SLD-L2S：用于高保真唇语到语音合成的分层子空间潜在扩散

梁一凡[1,2], **李安东** [1,2]、**杨康** [3]、**于国辰** [4], **刘方坤** [1,2]、**戴玲玲** [1,2]、**李晓东** [1,2], **Chengshi Zheng** [1,2]*

[1]中国科学院声学研究所，中国科学院，北京，中国 [2]中国科学院大学，北京，中国[3]人工智能与计算机科学学院，江南大学，无锡，中国[4]智谱 AI liangyifan@mail.ioa.ac.cn, liandong@mail.ioa.ac.cn, 6233110046@stu.jiangnan.edu.cn, guochen.yu@aminer.cn, liufangkun@mail.ioa.ac.cn, dailingling@mail.ioa.ac.cn, lxd@mail.ioa.ac.cn, cszheng@mail.ioa.ac.cn

arXiv:2602.11477v1 [eess.AS] 12 Feb 2026

摘要

尽管近年来唇语到语音合成（L2S）取得了显著进展，但当前最先进的方法通常依赖梅尔频谱图或离散自监督学习（SSL）词元等中间表征。潜在扩散模型（LDM）在这一任务中的潜力仍未得到充分探索。本文提出 SLD-L2S，这是一种基于分层子空间潜在扩散模型的新型唇语到语音框架。我们的方法旨在将视觉唇部运动直接映射到预训练神经音频编解码器的连续潜在空间，从而避免传统中间表征固有的信息损失。我们方法的核心是一种分层架构，它通过多个并行子空间处理视觉表示，并由子空间分解模块启动。为有效增强这些子空间内部及彼此之间的交互，我们设计了扩散卷积块（DiCB）作为网络骨干。此外，我们采用重参数化流匹配技术，直接生成目标潜在向量。这样便能在训练过程中有理有据地引入语音语言模型（SLM）和语义损失，突破传统流匹配目标的限制，并提升合成语音质量。实验结果表明，SLD-L2S 在多个基准数据集上实现了最先进的生成质量，在客观评估和主观评估中均优于现有方法。

引言

唇语到语音（L2S）合成是仅根据视觉唇部运动生成可听语音的任务（Milner 和 Le Cornu 2015年），在多种应用中具有潜力，包括自动视频配音、在复杂声学环境中的稳健通信，以及为发声困难和失声等发声障碍人士提供的辅助技术。L2S 合成的主要挑战在于，将模糊的视觉线索映射到复杂的语音声学特征这一问题本身具有不适定性。传统的 L2S 方法（Prajwal 等人 2020年；Wang 和 Zhao 2022年；Kim、Kim 和 Chung 2024年）大多局限于受约束的场景，其特点是词汇有限且依赖说话人依赖模型，因此在不受约束的真实场景中实用性有限。

自监督学习（SSL）（Hsu 等，2021年）和语音合成（Ren 等，2020年；Polyak 等，2021年）的最新进展推动了唇语到语音（L2S）领域的显著发展，使系统能够根据对应的视觉输入生成高保真语音。该领域已从早期受限的设置（Cooke 等，2006年；Harte 和 Gillen，2015年）发展到能够在开放的真实环境中有效运行的稳健多说话人系统（Chung 等，2017年；Afouras、Chung 和 Zisserman，2018年）。自监督学习的融入尤其带来了变革，使模型能够从海量无标签数据中学习丰富的语义表征。例如，近期研究（Hsu 等，2023年；Choi、Kim 和 Ro，2023年）利用 AV-HuBERT（Shi 等，2022年）等强大的视听自监督学习模型，显著提升了合成语音的可懂度。与此同时，基于扩散的生成模型（D´efossez 等，2022年；Lipman 等，2022年）已成为唇语到语音合成的强大范式，为合成质量树立了新的基准（Choi、Hong 和 Ro，2023年；Yemini 等，2023年；Choi 等人，2025年；Kim 等人，2025年）。

尽管取得了显著进展，现有的唇语到语音方法仍主要依赖生成中间表征，例如梅尔频谱图或离散自监督学习令牌。这些表征虽然能有效传递语义内容，却往往无法捕捉高保真波形生成所需的细粒度声学细节。这一局限主要源于两个因素。首先，视觉到音频映射本质上是一个定义不适定的一对多问题：同一段唇部运动序列可能对应多种有效的语音呈现形式，其韵律、情绪和说话风格各不相同。作为一种基于物理定义的声学表征，梅尔频谱图缺乏对这种多样性的建模能力。其次，尽管离散自监督学习令牌在语义表征方面很有效，但其本质上较为粗糙，可能会丢弃感知质量所必需的关键声学细节。其他方法（Liang 等，2025a,b）采用端到端训练，绕过显式映射策略，以避免潜在的信息损失。然而，生成的语音质量仍然有限，这表明此类方法需要高质量的大规模音频——

*通讯作者：Chengshi Zheng 版权所有 © 2026年，人工智能促进协会 (www.aaai.org)。保留所有权利。

视觉数据集来重新训练声码器。这些缺陷促使我们提出一个核心问题：我们能否找到一种替代表征，更有效地捕捉高保真唇语到语音合成所需的详细声学特征？

幸运的是，文本到语音（TTS）合成领域的最新进展（Wang 等人，2023；Du 等人，2024）已证明神经音频编解码器（Zeghidour 等人，2021；D´efossez 等人，2022）的有效性。这些编解码器能够学习将音频波形编码为离散潜在表示，并从中解码出音频波形。它们通过捕捉细致的声学特征，在生成高保真语音方面表现出色。这些编解码器在 TTS 任务中的成功表明，它们很可能也有助于唇语到语音合成。然而，现有的唇语到语音方法尚未充分探索音频编解码器潜在向量在唇语到语音任务中的潜力。尽管 Uni-Dubing（Lei 等人，2024）已尝试将 Hifi-Codec（Yang 等人，2023）用于语音合成，但它将问题设定为在语言建模范式下预测多个残差离散词元。这种方法在唇语到语音任务中面临重大挑战，因为有限的视觉输入不足以支持复杂的语言建模任务。

