arXiv:2512.15262v1 [eess.IV] 17 Dec 2025

面向生成式人脸视频编码的视听跨模态压缩

许佑民[‡§]，郭梦溪[†§]，赵世杰[†*]，李伟奇[‡]，李俊林[†]，张莉[†]，张建[‡]

[†]字节跳动，中国深圳及美国加利福尼亚州圣迭戈
[‡]北京大学电子与计算机工程学院

{guomengxi.qolab, zhaoshijie.0526, lijunlin.li, lizhang.idm}@bytedance.com
youmin.xu@stu.pku.edu.cn, liweiqi@stu.pku.edu.cn, zhangjian.sz@pku.edu.cn

摘要

生成式人脸视频编码（GFVC）对于视频会议等现代应用至关重要，然而现有方法主要关注视频中的运动，而忽视了音频所占据的显著比特率。尽管音频和唇部运动之间存在广为人知的相关性，但这种跨模态一致性尚未被系统地用于压缩。为此，我们提出了一种**视听跨模态压缩（AVCC）**框架，对音频和视频流进行联合压缩。我们的框架从视频中提取运动信息，并将音频特征标记化，然后通过统一的音视频扩散过程对二者进行对齐。借助共享表征，该框架能够同步重建两种模态。在极低比特率场景下，AVCC甚至可以利用一种模态重建另一种模态。实验表明，AVCC在率失真性能方面显著优于多功能视频编码（VVC）标准和最先进的生成式人脸视频编码方案，为构建更高效的多模态通信系统铺平了道路。

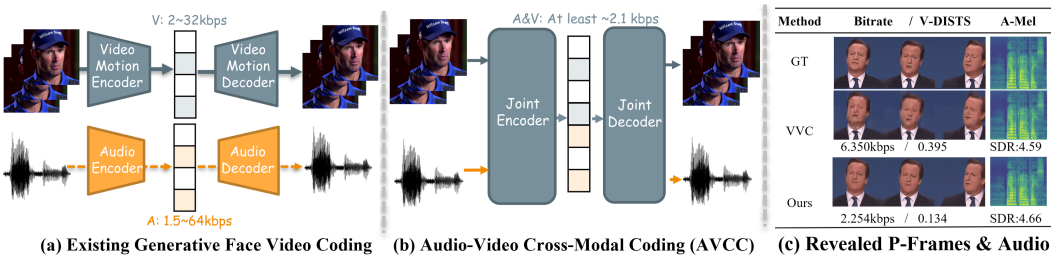
引言

互联网视频的激增使高效压缩变得至关重要。H.264/AVC [1] 和VVC [2] 等传统编解码器在信号层面运行，而近期的生成式视频编码（GVC）方法则利用语义表征来实现更优的率失真（RD）性能，尤其适用于具有强先验模式的人脸视频 [3]。当前的生成式人脸视频编码（GFVC）框架 [4, 5, 6] 将视频转换为关键点或特征等紧凑表征，然后由解码器利用这些表征重建逼真的视频。

然而，这些方法忽略了超低比特率通信中的一个关键方面：音频流。先进的音频编解码器 [7, 8] 降低了音频比特率，但音频仍占总带宽的很大一部分。现有的生成式人脸视频编码框架对音频和视频分别进行压缩，未能利用二者之间固有的强相关性。唇读 [9] 和说话人头生成 [10] 领域的研究表明，可以通过唇部运动推断语音，反之亦然。

为弥合这一差距，我们提出**视听跨模态压缩（AVCC）**，这是一种新颖的联合视听压缩框架。如图 1 所示，

[§]贡献相同。
[*]通讯作者。



(a) Existing Generative Face Video Coding (b) Audio-Video Cross-Modal Coding (AVCC) (c) Revealed P-Frames & Audio

图1：对比（a）现有的生成式人脸视频编码方法（分别处理音频和视频）与（b）我们提出的 AVCC 框架。AVCC 从音频和视频的时序特征中构建互表示，以辅助两个流的联合编码与解码。

