arXiv:2603.08126v1 [cs.CV] 9 Mar 2026

Foley-Flow：基于掩码视听对齐与动态条件流的协同视频到音频生成

Shentong Mo CMU / MBZUAI 达摩院，阿里巴巴集团shentongmo@gmail.com

Yibing Song*达摩院，阿里巴巴集团 湖畔实验室

songyibing.syb@alibaba-inc.com

摘要

基于视频输入的协同音频生成通常需要严格的音视频（*AV*）对齐，即生成音频片段的语义与节奏均应与视频帧中的语义和节奏相对应。以往研究采用两阶段设计：首先通过对比学习对齐音视频编码器，再由编码后的视频表征引导音频生成过程。我们观察到，对比学习与全局视频引导虽能有效对齐整体视听语义，却限制了时间维度上的节奏同步。在本工作中，我们提出 Foley-Flow，首先通过掩码建模训练对齐单模态音视频编码器，使掩码音频片段在相应视频片段的引导下得以恢复。训练之后，这些仅使用单模态数据分别预训练的音视频编码器实现了语义与节奏的一致性对齐。随后，我们为最终的音频生成设计了动态条件流。该方法建立在高效的速度流生成框架之上，动态条件流以随时间变化的视频特征作为动态条件，引导相应音频片段的生成。为此，我们在掩码视听对齐过程中提取连贯的语义与节奏表征，并利用视频片段的这一表征在时间上引导音频生成。我们的音频结果在标准基准测试上进行了评估，并在多项指标上大幅超越现有结果。这一优异表现表明，Foley-Flow 能够有效生成在语义和节奏上均与各类视频序列协调一致的协同音频。

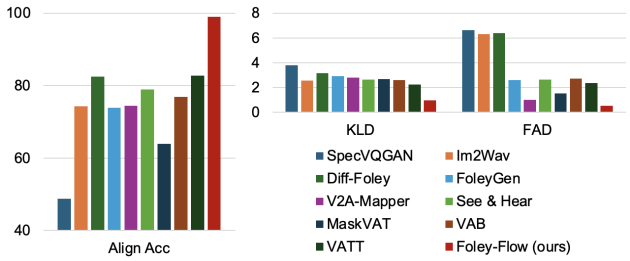

图 1. 与最先进方法的比较。我们在语义指标（AlignAcc）和节奏指标（KLD、FAD）方面取得了最佳结果。

1. 引言

视频序列通常是在伴有音频信号的情况下自然采集的。音频反映了视频中物体类型及其运动所产生的声音 [21–23, 25, 27, 29, 30, 32]。对于缺少伴随音频的视频，协同音频生成可为视频流补充音频，从而形成完整的人类感知。根据视频输入生成协调的音频，通常需要从两个角度考虑。首先，音频应反映视频内容独特的语义（例如，鸟鸣声不同于狗吠声）。其次，音频中的声音应与视频序列中的物体运动相协调（例如，马蹄哒哒声对应于马在路上行走。只有马蹄触地时才会发出哒哒声）。因此，生成音频片段的语义和节奏应与视频内容相对应。

现有的视频转音频生成方法 [10, 15, 17, 46]可概括为两步设计。第一步，将视觉编码器（例如ViT）与音频编码器（即用于处理音频频谱图的 ViT）对齐。视觉/音频编码器仅使用单模态数据进行预训练，采用自监督或监督预训练方案。

*Y. Song 是通讯作者。

对于视听编码器对齐，通常会引入视听数据对上的对比学习，以构建统一的视听表征。第二步，使用音频生成框架，将初始音频噪声逐步细化为最终的音频信号。在这一过程中，视频表征作为音频生成的条件输入。由于视频表征来自第一步中与音频对齐的视觉编码器，因此视频内容条件能够有效引导音频生成。

用于视听对齐的对比学习 [18, 24, 26, 28, 31, 34, 35, 42] 和视频引导的音频生成 [36, 39, 54, 55], 虽然有助于实现全局视听语义对齐，却限制了局部节奏对应方面的性能。这是因为对比学习将视频和音频视为一个整体，而每个视听数据对中的时间片段并未加以区分。此外，视频数据通常作为一个完整样本用于引导音频生成，却没有明确指定每个音频片段中的声音表现（即，节奏和振幅）。实际上，协调的视频到音频生成应同时捕捉高层语义线索（例如，与视觉内容相关联的声音类型）和与视频中事件时序相对应的低层节奏模式。这些挑战限制了现有方法生成自然且同步的音频；处理视听数据对时，应逐个处理其中的片段，而非将其作为一个整体。

在本文中，我们提出*Foley-Flow*来应对这些挑战，旨在推进对齐和生成过程中时间片段级别的视听对应关系。给定一个视觉编码器和一个音频编码器，我们提出了一种掩码视听对齐方案：训练音频编码器，根据时间顺序排列的视频片段恢复被掩码的音频片段。我们的掩码视听对齐使表示同时包含语义信息和同步信息，这为最终的音频生成奠定了基础。在下一步中，我们提出了一种建立在基于速度流的生成框架之上的动态条件流。我们的条件可以根据随时间变化的视频片段进行动态更新。为此，通过第一步与音频表示对齐的视频表征中的语义和节奏信息，将被充分用于基于流的音频生成。随时间变化的视频特征通过我们循环式基于流的模型的高效推理，逐步引导相应音频片段的生成。我们在 VGGSound 数据集上验证了我们的 *Foley-Flow* 。该数据集全面收集了包含多样视听内容的视频。实验结果表明， *Foley-Flow* 在包括弗雷歇音频距离（FAD）、库尔贝克-莱布勒（KL）散度和对齐准确率在内的关键指标上均实现了最先进性能，这些指标凸显了我们的模型生成同步且符合上下文的音频输出的能力，为视频到音频生成树立了新的基准。

我们总结如下贡献：

•我们提出了 *Foley-Flow* ，并采用音视频掩码对齐，以增强音视数据对中语义和节奏的对应关系。

• 我们提出了一种动态条件流，根据随时间变化的视频表征逐步生成最终音频结果。

• 我们通过广泛的评估，证明了所提方法在多项音频质量指标上达到最先进性能的有效性。

2. 相关工作

2.1. 视频到音频生成

视频到音频生成旨在将视觉信息转换为相应的音频输出，近年来，创新的模型架构推动了该领域的快速发展。SpecVQGAN [10] 采用向量量化生成对抗网络，将视觉特征转换为音频频谱图。Im2Wav [46]利用 CLIP 嵌入，直接根据图像生成音频波形。Diff-Foley [17] 使用对比式音频-视觉预训练（CAVP）对齐视频和音频嵌入，并通过扩散模型进一步优化结果。FoleyGen [19]采用语言建模方法，并结合神经音频编解码器进行波形到令牌转换。

