# TempoSyncDiff：用于低延迟音频驱动说话人头部生成的时间一致性蒸馏扩散模型

**Soumya Mazumdar**
计算机科学与商业系统系，Gargi Memorial Institute of Technology，印度西孟加拉邦加尔各答巴鲁伊普尔，邮编 700144，reachme@soumyamazumdar.com；Vineet KumarRakesh，计算机与信息学组，变能回旋加速器中心，印度西孟加拉邦加尔各答比丹纳加尔 1/AF，邮编 700064，vineet@vecc.gov.in

2026年3月9日

arXiv:2603.06057v1 [cs.CV] 6 Mar 2026

**摘要**

近年来，扩散模型推动了照片级真实感人像合成的发展，但实际应用中的说话人头部生成（THG）仍受诸多因素制约，包括推理延迟较高、闪烁和身份漂移等时间不稳定问题，以及在具有挑战性的语音条件下视听对齐不够理想。本文提出 TempoSyncDiff，这是一种参考图像条件控制的潜在扩散框架，旨在探索高效音频驱动说话人头部生成的少步推理。该方法采用师生蒸馏方案：使用基于标准噪声预测目标训练的扩散教师指导轻量级学生去噪器，使其能够以显著更少的推理步数运行，从而提升生成稳定性。该框架还引入身份锚定和时间正则化，以减轻合成过程中的身份漂移和帧间闪烁；同时，通过基于视素的音频条件控制，实现粗略的嘴部运动控制。在 LRS3数据集上进行的实验报告了去噪阶段组件级指标（相对于变分自编码器重建），并对延迟进行了初步测定，包括仅 CPU 和边缘计算环境下的测量，以及边缘部署的可行性估算。结果表明，蒸馏扩散模型能够保留较强教师模型的大部分重建能力，同时显著降低推理延迟。本研究是面向计算资源受限环境下实用化扩散模型说话人头部生成的初步探索。GitHub：https://mazumdarsoumya.github.io/TempoSyncDiff

**关键词：**说话人头部生成、潜在扩散、蒸馏、时间一致性、边缘推理

## 1 引言

计算能力的进步（例如现代图形处理器（图形处理器））以及大规模数据集的普及，显著推动了深度学习（深度学习）领域的发展，使复杂模型能够学习到恰当的表征，并由此催生了多种先进的深度学习应用 [30]。深度学习最引人注目的应用之一是说话人头部生成（THG），旨在生成逼真的人脸和人类虚拟形象视频，使其能够利用语音、文本、图像、视频及其他视觉数据等多种参考图像，同步模仿人类语音 [22]。在说话人头部生成中，基于扩散的模型提升了人类合成的照片真实感，以及对外观变化的鲁棒性 [20] ；它们也被视为少数几种最优秀的方法之一，适用于通过逐步去噪过程生成图像等任务，使模型能够学习

复杂的数据分布，从而更有效地生成高度逼真的输出 [28]。然而，基于扩散的方法在实际应用中存在多种限制，例如计算成本高、延迟大。

所提出的方法 TempoSyncDiff 是一种基于扩散的方法，旨在根据参考身份图像和语音音频合成人像视频，在保留身份特征的同时生成与所提供语音相符的口部动作，并保持时间连贯性，且能够以更少的推理步数运行。该方法避免了各种扩散模型中常用的多步去噪过程，而这种过程限制了其实时应用能力。此外，常见的基于扩散的方法也可能在生成的视频中产生时间伪影。虽然单帧看起来很逼真，但面部细微纹理（尤其是口腔内部）的细微不一致可能导致闪烁和逐渐出现的身份漂移。此外，扩散模型的语音条件对对齐十分敏感，因此，即使视觉效果逼真，输出仍可能在视素转换中出现细微的时序误差，尤其是在存在噪声音频或语速较快的情况下。为应对这些挑战，所提出的 TempoSyncDiff 采用教师-学生框架，旨在探索少步扩散去噪用于 THG 的可行性。其核心思路是在潜在空间中训练扩散教师，并将其去噪行为蒸馏到轻量级学生模型中，使其相比大多数扩散模型能够以更少的推理步数运行。该方法的条件信息由一组精简输入构成，仅包括参考身份图像和根据所提供音频生成的逐帧视素令牌序列，以减少视觉伪影。模型的时间一致性和身份锚定正则化项被纳入去噪阶段目标，并结合跨身份参考，在训练期间为模型提供支持，促使其稳健地利用参考条件信号。

### 1.1 贡献

1. **采用一致性蒸馏的少步扩散：** 模型通过学生采样器进行训练，该采样器使用多噪声一致性目标近似教师模型的去噪轨迹，使学生能够通过自适应步数计划，仅用极少步数完成采样。2. **结合口腔内部稳定化的时序身份锚定：** 该方法还在规范潜在空间中引入基于身份的锚点，并施加嘴部感兴趣区域（ROI）约束，以稳定不同时刻的牙齿和舌头。3. **通过同步正则化进行视素控制：** 所提出的条件机制基于与视频时间戳对齐的音素和视素词元，并应用视听同步正则化项，以提高嘴唇运动时序的精准度。

## 2 相关工作

将音频作为主要参考图像的音频驱动说话人头部生成模型（THG）涵盖了经典的三维可变形模型（3 DMM）以及各种现代深度生成器。基于音频的模型通过音频视觉对应网络来测量并正则化唇形同步，所需计算资源相对较少；而基于扩散的模型则能带来更逼真的效果和更精准的视素时序 [1, 2]。

扩散概率模型具有更强的去噪能力 [6] ，后续改进 [27, 28] 虽能提升扩散模型的性能，但推理速度较慢 [21] 。因此，更快的求解器和轨迹近似 [25, 26] ，以及一致性风格训练 [24] ，共同推动了少步生成模型的提出。由此，扩散模型通过潜空间建模和时间结构 [23]为视频合成带来益处。扩散概率模型通过逐步反转加噪过程 [6] 来生成样本。去噪扩散隐式模型（DDIM）引入了非马尔可夫采样，以加速推理 [7] ；潜在扩散模型（ LDM）则在学习到的潜空间中运行，从而降低计算量 [8]。DiffTalk [5] 将音频驱动动画肖像概括为一种条件潜扩散过程。