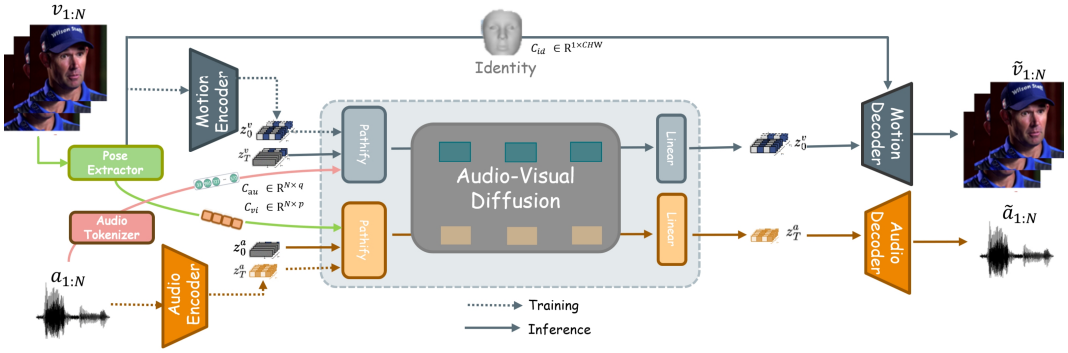


图2：我们提出的 AVCC 方法的整体框架。编码器利用跨模态信息对音频和视频输入进行联合压缩。解码器使用统一扩散过程重建同步的音频和运动向量，并以此驱动从传输的关键帧生成视频帧。

AVCC 从视频中提取运动信息，并从音频中提取令牌化特征。随后，这些表征在统一的**视听扩散（AVDiff）**模型中融合，该模型从噪声中生成同步的音频和视频输出。通过利用这种跨模态冗余，AVCC 相较于独立编码方法实现了更高的压缩效率。我们的主要贡献如下：

- 我们提出了 AVCC，这是首个将音频和视频整合到统一压缩流程中的生成式人脸视频编码框架，并利用二者的时序相关性提升效率。

- 我们引入了 AVDiff 模型，它能够有效地从共享潜在空间中对齐并重建音频和视频流，即使在极低比特率下也能实现高质量合成。

- 实验结果表明，AVCC 的性能显著优于 VVC 和其他最先进的生成式人脸视频编码方法，充分体现了跨模态压缩对下一代通信系统的价值。

相关工作

人脸视频编码

生成式视频编码不同于传统的混合编码框架，其重点在于实现高效、灵活的数据表示。受早期基于模型的编码启发，这些方法着重提取紧凑的空间和时间表示。与早期方法不同，当前的生成式人脸视频编码技术主要基于先进的人脸动画和重演模型，依靠深度学习强大的推理能力来确保高质量的视频重建。尽管紧凑表示效率较高，但其输出维度可能固定，难以适应不断变化的网络状况，从而降低编码灵活性。

视听表征

跨模态数据压缩旨在去除不同媒体类型之间的冗余。常见做法是将数据映射到共享的语义空间中，以获得统一且紧凑的表示。视听隐藏单元BERT（AV- HuBERT）[15] 是一种自监督模型，旨在从音频和视觉数据中学习鲁棒表征。该模型处理来自口部区域的视觉输入和作为梅尔频率倒谱系数的音频输入，并利用Transformer模块将二者融合，从而学习全面的视听语音表征。与SyncNet等早期模型相比，[16],AV-HuBERT的表现更加稳定，因此成为实现高质量视听同步的可靠选择，也为我们的联合压缩方法提供了启发。

方法

总体编码/解码流程

如图 2 所示，我们的 AVCC框架将视频和音频编码为一个统一的比特流，其中包含三个部分：身份码（C_{id}）、姿态码（C_{vi}）和音频词元（C_{au}）。

