arXiv:2602.04680v1 [cs.SD] 4 Feb 2026

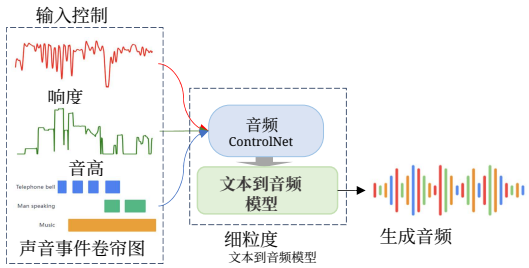
Audio ControlNet：细粒度音频生成与编辑

朱海娜[1,2], 肖瑶[1],Xiquan Li[1],马子扬[1],于建伟[4], 张博文[3],杨明琪[3], XieChen[1], 2*, [1]X-LANCE实验室，计算机科学学院，上海交通大学，[2]上海创新研究院，[3]MiniMax，[4]独立研究者 {hainazhu, chenxie95}@sjtu.edu.cn

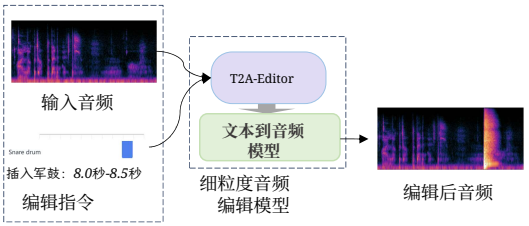
代码： https://github.com/juhayna-zh/AudioControlNet
演示： https://audio-controlnet.github.io

摘要

我们研究细粒度文本到音频（T2A）生成任务。尽管近期模型能够根据文本描述合成高质量音频，但它们往往无法精确控制响度、音高和声音事件等属性。不同于以往针对特定控制类型重新训练模型的方法，我们提出在预训练T2A骨干网络之上训练ControlNet模型，从而实现对响度、音高和事件卷帘的可控生成。我们提出了两种设计：T2A-ControlNet和T2A-Adapter，并展示了T2A-Adapter模型以更高效的结构实现了强大的控制能力。仅增加3800万个参数，T2A-Adapter便在AudioSet-Strong上取得了最先进的性能，无论是事件级还是片段级F1分数均表现出色。我们还将这一框架扩展至音频编辑，提出T2A-Editor，可根据指令指定的时间位置移除或插入音频事件。我们将发布模型、代码、数据集处理流程和基准测试，以支持未来关于可控音频生成与编辑的研究。



（a）细粒度文本到音频生成



（b）细粒度音频编辑

图 1：细粒度文本到音频（T2A）生成任务和细粒度音频编辑任务，以及所提出的 Audio ControlNet 的示意图。

1 引言

文本到音频（T2A）生成旨在根据自然语言描述合成通用音频和音效。由于其在声音设计（Evans 等，2025年）、音乐创作（Copet 等，2023）和视频制作（Cheng 等，2025年）中的潜在应用，这一领域近来受到越来越多的关注。随着生成建模技术的迅速发展，尤其是基于扩散的模型，近期的 T2A 系统（Evans 等，2025年；Liu 等，2023，2024；Hai 等，2024）在音频质量和文本-音频对齐方面均取得了显著进展。

然而，尽管取得了这些进展，大多数文本到音频模型仅通过文本提供粗粒度控制并且难以强制执行精确的时间事件结构，例如指定某种声音（如狗叫声）应出现在2到4秒之间，或多个声音事件应同时处于活动状态，如图1所示。它们还仅提供有限的信号级控制，例如响度力度、音高轮廓或特定音高成分。

近年来，多种方法尝试在文生音频生成中实现更细粒度的控制。这些方法通常分为两类。第一类通过文本描述指定应在何时出现哪些声音（Wang 等，2025年；Jiang 等，2025a）。这种方式直观易懂，但对模型的语言理解能力要求较高，而且对时间信息的误解可能导致控制不够精确。第二类采用结构化嵌入来显式表示事件发生（Zheng 等，2025年；Xie 等，2024），从而实现更精确的时间控制，但

* 通讯作者。

通常需要从头开始训练，并搭建复杂的数据模拟流程。

同样值得注意的是，大多数方法在将控制扩展到更精细的信号级属性方面仍然受限，例如精确的响度与音高变化。这促使我们需要一种框架，能够实现精确的细粒度控制，同时降低训练复杂度并保留预训练模型的灵活性。

我们并不重新训练整个模型，而是聚焦于轻量级辅助网络，使现有文本到音频模型能够接受细粒度控制信号，在保留生成能力的同时降低训练复杂度。受文本到图像生成中 ControlNet 的启发（Zhang 等人，2023；Ye 等人，2023；Mou 等人，2024；Zhao 等人，2023），我们为文本到音频模型引入 **Audio ControlNet**模型，并探索两种互补设计：基于复制网络的架构**T2A-ControlNet**，以及基于轻量级编码器的架构**T2A-Adapter**。

我们进一步认为，细粒度控制信号应当是结构化的，而非基于文本的，这需要一种既灵活又可扩展的表示，以容纳多样的控制条件。为此，我们将所有控制输入表述为与音频时间轴对齐的时间序列。在这一统一框架内，我们为响度、音高和声音事件设计了**结构化表示和特征提取器**。

我们进一步将框架扩展到细粒度音频编辑，其中指令同时指定内容和精确时间跨度（例如，“在3.0秒到4.0秒插入掌声”，“在6.0秒到8.0秒移除说话声”）。遵循相同的辅助网络范式，我们提出**T2A-Editor**，它通过引入参考音频和基于事件卷轴的编辑指令，为现有文本到音频模型配备精确的插入和移除能力。

总之，我们的主要贡献如下：

- 我们提出用于细粒度文本到音频（T2A）生成的Audio ControlNet框架。无需重新训练骨干模型，我们的方法即可实现在响度、音高和声音事件方面可控的文本到音频生成。
- 我们设计了两种针对精细化文生音频任务的 ControlNet 变体，分别为 T2A-ControlNet 和 T2A-Adapter。实验结果表明，T2A-Adapter 在保持轻量级架构的同时，取得了最先进的性能。
- 我们提出将所有控制信号表示为统一的时间序列以提高可扩展性，并为响度、音高和声音事件设计结构化表示及专用特征提取器。
- 我们进一步提出 T2A-Editor，它采用类似 ControlNet 的范式，将现有的文本到音频模型转化为音频编辑模型，支持在时间上局部定位的插入和移除。

我们的实验表明，T2A-ControlNet 和 T2A-Adapter 在控制响度、音高和声音事件方面均取得了稳定的提升。值得注意的是，T2A-Adapter 在 AudioSet-Strong 上达到了最先进的表现，取得了 54。36 $F1_{events}$ 和 68。26 $F1_{seg}$ 同时仅引入了 3800 万个额外参数。主观评估进一步表明，我们的方法优于所有基线模型（Wang 等，2025年；Hai 等，2024）。为促进未来研究，我们将发布代码、模型和基准测试。

