arXiv:2611.32458v1 [cs.SD] 24 Nov 2026

VividAudio：多事件参考条件视频到音频生成

You Li[1,2,3] Dewei Zhou[1,2] Fan Ma[1,2] Dongliang He[3] Yi Yang[1,2,†]

[1]脑机智能国家重点实验室，浙江大学[2]ReLER，CCAI，浙江大学，中国[3]字节跳动，中国 †通讯作者 {uli2000, zdw1999, mafan, yangyics}@zju.edu.cn hedlcc@126.com



图 1：**VividAudio 实现了超越语义视频到音频的参考条件拟音生成**。音频示例可引导同一视觉事件的声学特征；在多事件场景中，不同参考分别控制不同事件。**摘要**

现有视频到音频（V2A）方法在合成高质量音频方面已取得令人瞩目的成果。然而，随着 V2A 朝着更灵活、由用户主导的创作方向发展，现有方法在控制视频中多个事件各自不同的声学特征方面仍存在明显局限。因此，我们提出多事件参考条件视频到音频（Mr-V2A），为每个事件指定特定的音频参考和目标时间段。Mr-V2A 面临三项关键挑战。首先，事件级参考与目标训练配对十分稀缺。为此，我们提出视频内事件变化（IntraEvtVar），通过在每个视频中挖掘事件的四种变化模式，构造非预言配对。其次，我们将参考组织为时间对齐复合参考（TACO-Ref），并通过局部和全局路径注入：局部路径保留随时间变化的事件线索，而音频 Q-Former 则提供片段级声学条件。第三，同时以多个参考作为条件可能导致声学属性在不同事件间泄漏。局部化参考注意力和 TACO-Ref 路由器（TRR）明确约束事件与参考的对应关系，以减少此类干扰。在 VGGSound-Ref 和 VividBench 上进行的实验表明，与现有 V2A 方法及基于示例的方法相比，该方法具有更强的参考可控性，同时保持了有竞争力的拟音生成质量。

1 引言

生成模型的最新进展（Wan 等，2025年；Yang 等，2024年；BlackForest，2024年；Rombach 等人，2022年）已实现逼真的图像和视频合成，而视频到音频（V2A）模型（Cheng 等，2025；Shan 等，2025年；Wang 等，2025年；Li 等，2025年；Liu 等，2025年b）在合成与视觉内容在语义和时间上保持一致的拟音方面取得了显著进展。

随着 V2A 朝着更加灵活、由用户指导的创作方向发展，仅生成合理的拟音已不再足够：创作者可能还希望指定应当出现<\style id='1'>哪些声音事件、应当<\style id='3'>何时出现，以及每个事件应当听起来<\style id='5'>如何。现有的可控视频到音频方法已实现了这一目标的部分功能。Video-Foley（Lee 等人，2025年）和 FoleyDirector（Li 等，2026年）通过事件条件、结构化脚本和时间线索提升了语义和时间可控性，而 AC-Foley（Fang 等人，2026年）则在单事件拟音生成中引入参考音频，以实现声学可控性。然而，现实世界的视频通常包含多个事件，这些事件在同一条同步音轨中需要具有不同的声学身份。现有方法在这种多事件场景下仍存在明显局限。

Mr-V2A。因此，我们提出**多**事件**参**考条件视频到音频（Mr-V2A）。给定一个视频，以及一组与目标时间段配对的音频参考，Mr-V2A

会生成同步音轨，其中每个事件都遵循指定参考的声学特征，如图 1 所示。这一设定带来了三项挑战。**1）数据局限**。 现有数据集很少提供（视频、事件级参考、目标音频）三元组，而将目标音频本身用作参考则会造成预言机捷径。**2）多音色表示与注入**。如何有效编码多种音色并将其整合到生成模型中，同时不损害视频对齐？**3）音色解耦**。 对多个参考进行联合条件控制可能导致声学属性在事件之间泄漏，使某个事件遵循错误的参考。

我们的方法。 为应对这些挑战，我们提出了 **VividAudio**，包含以下组件。 ***i)***视频内事件变化（*IntraEvtVar*）。从同一视频中挖掘事件，并通过组替换、速度变化、混合和声学退化构建非预言参考-目标对。 ***ii)*** 双路径 *TACO-Ref* 注入。 我们将事件参考置于其目标位置，形成**时**间**对**齐**复**合**参**考（TACO-Ref）。其经过时间分辨处理的特征和经 Audio-Q-Former 聚合的特征分别通过局部和全局路径注入，以建模事件特定声学身份和共享声学上下文。 ***iii)*** 属性解耦。 局部化参考注意力限制无关的时间交互，而 TACO-Ref 路由器（TRR）则将每个事件特定响应显式路由至其对应的时间跨度，从而减少跨事件泄漏。

基准测试。 我们建立了两个互补的基准测试。VGGSound-Ref 是一个带有事件标注的 VGGSound 留出数据集划分，用于评估训练分布下的参考条件拟音音效生成。VividBench 评估更贴近现实的检索场景：我们对 50 个视频中的事件类别和时间跨度进行人工标注，从 Freesound 检索外部参考，并衡量生成的每个事件是否遵循其指定参考。

评估。 在 VGGSound-Ref 和 VividBench 上进行的实验表明，与纯 V2A 和采用单一嵌入示例的基线方法相比，VividAudio 显著提升了参考可控性。在 VGGSound-Ref 上，它将 BEATs-Cos/CLAP-Cos 提升至 $0.695/0.629$ ，同时在 $\mathrm{FAD}_{\mathrm{VGG}}$/ $\mathrm{KL}_{\mathrm{PANNs}}$ $(0.839/1.62)$ 指标上保持了有竞争力的质量。在 VividBench 上，随着声学身份数量增加，事件级增益愈发显著。

我们的主要贡献总结如下。

- 我们提出 Mr-V2A，将参考条件拟音生成从单个示例扩展到对同一同步音轨中多个事件进行事件特定声学控制。
- 我们提出 VividAudio，通过 IntraEvtVar、双路径 TACO-Ref 注入、局部参考注意力和 TRR，解决监督数据稀缺、多尺度参考条件以及跨事件干扰问题。
- 我们开发了一套数据构建流程，并推出 VGGSound-Ref 和 VividBench，展现出更优且均衡的多事件参考可控性，同时具备具有竞争力的生成质量。

2 相关工作

2.1 视频到音频

视频到音频（V2A）通过引入视觉语义和时间结构扩展了音频生成。FoleyCrafter（Zhang 等，2024年）、Diff-Foley（Luo 等人，2023年）、Frieren（Wang 等，2024年）、VTA-LDM（Xu 等，2024年）和 V-AURA（Viertola、Iashin 和 Rahtu，2025年）通过多模态扩散或基于流的融合生成同步声音。Multi-Foley（Chen 等，2025年）和 MMAudio（Cheng 等，2025）结合文本、音频和视频条件，实现灵活合成；与此同时，ThinkSound（Liu 等人，2025b）、HunyuanVideo-Foley（Shan 等，2025年）和 Kling-Foley（Wang 等，2025年）推动 V2A 朝向推理或统一多模态生成发展。然而，大多数系统仍依赖全局文本和视觉证据。尽管这些信息对于语义对齐十分有效，但在指定音色、材质、录音条件和场景氛围方面仍然能力有限。

DreamFoley（Li 等人，2025年）进一步探索了借助视觉语言模型进行自回归视频到音频生成，通过离散音频标记联合建模视频、音频和文本。

2.2 可控视频到音频

除了无条件或文本条件的视频到音频生成之外，近期方法还提供了对拟音生成的更直接控制。Video-Foley（Lee 等人，2025年）和 FoleyDirector（Li 等人，2026年）规定应出现哪些声音事件以及应在何时进行渲染，而 MultiSoundGen（Yang 等人，2025年）、SelVA（Lee、Nam 和 Lee，2026年）以及 Foley-Flow（Mo 和 Song，2026年）则改进了多事件生成、选择性声源合成和语义节奏同步。这些方法增强了对生成什么以及何时生成的控制，但并未明确通过参考音频控制每个事件的声学特征。音频条件视频到音频生成进一步研究了声音应呈现怎样的听感。Tri-Ergon（Li 等人，2024年）使用听觉提示控制全局风格和响度，FoleyGenEx（Wang 等人，2026年）在统一的多模态框架中引入参考音频注入，而 AC-Foley（Fang 等人，2026年）则使用音频示例进行声学迁移。然而，它们的音频条件主要被用作全局风格或声学提示，其中音色、事件语义和时间模式可能仍然相互纠缠。相比之下，Mr-V2A 为各个独立事件分配不同的参考音频，并通过时间局部化参考路由解耦多种声学身份。

2.3 相关视觉与媒体任务

视觉生成领域也对细粒度控制进行了研究。MIGC（Zhou 等，2024b）和 MIGC++ （Zhou 等，2024a）研究了多实例图像生成，而 RefineAnything（Zhou 等，2026）则研究了在背景保留条件下的特定区域图像细化。生成模型的互补用途还包括 AnySynth（Li、Ma 和 Yang，2024）中用于视觉语言任务的合成数据构建，以及 Imagine and Seek（Li、Ma 和 Yang，2025）中用于组合图像检索的想象代理。在媒体制作领域，RoughCut-Bench（Zhao 等，2026）评估视觉语言模型在自动视频粗剪和