多种模型通过改进求解器或蒸馏来减少采样步数。扩散概率模型求解器（DPM-Solver）和 DPM-Solver++ 为扩散常微分方程（ODE）提供了快速的高阶采样器 [9, 10]。这些模型持续学习一系列映射，从而支持一步和少步生成 [11]。分布匹配蒸馏通过蒸馏提供了一步生成器 [12, 13]。

采用该方法进行视频生成时，持续存在的问题是时间稳定性。通过 Fr´echet 视频距离（FVD）评估视频生成 [17]，而学习感知图像块相似度（LPIPS）则用于衡量图像之间的差异 [16], ，这是更广泛的评估稳定性研究的一部分，该研究侧重于去噪阶段代理指标。

尽管以往的工作提升了扩散效率和视觉真实感，但低延迟推理、稳定的时间动态以及可靠的视听同步所带来的综合挑战仍未得到充分解决。

为缩小这一差距，本文提出了*TempoSyncDiff*，这是一种经过蒸馏、具有时间一致性的扩散框架，旨在高效实现音频驱动的说话人头部生成。

## 3 方法

本文提出了一种新颖的方法，采用基于教师-学生的扩散方法，旨在尽量精简符号，并通过各个方程解决实际问题，包括：(i) 扩散采样缓慢，(ii) 输出随时间闪烁，(iii) 身份逐渐偏离参考人物，以及 (iv) 以稳定的方式注入语音控制。

方法中提供一张身份参考图像 $I_{\text{ref}}$（即需要保持身份的人物），以及语音音频 $a\,(t)$。随后，将音频转换为逐帧的 viseme 令牌序列 $\{v_t\}_{t=1}^{T}$，其中 $v_t$ 表示帧 $t$ 处预期的嘴部形状；如果没有可用的 viseme，则将 $v_t$ 设为全零令牌，以避免失败。

THG 模型需要两　个控制信号：(1) 像谁 （$I_{\text{ref}}$），以及 (2) 如何 在每一　h 时间（$v_t$）。因此，　模型在时间 $t$ 使用的条件定义于方程 1 中

$$\kappa_t = (I_{\text{ref}},\, v_t). \tag{1}$$

直接在像素空间上运行扩散模型的成本相当高，因此使用轻量级自编码器将每个训练帧 $I_t$ 压缩为更小的潜在表示 $z_t$，并在需要时根据方程 2 进行解码。

$$z_t = E(I_t), \qquad \tilde{I}_t = D(z_t). \tag{2}$$

与此同时，使用潜变量改进扩散模型需要进行多次神经网络评估，但与在全分辨率图像上运行相比，每次评估的成本都低得多。在可行性实验期间，将去噪质量与自编码器重建 $I_t$ 进行比较，以便将去噪器的行为与解码器的局限性区分开来。

所提出的扩散训练基于这样一种思想：将干净的潜变量$z_t$逐步加入噪声，而不是训练网络去移除同类噪声。使用方程 3，通过随机高斯噪声创建 $z_t$ 的带噪版本，其中 $s$表示噪声强度；较小的 $s$ 表示轻度污染，较大的 $s$表示重度污染。

$$z_t^{\text{noisy}} = z_t + s\,\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{3}$$

教师去噪器（$\theta$）接收含噪条件潜变量，并使用方程 4 预测添加的噪声。

$$\hat{\epsilon}_\theta = \epsilon_\theta(z_t^{\text{noisy}},\, \kappa_t). \tag{4}$$

然后，使用方程 5 训练教师模型，使预测的噪声与真实噪声相匹配，从而提供更强的去噪器（更高的质量），即使完成推理需要多个推理步数。

$$\mathcal{L}_{\text{teach}} = \mathbb{E}\left[\|\epsilon - \hat{\epsilon}_\theta\|_2^2\right]. \tag{5}$$

之后，使用方程 7 直接训练学生去噪器（$\phi$），使其模仿教师去噪器。

$$\hat{\epsilon}_\phi = \epsilon_\phi(z_t^{\text{noisy}},\, \kappa_t), \tag{6}$$

并优化学生，使其匹配教师的输出：

$$\mathcal{L}_{\text{dist}} = \mathbb{E}\left[\|\hat{\epsilon}_\phi - \hat{\epsilon}_\theta\|_2^2\right]. \tag{7}$$

在大多数情况下，常用的扩散模型相对较慢，因为需要多个去噪步数才能完成处理；但所提出的学生去噪器通过模仿教师模型，以少得多的步数实现了更优质、可接受的去噪，最终降低了延迟。

除此之外，生成模型的一个常见问题是身份漂移：即使每一帧看起来都很逼真，人物身份也会随时间逐渐改变。这是因为模型“偏好”长相平均化的面孔。因此，身份编码器 $f_{\text{id}}(\cdot)$ 将图像映射为身份特征向量。生成的帧 $\hat{}I_t$ 随后将在方程 8 中与参考身份匹配。

$$\mathcal{L}_{\text{id}} = \sum_{t=1}^{T} \left(1 - \cos\left(f_{\text{id}}(\hat{I}_t),\, f_{\text{id}}(I_{\text{ref}})\right)\right). \tag{8}$$

另一个主要问题是闪烁，即使是细小纹理和口腔内部，也会在帧与帧之间出现不自然的变化。因此，该方法通过强制实现时间一致性，确保连续帧之间不会发生过于突兀的变化。该方法使用变形函数 $\mathcal{W}(\cdot)$，利用方程 9 中所述的运动信息，将前一帧大致对齐到下一帧。

$$\mathcal{L}_{\text{temp}} = \sum_{t=2}^{T} \left\| \hat{I}_t - \mathcal{W}(\hat{I}_{t-1}) \right\|_1 . \tag{9}$$

现在，教师训练使用方程 10 最小化去噪损失和正则项，而学生去噪器的训练则使用方程 11 中的蒸馏损失替代教师去噪器损失。

$$\mathcal{L}^{\text{teacher}} = \mathcal{L}_{\text{teach}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}}. \tag{10}$$

$$\mathcal{L}^{\text{student}} = \mathcal{L}_{\text{dist}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}}. \tag{11}$$

