arXiv:2512.14938v1 [cs.CV] 16 Dec 2025

TalkVerse：让分钟级音频驱动视频生成惠及大众

Zhenzhi Wang[1], Jian Wang[2],Ke Ma[2],Dahua Lin[1],Bing Zhou[21]香港中文大学， [2]Snap Inc.
wz122@ie.cuhk.edu.hk

摘要

我们推出了 *TalkVerse*，这是一个大规模开放语料库，专为单人音频驱动说话视频生成而设计，旨在实现不同方法间公平、可复现的比较。当前最先进的系统依赖封闭数据或计算密集型模型，而 *TalkVerse* 提供了 *230* 万个高分辨率（*720p/1080P*）音视频同步片段，总时长达 *6,300* 小时。这些片段从超过 *6* 万小时的视频中筛选而来，经过透明的处理流程，包括场景切换检测、美学评估、严格的音视频同步检查，以及全面的标注，其中包含二维骨架和结构化的视觉/音频风格字幕。基于 *TalkVerse*，我们提出了一个可复现的 *5B* 扩散变换器基线模型，该模型基于 *Wan2.2-5B* 构建。通过采用具有高下采样率的视频*VAE*，以及结合运动帧上下文的滑动窗口机制，我们的模型能够实现分钟级生成，同时保持低漂移。其唇形同步效果和视觉质量可媲美 *14B Wan-S2V* 模型，同时 *10*× 推理成本更低。为提升长视频的叙事效果，我们集成了一个多模态大语言模型导演，根据音频和视觉线索改写提示词。此外，我们的模型还支持通过受控的潜在噪声注入实现零样本视频配音。我们将数据集、训练方案和 *5B* 检查点开源，以降低音频驱动人体视频生成研究的门槛。项目页面：*https://zhenzhiwang.github.io/talkverse/*。

1. 引言

基于音频和姿势条件的人体视频生成，随着预训练视频扩散模型的发展而迅速进步 [13, 14, 21, 28, 36, 41, 42, 55, 57, 62, 73, 74]；这些模型 [4, 6–8, 26, 27, 30, 40, 43, 53, 60, 63, 66, 86]，使富有表现力的说话／表演肖像及更多应用成为可能。然而，该领域的高质量研究与应用仍受到两方面阻碍：(1) 缺乏具有可靠视听同步效果的开放训练数据；(2) 计算资源消耗巨大的预训练模型（例如 Wan2.1-14B）妨碍了广泛参与。在这项工作中，我们展示了，构建一个大规模的——全身、开放透明且以人为中心的视频语料库，并配备高质量的同步音频，是切实可行的；借助有限的计算资源（例如使用 64 个图形处理器运行一周），它可以促进轻量级模型的训练，使其唇形同步和视觉质量达到商业级模型的水平。我们的目标是在适度计算资源下实现可复现性，而非不惜代价追求最先进的性能。

尽管许多现有方法已能生成全身音频驱动视频 [13, 21, 23, 24, 42, 58], ，但其训练数据集要么是内部数据 [13, 21, 37, 42]，要么由开源说话头像数据集 [18, 22, 84, 87]与少量自行收集的全身数据构成。此外，近期的大规模以人为中心的数据集 OpenHumanVid [39] 专注于文本生成视频，忽略了人体图像动画中的音频同步属性。我们发现现有开源数据存在若干挑战：**(1) 视频与音频同步**：与可通过姿态检测或视频描述生成从视频中推断出的姿势和文本条件不同，音频模态是在数据采集期间录制的，互联网视频无法保证其同步性；**(2) 外观一致性与消歧**：当存在多个主体时，当前说话者可能存在歧义，这使得训练音频驱动视频生成模型变得困难；**(3) 数据规模**：由于我们希望根据真实场景音频生成逼真的唇部动作和全身动作，因此数据的总时长和语言覆盖范围对这项任务至关重要。其基本假设是，训练语料库应覆盖多种语言中的大多数音素，以便模型能够泛化。因此，高质量、大规模的数据集对于推动这一领域的发展和实现端到端可复现性不可或缺。