编码： 身份编码 C_{id} 从第一帧（I 帧）中提取，用于表示静态面部特征。对于后续帧（P 帧），**姿态提取器**会计算运动参数，并将其编码为姿态编码 C_{vi}。与此同时，**音频分词器** 会将音频流转换为离散的音频词元 C_{au}。

解码： 解码器使用身份编码 C_{id} 来初始化面部结构。姿态编码 C_{vi} 和音频词元 C_{au} 会输入我们的**视听扩散（AVDiff）**模型。该模型联合重建音频信号和稠密运动场，随后利用稠密运动场驱动初始身份帧并生成最终的视频输出。这一联合过程可确保同步，并利用跨模态信息实现更高质量的重建。

音频分词器和视频姿态提取器

音频分词器。 我们的音频编码方法采用双编码器架构，以高效提取多样的音频特征。语义编码器基于

自监督音频掩码自编码器（AudioMAE），用于提取高层音频特征。随后，采用集成式 k 均值聚类方法将这些特征量化为离散的语义标记。声学编码器则通过提取语义特征未能表示的残余音频细节，对此进行补充。两个编码器的输出经过拼接，形成全面的表征，并转换为离散索引。重建时，采用基于扩散模型的解码器，根据这些索引生成清晰的音频。

视频姿态提取器。我们的姿态提取器分别对静态身份和动态运动进行编码。从 I 帧中提取身份编码 C_{id}。对于 P 帧，则计算描述头部姿态、唇部运动和眼部运动的运动参数，并将其编码为姿态编码 C_{vi}。我们采用潜在图像动画器（LIA），将运动分解为特定于身份和特定于运动的潜在编码，从而更稳健、紧凑地表示面部动态。

用于联合解码的音视频扩散

解码器的核心是 AVDiff 模型，它通过联合生成压缩表示中的音频和视频，实现时序多模态对齐。我们针对这一任务改造了扩散 Transformer（DiT）架构。模型从高斯噪声开始，迭代地对音频（z_a）和视频（z_v）的潜在表示进行去噪。具体而言，给定成对噪声（z_a^T， z_v^T），联合去噪网络 θ_{av} 对逆向过程进行建模：

$$p_{\theta_{av}}(z_a^{t-1}|z_a^t, z_v^t) = \mathcal{N}(z_a^{t-1}; \mu_{\theta_{av}}(z_a^t, z_v^t, t), \sigma_t^2 I), \tag{1}$$
$$p_{\theta_{av}}(z_v^{t-1}|z_v^t, z_a^t) = \mathcal{N}(z_v^{t-1}; \mu_{\theta_{av}}(z_v^t, z_a^t, t), \sigma_t^2 I), \tag{2}$$

其中 t 表示扩散步骤。训练目标是最小化预测噪声与真实噪声之间的误差：

$$\mathcal{L}_{\theta_{av}} = \mathbb{E}_{z_v^t, z_a^t, t, \epsilon_v, \epsilon_a} \left[\|\epsilon_v - \epsilon_{\theta_{av}}(z_v^t, z_a^t, t)\|_2^2 + \|\epsilon_a - \epsilon_{\theta_{av}}(z_a^t, z_v^t, t)\|_2^2 \right]. \tag{3}$$

音视频对齐模块。 为促进有效的跨模态交互，我们在 DiT 模块中引入了一个 **AV-CrossAttn**模块。该模块与多头自注意力（MHSA）模块具有类似的结构，但拥有独立的可训练参数。在交叉注意力机制中，音频特征 x_a 和视频特征 x_v 被用作查询、键和值，使每种模态都能够关注另一种模态的特征。在每个去噪步骤中进行这种显式交互，使模型能够学习语音与面部运动之间复杂的相关性，从而生成高度同步且自然的输出。

熵编码。 我们采用基于区间的算术编码器，根据量化令牌的预测概率对其进行压缩，生成最终的比特流。

损失与调优

为确保输出质量高且同步，我们采用多分量损失函数。我们使用预训练的 AV-HuBERT 模型计算唇形同步损失 L_{sync}，以衡量音频与视觉（唇部区域）特征之间的余弦相似度，从而确保

