用于联合音视频生成的扩散模型

Alejandro Paredes La Torre杜克大学
aleja
ndro.paredeslatorre@duke.edu

arXiv:2603.16093v1 [cs.SD] 17 Mar 2026

摘要—多模态生成模型在单模态视频生成和音频合成方面已取得显著进展，但真正的联合音视频生成仍然是一项开放性挑战。本文探讨了推动该领域发展的四项关键贡献。首先，我发布了两个高质量的音视频配对数据集。这些数据集分别包含13小时的电子游戏片段和64小时的音乐会演出，并将每个数据集切分为时长一致的34秒采样点，以促进可复现研究。其次，我在我们的数据集上从头训练了MM-Diffusion架构，展示其生成语义连贯音视频对的能力，并对快速动作和音乐线索上的对齐程度进行了定量评估。第三，我利用预训练的视频和音频编码器-解码器研究联合潜在扩散，揭示了多模态解码阶段存在的挑战和不一致性。最后，我提出了一种顺序式两步文本到音视频生成流程：首先生成视频，然后基于视频输出和原始提示词进行条件控制，以合成时间同步音频。实验表明，这种模块化方法能够生成高保真的音视频生成结果。代码可在 https://github.com/AlejandroParedesLT/audioVideo-GenAI 获取

索引词——多模态生成、音视频生成



图 2：使用提示词“霓虹灯下，一场热闹的街舞对决正在上演，舞者随着充满活力的嘻哈节拍展示令人惊叹的舞步。人群欢呼……”进行两步序列生成。可以观察到，舞蹈动作与频谱图所反映的节奏模式之间有着清晰的对齐。

I. 引言

多模态生成模型已成为一个充满活力的研究领域，这得益于扩散技术的进步 [1]–[3]。视频生成模型的研究在大规模长上下文生成方面取得了普遍进展，越来越多的工作致力于开发能够根据范围极广的主题有效生成视频的基础模型[2], [4]。随着人们对大语言模型和扩散模型的研究日益关注，多模态生成模型也越来越受到重视。在视频生成领域的重要贡献基础上 [4], ， Stable Video diffusion [5] 和 Sora [3] 等强大的模型能够利用文本引导生成视频；这基于广泛用于图像生成的扩散技术 [6], ， 而 AudioLDM [7],等模型 MusicGen [8] 和 AudioGen [9]采用高效技术生成音频和音乐效果。

这些模型仅限于单模态生成，生成内容要么是视觉，要么是音频。近期研究开始探索音视频联合生成方面的研究空白。MM-diffusion[10]是一项值得关注的工作，它通过采用双 U-Net，并在音频与视频之间使用交叉注意力，实现音视频联合生成的对齐；随后， [11],对其进行了改进，提出了一种在音频和视频的对齐潜在表示上执行扩散的框架，从而优化了训练序列。其他方法，例如 Any-to-Any [12]，会将任意输入模态的潜在编码投影到共享多模态潜在空间中，但其缺点是动作与音频之间存在错位。解决这一任务的另一种方法是扩散变换器 [13]，它展现出了出色的性能。在此基础上，AV-DiT [14] 提出采用预训练图像变换器，仅使用少量可训练参数进行适配，以实现图像和音频生成。

该研究领域的发展为高效训练方面的实验和改进带来了潜在机遇。这一领域的进展受到一些限制，包括计算需求极高以及无条件采样，这为进一步改进留下了空间。受该领域令人振奋的研究启发，本文评估了以下贡献和实验，以拓展现有研究。

- 我发布了两个全新的高质量配对音视频数据集：电子游戏玩法数据集（共 13 小时的采样点，切分为时长一致的 34 秒片段）和音乐会数据集（共 64 小时的采样点，切分为时长一致的 34 秒片段）。两者均可应要求用于非商业研究。这些贡献旨在鼓励研究界继续探索联合音视频生成。
- 我通过采用预训练的视频编码器-解码器和预训练的音频编码器-解码器，探索了联合视频音频潜扩散，并发现联合音视频解码步骤中存在不一致之处和挑战，这与以往的技术形成对比
- 我使用新发布的数据集和 MM-Diffusion架构进行了从头训练，展现了视频与音乐生成的一致性。我评估了在短促声音和快速动作上的有效对齐，这正是第二个数据集——电子游戏的特点
- 我提出了一种顺序分两步的文本到音视频生成流程，利用两个强大的视频生成模型和音频生成模型。通过将文本到视频生成模型 [15] 的结果串接到视频转