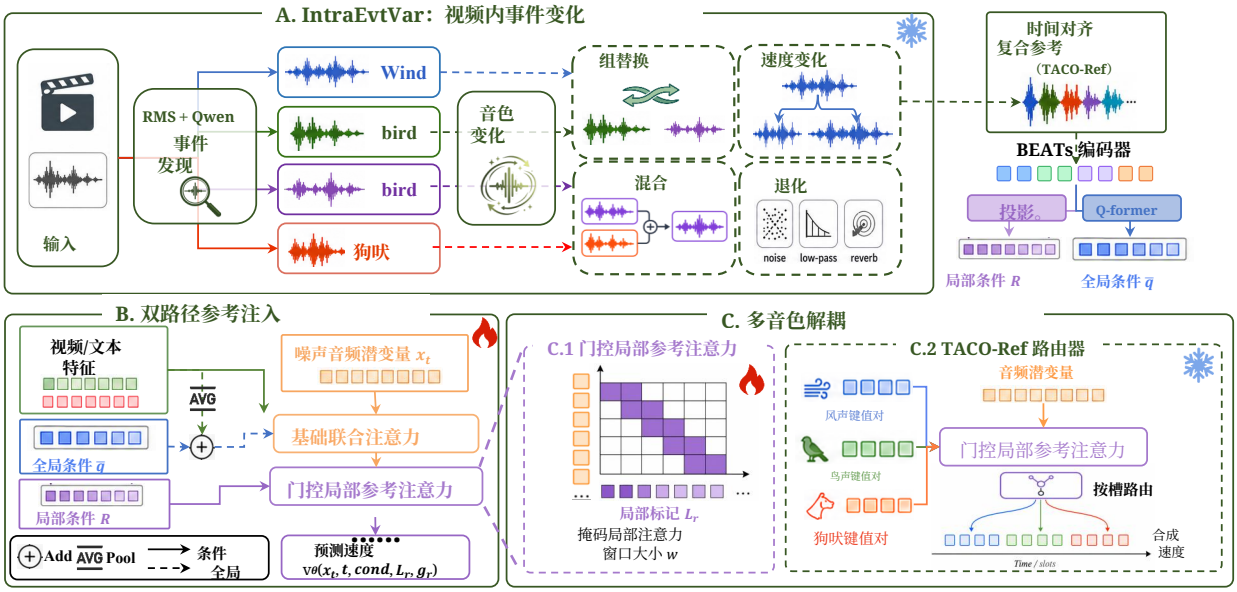

图2：**概览**。 我们首先发现每段视频中的音频事件，并探索四种模式来构建参考-目标对。随后，通过双路径将生成的 TACO-Ref 注入预训练的视频转音频 DiT。局部参考注意力和 TACO-Ref 路由器（TRR）可将事件特定声学属性解耦。

结构化剪辑规划方面的表现。这些任务为可控媒体创作提供了更广泛的背景；我们的重点则是视频同步音轨中的事件特定声学身份。

3 方法

3.1 定义与概述

定义。 给定一个视频 V、一个可选的文本提示 c 以及一组事件级参考条件 $\mathcal{R} = \{(a_k^r,\ \tau_k)\}_{k=1}^K$，其中 a_k^r 是参考音频片段，τ_k 指定了应表现其声学特征的目标时间段，VividAudio 会生成一个波形 $\hat{a}$，使其与视觉事件保持同步，同时在每个对应的时间段内保留所分配的参考音色。我们基于 MMAudio 风格（Cheng 等，2025）的多模态 DiT，并在 44 kHz VAE 的潜在空间中使用整流流匹配对其进行训练：

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{\mathbf{x}_0,\mathbf{x}_1,t}\left[\|v_\theta(\mathbf{x}_t, t, \mathbf{c}) - (\mathbf{x}_1 - \mathbf{x}_0)\|_2^2\right],$$

其中 $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$, ， $\mathbf{c}$ 表示视频、文本和参考条件。视频特征来自 CLIP（Radford 等，2021）和 Synchformer（Iashin 等，2024）；音频由 BigVGAN（gil Lee 等，2023）解码。

概述。 VividAudio 由三个部分组成。首先，由于现有的 MR-V2A 训练数据集均未提供同时配有真实音频和相应参考声音的视频，第 3.2 节在每个视频中挖掘声音事件，并采用四种变化模式构建非 oracle 参考-目标对。其次，第 3.3 节将得到的事件参考整理为**时间**对**齐**合成**参**考（TACO-Ref），并将其集成到 DiT 中以进行参考引导生成。第三，由于注入合成参考特征会引入跨事件声学属性泄漏，第 3.4 节采用局部参考注意力，并结合 TACO-Ref 路由器（TRR），以解耦单个声音的事件特定声学属性。

3.2 IntraEvtVar：视频内事件变化

如图 2A 所示，Mr-V2A 需要事件级参考-目标音频配对，但现有数据集中很少有这类配对。由于同一视频中的多个声音事件是在相似的声学环境和录音条件下录制的，因此在同一视频内构建这类配对更为自然。然而，直接将目标事件用作参考会引入预言机捷径，使模型能够复制目标特定信息，而不是学习声学属性。因此，我们利用视频内事件变化来构建多样化的非预言参考-目标对。

事件发现。 我们首先识别每个视频中的声音事件。均方根能量检测器会去除长时间静音或低能量区域，并将剩余音频划分为候选事件片段。随后，Qwen2.5-Omni（Xu 等，2025 年）会分析每个视听片段，并生成声音类别、细粒度描述和置信度分数。标注置信度较低的事件将被丢弃。

音色变化。 对于每个保留的事件 E_i，我们使用四种变化模式构造其参考音频：**i) 组替换**。 我们在 CLAP-HTSAT（Wu 等人，2023）音频嵌入空间中，使用 K 均值对同一视频中的所有事件进行聚类，并通过肘部准则自动选择聚类数量。与 E_i 被分配到同一聚类的事件构成候选邻居集 $\mathcal{N}(E_i)$，随后可从中采样一个语义相关但并不完全相同的事件，作为其音色参考。**ii) 速度变化**。 选定的参考事件将按一个因子进行重采样

该因子从 0。7、 0。8、1。2、 1。5 中采样，在保留其底层声学特征的同时，引入受控的时间和音高变化。**iii) 混合**。 从训练批次中的另一个样本随机采集一个无关波形，并以从 $[-5, 15]$ dB 均匀采样的信噪比与参考音频混合，以模拟受到干扰声音污染的参考音频。**iv) 声学退化**。随后通过添加加性噪声、低通滤波或短时混响进一步扰动参考音频，以模拟不完美的录音和检索条件。**v) 参考音频丢弃**。 每个事件的参考音频可能以 p_{empty} 的概率被静音替代。上述增强模式可以组合使用，以构造多样化的参考–目标对，并在第 4.3 节评估各模式的单独贡献。

3.3 TACO-Ref 编码与双路径注入时间对齐复合参考。

我们将经过扰动的参考事件放置在共享静音波形中各自对应的目标时间位置上，形成**T**时间-**A**对齐**CO**复合**Ref**参考（TACO-Ref）。当某个参考被采样为空时，其分配的时隙保持静音，并提供空音色条件。随后，冻结的 BEATs 音频编码器（Chen 等，2022）从完整的 TACO-Ref 中提取按时间索引的特征序列$\mathbf{R} \in \mathbb{R}^{T_r \times 768}$，保留多事件条件的时间组织结构。

双路径注入。 在图 2B 中，我们通过$\mathbf{R}$ 双路径进行注入。局部路径保留其时间分辨率，并向音频潜变量提供特定于事件的参考线索；而全局路径则将完整的 TACO-Ref 聚合为共享的声学上下文条件。

局部路径：时间对齐的参考注入。 我们将 TACO-Ref 特征投影到 DiT 隐空间中，并在原始潜变量–视频–文本联合注意力之后引入一个额外的参考注意力模块。给定音频潜变量$\mathbf{h}_a$，我们得到查询$\mathbf{Q}_a = \text{Proj}_q(\mathbf{h}_a)$，并根据$\mathbf{R}$计算参考键和值：

$$\mathbf{A}_{\text{ref}} = \text{softmax}\left(\frac{\mathbf{Q}_a \mathbf{K}_r^\top}{\sqrt{d}}\right)\mathbf{V}_r,$$
$$(\mathbf{K}_r, \mathbf{V}_r) = \text{Proj}_{kv}(\mathbf{R}).$$

所得的更新被加到音频潜变量上：

$$\mathbf{h}'_a = \mathbf{h}_a + \gamma_{\text{ref}}\mathbf{A}_{\text{ref}}\mathbf{W}_o,$$

其中，γ_{ref} 初始化为零。保留$\mathbf{R}$ 的时间结构，使后续的局部注意力和 TRR 模块能够将每个音频潜变量区域与其指定的参考事件关联起来。

全局：声学上下文注入。 与此同时，音频 Q-Former 使用 $N_q = 32$ 个可学习查询和 $L = 8$层对完整的 TACO-Ref 进行聚合：

$$\mathbf{Q} = \text{QFormer}(\mathbf{R}) \in \mathbb{R}^{N_q \times d_h}, \qquad \bar{\mathbf{q}} = \frac{1}{N_q}\sum_{i=1}^{N_q}\mathbf{Q}_i.$$

我们将 $\bar{\mathbf{q}}$投影到池化视频–文本条件空间：

$$\tilde{\mathbf{c}}_{v,t}^{\text{pool}} = \mathbf{c}_{v,t}^{\text{pool}} + \mathbf{W}_q\bar{\mathbf{q}}.$$

更新后的全局条件通过 DiT 的自适应归一化层注入，使模型能够调节各事件槽位共享的声学属性，例如背景纹理和整体频谱特征。

3.4 多音色解耦

当 TACO-Ref 包含对多个声音事件的参考时，不受限制的参考注意力可能会混合跨越事件边界的特征，导致不同事件之间的声学属性发生非预期泄漏。因此，我们对事件特定声学属性进行解耦：局部参考注意力将每个音频查询限制在附近的时间邻域内，而 TACO-Ref 路由器则根据软时间槽位掩码，将每个事件特定的参考响应路由至其分配的音频区域，从而建立显式的事件到槽位的对应关系。

局部参考注意力。 如图 2 C.1 所示，允许每个音频潜变量在整个 TACO-Ref 上进行全局注意，可能会使其接触到来自时间上不相关事件的参考特征。因此，我们施加一个局部注意力窗口，使参考条件与对应的音频区域对齐。具体而言，对于位置为 i 的音频潜变量查询，仅可访问 $\pm w$帧范围内的参考键和值：

$$M_{ij} = \begin{cases} 0, & |i-j| \le w, \\ -\infty, & |i-j| > w. \end{cases}$$

T局部参考注意力随后计算为 ted as

$$\mathbf{A}_{\text{loc}} = \text{softmax}\left(\frac{\mathbf{Q}_a \mathbf{K}_r^\top}{\sqrt{d}} + \mathbf{M}\right)\mathbf{V}_r.$$

训练期间，我们使用固定的窗口大小 w=50，约对应 $\pm$0.5 秒。在推理阶段， w 可在无需重新训练的情况下调整，以控制属性解耦的粒度：较小的窗口会加强时间隔离，而较大的窗口则允许更广泛的参考上下文。

TACO-Ref 路由器（TRR）。局部参考注意力能够抑制时间上相距较远的参考帧所造成的干扰，但固定的局部窗口仍可能包含同时发生或时间间隔很短的事件。因此，如图 2 C.2 所示，我们引入 TACO-Ref 路由器（TRR），在参考特征与音频潜在区域之间建立显式的事件到槽位的对应关系。对于每个事件 k，TRR 仅将其参考片段放置在指定的目标跨度处，并对生成的逐事件画布进行编码，从而构建一个事件专属 TACO-Ref 库，得到 $\mathbf{R}_k$。给定音频潜变量查询$\mathbf{Q}_a$，TRR 会计算事件专属参考注意力：

$$(\mathbf{K}_k, \mathbf{V}_k) = \text{Proj}_{kv}(\mathbf{R}_k), \qquad \mathbf{A}_k = \text{Attn}(\mathbf{Q}_a, \mathbf{K}_k, \mathbf{V}_k).$$