$\mathcal{L}_{\text{teach}}$ 和 $\mathcal{L}_{\text{dist}}$ 的主要目标是确保去噪能够解决上述两个明显问题：身份漂移（$\mathcal{L}_{\text{id}}$）和闪烁（$\mathcal{L}_{\text{temp}}$）。随后，权重 $\lambda_{\text{id}}$ 和 $\lambda_{\text{temp}}$ 会恰当地平衡这两个问题。
核心目标（$\mathcal{L}_{\text{teach}}$ 或 $\mathcal{L}_{\text{dist}}$）确保去噪能够正常进行。额外的项用于解决 THG 的两个明显问题：身份漂移（$\mathcal{L}_{\text{id}}$）和闪烁（$\mathcal{L}_{\text{temp}}$）。权重 $\lambda_{\text{id}}$ 和 $\lambda_{\text{temp}}$ 则用于平衡对这两个问题的修正力度。
训练期间，有时会将参考图像替换为具有不同身份的图像，以增强条件信息的鲁棒性。假设 $p_{\text{mismatch}}$ 是使用不匹配参考图像的概率，如方程 12 所述；因此，参考图像始终来自同一段视频片段，以迫使模型依赖条件输入，并在训练中使用不匹配参考图像的设置来对身份控制进行压力测试。

$$p_{\text{mismatch}} = 0.5. \tag{12}$$

## 4 架构与推理模式

所提出方法的详细架构总结如图 1 所示。该方法将视素词元转换后，与参考身份帧编码嵌入身份信息中，并以两者作为潜在扩散去噪器的条件信息，同时生成由 VAE解码器 解码的去噪器潜变量。

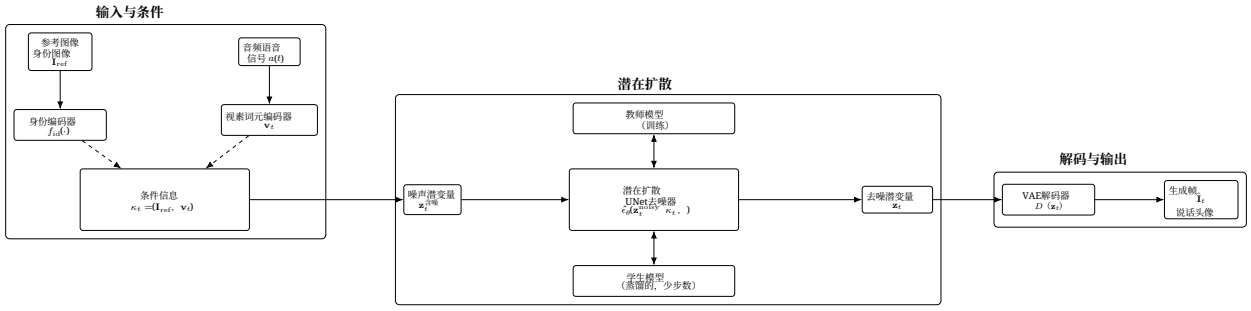


图1：TempoSyncDiff概述。参考身份图像与语音音频被编码为身份特征和视素词元，形成条件信息 $\kappa_t$。潜在扩散UNet在该条件信息下对含噪潜变量$\mathbf{z}_t^{\text{noisy}}$ 进行去噪。在训练过程中，教师模型指导蒸馏学生模型，从而实现少步采样。去噪潜变量$\mathbf{z}_t$^由变分自编码器解码，生成说话人头部帧$\mathbf{I}_t$。

师生关系的设置如图2所示：先训练教师模型，再优化学生以匹配教师模型的去噪预测。

图3中所构想的边缘部署流水线中的推理。
模型的训练与推理在数学形式上与第 3 节中定义的内容相似。编写这些算法是为了描述模型的运行方式，并探究：(i) 输入系统的数据是什么；(ii) 前向传播会计算什么；(iii) 针对模式失效会采用哪些损失；以及 (iv) 发生失效时会更新哪些参数。首先训练教师去噪器，因为如算法 1 所定义，它为学生去噪器规定了目标行为。