2 相关工作

2.1 细粒度文本到音频生成

文本到音频生成（T2A）旨在根据文本描述生成音频，近期的进展主要得益于 AudioLDM（Liu 等，2023）、Tango（Ghosal 等，2023）和 AudioBox（Vyas 等，2023）等模型。近期方法还包括 Stable Audio Open（Evans 等，2025年）、EzAudio（Hai 等，2024）、TangoFlux（Hung 等，2024）和 MeanAudio（Li 等，2025年），这些方法利用扩散变换器（Peebles 和 Xie，2023）高效生成高保真、可变长度的音频。

近年来，许多研究致力于为文本到音频模型引入细粒度控制。早期工作，如 MC-Diffusion（Guo 等，2024）和 PicoAudio（Xie 等，2024；Zheng 等，2025年），训练模型以实现声音事件级控制，但仅适用于固定的事件类别或互不重叠的声音。AudioComposer（Wang 等，2025年）通过以语言驱动的形式注入控制信息，将这一能力扩展到开放词汇类别。FreeAu- dio（Jiang 等，2025b）提出了一种用于细粒度可控文生音频生成的免训练方法。ControlAudio（Jiang 等，2025a）进一步

将文生音频的可控性扩展到语音，并采用渐进式建模。

2.2 通过 ControlNet 实现可控生成

生成模型的可控性已得到广泛研究。在文本到图像生成领域，ControlNet (Zhang 等人，2023) 是一种代表性方法，它通过辅助网络引入控制信号。后续工作，如 IP-Adapter (Ye 等人，2023)、T2I-Adapter (Mou 等人，2024)、Uni-ControlNet (Zhao 等人，2023) 和 ControlNeXt (Peng 等人，2024)，则在 ControlNet 框架基础上进一步发展，使条件控制更加轻量、高效，并支持多条件控制。

类似思路也已在音频领域得到探索。在音乐生成领域，若干研究 (Wu 等人，2024；Tsai 等人，2025年；Baker 和 Nistal，2025年；Hou 等人，2025年) 对 ControlNet 进行了改造，使其能够控制旋律、力度、和弦和节奏等音乐属性。最近，SpecMaskFoley (Zhong 等人，2025年) 将 ControlNet 应用于视频同步拟音合成。

尽管取得了这些进展，文本到音频生成中的细粒度可控性仍未得到充分探索，尤其是在声音事件层面的控制方面。

2.3 细粒度音频编辑

音频编辑旨在根据编辑指令修改输入音频，这些指令通常以动作-类别对的形式表达（例如“添加笑声”）。最早的相关工作之一 AUDIT（Wang 等，2023）在合成数据上训练扩散模型以执行编辑任务。然而，AUDIT 对编辑发生时间的控制有限。Recomposer（Ellis 等人，2025年）通过引入事件卷来解决这一局限，从而能够在指定的开始和结束时间戳处精确应用编辑。尽管 Recomposer 与我们对细粒度音频编辑的定义高度吻合，但它仍需从头开始训练模型。

3 FluxAudio

在介绍我们提出的 AudioControlNet 模型之前，我们先介绍 FluxAudio。这是一种文本到音频模型，作为本研究的预训练基础模型，最初见于（Li 等，2025年）。FluxAudio 基于 FLUX（Labs 等，2025年）架构构建，并采用 MMDiT（Esser 等，2024）和 DiT（Peebles 和 Xie，2023）作为其骨干网络。

FluxAudio 的结构如图2所示。FluxAudio 接收由以下编码器编码的文本输入：CLAP 文本编码器（Wu 等人，2023）和 FLAN-T5（Chung 等，2024），这些输入首先经过多层感知机层处理，并与正弦时间嵌入以捕捉时间信息。采样率为 44.1 kHz 的输入音频首先由变分自编码器编码为潜在序列 x_0，帧率为 43.0 Hz，维度为 40。随后对潜在表示 x_0 加噪，并依次通过 N_1 个 MMDiT 块和N_2 个 DiT 块，最后经过一个卷积多层感知机生成输出。FluxAudio 采用流匹配目标进行训练，其定义为：

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{x_0,t,\epsilon}\left\| f_\theta(x_t, t) - \frac{dx_t}{dt} \right\|^2 \qquad (1)$$

其中 $x_t = (1-t)x_0 + t\epsilon$ 是时刻 t的加噪音频，f_θ 用于预测速度场。

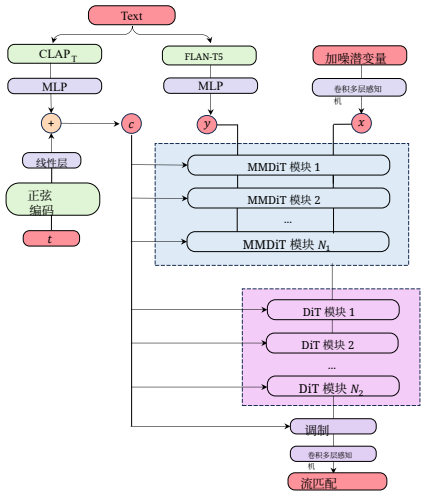


图 2：FluxAudio 的架构，由融合 MMDiT 与 DiT 的混合骨干网络组成，并采用流匹配目标进行训练。

4 Audio ControlNet

我们提出了 Audio ControlNet，这是一种用于文本到音频生成细粒度控制的框架。其核心思想是引入一个辅助条件网络，将条件注入预训练的 FluxAudio，从而无需重新训练骨干网络即可实现可控音频生成。

4.1 用于文本到音频生成的 ControlNet

我们提出了 Audio ControlNet 的两种变体：T2A-ControlNet 和 T2A-Adapter，其灵感来自 ControlNet（Zhang 等人，2023；Ye 等人，2023）在文本到图像领域所取得的成功。我们对它们进行了改造，以适配 FluxAudio 的 MMDiT与DiT混合骨干网络，从而得到 T2A-

ControlNet 和 T2A-Adapter，专为文生音频模型定制。

4.1.1 T2A-ControlNet

T2A-ControlNet 将时变条件（例如第 4.2 节中的响度、音高和声音事件滚动）作为输入，并为 FluxAudio 的每一层潜在表示提供残差分量，如图 3(a) 所示。其架构复制了 l 层 FluxAudio，包括 MMDiT 和 DiT 层，其中 l 是一个超参数。

与 FluxAudio 类似，T2A-ControlNet 接收文本提示和时间步嵌入作为输入；但其潜在输入被控制条件取代，并在输入处应用零卷积（Zhang 等人，2023）。与标准 ControlNet（Zhang 等人，2023）不同，由于 FluxAudio 使用 MMDiT 与 DiT 混合骨干网络，T2A-ControlNet 会将处理后的潜在表示直接加到 FluxAudio 的每一层中。对于 DiT 层，添加操作直接作用于对应的潜在表示。对于 MMDiT 层，残差仅添加到音频潜在分量中。

