arXiv:2301.02379v2 [cs.CV] 3 Apr 2023

CodeTalker：基于离散运动先验的语音驱动三维面部动画

Jinbo Xing[1] Menghan Xia[2], * Yuechen Zhang[1] Xiaodong Cun[2] Jue Wang[2] Tien-Tsin Wong[1]香港中文大学 [2]腾讯人工智能实验室
{jbxing,yczhang21,ttwong}@cse.cuhk.edu.hk {menghanxyz,vinthony,arphid}@gmail.com

摘要

语音驱动的三维面部动画已得到广泛研究，但由于问题本身高度不适定，且音视频数据稀缺，实现逼真、生动的效果仍有差距。现有工作通常将跨模态映射建模为回归任务，因而受到回归均值问题的影响，生成的面部运动过于平滑。本文提出将语音驱动的面部动画转化为在学习所得码本的有限代理空间中执行代码查询任务，通过降低跨模态映射的不确定性，有效提升生成运动的生动性。该码本通过对真实面部运动进行自重建而学习得到，因此蕴含了真实的面部运动先验。在离散运动空间中，我们采用时间自回归模型，根据输入语音信号依次合成面部运动，从而保证唇形同步和合理的面部表情。实验表明，我们的方法在定性和定量评估中均优于当前最先进方法。此外，用户研究进一步证实了我们在感知质量方面的优势。代码和视频演示见：
h
t
tps://doubiiu.github.io/projects/codetalker。

1. 引言

数十年来，三维面部动画一直是活跃的研究课题，这得益于它在虚拟现实、电影制作和游戏中的广泛应用。语音与面部动作（尤其是唇部运动）之间高度相关，因此可以利用语音信号驱动面部动画。早期尝试主要致力于建立音素与其视觉对应物之间复杂的映射规则，但通常效果有限 [55,67]。随着深度学习的发展，近期的语音驱动面部动画技术显著推动了该领域的最新进展。然而，生成逼真的人类运动仍然颇具挑战。

作为一个适定性不佳的问题，语音驱动面部动画通常会对每个输入产生多个合理的输出。这种模糊性往往会导致结果过于平滑。无论如何，个体特定方法 [31, 51] 通常能够生成不错的面部运动，因为说话风格相对一致，但在通用应用中的可扩展性较差。最近，VOCA [10] 将这些方法扩展到不同身份，使其能够泛化；然而，这类方法生成的上半脸表情通常较弱或静止。这是因为VOCA将语音到运动的映射形式化为回归任务，从而倾向于生成平均化的运动，尤其是上半脸运动，因为它与语音信号的相关性较弱，甚至毫无相关性。为减少不确定性，FaceFormer [16] 利用基于Transformer的模型来整合长期音频上下文，并以自回归方式合成连续运动。尽管其性能有了显著提升，但它仍沿袭了一对一映射形式化定义的缺陷，并且缺少细腻的高频运动。与之不同，MeshTalk [52]为面部动画构建了类别型潜在空间，将音频相关信息与音频无关信息解耦，从而能够妥善处理这两个方面。然而，所采用的量化方法和类别型潜在空间表示并不适合运动先验学习，使训练过程变得困难，进而影响了模型性能。

我们从三维可变形人脸模型（3DMM） [37], 中获得灵感：该模型在低维空间中表示一般的面部表情。据此，我们提出将语音驱动面部动画表述为在学习得到的离散码本先验所构成的有限代理空间中进行代码查询任务。该码本通过使用向量量化自编码器（VQ-VAE） [61], 对真实面部运动进行自重建而学习得到；它与解码器共同存储真实的面部运动先验。与3DMM 的连续线性空间不同，码本项的组合构成了一个基数有限的离散先验空间。不过，在解码器的作用下，代码表示仍具有很强的表达能力。通过将语音映射到有限代理空间，语音到运动映射中的不确定性显著降低，从而提升运动合成的质量，

* 通讯作者。

。从概念上说，代理空间近似于面部运动空间，而学习得到的码本项则充当离散运动基元。

基于学习到的离散码本，我们提出了一种基于代码查询的时间自回归模型，用于语音条件下面部运动的合成，称为*CodeTalker*。具体而言，模型以语音信号为输入，以时间递归的方式预测动作特征标记。随后，这些特征标记用于查询离散空间中的代码序列，并据此重建面部运动。得益于对历史运动的上下文建模和跨模态对齐，所提出的 CodeTalker 能够生成准确的唇部动作和自然表情。大量实验表明，所提出的 CodeTalker 在现有数据集上表现出优越的性能。我们还开展了系统研究和实验，以验证本方法相较于以往工作的优势。本文的贡献如下：

- 我们以一种新颖的方式，使用离散基元对面部运动空间进行建模，从而有助于提升运动合成的真实感，抵御跨模态不确定性的影响。
- 我们提出了一种基于离散运动先验的时间自回归模型，用于语音驱动面部动画，其性能优于现有最先进方法。

2. 相关工作

2.1. 语音驱动的三维面部动画

计算机人脸动画是一项历史悠久的任务 [46]，过去几十年来受到的关注迅速增加 [5, 20, 32, 34, 36, 56, 66, 74]。作为一个分支，语音驱动的面部动画旨在根据输入语音序列，同步重现人物。尽管该领域的大量研究都聚焦于二维说话头部 [1, 7–9, 11, 23, 28, 29, 38, 39, 45, 49, 53, 63, 65, 69–71], 本文聚焦于三维模型上的面部动画，大致可分为基于语言学的方法和基于学习的方法。

基于语言学的方法。通常，基于语言学的方法 [12, 42, 55, 67] 会建立一套复杂的映射规则，将音素与其视觉对应物（即视觉音素） [19,35,43]联系起来。例如，支配函数 [42] 用于确定音素对相应面部动画控制参数的影响。Xu 等. [67] 为构建的标准视觉音素集合定义动画曲线，以生成同步的口部动作。此外，还有一些方法考虑了音素与视觉音素之间的多对多映射，如动态视觉音素模型 [55]以及近期提出的 JALI [12]。基于心理语言学考量，并以面部动作编码系统（FACS） [13], 为基础，JALI 将口部动作分解为唇部与下颌绑定动画，并生成逼真的协同发音效果。尽管这些方法能够对动画进行显式控制，但其流程复杂，且缺乏一种有原则的方法来驱动整张脸的动画。

基于学习的方法。基于学习的方法 [6, 10,16,17,24,31,40,48,54,64]采用数据驱动的框架。Cao 等人。 [6] 通过提出约束搜索和 Anime Graph 结构，实现情感唇形同步。最近，Taylor 等人。 [54] 提出了一种基于深度学习的模型，该模型在转录后的音素序列输入上采用滑动窗口方法。Karras 等人。 [31] 提出了一种带有可学习的情感数据库的基于卷积的网络，用于生成语音驱动的三维网格动画。最近，VisemeNet [73] 采用三阶段长短期记忆（LSTM）网络，预测下脸部唇部模型的动画曲线。

我们在此更具体地回顾与本工作最相关的研究，因为它们与本工作设定相同，即：在高分辨率配对音频-网格数据上训练，并以说话者独立的方式在顶点空间中驱动整个面部网格。MeshTalk [52] 借助类别型潜在空间，成功地将与音频相关和无关的面部信息解耦。然而，所采用的潜在空间并非最优，可表达能力有限，因此在数据稀缺场景下，动画质量不够稳定。VOCA [10]采用强大的音频特征提取模型，能够生成具有不同说话风格的面部动画。此外，FaceFormer [16] 利用 Transformer [62]考虑长期音频上下文，从而生成时间上稳定的动画。尽管生成的动画颇具吸引力，但二者都存在过度平滑问题，因为它们直接在高度病态、具有很大不确定性和歧义的音频-视觉映射中回归面部运动。

2.2. 离散先验学习

过去几十年来，基于学习字典的离散先验表示在图像修复任务中展现出优越性 [14, 22, 30, 57, 58], ，因为字典能够很好地保留清晰的图像细节。这类技术进一步启发了高容量、高压缩率的离散先验学习。VQ-VAE [61] 首次提出学习图像的离散表示（码本），并以自回归方式对其分布建模，以实现图像合成。后续工作 VQ-VAE2 [50]和 VQ-GAN [15] 进一步提升了高分辨率图像合成的质量。近年来，离散先验学习已应用于图像着色 [26], 、图像修复 [47], 、盲人脸修复 [72], 、文本到图像合成 [21],等。

除了图像模态之外，近期大多数研究也探索了离散先验学习在其他模态任务中的潜力，例如双人面部运动生成 [44], co-

语音手势合成 [2], 语音增强 [68]。受码本学习的启发，本研究探讨了如何为语音驱动的三维面部动画学习离散运动先验。不同于 [44], 我们以富含上下文信息的方式利用离散运动基元来表示面部运动，这种方式更有利于学习通用先验。

3. 方法

我们的目标是从语音信号合成连续的三维面部运动，使任意中性面部网格都能被驱动，生成唇形同步的说话面部。然而，这是一个不适定问题，因为同一段语音可能对应多种不同的面部动画。这种歧义往往会导致跨模态学习产生平均化的运动，且缺乏细微变化。为突破这一障碍，我们提出先利用学习到的离散运动先验对面部运动空间建模，再在该空间上学习语音条件时序自回归模型，从而提高对跨模态不确定性的鲁棒性。