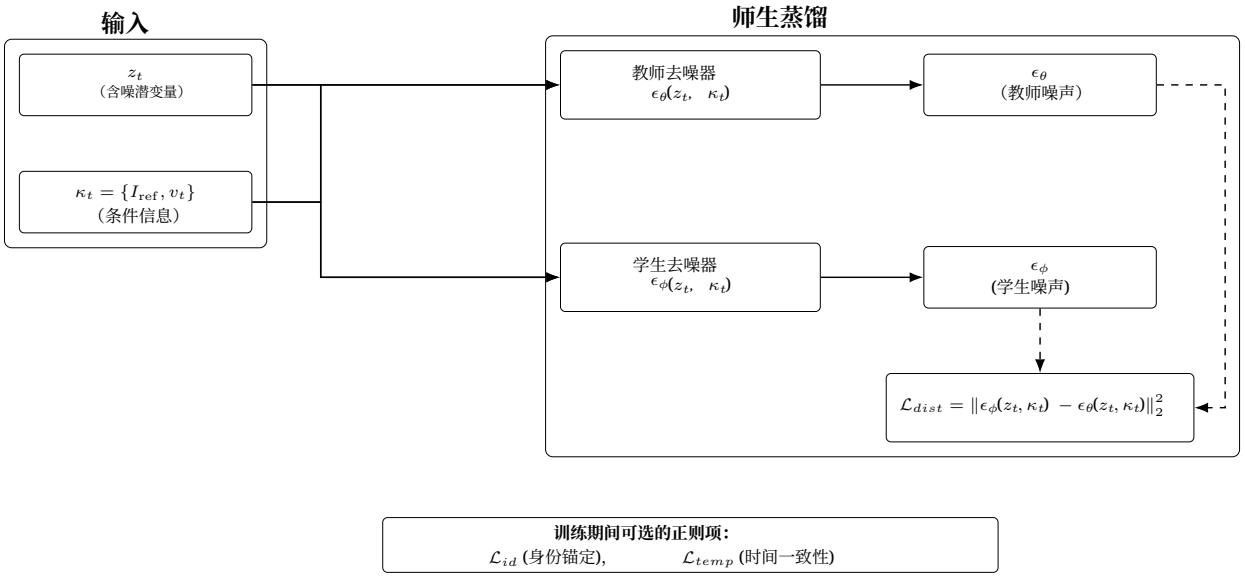


图2：TempoSyncDiff 中采用的师生蒸馏。冻结的教师去噪器 $\epsilon_\theta$ 根据条件 $\kappa_t = \{I_{\text{ref}}, v_t\}$，为含噪潜变量 $z_t$ 预测噪声。学生去噪器 $\epsilon_\phi$ 接收相同的输入，并通过蒸馏损失 $\mathcal{L}_{dist}$进行训练，以匹配教师模型的预测。训练期间还可选择使用身份正则项和时间正则项。

**Algorithm 1** Teacher Training (TempoSyncDiff)

**要求：** 训练视频片段 $\{I_{1:T}\}$、参考帧 $I_{\text{ref}}$、音频 $a(t)$、噪声强度（或调度）参数、权重 $\lambda_{\text{id}}, \lambda_{\text{temp}}$

将音频 $a(t)$ 转换为视素词元 $\{v_t\}_{t=1}^{T}$（若缺失，则使用全零词元）

对于每个小批次

采样帧索引 $t$（或一小段帧窗口）

编码潜变量 $z_t \leftarrow E(I_t)$

采样噪声 $\epsilon \sim \mathcal{N}(0, I)$

创建含噪潜变量 $z_t^{\text{noisy}} \leftarrow z_t + s\epsilon$ ▷ 扩散式扰动；方程 (3)

构造条件信息 $\kappa_t \leftarrow (I_{\text{ref}}, v_t)$ ▷ 方程 (1)

预测添加的噪声 $\hat{\epsilon}_\theta \leftarrow \epsilon_\theta(z_t^{\text{noisy}}, \kappa_t)$

计算教师去噪损失 $\mathcal{L}_{\text{teach}} \leftarrow \|\epsilon - \hat{\epsilon}_\theta\|_2^2$ ▷ 方程 (5)

去噪并解码以获得 $\hat{I}_t$ ▷ 用于时间与身份正则项

使用方程 (8) 计算身份损失 $\mathcal{L}_{\text{id}}$

使用方程 (9) 计算时间损失 $\mathcal{L}_{\text{temp}}$（需要相邻帧）

总损失 $\mathcal{L} \leftarrow \mathcal{L}_{\text{teach}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}}$ ▷ 方程 (10)

通过梯度下降更新教师模型参数 $\theta$

结束循环

教师去噪器训练完成后，便开始训练学生。训练的目标不是超越教师模型，而是使用复杂度更低的网络和更少的推理步数来近似教师预测，如算法2所述。

推理使用学生去噪器，因为它是采用该算法的低延迟模型 3.

## 5 实验

### 5.1 数据集

所提出模型的训练使用 LRS3-TED [14]，该数据集包含相对丰富多样的语音内容，并提供人脸轨迹和对齐的转录文本。处理流程中使用的场景时长约为 6 秒，帧率为 25 帧每秒；每个场景均提供包含文件路径的清单格式，以及视素令牌文件。此外，

