SingingBot：一种用于机器人面部歌唱表演的虚拟形象驱动系统

Zhuoxiong Xu、Xuanchen Li、Yuhao Cheng、Fei Xu、Yichao Yan、Xiaokang Yan g

arXiv:2601.02125v1 [cs.RO] 5 Jan 2026

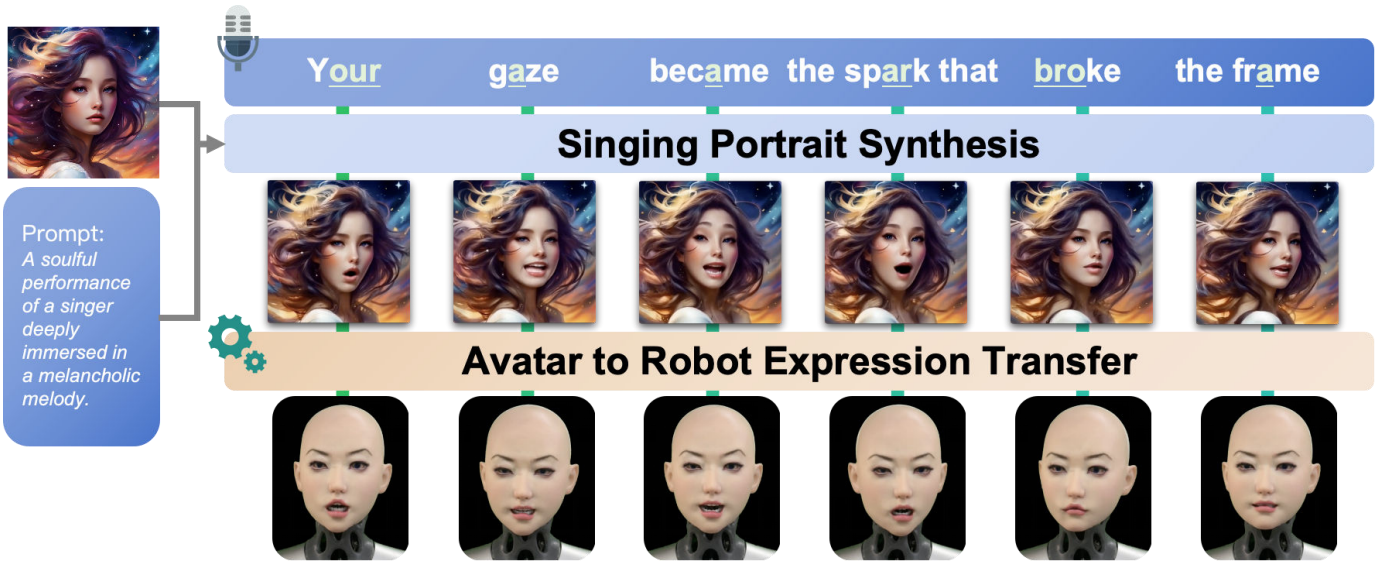

图 1：我们提出了**SingingBot**，一种基于合成肖像动画生成机器人面部歌唱表演的框架。我们的方法在保持明显唇音同步的同时，传达歌唱中丰富的情感。

摘要——赋予机器人面部歌唱能力，对于实现富有同理心的人机交互至关重要。然而，现有的机器人面部驱动研究主要聚焦于对话或静态表情模仿，难以满足歌唱对连续情感表达与连贯性的高要求。为此，我们提出了一种新颖的虚拟形象驱动框架，以实现富有感染力的机器人歌唱。我们首先利用嵌入了丰富人类先验的肖像视频生成模型，合成生动的歌唱虚拟形象，为表情和情感引导提供可靠依据。随后，通过覆盖广泛表情空间的语义导向映射函数，将这些面部特征传递给机器人。此外，为了定量评估机器人歌唱的情感丰富度，我们提出了情感动态范围指标，用于衡量效价-唤醒度空间中的情感广度；结果表明，广泛的情感谱对于富有感染力的表演至关重要。全面的实验表明，我们的方法在保持唇音同步的同时实现了丰富的情感表达，显著优于现有方法。项目主页：
https://droshown.github.io/SingingBot/

索引词—**机器人面部、面部动画、运动控制**

一、引言

歌唱是一种超越语言与文化、表达情感的普遍方式。与言语不同，歌唱要求发音连贯，并受旋律和节奏约束，因此需要延长元音，并采用特定的歌词表达。因此，赋予人形机器人歌唱能力，是评估其复现人类行为能力的一项重要基准。这也代表着迈向自然人机交互的重要一步，可应用于陪伴、接待和娱乐等场景。然而，对表达连贯性和情感广度的更高要求，使这项任务极具挑战性。

早期方法 [1], [2] 主要依赖在预设表情集或预编程硬件之间进行插值来实现机器人面部动画 [3]。这些方法受表情库规模的限制，难以捕捉歌唱中细腻的表情与情感。为解决这些问题，近期的数据驱动研究 [4]–[7] 聚焦于通过配对数据学习实现自动运动控制。其中一类方法 [8], [9] 利用运动学生成音频同步动画，但主要面向说话，无法生成富有旋律性的歌唱动画。另一类方法 [10], [11] 将图像空间中的人类表情迁移到机器人上，但通常侧重于静态表情匹配，忽视了歌唱表演所需的连贯情感表达。这两类方法都难以直接应用于机器人歌唱。