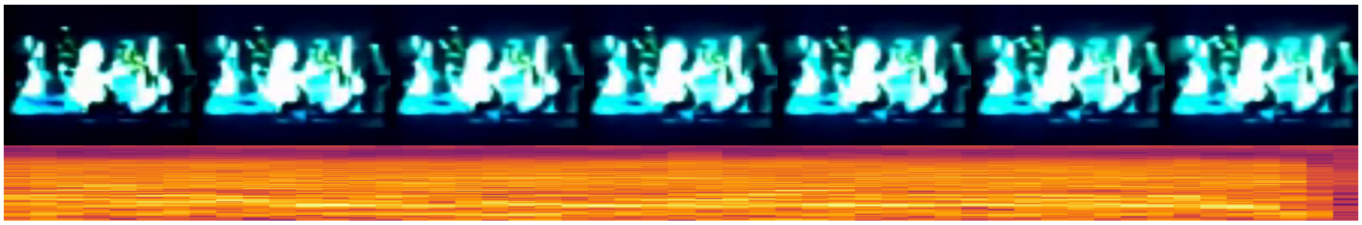


图1：MM-Diffusion 无条件生成。模型在音乐会数据集上从头训练（20k 步）。该模型能够捕捉数据集的语义特征（灯光和人物）。进一步训练可获得更好的结果。

音频 [16] 生成模型的输出中，并使用详尽的长文本提示对两个模型进行条件控制，我展示了高质量的生成结果。此外，在音频视频对齐中使用负向提示还能进一步提升任务效果。最后，我提供了一组提示词示例，以展示该方法的有效性。

II. 相关工作

A. 扩散概率模型

扩散概率模型是一种广泛应用于图像生成和声音生成的模型。它包含一个将信号映射为噪声的前向过程，以及一个将噪声映射回信号的反向过程 [17]。事实证明，这类模型在图像修复 [18], 、超分辨率 [19], 、图像到图像转换 [20] 等任务中表现尤为出色。在这一机制的基础上，研究者提出了去噪扩散隐式模型 [21] ，作为一种通过扩散概率模型以隐式方式进行采样的方法，从而提高采样速度。

B. 视频生成

近年来，随着高清扩散框架的出现，视频生成取得了飞速进步 [22], [23]。从 [5] 提出的基础性工作开始，他们推出了 Stable Video Diffusion (SVD)，这是一种用于高分辨率文本到视频和图像到视频生成的潜空间视频扩散模型，建立在基础潜扩散模型 (LDMs) [24] 之上，并扩展了 Align Your Latents 等早期视频适配方法。其三阶段训练流程包括：使用 Stable Diffusion v2.1 图像潜扩散模型进行文本到图像预训练 [5], ；通过系统化的数据整理流程，在精选的大型视频数据集 (LVD) 上进行大规模视频预训练，其中包括基于光流的运动筛选和基于OCR的文本移除；最后，在规模较小、经过精心筛选的数据集上进行高质量视频微调。在架构方面，SVD通过在U-Net骨干网络中加入时间卷积层和时间自注意力模块，在潜空间中建模时空动态，同时保留对CLIP文本嵌入的交叉注意力条件控制。作者采用连续时间去噪扩散目标，并通过线性引导调度实施无分类器引导，以平衡样本保真度与多样性。

在取得最先进的结果后， [15]提出了 CogVideoX，这是一种大规模文本到视频生成模型，利用扩散Transformer架构合成时长为10秒、帧率为每秒16帧、分辨率为 768×1360像素的视频。为应对高维视频数据建模的挑战，作者引入了一种三维变分自编码器 (VAE)，可沿空间和时间维度压缩视频输入。这种压缩有助于高效学习表征，同时保留视频保真度。该VAE采用因果卷积来维持时间一致性，确保每帧的生成仅依赖过去和当前的信息，这与自然语言处理中的自回归模型类似。为增强文本-视频对齐，模型采用配备专家自适应LayerNorm的专家Transformer，从而实现文本和视觉模态之间的深度融合。

C. 声音生成

近年来，音频生成领域的发展以 AudioLDM [7],、MusicGen [8],和 AudioGen [9], 等模型的问世为标志，它们各自引入了创新方法，以提升音频合成的质量和可控性。AudioLDM [7] 采用基于 CLAP嵌入训练的潜在扩散模型，可根据文本描述生成音频，在降低计算需求的同时实现高质量输出。MusicGen [8]采用单阶段Transformer语言模型，处理压缩离散音乐表示，从而能够根据文本或旋律输入高效、可控地生成音乐。AudioGen [9]采用自回归Transformer框架，根据文本输入生成音频采样点，并利用离散音频表示和无分类器引导来提升保真度及对文本提示的遵循程度。总的来说，这些模型代表了文本到音频生成领域的重大进展，为根据文本描述合成音频内容提供了多种不同方法。

