arXiv:2601.19112v1 [cs.AI] 27 Jan 2026

基于情感先验蒸馏的、具备不确定性感知能力的三维情感说话人脸合成

Nanhan Shen 和 *Zhilei Liu**

天津大学人工智能学院，中国天津

摘要

情感说话人脸合成在多媒体和信号处理中至关重要，但现有三维方法面临两大关键挑战：音频-视觉情感对齐效果不佳，表现为音频情感提取困难，且对情感微表情的控制不足；以及采用一刀切的多视图融合策略，忽视不确定性和特征质量差异，进而损害渲染质量。我们提出 UA-3DTalk，即基于情感先验蒸馏、具备不确定性感知能力的三维情感说话人脸合成方法，其包含三个核心模块：先验提取模块将音频解耦为用于对齐的内容同步特征，以及用于个性化的个体特定互补特征；情感蒸馏模块引入多模态注意力加权融合机制和采用多分辨率码本的4D高斯编码，从而实现细粒度的音频情感提取，并精准控制情感微表情；基于不确定性的变形模块部署不确定性模块，以估计视角特定的偶然不确定性（输入噪声）和认知不确定性（模型参数），实现自适应多视图融合，并引入多头解码器来优化高斯基元，从而缓解均匀权重融合的局限性。在常规数据集和情感数据集上进行的大量实验表明，UA-3DTalk 的表现优于 DEGSTalk 和 EDTalk 等先进方法：在情感对齐方面，E-FID 提升 5.2%；在唇部同步方面，SyncC 提升 3.1%；在渲染质量方面，LPIPS 提升 0.015。项目页面：https://mrask999.github.io/UA-3DTalk/。

索引词— 说话人脸合成、不确定性估计、3D高斯泼溅、情感合成、唇部同步

1. 引言

情感说话人脸合成因其广泛的应用价值[1]，已成为多媒体和信号处理领域的热门研究方向。现有方法主要面临两方面的技术挑战，且各自存在不容忽视的局限性。情感的引入带来了第一个挑战。首先，与传统的说话人脸不同，在这一场景中，夸张的嘴部动作或表情十分常见，因此需要更高的同步精度[2]。其次，音频情感提取本身就是一项挑战，因为这也是语音情感识别（SER）中的热门课题。现有工作主要依赖基于参考的情感控制（例如 EmotiveTalk[3]），这不仅依赖高质量的参考视频片段，而且在纯音频驱动场景中效果并不理想。最后，与传统的说话人脸相比，该任务要求从音频中实现更强的表情控制，而以往方法无法完全满足这一需求。第二个挑战在于多视图特征融合，而这一问题始终被忽视。通常，研究人员采用存在缺陷的“一刀切”策略，将所有特征拼接在一起 [4][5][6], ，从而可能生成长度达数百维的特征输入。在实际应用中，由于遮挡、表征学习问题和噪声，不同模态的质量存在显著差异；忽略这些差异以及多视图数据中固有的不确定性，会导致次优结果。为应对这些挑战，我们提出了 UA-3DTalk（基于不确定性的 3D 情感说话人脸合成与情感先验蒸馏）框架。该框架采用 3D 高斯泼溅（3DGS）作为渲染骨干网络，并包含三个核心模块：先验提取、情感蒸馏和基于不确定性的变形。

先验提取和情感蒸馏的实现是解决音频与视觉情感对齐不佳问题的基础。在先验提取中，我们将音频解耦为f_{exp} 和 f_{tone} ，分别作为 [7]。 f_{exp} 是用于音视对齐的内容同步特征， f_{tone} 是用于个性化的个体特定互补特征，二者共同确保同步性和身份精度。在情感蒸馏中，我们扩展音频窗口并采用多模态融合，以捕获充足的情感信息，同时避免特征稀释。我们从音频中提取频谱图、梅尔频率倒谱系数和音频波形，将前两者用作注意力键，对第三者进行重新加权，以进行情感分类和表情预测 [8], ；随后，通过多平面网格化 [9] 和多分辨率码本 [10] 对情感特征进行编码，从而精确控制情感表情。

基于不确定性的变形模块旨在解决现有三维情感说话人脸合成方法中次优的“一刀切”多视图融合问题。该模块为每个特征视图设置不确定性模块，共同生成每个高斯基元的状态向量。不同于 [11], 以往的工作，这些模块分别估计由输入噪声引起的偶然不确定性（AU）和由模型参数引起的认知不确定性（EU），并预测视图特定的状态向量 [12, 13]。随后，根据这些信息实施自适应融合，遵循不确定性越高、视图权重越低的原则。生成的状态向量由多头高斯变形解码器 [9] 解码，以输出对高斯基元的修改。此外，该模块还推动了不确定性建模首次在说话人脸生成领域得到系统性集成。