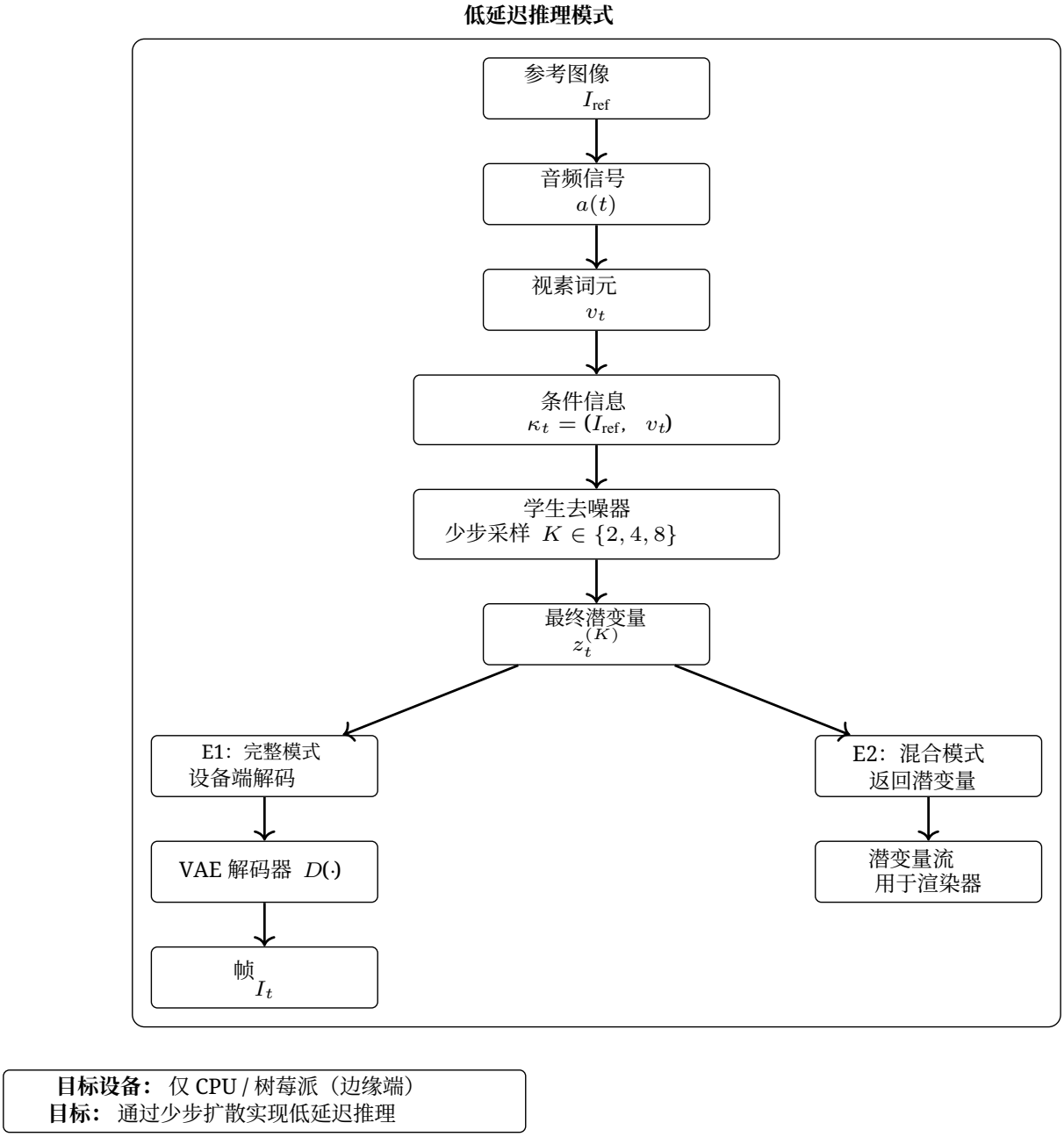


图 3：面向边缘部署的推理。学生模型在潜在空间中执行 $K$ 步去噪。在 **E1（完整模式）**中，VAE 解码器在设备端运行并输出帧。在 **E2（混合模式）**中，返回潜变量，以便延后解码或进行外部渲染。

HDTF [15]数据集被视为自然场景中的说话人脸常用评估数据集；全面的跨数据集评估将留待未来工作开展。

### 5.2 训练细节

教师模型训练使用批量大小 16、学习率 $2 \times 10^{-5}$ 和早停耐心值 8。教师模型日志显示，最佳验证损失约为 0.047；而在当前设置下，学生模型蒸馏训练的最佳验证损失约为 0.078。

**算法2** 学生蒸馏 (TempoSyncDiff)

**要求：** 冻结的教师模型 $\epsilon_\theta$、学生 $\epsilon_\phi$、视频片段 $\{I_{1:T}\}$、参考帧 $I_{\text{ref}}$、音频 $a(t)$、权重 $\lambda_{\text{id}}, \lambda_{\text{temp}}$

1: 将音频 $a(t)$ 转换为视素词元 $\{v_t\}_{t=1}^{T}$ （若缺失则为零）
2: **对**每个小批次**执行**
3: 采样一个帧索引 $t$（或一个短窗口）
4: 编码潜变量 $z_t \leftarrow E(I_t)$
5: 采样噪声 $\epsilon \sim \mathcal{N}(0,\ I)$ 并构造 $z_t^{\text{noisy}} \leftarrow z_t + s\epsilon$
6: 构造条件信息 $\kappa_t \leftarrow (I_{\text{ref}},\ v_t)$
7: 教师预测 $\hat{\epsilon}_T \leftarrow \epsilon_\theta(z_t^{\text{noisy}},\ \kappa_t)$ ▷ 教师目标
8: 学生预测 $\hat{\epsilon}_S \leftarrow \epsilon_\phi(z_t^{\text{noisy}},\ \kappa_t)$
9: 蒸馏损失 $\mathcal{L}_{\text{dist}} \leftarrow \|\hat{\epsilon}_S - \hat{\epsilon}_T\|_2^2$ ▷ 方程（7）
10: 去噪并解码以获得 $\hat{I}_t$（可选但推荐）
11: 使用方程（8）计算 $\mathcal{L}_{\text{id}}$，并使用方程（9）计算 $\mathcal{L}_{\text{temp}}$
12: 总损失 $\mathcal{L} \leftarrow \mathcal{L}_{\text{dist}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}}$ ▷ 方程 (11)
13: 通过梯度下降更新学生参数 $\phi$
14: **结束循环**

**算法 3** 低步数推理（模式 E1 / E2）

**要求：** 参考图像 $I_{\text{ref}}$、音频 $a(t)$、学生去噪器 $\epsilon_\phi$、步数 $K \in \{2, 4, 8\}$、模式 $\in \{\text{E1},\ \text{E2}\}$

1: 将音频 $a(t)$转换为视素词元 $\{v_t\}_{t=1}^{T}$ （若缺失则填零）
2: 为每一帧 $t$ 形成条件流 $\kappa_t \leftarrow (I_{\text{ref}}, v_t)$
3: **如果**模式 =E1 **则**
4: 从编码器 $E(\cdot)$ 获取初始潜变量 $\{z_t^{(0)}\}$ （或用噪声初始化）
5: **否则**
6: 从上游潜变量源或缓存编码初始化潜变量 $\{z_t^{(0)}\}$
7: **结束如果**
8: **对于** $k = 1$到 $K$ **执行**
9: **对于** $t = 1$到 $T$ **执行**
10: 预测噪声 $\hat{\epsilon} \leftarrow \epsilon_\phi(z_t^{(k-1)},\ \kappa_t)$
11: 更新潜变量 $z_t^{(k)} \leftarrow z_t^{(k-1)} - \eta_k\hat{\epsilon}$ ▷ 简单去噪更新（步长 $\eta_k$）
12: **结束循环**
13: **结束循环**
14: **如果**模式 = E1 **则**
15: 对所有 $t$解码帧 $\hat{} I_t \leftarrow D(z_t^{(K)})$
16: 返回 $\{\ \hat{I}_t\}_{t=1}^{T}$
17: **否则**
18: 返回最终潜变量 $\{z_t^{(K)}\}_{t=1}^{T}$ （延后解码或选择性解码）
19: **结束 if**

### 5.3 指标

THG 评估涵盖视觉真实感、身份保持、时间稳定性以及音频-视觉对齐等多个方面，常用指标包括用于感知相似度评估的**LPIPS** [16] 、通过嵌入网络计算的**身份余弦相似度** 、用于衡量**时间稳定性** 的指标（如相邻帧之间的闪烁或形变误差），以及采用 SyncNet 类方法评估的**唇形同步准确度** [1]。此外，**延迟**也是实时部署时需要考虑的实际因素。