在本研究中，我们提出了 SLD-L2S，这是一种新颖的唇语到语音框架，利用分层子空间潜在扩散模型，直接根据视觉输入生成高保真语音。我们框架的一个关键组成部分是将视觉表示分解为多个子空间，从而实现更稳健的跨模态映射。首先，通过预训练视觉编码器提取视觉表示，随后由我们所提的子空间分解模块进行处理。为有效建模这些分解后的表示，我们引入扩散卷积块（DiCB）作为生成模型的骨干网络。不同于扩散变换器（DiT）架构（Peebles 和 Xie，2023年），我们的 DiCB 利用卷积的归纳偏置来捕获时间依赖关系和跨子空间依赖关系。此外，我们采用重参数化流匹配（Lipman 等人，2022年）目标，以实现对连续潜变量的直接预测。为增强语音的可理解性，我们的训练策略超越了标准的流匹配损失，引入了两个辅助损失：施加于潜在向量上的语义一致性损失，以及在最终合成波形上评估的语音语言模型（SLM）损失。在推理过程中，生成的潜在向量通过预训练神经音频编解码器解码为最终的音频波形。借助这些架构和方法上的创新，SLD-L2S 在基准数据集上取得了当前最优的结果。本研究的主要贡献总结如下：

❑我们提出了一种新颖的唇语到语音框架，通过分层子空间潜在扩散模型，根据视觉输入直接生成预训练音频编解码器的潜在向量。

❑ 我们设计了扩散卷积块、子空间分解模块和重组模块，以有效地在子空间中映射不同模态的表征。

❑我们提出一种用于直接生成潜向量的重参数化流匹配技术，以提升唇语到语音合成的生成质量。

相关工作

多说话人唇语到语音合成

深度学习的最新进展改变了唇语到语音合成，推动该领域从受限场景下的说话人依赖模型，发展到适用性更广的多说话人系统。一项关键进展是引入自监督学习技术，通过利用从海量无标注音视频数据中学习到的强大语义表征，提升了系统性能。例如，近期研究（Hsu 等，2023；Choi、Kim 和 Ro，2023；Lei 等，2024；Liang 等，2025a）有效采用 AV-Hubert 等预训练自监督学习模型作为强大的特征提取器，显著提升了语音可懂度。与此同时，生成式扩散模型已成为主流范式。该领域的早期方法，如 DiffV2S（Choi、Hong 和 Ro，2023）和 LipVoicer（Yemini 等，2023），采用去噪扩散概率模型（DDPM）（Ho、Jain 和 Abbeel，2020），根据视觉特征以及视觉引导说话人嵌入、唇读模型提供的语义信息等辅助输入，合成梅尔频谱图。较新的研究，包括 V2SFlow（Choi 等人，2025）和 From Faces to Voices（Kim 等人，2025），引入了流匹配（Lipman 等人，2022），进一步提升合成质量，并展现了扩散模型在唇语到语音领域的巨大潜力。

除了生成梅尔频谱图或离散自监督学习令牌等中间表征外，Uni-Dubbing（Lei 等，2024年）还引入了 Hifi-Codec，将该问题转化为离散残差词元预测。这与语音合成领域的近期趋势一致，即越来越多地使用音频编解码器来提升语音质量。然而，信息稀疏的视觉模态限制了对离散词元进行高精度类别预测的能力。相比之下，我们提出的 SLD-L2S 框架利用分层子空间潜在扩散模型，根据视觉输入直接生成预训练音频编解码器的潜在向量。

神经音频编解码器

神经音频编解码器是一类为实现高保真音频压缩而设计的生成模型（Zeghidour 等，2021年；D´efossez 等，2022年；Kumar 等，2023年）。从架构上看，它们由一个编码器和一个对应的解码器组成：编码器将连续的原始波形映射为一串离散令牌，解码器则根据这些令牌合成高质量波形。编码器输出通常通过残差向量量化（RVQ）进行量化。该方法对前一量化步骤产生的残差信号进行递归量化，每个阶段使用不同的码本，逐步捕捉更多细节。

除了压缩之外，音频编解码器在生成式语音任务中也发挥着重要作用。通过将预训练编解码器的编码器重新用作声学分词器，语音合成可以转化为条件序列到序列建模问题，类似于自然语言处理（NLP）中的任务。一个典型例子是 VALL-E

（Wang 等人，2023），它根据文本和声学提示自回归地预测离散声学标记。然而，正如前文所述，这种标记预测方法并不适用于唇语到语音合成。受 NaturalSpeech2（Shen 等人，2023）方法的启发，该方法对量化前的连续潜在空间进行建模，我们摒弃了有问题的离散预测任务。相反，我们的方法旨在根据视觉输入直接生成预训练音频编解码器的连续潜向量。

流匹配

流匹配（Lipman 等人，2022）已成为一种强大的生成建模技术，在图像（Esser 等人，2024）、视频（Zheng 等人，2024）和语音生成（Chen 等人，2024）等多个领域广受关注。近期研究如 V2SFLow（Choi 等人，2025）和 From Faces to Voices（Kim 等人，2025）已展示了其在唇语到语音合成中的有效性。具体而言，V2SFLow 通过多种声学属性生成语音，而 From Faces to Voices 则通过分层视觉编码器映射这些属性，从而实现唇语到语音合成。

给定来自目标数据分布的样本 x_1 和来自标准高斯分布的样本 x_0，流匹配的核心目标是学习一个连续向量场 $v_\theta(x_t, t)$，该向量场沿着由 $dx_t = v_\theta(x_t, t)dt$ 定义的常微分方程（ODE）路径，将 x_0 映射到 x_1。一种常见的训练方法采用简单的线性流，其路径定义为

$$x_t = (1-t)x_0 + tx_1, t \in [0,1]. \quad (1)$$

对于这一特定的线性路径，真实速度场可解析地表示为 $v(x_t, t) = \frac{dx_t}{dt} = x_1 - x_0$。我们可以训练一个神经网络 v_θ，使其在时间 $t \in [0,1]$ 沿着从 x_0 到x_1 的路径上的任意连续点 x_t 近似真实速度，方法是最小化以下目标函数：

$$\mathbb{E}_{x_0 \sim p_0, x_1 \sim p_1, t}\left[||v_\theta(x_t, C, t) - (x_1 - x_0)||_2^2\right], \quad (2)$$