总而言之，我们工作的主要贡献可概括如下：

- 为解决多视图融合存在的缺陷，我们首次在三维情感说话人脸合成中系统地建模了偶然不确定性和认知不确定性。
- 我们提出了一种不确定性感知自适应融合策略，根据特征质量动态调整视图权重，优先采用不确定性较低的信息。
- 情感先验蒸馏以协作方式结合先验提取模块和情感蒸馏模块，以解决情感对齐问题。

*通讯作者（zhileiliu@tju.edu.cn）

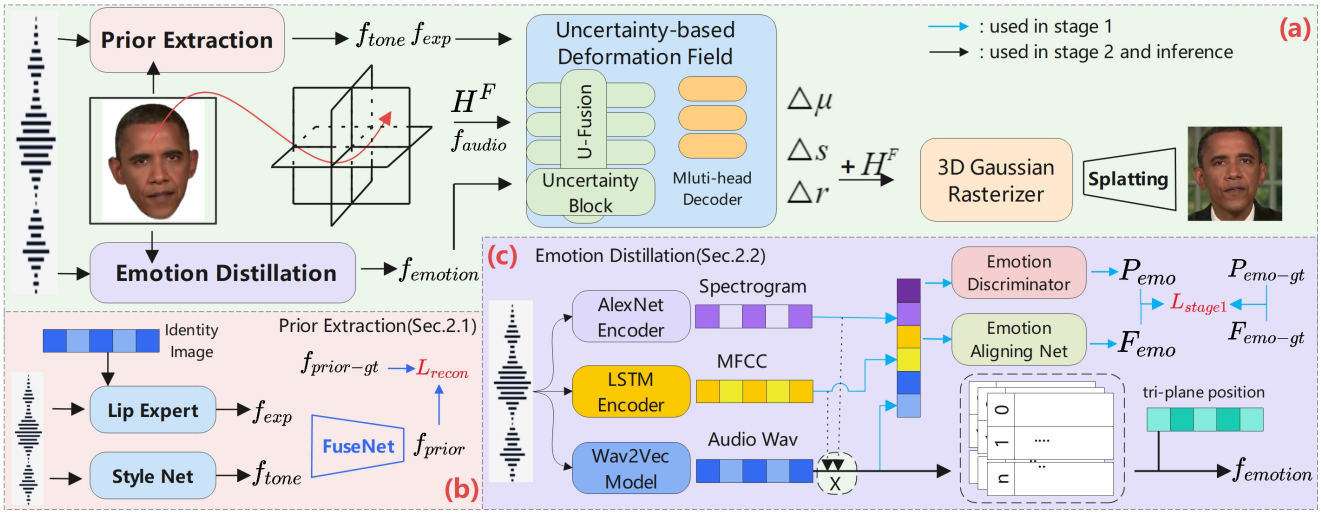

图1：所提出的 UA-3DTalk 框架整体架构：(a) 人脸分支的端到端流程，整合先验提取、情感蒸馏和基于不确定性的形变，用于调整和渲染 3D 高斯基元；(b) 先验提取模块的详细框架；(c) 情感蒸馏模块的技术流程。

2. 所提出的方法

我们的 UA-3DTalk 架构如图1所示，其中采用 3DGS 作为渲染骨干。输入图像和音频被输入先验提取模块（图1(b)），以生成 f_{exp}, f_{tone} 1 ，并被输入情感蒸馏模块（图(c)），以获得 $f_{emotion}$。这些特征与静态高斯基元 H以及基线简单编码的音频特征 f_{audio} [5]一同输入基于不确定性的变形模块，以计算对高斯基元的修改。随后，高斯光栅化器使用这些修改后的基元渲染动态面部帧。与 [5], 一致，我们将头部分解为人脸分支和嘴部分支；上述流程对应人脸分支（图1(a)）。

在 3DGS 中，用于表征人头的静态高斯基元由一组点云参数定义：三维中心$\mu \in \mathbb{R}^3$、三维缩放因子 $s \in \mathbb{R}^3$、四维旋转四元数 $r \in\mathbb{R}^4$、标量不透明度值 $\alpha \in \mathbb{R}$，以及 Z 维颜色特征 $c \in \mathbb{R}^Z$。协方差矩阵 Σ 由 s 和 r推导得出。进行像素级颜色渲染时，按照公式 1 和 2，将与目标像素重叠的 N 个高斯基元按顺序混合其贡献，从而计算出目标像素的颜色 C 。

$$C = \sum_{i\in\mathcal{N}} c_i\alpha_i \prod_{j=1}^{i-1}(1-\alpha_j), \qquad \alpha_i = \sigma_i \mathcal{G}_i(\mathbf{x}) \tag{1}$$

$$\mathcal{G}_i(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)} \tag{2}$$