与以往工作不同，我们提出了一种框架，利用视频扩散模型中强大的人类先验，实现逼真的机器人歌唱。具体而言，我们采用以人为中心的视频扩散Transformer [12] 来合成可控的二维肖像动画，作为机器人式表现的驱动源。随后，我们通过应用

语义导向的分段映射策略来生成电机控制值。与现有的数据驱动方法相比 [4]–[7], ，基于语义的映射覆盖更大的表情空间，确保最终表现兼具情感细微差异和唇形同步准确性。

此外，在感知层面量化机器人歌唱表演仍是一项重大挑战。现有的音频驱动动画方法侧重于唇形同步，却忽视了情感丰富度。受情感计算的启发，我们利用基于情感环状模型的效价-唤醒度（VA）空间 [13]来捕捉细微的情感差异，其中任何情绪都可以由两个连续且正交的维度表示。具体而言，我们提出情感动态范围（EDR），作为量化情感广度的指标。通过比较 VA 空间中情感轨迹形成的凸包面积，我们证明了我们的方法在实现富有表现力的表演的同时，仍能保持精准的唇形同步，其表现显著优于基线方法。

我们的贡献总结如下：（1）我们提出一种框架，将由提示控制的虚拟形象动画迁移至机器人面部，以生动的唇部动作和情绪弥合数字人与实体机器人之间的差距。（2）我们揭示了评估歌唱表演情感丰富度的重要性，并提出情感动态范围（EDR），作为基于 VA 空间衡量情感广度的定量指标。（3）我们开展了大量实验，对比了各种歌唱策略，并通过定量和定性分析证明，我们的系统达到了最先进的性能。

II. 相关工作

仿生机器人面部控制。机器人表情的生成受到生物肌肉与稀疏电机之间结构差异的阻碍，同时还面临硅胶皮肤复杂的非线性形变问题 [14]–[16]。传统方法 [1], [2], [17] 依赖对预定义表情基进行插值，但缺乏表现力和泛化能力。近期的基于学习的方法 [4], [7]利用神经网络将音频或图像映射为电机参数，既可以直接映射，也可以通过训练好的模仿器实现 [10], [18]。然而，由于配对数据稀缺，这些工作通常面向简单的对话场景，往往忽视时间连贯性和连续情感表达。尽管 UGotMe [9]等工作引入了情绪分类以实现共情交互，但它们依赖固定的情绪基，无法在歌唱场景中联合生成情感表达与同步的唇部动作。我们的工作则着眼于机器人歌唱，利用大规模人类先验实现高度真实的表现和丰富的情感。

音频驱动的虚拟形象面部动画。音频驱动的虚拟形象与机器人动画面临的一项根本挑战，是如何对音频与面部动态之间的复杂映射进行建模。由于标注数据有限，三维说话头方法 [19], [20] 往往难以泛化到夸张表情或歌唱场景。相反，二维方法 [12], [21] 利用网络规模数据提取稳健的人类先验，从而生成具有丰富情感、微表情和基于提示词的控制功能的逼真视频。我们利用这些预训练二维人类先验，将其作为高保真且灵活的驱动源，弥合虚拟形象与机器人之间的鸿沟。通过将这些虚拟形象面部特征迁移至实体机器人领域，我们赋予机器人歌唱能力，满足其持续表达情感和富有表现力的发声的高要求。

III. 方法

我们的目标是生成与歌曲人声同步的抒情机器人面部动画。具体而言，给定$\mathbf{T}$帧人声音频$\mathbf{A}_{0:T-1}$，我们的框架会生成机器人面部运动$\mathbf{M}_{0:T-1} = \{\mathbf{m}_i \in \mathbb{R}^d\}_{i=0}^{T-1}$，其中每个运动向量具有 d 个自由度（DoFs），并由取值范围为0至1的控制值表示。

为此，我们提出了**SingingBot**，如图2所示。首先，我们采用一个基于大量人类先验的预训练视频扩散模型，根据人声音频合成逼真的二维肖像视频（第III-A节）。这些生成的视频将作为机器人面部的中间驱动源。随后，我们的方法利用面部特征提取器从合成视频中提取表情向量。接着，我们通过精心设计的分段映射函数，将这些表情向量转换为电机驱动参数，最终生成机器人面部动画（第III-B节）。

A. 肖像动画合成

直接回归电机参数面临着巨大挑战，因为大规模配对的电机与表情数据十分稀缺。因此，许多先前的研究借助面部标志点或可驱动机器人仿体等中间表示来驱动机器人面部。然而，这些方法难以应用于歌唱场景。由于歌唱表演需要连续且饱含情感的表情，基于标志点的方法因其稀疏且缺乏外观信息，难以传达情感。同样，受限于训练数据有限，虚拟机器人仿体无法泛化到逼真歌唱所需的多样情绪和复杂表情。受近期在大规模以人为中心的数据上预训练的视频扩散变换器进展的启发，我们认为，其中蕴含的丰富人类先验能够为机器人面部歌唱动画提供充分的面部表情和情感引导。具体而言，我们采用预训练的视频扩散模型 $\mathcal{D}$ 合成与输入人声对齐的二维演唱视频$\mathbf{V}_{0:T-1} = \{\mathbf{v}_i \in \mathbb{R}^{h\times w\times 3}\}_{i=0}^{T-1}$。为控制歌唱表演的风格，我们使用参考肖像图像$\mathbf{I}^{ref}$ 和文本提示$\mathbf{p}^{ref}$ 作为条件：

$$\mathbf{V}_{0:T-1} = \mathcal{D}(\mathbf{A}_{0:T-1}, \mathbf{I}_{ref}, \mathbf{p}_{ref}). \tag{1}$$