近期研究探索了多模态预训练和微调策略。Seeing & Hearing [53] 利用预训练的 ImageBind 模型 [3]作为潜在对齐器，用于跨模态扩散生成。VAB [48] 提出了一种视觉条件化的掩码音频令牌预测任务，用于无需扩散模型的预训练。Mask-VAT [41] 将序列到序列掩码生成模型与高质量音频编解码器相结合，以实现时间同步性。VATT [16]将 Gemma-2B [50]和 LLama-2-7B [51] 等大型语言模型与投影层相结合，将视频特征映射为音频令牌。尽管这些方法解决了语义对齐问题，却往往忽视节奏同步，导致生成的音频不同步。我们的方法通过视频-音频掩码对齐和通用视频音频流，显式建模语义与节奏连贯性，从而弥补这一不足。

2.2. 音视掩码自编码器

掩码建模已被证明有助于学习音频-视觉表示，先前的研究提出了多种流程 [4, 8]。例如，CAV-MAE [4] 将对比学习与掩码建模相结合，以融合跨

模态信息，并通过联合编码器进行处理。MAViL [8]通过引入模态内和模态间对比学习以及掩码重建来扩展这一方法，从而捕获上下文化表示。MBT [40] 通过瓶颈潜变量强制实现跨模态信息共享，以提升表征学习效果。

与这些方法不同，我们的框架直接利用掩码建模进行视频到音频生成。通过掩码音频片段，并要求模型依据视频特征进行重建，我们的视频-音频掩码对齐模块能够学习同步模式和实现精确的时间对齐。这种设计不仅着眼于表征学习，还解决了视频到音频任务中的生成质量和节奏连贯性挑战。我们的 *Foley-Flow* 通过将扩散模型、视频到音频生成和掩码自编码器方面的创新整合到统一框架中，推动了该领域的最新进展。与现有方法不同，所提出的 *Foley-Flow* 通过视频-音频掩码对齐模块实现语义和节奏上的对齐，并利用广义视频音频流模块提高生成效率，为视频到音频生成任务树立了新的基准。

2.3. 扩散模型

扩散模型已成为多个领域中执行生成任务的强大范式。去噪扩散概率模型（DDPM）的奠基性工作 [6, 47] 提出了一种框架：在前向过程中，逐步向数据添加高斯噪声，并训练模型逆转这一过程，从而恢复原始数据。这些模型在图像生成 [33, 44], 、图像恢复 [43], 、语音合成 [13],、点云 [20, 37, 38], 和视频生成 [7]等应用中取得了显著成功。在此基础上，我们的工作将基于流的目标整合到视频到音频生成流水线中，以加速推理并提升输出质量，解决传统基于扩散的方法在效率和可扩展性方面的局限。

3. 提出的方法

本节介绍我们提出的协同视频到音频生成框架，即 *Foley-Flow*。该框架通过利用掩码音视频建模和广义流，实现语义与节奏连贯性。我们首先在第 3.1 节介绍预备知识，随后在第 3.2 节介绍视频-音频掩码对齐，以学习语义与节奏连贯性，最后在第 3.3 节介绍广义视频-音频流，以加速视频-音频去噪过程，实现快速的视频到音频生成。

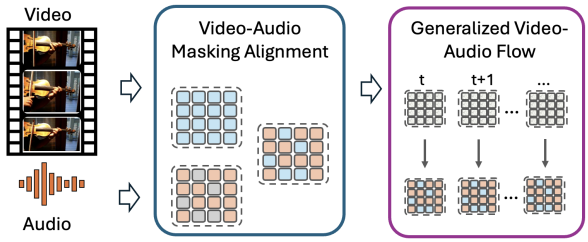

图 2. 所提出的 *Foley-Flow* 框架示意图。

3.1. 预备知识

本节首先介绍问题设定和符号，然后回顾对比式视频-音频预训练以及去噪扩散概率模型。

问题设置与符号。 视频到音频生成任务旨在根据连续视频帧$\mathbf{A}$ 合成时间上对齐的音频$\mathbf{V}$。令$\mathbf{F}^v \in \mathbf{R}^{T_v \times D_v}$和$\mathbf{F}^a \in \mathbf{R}^{T_a \times D_a}$ 分别表示从视频和音频模态中提取的特征，其中 T_v 和 T_a 为时间长度， D_v和 D_a 为特征维度。目标是确保生成音频$\mathbf{A}$在语义上连贯，并与$\mathbf{V}$在节奏上对齐。

掩码自编码器。 广义掩码自编码器 [5] 已展现出其在单模态表征学习中的有效性。在视频和音频场景中，单模态掩码是指随机遮蔽一部分输入词元$\mathbf{X}_{\text{mask}}$，并利用其余未掩码词元$\mathbf{X}_{\text{unmask}}$对其进行重建。对于视频帧$\mathbf{V}$和音频频谱图$\mathbf{A}$，重建目标为：

$$\mathcal{L}_{\text{recon}} = \|\mathbf{X}_{\text{mask}} - \hat{\mathbf{X}}_{\text{mask}}(\mathbf{X}_{\text{unmask}})\|^2. \tag{1}$$

这种自监督方法使模型能够学习丰富的单模态表征。

基于流的生成。 基于流的生成模型学习从简单噪声分布到复杂目标分布的可逆映射。给定一个潜变量 $\mathbf{z}$，流变换 f_ϕ 将 $\mathbf{z}$ 映射到目标分布：

$$\mathbf{x} = f_\phi(\mathbf{z}), \tag{2}$$

其中 $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$。该目标旨在最小化生成分布与目标分布之间的 Kullback-Leibler 散度：

$$\mathcal{L}_{\text{flow}} = D_{\text{KL}}(p(\mathbf{x}) \| p_{\text{target}}(\mathbf{x})). \tag{3}$$

在 *Foley-Flow* 中，我们构建了一种动态条件流，其条件是随时间变化的视频特征。

3.2. 视频-音频掩码对齐

传统的视频到音频生成方法通常依赖对比学习来对齐音频和视频表征。尽管这类方法在全局语义对齐方面效果显著，但由于没有显式建模音频片段与视频片段之间的时间依赖关系，因此难以捕捉节奏同步。为解决这一局限，我们提出了视频-音频掩码对齐（VAMA）策略，迫使模型利用时间上对应的视频特征重建掩码音频片段。这不仅能在语义上对齐音频和视频表征，还能确保二者的时间节奏保持同步。