其中，C 表示引导生成过程的上下文信息。在我们的具体情形中，C 对应于从输入视频帧中提取的视觉表征。

在推理过程中，为了从目标分布生成新样本，我们首先从源分布中采样得到 x_0。然后，使用学习到的速度函数 $v_\theta(x_t, C, t)$ 迭代更新样本 x_t，直至其近似于目标分布中的 x_1。这一过程本质上是使用数值方法求解常微分方程，通常采用欧拉方法：

$$x_{t+1} = x_t + v_\theta(x_t, C, t)\Delta t, \quad (3)$$

其中，Δt 为步长。通过遵循学习到的速度场，这一迭代过程能够从目标分布生成样本；该速度场有效地引导样本从源分布出发。

所提方法

概述

我们所提唇语到语音框架 SLD-L2S 的整体架构如图 1 所示。我们方法的核心原则是学习从视觉表示到预训练神经音频编解码器连续潜在空间的直接映射。为此，我们采用一种以视觉表示为条件、通过多个并行子空间运作的层次化子空间潜在扩散模型。除传统的去噪损失外，整个框架还通过多层级损失函数进行优化，以确保合成语音兼具声学保真度和语义一致性。

视觉前端

我们框架中的视觉前端负责从输入视频序列中提取语言特征。为此，我们采用 AV-Hubert Large（Shi 等人（2022））的视觉前端；该模型是一种自监督音视模型，预训练所用的数据集为大规模的 LRS3（Afouras、Chung 和 Zisserman，2018）和 VoxCeleb2（Chung、Nagrani 和 Zisserman，2018）。与以往研究一致，我们发现 AV-Hubert 非常有效，因为它提供了富含语音内容的视觉表征，非常适用于唇语到语音合成。

层次子空间潜在流匹配

子空间分解 我们的潜变量流匹配模型的初始阶段是将视觉表征分解为多子空间表示。首先，对从 AV-Hubert 编码器提取的视觉特征进行时间上采样，使其分辨率与目标音频编解码器的潜在空间相匹配。随后，我们不采用单一映射，而是将上采样后的特征输入一组并行的卷积子空间。每个子空间由层归一化和一维卷积组成，旨在学习一种独特的子空间表示。然后，将这些并行子空间表示拼接起来，形成一个统一张量，并将其作为 DiCB 主干网络的输入。遵循扩散模型中的标准做法（Peebles 和 Xie，2023），时间步 t 和说话人嵌入 c_s 通过自适应层归一化（AdaLN）模块进行投影，以生成网络内用于条件调制的缩放和平移参数。

扩散卷积块 从视觉特征到声学特征的映射是一种复杂的跨模态变换，具有精细的多尺度依赖关系。尽管 DiT 等基于 Transformer 的架构在许多生成任务中表现出色，但其单一的自注意力机制可能不利于捕捉这一问题中固有的局部层次模式。为此，我们引入 DiCB，作为生成模型的基本构建模块。DiCB 的设计汲取了 DiT 和 Conv2Former（Hou 等，2024）的灵感，旨在高效处理分解后的子空间表示。

与 DiT 有显著不同的是，我们用专门的卷积模块取代标准的自注意力和多层感知机层，并借鉴了 Conv2Former 的设计原则。这使 DiCB架构天生不受序列长度变化的影响，这一特性对唇语到语音应用十分理想。每个模块的核心是一个卷积注意力模块。在这里，深度卷积

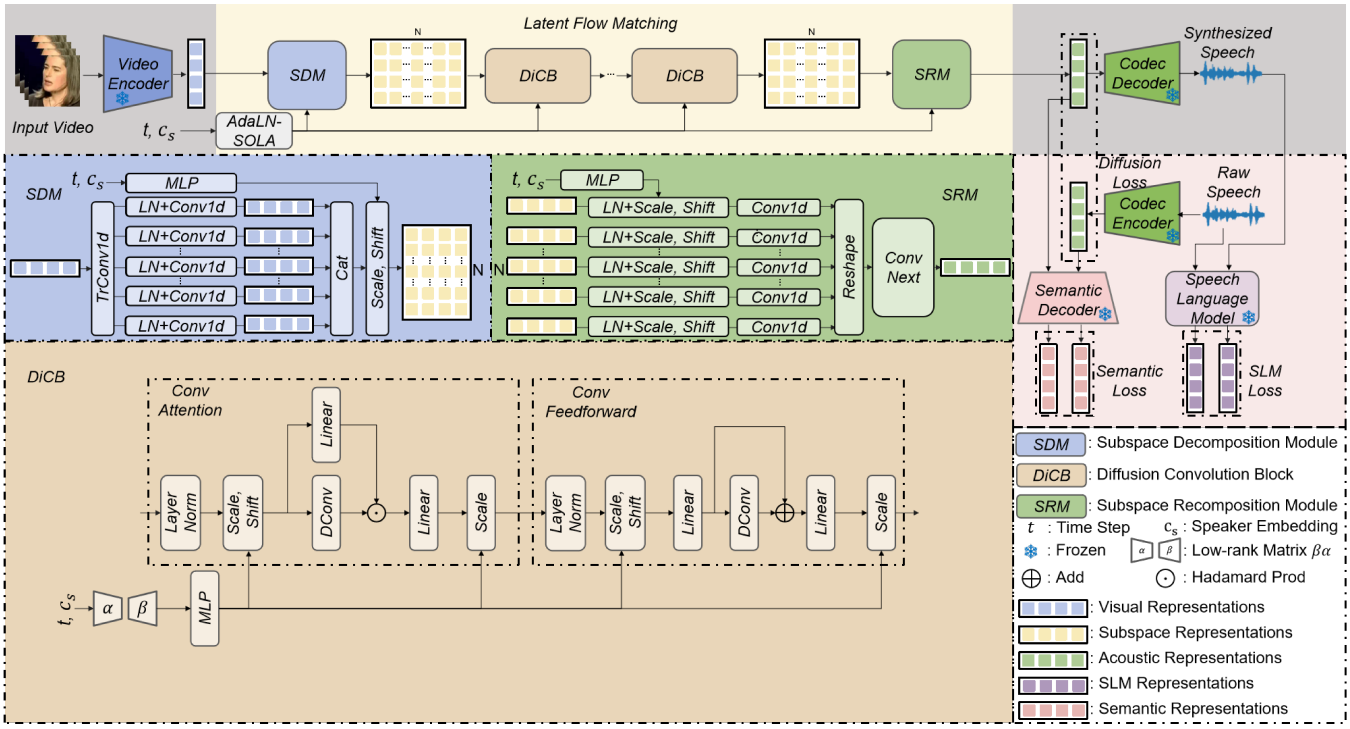

图 1：该模型采用潜变量流匹配方法，将视觉特征直接映射到连续声学潜空间。其架构由三个主要阶段组成：子空间分解模块将视觉特征处理为并行路径；采用自适应层归一化-SOLA 进行条件调制的 DiCB 骨干网络，有效地细化这些表征；子空间重组模块则将它们融合为最终的潜向量。模型采用多目标函数进行训练，包括用于声学重建的流匹配损失、针对潜在空间的语义损失，以及针对最终合成波形的语音语言模型（SLM）损失。推理时，模型生成潜在向量，供编解码器合成为语音。