其中，$\mathcal{G}_i$(x)表示像素位置 x处第 i个高斯基元的高斯分布值，σ_i 是第 i个高斯基元不透明度的缩放系数。为了渲染动态面部图像，我们引入了变形场 $\delta_i = \{\Delta\mu_i, \Delta s_i, \Delta r_i\}$，分别表示对高斯基元的位置、缩放和旋转所做的修改。该变形场定义如下式3所示：

$$\delta_i = Deformation(\mathcal{H}, f_{audio}, f_{exp}, f_{tone}, f_{emotion}) \tag{3}$$

值得注意的是，嘴部分支的流程与人脸分支略有不同：它不会将情感特征$f_{emoiton}$ 作为变形场的输入，且变形场仅输出高斯基元的位置修改 $\triangle\mu$ 。其他所有组件均与人脸分支保持一致。

2.1. 先验提取

为满足情感说话人脸对音视同步和身份保持的更高要求，我们使用先验提取模块生成音视同步度高的表情特征 f_{exp} 和个体特定互补信息 f_{tone}，其流程如图1(b)所示。该模块分三个连续步骤运行：**(1)唇部专家训练**：使用预训练的 SyncNet [14][7], 计算同步损失，据此训练唇部专家 [14] ，以生成唇形波形图像，初步建立音频唇形对齐。(2) **表情特征提取**：将这些唇形波形图像输入三维先验提取子模块 [15], ，以提取表情特征 f_{exp}，为后续处理持续保持音视同步奠定基础。(3) **个体特定特征融合**：引入基于StyleNet的音频编码器 [7] ，从输入音频中提取个体特定互补特征 f_{tone}。随后将 f_{tone} 与 f_{exp} 融合，以重建三维面部先验信息 $f_{gen-exp}$，确保生成的特征在保持同步的同时保留个体身份。本部分的训练由式4定义的重建损失监督，其中 f_{gt-exp} 为真实值。

$$\mathrm{g}_{L_{\text{recon}}} = \|f_{gen-exp} - f_{gt-exp}\|^2 \tag{4}$$

2.2. 情感蒸馏

情感蒸馏模块旨在实现高精度的音频情感提取，并对情感表情进行有力控制。如图1.(c)所示，该模块由两部分组成：提取和编码。

(1) **提取组件**：为捕获充分的情感信息而避免特征稀释，我们将原始音频预处理为频谱图、梅尔频率倒谱系数和音频波形特征 [16], ，借鉴语音情感识别任务 [17]。我们通过 AlexNet 提取二维频谱图特征，并通过长短期记忆网络提取时序梅尔频率倒谱系数特征，将这两类特征作为注意力键，对音频波形特征进行重新加权，使其具备情感感知能力$f_{emo-attn}$，从而支持下游任务中的情感分类和表达特征预测 [8, 3]。该组件经历两个训练阶段：在 IEMOCAP数据库 [18] 上进行预训练，由情感标签的交叉熵损失监督；随后在个体特定数据集上进行微调。微调损失结合了情感标签交叉熵损失，以及预测的表达特征与

真实表达特征 f_{gt-emo}之间的 L2预测损失 f_{emo}，具体定义见 E .5。

q

$$L_{\text{stage2}} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) + \|f_{emo} - f_{gt-emo}\|^2 \tag{5}$$

C 是情绪类别的数量， y_i 是独热真实标签，$\hat{y}_i$ 是预测概率。在后训练阶段，编码器参数被冻结，以保留提取能力。

(2) **编码组件**：为处理无法直接进行平面网格化的高维情绪特征 [9],，我们采用多分辨率码本 [10] 进行离散化，并以哈达玛积取代拼接。 [19] 验证了这种方法能够在避免维度爆炸的同时保留高维信息。最终编码特征 $f_{emotion}$ 定义见公式 6。

$$f_{emotion} = \bigcup_{(s)} \prod Interp(R^s(i,j)) \odot f^s_{emo-attn} \tag{6}$$

其中 R是平面，$(i,j) \in \{(X,Y),(X,Z),(Y,Z)\}$， $s \in \{1,2,4\}$表示用于码本离散化的多分辨率尺度，基础分辨率为 64。这些尺度能够为后续的高斯基元变形提供精确的微表情控制。

2.3. 基于不确定性的变形

如图 2 所示，基于不确定性的变形模块由特征融合和解码组件构成。该模块通过考虑模态质量差异，解决存在缺陷的“一刀切”多视图融合问题，旨在提升渲染质量和情绪表达精度。