为确保语义和节奏连贯性，我们提出了一种新颖的跨模态掩码策略，利用视频和音频信息的互补特性。该策略会对部分音频片段进行掩码$\mathbf{F}^a_{\text{mask}}$，并要求模型利用逐帧视频特征$\mathbf{F}^v$以及未掩码的音频输入$\mathbf{F}^a_{\text{unmask}}$重建这些片段。这种设置模拟了现实场景：在这些场景中，必须根据视频线索和剩余的音频上下文推断不完整或含噪的音频。自监督重建目标定义如下：

$$\mathcal{L}_{\text{VAMA}} = \|\mathbf{F}^a_{\text{mask}} - \hat{\mathbf{F}}^a_{\text{mask}}(\mathbf{F}^v, \mathbf{F}^a_{\text{unmask}})\|^2, \tag{4}$$

ˆ其中$\mathbf{F}^a_{\text{mask}}$表示模型重建的音频片段。掩码沿时间维度执行，以确保模型学会捕捉视频帧与其对应音频片段之间的同步模式。

为此，视频编码器提取时间特征$\mathbf{F}^v$，以表征视频帧中的空间和上下文信息。这些特征通过交叉注意力机制与潜在空间中的音频特征$\mathbf{F}^a$对齐，从而确保音频的时间动态与视频特征同步。掩码策略促使模型关注节奏模式，因为必须利用视觉和音频线索重建被掩码音频片段的时间连续性。通过结合跨模态掩码与时间一致性，视频-音频掩码对齐模块使模型能够学习兼具语义和节奏连贯性的鲁棒表征。

与仅对齐全局表征的传统对比方法不同，我们的视频-音频掩码对齐明确对齐局部和时间特征。这种方法确保模态之间的节奏一致性，这对于同步的视频到音频生成至关重要。在跨模态语境中创新性地采用片段级掩码，也为训练具有时间对齐能力的视听模型提供了新方法。

3.3. 广义视频-音频流

基于流的模型以能够通过高效采样生成高质量输出而著称。然而，现有的生成任务中基于流的方法往往使用静态条件，无法充分捕捉视频到音频生成的动态特性。视频特征会随时间演变，反映场景内容、运动和节奏的变化，因此应动态融入音频生成过程。为此，我们提出一种动态条件流模块，根据随时间变化的视频片段对音频生成进行条件控制，从而确保生成音频在语义和节奏上均保持对齐。

动态条件流。我们的基于流的模型建立在速度流框架之上，将随时间变化的视频特征$\mathbf{F}^v_t$作为动态条件纳入其中。流变换定义为：

$$\mathbf{F}^a_t = f_\phi(\mathbf{z}_t, \mathbf{F}^v_t), \tag{5}$$

其中$\mathbf{z}_t$是从高斯分布中采样的噪声向量。与静态流不同，引入时间条件使模型能够调整生成音频，以反映视频动态随时间的变化。

为了有效建模时间同步，我们使用视频片段特征$\mathbf{F}^v_t$作为基于流的生成器的条件。这些特征同时捕捉高层语义（例如物体动作）和低层节奏（例如运动模式），它们在不同视频帧之间发生变化。通过以这些动态特征对流模型进行条件化，生成音频$\mathbf{F}^a_t$以时间连贯的方式演化，从而确保与视频的时间结构对齐。

动态条件流模块提出了一种将随时间变化的条件纳入基于流的生成模型的新方法。与静态条件化方法不同，我们的方法动态调整音频生成过程，以反映视频片段的变化，从而实现细粒度同步。这种时间适应性尤其适用于视频到音频任务，其中音频的节奏和语义必须紧密跟随视频的进展。

广义动态条件流。为克服传统去噪过程效率低下的问题，并实现快速、高质量的视频到音频生成，我们提出了一种广义动态条件流模块。该模块利用归一化流，提供从简单噪声分布到复杂目标音频分布的连续可逆映射。基于流的方法不仅提升了推理速度，还增强了模型生成高保真音频输出的能力。

基于流的映射。给定视频的潜在表示$\mathbf{z}_v$，基于流的生成器学习一种变换

f_ϕ 将 $\mathbf{z}_v$ 和一个噪声向量 $\mathbf{z}_{\text{noise}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 映射到对应的音频潜变量 $\mathbf{z}_a$：

$$\mathbf{z}_a = f_\phi(\mathbf{z}_v, \mathbf{z}_{\text{noise}}). \tag{6}$$

f_ϕ 的可逆性确保能够有效地将生成音频追溯至其原始潜变量分布，从而实现与视频特征的精确对齐。

训练目标。 广义视频-音频流（GVAF）模块通过最小化生成音频特征 $p(\mathbf{z}_a)$的分布与目标分布$p_{\text{target}}(\mathbf{z}_a)$之间的库尔贝克-莱布勒散度进行训练：

$$\mathcal{L}_{\text{GVAF}} = D_{\text{KL}}(p(\mathbf{z}_a) \| p_{\text{target}}(\mathbf{z}_a)). \tag{7}$$

这一目标确保生成音频在分布特性和时间特性上都与目标音频高度吻合。

利用条件流进行高效推理。 在推理过程中，基于流的模块通过减少合成所需的迭代步数，加速视频到音频的生成过程。它不依赖耗时的去噪过程，而是通过条件流直接一步生成高质量音频：

$$\mathbf{z}_a = f_\phi^{-1}(\mathbf{z}_v, \mathbf{z}_{\text{noise}}). \tag{8}$$

这种方法计算效率高，非常适合实时应用，因为它绕过了传统扩散模型的迭代过程。

与掩码对齐的集成。为进一步提升生成质量，基于流的模块以视频-音频掩码对齐模块学习到的特征为条件。该条件信息为流提供了语义丰富且节奏连贯的视频表征，从而确保生成音频在上下文中与视觉输入对齐。

通过将基于流的模块与掩码对齐策略相结合，我们的 *Foley-Flow* 兼顾了效率与质量，能够在实时环境中生成同步且高保真的音频输出。

4. 实验

4.1. 实验设置

数据集。 我们在两个广泛使用的数据集上进行实验：VGGSound [1]由 20 万个 YouTube 视频片段组成，每段长 10 秒，涵盖 309 种多样的声音类别，例如动物、车辆、人类语音、舞蹈和乐器。AudioSet [2] 是一个规模更大的视听数据库，包含约 200 万个 YouTube 视频，内容丰富多样。

表1. 视频到音频生成在VGGSound测试集上的比较结果。报告了KLD、FAD和对齐准确率的数值。

方法	KLD ↓	FAD ↓	对齐准确率 ↑
SpecVQGAN [10]	3.78	6.63	48.79
Im2Wav [46]	2.54	6.32	74.31
Diff-Foley [17]	3.15	6.40	82.47
FoleyGen [19]	2.89	2.59	73.83
V2A-Mapper [52]	2.78	0.99	74.37
Seeing & Hearing [53]	2.62	2.63	78.95
MaskVAT [41]	2.65	1.51	63.87
VAB [48]	2.58	2.69	76.83
VATT [16]	2.25	2.35	82.81
Foley-Flow （我们的方法）	**0.97**	**0.52**	**98.97**