令$\mathbf{m}_k^a(t)$表示事件 k的音频潜变量槽位掩码，该掩码通过将其目标时间段投影到潜变量时间网格上得到。随后，将事件特定响应与槽位掩码融合：

$$\mathbf{A}_{\text{TRR}}(t) = \sum_{k=1}^{K}\mathbf{m}_k^a(t)\mathbf{A}_k(t).$$

模型	#参数量	质量 / 语义					可控性（相对于参考）			
		FAD_{VGG} ↓	$KL_{散度_{PANNs}}$ ↓	IS_{PaSST} ↑	IB-Score ↑	Desync ↓	BEATs-Cos ↑	CLAP-Cos ↑	LSD ↓	MCD ↓
MMAudio-L	1.03B	2.016± 0.081	2.087± 0.015	11.70± 0.09	**0.3468**± 0.0009	0.427± 0.012	0.6254± 0.0017	0.5320± 0.0013	20.06± 0.09	592.8± 2.7
MMAudio-M	621M	0.911± 0.072	2.082± 0.018	10.03± 0.07	0.3199± 0.0022	0.428± 0.005	0.6207± 0.0006	0.5213± 0.0017	20.91± 0.21	610.8± 3.1
混元视频拟音	5.13B	2.485± 0.098	2.562± 0.025	6.69± 0.06	0.3047± 0.0012	0.556± 0.004	0.6192± 0.0011	0.5150± 0.0013	21.11± 0.10	611.4± 2.9
ThinkSound	1.28B	2.424± 0.087	2.410± 0.018	8.67± 0.06	0.2433± 0.0018	0.670± 0.009	0.6078± 0.0007	0.4966± 0.0023	24.15± 0.14	641.3± 5.3
PrismAudio	674M	1.714± 0.035	2.527± 0.022	6.82± 0.08	0.2429± 0.0006	1.117± 0.006	0.6055± 0.0008	0.4945± 0.0016	21.25± 0.16	597.0± 1.6
FoleyDirector	681M	0.866± 0.069	2.071± 0.015	**13.75**± 0.28	0.3124± 0.0017	**0.423**± 0.005	0.6221± 0.0007	0.5180± 0.0029	21.05± 0.56	606.5± 3.9
AC-Foley（随机）	1.08B	4.657± 1.753	1.772± 0.104	8.75± 0.30	0.2551± 0.0165	0.764± 0.047	0.6240± 0.0140	0.4741± 0.0323	26.68± 2.43	519.4± 22.0
AC-Foley（复合事件）	1.08B	5.746± 1.568	2.273± 0.077	8.55± 0.40	0.2331± 0.0096	0.737± 0.016	0.6011± 0.0083	0.4560± 0.0225	27.57± 1.72	597.4± 13.9
AC-Foley（Tail。）	1.08B	**0.701**± 0.047	**1.419**± 0.019	7.88± 0.15	0.3020± 0.0017	0.545± 0.009	0.6637± 0.0013	0.5767± 0.0023	17.93± 0.24	475.3± 2.6
VividAudio（本文方法）	761M	0.839± 0.026	1.618± 0.018	10.66± 0.19	0.2964± 0.0014	0.476± 0.009	**0.6947**± 0.0008	**0.6286**± 0.0013	**16.92**± 0.04	**460.2**± 1.6

表1：VGGSound-Ref 上的主要对比。左侧区块：标准拟音生成质量和语义指标；右侧区块：根据指定参考计算的参考可控性指标。

模型	2-事件				3-事件			
	平均值 ↓	最差值 ↓	差距 ↓	DS ↓	平均值 ↓	最差值 ↓	差距 ↓	DS ↓
V-Foley	672.6	779.8	214.6	1.320	718.1	884.7	336.6	1.252
FoleyDir.	706.1	799.6	187.0	**0.945**	759.2	916.2	303.9	0.940
AC-Foley-S	544.6	712.3	335.4	1.021	640.1	867.3	463.7	1.010
AC-Foley-C	608.8	733.2	248.9	1.008	654.7	883.2	478.1	1.038
Ours	**512.6**	**578.0**	**130.7**	1.006	**504.8**	**599.0**	**185.6**	**0.932**

表 2：VividBench 在双事件和三事件组合下的多事件参考控制。V-Foley 和 FoleyDir. 分别表示 Video-Foley 和 FoleyDirector；AC-Foley-S 和 AC-Foley-C 分别表示单参考和复合参考的 AC-Foley。Avg 表示事件级梅尔倒谱距离的均值，Worst 表示每种情况下最差事件的 MCD 平均值，Gap 表示每种情况下的 MCD 范围，DS 表示 Desync。

Exp	局部	全局	FAD_{VGG} ↓	KL ↓	BEATs-Cos ↑
①	✓	–	0.977	1.98	0.652
②	–	Q-former	1.614	2.14	0.669
③	✓	mean	1.009	1.91	0.652
④	✓	Q-former	**0.839**	**1.62**	**0.695**

表3：VGGSound-Ref 上双路径注入的消融实验。

掩码在事件边界处较为平滑，从而允许相邻槽位之间平滑过渡。通过这种方式，局部注意力在每个事件内提供细粒度的时间局部性，而 TRR 则使各音频区域倾向于其指定的参考事件，并减少声学属性泄漏到无关区域的情况。**更多详情见附录**。

4 实验

4.1 数据集、指标和基线方法

训练数据。我们在 VGGSound（Chen 等人，2020 年）、AudioCaps（Kim 等人，2019 年）和一个内部长视频语料库上进行训练。每个语料库都被转换为面向 IntraEvtVar 的事件级参考-目标样本，具体实现细节见补充材料。

基准测试。**1）VGGSound-Ref**。该基准使用第 3.2 节介绍的方法，从 VGGSound 测试集的留出部分构建，并将视频内不重叠的事件片段作为参考

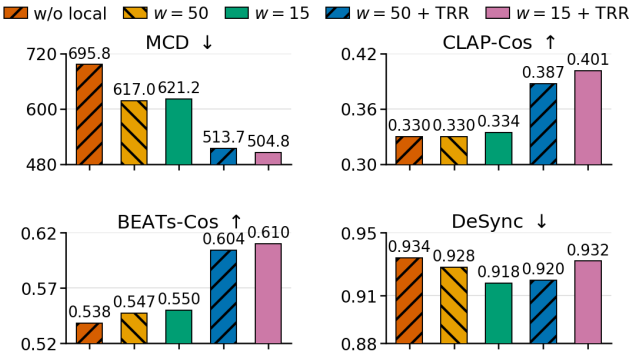

图3：VividBench 中局部参考注意力和 TACO-Ref 路由器（TRR）的三事件消融实验。

。这样可以避免使用目标事件本身作为参考这一预言机捷径。它评估 VGGSound 分布下的参考条件拟音生成，并支持标准生成指标和事件级声学身份指标。**2）VividBench**。为评估更贴近现实参考条件下的可控性，我们人工挑选了 50 个视频，并标注其声音事件类别（每个视频含三种音色）和时间跨度。对于每个事件，我们从 Freesound 检索一个外部音频参考（ Fonseca 等，2017）。详情请参阅补充材料。

指标。我们在 VGGSound-Ref 上报告 FAD_{VGG} （Kilgour 等，2019；Gemmeke 等，2017）、KL_{PANNs} （Liu 等人，2023；Kong 等，2020）、IS_{PaSST} （Salimans 等，2016；Koutini 等，2022）、IB-Score（Girdhar 等，2023）和 Desync（Iashin 等，2024），用于评估生成质量和视频-事件对齐。参考可控性通过 BEATs-Cos、CLAP-Cos、LSD（对数频谱距离，即生成音频与参考音频之间的低层频谱偏差）以及 AC-Foley 中使用的 MCD 进行测量；在 Vivid-Bench 上，声学指标基于标注目标区间在事件级别进行计算。

基线方法。我们比较了以下方法：MMAudio-L/M、混元视频拟音、ThinkSound、Pris-mAudio（Liu 等人，2025a）、FoleyDirector 和 AC-Foley。此外，我们还在 VividBench 上与 Video-Foley 进行了比较。*Tri-Ergon* 和 *FoleyGenEx* 未被纳入，因为其官方实现尚未公开。对于

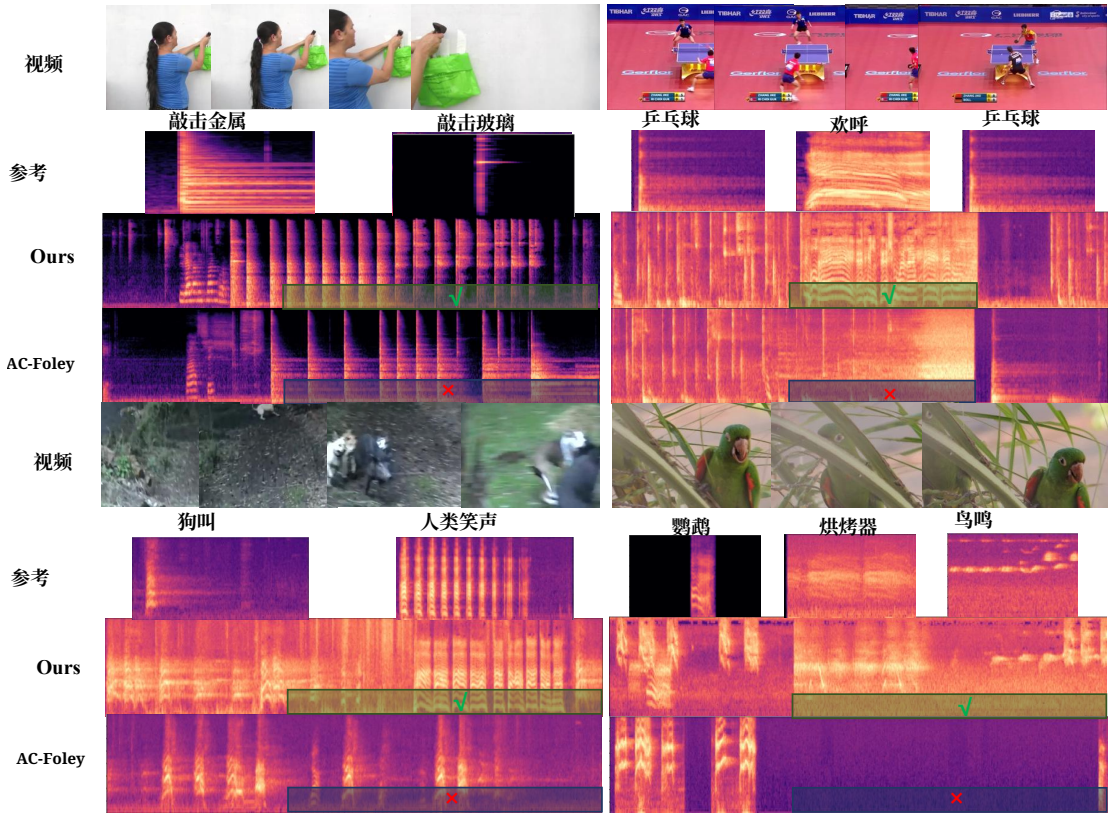

图4：**在 VividBench 上的定性比较**。从上到下，各行分别展示视频帧、事件级参考频谱图、VividAudio 的输出以及 AC-Foley 的输出。