训练期间，预训练的 FluxAudio 处于冻结状态，仅更新 T2A-ControlNet 的参数。这确保 T2A-ControlNet 能提供必要的条件控制信息，而不会改变原始的 FluxAudio。

4.1.2 T2A-Adapter

如图 3(b) 所示，T2A-Adapter 使用轻量级编码器，从随时间变化的条件中提取特征。与 T2A-ControlNet 不同，T2A-Adapter 仅将控制条件作为输入。

轻量级编码器由一维卷积和 SiLU激活函数（Elfwing 等人，2018）组成，用于提取条件特征，并通过零卷积和交叉注意力将其注入 FluxAudio 前 l 层的潜在表示中。对于 DiT 层，交叉注意力作用于整个潜在表示。对于 MMDiT 层，交叉注意力仅应用于音频潜在表示。

该过程可正式表示为：

$$\mathbf{Q}^{(i)} = \mathbf{z}^{(i)} \tag{2}$$

$$[\mathbf{K}^{(i)}, \mathbf{V}^{(i)}] = \text{轻量级编码器}(\tilde{\mathbf{c}}^{\text{cond}}) \tag{3}$$

$$\mathbf{H}^{(i)} = \text{CrossAttn}(\mathbf{Q}^{(i)}, \mathbf{K}^{(i)}, \mathbf{V}^{(i)}) \tag{4}$$

$$\mathbf{z}_{\text{out}}^{(i)} = \mathbf{z}^{(i)} + \text{ZeroConv}(\mathbf{H}^{(i)}) \tag{5}$$

其中，$\mathbf{z}^{(i)}$ 表示 FluxAudio 第 i 层的潜变量，在交叉注意力中充当查询。条件 $\tilde{\mathbf{c}}^{\text{cond}}$ 经轻量级编码器处理后，生成键 $\mathbf{K}^{(i)}$ 和值$\mathbf{V}^{(i)}$。

4.2 细粒度条件控制

在本节中，我们将展示如何向 Audio ControlNet 注入条件。我们重点关注三种时变条件，即响度、音高和声音事件。我们将介绍每种条件信号及其特征提取器的提取流程，如图所示4。

4.2.1 条件与特征提取器

响度 响度定义为音频信号随时间变化的能量曲线。我们计算帧级均方根能量，并将其转换为分贝尺度。形式化地，给定一个音频信号 x，有

$$\mathbf{c}^{\text{loud}} = 20 \log_{10} \left(E_{\text{rms}}(x) + \epsilon \right), \tag{6}$$

其中，$E_{\text{rms}}(x)$ 表示均方根能量，ϵ 是一个用于保证数值稳定性的小常数。

如4（a）所示，为了获得稳定的响度控制信号，我们随后应用萨维茨基-戈莱平滑（Savitzky 和 Golay，1964）：

$$\tilde{\mathbf{c}}^{\text{loud}} = \text{SavGol}\left(\mathbf{c}^{\text{loud}} \right). \tag{7}$$

提取响度曲线后，我们沿特征维度重复该曲线，并通过广播使其匹配音频 Control Net 的输入维度。这样，我们使用 Savitzky-Golay 滤波和广播，以一种简单而有效的方式实现了响度特征提取器。

音高 音高被定义为音频信号随时间变化的基频。给定音频信号 x，我们首先估计连续基频$\mathbf{f}_0(x)$，并将其转换到对数频率域。

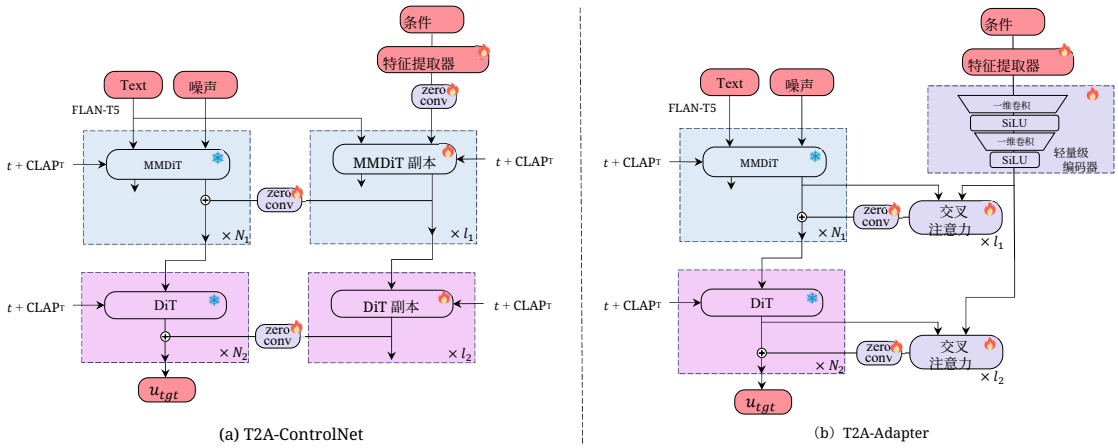
为了捕捉局部变化和多尺度变化，我们应用连续小波变换（CWT）：

$$\mathbf{c}_{\text{cwt}}^{\text{pitch}} = \text{CWT} \left(\log \mathbf{f}_0(x) \right), \tag{8}$$

从而得到连续的音高表示。如图4（b）所示，这些特征被量化为离散区间，并通过可学习的码本嵌入映射到模型输入中：

$$\tilde{\mathbf{c}}^{\text{pitch}} = \mathcal{E} \left(\text{Quantize} \left(\mathbf{c}_{\text{cwt}}^{\text{pitch}} \right); \mathcal{Z} \right). \tag{9}$$

其中，Quantize(·) 表示音高量化操作，而 $\mathcal{E}(\cdot; \mathcal{Z})$ 表示由码本 $\mathcal{Z}$ 参数化的可学习嵌入查找。



(a) T2A-ControlNet (b) T2A-Adapter

图 3：T2A-ControlNet 和 T2A-Adapter 的框架。

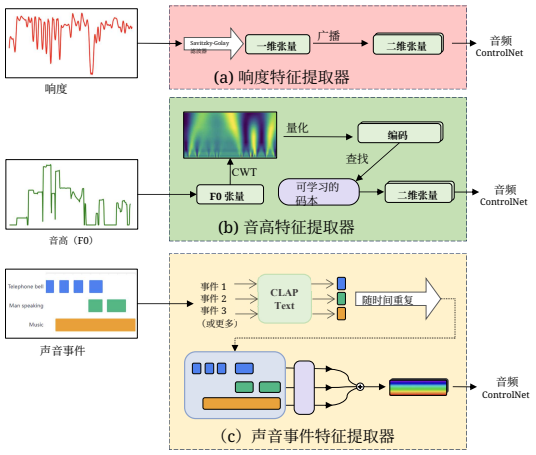


图 4：条件及其特征提取器。

声音事件 声音事件以事件卷帘表示，其中每个 N_e 声音事件类别的起始时间和结束时间沿时间轴标注，因此多个事件可以重叠（图 4(c)）。

对于每个事件 s_i，我们首先使用 CLAP 文本编码器（Wu 等人，2023）从其文本标签中提取语义嵌入：

$$\mathbf{e}_i = \mathrm{CLAP}_{\mathrm{T}}(s_i), \quad i = 1, \ldots, N_e, \tag{10}$$