评估指标。 为了评估视频到音频生成的质量，我们采用以下指标：库尔贝克-莱布勒散度（KLD）通过比较 PaSST 特征的分布，衡量生成音频与真实值（GT）的匹配程度 [14], ，反映音频对视频中概念的呈现效果。Fr´echet 音频距离（FAD） [11]评估生成音频相较于真实音频的分布质量，从而衡量整体质量。对齐准确率（Align Acc） [17] 评估生成音频与相应视频帧之间的时间对齐和相关性。生成速度通过计算生成一个波形样本所需的时间来衡量效率。

实现。 输入视频帧被调整为224×224 分辨率。音频输入表示为从以 8000Hz 采样的 10 秒片段中提取的对数频谱图。遵循先前的工作 [21], ，音频频谱图使用带有 50ms 窗长和 25ms hop size 的短时傅里叶变换（STFT）生成，从而得到大小为 128×128 的输入张量（128 个频带、128 个时间步）。对于音频和视觉编码器，我们使用单模态编码器 [9, 49] ，利用在 LAION-400M 上预训练的权重初始化视觉编码器 [45]，并使用在 AudioSet 上预训练的音频编码器 [2]。模型使用 Adam 优化器 [12]进行了 100 个 epoch 的训练，学习率为 $1e-4$ ，批量大小为 128。

4.2. 与先前工作的比较

在本文中，我们提出了一种新颖且有效的框架 *Foley-Flow*，用于视频到音频生成。为展示所提出 *Foley-Flow*的有效性，我们针对一系列成熟的视频到音频生成模型，对其性能进行了广泛评估。SpecVQGAN （BMVC 2021）[10]是一种用于音频频谱图生成的矢量量化生成对抗网络；Im2Wav（ICASSP 2023） [46]使用 CLIP 嵌入直接从图像生成波形；Diff-Foley （ NeurIPS 2023） [17] 将对比预训练与

用于音频生成的扩散模型；FoleyGen（arXiv 2023）[19]利用神经音频编解码器和语言建模范式，实现双向的波形到令牌转换；Seeing & Hearing（CVPR 2024）[53] 采用由 ImageBind 引导的多模态潜在对齐器 [3]进行基于扩散的跨模态生成；VAB（ICML 2024） [48] 预训练视觉条件掩码音频令牌预测，以在无需扩散的情况下进行表征学习；MaskVAT（ECCV 2024） [41] 将高质量通用音频编解码器与掩码生成模型相结合，实现语义和时间对齐；VATT（NeurIPS 2024） [16]利用经过指令微调的大语言模型，通过并行解码将视频特征映射到音频令牌空间。

表1 全面比较了*Foley-Flow* 与当前最先进的视频转音频生成模型在 VGGSound测试集上的表现。KLD 和 FAD 值越低，表示语义和整体音频质量越好；Align Acc 越高，则表示视频与生成音频之间的时间同步效果越佳。KLD 用于衡量生成音频与真实音频分布之间的语义相似度。 *Foley-Flow*取得了最低的 KLD 分数 0.97，显著优于次佳模型 VATT（KLD 为 2.25）。这一结果表明， *Foley-Flow* 在捕捉生成音频的语义一致性方面表现出色，能够使其与视频帧内容紧密对应。与 Diff-Foley（3.15）和 Mask-VAT（2.65）等模型相比，性能的提升凸显了我们的视频-音频掩码对齐模块在训练期间对齐语义表示的有效性。FAD 用于评估生成音频的整体质量及其与真实音频的分布相似度。 *Foley-Flow* 取得了 0.52 的 FAD，在所有模型中表现最佳。与此前表现最佳的模型 FoleyGen（FAD 为 2.59）相比，这一结果有了大幅提升。结果表明， *Foley-Flow* 采用的动态条件流不仅能生成高质量音频，还能在各时间片段间保持连贯性，确保平滑过渡并生成逼真的音频输出。Align Acc 用于衡量视频与生成音频之间的时间同步程度。*Foley-Flow* 取得了 98.97 的 Align Acc，创下新基准，并显著超越最接近的竞争模型 Diff-Foley（Align Acc 为 82.47）。这一提升凸显了 *Foley-Flow*有效建模节奏依赖关系的能力，并利用动态视频条件引导音频生成。在基于流的生成过程中融入随时间演变的视频特征，确保生成音频与视频中的事件时序相匹配。

SpecVQGAN 在所有指标上的表现都较差，KLD（3.78）和 FAD（6.63）较高，而对齐准确率较低（48.79）。这凸显了生成对抗网络在捕捉跨模态依赖关系方面的局限性。

表2. 针对视频到音频生成的视频-音频掩码对齐（VAMA）与广义视频-音频流（GVAF）消融实验。报告了KLD、FAD和对齐准确率。

VAMA	GVAF	KLD ↓	FAD ↓	对齐准确率 ↑
✗	✗	3.15	6.40	82.47
✓	✗	1.68	1.57	96.29
✗	✓	1.92	1.89	93.86
✓	✓	**0.97**	**0.52**	**98.97**

关系方面的局限性。Im2Wav 和 V2A-Mapper 的表现优于 SpecVQGAN，但在节奏同步方面仍有不足，对齐准确率分别为 74.31 和 74.37。Diff-Foley 的对齐准确率表现出色（82.47），但其较高的 KLD（3.15）和 FAD（6.40）表明音频质量和语义对齐效果欠佳。FoleyGen、Seeing & Hearing 和 VAB 的表现较为均衡，但其对齐准确率较低、KLD 较高，未能达到 *Foley-Flow* 在语义和节奏对齐方面的水平。尽管 MaskVAT 和 VATT 采用了先进的视听编码策略，但静态条件和时间适应性不足，导致其结果逊于 *Foley-Flow*。结果表明， *Foley-Flow* 在全部三项指标上均显著优于当前最先进的方法。视频-音频掩码对齐模块与动态条件流模块相结合，使 *Foley-Flow* 能够生成语义准确、时间同步且质量上乘的音频。这些进展使*Foley-Flow* 成为一个稳健的视频转音频生成框架，为该领域的性能树立了新标准。

我们在补充材料中提供了定性结果，将 *Foley-Flow* 与基线方法进行比较。 *Foley-Flow* 的生成音频输出在语义上更丰富，节奏同步更好，并且与视频内容在时间上更加对齐。值得注意的是，*Foley-Flow*能够以比竞争方法更高的保真度和自然度处理复杂场景，例如对视觉动作的延迟听觉响应。