采用卷积核大小为 $t \times k$（其中 t 和 k 分别是沿时间和子空间方向的卷积核维度）来建模局部特征依赖，并通过与值投影进行 Hadamard 积得到输出。这种设计能够高效地建模局部和跨子空间交互。随后是一个卷积前馈模块，用于促进通道级信息混合，其作用类似于标准 Transformer 模块中的前馈网络。

DiCB 的另一个关键特性是其条件机制。与 DiT 类似，我们使用自适应层归一化模块，将时间步 t 和说话人嵌入向量 c_s 融合起来。具体而言，我们采用 AdaLN-SOLA（Hai 等，2024），通过在所有块和时间步之间共享单个 AdaLN 模块，减少参数化规模和计算开销。此外，它还引入了一种低秩调整机制，根据条件变量调节归一化参数；我们发现，这种机制能够稳定训练并保持性能。

子空间重组 潜变量流匹配模型的最后一个阶段——子空间重组模块，会将 DiCB 主干网络处理后的特征转换为音频编解码器所需的目标潜空间格式。主干网络的输出此时已包含高度集成的特征表示，首先使用 AdaLN 层根据时间步 t 和说话人嵌入 c_s 进行条件化。

为解决 DiCB 输出与音频编解码器潜在向量之间的维度不匹配问题，我们采用一维卷积和重塑操作，将张量投影到正确的输出维度。最后，我们引入 ConvNeXt 块（Liu 等，2022），确保信息在各特征通道间稳健融合。该模块能够实现高阶交互，并使模型学会对多子空间特征进行最优组合。

训练目标

重参数化流匹配损失 标准的流匹配目标通过最小化均方误差（均方误差）损失来估计速度场v_θ，其定义见公式 2。然而，在我们的实践中，这种表述可能不稳定，因为它要求网络回归依赖噪声的随机导数。目标的随机性使得无法使用辅助损失，也让优化更加复杂。受（Fu 等，2025）的启发，我们采用重参数化目标。我们不直接预测速度场 v_θ，而是将模型的输出重新表述为对目标数据点 x_1 的直接估计。这种重参数化不仅提高了训练稳定性，还为引入作用于数据空间的辅助损失提供了更直观的框架。具体而言，我们将模型的输出 $d_\theta(x_t,\ c_s,\ t)$ 定义为对目标数据点的直接预

方法	NFE	LRS3-TED					LRS2-BBC				
		UTMOS↑	SCOREQ↑	D-BERT↑	词错误率↓	SECS↑	UTMOS↑	SCOREQ↑	D-BERT↑	词错误率↓	SECS↑
真实值	–	3.5721	3.8105	–	0.71	–	3.0538	3.3008	–	1.35	–
Intelligible	1	2.7019	2.7024	0.8231	27.73	0.761*	2.3315	2.3437	0.7932	36.50	0.701*
DiffV2S	1000	3.0681	3.3811	0.7408	35.81	0.627	2.9575	3.2721	0.7103	47.58	0.569
LipVoicer	400	2.4540	2.6598	0.7132	**20.62**	0.588	2.3756	2.7973	0.6906	**24.92**	0.578
NaturalL2S	1	3.6598	3.7695	0.8068	30.37	0.628	3.5014	3.6702	0.7838	39.19	0.562
V2SFlow	30	3.6939	4.0710	**0.8306**	27.55	**0.851***	3.4556	3.9157	**0.8064**	34.17	**0.820***
所提方法	10	**4.2096**	**4.6075**	0.8284	30.22	0.804*	**4.1664**	**4.5706**	0.8005	39.54	0.756*

表 1：LRS3-TED 和 LRS2-BBC 数据集上的唇语到语音合成性能比较。↑：越高越好，↓：越低越好。* 表示使用参考说话人嵌入的方法。

测 x_1。根据公式 1 中定义的线性概率路径，该预测与速度场 v_θ 之间的关系为：

$$d_\theta(x_t, c_s, t) = x_t + (1-t)v_\theta(x_t, c_s, t). \tag{4}$$

因此，速度场 v_θ 是通过隐式方式学习的。随后，将重参数化流匹配损失 $\mathcal{L}_{FM}$表示为模型预测值与目标 x_1之间的加权均方误差：

$$\mathcal{L}_{FM} = \mathbb{E}_{x_0,x_1,t}\left[\frac{||d_\theta(x_t, c_s, t) - x_1||_2^2}{(1-t)^2}\right], \tag{5}$$

其中 $x_0 \sim p_0$、 $x_1 \sim p_1$和 $t \sim U$∈(0，1)。

辅助损失为提升合成音频的感知质量和内容准确性，我们首先引入基于预训练SLM的 SLM 损失。该 SLM 损失通过在高层感知特征空间中强制保持一致性，促进知识迁移。具体而言，我们使用音频编解码器的解码器 $D_a(\cdot)$ 将预测的潜向量 $d_\theta(x_t, c_s, t)$ 和目标潜向量 x_1 解码为波形。随后，使用SLM从两段波形中提取特征表示，并计算它们之间的平方L2距离：

$$\begin{aligned}\mathcal{L}_{SLM} =& \mathbb{E}_{x_0,x_1,t}\left[\|SLM\left(D_a\left(d_\theta(x_t, c_s, t)\right)\right) - \right. \\ & \left. SLM\left(D_a(x_1)\right)\|_2^2\right],\end{aligned} \tag{6}$$