公平比较，我们在多种参考构建设置下评估 AC-Foley。在 VGGSound-Ref 上，我们报告以下结果：1）默认设置，记为 *Tail*，使用每段 10 秒真实音频片段最后 2 秒作为参考条件；2）随机。随机选择一个事件参考；3）*Comp*。按照时间顺序连接所有参考，形成复合参考。在 VividBench 上，由于无法获得真实音频，*Tail* 设置不适用。我们评估两种设置：*Sing*。仅使用与第一个事件关联的参考，以及 *Comp*。按照时间顺序连接所有事件参考。

实现。骨干网络是一个中型 MMAudio DiT，位于 44 kHz 音频自编码器的潜在空间中。训练在八个 40GB 图形处理器上进行，使用 1×10^{-5} 的学习率、余弦学习率调度、有效批量大小 48，以及 4×10^{5} 个监督微调步数。推理使用 $s = 4.5$ 的引导尺度。所有实验均使用五个随机种子进行：42、43、44、666 和 888。我们将 Q-former 设置为 $N_q = 32$、$L = 8$。

4.2 主要对比

VGGSound-Ref。在表1中，VividAudio实现了最强的参考可控性，BEATs-Cos为0。695，CLAP-Cos为0。629，LSD为16。92，MCD为460。2。同时，其生成质量也具有竞争力，FAD_{VGG} 为0。839，KL_{PANNs} 为1。618，同时保持了时间对齐，Desync为0。476。AC-Foley仅在其默认参考设置下取得了具有竞争力的表现，这表明其对事件级参考控制的泛化能力有限，也凸显了我们的方法在处理多种不同音色方面的优势。

VividBench。表 2 使用事件级梅尔倒谱距离报告了双事件和三事件设置下的结果。**1）最强的控制能力**。VividAudio 在两种设置下都取得了最低的平均 MCD（512.6/504.8）。**2）稳健的多事件音色控制**。它还将最差事件的 MCD 降至 578.0/599.0，并取得最低的案例内梅尔倒谱距离差（130.7/185.6），这表明其能够对各个事件进行一致控制，而非仅在某个简单事件上取得提升。相比之下，AC-Foley 的 MCD 差距显著更大，反映出其控制多个事件特定音色的能力有限。**3）有效的时间对齐**。VividAudio 取得了最佳的三事件 Desync 得分 0。932，而 FoleyDirector 在双事件设置下表现最佳，得分为 0。945。

定性分析。图4展示了两个具有代表性的 VividBench 案例。**1）多事件可控性**在动物事件和异质户外场景案例中，VividAudio 都能更稳定地保留指定参考对象在连续、重复及异质事件中的声学特征。相比之下，AC-Foley 的逐事件对应关系较弱，尤其是在

采样概率	KL ↓	IB-Score ↑	BEATs-Cos ↑	LSD ↓
(1，0，0，0，0)（分组）	1.77	0.2740	0.682	17.43
(0，1，0，0，0)（速度）	1.65	0.2775	0.697	17.06
(0,0,1,0,0)（混合）	1.96	0.2347	0.729	16.51
(0,0,0,1,0)（度）	1.97	0.1864	**0.756**	**15.87**
(0，0。8，0，0，0。2)	1.64	0.2842	0.693	17.27
(0.2,0.2,0.2,0.2,0.2)	1.80	0.2838	0.697	17.36
(0.4,0.4,0,0,0.2)	**1.62**	**0.2964**	0.695	16.92
(0.4,0.4,0,0.1,0.1)	1.73	0.2852	0.704	16.79
(0。4，0。4，0。1，0，0。1)	1.77	0.2828	0.699	17.08

表4：IntraEvtVar 消融。采样概率为（p_{group},p_{speed}, p_{mix}, p_{deg},{v13>），分别对应四种变化模式和空丢弃。

高亮区间内。**2）并非简单复制参考内容**。生成的频谱图并非只是复现或拼接参考音频。这一观察也证明了我们采用的非预言式参考构建方法是有效的。

4.3 消融研究

TACO-Ref 特征注入。 表3评估了局部和全局 TACO-Ref 注入路径。加入池化均值条件后，KL散度从 1.98 提升至 1.91，同时 FAD_{VGG} 相近（0.977 对比 1.009）。将均值池化替换为 AudioQ-Former 后，FAD_{VGG} 进一步提升至 0.839，BEATs-Cos 则从 0.652 提升至 0.695。关于 Q-former 的更多消融结果见补充材料。

多音色解耦。 图3评估了 Vivid-Bench 三事件划分上的局部参考注意力和 TRR。首先，移除局部注意力窗口会导致最高的 MCD，以及最低的 CLAP-Cos 和 BEATs-Cos 分数。引入局部参考注意力可显著提升音色相似度，而 TRR 则在不明显降低音频-视频对齐的情况下带来进一步提升。其次，在将训练窗口大小固定为 50 的情况下，我们研究了推理窗口大小 w的影响。结果表明，局部注意力是性能提升的主要来源，而在推理阶段调整 w 只能带来有限的额外收益。

路径分析。 我们通过在 VividBench 的三事件划分上进行成对反事实干预，分析局部和全局参考条件的作用。对于每种情况，我们将相同的三个参考事件从原始槽位顺序 (A, B, C) 循环重新分配到 (B, C, A)，分别仅在全局条件中进行（GS）、仅在局部条件中进行（LS），或同时在两种条件中进行（BS）；同时固定视频、文本提示、随机种子、参考集合和目标事件区间。图 5(a) 中，基线呈现出预期的对角线模式，

这意味着每个生成槽位都最接近最初分配给它的参考。仅交换全局条件会使这种对角结构基本保持不变，这表明仅靠全局路径无法决定为每个槽位选择哪个事件级音色。相比之下，交换局部条件或同时交换两条路径会将最近参考模式转变为循环非对角结构。

然后，我们固定所有局部事件参考，仅替换

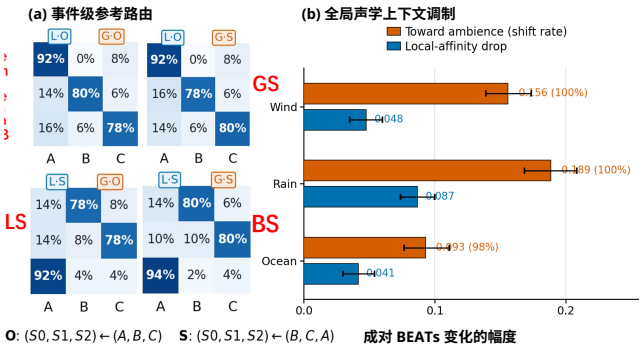

图5：VividBench 三事件划分上的因果路径分析。**(a)** 在循环参考重新分配下，O/S 表示原始/交换后的槽位分配，L/G 表示局部/全局路径。热图通过 MCD 报告每个生成槽位对应的最近参考事件。**(b)** 在固定局部参考的情况下，通过 BEATs 向环境声的偏移以及局部参考亲和度损失来衡量全局环境声注入。误差条表示自助法95%置信区间；括号表示跟随环境声的比率。

全局条件，将其替换为一段 8 秒的风、雨声或海洋环境声录音。图 5(b) 显示，与局部参考亲和度随之下降的程度相比，每种环境声都会使生成结果明显更大幅度地向注入的全局上下文偏移。在 98%–100% 的情况下，生成的声轨都会趋向于注入的环境声。这些结果表明，全局条件是一条片段级声学上下文路径，能够控制共享背景纹理和频谱特征，同时不会覆盖事件级分配。

视频内事件变化。 在表4中，组替换和速度变化提高了音色一致性，同时基本保持了生成质量。相比之下，混合和声学退化带来了更大的音色一致性增益，但也引入了明显的质量退化。随机丢弃部分事件参考有助于保持模型原有的质量和语义分布，因为这能防止模型过度依赖参考条件。结合互补模式可提高参考多样性，并在音频质量与音色一致性之间实现更好的平衡。

4.4 MLLM偏好评估

我们使用 Gemini 2.5 Pro（Gemini Team2025）在图 6 中对 VividBench 进行基于多模态大语言模型的偏好评估。具体而言，我们从 50 个视频中评估 100 个案例，其中两事件和三事件设置各包含 50 个案例，由表 2 中报告的五种方法生成。我们对方法身份进行匿名处理，并以随机顺序展示候选结果，随后由 Gemini 独立地根据视频–事件对齐、音频质量、参考音色可控性和总体偏好进行排序。VividAudio 在全部四项指标上的平均排名均为最佳。**详情见附录**。

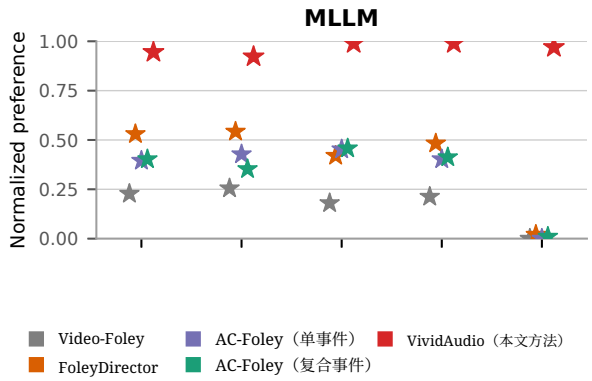

图6：在 VividBench 上进行的多模态大语言模型偏好比较。平均排名被转换为数值越高越好的偏好分数，计算方式为（5 – 排名）/4，而第一名比例仍以 $[0, 1]$ 表示。