4.3. 实验分析

在本节中，我们进行了详细的消融实验，以证明在我们提出的框架中引入视频-音频掩码对齐（VAMA）和广义视频-音频流（GVAF）的益处。此外，我们开展了大量实验，以探究视频-音频编码器的选择、视频-音频流目标以及掩码比例的影响。

视频-音频掩码对齐与广义视频-音频流。为评估视频-音频掩码对齐（VAMA）和广义视频-音频流（GVAF）模块对 *Foley-Flow* 的贡献，我们通过分别移除这两个组件以及同时移除它们，开展了消融研究

。表 2 所示的定量结果表明，这两个模块对于在视频到音频生成中实现最先进性能至关重要。移除 VAMA 后，对齐准确率从 *98.97%* 降至 *93.86%*，表明生成音频与视频帧之间的时间同步有所减弱。此外，KLD 从 *0.97* 升至 *1.92*，说明视频与音频之间的语义对齐不再那么稳健。这些结果凸显了 VAMA 的跨模态掩码策略对于实现节奏同步、确保音频和视频的语义表示有效对齐至关重要。移除 GVAF 后，FAD 分数从 *0.52* 大幅升至 *1.57*，表明生成音频的质量显著下降。这是因为缺少 GVAF，模型在音频生成过程中无法利用时间动态视频条件。对齐准确率也略微降至 *96.29%*，反映出时间对齐能力减弱。这些结果凸显了 GVAF 在确保高效、高保真音频生成方面的作用，尤其是通过其动态条件流机制。当两个模块都被移除时，所有指标的性能均显著恶化。KLD 升至 *3.15*，FAD 升至 *6.40*，对齐准确率则骤降至 *82.47%*。这一显著下降表明，VAMA 和 GVAF 对于实现语义连贯性、节奏同步以及高质量音频输出缺一不可。缺少这两个模块时，模型难以有效对齐视频输入并生成与之相符的音频。加入 VAMA 和 GVAF 后， *Foley-Flow* 在所有指标上均取得最佳性能：KLD 为 *0.97*，FAD 为 *0.52*，对齐准确率为 *98.97%*。这些结果验证了两个模块协同解决视频到音频生成中的语义和节奏难题的能力。VAMA 通过跨模态掩码策略，在训练过程中确保稳健的跨模态对齐；GVAF 则通过时间自适应流式方法，实现高效且高质量的生成。消融研究证实，VAMA 和 GVAF 对于 *Foley-Flow* 都必不可少。VAMA 通过跨模态掩码策略确保语义和节奏连贯性，而 GVAF 则提供动态时间条件控制，以提升音频生成的质量和效率。这些模块共同助力 *Foley-Flow* 在视频到音频生成方面超越现有方法。

视频-音频编码器类型。 为探究视频和音频输入采用不同编码器组合所产生的影响，我们使用 ImageBind [3], MAE [5], 和 EVA-CLIP [49] 进行视频编码，并使用 ImageBind 和 AudioMAE [9] 进行音频编码。定量结果见表 3。在三个视频编码器中，与任一音频编码器搭配时，EVA-CLIP 在所有指标上均始终取得最佳结果

表3. 视频到音频生成中视频-音频编码器的消融实验。报告了KLD、FAD和对齐准确率。

视频	音频	KLD ↓	FAD ↓	对齐准确率 ↑
ImageBind	ImageBind	1.38	1.27	97.29
MAE	ImageBind	1.17	1.06	98.32
EVA-CLIP	ImageBind	1.01	0.73	98.52
ImageBind	AudioMAE	1.36	1.13	97.74
MAE	AudioMAE	1.06	0.65	98.62
EVA-CLIP	AudioMAE	**0.97**	**0.52**	**98.97**

表 4. 视频到音频生成中动态条件流（DCF）和广义动态条件流（GDCF）的消融实验。报告了 KLD、FAD 和对齐准确率。

流类型	KLD ↓	FAD ↓	对齐准确率 ↑
DCF	1.28	0.85	97.73
GDCF	**0.97**	**0.52**	**98.97**

编码器。例如，与 AudioMAE 结合时，EVA-CLIP 的 KLD 最低（*0.97*）、FAD 最低（*0.52*），对齐准确率最高（*98.97%*）。这表明 EVA-CLIP 能够捕捉丰富的语义表示和稳健的时间特征，而这对于将视频帧与相应的音频片段对齐至关重要。相比之下，ImageBind 和 MAE 的表现略逊一筹。例如，ImageBind 与 AudioMAE 搭配时，KLD 较高，为 *1.36* ，对齐准确率较低，为*97.74%*，说明其语义和节奏对齐能力较弱。在所有情况下，AudioMAE 的表现均优于 ImageBind，展现了其学习细致音频表示的优势。例如，与 EVA-CLIP 搭配时，AudioMAE 的 FAD 为*0.52*；而使用 ImageBind 作为音频编码器时，FAD 为*0.73* 。这表明 AudioMAE 更适合生成高保真且与视频在时间上对齐的音频。采用 EVA-CLIP 进行视频编码、AudioMAE 进行音频编码的组合，在所有指标上均取得最佳结果，KLD 为*0.97*，FAD 为*0.52*，对齐准确率为*98.97%*。这表明，EVA-CLIP 在大规模图像文本数据集上的预训练与 AudioMAE 对细致频谱图重建的专注相辅相成，从而使生成音频实现出色的语义和节奏对齐。这些结果凸显了为视频到音频生成选择稳健编码器组合的重要性。EVA-CLIP 与 AudioMAE 的组合在语义丰富度和时间对齐之间实现了最佳平衡，从而达到最先进性能。

视频音频流的类型。为评估不同流目标的影响，我们比较了动态条件流（DCF）和广义动态条件流（GDCF）在视频-音频生成任务中的表现。定量结果汇总于表4中。DCF方法取得了优异的表现，其

表 5. 视频到音频生成中掩码比例的消融实验。报告了 KLD、FAD 和对齐准确率。

掩码比例	KLD ↓	FAD ↓	对齐准确率 ↑
0.3	1.25	1.06	98.12
0.5	1.16	0.87	98.37
0.7	1.09	0.65	98.78
0.8	**0.97**	**0.52**	**98.97**
0.9	1.02	0.56	98.96