与此同时，受近期关于 X-codec（Ye 等，2025）等具备语义感知能力的音频编解码器的研究启发，我们引入语义损失，直接在潜变量层面强制保持内容一致性。该损失确保生成的编解码器潜变量保留与目标潜变量相同的核心语义内容。为此，我们使用单独预训练的语义解码器 $D_s(\cdot)$；该解码器通过自编码器架构（配有相应的语义编码器）进行训练，以便从编解码器潜变量重建语义特征（例如 HuBERT（Hsu 等，2021）或 WavLM（Chen 等人（2022））的特征）。语义损失是预测潜变量与目标潜变量重建出的语义特征之间的平方L2距离：

$$\mathcal{L}_{sem} = \mathbb{E}_{x_0,x_1,t}\left[||D_s(d_\theta(x_t, c_s, t)) - D_s(x_1)||_2^2\right], \tag{7}$$

其中， $D_s(\cdot)$ 是预训练的语义解码器，该解码器由多层卷积网络实现。

我们的 SLD-L2S 框架最终的训练目标是以下多目标损失的加权组合：

$$\mathcal{L} = \mathcal{L}_{FM} + \lambda_1\mathcal{L}_{SLM} + \lambda_2\mathcal{L}_{sem}, \tag{8}$$

其中， λ_1 和 λ_2 是用于平衡各辅助目标贡献的超参数。这种多目标损失结构使模型能够同时优化感知质量和语义内容，从而提升整体合成性能。

实验设置

数据集

我们的实验在多说话人音视频基准测试上进行：LRS3-TED（Afouras、Chung 和 Zisserman，2018年）和 LRS2-BBC（Chung 等，2017年）。LRS3-TED 数据集源自 TED 和 TEDx 演讲，包含约 439 小时的片段，涵盖 5,000 多位说话人的 150,000 条话语。LRS2-BBC 数据集由英国广播公司的广播节目中 200 小时的数据组成，包含来自不同说话人的 144,000 个片段。我们仅使用 LRS3-TED 语料库进行训练，并遵循其官方预定义的训练、验证和测试划分。为评估各方法的泛化能力，我们还在 LRS2-BBC 测试集上进行评估。

实现细节

我们的数据预处理遵循 Shi 等人（2022）中的流程。对于每段视频，我们使用 dlib（King（2009））库提取面部关键点，裁剪出嘴部感兴趣区域（ROI），将其调整为 88x88 分辨率并转换为灰度图。随后，将这些帧输入预训练的 AV-Hubert Large 模型处理，该模型作为我们的视觉编码器，可生成 1024 维特征向量。对于音频合成，我们采用 X-Codec-hubert 作为神经音频编解码器，其采样率为 16 kHz，并在 LibriSpeech 语料库上进行了预训练。说话人身份由一个 256 维嵌入向量 c_s 提供，该向量通过 GE2E（Wan 等人（2018））从参考话语中提取。对于辅助损失，我们使用预训练的 HuBERT-base-ls960 模型提取语义特征，并使用 WavLM（Chen 等人（2022））计算 SLM 损失。为了

方法	自然度↑	Intelligibility↑	相似度↑
真实值	4.36 ±0.15	4.48 ±0.15	–
Intelligible	2.82 ±0.25	3.37 ±0.26	3.02 ±0.17
DiffV2S	2.95 ±0.27	2.98 ±0.27	2.80 ±0.23
LipVoicer	2.51 ±0.23	2.97 ±0.29	2.59 ±0.23
NaturalL2S	3.54 ±0.32	3.43 ±0.27	3.21 ±0.21
V2SFLow	3.88 ±0.25	3.69 ±0.25	3.90 ±0.14
所提方法	4.17 ±0.24	3.65 ±0.25	3.77 ±0.15

表2：LRS3 数据集上的 MOS 结果。

方法	UTMOS↑	SCOREQ↑	D-BERT↑	词错误率↓	SECS↑
Ours	4.2096	4.6075	0.8284	30.22	0.804
不含 DiCB	4.0835	4.5056	0.8184	30.70	0.618
不使用 Repara	3.9540	4.2600	0.8249	29.70	0.809
不使用 L_{sem}	4.1859	4.5830	0.8272	30.21	0.795
不使用 L_{SLM}	4.2885	4.6958	0.8207	29.99	0.776
不使用 L_{sem}	4.2507	4.6494	0.8157	30.53	0.771

表3：LRS3 数据集上的消融研究结果。Repara 表示采用了我们提出的重参数化流匹配目标。DiCB 表示将我们提出的 DiCB 主干网络替换为 DiT，同时使参数量保持接近（DiCB：57.7M，DiT：57.8M）。

弥合视频与音频之间的时间分辨率不匹配，我们使用转置卷积层对视觉特征进行上采样。

我们的 SLD-L2S 模型架构配置如下。子空间分解模块将 1024 维视觉特征划分为 8 个子空间。每个子空间包含一个层归一化层和一个 1D 卷积层，在特征拼接前对其进行处理。骨干网络由 12 个 DiCB 组成。在每个 DiCB 中，卷积注意力模块分别在时间维度和子空间维度上采用大小为 (5, 7) 的卷积核。卷积前馈模块使用大小为 3 的卷积核，并将通道维度扩展至 1024。条件输入通过 AdaLN-SOLA 模块进行整合，其低秩调整矩阵的秩设为 32。最后，子空间重组模块将 8 个子空间流各自的输出投影到 128 维，然后将它们拼接并融合为一个 1024 维向量。该融合表示经过 ConvNeXt 模块处理；这些模块由 3 层组成，包含 512 个隐藏通道，以便在生成输出前实现最终的跨通道交互。训练时，我们使用 AdamW 优化器，初始学习率设为 2e-4，并采用余弦学习率衰减策略。批次大小设为 16，模型在单块 NVIDIA H100 GPU 上训练 150K 次迭代，耗时约五天。AdamW 方案配置为 $\beta_1 = 0.8, \beta_2 = 0.99$辅助损失的超参数设为 $\lambda_1 = 1$ 和 $\lambda_2 = 100$。

评估指标

我们使用一组客观指标和主观指标，对所提 SLD-L2S 框架进行了全面评估。评估旨在考察合成语音的三个关键方面：感知质量、内容可懂度和说话人相似度。对于客观评估，我们采用以下指标：