这些方法依赖在潜在空间中进行高效编码。为改进这一任务，已有一些进展，例如 Music2latent [25], ，这是一种新颖的端到端一致性自编码器，用于潜在音频压缩。它将复数值 STFT 频谱图编码到高度压缩的连续潜在空间中，并通过由统一损失函数训练的一致性模型实现单步高保真重建。其架构包括一个编码器，用于将输入频谱图投影到低维潜在表示中；

一个解码器，用于将这些潜变量上采样回原始分辨率；以及各层级上的交叉连接，用中间编码器输出为一致性模型提供条件，以保留细粒度细节。

D. 音视频生成

音视频生成的联合对齐取得了令人鼓舞的进展。 [10] 提出了 MM-Diffusion，这是首个能够从纯高斯噪声中合成语义对齐的音视频对的联合音视频生成框架。其核心架构是一个顺序式多模态 U-Net，由两个耦合的去噪自编码器组成，分别用于音频频谱图和视频帧；它们在每个扩散时间步同步迭代地对含噪输入进行去噪。该模型为音频和视频定义了独立的前向过程，并采用共享的线性噪声调度，同时使用统一的反向模型，根据时间步 t 的音频和视频预测噪声残差。作者使用了 MM-Block，其中包含用于音频的一维膨胀卷积和用于视频的分解式 2D+1D 时空卷积，并交替结合模态专属的自注意力和交叉连接，以保留细粒度特征。为确保跨模态一致性，作者提出了一种基于随机移位的注意力模块，用于连接音频和视频子网络，使每种模态的去噪过程都能动态关注另一种模态，从而增强视频与音频在时间和语义方面的保真度。

在此贡献的基础上， [11] 提出了一种模型，在扩展 MM-Diffusion 方法论的同时，采用分层自编码器构建模态特定的低层感知潜在空间。这些空间在感知上与原始信号等效，却能显著降低维度；此外，模型还构建了一个共享的高层语义特征空间，以弥合不同模态之间的信息鸿沟。

同期研究采用了扩散过程的替代方案，例如 [26] 提出的 UniForm。这是一种统一的多任务扩散变换器，通过在共享潜在空间中拼接听觉和视觉词元，联合生成音频和视频，从而隐式学习跨模态相关性，并确保不同模态同步去噪。在架构上，该模型以单一的基于变换器的主干网络取代双 U-Net 主干网络，并采用共享权重，从而减少参数冗余，更充分地利用内在的视听协同效应。

最后，也有其他处理多模态生成的方法，尤其是视频转音频。MMAudio[16] 提出了一种新颖的多模态联合训练框架，旨在根据视频输入合成高质量、时间同步的音频，并可选择通过文本提示进行引导。不同于传统的单模态训练范式，MMAudio利用大量音频-文本数据集和视频音频对，构建统一的语义空间，从而提升音频生成的丰富性和多样性。其架构的核心是一个多模态变换器，用于整合视频、音频和文本模态，并辅以条件同步模块，在视频帧之间进行对齐与音频潜变量在帧级别的对应关系，从而确保精确的时间一致性。该模型在训练期间采用流匹配目标，以促进跨模态的高效学习。

III. 方法论

A. 扩散模型

扩散模型 [27] 是一类生成式框架，能够学习逆转施加于数据的渐进式加噪过程。前向过程会在 T 个时间步内逐步向数据添加高斯噪声，将干净样本 $\mathbf{x}_0$ 转换为带噪样本$\mathbf{x}_T$。该过程定义如下:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \quad t \in [1, T], \quad (1)$$

其中 β_t 表示用于控制每个时间步 [27]噪声大小的方差调度。反向过程旨在通过学习参数化模型p_θ（$\mathbf{x}_{t-1}|\mathbf{x}_t$），将$\mathbf{x}_t$ 去噪回$\mathbf{x}_0$ ，通常建模为:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)). \quad (2)$$

其中， μ_θ 表示由 θ预测的高斯均值。训练过程需要最小化负对数似然的变分界，这可简化为去噪得分匹配目标:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{\mathbf{x}_0,\epsilon,t}\left[\left\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\right\|_2^2\right], \quad (3)$$