本文介绍了 TalkVerse，这是一个大规模音视频语料库，用于以音频或姿势为条件的单人全身人物视频生成。该语料库包含*230万* 个高分辨率（720p 或 1080P）的音频同步片段，总时长超过*6.3千* 小时，素材来自两个以文本驱动的人体视频生成来源 [12, 39]，初始包含 3000 万个人物相关片段，如图 1所示。该语料库通过一套流程进行整理，流程包括场景切换检测、人物与字幕筛选、美学和图像质量评估、动作筛选以及音视频同步检查，随后进行语



图 1. TalkVerse 概览：包含 230 万个 720p 单人片段（6300 小时），并配有同步语音音频、二维骨骼序列和结构化描述（视觉+音频风格）。音频驱动的视频生成 50 亿基线模型的数据集和训练方案将开源。它还可用于音视频联合生成、姿态驱动视频生成和视频个性化。（右上）配有视觉+音频字幕的视频示例，以及自动语音识别结果。（右下）姿态可视化。

言标注。它强调经过验证的音视频同步，以及高质量且身份和动作多样的内容。为了在基于预训练 Wan DiT [61], 训练音频条件注入时，尽可能保留文本条件控制能力，我们还使用微调后的 Qwen2.5-VL [2] 标注了包含背景外观、摄像机运动、主体外观和主体运动的结构化视频描述，并使用 Qwen3-Omni [72]生成音频风格描述。我们利用姿态检测器 [76] 确保画面中只有一个主体，并提取姿势条件。总之，我们提供了多种人物相关标注，包括视频描述、同步音频、二维人体姿态、音频风格描述和语言类型。鉴于音视频同步这一特性，该数据集也适用于其他以人为中心的视频任务，例如音视频联合生成、姿态驱动视频生成，以及基于人物图像的视频个性化。

为展示数据的效用，我们采用 Wan-S2V 架构 [24]，并使用更高的变分自编码器下采样率训练 5B 模型。尽管该基线模型在推理时的成本比强大的 Wan-S2V-14B 低 10×，但其感知质量和唇形同步效果与之相当。总之，我们的贡献如下：**(1) TalkVerse- 230万**。我们发布一个大规模、同步的单人说话/表演视频语料库（包含*230万*个片段和*6300+*小时），该语料库通过透明的流程整理而成，并采用可复现的筛选标准评估视觉质量、动作和音视频对齐。**(2) TalkVerse-5B 模型和开放方案**。我们开源数据处理方案、预处理工具、训练脚本和推理代码，并提供检查点；该紧凑型 5B Wan-S2V 基线模型基于 TalkVerse 进行端到端训练。

2. 相关工作

人体动画模型可根据文本、参考图像、人体姿势或音频合成人体视频。早期基于 GAN 的方法 [51, 52, 65, 85], typically 在相对较小的数据集上训练 [47, 51, 71,87], , 致力于自监督姿势迁移。近期基于扩散的方法 [33, 50, 67, 68, 83] 通过以二维骨架、估计深度或网格序列为条件，超越了生成对抗网络。音频驱动的人像动画 [18, 36, 57, 77,82] 已通过两阶段流程 [17, 31, 45, 56] 或统一的单阶段设计 [13, 21, 23, 24, 41, 42, 58, 69]拓展至全身运动。然而，许多以往的研究成果仍仅供内部使用，或只以检查点形式发布。相比之下，我们是首个公开训练高质量人体动画模型的完整数据整理流程的团队。据我们所知，我们的数据集是该领域首个百万级开源资源。