按照事件卷帘，对每个嵌入沿时间轴进行扩展；事件未激活时则使用零向量：

$$\mathbf{E}_i(t) = \begin{cases} \mathbf{e}_i, & \text{if event } i \text{ is active at time } t, \\ \mathbf{0}, & \text{otherwise.} \end{cases} \tag{11}$$

为了聚合多个同时发生的事件，会应用一个共享的线性层投影$\mathbf{W}$，并将嵌入向量在各事件之间求和：

$$\tilde{\mathbf{c}}^{\text{event}}(t) = \sum_{i=1}^{N_e} \mathbf{W}\mathbf{E}_i(t) \tag{12}$$

$$\tilde{\mathbf{c}}^{\text{event}} = \left[\tilde{\mathbf{c}}^{\text{event}}(1), \ldots, \tilde{\mathbf{c}}^{\text{event}}(T)\right], \tag{13}$$

其中 $\mathbf{W}$ 表示一个可学习的线性变换。所得的固定维度时间表示 $\tilde{\mathbf{c}}^{\text{event}}(t)$ 用作 Audio ControlNet 模型的条件输入。

4.3 音频编辑

我们进一步扩展 Audio ControlNet 框架，以支持音频编辑任务。我们的动机是使文本到音频模型能够执行有针对性的音频修改。这一扩展涉及两个关键方面：

- 首先，它应向文本到音频模型引入参考音频输入。这意味着将文本到音频模型转换为音频到音频生成模型。

- 其次，它应将编辑指令作为输入，例如“插入鼓掌声音：从 2.0 秒到 2.5 秒”，“移除人声说话：从 5.0 秒到 8.0 秒”。

综合考虑上述因素，我们设计了用于音频编辑的 T2A-Editor 框架，如图5所示。T2A-Editor 采用了源自 T2A-Adapter 的架构。具体而言，我们采用了一个额外的外部条件网络，用于注入参考音频和编辑事件信息，从而在不修改核心架构的情况下，将文本到音频模型无缝转换为音频编辑模型。

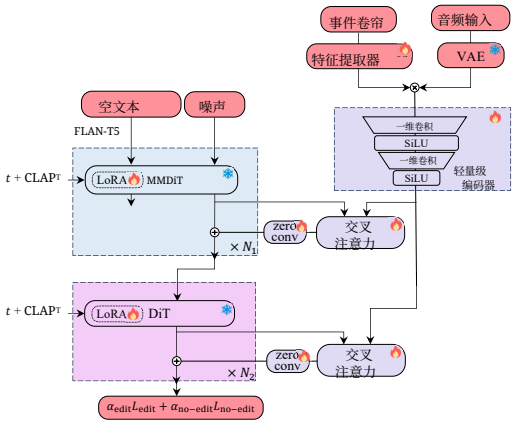


图 5：T2A-Editor 的架构。参考音频和编辑事件信息通过外部网络注入，以扩展文本到音频模型用于音频编辑任务。

首先，我们将编辑指令表示为事件卷帘，为目标音频事件提供结构化的时间描述。通过 CLAP 编码器提取文本嵌入，并使用相同的声音事件特征处理流程（第4.2.1节）进行处理，以保持一致性。

参考音频VAE潜变量与事件特征拼接后，输入轻量级编码器，再通过交叉注意力将编码器输出注入FluxAudio。T2A-Editor还可选择为DiT和MMDiT模块配备LoRA（Hu 等，2022），以共同提升可控性和音频到音频生成质量。

我们将FluxAudio中的原始文本输入替换为空字符串。训练期间，我们采用强调编辑区域的加权损失（α_{edit}：$\alpha_{\text{no-edit}}$ =10 : 1），以促进模型学习所需的编辑操作。

5 实验设置

5.1 数据集

我们在AudioSet-Strong数据集（Hershey 等，2021）上训练细粒度控制模型。该数据集包含约80,000个高质量、经人工标注的声音事件，并提供精确的起始时间戳和结束时间。我们采用与（Guo 等，2024）相同的数据划分构建训练集和验证集，并在测试集上报告结果。

对于音频编辑，我们使用 AudioSet-Strong 作为背景音频，并使用 FSD50K（Fonseca 等人，2021）作为目标音频。我们首先应用能量阈值对 FSD50K 进行筛选并分割音频，保留时长在 0.5 到 4 秒之间的片段。目标片段随后会随机插入背景音频中的随机位置，同时避免目标类别与音频描述之间发生重叠。

5.2 模型配置

FluxAudio 包含 $N_1 = 4$ 个 MMDiT 层和$N_2 = 8$ 个 DiT 层（460M 参数），以 44.1 kHz 生成 10 秒音频片段，并由变分自编码器编码为 43.0 Hz 下潜在维度为 40 的表示。该模型在 7.8 千小时音频上预训练，音频来自 AudioCaps（Kim 等人，2019）、AudioSet（Gemmeke 等人，2017；Bai 等人，2025年）、FreeSound（Mei 等人，2024）以及其他数据集（Chen 等人，2020；Doh 等人，2023；Hershey 等人，2021）。我们确保其与第 5.1 节中的测试集没有重叠。

对于 ControlNet 深度超参数 l，我们实验了 $l \in \{4, 8, 12\}$，并报告这些配置中的最佳结果（见第 6.5 节）。

响度特征通过 RMS 能量计算，帧长为 4096，跳长为 1025（43.0 fps），并使用萨维茨基-戈莱滤波进行平滑（窗口为 11，多项式阶数为 3）。音高特征使用 pYIN 和尺度为 1–32 的里克波小波提取，并量化为 256 个区间。

轻量级编码器包含 3 层带有 SiLU 的一维卷积，参数量仅为 0.77M。LoRA 以 64 的秩注入 T2A-Editor。

5.3 训练与推理

对于以响度、音高和声音事件为条件的音频 ControlNet，我们使用 64 的批量大小和 1×10^{-4} 的学习率。对于插入与移除编辑模型，所有配对训练数据均在训练期间在线提取，并采用更大的批量大小 128。所有模型均使用 AdamW 优化，在四张 GPU 上训练 120k 步，每张 GPU 的显存为 48 GB。