5 结论

我们提出了多事件参考条件视频到音频（Mr-V2A）任务，并介绍了 VividAudio，用于控制同步音轨中多个事件的声学身份。针对事件级监督稀缺的问题，IntraEvt-Var 从同一视频中挖掘出的事件构建非预言参考-目标对。随后，TACO-Ref 将参考组织在共享的时间画布上，并通过局部和全局路径注入冻结的 BEATs 特征；与此同时，局部参考注意力和 TRR 通过显式的事件到槽路由减少跨事件属性泄漏。在 VGGSound-Ref 和 VividBench 上的实验表明，与传统视频到音频基线相比，该方法具有更强且更均衡的参考可控性，同时保持了有竞争力的拟音生成质量和时间对齐。消融实验验证了双路径条件控制、局部注意力、TRR 和视频内事件变化的作用。这些结果表明，Mr-V2A 是迈向更灵活的用户定向音频创作的一步。未来工作包括将控制范围扩展到音频示例之外，并将事件特定的参考条件控制扩展至更长、更复杂的视频。

致谢

本研究得到中央高校基本科研业务费（226-2025-00080）、国家自然科学基金委员会（62293554）的资助，并部分得到浙江省自然科学基金（LDT23F02023F02）的资助。

参考

BlackForest . 2024. Black Forest Labs；Frontier AI Lab . Chen, H.; Xie, W.; Vedaldi, A.; and Zisserman, A. 2020. VGGSound：大规模视听数据集。载于 国际声学、语音与信号处理会议（*ICASSP*）。Chen, S.; Wu, Y.; Wang, C.;Liu, S.; Tompkins, D.; Chen,Z.; and Wei, F. 2022. BEATs：使用声学分词器进行音频预训练。arXiv:2212.09058。

Chen, Z.; Seetharaman, P.; Russell, B.; Nieto, O.; Bourgin, D.; Owens, A.; and Salamon, J. 2025年。使用多模态控制的视频引导的拟音音效生成。载于 *IEEE / CVF* 计算机视觉与模式识别会议（*CVPR*）。

Cheng, H. K.；Ishii, M.；Hayakawa, A.；Shibuya, T.；Schwing, A.；以及 Mitsufuji, Y. 2025年。MMAudio：驾驭多模态联合训练，实现高质量视频到音频合成。发表于 *CVPR*。

Fang, P.；He, Y.；Xing, Y.；Chen, Q.；Lim, S.-N.；以及 Yang, H. 2026年。AC-Foley：利用声学迁移实现参考音频引导的视频到音频合成。arXiv:2603.15597。

Fonseca, E.; Pons, J.; Favory, X.; Font, F.; Bogdanov, D.; Ferraro, A.; Oramas, S.; Porter, A.; and Serra, X. 2017. Freesound 数据集：一个用于创建开放音频数据集的平台。载于国际音乐信息检索学会第 *18* 届会议论文集。Gemini 团队. 2025年. Gemini 2.5：凭借先进推理、多模态、长上下文和新一代智能体能力推动前沿发展。*arXiv* 预印本 *arXiv:2507.06261*。Gemmeke, J. F.; Ellis, D. P. W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio Set：一种用于音频事件的本体和人工标注数据集。载于 *2017 IEEE*国际声学、语音与信号处理会议（*ICASSP*），776–780。IEEE 出版社。gil Lee, S.; Ping, W.; Ginsburg, B.; Catanzaro, B.; and Yoon, S. 2023. BigVGAN：一种采用大规模训练的通用神经声码器。arXiv:2206.04658。Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. ImageBind：将万物绑定到同一嵌入空间。载于 *CVPR*。Iashin, V.; Xie, W.; Rahtu, E.; and Zisser man, A. 2024. Synchformer：基于稀疏线索实现高效同步。载于*ICASSP 2024-2024 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE。Kilgour, K.; Zuluaga, M.; Roblek, D.; and Sharifi, M. 2019. Fréchet 音频距离：一种用于评估音乐增强算法的度量。arXiv:1812.08466。Kim, C. D.; Kim, B.; Lee, H.; and Kim, G. 2019. AudioCaps：为真实场景中的音频生成描述文本。载于 *NAACL-HLT*。Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; and Plumbley, M. D. 2020. PANNs：用于音频模式识别的大规模预训练音频神经网络。arXiv:1912.10211。Koutini, K.; Schlüter, J.; Eghbal-zadeh, H.; and Widmer, G. 2022. 使用 Patchout 高效训练音频 Transformer。载于 *Interspeech 2022*，国际语音通信协会第 *23* 届年会，韩国仁川，*2022*年*9*月*18*日至*22*日，2753–2757。ISCA。Lee, J.; Im, J.; Kim, D.; and Nam, J. 2025. Video-Foley：通过时间事件条件实现拟音音效的两阶段视频到声音生成。arXiv:2408.11915。

Lee, J.; Nam, J.; and Lee, J. 2026. 听听哪些内容重要！文本条件选择性视频到音频生成。arXiv:2512.02650。Li, B.; Yang, F.; Mao, Y.; Ye, Q.; Chen, H.; and Zhong, Y. 2024. Tri-Ergon：通过多模态条件和 LUFS 控制实现细粒度视频到音频生成。arXiv:2412.20378。Li, F.; Z hao, W.; Li, Y.; Zhou, Z.; and He, D. 2025. DreamFo-ley：用于高保真视频到音频生成的可扩展视觉语言模型。arXiv:2512.06022。Li, Y.; Ma, F.; and Yang, Y. 2024. AnySynth：利用图像合成数据生成的能力，完成通用视觉语言任务。*arXiv*预印本 *arXiv:2411.16749*。Li, Y.; Ma, F.; and Yang, Y. 2025. Imagine and Seek：借助想象代理改进组合图像检索。载于计算机视觉与模式识别会议论文集（*CVPR*），3984–3993。Li, Y.; Zhou, D.; Ma, F.; Li, F.; He, D.; and Yang, Y. 2026. FoleyDirector：通过结构化脚本对视频到音频生成进行细粒度时间引导。载于*IEEE/CVF*计算机视觉与模式识别会议论文集。Liu, H.; Chen, Z.; Yuan, Y.; Mei, X.; Liu, X.; Mandic, D.; Wang, W.; and Plumbley, M. D. 2023. AudioLDM：利用潜在扩散模型实现文本到音频生成。国际机器学习会议论文集，21450–21474。Liu, H.; Luo, K.; Wang, W.; Chen, Q.; Sun, P.; Huang, R.; Li, X.; Ye, J.; and Xue, W. 2025a. PrismAudio：通过分解式思维链和多维奖励实现视频到音频生成。arXiv:2511.18833。Liu, H.; Wang, J.; Luo, K.; Wang, W.; Chen, Q.; Zhao, Z.; and Xue, W. 2025b. ThinkSound：多模态大语言模型中的思维链推理，用于音频生成与编辑。*arXiv*预印本。Liu, H.; Yuan, Y.; Liu, X.; Mei, X.; Kong, Q.; Tian, Q.; Wang, Y.; Wang, W.; Wang, Y.; and Plumbley, M. D. 2024. AudioLDM 2：通过自监督预训练学习整体音频生成。*IEEE/ACM*音频、语音与语言处理汇刊，32: 2871–2883。Luo, S.; Yan, C.; Hu, C.; and Zhao, H. 2023. Diff-Foley：借助潜在扩散模型实现同步视频到音频合成。arXiv:2306.17203。Mo, S.; and Song, Y. 2026. Foley-Flow：通过掩码音视对齐和动态条件流实现协调的视频到音频生成。arXiv:2603.08126。Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. 通过自然语言监督学习可迁移的视觉模型。载于*ICML*。Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Om-mer, B. 2022. 利用潜在扩散模型进行高分辨率图像合成。Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Rad-ford, A.; and Chen, X. 2016. 改进 GAN 训练技术。arXiv:1606.03498。Shan, S.; Li, Q.; Cui, Y.; Yang, M.; Wang, Y.; Yang, Q.; Zhou, J.; and Zhong, Z. 2025年。混元视频拟音：通过表征对齐实现高保真拟音音频生成的多模态扩散。arXiv:2508.16930。Viertola, I.; Iashin, V.; and Rahtu, E. 2025年。通过自回归为视频生成时间对齐音频。载于*2025*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE。Wan, T.; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; Zeng, J.; Wang, J.; Zhang, J.; Zhou, J.; Wang, J.; Chen, J.; Zhu, K.; Zhao, K.; Yan, K.; Huang, L.; Feng, M.; Zhang, N.; Li, P.; Wu, P.; Chu, R.; Feng, R.; Zhang, S.; Sun, S.; Fang, T.; Wang, T.; Gui, T.; Weng, T.; Shen, T.; Lin, W.; Wang, W.; Wang, W.; Zhou, W.; Wang, W.; Shen, W.; Yu, W.; Shi, X.; Huang, X.; Xu, X.; Kou, Y.; Lv, Y.; Li, Y.; Liu, Y.; Wang, Y.; Zhang, Y.; Huang, Y.; Li, Y.; Wu, Y.; Liu, Y.; Pan, Y.; Zheng, Y.; Hong, Y.; Shi, Y.; Feng, Y.; Jiang, Z.; Han, Z.; Wu, Z.-F.; and Liu, Z. 2025年。Wan：开放且先进的大规模视频生成模型。*arXiv*预印本 *arXiv:2503.20314*。Wang, J.; Zeng, X.; Qiang, C.; Chen, R.; Wang, S.; Wang, L.; Zhou, W.; Cai, P.; Zhao, J.; Li, N.; 等。2025年。Kling-Foley：用于高质量视频到音频生成的多模态扩散Transformer。*arXiv*预印本。Wang, S.; Zeng, X.; Wang, H.; Zhao, S.; Deng, F.; Zhang, C.; and Qin, Y. 2026年。FoleyGenEx：具备多模态控制、时间对齐和语义精确性的统一视频到音频生成。arXiv:2606.14049。Wang, Y.; Guo, W.; Huang, R.; Huang, J.; Wang, Z.; You, F.; Li, R.; and Zhao, Z. 2024年。Frieren：采用整流流匹配的高效视频到音频生成网络。神经信息处理系统进展，37：128118–128138。Wu, Y.; Chen, K.; Zhang, T.; Hui, Y.; Berg-Kirkpatrick, T.; and Dubnov, S. 2023年。通过特征融合和关键词到字幕增强实现大规模对比式语言-音频预训练。载于*IEEE*国际声学、语音与信号处理会议，*ICASSP*。Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; Zhang, B.; Wang, X.; Chu, Y.; and Lin, J. 2025年。Qwen2.5-Omni：全模态大语言模型系列。*arXiv*预印本 *arXiv:2503.20215*。Xu, M.; Li, C.; Ren, Y.; Chen, R.; Gu, Y.; Liang, W.; and Yu, D. 2024年。通过隐式对齐实现视频到音频生成。arXiv:2407.07464。Yang, J.; Yang, X.; Zhang, L.; Guo, X.; Wang, Z.; and Huang, G. 2025年。MultiSoundGen：通过SlowFast对比式音视频预训练和直接偏好优化，实现多事件场景的视频到音频生成。arXiv:2509.19999。Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; 等。2024年。CogVideoX：采用专家Transformer的文本到视频扩散模型。*arXiv*预印本 *arXiv:2408.06072*。Zhang, Y.; Gu, Y.; Zeng, Y.; Xing, Z.; Wang, Y.; Wu, Z.; and Chen, K. 2024年。FoleyCrafter：用逼真且同步的声音让无声视频焕发生机。arXiv:2407.01494。