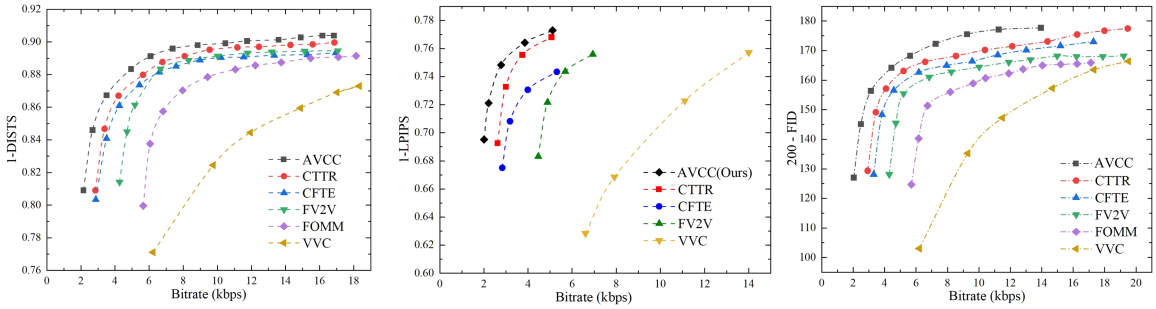


图 3：率失真（RD）性能对比。使用三种不同的感知指标将 AVCC 与 VVC 和最先进的生成式人脸视频编码方法进行比较：(a) 1-DISTS、(b) 1-LPIPS 和 (c) 200-FID。在所有指标下，我们的方法在比特率相近时始终能实现更高的重建质量（得分更高）。

嘴部运动准确。整体训练目标将其与感知损失 $\mathcal{L}_{per}$、对抗损失 $\mathcal{L}_{adv}$ 和特征匹配损失 $\mathcal{L}_{fea}$ 相结合，以监督端到端训练过程：

$$\mathcal{L}_{total} = \lambda_{per}\mathcal{L}_{per} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{fea}\mathcal{L}_{fea} + \lambda_{sync}L_{sync} \tag{4}$$

其中，权重参数通过经验设定，以平衡重建保真度、自然度和视听同步。

实验结果

实验设置

实现细节。 我们的框架使用 PyTorch 实现，并在 VoxCeleb2 数据集上训练 [17]，训练所用硬件为 NVIDIA A100 GPU。对于扩散过程，我们使用 Stable Diffusion 的预训练变分自编码器将视频帧编码到潜在空间，并使用预训练音频变分自编码器处理音频流。AVDiff 模型采用冻结的 DiT-XL/2 主干网络，并配备用于跨模态注意力的轻量级可训练层。

对比算法。 我们将 AVCC 与 VVC 标准 [2] 以及领先的 GFVC 方法进行比较：CTTR [18], CFTE [19], FV2V [**?**], 和 FOMM [4]。我们使用来自 VoxCeleb 测试集的 50 个测试视频，每个视频包含 250 帧，分辨率为 256×256。

评估指标。 我们使用感知指标 DISTS [20] 和 LPIPS [21],以及传统指标，如峰值信噪比和结构相似性指数。DISTS/LPIPS 越低表示质量越好。

性能比较

率失真性能。 图 3 展示了视频重建的率失真性能。AVCC 在各种感知指标上始终优于所有基线方法，包括 VVC 和其他生成式人脸视频编码。在极低比特率下（例如 2 千比特每秒），我们的方法仍能保持较高的感知质量（较低的 FID 分数），而竞争方法则会产生明显的伪影。这证明了