(1) **带不确定性估计的特征融合**：对于每个特征视图，我们部署一个由 N 不确定性网络构成的不确定性模块。每个网络以高斯基元和相应视图的特征为输入（ $f_{emotion}$ 除外，其仅使用输入），输出 (μ, σ)，以表示该视图状态向量的预测分布，如式 7 所示。

$$p(\hat{y}_q \mid x_q, \theta) = \mathcal{N}\big(\hat{y};\, \mathbf{F}^{\mu}_{\theta_1}(x),\, \mathbf{F}^{\sigma}_{\theta_2}(x)\big) \tag{7}$$

其中，$\mathbf{F}^{\mu}_{\theta_1}(.)$ 和 $\mathbf{F}^{\sigma}_{\theta_2}(.)$ 是分别输出均值 μ 和方差 σ的网络分支，其可训练参数分别为 θ_1 和 θ_2。

根据式 8 的定义， [12], 不确定性包括 AU（源自输入噪声）和 EU（源自模型参数）。

$$\mathrm{V}[\hat{y}_q|x_q] = \underbrace{\mathrm{V}_{q^*(\theta)}[\mathrm{E}(\hat{y}_q|x_q,\theta)]}_{\Delta_E[\hat{y}_q]} + \underbrace{\mathrm{E}_{q^*(\theta)}[\mathrm{V}(\hat{y}_q|x_q,\theta)]}_{\Delta_A[\hat{y}_q]} \tag{8}$$

其中，$\mathrm{E}(\hat{y}_q|x_q,\theta)$ 和 $\mathrm{V}(\hat{y}_q|x_q,\theta)$ 表示后验预测分布 $p(\hat{y}|x_q,\theta)$ 下预测状态向量$\hat{y}_q$ 的均值和方差，左侧表示 EU，右侧表示 AU。

AU 由不确定性网络直接估计，而 EU 则通过蒙特卡罗采样近似得到：我们为每个模块使用 $T = 10$ 个不确定性网络，并将其 μ 个输出的方差作为 EU。每个视图的总不确定性如式 9 所示，模块的最终输出（均值 $\hat{}\mu$ 和方差$\hat{\sigma}$）定义于式 10 和式 11。

$$\hat{\mathbf{V}}[\hat{y}_q \mid x_q] = \underbrace{\frac{1}{T}\sum_{t=1}^{T} \mathbf{F}^{\mu}_{\theta_1^t}(x_q)\mathbf{F}^{\mu}_{\theta_1^t}(x_q)^{\top} - \overline{\mathbf{F}}^{\mu}(x_q)\,\overline{\mathbf{F}}^{\mu}(x_q)^{\top}}_{\hat{\Delta}_{\mathrm{E}}(\hat{y}_q)} + \underbrace{\sum_{t=1}^{T} \mathbf{F}^{\sigma}_{\theta_2^t}(x_q)}_{\hat{\Delta}_{\mathrm{A}}(\hat{y}_q)} \tag{9}$$

其中，$\mathbf{F}^{\mu}(x_q) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{F}^{\mu}_{\theta_1^t}(x_q)$ 是 T 个网络输出的均值。

$$\hat{\mu}_{y|x} = \frac{1}{T}\sum_{t=1}^{T} \mu_t\big(x;\theta_1^t\big) \xrightarrow[T\to\infty]{} \mathbb{E}_{q^*_\phi(y|x)}[y] \tag{10}$$

$$\hat{\sigma}_{y|x} = \frac{1}{T}\sum_{t=1}^{T}\Big(\sigma_t(x;\theta_2^t) + \mu_t(x;\theta_1^t)\,\mu_t(x;\theta_1^t)^{\top}\Big) - \hat{\mu}_{y|x}\,\hat{\mu}^{\top}_{y|x} \tag{11}$$

其中，$\hat{}\mu$和$\hat{\sigma}$ 是当前视图的最终输出。

多视图特征通过高斯融合进行融合，如公式 12 所定义，其中权重根据不确定性动态调整（不确定性越高，权重越低）。

g g y

$$\mu = \Sigma\sum_{i=1}^{N}\hat{\sigma}_i^{-1}\hat{\mu}_i, \qquad \Sigma = \left(\sum_{i=1}^{N}\hat{\sigma}_i^{-1}\right)^{-1} \tag{12}$$

其中， N是输入视图的数量， μ_i 和 σ_i 分别是第 i个视图的均值和不确定性。特征融合的最终输出是预测的状态向量 μ，它表示情感空间中每个高斯基元的音频条件，其真实分布为 $q^*_\phi(y \mid x)$ 。