形式化定义。 令$\mathbf{M}_{1:T} = (\mathbf{m}_1, ..., \mathbf{m}_T)$表示一个面部运动序列，其中每一帧$\mathbf{m}_t \in \mathbb{R}^{V\times 3}$ 表示相对于中性面部网格模板$\mathbf{h} \in \mathbb{R}^{V\times 3}$的 V 个顶点的三维位移。进一步令$\mathbf{A}_{1:T} = (\mathbf{a}_1, ..., \mathbf{a}_T)$表示一个语音片段序列，其中每个$\mathbf{a}_t \in \mathbb{R}^d$ 包含d 个采样点，与对应的（视觉）帧$\mathbf{m}_t$对齐。于是，我们的目标是根据$\mathbf{A}_{1:T}$逐帧合成$\mathbf{M}_{1:T}$，使任意中性面部模板$\mathbf{f}$ 都能被驱动为$\mathbf{H}_{1:T} = \{\mathbf{m}_1 + \mathbf{h} ..., \mathbf{m}_T + \mathbf{h}\}$。

3.1. 离散面部运动空间

视觉上逼真的面部动画应呈现准确的唇部动作和自然表情。为了从语音信号实现这一点，需要额外的运动先验来减少不确定性并补足逼真的运动成分。近期的图像复原任务表明， [72], 离散码本先验 [61] 即使面对严重退化的输入，也能保证高保真结果，展现出显著优势。受此启发，我们提出将面部运动空间建模为离散码本，并通过跟踪得到的真实世界面部运动进行学习。

运动基元码本。 我们成功学习到了一个码本 $\mathcal{Z} = \{\mathbf{z}_k \in \mathbb{R}^C\}_{k=1}^N$，使任意面部运动$\mathbf{m}_t$ 都能由一组分配的码本项$\{\mathbf{z}_k\}_{k\in\mathcal{S}}$表示，其中 $\mathcal{S}$ 表示通过公式选出的索引集合。1从概念上讲，码本项充当面部运动空间的运动基元。为此，我们预训练了一个基于Transformer的VQ-VAE，其中包含一个编码器
E、一个解码器 D以及一个上下文丰富的码本 $\mathcal{Z}$，通过对真实面部运动进行自重建。如图 1所示，面部运动$\mathbf{M}_{1:T}$ 首先被嵌入为时间特征 $\hat{\mathbf{Z}} = E(\mathbf{M}_{1:T}) \in \mathbb{R}^{T'\times H\times C}$，其中 H是面部组件的数量， T' 表示编码时间单元的数量（$P = \frac{T}{T'}$ 帧）。随后，我们

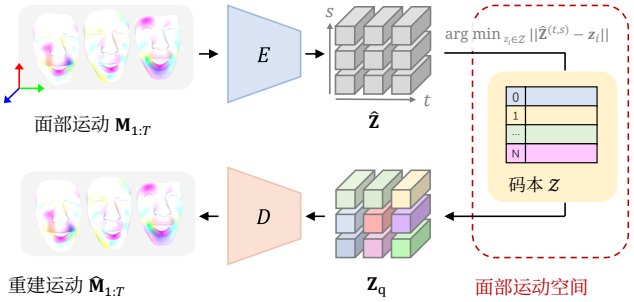


图1. 面部运动空间的学习框架。嵌入码本中的学习到的运动基元用于以空间和时间的方式表示面部运动。

通过逐元素量化函数 $Q(\cdot)$ 获得量化运动序列 $\mathbf{Z}_\mathbf{q} \in \mathbb{R}^{T'\times H\times C}$，该函数将$\mathbf{Z}$ 中的每个 $\hat{}$ 项映射到码本 $\mathcal{Z}$ 中与其最近的条目：

$$\mathbf{Z}_\mathbf{q} = Q(\hat{\mathbf{Z}}) := \arg\min_{\mathbf{z}_k\in\mathcal{Z}} \|\hat{\mathbf{z}}_t - \mathbf{z}_k\|_2. \tag{1}$$

然后，自重建表示为：

$$\hat{\mathbf{M}}_{1:T} = D(\mathbf{Z}_\mathbf{q}) = D(Q(E(\mathbf{M}_{1:T}))). \tag{2}$$

请注意，离散面部运动空间凭借有限的基数降低了映射歧义，同时得益于其作为潜在空间的丰富上下文表征，并未牺牲表达能力。

训练目标。 与 [61], 类似，为了监督量化自编码器的训练，我们采用运动级损失和两个中间编码级损失：

$$\begin{aligned}\mathcal{L}_{\text{VQ}} =& \|\mathbf{M}_{1:T} - \hat{\mathbf{M}}_{1:T}\|_1 \\ &+ \|\text{sg}(\hat{\mathbf{Z}}) - \mathbf{Z}_\mathbf{q}\|_2^2 + \beta\|\hat{\mathbf{Z}} - \text{sg}(\mathbf{Z}_\mathbf{q})\|_2^2,\end{aligned} \tag{3}$$

其中，第一项是重建损失，后两项用于通过缩小码本 $\mathcal{Z}$ 与嵌入特征$\hat{}\mathbf{Z}$之间的距离来更新码本项。sg(·) 表示停止梯度操作， β 是用于控制码本和编码器更新速率的加权因子。由于量化函数（公式1）不可微，因此采用直通梯度估计器 [4,61]将梯度从解码器输入复制到编码器输出。

讨论。 近期，Learn2Listen [44] 采用 VQ-VAE，根据给定的说话人头部合成面部表情，并利用二维单目视频获取3DMM系数。除了应用场景不同之外，我们还想强调两者的主要区别。首先，Learn2Listen构建说话人专属码本，而我们的方法使用通用码本，可用于表示任意面部运动。由于不同角色之间的运动被纳入其中，我们的码本天然蕴含了更加丰富的先验。其次，Learn2Listen利用

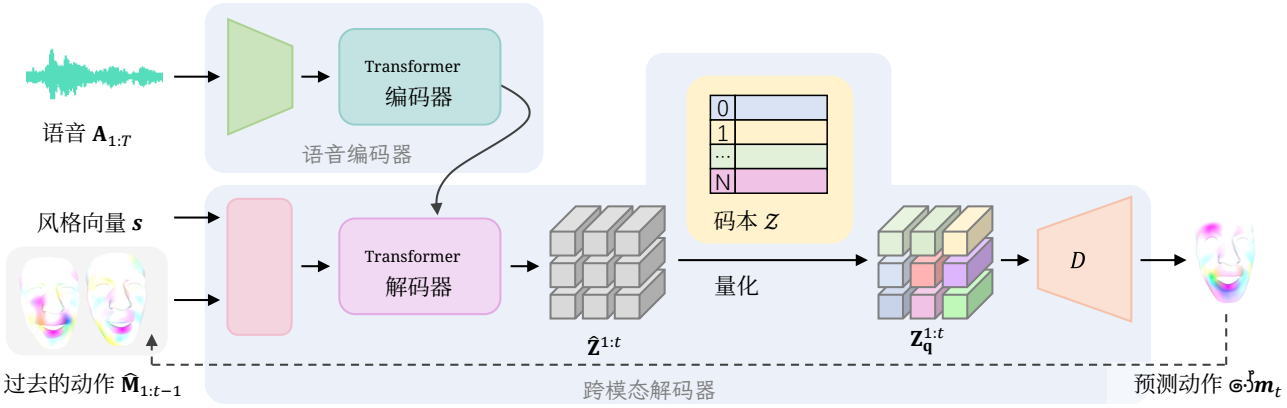


图2。我们的语音驱动运动合成模型示意图。给定语音$\mathbf{A}_{1:T}$和风格向量$\mathbf{s}$作为输入，模型通过预测运动代码，学习递归生成一系列面部运动。由于嵌入了学习良好的运动先验，预训练码本和解码器在训练期间保持冻结。

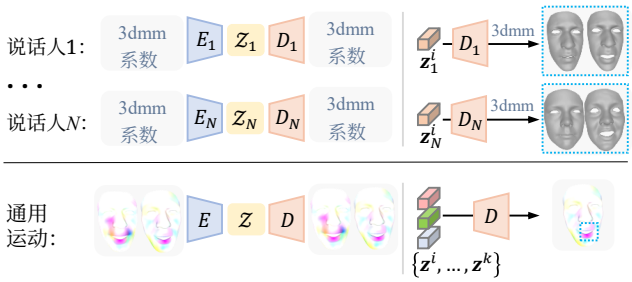


图3。与 Learn2Listen 的概念比较 [44]。（上）说话人特定的面部表情系数先验 [44], ，其中每个编码代表一系列面部表情系数。（下）我们的说话人无关的通用运动先验，其中每个编码代表面部组件的运动基元。蓝色虚线框表示每个编码在概念上可能代表的信息。

码本通过3DMM系数表示常见的面部表情序列， 即，每个编码表示一段（8帧）面部表情序列。相比之下，我们的码本旨在表示基于顶点的面部运动空间，其中的编码嵌入了面部各部分的逐顶点运动，并以丰富的上下文表示一个时间单元内的面部运动。如图3所示，Learn2Listen的码本旨在记忆三维形变模型空间中特定说话人的典型连续面部表情；由于三维形变模型的表现力有限，且受限于三维重建技术的精度，它无法合成具有细腻细节的逼真面部运动 [44]。作为首次尝试，我们的码本旨在通过运动基元表示采集到的面部网格数据所构成的通用面部运动空间，从而更有效地嵌入通用先验并保留生动的面部细节。

我们进一步讨论码本的超参数。首先，时间单元 P 的长度和面部组件 H 的数量分别决定了运动基元在时间和空间方面的复杂度。通常，复杂的运动基元灵活性和可重用性较低，从而妨碍表征效果。相反，过于简单的运动基元由于缺乏语义，会给运动预测带来困难。此外，码本大小 N 和特征维度 C 决定了表征能力，应根据数据集的复杂度并结合 P 和 H进行设定。在我们的实验中，我们根据数据集设置 $N = 256$， $P = 1$， $H = 8$或 $H = 16$，以及 $C = 64$或 $C = 128$，这些设置带来了高质量的结果，消融研究在第4.5节中对此进行了验证。更多细节请参见补充材料。