在本研究中，评估主要关注去噪阶段组件级指标，而非完整的端到端视频真实感，具体包括相对于VAE重建的峰值信噪比、相邻帧之间的时间L1差异，以及一种简单的闪烁统计量，即像素差异随时间变化的标准差。四个评估案例的均值±标准差汇总于表1。

与噪声基线相比，教师模型将去噪质量提升了约 5.24 dB 的峰值信噪比。尽管蒸馏学生模型保留了大部分提升，但与教师模型相比，重建质量仅小幅下降。然而，初步可行性实验中的时间代理指标尚未显示出相较于 VAE 重建更高的平滑度，

表1：相对于VAE重建测得的去噪阶段结果（四个评估案例的均值 ±标准差）。峰值信噪比以VAE重建为参照计算，以便将去噪行为与解码器局限性区分开来。时间L1和闪烁统计根据相邻帧像素差异计算，用于概括短期时间变化。

| 指标 | 噪声解码 | 教师去噪 | 学生去噪 |
|---|---|---|---|
| 峰值信噪比 vs. 变分自编码器重建（dB） | $25.71 \pm 0.02$ | $30.95 \pm 0.06$ | $29.97 \pm 0.06$ |
| 时间L1（相邻帧差异） | $0.0184 \pm 0.0036$ | $0.0365 \pm 0.0020$ | $0.0402 \pm 0.0021$ |
| 闪烁标准差（像素差异标准差） | $0.00325 \pm 0.00171$ | $0.00336 \pm 0.00159$ | $0.00337 \pm 0.00155$ |

这一结果在一定程度上是意料之中的，因为 VAE 重建往往会生成过度平滑的输出，从而人为地降低时间变化。

### 5.4 质量–延迟权衡（占位符）

图4预留用于展示不同步数（例如 $K \in \{2, 4, 8\}$）下的质量—延迟前沿，并将在相关数据可用后采用主要感知指标和时间指标。图4中已预留此图。

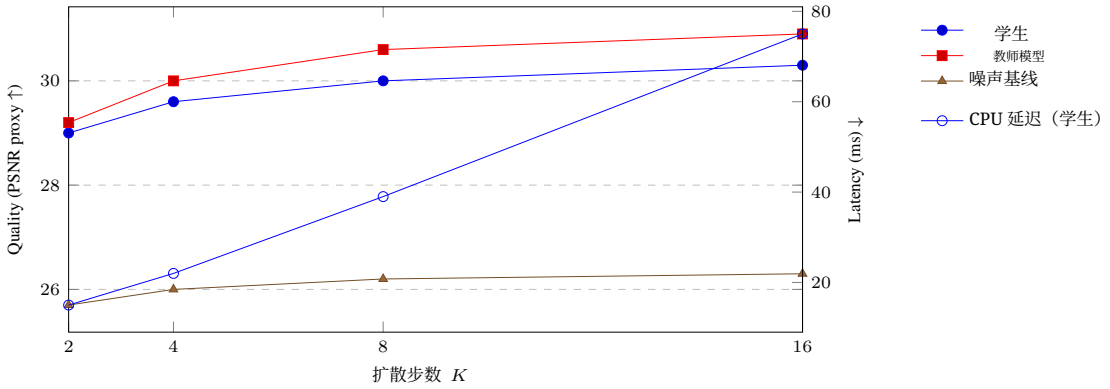


图4：不同步数下的质量–延迟权衡（模板）。在延迟轴上添加了树莓派 5 边缘推理延迟（E2混合模式）。预期效果是，对于较小的$K$，以显著更低的延迟实现接近教师模型的质量。

## 6 仅 CPU 与边缘计算机可行性

### 6.1 纯CPU推理

纳入纯CPU推理是为了验证轻量级可行性。在一台具有代表性的 x86 CPU 设备上测得的纯CPU延迟统计数据列于表 2。

表 2：纯CPU推理表征（实测结果）。

| 步数 | 分辨率 | 平均延迟（毫秒） | p95延迟（毫秒） | FPS |
|---|---|---|---|---|
| 2 | $128 \times 128$ | 13.21 | 16.58 | 75.72 |
| 4 | $128 \times 128$ | 21.56 | 25.28 | 46.38 |
| 8 | $128 \times 128$ | 38.61 | 43.11 | 25.90 |

### 6.2 边缘计算推理（RaspberryPi 4 Model B）

树莓派 4 Model B 为评估边缘计算部署的可行性提供了一个实际案例。该设备配备一颗 $64-bit$ 四核 Cortex-A72 (ARMv8) 处理器，内存为 $4\,GB$ LPDDR4-3200，运行频率为 $1.8\,GHz$ [18,19]。各类基准测试表明，轻量级卷积神经网络推理（例如 MobileNet 系列模型）可以实时运行

在经过优化的执行 [32,33]模式下，可在树莓派 4 和树莓派 5 上运行。根据以下公开基准测试，以及对所提出流水线阶段按 FLOP 缩放的核算，表 3 给出了降低分辨率和采用低步数推理的实际可行性估算。

表 3：树莓派 5 边缘推理表征（估算值）。为便于实现，降低了分辨率。若延后解码，E2 模式将输出潜变量。

| Mode | 步数 | 分辨率 | 平均延迟（毫秒） | p95延迟（毫秒） | FPS |
|---|---|---|---|---|---|
| E1 完整版 | 2 | $128 \times 128$ | 260.88 | 368.58 | 3.83 |
| E1 完整版 | 4 | $128 \times 128$ | 403.93 | 407.00 | 2.48 |
| E2 混合 | 2 | 潜变量 | 172.05 | 174.52 | 5.81 |
| E2 混合 | 4 | 潜变量 | 335.57 | 373.09 | 2.98 |

## 7 讨论

本提案的多项实验表明，蒸馏的学生去噪器能够近似表现得强于扩散教师，同时大幅减少推理所需的去噪步数。在相对重建评估中，教师模型相对于含噪潜变量基线在质量上有显著提升，而学生在重建精度仅有适度下降的情况下保留了大部分改进。因此，该结果表明蒸馏的作用在于近似表现得如同计算成本更高的模型，从而实现显著更快的推理。