（2）**高斯形变解码**：如图 2 所示，融合后的状态向量被输入多头高斯形变解码器，以取代传统的简单多层感知机，从而分别为高斯基元生成修改参数（$\triangle\mu$、 $\triangle r$、 $\triangle s$）（位置、旋转、缩放）。这种设计与动态高斯建模 [9], 相契合，使每个基元都能得到细粒度控制，从而提升渲染保真度。

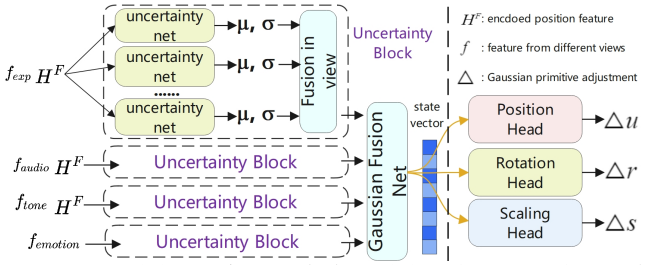

图 2：基于不确定性的变形。（左）不确定性感知多视图特征融合模块；（右）多头高斯变形解码器。

2.4. UA-3DTalk 训练策略

我们的 UA-3DTalk 框架的训练过程分为两个连续阶段，从而确保核心模块能够独立优化，并实现完整渲染流程的有效集成。

阶段 1：本阶段重点训练先验提取模块，并对情感蒸馏模块进行预训练。本阶段的详细实施步骤分别在对应章节（第 2.1 节和第 2.2 节）中说明。

阶段 2：本阶段针对渲染流程进行端到端优化，包含两个关键步骤：首先，分别针对面部和嘴部分支训练基于不确定性的变形模块，并使用真实掩码图像 $\mathcal{I}_{mask}$ 与渲染输出 $\mathcal{I}_{render}$ 之间的损失进行监督，如公式 13 所定义。其次，对面部和嘴部分支进行微调与融合，并使用真实图像 $\hat{}\mathcal{I}$ 与融合输出 $\mathcal{I}_{fuse}$ 之间的损失进行监督，如公式 14 所示。

$$\mathcal{L}_D = \mathcal{L}_1(\mathcal{I}_{render}, \mathcal{I}_{\text{mask}}) + \lambda\,\mathcal{L}_{\text{SSIM}}(\mathcal{I}_{render}, \mathcal{I}_{\text{mask}}) \tag{13}$$

$$\mathcal{L}_F = \mathcal{L}_1(\mathcal{I}_{\text{fuse}}, \hat{\mathcal{I}}) + \lambda\,\mathcal{L}_{\text{SSIM}}(\mathcal{I}_{\text{fuse}}, \hat{\mathcal{I}}) + \gamma\,\mathcal{L}_{\text{LPIPS}}(\mathcal{I}_{\text{fuse}}, \hat{\mathcal{I}}) \tag{14}$$

其中 λ和 γ 均为超参数 .

方法	常规 / 情感					
	LMD(↓)	峰值信噪比(↑)	LPIPS(↓)	结构相似性指数(↑)	Sync-C(↑)	E-FID(↓)
StableAvatar '25	4.117 / 7.150	18.403 / 19.290	0.258 / 0.228	0.480 / 0.619	4.421 / 3.972	0.546 / 0.430
TalkingGaussian '24	3.018 / 5.934	26.943 / 25.533	0.045 / 0.096	0.906 / 0.892	5.011 / 4.886	0.089 / 0.356
EDTalk '24	3.827 / 6.548	25.627 / 18.061	0.073 / 0.297	0.888 / 0.864	**6.173 / 7.550**	0.483 / 0.668
DEGSTalk '25	**1.960/ 3.923**	27.104 / 28.051	0.042 / 0.162	0.891 / 0.924	5.663/5.007	0.076 / 0.154
Our	2.492 / 5.407	**28.923 / 28.408**	**0.032 / 0.067**	**0.928/ 0.938**	5.750 / 5.152	**0.072 / 0.145**
GT	-	-	-	-	7.832 / 8.163	-

表1: 两个数据集上的定量结果。最佳方法以**粗体**标出，次佳方法以 $underline$标出。

3. 训练与评估

3.1. 实验设置

数据集。 为评估我们工作的有效性，我们在两种类型的数据集上进行了实验：常规说话人脸数据集 [20] （包括 Obama 和 May 子集）以及情感数据集 [21]（使用 M003 和 M030 子集）。对于常规数据集，输入视频片段的帧率为 25 FPS，分辨率为 256 × 256。对于情感数据集，输入视频片段的帧率为 30 FPS，分辨率为 512 × 512。