3.2. 语音驱动运动合成

利用学得的离散运动先验，我们可以构建从输入语音到目标运动代码的跨模态映射，再将其解码为逼真的面部运动。除了语音之外，我们还将说话风格作为控制输入，即一个风格向量$\mathbf{s} \in \mathbb{R}_{+}^{M} \cup \{0\}$，其中 M 是学得的风格空间的维度（见公式4）。以语音$\mathbf{A}_{1:T}$和风格向量$\mathbf{s}$为条件，我们采用由语音编码器 E_{speech} 和跨模态解码器 $D_{\text{cross-modal}}$组成的时间自回归模型，来学习面部运动空间，如图2所示。

遵循 FaceFormer [16], 的方法，我们的语音编码器采用了当前最先进的自监督预训练语音模型 wav2vec 2. 的架构。0 [3], 该模型由音频特征提取器和多层 Transformer 编码器组成。音频特征提取器通过时间卷积网络（TCN）将原始波形中的语音转换为特征向量。得益于高效的注意力机制，Transformer 编码器将音频特征转换为上下文化语音表示。除预训练码本和 VQ-VAE 解码器外，我们的跨模态解码器还包含一个嵌入模块和一个带因果自注意力的多层 Transformer 解码器。

嵌入模块通过以下方式组合过去的面部运动和风格嵌入：

$$\mathbf{F}_{\text{emb}}^{1:t-1} = \mathcal{P}_\theta(\hat{\mathbf{M}}_{1:t-1}) + \mathbf{B} \cdot \frac{\mathbf{s}}{\|\mathbf{s}\|_1}, \tag{4}$$

其中 $\mathcal{P}_\theta$ 是一个线性投影层，而 $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_M] \in \mathbb{R}^{C\times M}$ 表示 M 可学习的基向量，这些向量线性地张成风格空间。与 Face-Former 类似 [16], 我们为 Transformer 解码器配备因果自注意力，以学习上下文中各帧与过去面部运动序列之间的依赖关系；同时配备跨模态注意力，以对齐音频和运动模态。输出特征 $\hat{}\mathbf{Z}^{1:t}$ 随后通过量化得到 $\mathbf{Z}_\mathbf{q}^{1:t}$，具体见式1 ，再由预训练的 VQ-VAE 解码器解码。新预测的运动 $\hat{}\mathbf{m}_t$ 用于更新过去的 $\hat{}$运动，得到 $\mathbf{M}_{1:t}$，为下一次预测做准备。形式上，这一递归过程可写为：

$$\hat{\mathbf{m}}_t = D_{\text{cross-modal}}(E_{\text{speech}}(\mathbf{A}_{1:T}), \mathbf{s}, \hat{\mathbf{M}}_{1:t-1}). \tag{5}$$

训练目标。 我们训练 Transformer 编码器、解码器和用于跨模态映射的嵌入模块，同时冻结码本 $\mathcal{Z}$ 和动作解码器 D。为充分利用大规模语料中的语音表征学习成果，我们使用预训练的 wav2vec 2.0 权重初始化 TCN 和 Transformer 编码器。总体而言，自回归模型采用教师强制方案进行训练，并受两个损失项约束：（i）特征正则损失 $\mathcal{L}_{\text{reg}}$，用于衡量预测动作特征 $\hat{}\mathbf{Z}^{1:T}$ 与码本中的量化特征$\mathbf{Z}_\mathbf{q}^{1:T}$之间的偏差；（ii）动作损失 $\mathcal{L}_{\text{motion}}$，用于衡量预测动作$\mathbf{M}_{1:T}$ 与真实动作 $\mathbf{M}_{1:T}$之间的差异，这对稳定训练过程起着重要作用。最终损失函数为：

$$\begin{aligned}\mathcal{L}_{\text{syn}} &= \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{motion}} \\ &= \|\hat{\mathbf{Z}}^{1:T} - \text{sg}(\mathbf{Z}_\mathbf{q}^{1:T})\|_2^2 + \|\hat{\mathbf{M}}_{1:T} - \mathbf{M}_{1:T}\|_2^2.\end{aligned} \tag{6}$$

3.3. 训练细节

在第一阶段，我们在单张 NVIDIA V100 上训练 VQ-VAE 模型（图 1），共训练 200 个周期（∼2 小时），使用 AdamW [41]优化器（$\beta_1 = 0.9$， $\beta_2 = 0.999$ 和 $\epsilon = 1e-8$），学习率初始化为 10^{-4}，小批量大小设为 1。第二阶段，我们使用 Adam 优化器训练时间自回归模型 [33]。训练时长为 100 个周期（∼3 小时），其他超参数与第一阶段保持不变。

风格嵌入空间。 风格嵌入空间由 M 个学习得到的基向量线性张成，其中每种风格都由风格向量$\mathbf{s}$ 表示，该向量充当线性组合系数或坐标。在 VOCA [10], 中，它为每位说话人分配一个类似类别的独热风格向量。相反，我们提出一种通过若干可学习基向量（式 4）构建统一风格空间的概念，其中风格向量不再与说话人 ID 绑定，而是与每个说话样本相关联。训练期间，考虑到训练数据集的风格多样性有限，我们为每位说话人（例如编号 i）分配一个标准单位向量 $\mathbf{e}_i$ 作为风格向量，假设每位说话人对应一种独特且一致的风格。无论如何，推理时允许使用任意风格向量来插值生成新的说话风格。

4. 实验

4.1. 数据集与实现

我们采用两个广泛使用的数据集 BIWI [18] 和 VOCASET [10], , 在实验中训练和测试不同的方法。这两个数据集都包含 4D人脸扫描，以及以英语朗读的话语。BIWI 包含数据集中所有说话人共有的 40 个不同句子，而 VOCASET 包含 255 个不同句子，这些句子在不同说话人之间部分共享。

BIWI数据集。BIWI是一个用于研究情感交流的3D视听语料库，包含以稠密动态3D人脸几何形式呈现的情感语音和面部表情数据。语料库共有14名受试者（8名女性和6名男性）说出的40个句子。每个句子均录制两次：一次带有情感，另一次不带情感。每个句子的平均时长为4.67秒。3D人脸动态以25 fps的帧率捕捉，每个模型包含23370个顶点，并具有配准拓扑。我们遵循 [16] 中的数据划分，并使用情绪子集。具体而言，训练集（BIWI-Train）包含192个句子，验证集（BIWI-Val）包含24个句子。测试集分为两个：BIWI-Test-A包含由6名训练期间见过的受试者说出的24个句子，BIWI-Test-B包含由8名未见过的主体说出的32个句子。由于BIWI-Test-A中的受试者在训练期间出现过，因此该测试集可用于定量和定性评估；而BIWI-Test-B则更适合定性评估。

VOCASET数据集。 VOCASET由来自12名受试者的480段配对视听序列组成。面部运动以60帧/秒的速度采集，每段时长约为4秒。与BIWI不同，每个3D人脸网格都与具有5023个顶点的FLAME [37] 拓扑配准。为进行公平比较，我们采用与VOCA [10] 和FaceFormer相同的训练集（VOCA-Train）、验证集（VOCA-Val）和测试集（VOCA-Test）划分。

实现。 我们将工作与三种最先进方法进行比较：VOCA [10], MeshTalk [52] 和 FaceFormer [16]。我们使用官方代码库在 BIWI 上训练并测试 VOCA，同时直接测试在 VOCASET 上训练的已发布模型。对于 MeshTalk，我们使用官方实现在这两个数据集上进行训练和测试。为了与 FaceFormer 进行比较，我们直接使用预训练权重进行测试——

表 1. BIWI-Test-A 上的定量评估。两项指标的值越低越好。

方法	唇部顶点误差 ($\times 10^{-4}$ 毫米)	FDD ($\times 10^{-5}$ 毫米)
VOCA	6.5563	8.1816
MeshTalk	5.9181	5.1025
FaceFormer	5.3077	4.6408
CodeTalker（我们的方法）	**4.7914**	**4.1170**

。在这四种方法中，VOCA、FaceFormer 和我们的 CodeTalker 都要求测试时以一种训练说话风格作为条件。对于未见过的主体，我们以所有训练风格为条件生成面部动画。有关实现细节的更多信息，请参阅补充材料。

4.2. 定量评估

遵循 MeshTalk [52] 和 FaceFormer [16]，，我们采用唇部顶点误差来衡量唇形同步。据我们所知，这是目前唯一公开提出的用于评估语音驱动面部动画的指标。作为补充，我们引入了一种新的定量测量方法，即上脸部运动统计，用于评估整体面部动态。

唇部顶点误差。 该指标衡量一个序列相对于真实值的唇部偏差，即计算每帧所有唇部顶点的最大 L2 误差，并对所有帧取均值。

上脸部动态偏差。 上脸部表情与语音的相关性较弱，取决于个人的说话风格和语音内容的语义。基于这一认识，我们提出将运动序列的面部动态变化与真实值进行比较。具体而言，上脸部动态偏差（FDD）的计算方式如下：

$$\mathrm{FDD}(\mathbf{M}_{1:T}, \hat{\mathbf{M}}_{1:T}) = \frac{\sum_{v \in \mathcal{S}_U} (\mathrm{dyn}(\mathbf{M}^v_{1:T}) - \mathrm{dyn}(\hat{\mathbf{M}}^v_{1:T}))}{|\mathcal{S}_U|}, \tag{7}$$