所提方法中报告的时间代理指标并未表明，相对于VAE重建，逐帧变化的估计有显著降低，这一观察完全符合预期。VAE重建往往因解码器的平均化行为而产生更好的时间输出，即使感知伪影仍然存在，也会降低表观时间变化。因此，时间正则化带来的改进可能无法被简单的像素差异代理指标完全捕捉，这使得与感知更一致的时间指标（如运动补偿扭曲误差或视频感知指标）在未来评估中可能提供更可靠的时间稳定性指标。

该方案的跨身份参考替换机制（$p_{\text{mismatch}} = 0.5$）旨在防止条件崩溃，迫使方案仅依赖所提供的参考身份，而不是隐式地直接从驱动帧复制身份信息。尽管该机制已在训练流程中实现，但当前实验尚未单独量化其影响；因此，需要专门制定跨身份评估方案，以衡量参考身份与驱动身份有意分离时的身份保持效果。因此，仅 CPU 和边缘计算的可行性测量表明，低步数推理在降低的分辨率下具有计算可行性，支持了在资源受限环境中探索蒸馏扩散模型的动机。

## 8 局限性与伦理

所提出的方法探索了蒸馏扩散模型，该模型能够以相对较低的延迟生成合成视频；然而，仍存在若干局限性。与变分自编码器重建相比，本研究开展的实验主要聚焦于去噪的潜变量阶段，尚未充分衡量生成的渲染视频的感知质量。时间稳定性仅通过逐帧 L1 差异和闪烁统计等简单代理指标进行评估，这些指标可能无法捕捉更复杂的视觉问题，例如嘴部抖动或身份漂移。此外，论文尚未对标准唇形同步评估进行充分评估。仅 CPU 设备和边缘计算机上的可行性实验也在降低的分辨率下进行；要在更高分辨率下实现实时性能，可能需要进一步优化，例如模型压缩。

THG 系统也引发了严重的伦理担忧，因为这些系统可能被滥用于冒充、传播错误信息或侵犯隐私。因此，相关责任在于谨慎处理训练数据，包括确保获得适当的许可和同意；系统还应考虑采用水印或元数据等安全保障措施，以帮助识别合成内容。在现实场景中，明确告知用户内容是人工生成的至关重要。

## 9 结论

本文介绍了一种名为 TempoSyncDiff 的新型框架，这是一种基于参考条件的音频驱动扩散 THG 模型。该方法研究了师生蒸馏策略，使轻量级学生去噪器能够以显著更少的推理步数，近似表现相对更好的扩散教师模型。实验结果表明，在去噪阶段评估协议下，学生模型保留了大部分重建质量，同时达到了教师模型的表现。所提出的框架引入了身份锚定和时间正则化，以识别生成过程中的漂移和帧间闪烁。主要基于仅 CPU 的测量以及边缘计算测量的结果进一步表明，少步扩散推理在降低的分辨率下具有计算可行性。

尽管本研究主要关注可行性和去噪阶段指标，但也进一步评估了标准的感知视听同步指标。未来的研究方向是进一步评估完整的端到端视频质量，并改进时间稳定性指标和跨身份验证。

## 参考资料

[1] K. R. Prajwal、R. Mukhopadhyay、V. P. Namboodiri 和 C. V. Jawahar，《野外语音到唇形生成只需一位唇形同步专家》，载于*ACM*多媒体会议论文集，2020年。[2] A. Roy、B. S. Gyanchandani、J. Sahu 和 K. Mallikharjuna Rao，《SyncNet：协调节点以实现高效学习》，载于*2024*年软件、电信与计算机网络国际会议（*Soft-COM*），2024年，第1–6页。doi:10.23919/SoftCOM62040.2024.10721994。[3] Y. Zhou、X. Han、E. Xie 等，《MakeItTalk：具备说话人感知能力的说话头动画》，载于*ACM*多媒体会议论文集，2020年。[4] W. Zhang、X. Cun、X. Wang、Y. Zhang、X. Shen、Y. Guo、Y. Shan、F. Wang，《SadTalker：为风格化的音频驱动单图说话人脸动画学习逼真的3D运动系数》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），2023年。[5] S. Shen 等，《DiffTalk：为通用音频驱动肖像动画定制扩散模型》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），2023年。[6] J. Ho、A. Jain 和 P. Abbeel，《去噪扩散概率模型》，载于神经信息处理系统进展（*NeurIPS*），2020年。[7] J. Song、C. Meng 和 S. Ermon，《去噪扩散隐式模型》，载于国际学习表征会议（*ICLR*），2021年。[8] R. Rombach、A. Blattmann、D. Lorenz、P. Esser、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（*CVPR*），2022年。[9] C. Lu、Y. Zhou、F. Bao、J. Chen、C. Li、J. Zhu，《DPM-Solver：约10步完成扩散概率模型采样的快速ODE求解器》，载于神经信息处理系统进展（*NeurIPS*），2022年。[10] C. Lu、Y. Zhou、F. Bao、J. Chen、C. Li、J. Zhu，《DPM-Solver++：用于扩散概率模型引导采样的快速求解器》，载于国际学习表征会议（*ICLR*），2023年。[11] Y. Song、P. Dhariwal、M. Chen、I. Sutskever，《一致性模型》，*arXiv*预印本 *arXiv:2303.01469*，2023年。[12] T. Yin 等，《利用分布匹配蒸馏实现单步扩散》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），2024年。[13] Z. Zhou 等，《扩散模型的简单快速蒸馏》，载于神经信息处理系统进展（*NeurIPS*），2024年。[14] T. Afouras、J. S. Chung、A. Zisserman，《LRS3-TED：用于视觉语音识别的大规模数据集》，*arXiv*预印本 *arXiv:1809.00496*，2018年。[15] Z. Zhang、L. Li、Y. Ding、C. Fan，《利用高分辨率音视频数据集和光流引导的单样本说话人脸生成》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），2021年。[16] R. Zhang、P. Isola、A. A. Efros、E. Shechtman、O. Wang，《深度特征作为感知度量的惊人效能》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），2018年。