丰富的人类先验确保合成的歌唱视频具有丰富的情感保真度和连贯的表情。

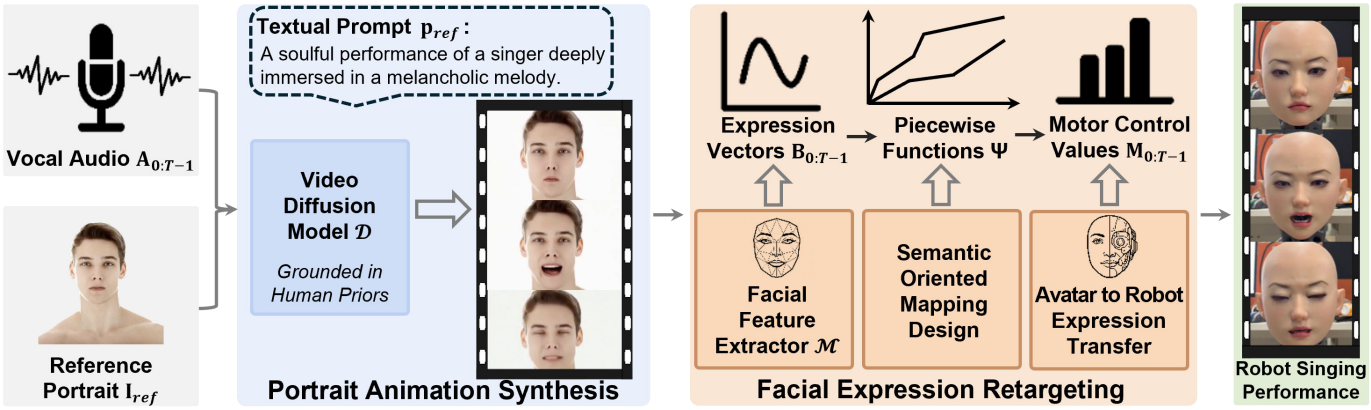

图 2：**SingingBot 用于机器人歌唱表演的整体流程**。给定人声音频和参考肖像，我们的方法首先利用预训练的视频扩散模型合成生动的头像歌唱动画。得益于模型中蕴含的丰富表情与情绪先验，头像动画可作为可靠的驱动源，为后续的机器人表演提供富有表现力的面部特征。随后，通过面向语义的分段函数，将头像的面部特征映射到实体机器人的运动空间，从而实现连贯且吸引人的机器人歌唱表演。

B. 面部表情重定向

我们以合成的肖像歌唱视频作为驱动源，进而将富有表现力的面部动态重定向到机器人的物理动作空间中。

头像面部动态提取。我们采用符合 ARKit 标准的 52 维混合形状系数向量 [22] 作为面部表情的参数化表示。这些系数用于量化特定肌肉动作的强度，例如 下颌张开 和 嘴部拉伸。不同于以往依赖原始二维面部关键点的方法，混合形状提供了一种具有语义意义的特征空间，该空间与身份无关，且与现有动画流程高度兼容。具体而言，我们使用 MediaPipe $\mathcal{M}$ [23]逐帧提取混合形状系数 $\mathbf{B}_{0:T-1} = \{\mathbf{b}_i \in \mathbb{R}^{52}\}_{i=0}^{T-1}$，提取自合成视频：

$$\mathbf{B}_{0:T-1} = \mathcal{M}(\mathbf{V}_{0:T-1}). \tag{2}$$

我们对提取的表情向量进行高斯平滑后处理，以减轻可能导致机器人动作生硬的时间抖动。这样既能确保运动平滑，同时保留显著的表情细节。

面向语义的虚拟形象到机器人表情迁移。驱动机器人面部的核心挑战在于将视觉表情特征映射到机器人的运动空间。为避免以往数据驱动方法固有的泛化局限，我们不再从配对数据集中学习黑箱映射。相反，我们提出一种**面向语义的分段函数设计**，根据语义对应关系将混合形状参数映射为电机控制值。形式上，给定第 j 个符合 ARKit 标准的语义表情的强度标量 $\beta_j \in \mathbf{b}$，我们将其对应函数定义为 $\Psi_j(\cdot)$：

$$\Delta\mathbf{m}_j = \Psi_j(\beta_j) = \mathbf{w}_{j,k} \cdot \beta_j + \mathbf{c}_{j,k}, \quad \text{for } \beta_j \in [\tau_{j,k}, \tau_{j,k+1}), \tag{3}$$

其中，$\Delta\mathbf{m}_j \in \mathbb{R}^d$ 表示对机器人 d 执行器产生作用的运动控制值。函数 Ψ_j 由一组 K 区间定义，其中$\mathbf{w}_{j,k}$ 和 $\mathbf{c}_{j,k}$ 分别表示由阈值 $[\tau_{j,k}, \tau_{j,k+1})$ 界定的第 k 个线性段的斜率向量和截距向量。我们为每种语义表情手动设计这些分段函数；高度非线性的表情基通常需要更密集的分段，以确保表现力。