其中， $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$，而 ϵ_θ用于预测添加的噪声。

B. 多模态扩散模型（MM-Diffu sion）

[10]中提出的架构在多模态音视频联合生成方面表现出色。该架构采用两个相互耦合的 U-Net 去噪自编码器。不同的数据类型需要采用不同的分布处理方式，并确保跨模态一致性。给定来自 1D 音频集 $\mathcal{A}$ 和视频集 $\mathcal{V}$的配对数据（$\mathbf{a}$，$\mathbf{v}$），每种模态的前向过程分别定义如下:

$$q(\mathbf{a}_t|\mathbf{a}_{t-1}) = \mathcal{N}(\mathbf{a}_t; \sqrt{1-\beta_t}\mathbf{a}_{t-1}, \beta_t\mathbf{I}), \quad (4)$$

$$q(\mathbf{v}_t|\mathbf{v}_{t-1}) = \mathcal{N}(\mathbf{v}_t; \sqrt{1-\beta_t}\mathbf{v}_{t-1}, \beta_t\mathbf{I}), \quad (5)$$

采用共享噪声调度 $\{\beta_t\}_{t=1}^T$ 以保持同步。为对生成过程中的相互依赖关系进行建模，采用统一反向过程。该联合模型 $p_{\theta_{av}}$ 并非使用独立的解码器，而是基于音频和视频的潜变量对各模态进行去噪:

$$p_{\theta_{av}}(\mathbf{a}_{t-1}|\mathbf{a}_t, \mathbf{v}_t) = \mathcal{N}(\mathbf{a}_{t-1}; \mu_{\theta_{av}}(\mathbf{a}_t, \mathbf{v}_t, t), \Sigma_{\theta_{av}}(\mathbf{a}_t, \mathbf{v}_t, t)), \quad (6)$$

$$p_{\theta_{av}}(\mathbf{v}_{t-1}|\mathbf{v}_t, \mathbf{a}_t) = \mathcal{N}(\mathbf{v}_{t-1}; \mu_{\theta_{av}}(\mathbf{v}_t, \mathbf{a}_t, t), \Sigma_{\theta_{av}}(\mathbf{v}_t, \mathbf{a}_t, t)). \quad (7)$$



图 3：使用提示词进行两步生成："一座繁华城市中的拥挤街头市场，摆满了色彩缤纷的水果和手工制品摊位。摊贩们高声叫卖价格，而....."可以观察到预期噪声与视频之间清晰的对齐关系。

训练目标是最小化两种模态的预测噪声与真实噪声之间的差异：

$$\mathcal{L}_{\theta_{av}} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\lambda(t)\,\|\epsilon - \epsilon_{\theta_{av}}(\mathbf{a}_t, \mathbf{v}_t, t)\|_2^2\right], \tag{8}$$

其中，$\lambda(t)$ 是一种加权函数，可强调特定时间步。这种联合建模方法确保生成的音频和视频不仅各自连贯，而且在语义上相互对齐，从而捕捉多模态数据中固有的复杂关联。

C. 音视频生成的顺序流程

为了执行文本条件下的音视频生成，我采用了两个大型预训练模型：用于视频生成的 [15] 和用于音频到视频生成的 [16]。首先介绍 *CogVideoX*，它由一个 3D因果变分自编码器、视频空间中的潜在扩散和一个专家 Transformer 构成。视频空间中的潜在扩散包括一种压缩表示 $\mathbf{z}_0$，以及一个应用于潜在空间的扩散过程。在 T 个时间步上的前向（加噪）过程定义为

$$q(\mathbf{z}_t \mid \mathbf{z}_{t-1}) = \mathcal{N}\big(\mathbf{z}_t;\ \sqrt{1-\beta_t}\,\mathbf{z}_{t-1},\ \beta_t\mathbf{I}\big), \quad \bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s). \tag{9}$$

模型学习一个反向去噪网络 ϵ_θ：

$$p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, \mathbf{y}) = \mathcal{N}\big(\mathbf{z}_{t-1};\ \mu_\theta(\mathbf{z}_t, \mathbf{y}, t),\ \Sigma_\theta\big), \tag{10}$$

其目标为

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{z}_0,\epsilon,t}\Big[\big\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \mathbf{y}, t)\big\|_2^2\Big], \tag{11}$$

其中，$\mathbf{y}$ 表示条件信号（文本嵌入和专家控制）。*CogVideoX*[15]的第二项重要贡献是专家变换器模块。专家专用适配器通过专家自适应层归一化（EA-LN）集成到变换器层中。给定第 ℓ 层的隐藏表示 $\mathbf{h}_\ell$ 和专家 e的专家权重（γ_e, β_e），EA-LN 的计算方式如下：

$$\text{EA-LN}(\mathbf{h}_\ell; e) = \gamma_e \odot \frac{\mathbf{h}_\ell - \mu(\mathbf{h}_\ell)}{\sigma(\mathbf{h}_\ell)} + \beta_e. \tag{12}$$