预处理。 我们遵循 [7], [17] 中的预处理流程以及 [5]中的初始化设置。为减少背景杂波干扰和相机运动的影响（这些问题在 MEAD 数据集 [21],中尤为突出），我们采用了 [22]中的人脸窗口裁剪方法。**实现细节**。 我们的方法基于 PyTorch 实现。与结构相关的超参数将在各小节中介绍。在训练过程中，我们首先并行训练面部和内部嘴部分支 50,000 次迭代，然后对它们进行 10,000 次迭代的联合微调。所有模块均使用 Adam [23] 和 AdamW [24] 优化器，最终损失函数中的 γ 和 λ 分别设为 0.2 和 0.5。

评估指标。我们采用多种成熟指标评估模型性能，涵盖任务的关键要求：用于评估面部几何精度的地标距离（LMD） [25]、唇形同步置信度分数（Sync-C） [26], 、用于评估高频细节保真度的 LPIPS [27] 、用于评估面部结构一致性的 SSIM [28] ，以及用于评估图像级情感特征对齐的 E-FID [29] 。

3.2. 对比实验

我们将 UA-3DTalk 与四种最先进方法进行比较，包括 TalkingGaussian [5], 、DEGSTalk [30], 、StableAvatar [31] 和 EDTalk [22]。表1中的定量结果表明，在常规数据集和情感数据集上，我们的方法在大多数指标上均优于这些方法：E-FID 比 EDTalk 高 5.2%，SyncC 比 DEGSTalk 高 3.1%，LPIPS 比其他方法低 0.015，图像质量更优。值得注意的是，LMD 最佳的 DEGSTalk 使用了若干真实3DMM参数，而 SyncC 具有竞争力的 EDTalk 依赖参考视频；相比之下，我们的 UA-3DTalk 在纯音频驱动场景中表现出色，且不依赖这些条件。

为进一步验证 UA-3DTalk 的定性性能，我们在图 3 中展示了两个数据集上的关键帧对比。我们突出显示了嘴部和眼部区域（各行中的位置保持一致，以便进行针对性评估）。可以看出，我们的方法在三个核心方面优于其他方法：唇形同步更准确、眼部运动控制更自然、与情绪相关的表情预测更精准。这些动态面部区域的一致性和精确性进一步证实了所提出框架的鲁棒性和有效性。

3.3. 消融研究

我们开展了消融研究，以量化 UA-3DTalk 中每个核心模块的贡献。测试配置包括：

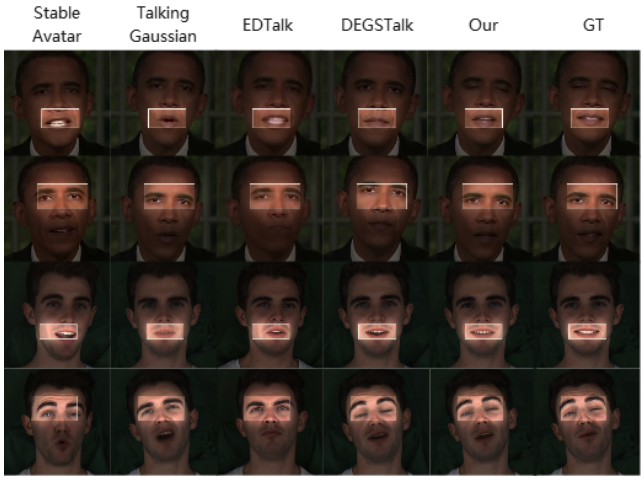

图 3: 对比实验的视觉结果。

基线（不包含任何提出的模块）、先验提取（P）、情感蒸馏（E）以及基于不确定性的变形（U）。结果见表 2。需要特别指出的是，U 无法单独应用于基线，这是因为基线仅使用单视图特征，而 U 依赖多视图特征进行不确定性估计。因此，我们在与 P 集成的配置基础上测试 U，以便与“w/ P”设置进行比较。

	LMD↓	峰值信噪比↑	LPIPS↓	结构相似性指数↑	Sync-C↑	E-FID↓
基线	5.934	25.53	0.096	0.892	4.886	0.356
w/ P	5.878	28.03	0.069	0.931	4.973	0.203
w/ E	5.872	25.52	0.095	0.894	4.897	0.312
w/ P,U	5.691	28.30	0.068	0.935	5.010	0.178
w/ P,E,U	**5.407**	**28.40**	**0.067**	**0.938**	**5.152**	**0.145**

表 2: MEAD 上的消融研究结果

如表 2 所示，P 通过生成音频同步动作单元和表情，显著提升了性能，验证了其有效性。U 在不引入额外信息的情况下，稳定提升了渲染质量，证实了我们提出的基于不确定性的融合策略是有效的。尽管模块 E 的贡献不如模块 P，但它显著增强了情绪相关微表情和最终帧；无论单独使用还是与模块 P 配合使用，情绪表达指标都充分证明了其有效性。