Zhao, R.; Li, Y.; Chen, S.; Cheng, A.; Ma, C.; Pei, S.; Yang, Z.; Fan, H.; He, D.; and Yang, Y. 2026年。RoughCut-Bench：基于大型视觉语言模型的自动视频粗剪基准。*arXiv*预印本 *arXiv:2611.24098*。Zhou, D.; Li, Y.; Ma, F.; Yang, Z.; and Yang, Y. 2024a。Migc++：用于图像合成的高级多实例生成控制器。*IEEE*模式分析与机器智能汇刊。Zhou, D.; Li, Y.; Ma, F.; Zhang, X.; and Yang, Y. 2024b。Migc：用于文本到图像合成的多实例生成控制器。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，6818–6828。Zhou, D.; Li, Y.; Yang, Z.; and Yang, Y. 2026年。RefineAnything：用于完善局部细节的多模态区域特定细化。*arXiv*预印本 *arXiv:2604.06870*。

A 补充数据构建细节

本节介绍主论文第 3.2 节所述数据流水线的实施细节。我们重点介绍语料处理、事件提取、图序列化和训练三元组导出；关于视频内事件变化增强和参考扰动策略的动机，请参阅主论文的方法部分。

A.1 语料库与事件提取

我们处理 VGGSound（Chen等人，2020年）、AudioCaps（Kim等人，2019年）和一个内部语料库。首先使用 RMS 能量事件检测器对源片段进行分段。自适应能量阈值会移除长时间静音或低能量区域，并将每个保留的事件裁剪为时长介于 0.2 秒和 8.0 秒之间的片段。

随后，将每个视听裁剪片段输入 Qwen2.5- Omni（Xu 等，2025），输出包含声音类别、细粒度描述和置信度分数的结构化元数据。置信度低于 0.6 的事件将被丢弃。其余记录保留源视频标识符、时间边界、音频路径、视觉裁剪片段，以及图构建和三元组生成阶段所需的注释字段。

A.2 特征提取与图序列化

对于每个保留的事件，我们将音频重采样至 48 kHz，并从 CLAP-HTSAT（Wu 等人，2023）中提取 512 维全局池化表示。我们在每个源视频内独立应用 K 均值。我们评估$K \in \{2, \ldots, 8\}$，并根据聚类惯性的局部变化选择肘点。这样，无需全局的类别特定聚类模型，即可为每个事件生成一个组标识符。

聚类完成后，每个事件都会获得两个图字段。第一个是组标识符，用于标记该事件所属的视频内声学簇。第二个是邻居列表，其中包含分配到同一簇的其他事件的索引。这些邻居构成了非预言机参考构建的候选池。如果聚类失败，或某个事件形成单例簇，我们会存储 group=-1 或空邻居列表。这样可使标注格式能够适应稀疏或模糊视频，并避免强行要求每个事件都有同组伙伴。

A.3 邻居记录与三元组导出

对于每个事件 E_i，该流程会将其分组标识符以及所有关联事件的索引序列化为邻居列表：

{ "label": "拟音音效", "desc": "木地板上的脚步声", "group": 2, "neighbors": [4, 7, 12]}

邻居列表由主论文第 3.2 节所述的扰动过程使用。具体而言，导出器从一个扰动系列中构建事件参考，该系列包括同组替换、速度变化、混合、声学退化，以及可选的空参考条件。当存在有效的同组伙伴时，同组替换会从邻居列表中抽取一个不同的事件。在空参考的情况下，不会将波形粘贴到相应时隙中，因此该时隙在 TACO-Ref 中保持静音，并作为明确的零参考条件，而非缺失数据错误。

为便于记录，导出的 JSONL 会保留参考侧和生成侧所用的音频路径及时间跨度。典型记录包含事件级参考所需的 rag_audio_list、rag_seg_in_ref 和 rag_seg_in_gen，以及可选修复/背景区域所需的 inp_audio_list、inp_seg_in_ref 和 inp_seg_in_gen。在最简单的导出器中，如果不存在有效的同组邻居，系统会为该事件回退到自重建，使样本仍可用；后续扰动阶段仍可将该参考转换为非预言或空条件。随后，导出器会将复合参考波形及生成的 (V， a_r， a_1) 路径写入训练 JSONL 文件。相同的元数据格式支持参考特征提取、TACO-Ref 构建以及可复现的样本级记录。

B 基于扩散的音频生成背景

现代音频生成系统通常沿用与图像潜在扩散相同的高级设计：它们在紧凑的潜在空间中学习生成过程，并将生成的潜变量解码回波形音频。早期的文本转音频系统通常在类似频谱图的表示或学习得到的音频潜变量上应用扩散，AudioLDM 表明，潜在扩散模型可以根据文本条件合成逼真的音频（Liu 等人 2023、2024）。视频转音频模型通过加入视觉条件和时间条件扩展了这一框架，因此生成的配乐不仅必须在声学上合理，还必须与输入视频保持同步（Luo 等人 2023；Wang 等，2024；Cheng 等，2025）。

B.1 潜在音频表示

直接生成波形的成本很高，因为音频具有较高的采样率和精细的时间结构。一种常见的解决方案是训练音频自编码器，将波形音频 a映射为低采样率的潜在序列 $\mathbf{x}$，然后从生成的潜在表示中重建波形。在这种设置下，扩散模型或流模型作用于 $\mathbf{x}$，而不是原始采样。这样既能缩短序列长度，又能保留感知上重要的频谱和时间信息。在 VividAudio 中，DiT 主干网络运行于 44 kHz 音频自编码器的潜在空间中，最终波形则通过神经声码器恢复，这遵循了 BigVGAN 等神经声码器所采用的常见音频解码实践（gil Lee 等，2023）。

B.2 扩散与整流流

扩散模型学习逆转一个逐步将干净数据映射为噪声的破坏过程。整流流采用了密切相关的连续时间视角，但训练网络来预测噪声样本$\mathbf{x}_0$与数据潜在表示$\mathbf{x}_1$之间一条简单路径上的速度场。通过线性插值

$$\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1, \qquad t \in [0,1],$$

训练目标是位移 $\mathbf{x}_1 - \mathbf{x}_0$。条件模型 $v_\theta(\mathbf{x}_t, t, \mathbf{c})$ 经过优化，能够根据当前潜变量、时间和条件信号预测该速度：

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{\mathbf{x}_0,\mathbf{x}_1,t,\mathbf{c}} \left\| v_\theta(\mathbf{x}_t, t, \mathbf{c}) - (\mathbf{x}_1 - \mathbf{x}_0) \right\|_2^2 .$$

推理阶段，生成从噪声开始，并沿着学习到的速度场积分，朝数据分布推进。与经典的去噪解释相比，这种表述便于将模型表示为从噪声到音频潜变量的条件轨迹，这是近期高效视频到音频系统所采用的表述（Wang 等，2024；Cheng 等，2025）。

B.3 音频与视频到音频的条件

对于文本到音频生成，条件 $\mathbf{c}$ 通常是文本嵌入。对于视频到音频，$\mathbf{c}$还必须包含视觉语义和时间对齐线索。因此，视频条件音频模型需要判断应该出现什么声音、何时出现，以及应具备何种声学纹理。现有的视频到音频模型主要使用全局视频/文本条件或同步视觉特征来引导潜变量生成器（Luo 等人，2023；Cheng 等，2025）。

VividAudio 将事件级参考音频纳入这一条件接口。视频和文本条件提供语义与时间锚定，而 TACO-Ref 提供参考声学特征。因此，正文中描述的局部和全局参考路径并非独立的生成器，而是额外的条件路径，用于引导同一个潜变量流模型生成所需的事件级和片段级声学属性。

C 补充方法细节

本节详细介绍 VividAudio 的实现细节，这些内容在正文中有所简略。我们重点说明如何将事件级参考转换为 TACO-Ref、局部和全局路径如何利用参考条件，以及存在多个参考时如何分离事件级声学身份。

C.1 TACO-Ref 构建

对于包含 K 个已标注目标事件的训练或评估样本，我们将参考集合表示为 $\mathcal{R} = \{(a_k^r,\ \tau_k)\}_{k=1}^{K}$，其中 a_k^r 是分配给事件 k 的参考波形，而 $\tau_k = [s_k, e_k]$ 是生成声轨中的目标跨度。TACO-Ref 初始化为与目标音频时长相同的静音波形。根据事件构造规则，对每个参考波形进行裁剪或拉伸，然后将其放置在时间位置 τ_k。如果存在多个事件，则将所有事件参考放入同一画布，同时保留其预期的目标时间对应关系。

当某个事件采样到空参考条件时，其分配的槽位将保持静音，而不会填入特殊标记或学习到的嵌入。这为该事件区间提供了明确的零音色条件，并防止模型假定每个事件始终都有有效的声学样本作为支持。空参考条件是在事件级别采样的，因此，多事件示例中的一个事件可以不受条件约束，而其余事件仍可获得有效参考。

C.2 参考扰动细节

IntraEvtVar 使用四种参考侧声学变化模式，以避免目标复制捷径并增加声学条件的多样性。组替换会从同一视频内的声学分组中选取一个不同的邻居样本。速度变化会使用 $\{0.7, 0.8, 1.2, 1.5\}$ 中的因子调整所选参考。混合会加入随机采样信噪比的无关波形，以模拟受到污染的检索参考。声学退化会应用加性噪声、低通滤波和短混响等简单信号扰动。此外，如上所述，一小部分事件参考会留空。这些操作共同构成了参考扰动族，用于在训练期间丰富事件级参考条件；消融表中的混合概率则概述了其整体采样方案。

C.3 双路径参考注入

局部路径使用按时间索引的 TACO-Ref 特征序列，并通过参考注意力将其注入。由于参考画布与目标音频时间线对齐，来自某个音频潜在区域的查询可以关注目标时间附近的参考帧。该路径提供事件特定声学身份线索，并负责在每个事件区间内实现细粒度音色控制。