专家通过门控机制，根据用户提供的控制令牌进行激活，使同一主干网络能够动态适应不同的模态或控制输入。

该框架的下一个连续步骤是使用文本提示词引导一个预训练的视频转声音模型。MMAudio [16] 被选为此任务的先进模型。该模型采用条件流匹配（CFM），这是一种连续时间生成范式。模型在潜在空间 $\mathbb{R}^d$上学习条件速度场 $v_\theta(t, C, x)$，其中 C 表示条件控制输入（例如视频或文本），而x 表示音频潜在特征。样本$\mathbf{x}_1$ 通过从高斯噪声 $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$积分得到，如下所示：

$$\frac{d\mathbf{x}_t}{dt} = v_\theta(t, C, \mathbf{x}_t), \quad t \in [0, 1]. \tag{13}$$

训练期间，模型最小化以下流匹配目标：

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1,C}\left[\|v_\theta(t, C, \mathbf{x}_t) - u(\mathbf{x}_t \mid \mathbf{x}_0, \mathbf{x}_1)\|_2^2\right], \tag{14}$$

其中，$\mathbf{x}_t = t\mathbf{x}_1 + (1-t)\mathbf{x}_0$ 是线性插值，而$u(\mathbf{x}_t \mid \mathbf{x}_0, \mathbf{x}_1) = \mathbf{x}_1 - \mathbf{x}_0$ 是流速度。

1）序列生成

给定一段长篇文本提示词$\mathbf{p}$ 和无分类器引导尺度 ω，视频生成阶段会在$\mathbf{v}_{1:T}$ 潜变量空间中合成一系列帧：

$$\mathbf{v}_{1:T} = \texttt{CogVideoXPipeline}(\mathbf{p}; \omega). \tag{15}$$

这些帧经过渲染后，会以固定帧率（例如 8 fps）导出为视频文件 $\mathcal{V}_{\text{out}}$。在第二阶段，MMAudio 模型用于执行视频转音频生成任务。该模型以视觉输入（通过 Synchformer 提取的片段级和帧对齐特征）以及文本提示为条件控制，并在学习到的速度场引导下生成音频潜变量轨迹。具体而言，给定来自视频$\mathbf{v}$，$\mathbf{p}\}$ 的条件控制信息$\mathbf{v}$ 和提示词$\mathbf{p}$，以及流匹配目标 $u(\cdot)$，模型会进行如下积分：

$$\mathbf{x}_t = \mathbf{x}_0 + \int_0^t v_\theta(\tau, C, \mathbf{x}_\tau)d\tau, \quad \text{with} \quad v_\theta \approx u(\mathbf{x}_t \mid \mathbf{x}_0, \mathbf{x}_1), \tag{16}$$

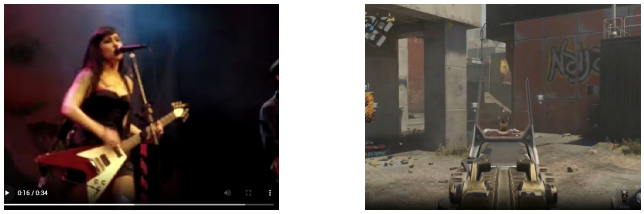
其中 $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0},\ \mathbf{I})$ 和 $\mathbf{x}_1$ 是目标潜变量。随后，生成的音频潜变量会被解码为梅尔谱图，再通过基于 BigVGAN 的解码器进行声码器转换，生成与视频流帧级同步的高保真音频波形。

IV. 实验

本节修订了实验设置的详细说明。

A. 数据集

为了获得有效的训练序列，音视频联合生成任务需要高质量的数据集。通过筛选 YouTube 8M [28] 数据集中的“音乐会”和“使命召唤”类别，我创建了一个爬取任务，以获取托管在 YouTube 上的不同视频。对这些信息进行进一步处理时，将每个视频裁剪为长度一致的 34秒片段，并去除可能包含与视频无关内容的开场部分，从而对视频进行整理。随后进一步检查音频模式，以筛除与该类别无关的视频。



（a）音乐会数据集　　（b）游戏数据集

图 4：数据集并排比较：音乐会和游戏。

因此，音乐会数据集包含超过 60 小时的视频和音频录制内容，共由 7200 个片段组成；而《使命召唤》（即游戏）数据集包含 1700 个片段，视频录制时长达 13 小时，且质量较高。