❑ **质量**：我们使用 UTMOS（Saeki 等人，2022年）和 SCOREQ（Ragano、Skoglund 和 Hines，2024年）评估感知质量。二者均为非侵入式、无需参考语音的模型，可在 1 至 5 分的范围内预测平均意见分（MOS），旨在与人类对质量的判断相关。

❑ **可懂度**：我们从两个角度衡量可懂度。首先，使用 AUTO-AVSR（Ma 等人，2023）计算词错误率（WER）。其次，我们采用 SpeechBERTScore（D-BERT）（Saeki 等人，2024），该指标利用自然语言处理评估指标，计算合成语音与其自监督学习嵌入之间的余弦相似度

以及目标语音，提供范围为 –1 至 1 的语义相似度分数。

❑ **说话人相似度**：我们使用说话人嵌入余弦相似度（SECS）量化说话人身份保持情况。该指标根据合成语音和目标语音的说话人嵌入计算得出，这些嵌入使用 Resemblyzer（Wan 等人（2018））库提取。

主观评估通过平均意见分测试进行。15名参与者聆听了30段合成语音样本，并被要求从三个标准对每段音频进行评分：自然度（语音听起来有多自然）、可懂度（内容有多容易理解）和说话人相似度（合成语音与目标说话人有多相似）。所有标准均采用5分制（从1分到5分）评分。

结果与讨论

与最先进方法的比较

我们将所提 SLD-L2S 框架与先进的唇语到语音方法进行比较：Intelligible（Choi、Kim 和 Ro，2023年）、DiffV2S（Choi、Hong 和 Ro，2023年）、LipVoicer（Yemini 等，2023年）、NaturalL2S（Liang 等，2025a）和 V2SFlow（Choi 等人，2025年）。LRS3 和 LRS2 数据集上的结果见表1。我们的分析重点关注感知质量、效率、可懂度和说话人相似度等关键方面。值得注意的是，训练过程中未使用 LRS2 数据集，以评估鲁棒性。

首先，结果明确确立了 SLD-L2S 在感知质量方面新的业界最佳水平。如表 1 所示，在无参考质量指标上，我们的模型在两个数据集上均显著优于所有基线模型。以主要的 LRS3-TED 基准测试为例，SLD-L2S 的 UTMOS 比最强的基于流匹配的竞争方法 V2SFlow 高出 0.5157。更为重要的是，这种卓越的合成质量是以极高的计算效率实现的。SLD-L2S 在推理时仅需 10 次函数评估（NFE），其效率高于其他基于扩散的方法。

在可懂度方面，LipVoicer 的词错误率最低，这是因为它依赖强大的预训练视觉语音识别（VSR）模型（Ma 等，2023）来提取语义特征。然而，这也导致语音质量下降，其 UTMOS 和 SCOREQ 得分较低便体现了这一点。此外，该方法还需要使用 VSR 模型

卷积核大小	UTMOS↑	SCOREQ↑	D-BERT↑	词错误率↓	SECS↑
所提方法（5×7）	4.2096	4.6075	0.8284	30.22	0.804
3×5	4.2120	4.6016	0.8292	30.34	0.801
7×9	4.1914	4.5733	0.8291	30.17	0.800
9×11	4.2008	4.5852	0.8288	30.10	0.803

表4：LRS3数据集上不同卷积核大小的消融研究结果。卷积核大小表示为（时间 × 子空间）。

子空间	UTMOS↑	SCOREQ↑	D-BERT↑	词错误率↓	SECS↑
本文（8）	4.2096	4.6075	0.8284	30.22	0.804
4	4.1797	4.5638	0.8278	29.94	0.802
16	4.1752	4.5683	0.8295	29.92	0.802
32	4.1288	4.5219	0.8268	30.35	0.799

表5：LRS3 数据集上不同子空间数量的消融研究结果。

这限制了它在无法获取文本转录的场景中的适用性。相比之下，SLD-L2S 实现了更出色的平衡。它生成的语音具有良好的可懂度，D-BERT 得分具有竞争力，优于几乎所有不使用专用视觉语音识别前端的方法，并且与表现最佳的 V2SFlow 相比也很有竞争力。这表明，我们的模型能够有效生成可懂的语音，同时保持其感知质量。

在说话人身份保持方面，V2SFlow取得了最高的SECS分数，因为其训练目标明确包含说话人嵌入损失，以强化说话人相似度。尽管未采用此类直接损失，我们的框架仍取得了极具竞争力的第二高SECS分数。这一结果凸显了我们所提DiCB模块的有效性：无需直接监督，即可在生成过程中根据说话人身份进行条件控制。

表2所示的主观评估结果进一步验证了我们框架的有效性。我们的模型在自然度方面表现出色，取得了4.17的优异分数，显著超过所有竞争方法，并接近真实值。在可懂度方面，我们的方法以3.65分排名第二。在说话人相似度方面，我们的模型取得了极具竞争力的3.77分，几乎与表现最佳的V2SFlow相当。这项评估还凸显了仅依赖客观指标的局限性。例如，LipVoicer在客观词错误率指标上表现良好，但在主观测试中得分较低，表明其较差的声学质量和不自然性会严重降低人类听众实际感知的可懂度。这种差异表明，感知质量是唇读语音合成系统的关键目标，因为它直接影响人类能否舒适地理解生成的语音。

消融研究

为分析各组件的贡献，我们进行了消融研究，结果见表3。

❑ **DiCB架构：**将我们的 DiCB 主干网络替换为参数量匹配的 DiT 后，几乎所有指标的性能均显著下降。这一结果凸显了我们的 DiCB 设计在实现高质量合成与有效说话人条件控制方面的关键作用。

❑**重参数化训练目标：**恢复使用标准的速度预测流匹配目标，会导致感知质量（UTMOS、SCOREQ）显著下降。这证实了我们的重参数化方法对于提升训练稳定性和合成质量至关重要。