全局路径使用音频 Q-Former 对完整的 TACO-Ref 进行汇总，并通过自适应归一化注入池化后的条件信息。该路径不会保留事件级身份槽位，而是提供片段级声学上下文，例如录音质感、背景环境声，以及贯穿整个配乐的整体频谱特征。正文中的路径作用分析通过实证支持了这种区分：交换局部参考会改变事件级身份，而替换全局环境声则会改变片段级声学上下文，同时大体保留局部事件分配。

C.4 局部参考注意力与 TRR

对整个 TACO-Ref 施加全局参考注意力，可能使音频查询暴露于无关的参考事件，尤其是在多事件案例中。因此，局部参考注意力会掩蔽每个音频潜在查询周围时间窗口之外的参考帧。该掩码在注意力 softmax 之前应用，因此不可访问的参考位置在归一化后贡献零概率。这一约束减少了来自时间上较远参考的泄漏，同时在允许的窗口内保留可微注意力。

TACO-Ref 路由器（TRR）进一步处理相邻或邻近事件落在同一局部注意力

$(N_q,\ L)$	$\text{FAD}_{\text{VGG}}\downarrow$	$\text{KL}_{\text{PANNs}}\downarrow$	BEATs-Cos ↑
(16, 4)	1.017	1.773	0.676
(32，4)	0.933	1.727	0.681
(32，8)（选定）	**0.839**	**1.618**	**0.695**
(64，8)	0.912	1.749	0.705

表5：用于池化 TACO-参考条件的 AudioQ-Former 尺寸比较。(32，8) 在测试配置中实现了最佳的总体权衡，因此我们在最终模型中采用了该配置。

窗口内的情况。对于每个事件 k，TRR 会构建一个该事件专属的参考画布，其中仅包含该事件在其指定目标跨度处的参考内容。模型会根据每个事件专属画布计算注意力响应$\mathbf{A}_k$，然后使用投影到音频潜变量时间网格上的软槽位掩码$\mathbf{m}_k^a(t)$融合这些响应：

$$\mathbf{A}_{\text{TRR}}(t) = \sum_{k=1}^{K} \mathbf{m}_k^a(t)\mathbf{A}_k(t).$$

我们在事件边界附近对掩码进行软化，以避免突变。实际上，TRR 充当的是一种软性的事件到槽位偏置，而非硬路由规则：每个目标区域都会被鼓励使用分配给相应事件的参考响应，而边界帧则可以在相邻槽位之间平滑插值。

D 补充实验细节

本节介绍正文中有所简略的实现与评估细节。

D.1 训练和推理配置

我们将片段转换为事件级参考-目标样例后，在 VGGSound、AudioCaps 和内部语料库上训练中等规模的 DiT 主干网络。模型使用 SFT 进行 4×10^5 步训练，学习率为1×10^{-5}，采用余弦调度，有效批量大小为 48。音频以 44kHz 表示。除非另有说明，推理时使用引导尺度 $s = 4.5$。

对于多事件参考条件控制，我们将事件参考置于指定的时间槽位中，以形成 TACO-Ref。我们将 TACO-Ref 重采样至 16kHz，并使用在 AS2M（Chen 等，2022）上预训练且冻结的 BEATs iter3+ 模型提取 768 维帧级特征；该编码器在训练和推理期间均保持固定。参考注意力作用于这一时间对齐表示，而经 Audio-Q-Former 聚合的条件则通过自适应归一化路径注入。局部注意力窗口使用 $w = 50$ 帧进行训练，并可在推理时调整；除非另有说明，最终的 VividBench 配置使用 $w = 15$以实现更强的事件级隔离。TACO-Ref 路由器利用事件槽位掩码来组合特定于事件的参考响应。

表 5 显示，将音频 Q-Former 的容量从 (16，4) 提高到 (32，4)，能够持续改善 FAD_{VGG}、KL_{PANNs} 和 BEATs-Cos，表明适度增加查询数量有助于概括复合参考。进一步将模块加深至 (32，8) 可实现整体最佳权衡，因此最终模型采用了这一配置。再次将查询数量增加至 (64，8) 并未带来额外收益，反而使所有三项指标略有下降，这表明，与其单纯扩大池化全局分支，不如保持适中的查询预算并确保足够的深度。

D.2 VGGSound-Ref 与基线协议

VGGSound-Ref 基于留出的 VGGSound 测试集划分构建，采用主论文中描述的相同事件发现与视频内参考构建流程。对于每个目标事件，我们从同一源视频中选取一个与其不重叠的事件片段作为参考条件，而不是重复使用目标事件本身。这样既保留了同一录音内的声学相关性，又避免了以完全相同的目标片段作为条件这一预言机捷径。由此得到的基准仍属于 VGGSound 域内基准，但会明确评估模型能否遵循事件级声学参考，而非仅生成听起来合理的通用拟音音效。

在 VGGSound-Ref 上，我们同时报告完整片段生成指标和参考跟随指标。FAD_{VGG}、KL_{PANNs}、IS_{PaSST}、IB-Score 和 Desync 按照各指标的标准协议，在生成的片段上计算。参考可控性通过比较生成音频与指定参考条件之间的 BEATs-Cos、CLAP-Cos、LSD 和 MCD 进行评估。在 VividBench 上，则使用人工标注的目标跨度，在事件级别评估这些可控性指标。

在计算指标时，VGGSound-Ref 使用标准的 av-benchmark 流程计算完整片段指标，并通过单独的参考跟随评估计算事件条件指标。在 VividBench 上，对于每个事件，我们根据其标注目标区间裁剪生成波形，并根据参考区间裁剪参考波形，然后比较配对片段。CLAP-Cos 基于 48 kHz 波形计算，而 BEATs-Cos 则基于 16 kHz 单声道波形计算，并将令牌轴池化取均值，得到每个片段的一个嵌入。LSD 根据配对波形片段的对数功率谱计算；MCD 则遵循 AC-Foley 评估协议，基于经 DTW 对齐的 MFCC 计算。对于多事件案例，我们先分别计算每个事件的这些指标，再对每个案例中的事件级数值取平均，最后对该划分中的所有案例取平均。每个生成片段的 Desync 指标只针对完整输入视频计算一次，而不是按事件计算。

公平比较。 在进行基线比较时，我们对所有方法采用相同的基准划分和输入视频，但会根据每个基线所支持的输入格式，调整其参考接口。**这对 AC-Foley 尤其重要，因为其原始设计假定使用单参考音轨，因此无法直接编码多个事件特定参考**。 因此，在每个基准测试中，我们在多种参考构建设置下评估 AC-Foley，以涵盖其默认用法，以及信息量更高的单参考和复合参考变体。在 VGGSound-Ref 上，我们报告三种设置：*Tail*，即使用真实片段最后 2 秒作为默认参考条件；

the final 2s of the ground-truth clip as the default reference condition; *random.*，随机选择一个事件参考；以及*Comp.*，按照时间顺序将所有事件参考连接成一个复合参考。在 Vivid-Bench 上，由于没有真实音频，因此不适用 *Tail* 设置。我们改为报告*Sing.*，即仅使用与第一个事件关联的参考；以及*Comp.*，即按照时间顺序连接所有事件参考。这些变体更全面、更公平地展现了 AC-Foley 在事件级和多事件参考条件下的表现。

D.3 VividBench 评估协议

VividBench 基于 VGGSound 视频构建，涵盖人类、动物、自然、动作和音乐五大主题。我们首先检索这些主题下的候选视频，然后手动筛选，选出 50 个具有清晰多事件结构且声源在视觉上可识别的源视频。对于每个选定的视频，我们手动标注候选音色主题及其对应的时间跨度，然后从 Freesound 检索每个已标注事件对应的外部参考波形。基于这些源视频，我们分别评估双事件和三事件组合。对于每个事件，我们使用 CLAP 和 BEATs 余弦相似度，将生成音频片段与指定参考进行比较；报告的分数是相应数据集划分中所有案例的事件级数值的平均值。Desync 根据已标注的视频事件时间计算，并在完整数据集划分上报告。

D.4 定性案例可视化

为完整起见，我们以补充图的形式提供论文中使用的完整定性可视化。该可视化对齐了采样视频帧、事件级参考频谱图和生成频谱图。**您还可以参阅媒体资料，了解更多示例**。

D.5 多模态大语言模型偏好评估

我们在主论文中使用的 100 个 VividBench 案例上评估多模态大语言模型偏好，其中包括 50 个双事件案例和 50 个三事件案例。每个案例都包含原始视频、特定事件的参考声音、目标时间跨度，以及由五种受评方法生成的匿名候选配乐。候选项的身份信息均被隐藏，且每个案例中的候选项顺序都独立随机排列。

我们使用 Gemini 2.5 Pro 作为多模态大语言模型评审，并分别针对四项标准进行查询：视频–事件对齐、音频质量、参考音色可控性和整体偏好。对于每项标准，模型都会在同一案例内对所有匿名候选项进行排序，而不是为其赋予绝对李克特评分。排序可避免不同案例间的校准偏差，也更符合主观偏好评估的比较性质。我们将排名 $r \in \{1, \ldots, 5\}$转换为归一化偏好分数（$5 - r$）$/4$，分数越高越好，同时报告整体第一名比例。报告中的数值是 100 个案例的归一化偏好分数的平均值。

每次排序时，评审都会收到一段时间对齐的复合参考音频，以及候选视频

方法	S_{atmos}（严格）↑	S_{atmos}（滑动）↑
MMAudio	0.603	0.721
VividAudio（本文方法）	**0.688**	**0.774**
Δ（95%置信区间）	$+0.085^{*}$ $[+0.035, +0.135]$	$+0.054^{*}$ $[+0.019, +0.088]$

表 6：在 55 个留出视频上进行带音频上下文的长时段视频到音频生成。两个模型分别为两个不连续的8 秒视频窗口合成音频，并将输出拼接成一段 16 秒配乐。 S_{atmos} 表示时间窗口之间的 CLAP 余弦相似度。 Δ表示配对逐视频提升及其 95% 置信区间。* 表示在配对 t-检验下的 $p < 0.01$。

以随机顺序排列。提示中提供案例元数据，包括事件标签及对应的目标片段，并要求模型按照对齐度、质量、音色控制和总体表现从最佳到最差排列所有候选ID。提示明确指示评审者，将每个目标片段与复合参考中相同时间区间的参考音频进行比较；将音色控制视为对参考音色的粗略感知跟随，而非精确的录音模仿；并避免并列。四个排名维度分别独立提问，以免质量、对齐度和音色控制方面的判断相互干扰。随后，我们将返回的排名列表转换为逐方法排名和 Top-1 指标，并在所有案例中汇总。