B. 从头开始训练 MM-Diffusion

为评估 MM-Diffusion框架的鲁棒性和泛化能力，我从头开始在一个新整理的音视频数据集上进行训练。MM-Diffusion架构由两个耦合的基于U-Net的去噪自编码器组成，用于联合生成时间对齐的音视频序列。该设计使模型能够有效学习跨模态相关性，从而确保连贯的视听输出。

训练配置经过精心设置，以适应自定义数据集的特性。训练采用线性噪声调度器，扩散过程包含 2000 个步骤。视频生成组件被配置为生成尺寸为 $16 \times 3 \times 64 \times 64$ 的序列，对应于由 64×64 RGB图像组成的 16 帧。同时，音频生成组件被设置为输出长度为 25,600 个采样点的波形，与视频序列的时间跨度保持一致。

关键架构参数包括在分辨率为 2、4 和 8 的情况下使用交叉注意力机制，并分别采用大小为 1、4 和 8 的注意力窗口。应用 0.1 的丢弃率以减轻过拟合。模型使用了 128 个基础通道，其中每个注意力头分配 64 个通道，并在每层加入两个残差块。为提高训练效率和稳定性，通过使用 FP16 计算启用混合精度训练，并在整个网络中应用缩放平移归一化。训练采用批量大小为 4，并使用四个 GPU，学习率设为 0.0001。

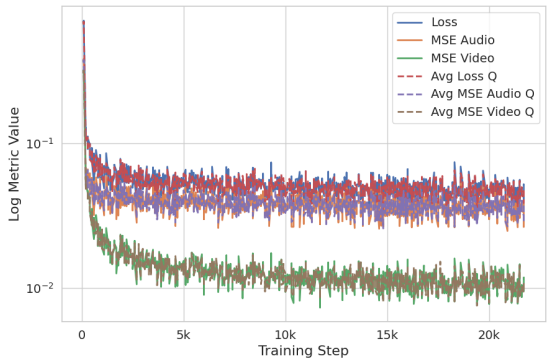


图 5：使用我自定义的音乐会数据集训练 MM-Diffusion 时的损失。图中列出了耦合的联合 U-Net 损失、音频损失和视频损失。该模型需要大量资源才能完成训练，损失函数波动很大，因此需要谨慎设置超参数。

C. 潜在 MM-Diffusion 联合音视频生成

为了缩短 MM-Diffusion 流程的训练时间，我测试了潜在扩散模型。借鉴 [11] I提出的架构，我使用了来自 stable diffusion 的预训练潜在图像扩散编码器-解码器 *sd-vae-ft-mse* [29]；对于音频编码器-解码器，我则使用了预训练音频编码器 *music2latent* [25]。通过匹配这两个编码器的长度和维度，我沿用了原始的 MM-Diffusion 训练流程来训练模型。然而，由于解码步骤中的架构差异很大，实验表现不佳。

D. T文本到音视频生成的两步序列生成

在这项实验中，我们使用 LLaMA 3.2 3B 语言模型生成了一个包含 100 条提示的数据集。每条提示平均约 50 个单词，旨在概括适用于音视频生成任务的详细描述。随后，我们将这些提示输入 CogVideoX 模型，这是一种先进的文生视频扩散模型。CogVideoX 使用三维变分自编码器（VAE）实现高效的时空压缩，并采用专家变换器架构增强文本-视频对齐能力。该模型能够生成连贯、时长较长且具有显著运动的视频，因此尤其适合这一应用。

视频生成完成后，我们使用 MM-Audio 模型处理每段视频及其对应的提示词。MM-Audio 专为根据视频和/或文本输入生成同步音频而设计。它采用多模态联合训练框架，能够从丰富多样的音频-视觉和音频-文本数据集中学习。这种方法



图6：使用提示词“正午时分，尘土飞扬的小镇上，一场狂野西部枪战一触即发。枪手们对峙时，马刺叮当作响.....”进行两步序列生成。

有助于生成高质量、语义对齐且与视觉内容保持时间一致性的音频。在视频和音频生成阶段整合详细的提示词与先进的建模技术，有助于提升系统的整体效果。

E. 评估指标

本研究采用的评估指标以弗雷歇距离（FD）为基础，用于量化高维嵌入空间中真实特征表示与生成特征表示的分布差异。具体而言，音频和视频模态均通过预训练神经模型映射为定长特征向量：基于频谱图的音频使用 ResNet18，视频使用 R(2+1)D。对于每种模态，均计算提取所得特征集的统计属性，包括经验均值和协方差矩阵。随后，利用闭式解计算真实数据分布与生成数据分布的高斯近似之间的弗雷歇距离，同时考虑均值和协方差：

$$\mathrm{FD}(\mu_1, \Sigma_1, \mu_2, \Sigma_2) = \|\mu_1 - \mu_2\|^2 + \mathrm{Tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{1/2})$$