我们采用无分类器引导（CFG），将字幕输入和控制条件视为统一的条件信号。在训练过程中，以 10% 的概率同时丢弃字幕和控制条件。推理时，我们采用 4.5 的 CFG 缩放系数和 25 个流匹配步骤。

5.4 评估指标

客观指标。对于响度和音高控制，我们通过计算生成音频与参考特征之间的平均绝对误差（MAE）来评估生成音频。对于声音事件控制，我们采用 PB-SED（Ebbers 和 Haeb-Umbach，2022）作为声音事件检测模型

并同时报告基于事件的 F1 分数（F_1^{event}）和基于片段的 F1 分数（F_1^{seg}）（Mesaros 等人，2016），以评估事件准确性。

对于细粒度音频编辑，我们发现序列级指标往往无法准确捕捉细粒度编辑，因为它们需要精确的时间定位。为解决这一问题，我们使用 FlexSED（Hai 等，2025年），这是一种语言查询声音事件检测模型。我们将目标类别名称（用于插入或移除）输入 FlexSED，并计算编辑时间区间内的平均输出，以衡量操作的正确性。分数越高，表示插入越成功；分数越低，表示移除越成功。

主观指标。我们从三个方面采用平均意见分（MOS）方案评估模型的可控性。 MOS_{loud} 衡量生成音频对目标响度的遵循准确度。
MOS_{pitch} 评估音高控制的正确性。
MOS_{events} 评估生成的声音事件与指定事件卷帘之间的一致性。对于每项任务，每个模型的5组测试样本由20位评估者评分，并计算平均分。

6 结果与分析

我们将所提出的 T2A-ControlNet 和 T2A-Adapter 与最先进的文本到音频模型进行比较，包括 TangoFlux（Hung 等，2024）、Stable Audio Open（Evans 等，2025年）、EzAudio（Hai 等，2024）和 FluxAudio（Li 等，2025年）。对于响度控制，我们还与 EzAudio 仓库提供的能量 ControlNet 进行了比较[1]。对于声音事件控制，我们将 AudioComposer（Wang 等，2025年）作为基线。对于音频编辑，由于现有相关方法（Ellis 等人，2025年）未提供公开可用的实现，我们将编辑结果与输入音频和真实目标进行比较。

6.1 响度和音高控制

如表1所示，我们提出的方法在响度和音高方面均实现了更优的控制。具体而言，T2A-Adapter仅使用38M可训练参数，在响度上取得了1.40的平均绝对误差。T2A-ControlNet在音高上表现更好（平均绝对误差119.28），但需要大得多的参数量（410M）。与FluxAudio相比，T2A-Adapter在保持紧凑模型大小的同时带来了明显提升。此外，T2A-Adapter 在响度指标上优于EzAudio-L-Control，展现出强大的响度控制能力。主观指标同样证明了T2A-ControlNet和T2A-Adapter的有效性。

表1：响度和音高控制下的文本到音频结果。

模型	可训练参数	响度 平均绝对误差↓	音高 平均绝对误差↓	主观评价 MOS_{loud}↑	主观评价 MOS_{pitch}↑
真实值	-	-	-	3.47	3.18
TangoFlux	819M	11.41	239.46	-	-
Stable AudioOpen	1050M	17.49	233.88	-	-
EzAudio-XL	875M	11.03	203.10	1.99	1.77
EzAudio-L-Energy	278M	2.22	-	3.28	-
FluxAudio	460M	13.36	251.89	2.81	1.94
T2A-ControlNet	410M	4.08	**119.28**	**3.65**	**3.04**
T2A-Adapter	38M	**1.40**	148.02	3.64	**3.04**

6.2 声音事件控制

如表 2所示，我们提出的方法在声音事件控制方面表现强劲。具体而言，T2A-Adapter 取得了最佳结果，其 $F1_{event}$ 得分为 54.36，$F1_{seg}$ 得分为 68.26，优于所有基线文本到音频模型。与 FluxAudio 相比，T2A-Adapter 在事件级和片段级指标上均有显著提升，同时仅引入 38M 可训练参数。

此外，T2A-Adapter 持续优于 T2A-ControlNet，以显著更少的可训练参数实现了更高的控制准确率。值得注意的是，T2A-Adapter 的表现也优于 AudioComposer-S 和 AudioComposer-L，表明与专用模型相比，其声音事件可控性更优。

主观评估表明，T2A-Adapter 和 T2A-ControlNet 均优于所有基线方法，其中 T2A-Adapter 的表现略胜一筹。在 MOS_{events} 上，我们的模型甚至取得了高于真实值的分数，这可能反映了在大规模数据集上预训练骨干网络带来的优势，而这也可能提升 Audio ControlNet 的性能。

6.3 多条件控制

尽管针对不同条件的音频 ControlNet 是独立训练的，但它们可以在推理时通过直接对其潜在表示求和来组合，从而实现多条件控制[2]。这得益于 ControlNet 强大的可组合性，正如（Zhang

[1] https://github.com/haidog-yaqub/EzAudio

[2] 更多多条件控制示例请见演示链接。

表2：声音事件控制下的文本到音频结果。AudioComposer-L 的结果直接取自论文（Wang 等，2025年），因为其权重尚未公开。

模型	可训练参数	PB-SED $F1_{event}\uparrow$	PB-SED $F1_{seg}\uparrow$	主观评价 $MOS_{\text{events}}\uparrow$
真实值	-	43.36	63.46	3.21
TangoFlux	820M	6.29	43.14	-
Stable Audio Open	1050M	6.05	31.62	-
EzAudio	875M	10.43	50.18	2.16
FluxAudio	460M	10.01	50.59	2.05
AudioComposer-S	272M	43.51	60.83	2.84
AudioComposer-L	743M	44.40*	63.30*	-
T2A-ControlNet	410M	47.98	67.92	3.71
T2A-Adapter	38M	**54.36**	**68.26**	**3.76**

等人，2023）中所讨论的那样，这使得多种条件能够以直接的方式联合应用。

6.4 音频编辑

如表3所示，我们使用 FlexSED 指标评估了插入和移除两种设置下的细粒度文本到音频编辑性能。与输入音频相比，T2A-Editor 在这两种情形下都显著提升了编辑质量，表明文本引导的音频操控效果良好。

具体而言，采用 LoRA 可进一步提升性能，使用 LoRA 的 T2A-Editor 在插入和移除任务中的 FlexSED 得分分别达到 0.1340 和 0.0429，均优于原始 T2A-Editor。值得注意的是，编辑结果显著接近真实值上界，表明我们的方法仅需少量额外的可训练参数即可发挥作用。

表3：细粒度文本到音频编辑结果。

模型	可训练参数	插入 FlexSED↑	移除 FlexSED↓
输入	-	0.0257	0.1946
真实值	-	0.1946	0.0257
T2A-Editor	14M	0.0989	0.0679
T2A-Editor 搭配 LoRA	14M+4M	**0.1340**	**0.0429**