其中，$\mathbf{M}^v_{1:T} \in \mathbb{R}^{3\times T}$ 表示第 v个顶点的运动， $\mathcal{S}_U$ 是上半脸顶点的索引集合。dyn(·)表示沿时间轴计算的逐元素 L2 范数的标准差。

我们计算 BIWI-Test-A 中所有序列的唇部顶点误差和上脸部动态偏差（FDD），并取平均值进行比较。根据表1，所提出的 CodeTalker 取得了比现有最先进方法更低的误差，表明它能够生成更加准确的唇部同步运动。此外，表1 显示，我们的 CodeTalker 在 FDD 方面取得了最佳性能。这表明预测的上脸部表情与真实值的面部动态变化趋势（以语音和说话风格为条件）具有高度一致性以及真实值的。

4.3. 定性评估

我们在图4中将我们的方法与其他竞争方法进行了可视化比较。为确保公平比较，我们为 VOCA、FaceFormer 和我们的 CodeTalker 分配了相同的说话风格作为条件输入，该输入是随机采样的。为检验唇部同步性能，我们展示了合成面部动画中的三个典型帧，这些动画分别在特定音节处发音，并与图4上半部分中的结果进行了比较。可以观察到，与其他竞争方法相比，我们的 CodeTalker 生成的唇部运动能够更准确地配合语音信号进行发音，并且与参考结果更加一致。例如，在发双唇辅音 /b/（即图4右上角案例中的“bedside”）时，与 VOCA 和 MeshTalk 相比，CodeTalker 能够通过适当的闭唇产生更好的唇形同步；对于需要噘嘴、难度更高的语音片段“waterproof”和“shaving”，CodeTalker 能够生成准确的唇形，而其他方法则受到过度平滑问题的影响，无法正确实现唇形同步（放大查看效果更佳）。

与唇部运动不同，面部表情与语音信号的相关性较弱，在面对跨模态映射歧义时往往趋于静态。为直观展示面部运动动态，我们计算序列中相邻帧面部运动的时间统计量。具体而言，我们首先计算帧间运动 L2距离，然后针对每个顶点计算整个序列上的均值和标准差（std）。均值越高，表示面部运动越强；标准差越高，则表明面部动态的变化越丰富。图4的最后两行展示了两个示例，证明我们的方法在实现更强的面部运动和更广泛的动态范围方面优于其他方法。这主要归功于离散面部运动空间的优越性，它能有效提高对跨模态不确定性的鲁棒性。建议读者观看补充视频中的动画比较。

说话风格插值。 我们的方法可以从学到的风格嵌入空间中合成新的说话风格。为考察其在 VOCA-Test 上的效果，我们选取两个说话风格向量，即，$\mathbf{e}_i$ 和$\mathbf{e}_j$，它们分别对应较大和较轻微的唇部发音动作，并通过插值生成新的说话风格向量$\mathbf{s}_{\text{new}} = \mathbf{B} \cdot [\omega \mathbf{e}_i + (1-\omega)\mathbf{e}_j]$，其中线性系数为 ω。对于采样序列合成的三维动画，我们在图5中绘制了每种风格下各帧的上下唇距离，由此可以观察到两种典型风格之间嘴部幅度的平滑过渡。这不仅有助于合成新的说话风格，也便于在训练期间匹配未见过的主体的特定说话表现。

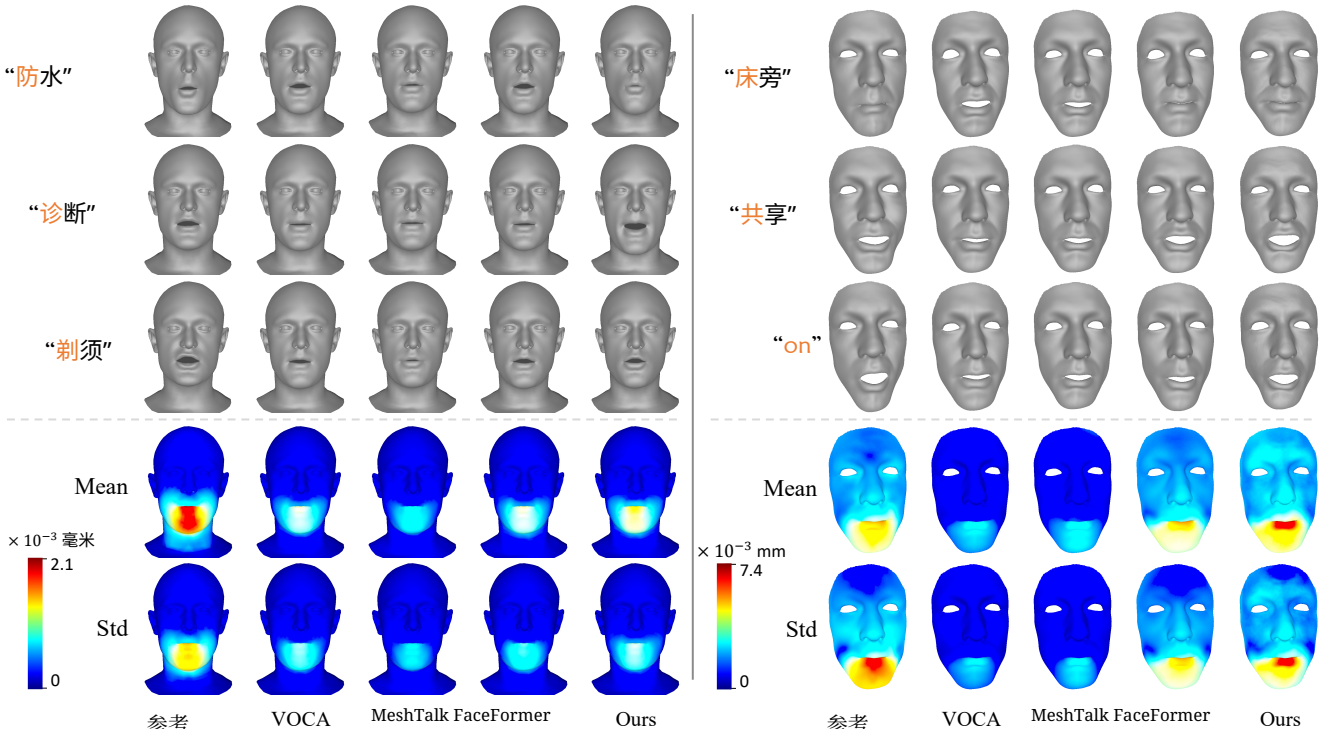


图4。在 VOCA-Test（左）和 BIWI-Test-B（右）上，不同方法生成的采样面部运动的视觉比较。上半部分展示了以不同语音片段为条件的面部动画，下半部分展示了序列中相邻帧运动变化的时间统计（均值和标准差）。

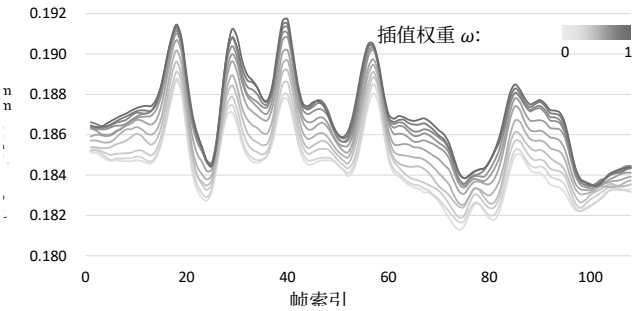


图5。以两个风格向量的不同加权线性组合为条件时，我们在序列中预测的下唇和上唇之间的距离。

4.4. 用户研究

人类感知系统经过进化，能够理解细微的面部运动并捕捉唇部同步。因此，在语音驱动面部动画任务中，它仍然是最可靠的评估方式。我们开展了一项用户研究，评估动画人脸在感知层面的唇部同步和真实感，并将其与 VOCA、MeshTalk、FaceFormer 以及真实值进行比较。我们针对每组比较采用 A/B测试，即，我们的方法与竞争方法在真实感面部动画和唇形同步方面的比较。对于 BIWI，我们分别从 BIWI-Test-B 中随机选取 30 个样本，得出四类比较结果。为使说话方面的变化达到最大程度

表2。BIWI-Test-B 和 VOCA-Test 上的用户研究结果。我们采用 A/B测试，并报告受访者更偏好 A 而非 B 的答案比例。

竞争方法	BIWI-Test-B		VOCA-Test	
	唇形同步	真实度	唇形同步	真实度
我们的方法对比 VOCA	92.47	89.25	86.02	84.95
我们的方法对比 MeshTalk	80.65	82.80	95.70	92.47
我们的方法对比 FaceFormer	53.76	56.99	70.97	69.89
我们的方法与真实值	43.01	49.46	43.01	43.01

风格，我们确保采样结果能够公平覆盖所有条件风格。因此，我们为 BIWI-Test-B 创建了 120 个 A与B 配对（30 个样本× 4 比较）。每个配对分别由至少 3 名不同参与者进行评判，最终共收集 372 条记录。对于 VOCASET 上的用户研究，我们采用与 BIWI 数据集相同的设置，即从 VOCA-Test 集中选取另外 120 个 A与B 配对，最终同样得到 372 条记录。在本研究中，31 名视力和听力良好的参与者顺利完成了评估。此外，每名参与者都参与了全部 8 类比较，以获得更充分的体验并覆盖偏好多样性。

BIWI-Test-B 上唇形同步和真实感方面的 A/B 测试结果百分比列于表 2，结果表明参与者更青睐 CodeTalker，而非其他竞争方法。根据第 4.3 节的视觉分析，我们认为这是因为