由于机器人的自由度远少于混合形状，因此某些语义表情（例如 鼓腮）没有直接的物理对应项，因此被排除在 $\Psi_j(\beta_j) = 0$ 之外。此外，为了处理机械约束（例如，某些非对称表情，如 左/右鼻翼皱缩，共用一个执行器），我们将它们合并为单一的对称映射，并取其平均值作为输入密度。设计详情请参阅补充材料。

最后，通过融合所有有效映射得到的控制值，我们得到最终控制值$\mathbf{m}$：

$$\mathbf{m} = \mathbf{m}_{rest} + \sum\nolimits_{\beta_j \in \mathbf{b}} \Psi_j(\beta_j), \tag{4}$$

其中，$\mathbf{m}_{rest}$ 表示初始静止姿态。此外，我们将通过 MediaPipe [23] 提取的三自由度姿态线性映射到三个颈部电机，以控制头部运动。

IV. 实验

A. 实现细节

机器人面部平台。我们在**Hobbs**机器人平台上测试了该方法。该平台具有32个自由度，其中29个电机用于面部运动，3个用于头部/颈部运动。电机通过铰接连杆与硅胶皮肤相连。肖像动画由远程服务器上的Hallo3 [12] 生成，该服务器配备一块NVIDIA A800 GPU；表情迁移模块则在机器人内置的RK3588处理器上本地运行。

数据准备。我们收集了一个测试集，其中包含40段歌唱人声片段，每段时长为3至4秒。该数据集涵盖多种语言和情感风格，以确保音素覆盖和情感表达的多样性。

基线方法。我们将所提出的方法与三种基于数据的基线方法进行比较：**(1)** 随机采样（RT），即从训练集中随机采样控制值；**(2)**最近邻检索（NNR），即根据混合形状在训练集中寻找最接近的样本；以及 **(3)** Zhu 等人提出的一种直接回归方法。 [4],该方法将混合形状系数直接映射为电机控制值。我们使用 EmoTalk 生成混合形状 [19]。所有基线方法均已重新实现，并在我们收集的数据集上进行训练，该数据集由 1 万组随机采样的配对数据组成。

B. 定量比较

我们采用广泛使用的唇形同步误差距离（LSE-D）和唇形同步误差置信度（LSE-C），衡量演唱音频与机器人式表现之间的同步程度 [24]。LSE-D 值越低，表示嘴唇与音频的一致性越高；LSE-C 值越高，则表示音频与表演的对齐效果越好。此外，我们提出情感动态范围（EDR）指标，基于效价-唤醒度（VA）空间量化情感广度。具体而言，我们使用一个预训练的情感识别模型 $\Phi(\cdot)$ [25], [26] ，从机器人式表现中提取逐帧情感坐标，得到一组离散的 VA 空间点 $\mathcal{P} = \{p_t \mid p_t = (v_t,\ a_t)\}_{t=0}^{T-1}$，其中 $v_t,\ a_t \in [-1, 1]$ 分别表示效价和唤醒度。形式上，令 $\mathcal{H}$ 为移除最高 5% 的离群点后，包含 $\mathcal{P}$ 中所有点的最小凸多边形。EDR 定义为 $\mathcal{H}$的几何面积。EDR 值越大，表示情感状态的分布越广，反映出表演更富有表现力。

定量结果见表I。在唇形同步和情感表现力方面，我们的方法都显著优于基线方法。值得注意的是，我们的方法取得的 EDR 比竞争方法高一个数量级。尽管 RT 涵盖了广泛的表情空间，但它往往会生成被误分类为“惊讶”的不自然表情，而 NNR 和 Zhu 等人 [4]则无法生成丰富的情感或富有表现力的唇部动作。我们的方法得益于人类先验和语义导向映射，能够在保持高质量唇形同步的同时展现丰富的情感。此外，通过使用照片级真实感虚拟形象提供可靠的表情与情感信号，与受数据限制的方法相比，我们的方法显著提升了泛化能力和情感多样性。

情感动态范围的可视化。情感动态范围（EDR）指标反映了歌唱表演中情绪在效价-唤醒度（VA）空间内的分布范围。为了直观展示 EDR 的几何解释，我们采样了两个片段并绘制了结果。如图3所示，不同方法的情绪范围以颜色各异的多边形表示，

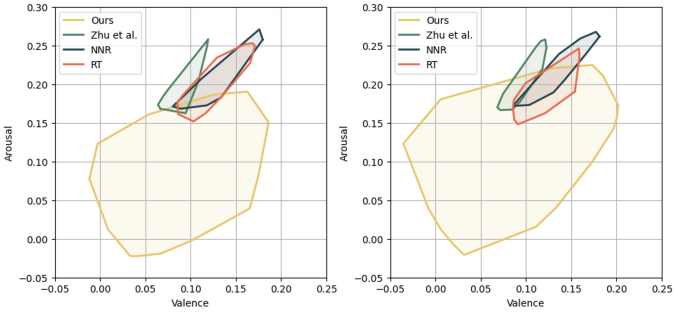

图 3：**情感动态范围可视化**。多边形区域代表歌唱表演的情感丰富度。我们的方法在情感表现力方面显著更优。

表 I：基于面部表情和情感丰富度与基线方法的定量比较。

方法	LSE-D ↓	LSE-C ↑	EDR ↑
RT	14.892	0.196	0.0044
NNR	12.834	1.481	0.0030
Zhu 等人 [4]	12.428	1.504	0.0021
Ours	**11.095**	**2.313**	**0.0389**