该指标同时捕捉一阶（均值）和二阶（协方差）差异，为生成模型对真实数据流形的拟合程度提供了一种有理论依据的衡量方式。

表I：无条件生成和文本-视频生成模型的 FAD 与 FVD 分数比较。针对音乐会数据集。 [10] 的基线结果。数值越低越好。

方法	FAD ↓	FVD ↓
无条件生成	9260.94	251.62
文本到视频生成	5020.37	206.49
[10](AIST++，舞蹈数据集)	10.69	75.71

V. 结论

本文介绍了使用 MM-Diffusion框架在自定义数据集上进行联合音视频生成的实验，并提出了一种两步顺序流程：使用 CogvideoX [15]进行视频生成，再使用 MM-Audio [16] 进行视频转音频生成。所提出的潜在音视频联合扩散实现颇具挑战，因为解码步骤会变得不稳定。研究还发现，从头开始训练 MM-Diffusion 需要大量资源，并且需要仔细设置超参数。未来在这一研究方向上的工作可以包括改进潜在 MM-Diffusion 方法，以更少的资源实现最先进的结果。

参考文献

[1] Z. Yang、L. Li、K. Lin、J. Wang、C.-C. Lin、Z. Liu 和 L. Wang，《大型多模态模型的曙光：使用 GPT-4V(ision) 进行的初步探索》，2023年。[2] NVIDIA：N. Agarwal、A. Ali、M. Bala、Y. Balaji、E. Barker、T. Cai、P. Chattopadhyay、Y. Chen、Y. Cui、Y. Ding、D. Dworakowski、J. Fan、M. Fenzi、F. Ferroni、S. Fidler、D. Fox、S. Ge、Y. Ge、J. Gu、S. Gururani、E. He、J. Huang、J. Huffman、P. Jannaty、J. Jin、S. W. Kim、G. Klár、G. Lam、S. Lan、L. Leal-Taixe、A. Li、Z. Li、C.-H. Lin、T.-Y. Lin、H. Ling、M.-Y. Liu、X. Liu、A. Luo、Q. Ma、H. Mao、K. Mo、A. Mousavian、S. Nah、S. Niverty、D. Page、D. Paschalidou、Z. Patel、L. Pavao、M. Ramezanali、F. Reda、X. Ren、V. R. N. Sabavat、E. Schmerling、S. Shi、B. Stefaniak、S. Tang、L. Tchapmi、P. Tredak、W.-C. Tseng、J. Varghese、H. Wang、H. Wang、H. Wang、T.-C. Wang、F. Wei、X. Wei、J. Z. Wu、J. Xu、W. Yang、L. Yen-Chen、X. Zeng、Y. Zeng、J. Zhang、Q. Zhang、Y. Zhang、Q. Zhao 和 A. Zolkowski，《面向物理人工智能的 Cosmos 世界基础模型平台》，2025年。[3] Y. Liu、K. Zhang、Y. Li、Z. Yan、C. Gao、R. Chen、Z. Yuan、Y. Huang、H. Sun、J. Gao、L. He 和 L. Sun，《Sora：大型视觉模型的背景、技术、局限与机遇综述》，2024年。[4] A. Blattmann、R. Rombach、H. Ling、T. Dockhorn、S. W. Kim、S. Fidler 和 K. Kreis，《Align Your Latents：利用潜在扩散模型进行高清视频合成》，2023年。[5] A. Blattmann、T. Dockhorn、S. Kulal、D. Mendelevitch、M. Kilian、D. Lorenz、Y. Levi、Z. English、V. Voleti、A. Letts、V. Jampani 和 R. Rombach，《Stable Video Diffusion：将潜空间视频扩散模型扩展至大型数据集》，2023年。[6] A. Ramesh、P. Dhariwal、A. Nichol、C. Chu 和 M. Chen，《利用 CLIP 潜变量进行分层文本条件图像生成》，2022年。[7] H. Liu、Z. Chen、Y. Yuan、X. Mei、X. Liu、D. Mandic、W. Wang 和 M. D. Plumbley，《AudioLDM：利用潜在扩散模型实现文本到音频生成》，2023年。[8] J. Copet、F. Kreuk、I. Gat、T. Remez、D. Kant、G. Synnaeve、Y. Adi 和 A. Défossez，《简单且可控的音乐生成》，2024年。[9] F. Kreuk、G. Synnaeve、A. Polyak、U. Singer、A. Défossez、J. Copet、D. Parikh、Y. Taigman 和 Y. Adi，《AudioGen：文本引导的音频生成》，2023年。[10] L. Ruan、Y. Ma、H. Yang、H. He、B. Liu、J. Fu、N. J. Yuan、Q. Jin 和 B. Guo，《MM-Diffusion：学习用于联合音频和视频生成的多模态扩散模型》，2023年。[11] M. Sun、W. Wang、Y. Qiao、J. Sun、Z. Qin、L. Guo、X. Zhu 和 J. Liu，《MM-LDM：用于有声视频生成的多模态潜在扩散模型》，2024年。[12] Z. Tang、Z. Yang、C. Zhu、M. Zeng 和 M. Bansal，《通过可组合扩散实现任意到任意生成》，2023年。[13] W. Peebles 和 S. Xie，《采用 Transformer 的可扩展扩散模型》，2023年。[14] K. Wang、S. Deng、J. Shi、D. Hatzinakos 和 Y. Tian，《AV-DiT：用于联合音频和视频生成的高效视听扩散 Transformer》，2024年。[15] Z. Yang、J. Teng、W. Zheng、M. Ding、S. Huang、J. Xu、Y. Yang、W. Hong、X. Zhang、G. Feng、D. Yin、Y. Zhang、W. Wang、Y. Cheng、B. Xu、X. Gu、Y. Dong 和 J. Tang，《CogVideoX：采用专家 Transformer 的文本转视频扩散模型》，2025年。[16] H. K. Cheng、M. Ishii、A. Hayakawa、T. Shibuya、A. Schwing 和 Y. Mitsufuji，《MMAudio：通过多模态联合训练实现高质量视频转音频合成》，2025年。[17] P. Dhariwal 和 A. Nichol，《扩散模型在图像合成方面胜过生成对抗网络》，2021年。[18] A. Lugmayr、M. Danelljan、A. Romero、F. Yu、R. Timofte 和 L. V. Gool，《RePaint：使用去噪扩散概率模型进行图像修复》，2022年。[19] Z. Qiu、H. Yang、J. Fu 和 D. Fu，《用于压缩视频超分辨率的时空频率 Transformer 学习》，2022年。[20] C. Saharia、W. Chan、H. Chang、C. A. Lee、J. Ho、T. Salimans、D. J. Fleet 和 M. Norouzi，《Palette：图像到图像扩散模型》，2022年。[21] J. Song、C. Meng 和 S. Ermon，《去噪扩散隐式模型》，2022年。[22] J. Ho、W. Chan、C. Saharia、J. Whang、R. Gao、A. Gritsenko、D. P. Kingma、B. Poole、M. Norouzi、D. J. Fleet 和 T. Salimans，《Imagen Video：利用扩散模型进行高清视频生成》，2022年。