表3. 码本表示空间上的消融研究。性能通过 VOCA-Test 和 BIWI-Test-A 上的重建误差（即，L2误差）和唇部顶点误差进行衡量。

变体	VOCA-Test	BIWI-Test-A	
	重建误差 ($\times 10^{-5}$ 毫米)	重建误差 ($\times 10^{-5}$毫米)	唇部顶点误差 ($\times 10^{-4}$ 毫米)
形状纠缠码本	2.75	4.07	6.41
运动码本（我们的方法）	**0.08**	**2.83**	**4.79**

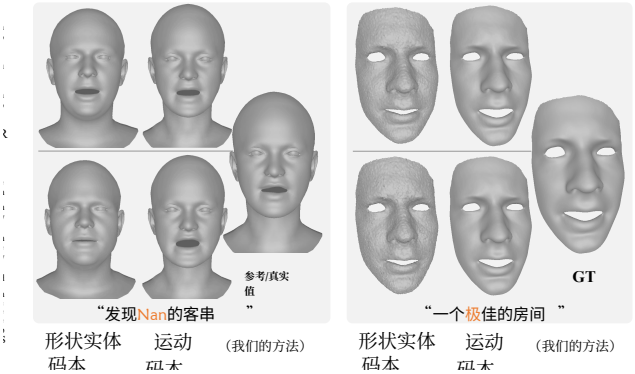


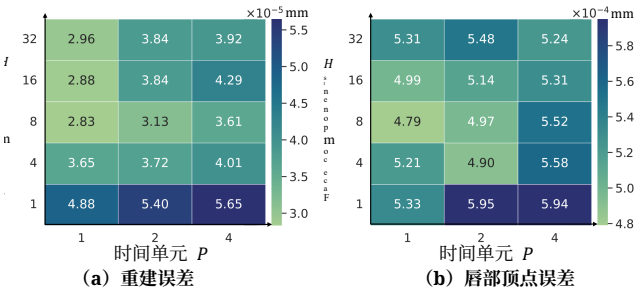
图6. VOCA-Test（左）和 BIWI-Test-A（右）上，采用不同表示空间进行重建和语音驱动运动合成的结果对比。

CodeTalker 合成的面部动画具有更丰富的面部运动、更准确的唇形以及同步效果更好的口部动作。对于上脸部运动较少的 VOCA-Test，表 2 中也能观察到类似的偏好。我们认为原因至少有三点：眼睛周围的细微运动、更准确的唇部运动，以及下脸部富有表现力的运动。尽管我们的预测与录制的表演（真实值）之间存在差距，但令人惊讶的是，我们仍获得了超过 40% 的偏好率。总体而言，用户研究证实，CodeTalker 生成的面部动画具有更优越的感知质量。

4.5. 消融研究

本节研究所提出方法的若干关键设计，包括表示空间以及码本构建的超参数。

表示空间。 为了研究我们基于运动的表示的优越性，我们构建了一个基线模型，该模型学习形状纠缠码本，即，编码表示形状（$\mathbf{m_t}+\mathbf{h}$），而非运动（$\mathbf{m_t}$）。如表3所示，该基线模型显著降低了重建准确率，图6中的可视化示例也证实了这一点。这主要是因为形状纠缠序列包含更多说话人特定信息，阻碍了编码的复用。其一个直接缺陷是自重建的泛化能力较差，进而进一步影响跨模态映射的正确性。相比之下，我们提出的说话人无关的运动表示能够更有效地表示个体间共享的通用运动先验



（a）重建误差　　（b）唇部顶点误差

图 7. 采用不同码本超参数时的模型性能比较，即时间单元长度 P和面部组件数量 H。我们在 BIWI-Test-A 上测量重建误差和唇部顶点误差。

从而提升自重建和语音驱动运动合成的质量。

码本构建。我们进一步研究了码本构建所用的超参数。我们通过测量不同设置的重建准确率和跨模态映射准确率（即唇部顶点误差），来评估其性能 $\langle P, H\rangle$。首先，我们评估了重建准确率，如图7(a) 所示。一方面，增大 P 会降低重建准确率，这可以用待表示运动的复杂性增加来解释。另一方面，增大 H 会使重建更容易，但也有过拟合风险，这解释了总体上的优势（$H < 8$），以及在 $H \geq 8$时性能较差的现象。值得注意的是，跨模态映射性能也呈现类似趋势，如图7(b) 所示。我们推测，复杂的运动基元会导致可复用性降低、冗余度升高，从而使跨模态代码查询过程产生歧义。

5. 讨论与结论

我们展示了将语音驱动的面部动画视为离散空间中的代码查询任务所带来的优势，这显著提升了动作合成质量，并缓解了跨模态歧义。与现有最先进方法相比，我们提出的方法在实现精准唇形同步和生动面部表情方面表现更优。然而，我们仍然假设面部运动与形状相互独立，这一假设是否合理值得进一步研究。此外，整体感知质量仍逊于真实值，主要原因是配对音视数据稀缺。另外，所获取的通用运动先验遵循训练集所描绘的运动分布，可能与真实世界中的面部运动有所偏差。未来值得探索利用大规模可用的说话人头部视频中的先验来引导三维面部动画。

致谢。本项目部分由香港创新及科技基金（创新及科技基金）（参考编号：ITS/307/20FP 和 ITS/313/20）资助。

参考

[1] Mohammed M Alghamdi、He Wang、Andrew J Bulpitt 和 David C Hogg。利用预训练图像生成器从语音音频生成说话人头部。载于 ACM多媒体国际会议（MM），2022年。2[2] Tenglong Ao、Qingzhe Gao、Yuke Lou、Baoquan Chen 和 Libin Liu。节律手势生成器：利用层次化神经嵌入进行节律感知的伴随语音手势合成。ACM图形学汇刊（TOG），第1–19页，2022年。3[3] Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli。wav2vec 2.0：语音表示自监督学习框架。载于神经信息处理系统大会（NeurIPS），2020年。4[4] Yoshua Bengio、Nicholas L´eonard 和 Aaron Courville。通过随机神经元估计或传播梯度以实现条件计算。arXiv预印本 arXiv:1308.3432，2013年。3[5] Chen Cao、Hongzhi Wu、Yanlin Weng、Tianjia Shao 和 Kun Zhou。基于图像的动态数字人实时面部动画。ACM图形学汇刊（TOG），35(4):1–12，2016年。2[6] Yong Cao、Wen C Tien、Petros Faloutsos 和 Fr´ed´eric Pighin。富有表现力的语音驱动面部动画。ACM图形学汇刊（TOG），24(4):1283–1302，2005年。2[7] Lele Chen、Guofeng Cui、Celong Liu、Zhong Li、Ziyi Kou、Yi Xu 和 Chenliang Xu。结合节律性头部运动的说话人头部生成。载于欧洲计算机视觉会议（ECCV），2020年。2[8] Lele Chen、Zhiheng Li、Ross K Maddox、Zhiyao Duan 和 Chenliang Xu。一瞥即知的唇部运动生成。载于欧洲计算机视觉会议（ECCV），2018年。2[9] Joon Son Chung 和 Andrew Zisserman。时不我待：真实场景中的自动唇形同步。载于亚洲计算机视觉会议（ACCV），2016年。2[10] Daniel Cudeiro、Timo Bolkart、Cassidy Laidlaw、Anurag Ranjan 和 Michael J Black。三维说话风格的捕捉、学习与合成。载于IEEE/CVF计算机视觉与模式识别会议（CVPR），2019年。1, 2, 5, 14[11] Dipanjan Das、Sandika Biswas、Sanjana Sinha 和 Brojeshwar Bhowmick。利用级联生成对抗网络学习运动和纹理的语音驱动面部动画。载于欧洲计算机视觉会议（ECCV），2020年。2[12] Pif Edwards、Chris Landreth、Eugene Fiume 和 Karan Singh。JALI：面向动画师的富有表现力的唇部同步视素模型。ACM图形学汇刊（TOG），35(4):1–11，2016年。2[13] Paul Ekman 和 Wallace V Friesen。面部动作编码系统。《环境心理学与非语言行为》，1978年。2[14] Michael Elad 和 Michal Aharon。通过学习字典和稀疏表示进行图像去噪。载于IEEE/CVF计算机视觉与模式识别会议（CVPR），2006年。2[15] Patrick Esser、Robin Rombach 和 Bjorn Ommer。驯服变换器以实现高分辨率图像合成。载于IEEE/CVF计算机视觉与模式识别会议（CVPR），2021年。2[16] Yingruo Fan、Zhaojiang Lin、Jun Saito、Wenping Wang和Taku Komura。Faceformer：基于变换器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（CVPR），2022年。1, 2, 4, 5, 6, 14[17] Yingruo Fan、Zhaojiang Lin、Jun Saito、Wenping Wang和Taku Komura。用于表现力语音驱动三维面部动画的联合音频文本模型。载于ACM交互式三维图形与游戏研讨会（i3D），2022年。2[18] Gabriele Fanelli、Juergen Gall、Harald Romsdorfer、Thibaut Weise和Luc Van Gool。情感交流的三维视听语料库。IEEE多媒体汇刊（TMM），12(6):591–598，2010年。5[19] Cletus G Fisher。视觉感知辅音之间的混淆。言语与听力研究期刊，11(4):796–804，1968年。2[20] Ohad Fried、Ayush Tewari、Michael Zollh¨ofer、Adam Finkelstein、Eli Shechtman、Dan B Goldman、Kyle Genova、Zeyu Jin、Christian Theobalt和Maneesh Agrawala。基于文本的说话人头部视频编辑。ACM图形学汇刊（TOG），38(4):1–14，2019年。2[21] Shuyang Gu、Dong Chen、Jianmin Bao、Fang Wen、Bo Zhang、Dongdong Chen、Lu Yuan和Baining Guo。用于文本到图像合成的向量量化扩散模型。载于IEEE/CVF计算机视觉与模式识别会议（CVPR），2022年。2[22] Shuhang Gu、Wangmeng Zuo、Qi Xie、Deyu Meng、Xiangchu Feng和Lei Zhang。用于图像超分辨率的卷积稀疏编码。载于IEEE/CVF国际计算机视觉会议（ICCV），2015年。2[23] Yudong Guo、Keyu Chen、Sen Liang、Yong-Jin Liu、Hujun Bao和Juyong Zhang。AD-NeRF：用于说话人头部合成的音频驱动神经辐射场。载于IEEE/CVF国际计算机视觉会议（ICCV），2021年。2[24] Ikhsanul Habibie、Weipeng Xu、Dushyant Mehta、Lingjie Liu、Hans-Peter Seidel、Gerard Pons-Moll、Mohamed El-gharib和Christian Theobalt。从视频中学习语音驱动的三维会话手势。载于ACM智能虚拟代理国际会议（IVA），2021年。2[25] Xun Huang和Serge Belongie。利用自适应实例归一化实现实时任意风格迁移。载于IEEE/CVF国际计算机视觉会议（ICCV），2017年。12[26] Zhitong Huang、Nanxuan Zhao和Jing Liao。Unicolor：基于Transformer的多模态着色统一框架。ACM图形学汇刊（TOG），页码1–16，2022年。2 – 。[27] Sergey Ioffe和Christian Szegedy。批归一化：通过减少内部协变量偏移加速深度网络训练。载于国际机器学习会议（ICML），2015年。12[28] Xinya Ji、Hang Zhou、Kaisiyuan Wang、Qianyi Wu、Wayne Wu、Feng Xu和Xun Cao。EAMM：通过基于音频的情感感知运动模型实现单样本情感说话人脸。载于ACM SIGGRAPH会议，2022年。2