表 1：与其他 GFVC方法相比，AVCC 在视频通道上的 BD-Rate 节省（%）。结果统一为 256×256 分辨率。

方法	MEAD		Facestar		VoxCeleb	
	DISTS	峰值信噪比	DISTS	峰值信噪比	DISTS	峰值信噪比
CTTR	16.6	9.8	22.1	15.4	18.8	10.7
IFVC	28.1	17.6	33.2	15.9	31.9	18.4
FV2V	52.9	33.6	58.3	38.5	49.7	31.7
VVC	93.4	5.9	96.4	9.3	94.3	7.2

表 2：与最先进的音频编解码器相比的音频通道性能。梅尔频谱/短时傅里叶变换/词错误率是误差指标（↓），而 VIS 是质量评分（↑）。

模型	Kbps	MEL↓	短时傅里叶变换↓	VIS↑	WER↓
真实值	-	0.0	0.0	4.99	2.09
Encodec [22]	1.50	5.39	4.47	3.04	5.0
HiFi-Codec [23]	2.00	4.35	3.61	3.11	3.6
语义 [24]	1.40	4.45	3.79	3.33	3.4
AVCC（高）	1.40	**4.43**	**3.71**	**3.38**	**3.22**
AVCC（低）	0.36	5.11	3.92	2.87	19.3

我们的联合音视方法在压缩率与视觉保真度之间实现更佳平衡的有效性。表 1 进一步证实了这些优势，显示出在多个数据集上相较于基线方法显著的 BD-Rate 节省。

音频性能。 如表 2 所示，我们的方法同样实现了最先进的音频压缩性能。在 1.40 千比特每秒下，AVCC 在评估频谱距离（梅尔频谱、短时傅里叶变换）、质量（VIS）和词错误率（WER）的指标上，优于 Encodec 和 Semantic Codec 等专用音频编解码器。这是因为我们的模型利用视觉线索（唇部运动）更准确地重建音频，而这是纯音频编解码器无法实现的优势。

主观质量与极端情况。 图 4 中的视觉对比验证了客观结果。在超低比特率下，AVCC 能够重建出表情精准、生动的面部，而基线方法则出现明显的伪影。此外，如图 6 所示，我们的方法在大角度头部旋转场景下展现出卓越的鲁棒性。不同于传统方法在此类条件下往往难以保持一致的面部结构，我们的方法能够确保较高的视觉保真度和时间连贯性，凸显了其处理复杂姿态的能力。

联合率失真分析。 图 5 展示了三维率失真曲面，呈现总比特率、视频质量（峰值信噪比）和音频质量（信号失真比）之间的相互作用。对于 CTTR，由于其对各模态进行独立压缩，曲面平滑，未体现模态间相关性。相比之下，AVCC 的曲面呈现出明显的曲率，表明跨模态信息有助于提升两种模态的压缩质量。右侧的等高线图进一步凸显了这一点：CTTR 未显示相关性，而 AVCC 则呈现出音频和视频质量之间明显的正相关，尤其是在低比特率下。这直观地证实了我们联合压缩方案的有效性。

Methods	V-Bitrate/ DISTS	GT	Methods	V-Bitrate/ DISTS	GT
FOMM	4.850kbps 0.345		VVC	4.321kbps 0.402	
F2F-RHO	4.530kbps 0.303		IFVC	2.830kbps 0.143	
FV2V	3.054kbps 0.223		Ours (8kp)	2.823kbps 0.129	
CFTE	2.983kbps 0.165		Ours (2kp)	0.821kbps 0.157	