6.5 ControlNet 深度的影响

图 6 分析了 ControlNet 深度对 T2A-ControlNet 和 T2A-Adapter 的响度控制与声音事件控制的影响。随着深度增加，T2A-ControlNet 的响度平均绝对误差有所改善，从深度为 4 时的 5.6 降至深度为 12 时的 4.0。然而，其在声音事件控制方面的表现基本趋于饱和，仅在 中出现了边际变化。在声音事件控制上的表现基本已饱和， $F1_{event}$ 中仅有微小变化。

对于 T2A-Adapter，增加深度可显著提升响度控制效果。对于声音事件控制，即使深度较浅，T2A-Adapter 也能取得高$F1_{event}$ 分，表明其效果显著。

总体而言，这些结果表明，T2A-Adapter 对 ControlNet 深度的敏感度较低，并且能够以较浅的深度实现出色的控制性能，凸显了其相较于 T2A-ControlNet 更高的参数效率。

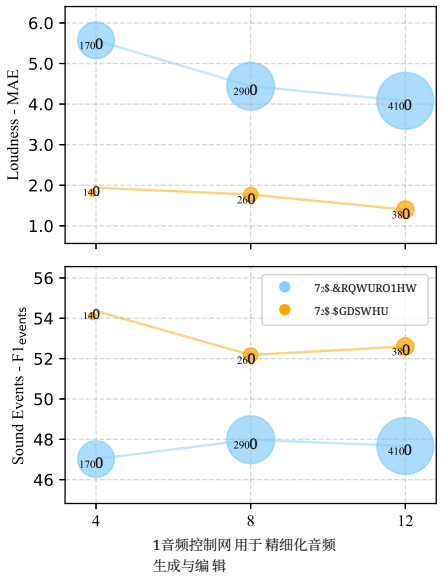


图 6：ControlNet 层对性能的影响。标记大小表示模型的可训练参数数量。

7 结论

在本研究中，我们提出了 Audio ControlNet，这是一种用于细粒度文本到音频生成和音频编辑的框架，通过轻量级控制网络增强预训练文本到音频模型。我们的设计 T2A-ControlNet 和 T2A-Adapter 能够精准控制响度、音高和声音事件。与 T2A-ControlNet 相比，T2A-Adapter 使用更少的参数，同时仍能取得出色的性能。我们进一步将这一方法扩展到音频编辑领域，提出 T2A-Editor，支持在时间上定位的音频事件插入和移除。总体而言，我们的结果表明，Audio ControlNet 模型能够为文本到音频模型提供精准、可扩展的控制与编辑能力。

局限性

本研究存在一些局限性。受计算资源限制，我们没有对超参数进行穷尽式探索，更细致的搜索可能会得到更轻量或性能更强的 Audio ControlNet。我们框架中的多条件控制主要依赖推理时的泛化能力。在训练期间引入联合多条件监督，可能会进一步提升多条件可控性。最后，我们仅研究了有限的一组条件；更丰富的控制信号，例如语音情感或歌唱旋律，将留待未来研究。

参考文献

Jisheng Bai、Haohe Liu、Mou Wang、Dongyuan Shi、Wenwu Wang、Mark D Plumbley、Woon-Seng Gan 和 Jianfeng Chen。2025年。Audiosetcaps：一种利用大型音频与语言模型，通过自动化生成流程构建的增强型音频描述数据集。*IEEE*音频、语音和语言处理汇刊。

Tom Baker 和 Javier Nistal。2025年。Lilac：一种用于音乐音频生成的轻量级潜空间ControlNet。*arXiv*预印本 *arXiv:2506.11476*。

Honglie Chen、Weidi Xie、Andrea Vedaldi 和 Andrew Zisserman。2020。VGGSound：一个大规模音视频数据集。载于*ICASSP 2020-2020 IEEE*声学、语音与信号处理国际会议（*ICASSP*），第721–725页。IEEE。

Ho Kei Cheng、Masato Ishii、Akio Hayakawa、Takashi Shibuya、Alexander Schwing 和 Yuki Mitsufuji。2025年。MMAudio：通过调控多模态联合训练，实现高质量视频到音频合成。载于计算机视觉与模式识别会议论文集，第28901–28911页。

Hyung Won Chung、Le Hou、Shayne Longpre、Barret Zoph、Yi Tay、William Fedus、Yunxuan Li、Xuezhi Wang、MostafaDehghani、Siddhartha Brahma 等人。2024。指令微调语言模型的扩展。机器学习研究期刊，25(70):1–53。

Jade Copet、Felix Kreuk、Itai Gat、Tal Remez、David Kant、Gabriel Synnaeve、Yossi Adi 和 Alexandre Défossez。2023。简单且可控的音乐生成。神经信息处理系统进展，36:47704–47720。

SeungHeon Doh、Keunwoo Choi、Jongpil Lee 和 Juhan Nam。2023。LP-MusicCaps：基于大语言模型的伪音乐描述。*arXiv* 预印本 *arXiv:2307.16372*。

Janek Ebbers 和 Reinhold Haeb-Umbach。2022年。家庭环境中的声音事件检测预训练与自训练。

Stefan Elfwing、Eiji Uchibe 和 Kenji Doya。2018 年。用于强化学习中神经网络函数逼近的Sigmoid加权线性单元。神经网络，107：3–11。

Daniel PW Ellis、Eduardo Fonseca、Ron JWeiss、Kevin Wilson、Scott Wisdom、Hakan Erdogan、John R Hershey、Aren Jansen、R Channing Moore 和 Manoj Plakal。2025年。Recomposer：基于事件卷轴引导的生成式音频编辑。*arXiv* 预印本 *arXiv:2509.05256*。

Patrick Esser、Sumith Kulal、Andreas Blattmann、Rahim Entezari、Jonas Müller、Harry Saini、Yam Levi、Dominik Lorenz、Axel Sauer、Frederic Boesel 等 1 人。2024。扩展用于高分辨率图像合成的整流流变换器。发表于第四十一届国际机器学习大会。

Zach Evans、Julian D Parker、CJ Carr、Zack Zukowski、Josiah Taylor 和 Jordi Pons。2025年。Stable Audio Open。发表于*ICASSP 2025——2025*年 *IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 1–5 页。IEEE。

Eduardo Fonseca、Xavier Favory、Jordi Pons、Frederic Font 和 Xavier Serra。2021。FSD50K：一个人工标注声音事件的开放数据集。*IEEE/ACM* 音频、语音与语言处理汇刊，30：829–852。

Jort F Gemmeke、Daniel PW Ellis、Dylan Freedman、Aren Jansen、Wade Lawrence、R Channing Moore、Manoj Plakal 和 Marvin Ritter。2017。Audio Set：一个用于音频事件的本体和人工标注数据集。载于*2017 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，第 776–780 页。IEEE。