❑ **辅助损失：** 移除语义损失（不含$\mathcal{L}_{sem}$）会导致质量和说话人相似度略有下降，表明它有助于对潜在空间进行正则化。更值得注意的是，移除 SLM 损失（不含 $\mathcal{L}_{SLM}$）揭示了一种权衡：虽然声学质量指标（UTMOS/SCOREQ）略有改善，但可懂度（D-BERT）和说话人相似度（SECS）均有所下降。这表明，即使 SLM 损失会对声学非侵入式指标造成轻微限制，它对于内容和身份保留仍至关重要。两种损失都移除后，性能下降更加明显，证实了它们共同发挥的重要作用。

DiCB 设计选择分析

我们进一步对 DiCB 模块的两个关键设计选择进行了消融实验：卷积核大小和子空间数量。结果总结如下。

❑ **卷积核大小**。如表4所示，DiCB 中卷积注意力模块的模型性能对卷积核大小表现出极强的鲁棒性。不同于计算机视觉中采用大卷积核的趋势，在我们的任务中，较小的卷积核同样具有竞争力。 5×7 卷积核在所有指标之间实现了最佳平衡，验证了我们的选择。

❑ **子空间数量**。事实证明，这是一个更为敏感的超参数。表5显示，在我们选择的 8 个子空间配置下，性能达到明显峰值。使用更少的子空间（4 个）或更多的子空间（16 个、32 个）都会导致感知质量（UTMOS/SCOREQ）显著下降。这表明，8 个子空间实现了最佳权衡，能够有效分解视觉特征，同时避免形成过度碎片化、难以整合的表征。

结论

本文提出了 SLD-L2S，这是一种新颖的框架，通过将视觉特征直接映射到音频编解码器的潜在空间，在唇语到语音合成领域达到了最先进的水平。我们的方法采用高效的重参数化流匹配技术来训练分层子空间网络。该网络基于我们提出的扩散卷积块构建，能够有效捕捉视觉和听觉模态之间复杂的关系。结合用于提升内容保真度的双重语义损失与 SLM 损失策略，SLD-L2S 在感知质量方面显著优于现有方法。我们的工作验证了潜在扩散在唇语到语音合成中的潜力，并为未来的高保真唇语到语音合成提供了可靠的蓝图。

参考文献

Afouras, T.; Chung, J. S.; and Zisserman, A. 2018年。LRS3-TED：用于视觉语音识别的大规模数据集。*arXiv* 预印本 *arXiv:1809.00496*。

Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; 等. 2022年. Wavlm：用于全栈语音处理的大规模自监督预训练。*IEEE* 信号处理专题期刊，16(6)：1505–1518。

Chen, Y.; Niu, Z.; Ma, Z.; Deng, K.; Wang, C.; Zhao, J.; Yu, K.; and Chen, X. 2024年. F5-tts：一位能通过流匹配合成流畅且忠实语音的童话讲述者。*arXiv*预印本 *arXiv:2410.06885*。

Choi, J.; Hong, J.; and Ro, Y. M. 2023年. DiffV2S：基于扩散的视觉引导说话人嵌入视频到语音合成。载于 *IEEE/CVF* 计算机视觉国际会议论文集，7812–7821。

Choi, J.; Kim, J.-H.; Li, J.; Chung, J. S.; 和 Liu, S. 2025年。V2sflow：结合语音分解与整流流的视频转语音生成。载于*ICASSP 2025-2025 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），1–5。IEEE。

Choi, J.; Kim, M.; 和 Ro, Y. M. 2023年。基于语音单元的可理解唇语转语音合成。载于*INTERSPEECH2023*，4349–4353。国际语音通信协会。

Chung, J. S.; Nagrani, A.; 和 Zisserman, A. 2018年。Voxceleb2：深度说话人识别。*arXiv*预印本 *arXiv:1806.05622*。

Chung, J. S.; Senior, A.; Vinyals, O.; 和 Zisserman, A. 2017年。野外场景中的句子唇读。载于*IEEE*计算机视觉与模式识别会议。

Cooke, M.; Barker, J.; Cunningham, S.; 和 Shao, X. 2006. 用于语音感知和自动语音识别的视听语料库。美国声学学会期刊，120(5): 2421–2424。

D´efossez,A.; Copet, J.; Synnaeve, G.; 和 Adi, Y. 2022年。高保真神经音频压缩。*arXiv* 预印本 *arXiv:2210.13438*。

Du, Z.; Chen, Q.; Zhang,S.; Hu, K.; Lu, H.; Yang, Y.; Hu, H.; Zheng, S.; Gu, Y.; Ma, Z.; 等。2024年。Cosyvoice：一种基于监督语义令牌的可扩展多语言零样本文本转语音合成器。*arXiv* 预印本 *arXiv:2407.05407*。

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; M¨uller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; 等。2024年。扩展整流流 Transformer，以实现高分辨率图像合成。载于第四十一届国际机器学习大会。

Fu, Y.; Yan, Q.; Wang, L.; Li, K.; and Liao, R. 2025年。Moflow：基于隐式最大似然估计的蒸馏方法，用于人体轨迹预测的单步流匹配。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，17282–17293。

Hai, J.; Xu, Y.; Zhang, H.; Li, C.; Wang, H.; Elhilali, M.; and Yu, D. 2024年。Ezaudio：利用高效扩散变换器增强文本到音频生成。*arXiv*预印本 *arXiv:2409.10819*。Harte, N.; and Gillen, E. 2015年。TCD-TIMIT：连续语音的音频-视觉语料库。*IEEE*多媒体汇刊，17(5)：603–615。Ho, J.; Jain, A.; and Abbeel, P. 2020年。去噪扩散概率模型。神经信息处理系统进展，33：6840–6851。Hou, Q.; Lu, C.-Z.; Cheng, M.-M.; and Feng, J. 2024年。Conv2former：一种用于视觉识别的简易变换器风格卷积网络。*IEEE*模式分析与机器智能汇刊，46(12)：8274–8283。Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021年。HuBERT：通过对隐藏单元进行掩码预测实现自监督语音表征学习。*IEEE/ACM*音频、语音与语言处理汇刊，29：3451–3460。Hsu, W.-N.; Remez, T.; Shi, B.; Donley, J.; and Adi, Y. 2023年。ReVISE：利用视觉输入进行通用化和泛化语音再生的自监督语音重合成。载于*IEEE/CVF*计算机视觉与模式识别会议，18795–18805。Kim, J.-H.; Choi, J.; Kim, J.; Jung, C.; and Chung, J. S. 2025年。从面孔到声音：学习用于高质量视频转语音的分层表征。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，15874–15884。Kim, J.-H.; Kim, J.; and Chung, J. S. 2024年。让声音响起：从静音视频重建高质量语音。载于人工智能促进协会会议论文集，第38卷，2759–2767。King, D. E. 2009年。Dlib-ml：一种机器学习工具包。机器学习研究杂志，10：1755–1758。Kumar, R.; Seetharaman, P.; Luebs, A.; Kumar, I.; and Kumar, K. 2023年。通过改进的RVQGAN实现高保真音频压缩。神经信息处理系统进展，36：27980–27993。