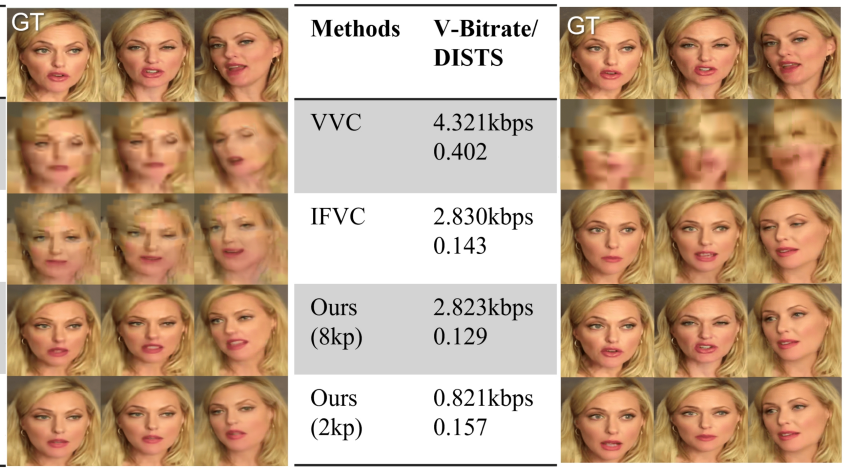


图 4：视觉质量对比。在 2 千比特每秒的超低比特率下，与基线方法相比，我们的 AVCC 方法能更有效地保留面部细节和表情，从而实现更好的感知质量（DISTS分数更低）。

表 3：针对 AV-CrossAttn模块的消融研究。比较在不同视频比特率（由令牌速率控制）下，采用（✓）和不采用（×）跨模态注意力时的性能。

令牌/秒 AV-	CrossAttn	速率-V	FID↓	唇形同步↑	峰值信噪比↑
8	×	1913	22.49	6.433	24.13
4	×	1109	22.87	4.230	23.13
2	×	578	23.14	3.453	20.23
8	✓	1913	**22.05**	**7.279**	**29.02**
4	✓	1109	**22.30**	**7.123**	**27.94**
2	✓	578	**22.43**	**6.890**	**22.41**

消融与复杂度分析

消融研究。 我们开展了消融研究，以验证框架中的关键组件。表 3 考察了 AV-CrossAttn模块的影响。我们将完整模型与禁用音视频条件的变体进行比较。结果表明，启用 AV-CrossAttn 可显著提升性能，尤其是在较低比特率下。例如，在视频速率为 578 千比特每秒时，跨模态注意力使 FID 分数提高了 0.71，并大幅改善了唇部同步（LipSync分数从 3.453 提升至 6.890）。这证实了利用音频信息对于重建高保真、同步的面部运动至关重要。

计算复杂度。 与传统编解码器相比，我们的方法存在计算复杂度较高这一局限。如表 4 所示，基于扩散的解码器（AVCC-Diffusion）速度较慢。不过，我们的多模态编码框架具有很强的灵活性。我们通过将解码器替换为传统的 3DMM模型（AVCC-3DMM）验证了这一点。该配置的实时编码/解码速度与以 CTTR 为代表的既有 GFVC方法相当，同时凭借联合音频-视觉运动表征的高效性，仍能保持更优的质量和比特率表现。