Deepanway Ghosal、Navonil Majumder、Ambuj Mehrish 和 Soujanya Poria。2023。使用指令引导的潜在扩散模型进行文本到音频生成。载于第 *31* 届 *ACM*多媒体国际会议论文集，第 3590–3598 页。

Zhifang Guo、Jianguo Mao、Rui Tao、Long Yan、Kazushige Ouchi、Hong Liu 和 Xiangdong Wang。2024年。基于多条件扩散模型的音频生成。载于人工智能 *AAAI* 会议论文集，第 38 卷，第 18153–18161 页。

Jiarui Hai、Helin Wang、Weizhe Guo 和 Mounya Elhilali。2025年。FlexSED：迈向开放词汇声音事件检测。*arXiv* 预印本 *arXiv:2509.18606*。

Jiarui Hai、Yong Xu、Hao Zhang、Chenxing Li、Helin Wang、Mounya Elhilali 和 Dong Yu。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*。

Shawn Hershey、Daniel PW Ellis、Eduardo Fonseca、Aren Jansen、Caroline Liu、R Channing Moore 和 Manoj Plakal。2021。时间强标签的优势

在音频事件分类中的应用。在 *ICASSP 2021-2021 IEEE*声学、语音与信号处理国际会议*(ICASSP)*，第 366–370 页。IEEE。

Siyuan Hou、Shansong Liu、Ruibin Yuan、Wei Xue、Ying Shan、Mangsuo Zhao 与 Chao Zhang。2025年。用旋律和文本编辑音乐：将 ControlNet 用于扩散变换器。载于*ICASSP 2025-2025 IEEE*声学、语音与信号处理国际会议（*ICASSP*），第 1–5 页。IEEE。

Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang、Weizhu Chen 以及其他 1 人。2022。LoRA：大语言模型的低秩适配。*ICLR*，1(2):3。

Chia-Yu Hung、Navonil Majumder、Zhifeng Kong、Ambuj Mehrish、Amir Ali Bagherzadeh、Chuan Li、Rafael Valle、Bryan Catanzaro 与 Soujanya Poria。2024。TangoFlux：基于流匹配与 CLAP 排序偏好优化的超快速、高保真文本到音频生成。*arXiv* 预印本 *arXiv:2412.21037*。

Yuxuan Jiang、Zehua Chen、Zeqian Ju、Yusheng Dai、Weibei Dou 和 Jun Zhu。2025年a。Controlaudio：通过渐进式扩散模型解决文本引导、时序指示和可理解音频生成问题。*arXiv* 预印本 *arXiv:2510.08878*。

Yuxuan Jiang、Zehua Chen、Zeqian Ju、Chang Li、Weibei Dou 和 Jun Zhu。2025年b。Freeaudio：无需训练的时序规划，实现可控的长篇文本转音频生成。收录于第*33*届 *ACM*多媒体国际会议论文集，第 9871–9880 页。

Chris Dongjoo Kim、Byeongchang Kim、Hyunmin Lee 和 Gunhee Kim。2019。Audiocaps：为真实场景中的音频生成描述。收录于 *2019*年美国计算语言学协会北美分会年会论文集：人类语言技术，第 *1*卷（长篇与短篇论文），第 119–132 页。

Black Forest Labs、Stephen Batifol、Andreas Blattmann、Frederic Boesel、Saksham Consul、Cyril Diagne、Tim Dockhorn、Jack English、Zion English、Patrick Esser、Sumith Kulal、Kyle Lacey、Yam Levi、Cheng Li、Dominik Lorenz、Jonas Müller、Dustin Podell、Robin Rombach、Harry Saini以及其他2人。2025年。Flux.1 Kontext：面向潜空间中上下文图像生成与编辑的流匹配。预印本，arXiv:2506.15742。

Xiquan Li、Junxi Liu、Yuzhe Liang、Zhikang Niu、Wenxi Chen和Xie Chen。2025年。MeanAudio：基于均值流的快速且忠实的文本到音频生成。*arXiv*预印本 *arXiv:2508.06098*。

Haohe Liu、Zehua Chen、Yi Yuan、Xinhao Mei、Xubo Liu、Danilo Mandic、Wenwu Wang 和 Mark D Plumbley。2023。Audioldm：使用潜在扩散模型进行文本到音频生成。*arXiv* 预印本 *arXiv:2301.12503*。

Haohe Liu、Yi Yuan、Xubo Liu、Xinhao Mei、Qiuqiang Kong、Qiao Tian、Yuping Wang、Wenwu Wang、Yux- uan Wang 和 Mark D Plumbley。2024。Audi- oldm 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。Xinhao Mei、Chutong Meng、Haohe Liu、Qiuqiang Kong、Tom Ko、Chengqi Zhao、Mark D Plumbley、Yuexian Zou 和 Wenwu Wang。2024。Wavcaps：一个借助 ChatGPT 构建的弱标注音频描述数据集，用于音频语言多模态研究。*IEEE/ACM* 音频、语音与语言处理汇刊，32：3339–3354。

Annamaria Mesaros、Toni Heittola和Tuomas Virtanen。2016。多音声音事件检测的评估指标。《应用科学》，6(6):162。

Chong Mou、Xintao Wang、Liangbin Xie、Yanze Wu、Jian Zhang、Zhongang Qi 和 Ying Shan。2024。T2i-adapter：通过学习适配器，挖掘文本到图像扩散模型更强的可控能力。发表于*AAAI* 人工智能会议论文集，第 38 卷，第 4296–4304 页。

William Peebles 和 Saining Xie。2023。基于变换器的可扩展扩散模型。发表于*IEEE/CVF* 国际计算机视觉会议论文集，第 4195–4205 页。

Bohao Peng、Jian Wang、Yuechen Zhang、Wenbo Li、Ming-Chang Yang 和 Jiaya Jia。2024。ControlNeXt：用于图像与视频生成的强大且高效的控制方法。*arXiv* 预印本 *arXiv:2408.06070*。

Abraham Savitzky 和 Marcel JE Golay。1964。通过简化的最小二乘程序对数据进行平滑与微分。《分析化学》，36(8)：1627–1639。

Fang-Duo Tsai、Shih-Lun Wu、Weijaw Lee、Sheng-Ping Yang、Bo-Rui Chen、Hao-Chung Cheng 和 Yi-Hsuan Yang。2025年。Musecontrollite：使用轻量级条件器进行多功能音乐生成。*arXiv* 预印本 *arXiv:2506.18729*。

Apoorv Vyas、Bowen Shi、Matthew Le、Andros Tjan- dra、Yi-Chiao Wu、Baishan Guo、Jiemin Zhang、Xinyue Zhang、Robert Adkins、William Ngan 等人。2023。AudioBox：使用自然语言提示的统一音频生成。*arXiv* 预印本 *arXiv:2312.15821*。