其中，多边形区域代表 EDR 指标。由于 NNR 和 Zhu 等人 [4] 生成的唇部动作模糊且缺乏明确的情绪状态，因此仅占据很小的区域。尽管 RT 覆盖了广阔的机器人表情空间，但其生成的表情通常缺乏有效的情感语义，导致 EDR 有限。相比之下，我们的方法展现出显著更丰富的情感表现力，这对于实现引人入胜的歌唱表演至关重要。

C. 用户研究

我们开展了一项用户研究，进一步评估观众对不同方法生成的机器人歌唱表演的偏好。具体而言，研究包含 16 个问题，每个问题都会向参与者展示一段时长为 7 秒的歌唱表演片段。参与者需要根据表演的表情真实感、情感共鸣和唇音同步情况进行评分，评分范围为 1 至 5 分。最终，我们收到了 40 份有效提交。如表II所示，我们的方法生成的歌唱表演最为真实，在情感表现力方面也表现出色，获得了最佳感知质量。

D. 定性比较

图4 展示了不同方法生成的机器人式表现的视觉对比。RT 从数据集中进行随机采样，产生强烈却不自然的表情，从而造成恐怖谷效应。尽管 NNR 和 Zhu 等人 [4] 生成的唇部动作与音频有一定的一致性，但结果过于保守，缺乏表现力。相比之下，我们的方法生成了更逼真的唇形，能够有效捕捉

表 II：用户研究。我们从真实感、情感共鸣和唇形同步三个方面评估受众偏好。

方法	真实感 ↑	共鸣 ↑	唇形同步 ↑
RT	1.21	1.32	1.33
NNR	1.50	1.57	1.87
Zhu 等人 [4]	1.94	1.73	2.21
Ours	**3.02**	**3.05**	**3.57**

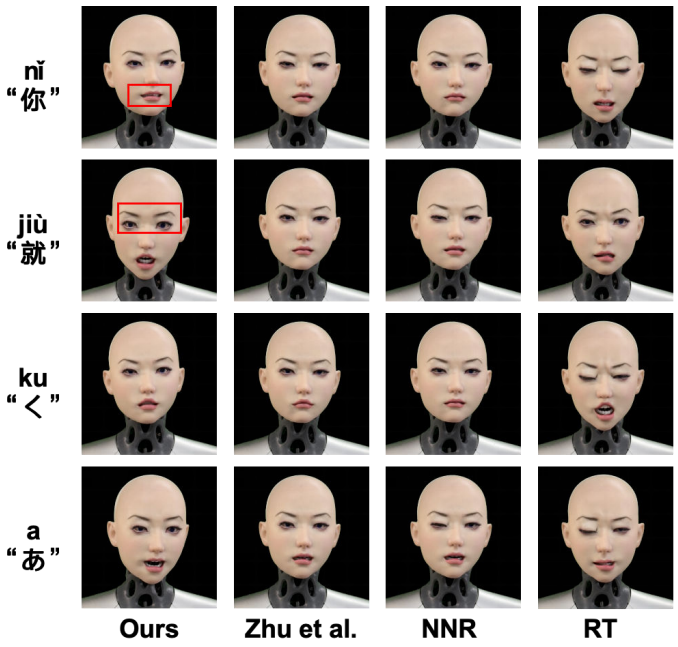

图4：**生成的机器人歌唱表演的视觉对比**。我们的方法在呈现合理唇形的同时，展现出丰富的情感。红框标出了表达细腻情感的微表情。

表 III：消融研究，分析不同驱动策略对我们方法的影响。

方法	LSE-D ↓	LSE-C ↑	EDR ↑
无虚拟形象驱动	13.069	1.023	0.0038
无扩散先验	14.387	1.048	0.0035
Ours	**11.095**	**2.313**	**0.0389**

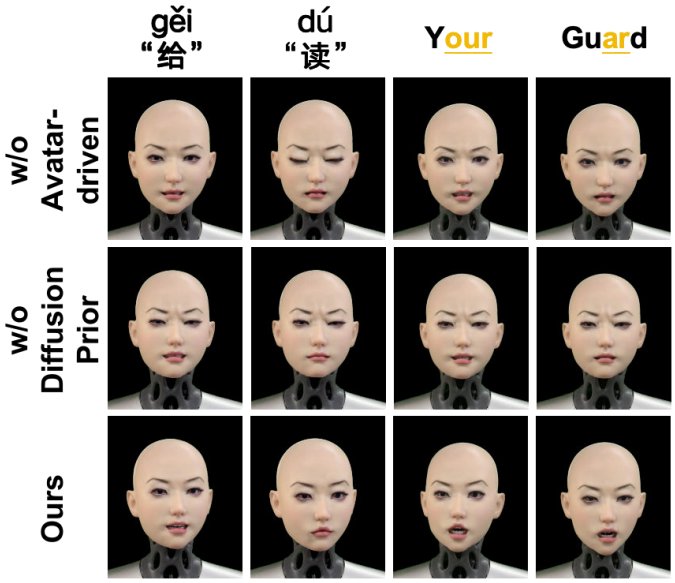

图 5：**不同歌唱策略的视觉对比**。使用嵌入扩散先验的虚拟形象视频作为驱动源，可显著提升机器人歌唱的动画表现。

歌唱表演中的大幅度动作和延长元音。此外，在保持唇形同步的同时，我们的方法还会呈现传递丰富情感的细微表情，例如嘴角微微上扬（第 1 行）和眼睛睁大（第 2 行）。

E. 消融研究

基于虚拟形象的驱动分析。为评估基于虚拟形象的驱动源的贡献，我们移除了中间虚拟形象视频生成阶段，改用 EmoTalk 预测的混合形状系数 [19]作为我们分段函数的输入（记为**无基于虚拟形象驱动**）。如表5所示，所有指标均显著下降。图5 中的结果表明，机器人只能生成模糊的嘴型和轻微的情感变化。相比之下，我们的方法能够传递肖像动画中蕴含的视觉表情和情感特征，实现富有表现力的表演。