KLD为*1.28*，FAD为*0.85*，对齐准确率为*97.73%*。尽管DCF能够有效捕捉音频生成过程中的时间依赖关系，但其依赖潜在空间中的静态变换，限制了其适应视频片段复杂时间变化的能力。这导致其在评估语义对齐（KLD）和音频质量（FAD）的指标上表现欠佳。GDCF通过引入更大的潜在变换灵活性扩展了DCF，使其能够动态适应视频不断变化的时间和语义特征。因此，GDCF在所有指标上均取得了显著提升，KLD为*0.97*，FAD为*0.52*，对齐准确率为*98.97%*。对复杂时间动态的建模能力使GDCF生成的音频输出不仅在时间上保持同步，而且与DCF相比具有更丰富的语义和更高的质量。GDCF的KLD显著低于DCF（*0.97* 对比*1.28*），表明其在生成音频与视频输入之间实现了更优的语义对齐。GDCF的FAD（*0.52*）明显优于DCF（*0.85*），说明GDCF生成的音频输出在分布上更接近真实世界的音频。GDCF在对齐准确率方面同样优于DCF，达到*98.97%*的近乎完美同步，而DCF为*97.73%*。这体现了GDCF有效处理细粒度时间依赖关系的能力。这些结果证明，在平衡质量与效率方面，GDCF优于DCF。通过在方差保持流的基础上引入额外的灵活性，GDCF能够更有效地捕捉视频-音频任务中的分布特征和时间特征，从而在所有评估指标上达到最先进的性能。这凸显了动态且自适应的基于流的目标在实现高质量、同步的视频到音频生成中的关键作用。

掩码比例的影响。为了分析掩码比例对视频-音频掩码对齐（VAMA）模块性能的影响，我们采用掩码比例{0.3、0.5、0.7、0.8、0.9}进行模型评估。定量结果见表5。掩码比例为 0.8 时，各项指标均达到最佳：KLD 为*0.97*，FAD 为*0.52*，对齐准确率为*98.97%*。这表明，当 80% 的音频片段被掩码时，模型能够有效学习稳健的跨模态依赖和时间同步关系。这一比例恰到好处，既能促使模型推断缺失信息，又能提供充足的上下文线索以有效训练。当掩码比例降至 0.3 和 0.5 时，各项指标的性能均有所下降。例如，掩码比例为 0.3 时，KLD 升至 *1.25*，FAD 升至*1.06*，对齐准确率则降至 *98.12%*。这表明，较低的掩码比例不足以促使模型学习稳健的表征，因为大部分音频信息都可直接获取，从而降低了模型对视频引导的依赖。将掩码比例提高至 0.9 会导致性能略微下降，KLD 为 *1.02*，FAD 为*0.56*，对齐准确率为 *98.96%*。尽管性能下降幅度不大，但过度掩码可能会导致关键音频上下文丢失，使模型更难准确重建缺失片段。结果凸显了不同掩码比例之间的权衡。较低比例提供的信息过多，限制了模型学习有意义依赖关系的能力；较高比例则会造成信息过度丢失。0.8 的最佳掩码比例平衡了这些因素，促使模型依赖跨模态线索进行重建，同时保留足够的上下文以实现有效学习。掩码比例对 VAMA 模块的有效性至关重要。比例为 0.8 时，各项指标均达到最佳性能，使 *Foley-Flow* 能够生成语义连贯、节奏协调且质量上乘的音频。这表明，为对齐模型训练选择适当的掩码比例十分重要。

5. 结论

本文提出了 *Foley-Flow*，这是一种新颖的视频到音频生成框架，通过创新的跨模态掩码和基于流的建模技术，实现语义与节奏连贯性。我们的 *Foley-Flow*通过引入视频-音频掩码对齐模块，强制视频帧与音频片段之间实现细粒度对齐；并引入广义视频-音频流（GVAF）模块，实现高效、高质量的音频生成，从而克服了以往方法的局限。通过在 VGGSound 和 AudioSet 数据集上进行大量实验，我们证明了 *Foley-Flow* 在多项指标上均达到最先进性能，包括库尔贝克-莱布勒散度（KLD）、Fr´echet 音频距离（FAD）和对齐准确率（Align Acc）。我们的框架生成的音频输出同步且符合上下文，同时大幅缩短推理时间，性能优于现有方法。消融实验进一步验证了所提组件的有效性。此外，视频和音频编码器的选择、广义方差保持流的使用以及最优掩码比例，也促成了 *Foley-Flow*的成功。