Lei, S.; Cheng, X.; Lyu, M.; Hu, J.; Tan, J.; Liu, R.; Xiong, L.; Jin, T.; Li, X.; and Zhao, Z. 2024年。Uni-dubbing：基于视觉发音动作的零样本语音合成。载于计算语言学协会第*62*届年会论文集（第*1*卷：长文），10082–10099。

Liang, Y.; Liu, F.; Li, A.; Li, X.; 和 Zheng, C. 2025a. NaturalL2S：采用差分数字信号处理的端到端高质量多说话人唇语到语音合成。*arXiv*预印本 *arXiv:2502.12002*。Liang, Y.; Yang, K.; Liu, F.; Li, A.; Li, X.; 和 Zheng, C. 2025b. LightL2S：面向多说话人场景的超低复杂度唇语到语音合成。载于*Interspeech 2025*论文集，3783–3787。Lipman, Y.; Chen, R. T.; Ben-Hamu, H.; Nickel, M.; 和 Le, M. 2022. 用于生成建模的流匹配。*arXiv*预印本 *arXiv:2210.02747*。

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; 和 Xie, S. 2022. 面向2020年代的卷积网络。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，11976–11986。Ma, P.; Haliassos, A.; Fernandez-Lopez, A.; Chen, H.; Petridis, S.; 和 Pantic, M. 2023. AUTO-AVSR：采用自动标签的视听语音识别。载于*ICASSP 2023- 2023*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），1–5。IEEE。Milner, B.; 和 Le Cornu, T. 2015年。从视觉语音特征重建清晰可懂的音频语音。载于*INTERSPEECH 2015*。Peebles, W.; 和 Xie, S. 2023. 采用Transformer的可扩展扩散模型。载于*IEEE/CVF*国际计算机视觉会议论文集，4195–4205。Polyak, A.; Adi, Y.; Copet, J.; Kharitonov, E.; Lakhotia, K.; Hsu, W.-N.; Mohamed, A.; 和 Dupoux, E. 2021. 基于离散解耦自监督表征的语音重合成。*arXiv*预印本 *arXiv:2104.00355*。Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V. P.; 和 Jawahar, C. 2020. 学习个体说话风格，实现准确的唇动到语音合成。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，13796–13805。Ragano, A.; Skoglund, J.; 和 Hines, A. 2024. SCOREQ：基于对比回归的语音质量评估。神经信息处理系统进展，37：105702–105729。Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; 和 Liu, T.-Y. 2020年。FastSpeech 2：快速、高质量的端到端文本转语音。*arXiv*预印本 *arXiv:2006.04558*。Saeki, T.; Maiti, S.; Takamichi, S.; Watanabe, S.; 和 S aruwatari, H. 2024. SpeechBERTScore：利用NLP评估指标进行参考感知的语音生成自动评估。*arXiv*预印本 *arXiv:2401.16812*。Saeki, T.; Xin, D.; Nakata, W.; Koriyama, T.; Takamichi, S.; 和 Saruwatari, H. 2022. UTMOS：Utokyo-Sarulab的VoiceMOS Challenge 2022系统。*arXiv*预印本 *arXiv:2204.02152*。Shen, K.; Ju, Z.; Tan, X.; Liu, Y.; Leng, Y.; He, L.; Qin, T.; Zhao, S.; 和 Bian, J. 2023. NaturalSpeech 2：潜在扩散模型实现自然的零样本语音与歌声合成。*arXiv*预印本 *arXiv:2304.09116*。Shi, B.; Hsu, W.-N.; Lakhotia, K.; 和 Mohamed, A. 2022. 通过掩码多模态聚类预测学习视听语音表征。*arXiv*预印本 *arXiv:2201.02184*。Wan, L.; Wang, Q.; Papir, A.; 和 Moreno, I. L. 2018年。用于说话人验证的广义端到端损失。载于*2018*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），4879–4883。IEEE。Wang, C.; Chen, S.; Wu, Y.; Zhang, Z.; Zhou, L.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; 等. 2023. 神经编解码器语言模型是零样本文本转语音合成器。*arXiv*预印本 *arXiv:2301.02111*。Wang, Y.; 和 Zhao, Z. 2022年。FastLTS：非自回归端到端无约束唇语到语音合成。载于 第*30*届计算机协会国际多媒体会议论文集，MM '22，5678–5687。计算机协会。国际标准书号 978-1-4503-9203-7。Yang, D.; Liu, S.; Huang, R.; Tian, J.; Weng, C.; 和 Zou, Y. 2023年。Hifi-Codec：用于高保真音频编解码器的组残差向量量化。*arXiv*预印本 *arXiv:2305.02765*。Ye, Z.; Sun, P.; Lei, J.; Lin, H.; Tan, X.; Dai, Z.; Kong, Q.; Chen, J.; Pan, J.; Liu, Q.; 等。2025年。编解码器确实重要：探索编解码器对音频语言模型的语义缺陷。载于 *AAAI*人工智能会议论文集，第39卷，25697–25705。Yemini, Y.; Sha msian, A.; Bracha, L.; Gannot, S.; 和 Fetaya, E. 2023年。LipVoicer：利用唇读引导从静音视频生成语音。*arXiv*预印本 *arXiv:2306.03258*。Zeghidour, N.; Luebs, A.; Omran, A.; Skoglund, J.; 和 Tagliasacchi, M. 2021年。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，30：495–507。Zheng, Z.; Peng, X.; Yang, T.; Shen, C.; Li, S.; Liu, H.; Zhou, Y.; Li, T.; 和 You, Y. 2024年。Open-sora：让所有人都能高效进行视频制作。*arXiv*预印本 *arXiv:2412.20404*。