扩散人体先验分析。视频扩散模型利用广泛的人类先验，实现了显著优于在有限数据集上训练的方法的肖像动画质量。为研究扩散先验对机器人动画质量的影响，我们使用由 EmoTalk 生成的渲染虚拟形象动画 [19] 作为驱动动画（记为 **无扩散先验**）。如表 III所示，在此设置下所有指标均有所下降。这种性能退化表明，尽管渲染的 3D 虚拟形象动画能够提供基本的视觉运动信号，但缺乏扩散先验中蕴含的细微微表情和外观特征。这些特征对于识别情绪状态和表情强度至关重要。因此，如图 5所示，最终的机器人式表现受到唇部动作受限和情绪错位的影响。

参考肖像分析。我们的方法能够受益于扩散生成模型的可控性，通过改变条件生成不同风格的歌唱表演，并进一步实现多样化的机器人面部动画。图6 展示了使用相同语音输入、以4种不同参考肖像合成驱动动画所生成的机器人面部表演。通过切换参考肖像，我们的方法能够实现多样化的歌唱表演风格。

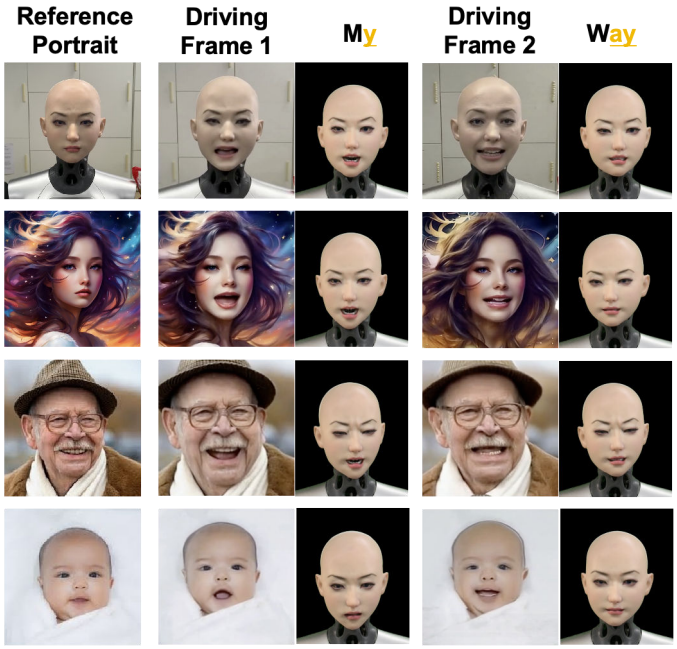

图 6：**歌唱风格控制**。我们的方法通过不同的参考肖像实现了由其驱动的风格化表演。

V. 结论

本文提出了SingingBot，一种用于机器人歌唱表演的新型框架。为了解决现有方法在歌唱场景中的局限性——这些场景对更出色的唇部发音和情感表现力提出了更高要求——我们的方法利用在大规模数据上预训练的肖像动画模型，生成逼真的虚拟形象歌唱视频作为驱动源，并借助其中蕴含的人类先验提供丰富的表情和情感引导。通过一组语义导向映射函数，我们的方法将虚拟面部特征传递到机器人实体运动空间，从而实现富有吸引力的机器人歌唱表演。此外，我们引入情感动态范围指标，以量化情感多样性。通过分析效价-唤醒度空间中的情感广度，我们揭示出丰富的情感谱对于表演的情感表现力至关重要。大量实验表明，与现有方法相比，我们的方法在唇音同步和情感丰富度方面均取得了更优的性能，为富有同理心的人机交互迈出了有意义的一步。

参考文献

[1] Chyi-Yeu Lin、Chun-Chia Huang 和 Li-Chieh Cheng，“拟人面部机器人设计中的一种表现力简化机构”，*Robotica*，2016年。[2] Wagshum Techane Asheber、Chyi-Yeu Lin 和 Shih Hsiang Yen，“具有可扩展面部表情的人形头部面部机构”，*International JournalofAdvanced Robotic Systems*，2016年。[3] Fuji Ren 和 Zhong Huang，“基于人形机器人 xin-ren 的自动面部表情学习方法”，*IEEE Transactions on Human-MachineSystems*，2016年。[4] Yongtong Zhu、Lei Li、Iggy Qian、WenBin Zhou、Ye Yuan、Qingdu Li、Na Liu 和 Jianwei Zhang，“唤醒人机交互中的面部情感表达”，发表于*IROS*，2025年，第 21511–21518 页。