参考文献

[1] Honglie Chen、Weidi Xie、Andrea Vedaldi 和 Andrew Zisserman。《VGGSound：大规模音视频数据集》。发表于国际声学、语音与信号处理会议（ICASSP），第 721–725 页。IEEE，2020 年。5[2] Jort F. Gemmeke、Daniel P. W. Ellis、Dylan Freedman、Aren Jansen、Wade Lawrence、R. Channing Moore、Manoj Plakal 和 Marvin Ritter。《AudioSet：用于音频事件的本体与人工标注数据集》。发表于 2017 年 IEEE 国际声学、语音与信号处理会议论文集（ICASSP），第 776–780 页，2017 年。5[3] Rohit Girdhar、Alaaeldin El-Nouby、Zhuang Liu、Mannat Singh、Kalyan Vasudev Alwala、Armand Joulin 和 Ishan Misra。《ImageBind：一个将所有模态绑定在一起的嵌入空间》。发表于 IEEE/CVF 计算机视觉与模式识别会议（CVPR）论文集，第 15180–15190 页，2023 年。2，6，7。[4] Yuan Gong、Andrew Rouditchenko、Alexander H. Liu、David Harwath、Leonid Karlinsky、Hilde Kuehne 和 James R. Glass。《对比式音视频掩码自编码器》。发表于第十一届国际学习表征会议（ICLR）论文集，2023 年。2[5] 何恺明、Xinlei Chen、Saining Xie、Yanghao Li、Piotr Doll´ar 和 Ross B. Girshick。《掩码自编码器是可扩展的视觉学习器》。arXiv 预印本 arXiv:2111.06377，2021 年。3，7[6] Jonathan Ho、Ajay Jain 和 Pieter Abbeel。《去噪扩散概率模型》。发表于神经信息处理系统进展（NeurIPS）论文集，2020 年。3[7] Jonathan Ho、William Chan、Chitwan Saharia、Jay Whang、Ruiqi Gao、Alexey Gritsenko、Diederik P. Kingma、Ben Poole、Mohammad Norouzi、David J. Fleet 和 Tim Salimans。《Imagen Video：基于扩散模型的高清视频生成》。arXiv 预印本 arXiv:2210.02303，2022 年。3[8] Po-Yao Huang、Vasu Sharma、Hu Xu、Chaitanya Ryali、Haoqi Fan、Yanghao Li、Shang-Wen Li、Gargi Ghosh、Jitendra Malik 和 Christoph Feichtenhofer。《MAViL：掩码音视频学习器》。arXiv 预印本 arXiv:2212.08071，2022 年。2，3。[9]Po-Yao Huang、Hu Xu、Juncheng Li、Alexei Baevski、Michael Auli、Wojciech Galuba、Florian Metze 和 Christoph Feichtenhofer。《能够聆听的掩码自编码器》。发表于神经信息处理系统进展（NeurIPS）论文集，2022 年。5，7[10] Vladimir Iashin 和 Esa Rahtu。《驾驭视觉引导的声音生成》。发表于英国机器视觉会议（BMVC），2021 年。1，2，5[11] Kevin Kilgour、Mauricio Zuluaga、Dominik Roblek 和 Matthew Sharifi。《Fr´echet 音频距离：一种评估音乐增强算法的指标》。arXiv 预印本 arXiv:1812.08466，2018 年。5[12] Diederik P. Kingma 和 Jimmy Ba。《Adam：一种随机优化方法》。arXiv 预印本 arXiv:1412.6980，2014 年。5。[13] Zhifeng Kong、Wei Ping、Jiaji Huang、Kexin Zhao 和 Bryan Catanzaro。《DiffWave：一种通用的扩散模型，用于音频合成。载于国际学习表征会议（ICLR），2021年。3[14] Khaled Koutini、Jan Schl¨uter、Hamid Eghbal-zadeh 和 Ger- hard Widmer。使用 Patchout 高效训练音频变换器。arXiv 预印本 arXiv:2110.05069，2022年。5[15] Felix Kreuk、Gabriel Synnaeve、Adam Polyak、Uriel Singer、Alexandre D´efossez、Jade Copet、Devi Parikh、Yaniv Taig- man 和 Yossi Adi。Audiogen：文本引导的音频生成。载于国际学习表征会议（ICLR），2023年。1[16] Xiulong Liu、Kun Su 和 Eli Shlizerman。从所见推知所闻——通过文本实现视频转音频生成。载于神经信息处理系统进展（NeurIPS）论文集，2024年。2, 5, 6[17] Simian Luo、Chuanhao Yan、Chenxu Hu 和 Hang Zhao。Diff-foley：通过潜在扩散模型实现同步的视频转音频合成。arXiv 预印本 arXiv:2306 17203，2023 1, 2, 5 . . [18] Tanvir Mahmud、Shentong Mo、Yapeng Tian 和 Di- ana Marculescu。MA-AVT：面向参数高效音视变换器的模态对齐。arXiv 预印本 arXiv:2406.04930，2024年。2[19] Xinhao Mei、Varun Nagaraja、Gael Le Lan、Zhaoheng Ni、Ernie Chang、Yangyang Shi 和 Vikas Chandra。Foley-gen：视觉引导的音频生成。arXiv 预印本 arXiv:2309.10537，2023 年。2, 5, 6[20] Shentong Mo。通过采用双向状态空间模型的扩散 Mamba 高效生成三维形状。arXiv 预印本 arXiv: 2406.05038，2024年。3[21] Shentong Mo 和 Pedro Morgado。轻松定位视觉声音。载于欧洲计算机视觉会议（ECCV）论文集，第218–234页，2022年。1, 5[22] Shentong Mo 和 Pedro Morgado。深入探究弱监督音视频源定位。载于神经信息处理系统进展（NeurIPS）论文集，2022年。[23] Shentong Mo 和 Pedro Morgado。弱监督音视频声音定位基准测试。载于欧洲计算机视觉会议（ECCV）研讨会，2022年。1[24] Shentong Mo 和 Pedro Morgado。通过掩码建模揭示具有密集交互的音视早期融合变换器的潜力。arXiv 预印本 arXiv:2312.01017，2023年。2[25] Shentong Mo 和 Pedro Morgado。面向定位、分离和识别的统一音视学习框架。载于国际机器学习会议（ICML）论文集，2023年。1[26] Shentong Mo 和 Pedro Morgado。轻松实现音视频广义零样本学习。arXiv 预印本 arXiv:2407.13095，2024年。2[27] Shentong Mo 和 Bhiksha Raj。弱监督音视频分割。arXiv 预印本 arXiv:2311 15080，2023 1 . . [28] Shentong Mo 和 Yibing Song。通过智能体工作流对齐音视联合表征。arXiv 预印本 arXiv:2410.23230，2024年。2