4. 结论

本文提出了 UA-3DTalk，旨在解决三维情感说话人脸合成中的两大关键挑战：音频情感提取困难且表情控制能力弱，以及存在缺陷的“一刀切”多视图融合。该方法首次在这一领域引入不确定性建模，其先验提取和情感蒸馏模块增强了音频-视觉情感对齐。实验表明，UA-3DTalk 达到了最先进性能，在 E-FID、SyncC 和 LPIPS 指标上分别比最先进方法提升了 5.2%、3.1%，并降低了 0.015。

5. 致谢

本研究得到非传染性慢性疾病国家科技重大专项（2024ZD0525800）的资助。

6. 参考文献

[1] Rebecca Mobbs、Dimitrios Makris 和 Vasileios Argyriou，《情绪识别与生成：面部、语音和文本模态的全面综述》，2025年。[2] Shuai Tan、Bin Ji 和 Ye Pan，《Flowvqtalker：通过归一化流和量化实现高质量情感说话人脸生成》，载于《*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2024年6月，第26317–26327页。[3] Haotian Wang、Yuzhe Weng、Yueyan Li、Zilu Guo、Jun Du、Shutong Niu、Jiefeng Ma、Shan He、Xiaoyan Wu、Qiming Hu、Bing Yin、Cong Liu 和 Qingfeng Liu，《Emotivetalk：通过音频信息解耦和情感视频扩散实现富有表现力的说话头生成》，载于《*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2025年6月，第26212–26221页。[4] Zhilei Liu、Xiaoxing Liu、Sen Chen、Jiaxing Liu、Longbiao Wang 和 Chongke Bi，《利用语音相关面部动作单元进行说话人脸生成的多模态融合》，《*ACM*多媒体计算、通信与应用汇刊》，第20卷，第9期，2024年9月。[5] Jiahe Li、Jiawei Zhang、Xiao Bai、Jin Zheng、Xin Ning、Jun Zhou 和 Lin Gu，《Talkinggaussian：通过高斯泼溅实现结构持久的三维说话头合成》，*arXiv*预印本 *arXiv:2404.15264*，2024年。[6] Xinya Ji、Hang Zhou、Kaisiyuan Wang、Qianyi Wu、Wayne Wu、Feng Xu 和 Xun Cao，《EAMM：基于音频情绪感知运动模型的单次情感说话人脸生成》，载于*ACM SIGGRAPH 2022*会议论文集，美国纽约州纽约，2022年，SIGGRAPH '22，计算机协会。[7] Xiaoxing Liu、Zhilei Liu 和 Chongke Bi，《Nerf-3dtalker：结合三维先验辅助音频解耦的神经辐射场说话头合成》，载于*ICASSP 2025——2025*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），2025年，第1–5页。[8] Duomin Wang、Yu Deng、Zixin Yin、Heung-Yeung Shum 和 Baoyuan Wang，《用于细粒度可控说话头合成的渐进式解耦表示学习》，载于《*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2023年6月，第17979–17989页。[9] Guanjun Wu、Taoran Yi、Jiemin Fang、Lingxi Xie、Xiaopeng Zhang、Wei Wei、Wenyu Liu、Qi Tian 和 Xinggang Wang，《用于实时动态场景渲染的四维高斯泼溅》，载于《*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2024年6月，第20310–20320页。[10] Shuai Shen、Wanhua Li、Yunpeng Zhang、Yap-Peng Tan 和 Jiwen Lu，《Audio-plane：用于实时说话头合成的音频因子分解平面高斯泼溅》，2025年。[11] Yifan Xie、Fei Ma、Yi Bin、Ying He 和 Fei Yu，《通过联合不确定性学习实现音频驱动的说话人脸视频生成》，2025年。[12] Riccardo Barbano、Simon Arridge、Bangti Jin 和 Ryutaro Tanno，《第26章——医学图像合成中的不确定性量化》，载于《生物医学图像合成与仿真》，MICCAI学会丛书，第601–641页。学术出版社，2022年。[13] Ryutaro Tanno、Aurobrata Ghosh、Francesco Grussu、Enrico Kaden、Antonio Criminisi 和 Daniel C. Alexander，，"载于《医学图像计算与计算机辅助干预——*MICCAI 2016*》，Cham，2016年，第265–273页，施普林格国际出版。[14] K R Prajwal、Rudrabha Mukhopadhyay、Vinay P. Namboodiri 和 C.V. Jawahar，《野外语音到唇形生成，只需一个唇形同步专家》，载于《第*28*届*ACM*国际多媒体会议论文集，美国纽约州纽约，2020年，MM '20，第484–492页，计算机协会。