[23] R. Villegas、M. Babaeizadeh、P.-J. Kindermans、H. Moraldo、H. Zhang、M. T. Saffar、S. Castro、J. Kunze 和 D. Erhan，《Phenaki：根据开放领域文本描述生成可变长度视频》，2022年。[24] R. Rombach、A. Blattmann、D. Lorenz、P. Esser 和 B. Ommer，《利用潜在扩散模型进行高清图像合成》，2022年。[25] M. Pasini、S. Lattner 和 G. Fazekas，《Music2latent：用于潜在音频压缩的一致性自编码器》，2024年。[26] L. Zhao、L. Feng、D. Ge、R. Chen、F. Yi、C. Zhang、X.-L. Zhang 和 X. Li，《UniForm：用于音频视频生成的统一多任务扩散变换器》，2025年。[27] J. Ho、A. Jain 和 P. Abbeel，《去噪扩散概率模型》，2020年。[28] S. Abu-El-Haija、N. Kothari、J. Lee、P. Natsev、G. Toderici、B. Varadarajan 和 S. Vijayanarasimhan，《YouTube-8M：大规模视频分类基准》，2016年。[29] R. Rombach、A. Blattmann、D. Lorenz、P. Esser 和 B. Ommer，《利用潜在扩散模型进行高清图像合成》，载于 *IEEE/CVF* 计算机视觉与模式识别会议（*CVPR*）论文集，第10684–10695页，2022年6月。

附录

从头训练 MM-Diffusion

下面展示了 MM-Diffusion 的更多示例。



图 A1：在音乐会数据集上从头训练 MM-Diffusion。

序列式音视频生成

