenFace 2.0: Facial Behavior Analysis

Toolkit. In International Conference on Automatic Face and Gesture Recognition, 59–66. IEEE.

Chen, L.; Li, Z.; Maddox, R. K.; Duan, Z.; and Xu, C. 2018. Lip movements generation at a glance. In European Conference on Computer Vision, 520–535.

Cho, K.; Lee, J.; Yoon, H.; Hong, Y.; Ko, J.; Ahn, S.; and Kim, S. 2024. GaussianTalker: Real-Time High-Fidelity Talking Head Synthesis with Audio-Driven 3D Gaussian Splatting. In ACM International Conference on Multimedia, 10985–10994.

Chu, X.; and Harada, T. 2025. Generalizable and Animatable Gaussian Head Avatar.

Chung, J. S.; and Zisserman, A. 2017a. Lip reading in the wild. In Asian Conference on Computer Vision, 87–103. Springer.

Chung, J. S.; and Zisserman, A. 2017b. Out of time: automated lip sync in the wild. In Asian Conference on Computer Vision, 251–263. Springer.

Gong, S.; Li, H.; Tang, J.; Hu, D.; Huang, S.; Chen, H.; Chen, T.; and Liu, Z. 2025. Monocular and Generalizable Gaussian Talking Head Animation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Guo, M.-H.; Liu, Z.-N.; Mu, T.-J.; and Hu, S.-M. 2022. Beyond self-attention: External attention using two linear layers for visual tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(5): 5436–5447.

Guo, Y.; Chen, K.; Liang, S.; Liu, Y.-J.; Bao, H.; and Zhang, J. 2021. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In IEEE/CVF International Conference on Computer Vision, 5784–5794.

Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29: 3451–3460.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics, 42(4).

Kvanchiani, K.; Petrova, E.; Efremyan, K.; Sautin, A.; and Kapitanov, A. 2023. EasyPortrait–Face Parsing and Portrait Segmentation Dataset.

Li, J.; Zhang, J.; Bai, X.; Zheng, J.; Ning, X.; Zhou, J.; and Gu, L. 2025. Talkinggaussian: Structure-persistent 3d talking head synthesis via gaussian splatting. In European Conference on Computer Vision, 127–145.

Li, J.; Zhang, J.; Bai, X.; Zhou, J.; and Gu, L. 2023. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In IEEE/CVF International Conference on Computer Vision, 7568–7578.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1): 99–106.

Paysan, P.; Knothe, R.; Amberg, B.; Romdhani, S.; and Vetter, T. 2009. A 3D face model for pose and illumination invariant face recognition. In IEEE International Conference on Advanced Video and Signal based Surveillance, 296–301. IEEE.

Peng, Z.; Hu, W.; Shi, Y.; Zhu, X.; Zhang, X.; Zhao, H.; He, J.; Liu, H.; and Fan, Z. 2024. Synctalk: The devil is in the synchronization for talking head synthesis. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 666–676.

Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020. A lip sync expert is all you need for speech to lip generation in the wild. In ACM International Conference on Multimedia, 484–492.

Prince, E. B.; Martin, K. B.; Messinger, D. S.; and Allen, M. 2015. Facial action coding system. Environmental Psychology & Nonverbal Behavior, 1.

Tang, J.; Wang, K.; Zhou, H.; Chen, X.; He, D.; Hu, T.; Liu, J.; Zeng, G.; and Wang, J. 2022. Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. arXiv preprint arXiv:2211.12368.

Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In European Conference on Computer Vision, 402–419. Springer.

Wang, J.; Qian, X.; Zhang, M.; Tan, R. T.; and Li, H. 2023. Seeing what you said: Talking face generation guided by a lip reading expert. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14653–14662.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4): 600–612.

Xie, Y.; Feng, T.; Zhang, X.; Luo, X.; Guo, Z.; Yu, W.; Chang, H.; Ma, F.; and Yu, F. R. 2025. PointTalk: Audio-Driven Dynamic Lip Point Cloud for 3D Gaussian-based Talking Head Synthesis. In AAAI Conference on Artificial Intelligence.

Ye, Z.; Jiang, Z.; Ren, Y.; Liu, J.; He, J.; and Zhao, Z. 2023. GeneFace: Generalized and High-Fidelity Audio-Driven 3D Talking Face Synthesis. In The Eleventh International Conference on Learning Representations.

Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; and Sang, N. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In European Conference on Computer Vision, 325–341.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 586–595.

Zhang, W.; Cun, X.; Wang, X.; Zhang, Y.; Shen, X.; Guo, Y.; Shan, Y.; and Wang, F. 2023a. SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Zhang, Z.; Hu, Z.; Deng, W.; Fan, C.; Lv, T.; and Ding, Y. 2023b. Dinet: Deformation inpainting network for realistic face visually dubbing on high resolution video. In AAAI Conference on Artificial Intelligence, 3543–3551.

Zhong, W.; Fang, C.; Cai, Y.; Wei, P.; Zhao, G.; Lin, L.; and Li, G. 2023. Identity-preserving talking face generation with landmark and appearance priors. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9729–9738.