[29] Xinya Ji、Hang Zhou、Kaisiyuan Wang、Wayne Wu、Chen Change Loy、Xun Cao和Feng Xu。音频驱动的情感视频肖像。载于IEEE/CVF计算机视觉与模式识别会议（CVPR），2021年。2[30] Younghyun Jo和Seon Joo Kim。利用查找表实现实用的单图像超分辨率。载于IEEE/CVF计算机视觉与模式识别会议（CVPR），2021年。2[31] Tero Karras、Timo Aila、Samuli Laine、Antti Herva和Jaakko Lehtinen。通过端到端联合学习姿态与情感实现音频驱动面部动画。ACM图形学汇刊（TOG），36(4):1–12，2017年。1, 2, 14[32] Hyeongwoo Kim、Pablo Garrido、Ayush Tewari、Weipeng Xu、Justus Thies、Matthias Niessner、Patrick P´erez、Chris-tian Richardt、Michael Zollh¨ofer和Christian Theobalt。深度视频肖像。ACM图形学汇刊（TOG），37(4):1–14，2018年。2[33] Diederik P Kingma和Jimmy Ba。Adam：一种随机优化方法。载于国际学习表征会议（ICLR），2014年。5[34] Avisek Lahiri、Vivek Kwatra、Christian Frueh、John Lewis和Chris Bregler。Lipsync3d：通过姿态和光照归一化，从视频中进行数据高效的个性化三维说话人脸学习。载于IEEE/CVF计算机视觉与模式识别会议（CVPR），2021年。2[35] John Lewis。自动唇形同步：背景与技术。《可视化与计算机动画期刊》，2(4):118–122，1991年。2[36] Hao Li、Jihun Yu、Yuting Ye和Chris Bregler。利用实时修正实现实时面部动画。ACM图形学汇刊（TOG），32(4):42–1，2013年。2[37] Tianye Li、Timo Bolkart、Michael J Black、Hao Li和Javier Romero。从四维扫描中学习面部形状与表情模型。ACM图形学汇刊（TOG），36(6):194–1，2017年。1, 5[38] Borong Liang、Yan Pan、Zhizhi Guo、Hang Zhou、Zhibin Hong、Xiaoguang Han、Junyu Han、Jingtuo Liu、Errui Ding和Jingdong Wang。通过细粒度视听控制生成富有表现力的说话头部。载于IEEE/CVF计算机视觉与模式识别会议（CVPR），2022年。2[39] Xian Liu、Yinghao Xu、Qianyi Wu、Hang Zhou、Wayne Wu和Bolei Zhou。语义感知的隐式神经音频驱动视频肖像生成。载于欧洲计算机视觉会议（ECCV），2022年。2[40] Yilong Liu、Feng Xu、Jinxiang Chai、Xin Tong、Lijuan Wang和Qiang Huo。视频音频驱动的实时面部动画。ACM图形学汇刊（TOG），34 10，2015年 6 1 2 ( ): – 。[41] Ilya Loshchilov和Frank Hutter。解耦权重衰减正则化。载于国际学习表征会议（ICLR），2019年。5[42] DW Massaro、MM Cohen、M Tabain、J Beskow和R Clark。动画语音：研究进展与应用。视听语音处理，页码309–345，2012年。2[43] Wesley Mattheyses和Werner Verhelst。视听语音合成：技术现状综述。《语音交流》，66:182–217，2015年。2

[44] Evonne Ng、Hanbyul Joo、Liwen Hu、Hao Li、Trevor Darrell、Angjoo Kanazawa 和 Shiry Ginosar。学习倾听：非确定性双人面部运动建模。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR），2022年。2、3、4、13[45] Youxin Pang、Yong Zhang、Weize Quan、Yanbo Fan、Xiaodong Cun、Ying Shan 和 Dong-ming Yan。DPE：面向通用视频肖像编辑的姿态与表情解耦。arXiv预印本 arXiv: *2301 06281*, 2023 2 。

[46] Frederic I Parke 和 Keith Waters。计算机面部动画。CRC出版社，2008 2 。Jialun Peng、Dong Liu、

[47] Songcen Xu 和 Houqiang Li。利用层次化 VQ-VAE 为图像修复生成多样化结构。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR），2021年。2[48] Hai Xuan Pham、Yuting Wang 和 Vladimir Pavlovic。端到端学习语音驱动的三维面部动画。载于 ACM多模态交互国际会议（ICMI），2018年。2[49] KR Prajwal、Rudrabha Mukhopadhyay、Vinay P Namboodiri 和 CV Jawahar。只需一个唇形同步专家，即可实现自然场景中的语音到唇形生成。载于 ACM多媒体国际会议（MM），2020 2 。Ali

[50] Razavi、Aaron Van den Oord 和 Oriol Vinyals。利用 VQ-VAE-2 生成多样化的高保真图像。载于神经信息处理系统大会（NeurIPS），2019年。2[51] Alexander Richard、Colin Lea、Shugao Ma、Jurgen Gall、Fernando De la Torre 和 Yaser Sheikh。由音频与视线驱动的编解码头像面部动画。载于 IEEE/CVF冬季应用计算机视觉会议（WACV），2021年。1[52] Alexander Richard、Michael Zollh¨ofer、Yandong Wen、Fernando De la Torre 和 Yaser Sheikh。MeshTalk：通过跨模态解耦实现语音驱动的三维面部动画。载于 IEEE/CVF国际计算机视觉会议（ICCV），2021年。1、2、5、6、14[53] Shuai Shen、Wanhua Li、Zheng Zhu、Yueqi Duan、Jie Zhou 和 Jiwen Lu。学习动态人脸辐射场，以实现少样本说话人头部合成。载于欧洲计算机视觉会议（ECCV），2022年。2[54] Sarah Taylor、Taehwan Kim、Yisong Yue、Moshe Mahler、James Krahe、Anastasio Garcia Rodriguez、Jessica Hodgins 和 Iain Matthews。一种通用语音动画的深度学习方法。ACM图形学汇刊（TOG），36(4):1–11，2017年。2[55] Sarah L Taylor、Moshe Mahler、Barry-John Theobald 和 Iain Matthews。视觉语音动态单元。载于 ACM/Eurographics计算机动画研讨会（SCA），2012年。1、2[56] Justus Thies、Mohamed Elgharib、Ayush Tewari、Christian Theobalt 和 Matthias Nießner。神经语音木偶：音频驱动的人脸重演。载于欧洲计算机视觉会议（ECCV），2020年。2[57]Radu Timofte、Vincent De Smet 和 Luc Van Gool。用于快速基于示例超分辨率的锚定邻域回归。载于 IEEE/CVF国际计算机视觉会议（ICCV），2013年。2