[17] T. Unterthiner、S. van Steenkiste、K. Kurach、R. Marinier、M. Michalski、S. Gelly，《FVD：一种用于视频生成的新度量》，载于国际学习表征会议（*ICLR*）研讨会，2019年。[18] Raspberry Pi Ltd.，《Raspberry Pi 4 Model B 产品简介》，产品简介，2026年2月。[19] Raspberry Pi Ltd.，《Raspberry Pi 4 Model B 产品页面》，访问于2026年。[20] J. Ho、A. Jain、P. Abbeel，《去噪扩散概率模型》，载于神经信息处理系统进展第*33*卷（*NeurIPS 2020*），2020年。[在线]。可访问：https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html[21] M.-M. Cheng、T. Hu、F. Khan、L. Li、S. Li、T. Liu、J. Van De Weijer、Y. Wang、J. Yang、S. Yang，《快速扩散：重新思考编码器在扩散模型推理中的作用》，载于*NeurIPS2024*，2024年。doi:10.52202/079017- 2706。[22] V. K. Rakesh、S. Mazumdar、S. Pal、A. Das、T. Samanta，《说话头生成的进展：技术、度量与挑战综述》，*TheVisual Computer*，第42卷，第1期，2025年12月。doi:10.1007/s00371-025-04232-w。[23] A. Blattmann、R. Rombach、H. Ling、T. Dockhorn、S. W. Kim、S. Fidler、K. Kreis，《对齐潜变量：利用潜在扩散模型进行高分辨率视频合成》，载于*CVPR 2023*，加拿大不列颠哥伦比亚省温哥华，2023年，第22563–22575页。doi:10.1109/CVPR52729.2023.02161。[24] Y. Song、P. Dhariwal、M. Chen、I. Sutskever，《一致性模型》，载于*ICML 2023*，美国夏威夷州檀香山，2023年，第202卷，第32211–32252页。[25] J. Song、C. Meng、S. Ermon，《去噪扩散隐式模型》，载于*ICLR2021*，线上会议，奥地利，2021年。[26] S. Koyejo、S. Mohamed、A. Agarwal、D. Belgrave、K. Cho、A. Oh（编），神经信息处理系统进展第*35*卷：*NeurIPS 2022*，美国路易斯安那州新奥尔良，2022年。[27] T. Karras、M. Aittala、T. Aila、S. Laine，《阐明基于扩散的生成模型设计空间》，载于*NeurIPS 2022*，美国路易斯安那州新奥尔良，2022年。[28] P. Dhariwal、A. Q. Nichol，《扩散模型在图像合成方面胜过生成对抗网络》，载于*NeurIPS2021*，第8780–8794页。[29] A. Q. Nichol、P. Dhariwal，《改进的去噪扩散概率模型》，载于*ICML 2021*，第139卷，第8162–8171页。[30] S. Issabayeva、A. Katyetova，《人工智能深度学习：哈萨克斯坦案例》，载于信息与通信进展，2025年，第586–602页。[31] T. Karras、M. Aittala、T. Aila、S. Laine，《阐明基于扩散的生成模型设计空间》，*CoRRabs/2206.00364*，2022年。doi:10.48550/ARXIV.2206.00364。[32] T. Rice，《在树莓派 4 和 5 上进行实时推理（40 帧每秒！）》，PyTorch 教程，2025年。[33] M. N. Jacob，《低成本边缘单板计算机上的功耗感知图像分类：使用 MobileNetV3-Small 和 ResNet-18 的可复现实测》，*TechRxiv*，2025年10月。doi:10.36227/techrxiv.176187020.04189831/v1。[34] 微软，《ONNX Runtime 性能分析工具与 onnxruntime perf test》，ONNX Runtime 文档，访问于2026年。

## 作者贡献

**SoumyaMazumdar**提出了研究方法，编写了核心算法，开展了实验，提出了理论，进行了测试，并撰写了论文初稿。**Vineet Kumar Rakesh** 独立验证了实验结果和分析，以确保技术准确性和一致性。所有作者均审阅并批准了最终稿。

## 致谢

作者衷心感谢可变能量回旋加速器中心（VECC）、原子能部（DAE）和印度政府（GoI）提供的支持，感谢他们准许作者使用开展本研究所需的基础设施和工具，并提供技术协助。我们还要感谢 VECC 图书馆的工作人员，在整个研究过程中给予的帮助与支持。

## 利益冲突声明

所有作者均声明，他们不存在任何可能被视为会影响本研究所述工作的财务或个人关系。不存在利益冲突。

## 数据可用性声明

本研究使用的数据集均可公开获取。主要实验使用 LRS3-TED 数据集进行，HDTF 数据集则作为说话人头部生成研究的标准基准。作者不再分发原始数据集；应根据相应的许可和使用条款，从其官方来源获取。与本研究相关的代码、预处理详情和实验配置现已提供，或将通过项目代码仓库提供：https://mazumdarsoumya.github.io/ TempoSyncDiff。

## 作者简介



**Soumya Mazumdar** 正在攻读双学位：在 Gargi Memorial Institute of Technology 攻读计算机科学与商业系统工学学士学位，并在印度理工学院马德拉斯分校攻读数据科学理学学士学位。他参与了跨学科研究，在爱思唯尔、施普林格、电气电子工程师学会、威立和 CRC 出版社出版的期刊及编著论文集中发表了 25 篇以上的论文。他的研究兴趣包括人工智能、机器学习、6G 通信、医疗保健技术和工业自动化。



**Vineet Kumar Rakesh** 是印度原子能部可变能量回旋加速器中心（VECC）的技术官员（科学类），在软件工程、数据库系统和人工智能领域拥有逾 22 年经验。他的研究重点包括说话人头部生成、唇读，以及用于实时远程会议的超低码率视频压缩。他正在孟买的霍米·巴巴国家研究所攻读哲学博士学位。Rakesh 先生曾参与 VECC 的办公自动化、OCR 系统和数字化转型项目。他是印度工程师学会的准会员，并曾获原子能部集体成就奖。