D.6 长视频转音频应用

我们使用了 55 个留出视频，每个视频时长至少为 16 s，每个子类别最多抽取五个视频。每个视频被划分为两个不重叠的片段，分别为 $V_A = [0, 8]$ s 和$V_B = [8, 16]$s。两个系统接收相同的视觉输入，并为每个片段合成一个 8 s 的音频片段。MMAudio 独立生成这两个片段。VividAudio 将从 V_A 生成的音频用作V_B 的音频上下文，通过其全局和条件级参考路径生成音频；此外，还将第一个片段最后 2 s 的音频作为边界处尾到头修复前缀。两个生成的片段直接拼接，不进行额外后处理。

我们使用生成的 16 s 配乐中不同时间窗口之间的 LAION-CLAP 余弦相似度来量化氛围一致性。严格分数比较两个不重叠的 8 s 片段（$[0, 8]$ s 和 $[8, 16]$ s）。滑动分数则对三个 8 s 时间窗口（分别从 0、4 和 8 s 开始）之间的两两相似度取平均，因此对单个边界不那么敏感。我们报告每种方法的平均分数，以及配对逐视频差值 $\Delta = S^{\text{Ours}} - S^{\text{MMAudio}}$ 及其 95% 置信区间。观察到的提升为：严格分数$+0.085$，置信区间为 $[+0.035, +0.135]$；滑动分数 $+0.054$，置信区间为 $[+0.019, +0.088]$；两项配对检验均满足$p < 0.01$。

E 训练数据标注摘要

本补充章节记录了 VividAudio 使用的训练时标注资源。我们的监督微调流程读取

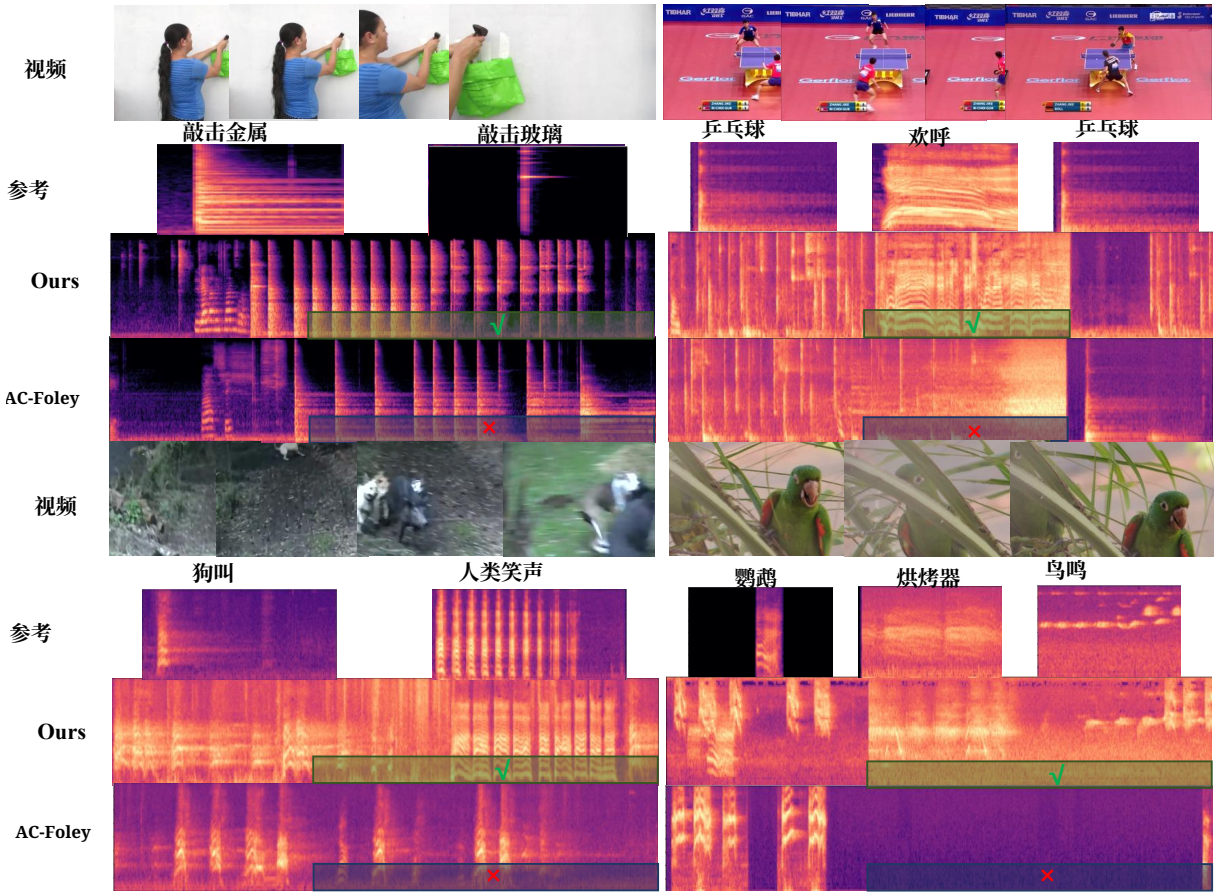

图7：**VividBench 上的定性可视化**。从上到下，各行依次展示采样视频帧、事件级参考频谱图、VividAudio 的输出以及 AC-Foley 的输出。对齐布局凸显了生成的声轨是否在正确的时间位置遵循指定的事件参考，以及相邻事件之间是否出现干扰。

当前代码库中的三个事件标注语料库：VG-GSound、AudioCaps 和一个内部语料库。对于每个源片段，标注文件都会存储事件边界、粗粒度声音标签、自由形式描述、置信度分数，以及用于构建参考-目标对的视频内分组字段。

表 7 显示，三个语料库都提供了大规模的事件级监督，但在数据干净程度和多样性之间各有取舍。VGGSound 和 AudioCaps 的有效样本保留率均超过 80%，表明经过筛选后，事件标注相对稳定。相比之下，内部语料库的覆盖率较低，反映出其声学构成更复杂，但它贡献的样本和事件总数最多。因此，尽管其原始标注噪声较多，它仍是多事件多样性的重要来源。我们还观察到，各语料库的有效事件密度始终较高，这意味着每个保留的样本通常仍包含多个可用事件，可用于构建参考—目标训练对，而非退化为近乎单事件的监督。

表 8 进一步验证了我们数据构建流程所采用的视频内分组在所有语料库中，同簇配对的相似度始终高于同一视频中的不同簇配对以及随机跨视频配对，这表明该分组捕捉到了有意义的声学亲和性，而非任意的时间共现。同时，内部与之间的间隔明显小于内部与跨视频之间的间隔，尤其是在内部语料库中。这表明，即使属于不同簇，同一视频中的事件仍共享大量背景或录音特征；而这正是需要事件级音色控制的棘手情形。最后，单例比例保持在 50% 左右，平均邻居数量接近 1，说明直接进行同组替换虽有用，但存在局限；因此，我们在训练期间通过额外加入速度、混合、退化和空参考扰动，使参考侧更加多样化。

E.1 标注模式

每个事件文件在所有语料库中都遵循相同的 JSON结构：

表 7：训练数据规模和标注覆盖率。覆盖率表示保留至少一个有效事件标注的训练样本所占的比例。

语料库	#样本	#已标注样本	#有效样本	#事件	#有效事件	每个样本的事件数	每个有效样本的有效事件数	平均时长	平均置信度	覆盖率
VGGSound	181,516	181,516	149,629	694,819	694,819	3.83	4.64	0.76	0.85	82.4%
AudioCaps	89,269	89,269	72,387	288,857	288,857	3.24	3.99	0.73	0.83	81.1%
内部	248,848	248,422	172,272	777,678	777,678	3.13	4.51	0.85	0.84	69.2%

表 8：视频内聚类后的事件组统计。相似度在 LAION-CLAP-HTSATaudio 嵌入空间中以余弦相似度进行测量（laion_clap、music_speech_audioset_epoch_15_esc_89.98 检查点），其中每个事件嵌入均通过输入事件区间内的音频片段获得。统计数据基于每个语料库中最多 500 个分组片段计算。内部：视频内的同簇事件；之间：视频内的不同簇事件；跨视频：从不同视频中随机采样的事件（5,000 个配对）。

语料库	#已分组	#簇	簇大小	单例	邻居	相似度（均值±标准差）			间隔		Top-1
						内部	之间	跨视频	内部–之间	内部–跨	
VGGSound	112,398	307,211	1.67	51.1%	1.12	0.773±0.132	0.702±0.157	0.312±0.138	0.071	0.461	0.785
AudioCaps	50,963	126,418	1.66	49.7%	1.08	0.751±0.135	0.659±0.159	0.338±0.146	0.092	0.412	0.756
内部	128,457	340,739	1.66	50.9%	1.10	0.808±0.118	0.753±0.134	0.396±0.149	0.055	0.413	0.812

{ "video": "...mp4", "audio": "...wav", "num_events": 3, "events": [{ "start": 1.18, "end": 1.98, "energy": 0.63, "label": "音乐", "desc": "柔和的氛围旋律", "confidence": 0.80, "group": 0, "neighbors": [3, 4]}]}

主要字段的用途如下：start/end 指定源片段中事件的时间跨度；energy 存储检测器使用的片段能量；label、desc 和 confidence 来自多模态标注器；group 是视频内聚类标识符；neighbors 则列出与该分组相关联、可重复用作音色一致参考的事件。

E.2 从事件标注到训练样本

上述事件标注不会直接输入生成器。相反，我们将其转换为经过事件条件化的训练记录，明确区分参考侧时序与目标侧时序。对于每个有效事件，目标跨度始终是该事件自身的已标注区间，而参考侧则根据正文中所述的同一扰动族构建：同组替换、速度变化、混合、声学退化，以及可选的空参考条件。同组替换会从邻居中选取一个项目；如果不存在有效邻居，实现会回退到另一条有效的构建路径。

空模式值得特别关注。当某个事件被采样为空时，我们不会插入占位符嵌入或特殊标记。相反，在构建 TACO-Ref 时，会将对应的参考槽位留为空声。因此，在一个多事件样本中，某个事件可能获得有效的声学参考，而同一样本中的另一个事件则获得明确的零参考条件。这会使模型了解到，参考支持可能是不完整的，并防止其过拟合于“每个目标事件始终都必须与信息丰富的示例配对”这一假设。

一条典型的导出训练数据包含 rag_audio_list、rag_seg_in_ref 和 rag_seg_in_gen 等字段，用于说明哪段波形被用作事件参考、从该波形中读取哪一片段，以及应在生成的配乐中的何处呈现所得参考。启用修复或背景锚定时，该条目还会为未标记区域存储 inp_audio_list、inp_seg_in_ref 和 inp_seg_in_gen。这种将参考侧跨度与生成侧跨度分开记录的做法，使我们能够构建时间对齐的复合参考，同时确保训练数据完全可复现。