[58]Radu Timofte、Vincent De Smet 和 Luc Van Gool。A+：改进的锚定邻域回归，用于快速超分辨率。载于亚洲计算机视觉会议（ACCV），2014年。2[59] Dmitry Ulyanov、Andrea Vedaldi 和 Victor Lempitsky。实例归一化：快速风格化不可或缺的要素。arXiv预印本 arXiv:1607.08022，2016年。12[60] Dmitry Ulyanov、Andrea Vedaldi 和 Victor Lempitsky。改进的纹理网络：提升前馈风格化与纹理合成的质量和多样性。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR），2017年。12[61] Aaron Van Den Oord、Oriol Vinyals 等。神经离散表示学习。神经信息处理系统大会（NeurIPS），2017年。1、2、3[62] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。注意力就是你所需要的一切。神经信息处理系统大会（NeurIPS），2017年。2[63] Konstantinos Vougioukas、Stavros Petridis 和 Maja Pantic。利用生成对抗网络实现逼真的语音驱动面部动画。《国际计算机视觉期刊》（IJCV），128(5):1398–1413，2020年。2[64] Qianyun Wang、Zhenfeng Fan 和 Shihong Xia。3D-TalkEmo：学习合成 3D情感说话头部。arXiv预印本 arXiv:2104.12051，2021年。2[65] Suzhen Wang、Lincheng Li、Yu Ding 和 Xin Yu。通过单说话人视听相关性学习实现单次说话人脸生成。载于 AAAI人工智能会议（AAAI），2022年。2[66] Thibaut Weise、Sofien Bouaziz、Hao Li 和 Mark Pauly。实时基于表演的面部动画。ACM图形学汇刊（TOG），30(4):1–10，2011年。2[67] Yuyu Xu、Andrew W Feng、Stacy Marsella 和 Ari Shapiro。一种实用且可配置的游戏唇形同步方法。载于《游戏动态会议论文集》，2013年。1、2[68] Karren Yang、Dejan Markovi´c、Steven Krenn、Vasu Agrawal 和 Alexander Richard。视听语音编解码器：通过重新合成重新思考视听语音增强。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR），2022年。3[69] Ran Yi、Zipeng Ye、Juyong Zhang、Hujun Bao 和 Yong-Jin Liu。利用基于学习的个性化头部姿态实现音频驱动的说话人脸视频生成。arXiv预印本 arXiv:2002.10137，2020年。2[70] Bowen Zhang、Chenyang Qi、Pan Zhang、Bo Zhang、Hsiang-Tao Wu、Dong Chen、Qifeng Chen、Yong Wang 和 Fang Wen。MetaPortrait：通过快速个性化适配实现保留身份的说话人头部生成。arXiv预印本 arXiv:2212.08062，2022年。2[71] Hang Zhou、Yasheng Sun、Wayne Wu、Chen Change Loy、Xiaogang Wang 和 Ziwei Liu。通过隐式模块化视听表示实现姿态可控说话人脸生成。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR），2021年。2[72] Shangchen Zhou、Kelvin CK Chan、Chongyi Li 和 Chen Change Loy。迈向稳健的盲人脸修复以及码本查找变换器。 神经信息处理系统大会（*NeurIPS*），2022年。 2，3，13[73] Yang Zhou、Zhan Xu、Chris Landreth、Evangelos Kalogerakis、Subhransu Maji 和 Karan Singh。VisemeNet：以音频驱动、以动画师为中心的语音动画。 *ACM* 图形学汇刊（*TOG*），37(4):1–10，2018年。 2[74]Michael Zollh¨ofer、Justus Thies、Pablo Garrido、Derek Bradley、Thabo Beeler、Patrick P´erez、Marc Stamminger、Matthias Nießner 和 Christian Theobalt。单目三维人脸重建、跟踪及应用的最新进展。 计算机图形学论坛（*CGF*），37(2):523–550，2018年。 2

附录

本文档包含四个部分：第 A 节介绍我们 CodeTalker 的实现细节；第 B节对所提出的方法进行了进一步讨论；第 C节介绍用户研究的详细内容；第 D 节简要介绍补充视频。论文发表后，我们还将发布源代码和训练模型。

A. 实现细节

A.1. 码本的超参数

我们在主论文的 BIWI 数据集“码本构建”第4.5 节中探讨并讨论了运动码本的重要超参数。在此，我们提供了针对在两个数据集上训练的 CodeTalker 所采用的更具体参数。对于 BIWI，我们在测试集 BIWI-Test-A 上有真实值可用于定量评估，据此确定一组参数 $P = 1$和 $H = 8$，以获得高质量结果（即主论文中“码本构建”第4.5 节）。此外，我们设置了码本项数量 $N = 256$ 以及码本项的维度 $C = 128$。尽管更多的码本项和更高的维度可以缓解重建问题，但冗余元素可能会导致语音驱动运动合成出现歧义。因此，我们没有对这些参数进行大幅调优，而是仅凭经验将其设置为能够获得良好视觉质量的值。对于 VOCASET，由于没有真实值供我们获取定量结果，因此我们凭经验选择了一组参数（即，$N = 256$，$P = 1$，$H = 16$，$C = 64$），在我们的实验中能够生成视觉上合理的面部动画。

A.2. 网络架构

为提高 CodeTalker 的可复现性，我们进一步说明了面部运动空间学习和语音驱动运动合成的详细网络架构（分别对应主论文第 3.1 和 3.2 节），如表4所示。

B. 关于 CodeTalker 的更多讨论

B.1. 自重建学习中的实例归一化

实例归一化 [59]（IN）已广泛用于风格迁移领域 [25, 60], 其定义如下：

$$\mathrm{IN}(x) = \gamma\left(\frac{x - \mu(x)}{\sigma(x)}\right) + \beta. \tag{8}$$

与批归一化 [27] 层不同，这里的 $\mu(x)$ 和 $\sigma(x)$ 是针对每个样本中的每个通道，沿时间维度独立计算的

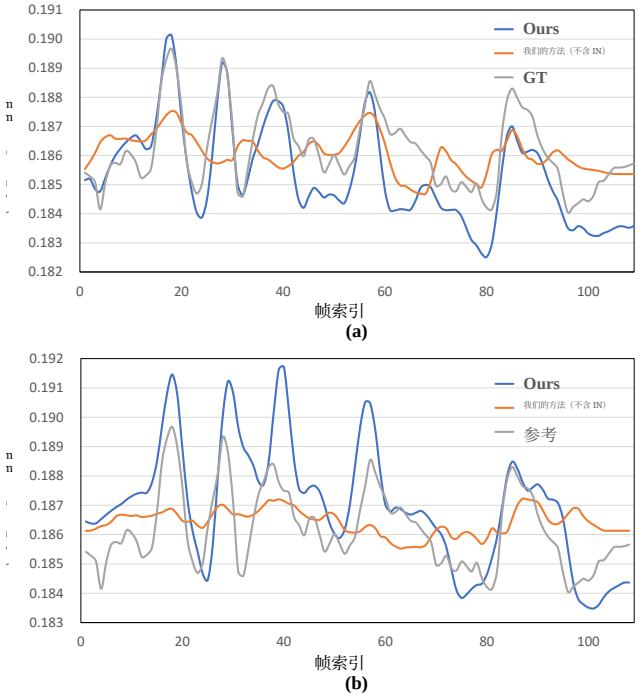


图 8. VOCA-Test 中采样序列的下唇和上唇之间的距离，其中分别展示了由不同变体生成的（a）重建结果和（b）语音驱动运动合成结果。

每个样本内的通道：

$$\mu_{nc}(x) = \frac{1}{T}\sum_{t=1}^{T} x_{nct} \tag{9}$$

$$\sigma_{nc}(x) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_{nct} - \mu_{nc}(x))^2 + \epsilon} \tag{10}$$

有趣的是，我们通过实验发现，在自重建学习中，使用 IN（而非 BN，因为小批量大小有限）对特征统计量（即，均值和方差）进行归一化，可以提升我们的 CodeTalker 的性能，如表 5所示。此外，这还能使自重建训练更加稳定。为了更直观地展示归一化带来的收益，我们还在图 8(a) 中可视化了 VOCA-Test 中一个采样序列重建结果的唇部距离。可视化结果表明，采用 IN 后，预测的唇部振幅更接近真实值；而消融变体（即，我们的方法（不含 IN））无法以准确的振幅重建唇部运动。如图 8(b) 所示，语音驱动的面部运动合成（第二阶段）也能受益于自重建学习中通过 IN 学到的面部运动码本。请注意，我们是在随机采样的说话风格条件下合成面部运动。我们推测，不同幅度的面部运动可以很好地封装进

表4. 网络架构参数说明。C(k,s,p,n) 表示一维卷积层，其卷积核大小为 k、步幅大小为 s、填充大小为 p、输出通道数为 n。T_{enc}(d1,d2,h,l) 表示 Transformer 编码器层，其基础通道数为 d1、前馈通道数为 d2、自注意力头数为 h、层数为 l；类似地，T_{dec} 表示 Transformer 解码器层。L(n) 表示输出通道数为 n 的线性层。CA[·] 表示 Transformer 解码器的额外交叉注意力输入。BIWI 的 $l_{CM} = 12$，VOCASET 的 $l_{CM} = 6$。$n \cdot T$ 表示为与视频帧对齐而插值后的音频特征长度，其中 BIWI 为 $n = 2$，VOCASET 为 $n = 1$。“+”表示按通道相加。“Drop”表示丢弃操作。