[5] Peizhen Li、Longbing Cao、Xiao-Ming Wu、Runze Yang 和 Xiaohan Yu，《X2c：一个收录细腻面部表情、用于逼真人形机器人模仿的数据集》，arXiv 预印本 arXiv:2505.11146，2025年。[6] Yuhang Hu、Boyuan Chen、Jiong Lin、Yunzhe Wang、Yingke Wang、Cameron Mehlman 和 Hod Lipson，《人机面部共同表情》，Science Robotics，第 eadi4724 页，2024年。[7] Dong Zhang、Jingwei Peng、Yuyang Jiao、Jiayuan Gu、Jingyi Yu 和 Jiahao Chen，《Exface：利用扩散变换器和自举训练实现人形机器人的表情控制》，arXiv 预印本 arXiv:2504.14477，2025年。[8] Boren Li、Hang Li 和 Hangxin Liu，《通过语音驱动仿生机器人面部表情》，载于 IROS，2024年。[9] Peizhen Li、Longbing Cao、Xiao-Ming Wu、Xiaohan Yu 和 Runze Yang，《UGotMe：用于情感化人机交互的具身系统》，载于 ICRA。IEEE，2025年，第 5542–5548 页。[10] Boyuan Chen、Yuhang Hu、Lianfeng Li、Sara Cummings 和 Hod Lipson，《Smile like you mean it：利用学习模型驱动仿生机器人面部》，载于 ICRA，2021年。[11] Xiaofeng Liu、Rongrong Ni、Biao Yang、Siyang Song 和 Angelo Cangelosi，《解锁人形机器人的类人面部表情：一种基于动作单元的面部表情解耦合成新方法》，IEEE Transactions on Robotics，第 3850–3865 页，2024年。[12] Jiahao Cui、Hui Li、Yun Zhan、Hanlin Shang、Kaihui Cheng、Yuqi Ma、Shan Mu、Hang Zhou、Jingdong Wang 和 Siyu Zhu，《Hallo3：利用视频扩散变换器实现高度动态、逼真的肖像图像动画》，载于 CVPR，2025年，第 21086–21095 页。[13] James A Russell，《情感环状模型》，Journal of Personality and Social Psychology，1980年。[14] Qincheng Sheng、Wei Tang、Hao Qin、Yujie Kong、Haokai Dai、Yiding Zhong、Yonghao Wang、Jun Zou 和 Huayong Yang，《人工智能驱动的面部机器人在人机表情交互中的应用综述》，Science China Technological Sciences，第 2010301 页，2025年。[15] Gaoyang Pang、Geng Yang 和 Zhibo Pang，《机器人皮肤综述：赋能未来协作机器人安全协作、沉浸式远程操作和情感交互的潜在技术》，IEEE Transactions on Medical Robotics and Bionics，第 681–700 页，2021年。[16] Zanwar Faraj、Mert Selamet、Carlos Morales、Patricio Torres、Maimuna Hossain、Boyuan Chen 和 Hod Lipson，《具有面部表情的人形机器人脸部》，2021年。[17]Toshio Fukuda、Jun Taguri、Fumihito Arai、Masakazu Nakashima、Daisuke Tachibana 和 Yasuhisa Hasegawa，《用于人机双向交流的机器人面部表情》，载于 ICRA，2002年。[18] Zongzheng Zhang、Jiawen Yang、Ziqiao Peng、Meng Yang、Jianzhu Ma、Lin Cheng、Huazhe Xu、Hang Zhao 和 Hao Zhao，《Morpheus：一种采用混合驱动并支持多样化情感控制的神经网络驱动仿生面部》，arXiv 预印本 arXiv:2507.16645，2025年。[19] Ziqiao Peng、Haoyu Wu、Zhenbo Song、Hao Xu、Xiangyu Zhu、Jun He、Hongyan Liu 和 Zhaoxin Fan，《EmoTalk：用于三维人脸动画的语音驱动情感解耦》，载于 ICCV，2023年。[20] Chaeyeon Chung、Ilya Fedorov、Michael Huang、Aleksey Karmanov、Dmitry Korobchenko、Roger Ribera、Yeongho Seol 等，《Audio2face-3d：面向数字虚拟形象的音频驱动逼真面部动画》，arXiv 预印本 arXiv:2508.16401，2025年。[21] Wenxuan Zhang、Xiaodong Cun、Xuan Wang、Yong Zhang、Xi Shen、Yu Guo、Ying Shan 和 Fei Wang，《SadTalker：学习逼真的三维运动系数，以实现风格化音频驱动单图说话人脸动画》，载于 CVPR，2023年，第 8652–8661 页。[22] 苹果公司，《ARKit》，https://developer.apple.com/augmented-reality/ arkit/，2025年。[23] Camillo Lugaresi、Jiuqiang Tang、Hadon Nash、Chris McClanahan、Esha Uboweja、Michael Hays、Fan Zhang、Chuo-Ling Chang、Ming Guang Yong、Juhyun Lee 等，《MediaPipe：用于构建感知流水线的框架》，arXiv 预印本 arXiv:1906.08172，2019年。[24]KR Prajwal、Rudrabha Mukhopadhyay、Vinay P Namboodiri 和 CV Jawahar，《野外语音到唇形生成，只需一个唇形同步专家》，载于 ACM MM，2020年，第 484–492 页。[25] Andrey V Savchenko、Lyudmila V Savchenko 和 Ilya Makarov，《基于单个面部表情识别神经网络对在线学习中的情绪和参与度进行分类》，IEEE Transactions on Affective Computing，2022年。[26] Andrey Savchenko，《基于多重检验校正和自适应帧率的面部表情识别》，载于 ICML，2023年。