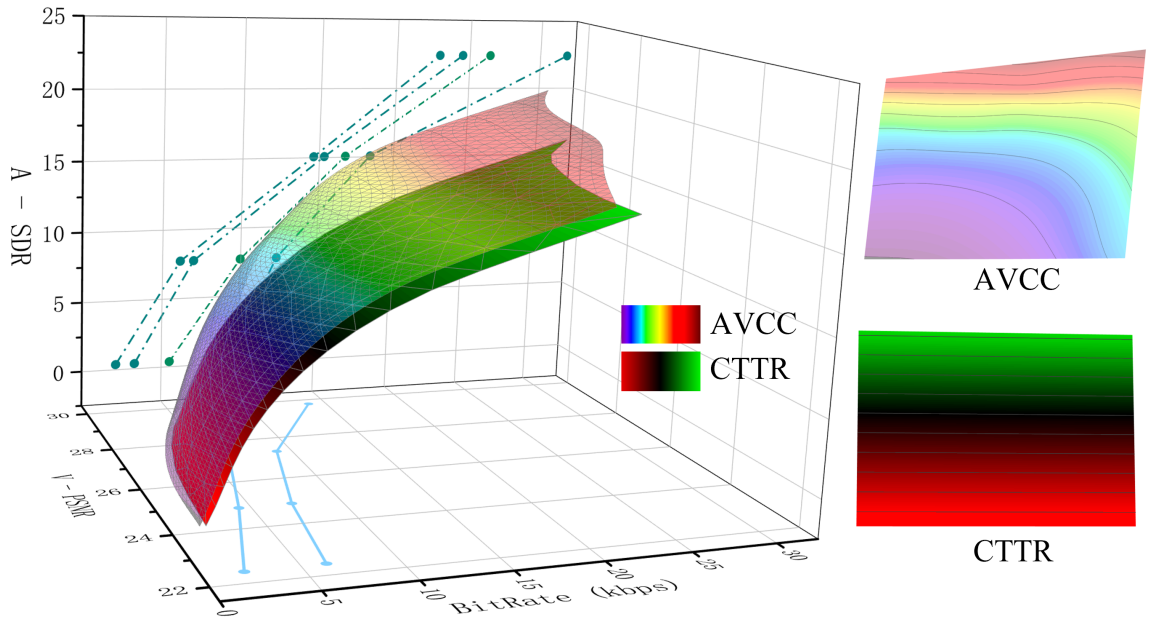


图 5：展示比特率（X）、视频峰值信噪比（Y）和音频信号失真比（Z）之间关系的三维率失真（RD）曲面。AVCC 曲面的扭曲表明音频和视频质量之间存在很强的相关性。等高线图（右侧）证实了这一点：CTTR 显示两者之间没有相关性，而 AVCC 则显示出明显的正相关关系，尤其是在低比特率下（红色区域）。

表 4：各方法在时间开销（帧率）、感知质量（1-LPIPS）和比特率方面的比较。

方法	帧率 1-LPIPS↑		比特率（千比特每秒）
VVC	0.67	0.619	7230
CTTR	0.22	0.695	3293
AVCC-Diffusion	0.13	**0.713**	**2530**
AVCC-3DMM	**0.23**	0.702	2854

潜在应用场景

所提出的多模态编码框架支持多种先进应用。在视频会议中，它可以利用可用模态进行补偿，确保传输不中断并应对丢包——如果音频有噪声，视频可以帮助预测音频词元，反之亦然。在直播中，它可以根据音频流估算缺失的视频数据，从而灵活应对网络波动。此外，它还为无障碍应用带来了新的可能，例如结合音频和视觉唇读线索生成实时字幕，以及创建逼真的虚拟化身，使其面部表情能够由音频或文本输入同步驱动。



图 6：大角度旋转下的视觉质量对比。与其他方法可能失效的情况不同，我们的方法展现出更强的一致性和更优的视觉质量，能够有效保留面部结构和运动动态。

结论

我们提出了 AVCC，这是一种新颖的生成式人脸视频编码框架，首次将音频和视频流整合到统一的压缩流水线中。通过共享扩散模型利用音频与视觉模态之间的内在相关性，我们的方法显著提升了压缩效率和重建质量，尤其是在超低码率场景下。实验结果表明，在客观和主观评价中，AVCC 均优于最新的 VVC 标准和其他先进的生成式人脸视频编码算法。本研究凸显了跨模态学习在未来视频通信中的潜力，并为开发更先进的多模态压缩系统奠定了坚实基础。

参考文献

[1] G. Sullivan 等，“高效视频编码（HEVC）标准概述”，*IEEE*视频技术电路与系统汇刊，第22卷，第12期，第1649–1668页，2012年。[2] B. Bross 等，“多功能视频编码（VVC）标准概述及其应用”，*IEEE*视频技术电路与系统汇刊，第31卷，第10期，第3736–3764页，2021年。[3] B. Chen等，“生成式人脸视频编码技术与标准化进展：综述”，载于数据压缩会议，2024年，第103–112页。[4] A. Siarohin 等，“用于图像动画的一阶运动模型”，神经信息处理系统进展，第32卷，第7137–7147页，2019年。[5] T. Wang等，“用于视频会议的单样本自由视角神经说话人头部合成”，载于*IEEE*计算机视觉与模式识别会议论文集，2021年，第10039–10049页。[6] B. Chen 等，“用于说话人脸视频压缩的紧凑时序轨迹表示”，*IEEE*视频技术电路与系统汇刊，第33卷，第11期，第7009–7023页，2023年。