阶段	模块	输入 → 输出	层操作
I	编码器	$\mathbf{M}(T,V,3) \rightarrow \mathbf{M}(T,V\cdot 3)$ $\mathbf{M}(T,V\cdot 3) \rightarrow \mathbf{Z}_e^1(T,1024)$ $\mathbf{Z}_e^1(T,\ 1024) \rightarrow \mathbf{Z}_e^2(T,\ H\cdot C)$ $\mathbf{Z}_e^2(T,\ H\cdot C) \rightarrow \mathbf{Z_q}(T,\ H,\ C)$	重塑 L(1024) → 带泄漏的修正线性单元 → C(5,1,2,1024) → 带泄漏的修正线性单元 → 实例归一化 L(1024) → T_{enc}(1024,1536,8,6) → L($H\cdot C$) 重塑 →量化
	解码器	$\mathbf{Z_q}(T,\ H,\ C) \rightarrow \mathbf{Z_q}(T,\ H\cdot C)$ $\mathbf{Z_q}(T,\ H\cdot C) \rightarrow \mathbf{Z}_d^1(T,\ 1024)$ $\mathbf{Z}_d^1(T,\ 1024) \rightarrow \hat{\mathbf{M}}(T,\ V\cdot 3)$	重塑 L(1024) → C(5,1,2,1024) → 带泄漏的修正线性单元 → 实例归一化 L(1024) → T_{enc}(1024,1536,8,6) → L($V\cdot 3$)
II	语音编码器	$\mathbf{A}(T,\ d) \rightarrow \mathbf{F}_e^1(T',\ 512)$ 特征$_e^1(T',\ 512) \rightarrow$ 特征$_e^2(n\cdot T,\ 768)$ $\mathbf{F}_e^2(n\cdot T,768) \rightarrow \mathbf{F}_e^3(n\cdot T,1024)$	C(10,5,0,512) → 组归一化 → GeLU → C(3,2,0,512) → 组归一化 → GeLU → C(3,2,0,512) → 组归一化 → GeLU → C(3,2,0,512) → 组归一化 → GeLU → C(3,2,0,512) → 组归一化 → GeLU → C(3,2,0,512) → 组归一化 → GeLU → C(2,2,0,512) → 组归一化 → GeLU → C(2,2,0,512) → 组归一化 → GeLU 插值 →层归一化 →L(768) →Drop T_{enc}(768,3072,12,12) → L(1024)
	跨模态解码器	$\hat{\mathbf{M}}^{past}(T,V\cdot 3) \rightarrow \mathbf{F}^{past}{}_{emb}(T,1024)$ $\mathbf{F}^{past}{}_{嵌入}(T,\ 1024) \rightarrow \hat{\mathbf{Z}}_d^1(T,\ 1024)$ $\hat{\mathbf{Z}}_d^1(T,H\cdot C) \rightarrow \hat{\mathbf{Z}}_q(T,H,C)$ $\hat{\mathbf{Z}}_q(T,\ H,\ C) \rightarrow \hat{\mathbf{Z}}_q(T,\ H\cdot C)$ $\hat{\mathbf{Z}}_q(T,\ H\cdot C) \rightarrow \hat{\mathbf{Z}}_d^2(T,\ 1024)$ $\hat{\mathbf{Z}}_d^2(T,1024) \rightarrow \hat{\mathbf{M}}(T,V\cdot 3)$	L(1024) → +风格向量 T_{dec}(1024,2048,4,l_{CM})，使用交叉注意力[$\mathbf{F}_e^3$] → L($H\cdot C$) 重塑 →量化 重塑 L(1024) → C(5,1,2,1024) → 带泄漏的修正线性单元 → 实例归一化 L(1024) → T_{enc}(1024,1536,8,6) → L($V\cdot 3$)

表5. 自重建学习中实例归一化（IN）的消融研究。性能通过 VOCA-Test 和 BIWI-Test-A 上的重建误差进行评估。

变体	重建误差	
	VOCA-Test（$\times 10^{-5}$ 毫米）	BIWI-Test-A（$\times 10^{-5}$ 毫米）
我们的方法（不含 IN）	0.12	3.27
Ours	**0.08**	**2.83**

离散运动先验中，其方法是对每个通道内的时间元素进行归一化。IN 的合理性及其效果值得进一步研究，这也是我们未来的一个研究方向。

B.2. 替代数据流与监督

正如我们在论文正文第2.2节中所总结的，近期研究在各种任务中探索了离散先验学习的潜力，其中大多数现有的基于向量量化（VQ）的方法 [44, 72] 采用类别交叉熵（CE）损失来监督其词元预测。

表6. 唇形同步误差对比。我们在 BIWI-Test-A 上比较了不同方法。数值越低越好。λ 为加权因子。

方法	唇部顶点误差（$\times 10^{-4}$ mm）
替代方法（$\mathcal{L}_{ce}$）	9.6356
替代方法（$\lambda\mathcal{L}_{ce}+\mathcal{L}_{reg}$）	5.1138
替代方法（$\lambda\mathcal{L}_{ce}+\mathcal{L}_{reg}+\mathcal{L}_{motion}$）	5.0254
CodeTalker（我们的方法）	**4.7914**

因此，我们也探索了一些替代的数据流和监督框架，将其用作跨模态解码器，如图9所示。值得注意的是，为简洁起见，图中省略了风格向量和音频特征。

与正文中的跨模态解码器不同，该替代方案以过去动作编码为输入，然后以 n 类分类的形式自回归地预测代码序列。预测得到的代码序列随后从学习得到的码本 $\mathcal{Z}$ 中检索相应的代码项，并

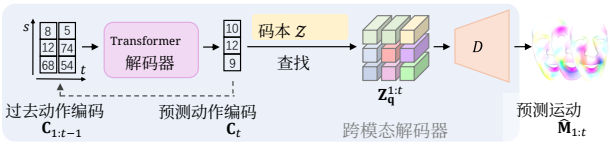


图9。我们跨模态解码器的替代数据流与监督框架。为简洁起见，省略了风格向量和音频特征输入。以过去动作编码为输入时，替代跨模态解码器首先自回归地预测运动代码，然后使用预训练码本和解码器将其解码为动作。

通过固定的解码器 D 进一步生成面部动作序列。采用交叉熵损失来惩罚预测代码序列 $\hat{\mathbf{c}} \in \{0, \ldots, |N|-1\}^{T' \cdot H}$ 与预训练编码器 E生成的真实值 $\mathbf{c}$ 之间的误差：

$$\mathcal{L}_{\text{ce}} = \sum_{i=0}^{T' \cdot H} -\mathbf{c}_i \log(\hat{\mathbf{c}}_i). \tag{11}$$

我们采用与论文正文（第3.3节）相同的设置训练替代方案。唇形同步评估结果列于表6中。仅使用 $\mathcal{L}_{\text{ce}}$ 的替代模型由于词元预测中的跨模态映射较为困难，无法很好地收敛。虽然加入更多约束（即， $\mathcal{L}_{\text{reg}}$ 和 $\mathcal{L}_{\text{motion}}$，见论文正文公式6）可以降低词元预测学习的难度，但采用这一词元预测框架的性能仍然有限。总体而言，我们的 CodeTalker 达到了更低的平均唇部误差，这表明其框架在唇部运动精度方面更具优势。

C. 用户研究

设计的用户研究界面如图 10所示。用户研究预计在 5–10 分钟内完成（24 个视频对 $\times$ 5 秒 $\times$ 3 观看次数）。为排除随机选择的影响，我们剔除了完成时间少于两分钟的比较结果。对于每位参与者，用户研究界面会显示 24 个视频对，并要求参与者根据以下两个问题分别对视频进行两次判断：“比较两张脸的嘴唇，哪一张与音频更加同步？”以及“比较两张完整的脸，哪一张看起来更加逼真？”。

D. 视频比较

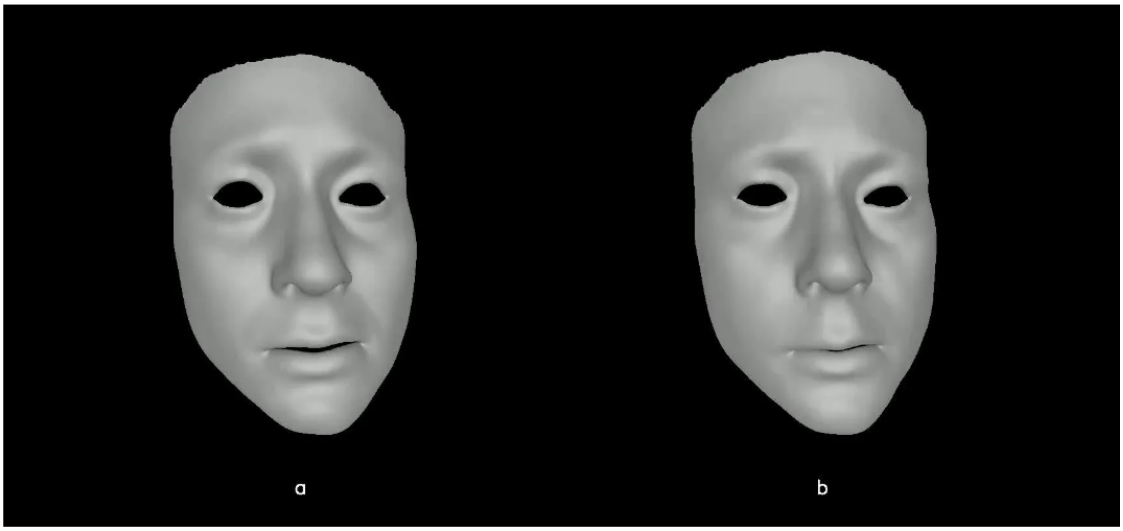
为了更好地评估其他方法 [10, 16, 31, 52]和我们的 CodeTalker 所生成的定性结果，我们提供了补充视频* 用于演示和比较。具体而言，我们使用各种音频片段测试了模型，包括从 TED 和 TEDx 视频中提取的音频片段、来自 VOCASET 和 BIWI数据集的音频序列，以及先前方法的补充视频中的语音。视频展示了 CodeTalker 能够合成自然且逼真的面部动画，并实现同步良好的唇部运动。值得注意的是，与受过度平滑问题困扰的其他方法（即 VOCA、MeshTalk 和 FaceFormer）相比，我们的 CodeTalker 能生成更加生动、逼真的面部动作，并实现更好的唇形同步。此外，我们还展示了说话风格插值结果，以及使用不同语言说话时的面部动画。

* https://doubiiu.github.io/projects/codetalker

Instructions:

Please watch the short videos (duration ~5s) of two animated talking heads. You need to choose the talking head (a or b) that moves **more naturally** in terms of the **full face and the lips (two questions for each video)**. The total duration for this survey is about 5-10mins.

Reminder: Please turn on the **sound** on your computer when watching.



1.1 Comparing the lips of two faces, which one is more in sync with the audio?

- a
- b

1.2 Comparing the two full faces, which one looks more realistic?

- a
- b

图10. 设计的用户研究界面。每位参与者需要回答24组视频对，这里由于页面限制仅展示一组视频对。