SingingBot：一种用于机器人面部歌唱表演的虚拟形象驱动系统

补充材料

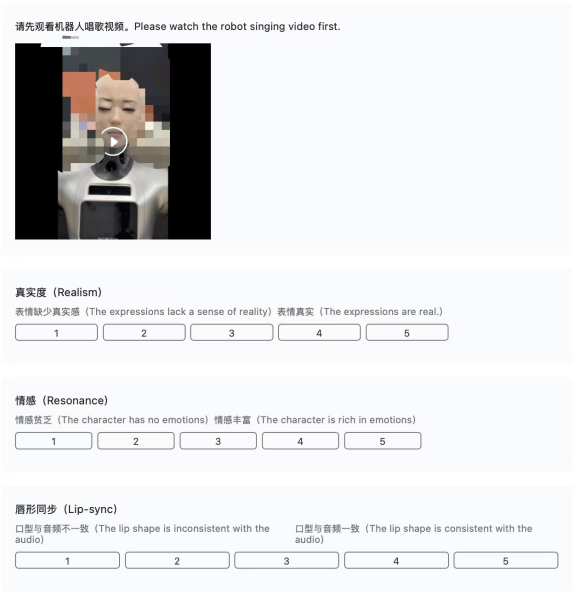

图 A：用户研究界面设计示例。

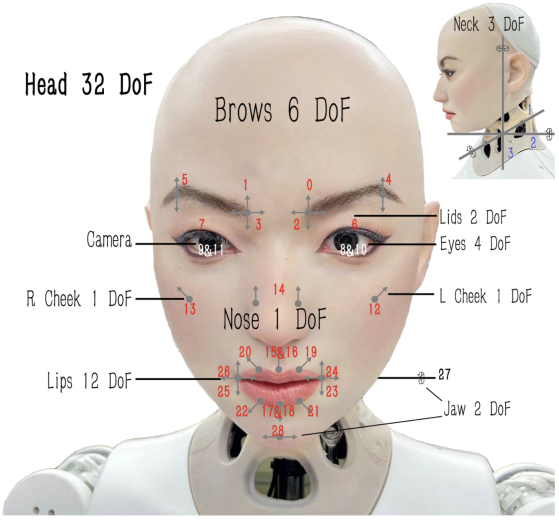

图 B：机器人面部控制点的自由度。

在补充材料中，我们提供了有关本研究的更多信息，包括用户研究的具体设计、人形机器人的硬件设计，以及映射函数的更多细节。我们还简要介绍了补充视频和映射函数配置数据。代码和配置数据将在论文发表后公开。

A. 用户研究详情

图 A 展示了设计的用户研究界面。该研究包含 16 个问题，每个问题都会向参与者展示一段随机的 7 秒片段，内容是由基线方法和我们的方法驱动的机器人歌唱表演。不同方法的呈现顺序经过打乱。对于每段视频，参与者需要根据真实值，对匿名动画进行 1 到 5 分的评分，分数越高表示表现越好。每个问题包含三个子项：“真实感：表情真实吗，还是缺乏真实感？”“情感丰富度：角色是否富有情感？”以及“唇形同步：唇形是否与音频一致？”为确保参与者充分参与用户研究，只有观看完整段视频并为所有子项评分后，才能进入下一题；否则会弹出警告信息。只有回答完所有问题后才会将其纳入最终结果。最后，我们收集了 40 名参与者的有效提交。

B. 硬件设计

为评估我们的SingingBot框架的实体表现，我们采用高保真人形机器人平台作为硬件后端。机器人头部经过专门设计，可实现富有表现力的面部动画，其仿生机械结构覆盖着柔性硅胶皮肤。

机器人头部集成了32个独立自由度，能够实现唱歌所需的复杂微表情和高频唇部动作。这些执行器的分布经过精心设计，旨在模拟人类面部肌肉组织，如图B所示。

C. 映射函数详情

为弥合生成的数字面部系数与实体机器人硬件之间的差距，我们实现了一套稳健的映射流程，称为BS2Action。该流程将52种兼容ARKit的标准Blendshape转换为控制信号，以驱动机器人平台的32个自由度。

该映射基于一组高保真锚点对，这些锚点对是在专业动画师的协助下创建的。对于每个关键 Blendshape（例如下颌张开、嘴部收拢、BrowIn- nerUp），我们选取一个或多个强度，并由一位动画师

手动调整机器人的 32 个执行器，以获得最精确的相似度。给定这组手动定义的锚点，我们采用分段插值策略，确定任意一组 Blendshape 权重对应的执行器值。

完整的配置文件和映射字典已提供于随附的补充 ZIP 压缩包中，供进一步参考。在 YAML 文件中，每个 Blendshape 都对应一个或多个具有不同 Blendshape 强度的舵机参数文件，其中包含每个舵机的旋转角度。

D. 视频动态结果

我们在补充视频中提供了更多动态实验结果，以更好地展示我们的工作。具体而言，我们展示了与基线方法和消融变体的视觉对比。此外，视频还展示了通过更改参考肖像，我们的方法生成不同风格表演的动态结果。最后，我们包含了 3 段时长约 30 秒的连续歌唱表演，以展示我们方法的时间稳定性。