[15] Wenxuan Zhang、Xiaodong Cun、Xuan Wang、Yong Zhang、Xi Shen、Yu Guo、Ying Shan 和 Fei Wang，《SadTalker：为风格化音频驱动单图说话人脸动画学习逼真的三维运动系数》，载于《*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2023年6月，第8652–8661页。[16] Alexei Baevski、Henry Zhou、Abdelrahman Mohamed 和 Michael Auli，《wav2vec 2.0：自监督语音表征学习框架》，载于第*34*届神经信息处理系统国际会议论文集，美国纽约州雷德胡克，2020年，NIPS '20，Curran Associates Inc.。[17] Heqing Zou、Yuke Si、Chen Chen、Deepu Rajan 和 Eng Siong Chng，《基于协同注意力的多层次声学信息语音情感识别》，载于*ICASSP 2022——2022*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），2022年，第7367–7371页。[18] Carlos Busso、Murtaza Bulut、Chi-Chun Lee、Abe Kazemzadeh、Emily Mower、Samuel Kim、Jeannette N. Chang、Sungbok Lee 和 Shrikanth S. Narayanan，《IEMOCAP：交互式情感双人动作捕捉数据库》，《语言资源与评估》，第42卷，第4期，第335–359页，2008年。[19] Sara Fridovich-Keil、Giacomo Meanti、Frederik Rahbæk Warburg、Benjamin Recht 和 Angjoo Kanazawa，《K-Planes：空间、时间和外观中的显式辐射场》，载于《*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2023年6月，第12479–12488页。[20] Yudong Guo、Keyu Chen、Sen Liang、Yong-Jin Liu、Hujun Bao 和 Juyong Zhang，《AD-NeRF：用于说话头合成的音频驱动神经辐射场》，载于《*IEEE/CVF*计算机视觉国际会议（*ICCV*）论文集，2021年10月，第5784–5794页。[21] Kaisiyuan Wang、Qianyi Wu、Linsen Song、Zhuoqian Yang、Wayne Wu、Chen Qian、Ran He、Yu Qiao 和 Chen Change Loy，《MEAD：用于情感说话人脸生成的大规模音视频数据集》，载于*ECCV*，2020年8月。[22] Shuai Tan、Bin Ji、Mengxiao Bi 和 Ye Pan，《EDTalk：用于情感说话头合成的高效解耦》，载于《计算机视觉——*ECCV 2024*》，Cham，2025年，第398–416页，施普林格·自然瑞士。[23] Diederik P. Kingma 和 Jimmy Ba，《Adam：一种随机优化方法》，2017年。[24] Ilya Loshchilov 和 Frank Hutter，《解耦权重衰减正则化》，2019年。[25] Lele Chen、Zhiheng Li、Ross K Maddox、Zhiyao Duan 和 Chenliang Xu，《唇部动作生成一览》，载于《欧洲计算机视觉会议（*ECCV*）论文集，2018年9月。[26] Joon Son Chung 和 Andrew Zisserman，《超越时间：野外自动唇形同步》，载于《计算机视觉——*ACCV 2016*研讨会，Cham，2017年，第251–263页，施普林格国际出版。[27] Richard Zhang、Phillip Isola、Alexei A. Efros、Eli Shechtman 和 Oliver Wang，《深层特征作为感知度量为何如此有效》，载于《*IEEE*计算机视觉与模式识别会议（*CVPR*）论文集，2018年6月。[28] Zhou Wang、A.C. Bovik、H.R. Sheikh 和 E.P. Simoncelli，《图像质量评估：从误差可见性到结构相似性》，《*IEEE*图像处理汇刊》，第13卷，第4期，第600–612页，2004年。[29] Linrui Tian、Qi Wang、Bang Zhang 和 Liefeng Bo，《EMO：弱条件下利用音频转视频扩散模型生成富有表现力的肖像视频》，载于《计算机视觉——*ECCV 2024*》，Cham，2025年，第244–260页，施普林格·自然瑞士。[30] Kaijun Deng、Dezhi Zheng、Jindong Xie、Jinbao Wang、Weicheng Xie、Linlin Shen 和 Siyang Song，《DEGSTalk：用于保发型说话人脸合成的逐嵌入高斯场分解》，载于*ICASSP 2025——2025*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），2025年，第1–5页。[31] Shuyuan Tu、Yueming Pan、Yinming Huang、Xintong Han、Zhen Xing、Qi Dai、Chong Luo、Zuxuan Wu 和 Yu-Gang Jiang，《StableAvatar：无限时长音频驱动虚拟形象视频生成》，2025年。