[29] Shentong Mo 和 Yapeng Tian。用于弱监督音视频解析的多模态分组网络。载于神经信息处理系统进展（NeurIPS）论文集，2022年。1[30] Shentong Mo 和 Yapeng Tian。用于弱监督音视频解析的语义感知多模态分组。载于欧洲计算机视觉会议（ECCV）研讨会，2022年。1[31] Shentong Mo 和 Yapeng Tian。用于混合音频声音定位的音视频分组网络。arXiv 预印本 arXiv:2303.17056，2023年。2[32] Shentong Mo 和 Yapeng Tian。AV-SAM：通用分割模型与音视频定位和分割相结合。arXiv 预印本 arXiv:2305.01836，2023年。1[33] Shentong Mo 和 Yapeng Tian。通过双向状态空间模型扩展扩散 Mamba，以高效生成图像和视频。arXiv 预印本 arXiv: 2405.15881，2024年。3[34] Shentong Mo 和 Yapeng Tian。用于音源分离的语义分组网络。arXiv 预印本 arXiv:2407.03736，2024年。2[35] Shentong Mo、Weiguo Pian 和 Yapeng Tian。用于持续音视频学习的类增量分组网络。arXiv 预印本 arXiv:2309.05281，2023年。2[36] Shentong Mo、Jing Shi 和 Yapeng Tian。DiffAVA：通过视觉对齐实现个性化文本转音频生成。arXiv 预印本 arXiv:2305.12903，2023年。2[37] Shentong Mo、Enze Xie、Ruihang Chu、Lewei Yao、Lanqing Hong、Matthias Nießner 和 Zhenguo Li。DiT-3D：探索用于三维形状生成的朴素扩散变换器。载于神经信息处理系统进展（NeurIPS）论文集，2023年。3[38] Shentong Mo、Enze Xie、Yue Wu、Junsong Chen、Matthias Nießner 和 Zhenguo Li。通过极端掩码快速训练扩散变换器以生成三维点云。arXiv 预印本 arXiv: 2312.07231，2023年。3[39] Shentong Mo、Jing Shi 和 Yapeng Tian。与视频同步的文本转音频生成。arXiv 预印本 arXiv:2403.07938，2024 年。2[40] Arsha Nagrani 、Shan Yang、Anurag Arnab、Aren Jansen、Cordelia Schmid 和 Chen Sun。用于多模态融合的注意力瓶颈。载于神经信息处理系统进展（NeurIPS）论文集，2021年。3[41] Santiago Pascual、Chunghsin Yeh、Ioannis Tsiamas 和 Joan Serr`a。具有增强同步性的掩码生成式视频转音频变换器。载于欧洲计算机视觉会议（ECCV）论文集，2024年。2, 5, 6[42] Weiguo Pian、Shentong Mo、Yunhui Guo 和 Yapeng Tian。音视频类增量学习。arXiv 预印本 arXiv:2308.11073，2023年。2[43] Chitwan Saharia、Jonathan Ho、William Chan、Tim Salimans、David J. Fleet 和 Mohammad Norouzi。通过迭代细化实现图像超分辨率。arXiv 预印本 arXiv:2104.07636，2021年。3[44] Chitwan Saharia、William Chan、Saurabh Saxena、Lala Li、Jay Whang、Emily Denton、Seyed Kamyar Seyed Ghasemipour、Burcu Karagol Ayan、S. Sara Mahdavi、Rapha Gontijo Lopes、Tim Salimans、Jonathan Ho、David J Fleet，以及 Mohammad Norouzi。具有深度语言理解能力的逼真文本到图像扩散模型。arXiv预印本 arXiv:2205.11487，2022年。3[45] Christoph Schuhmann、Richard Vencu、Romain Beaumont、Robert Kaczmarczyk、Clayton Mullis、Aarush Katta、Theo Coombes、Jenia Jitsev 和 Aran Komatsuzaki。Laion-400m：由 CLIP 筛选的 4 亿图文对开放数据集。arXiv预印本 arXiv:2111.02114，2021年。5[46] Roy Sheffer 和 Yossi Adi。我听见你真实的色彩：图像引导的音频生成。载于国际声学、语音与信号处理会议（ICASSP），2023年。1、2、5[47] Yang Song、Jascha Sohl-Dickstein、Diederik P Kingma、Abhishek Kumar、Stefano Ermon 和 Ben Poole。通过随机微分方程进行基于分数的生成建模。载于国际学习表征会议（ICLR），2021年。3[48] Kun Su、Xiulong Liu 和 Eli Shlizerman。从视觉到音频及更远：用于音频-视觉表征与生成的统一模型。载于国际机器学习会议（ICML）论文集，2024年。2、5、6[49] Quan Sun、Yuxin Fang、Ledell Wu、Xinlong Wang 和 Yue Cao。EVA-CLIP：改进 CLIP 的大规模训练技术。arXiv预印本 arXiv:2303.15389，2023年。5、7[50] Gemma Team、Thomas Mesnard、Cassidy Hardin、Robert Dadashi、Surya Bhupatiraju、Shreya Pathak、Laurent Sifre、Morgane Rivi`ere、Mihir Sanjay Kale、Juliette Love、Pouya Tafti、L´eonard Hussenot、Pier Giuseppe Sessa、Aakanksha Chowdhery、Adam Roberts、Aditya Barua、Alex Botev、Alex Castro-Ros、Ambrose Slone、Am´elie H´eliou、Andrea Tac-chetti、Anna Bulanova、Antonia Paterson、Beth Tsai、Bobak Shahriari、Charline Le Lan、Christopher A. Choquette-Choo、Cl´ement Crepy、Daniel Cer、Daphne Ippolito、David Reid、Elena Buchatskaya、Eric Ni、Eric Noland、Geng Yan、George Tucker、George-Christian Muraru、Grigory Rozhdestvenskiy、Henryk Michalewski、Ian Tenney、Ivan Grishchenko、Ja-cob Austin、James Keeling、Jane Labanowski、Jean-Baptiste Lespiau、Jeff Stanway、Jenny Brennan、Jeremy Chen、Jo-han Ferret、Justin Chiu、Justin Mao-Jones、Katherine Lee、Kathy Yu、Katie Millican、Lars Lowe Sjoesund、Lisa Lee、Lucas Dixon、Machel Reid、Maciej Mikuła、Mateo Wirth、Michael Sharman、Nikolai Chinaev、Nithum Thain、Olivier Bachem、Oscar Chang、Oscar Wahltinez、Paige Bailey、Paul Michel、Petko Yotov、Rahma Chaabouni、Ramona Co-manescu、Reena Jana、Rohan Anil、Ross McIlroy、Ruibo Liu、Ryan Mullins、Samuel L Smith、Sebastian Borgeaud、Ser-tan Girgin、Sholto Douglas、Shree Pandya、Siamak Shakeri、Soham De、Ted Klimenko、Tom Hennigan、Vlad Feinberg、Wojciech Stokowiec、Yu hui Chen、Zafarali Ahmed、Zhitao Gong、Tris Warkentin、Ludovic Peran、Minh Giang、Cl´ement Farabet、Oriol Vinyals、Jeff Dean、Koray Kavukcuoglu、Demis Hassabis、Zoubin Ghahramani、Douglas Eck、Joelle Barral、Fernando Pereira、Eli Collins、Armand Joulin、Noah Fiedel、Evan Senter、Alek Andreev 和 Kathleen Kenealy。Gemma：基于 Gemini 研究与技术的开放模型。arXiv预印本 arXiv:2403.08295，2024年。2[51]Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timoth´ee Lacroix、Baptiste

Rozi`ere、Naman Goyal、Eric Hambro、Faisal Azhar、Aure-lien Rodriguez、Armand Joulin、Edouard Grave 和 Guil-laume Lample。Llama：开放且高效的基础语言模型。*arXiv*预印本 *arXiv:2302.13971*，2023年。2[52] Heng Wang、Jianbo Ma、Santiago Pascual、Richard Cartwright 和 Weidong Cai。V2A-Mapper：通过连接基础模型实现视觉到音频生成的轻量级解决方案。载于*AAAI*人工智能大会论文集，2024年。5[53] Yazhou Xing、Yingqing He、Zeyue Tian、Xintao Wang 和 Qifeng Chen。看见与听见：借助扩散潜在对齐器实现开放域视觉-音频生成。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，第7151–7161页，2024年。2、5、6[54] Lin Zhang、Shentong Mo、Yijing Zhang 和 Pedro Morgado。音频同步视觉动画。*arXiv*预印本 *arXiv:2403.05659*，2024年。2[55] Lin Zhang、Zefan Cai、Yufan Zhou、Shentong Mo、Jinhong Lin、Cheng-En Wu、Yibing Wei、Yijing Zhang、Ruiyi Zhang、Wen Xiao、Tong Sun、Junjie Hu 和 Pedro Morgado。扩展音频同步视觉动画：一种高效训练范式。*arXiv*预印本 *arXiv:2508.03955*，2025年。2