Yuancheng Wang、Zeqian Ju、Xu Tan、Lei He、Zhizheng Wu、Jiang Bian 等人。2023。AUDIT：通过遵循指令，使用潜在扩散模型进行音频编辑。神经信息处理系统进展，36：71340–71357。

Yuanyuan Wang、Hangting Chen、Dongchao Yang、Zhiy- ong Wu 和 Xixin Wu。2025。Audiocomposer：通过自然语言描述实现细粒度音频生成。载于*ICASSP 2025-2025 IEEE*声学、语音与信号处理国际会议（*ICASSP*），第 1–5 页。IEEE。

Shih-Lun Wu、Chris Donahue、Shinji Watanabe 和 Nicholas J Bryan。2024。Music controlnet：用于音乐生成的多种时变控制。*IEEE/ACM 音频、语音与语言处理汇刊*，32：2692–2703。

Yusong Wu、Ke Chen、Tianyu Zhang、Yuchen Hui、Tay- lor Berg-Kirkpatrick 和 Shlomo Dubnov。2023。通过特征融合和关键词到描述的增强进行大规模对比式语言-音频预训练。载于*ICASSP 2023-2023 IEEE 国际声学、语音与信号处理会议（ICASSP）*，第 1–5 页。IEEE。

Zeyu Xie、Xuenan Xu、Zhizheng Wu 和 Mengyue Wu。2024。Picoaudio：实现文本到音频生成中音频事件的精确时间戳和频率可控性。*arXiv 预印本 arXiv:2407.02869*。

Hu Ye、Jun Zhang、Sibo Liu、Xiao Han 和 Wei Yang。2023。Ip-adapter：兼容文本的图像提示适配器，用于文本到图像扩散模型。*arXiv 预印本 arXiv:2308.06721*。

Lvmin Zhang、Anyi Rao 和 Maneesh Agrawala。2023。为文本到图像扩散模型添加条件控制。载于*IEEE/CVF 国际计算机视觉会议论文集*，第 3836–3847 页。

Shihao Zhao、Dongdong Chen、Yen-Chun Chen、Jianmin Bao、Shaozhe Hao、Lu Yuan 和 Kwan-Yee K Wong。2023。Uni-ControlNet：面向文本到图像扩散模型的一体化控制。神经信息处理系统进展，36:11127–11150。

Zihao Zheng、Zeyu Xie、Xuenan Xu、Wen Wu、Chao Zhang 和 Mengyue Wu。2025年。Picoaudio2：基于自然语言描述的时间可控文本到音频生成。*arXiv 预印本 arXiv:2509.00683*。

Zhi Zhong、Akira Takahashi、Shuyang Cui、Keisuke Toyama、Shusuke Takahashi 和 Yuki Mitsufuji。2025年。SpecMaskFoley：通过 ControlNet 引导预训练频谱掩码生成式 Transformer 实现同步视频到音频合成。*arXiv 预印本 arXiv:2505.16195*。

A 附录：主观评价

在本附录中，我们提供主观评估方案的更多细节。所有评估均通过统一的网页界面进行，评估者会看到输入字幕及相应的控制条件（即响度曲线、音高轮廓或事件卷帘），并需按照 5 点平均意见分量表对生成音频进行评分。图 7 89 分别展示了响度控制、音高控制和声音事件控制的评估界面截图。

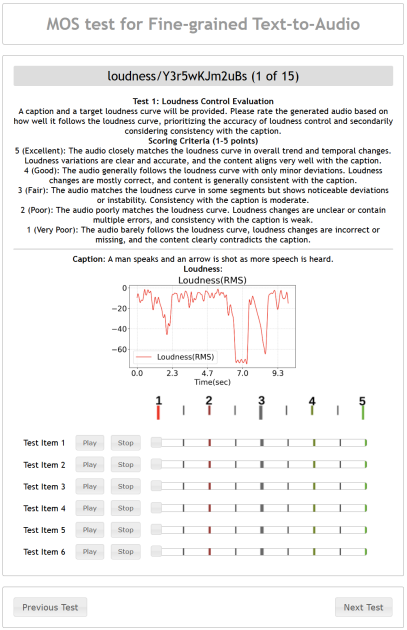


图7：响度控制主观评价所用网页用户界面的截图。

B 伦理考量

我们的工作聚焦于可控音频生成与编辑。尽管模型是在公开可用数据集上训练的，但生成音频仍可能被滥用于冒充、传播虚假信息或制作令人不安的内容。为降低此类风险，我们建议谨慎考虑应用场景、遵守相关版权与隐私法规，并避免涉及敏感或有害的使用场景。我们还强调，应透明地报告模型性能，并鼓励研究社区采用负责任的使用指南。

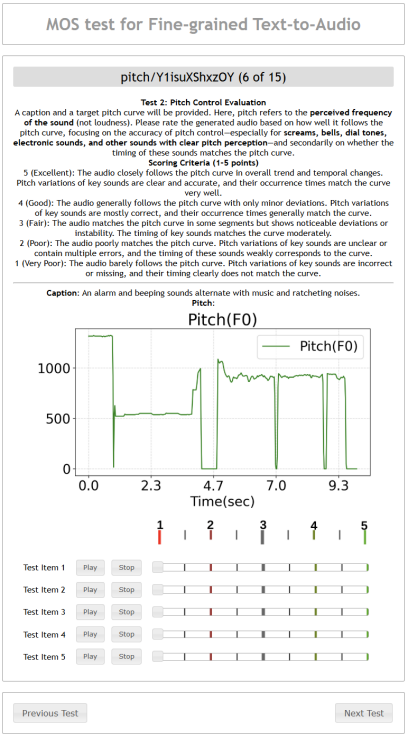


图 8：用于音高控制主观评价的网页用户界面截图。

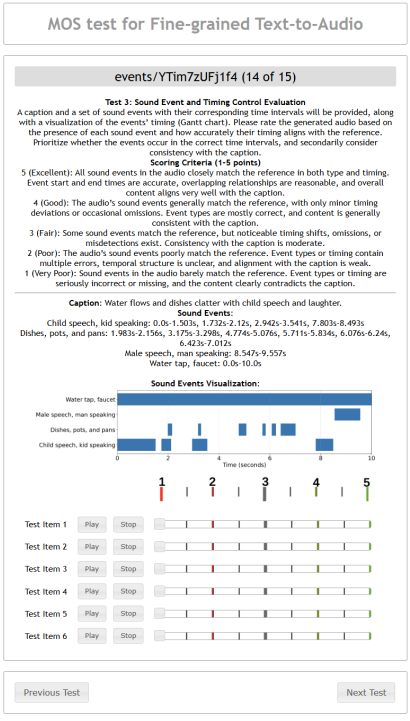


图 9：用于声音事件控制主观评价的网页用户界面截图。