[7] A. Van Den Oord、O. Vinyals 等，“神经离散表示学习”，《神经信息处理系统进展》，第30卷，2017年。[8] N. Zeghidour、A. Luebs、A. Omran、J. Skoglund 和 M. Tagliasacchi，“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》，2021年。[9] Y. M. Assael、B. Shillingford、S. Whiteson 和 N. de Freitas，“Lipnet：端到端句子级唇读”，arXiv预印本 arXiv:1611.01599，2016年。[10] S. Chen、Z. Liu、J. Liu 和 L. Wang，“Dae-talker：利用扩散自编码器实现高保真语音驱动的说话人脸生成”，arXiv预印本 arXiv:2303.17550，2023年。[11] W. Schreiber 等，“合成高频——一种实验性电视带宽缩减系统”，《SMPTE期刊》，第68卷，第8期，第525–537页，1959年。[12] F.-T. Hong 等，“用于说话头视频生成的深度感知生成对抗网络”，载于 IEEE计算机视觉与模式识别会议论文集，2022年。[13] R. Rombach、A. Blattmann、D. Lorenz、P. Esser 和 B. Ommer，“利用潜在扩散模型进行高分辨率图像合成”，载于 IEEE/CVF计算机视觉与模式识别会议论文集，2022年，第10 684–10 695页。[14] R. Ming、Z. Huang、Z. Ju、J. Hu、L. Peng 和 S. Zhou，“视频预测综述：从确定性方法到生成式方法”，arXiv预印本 arXiv:2401.14718，2024年。[15] B. Shi、W.-N. Hsu、K. Lakhotia 和 A. Mohamed，“通过掩码多模态聚类预测学习音视频语音表示”，arXiv预印本 arXiv:2201.02184，2022年。[16] K. Prajwal、R. Mukhopadhyay、V. P. Namboodiri 和 C. Jawahar，“用于真实场景语音到唇形生成的唇音同步专家就够了”，载于第28届 ACM多媒体国际会议论文集，2020年，第484–492页。[17] A. Nagrani、J. S. Chung 和 A. Zisserman，“VoxCeleb：大规模说话人识别数据集”，载于 INTERSPEECH，2017年。[18] B. Chen、Z. Wang、B. Li、S. Wang 和 Y. Ye，“用于说话人脸视频压缩的紧凑时序轨迹表示”，《IEEE视频技术电路与系统汇刊》，第33卷，第11期，第7009–7023页，2023年。[19] B. Chen 等，“超越关键点编码：利用紧凑特征表示进行时间演化推断，以实现说话人脸视频压缩”，载于数据压缩会议，2022年，第13–22页。[20] K. Ding 等，“图像质量评估：统一结构与纹理相似性”，《IEEE模式分析与机器智能汇刊》，2020年。[21] R. Zhang 等，“深层特征作为感知度量的惊人有效性”，载于其他会议论文集，2018年。[22] A. D´efossez、J. Copet、G. Synnaeve 和 Y. Adi，“高保真神经音频压缩”，arXiv预印本 arXiv:2210.13438，2022年。[23] P. Andreev、A. Alanov、O. Ivanov 和 D. Vetrov，“Hifi++：神经声码器合成、带宽扩展和语音增强的统一框架”，arXiv预印本 arXiv:2203.13086，2022年。[24] H. Liu、X. Xu、Y. Yuan、M. Wu、W. Wang 和 M. D. Plumbley，“Semanticodec：面向通用声音的超低比特率语义音频编解码器”，《IEEE信号处理精选主题期刊》，2024年。