以人为中心的视频数据集 包含成对且在时间上同步的模态（例如姿势、音频和文本），对人体图像动画的发展至关重要。以往以人为中心的数据集要么侧重姿势而不包含音频 [5, 35, 67, 79] ，要么仅限于说话头部视频 [18,84, 87]。与此同时，UCF101 [54], NTU RGB+D [49], ActivityNet [9], 和 AVA [25] 等动作识别数据集，分辨率较低，视觉质量也不符合我们的需求。通用文本生成视频语料库，例如 HD-VILA-100M [75], HowTo100M [46],WebVid-10M [3], Panda-70M [12], 和 OpenVid-1M [48],，在人类类别中的多样性（身份和动作）不足，通常也无法提供准确的字幕、骨架表示，或作为运动条件的同步语音音频。近期，OpenHumanVid [39]收录了 1.67万小时的以人为中心的视频，但其时长和视觉质量仍有限，且无法保证音频与动画训练同步。TalkCuts [11]收集了多镜头长篇访谈节目，但其规模仍约为 500 小时。尽管这些资源很有价值，但由于多样性、规模、音频同步和骨架质量方面的限制，仍不足以训练人体图像动画模型。在 Panda-70M 和 OpenHumanVid 等通用文本生成视频语料库的基础上，我们整理了一个高

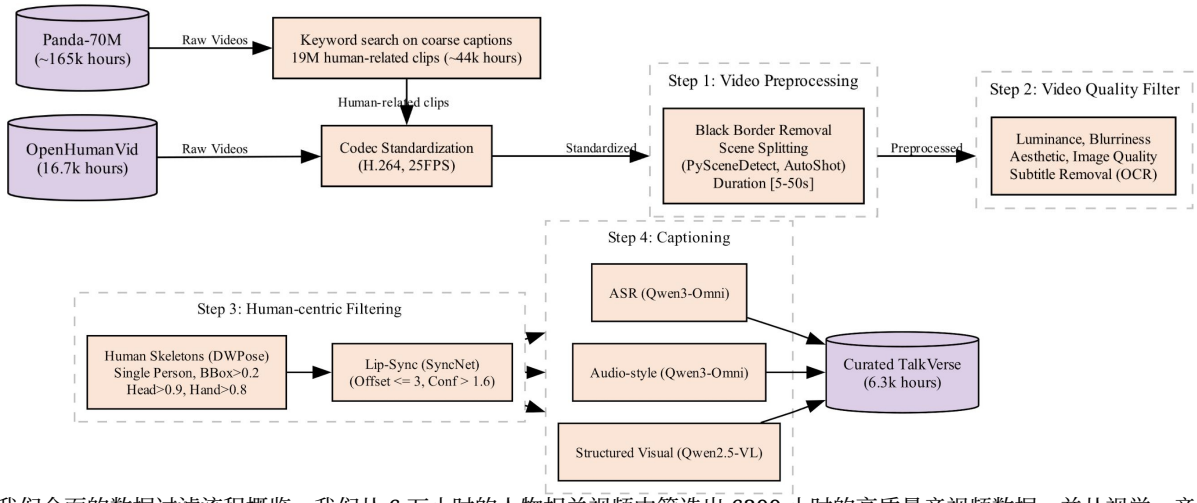


图 2. 我们全面的数据过滤流程概览。我们从 6 万小时的人物相关视频中筛选出 6300 小时的高质量音视频数据，并从视觉、音频风格和自动语音识别内容三个方面提供二维骨架和丰富的文本标注。

质量优良的单人全身人物中心数据集，并验证了姿态质量和音频与姿态同步情况；实验结果表明，该数据集适用于训练人体图像动画模型。

3. TalkVerse 数据集

TalkVerse 包含**230万** 个短片段（5–50 秒），总时长达*6300*小时。每个片段都包含：(i) 720p/1080p 帧序列，(ii) 与时间对齐的语音音频，(iii) 二维姿态骨架，(iv) 视觉与音频描述，以及 (v) 用于分析和评估的元数据，例如时长、语言类型和质量评分。我们主要从 *OpenHumanVid* [39] 和 *Panda-70M* [12]中整理数据。所有片段在下游处理前均经过标准化、去重，并验证音视频对齐情况。为减少音频驱动生成中的当前说话人歧义，我们仅保留画面中仅有一人的片段以及说话／表演片段。

3.1. 数据处理

如图 2所示，我们构建了一套全面的数据筛选流程，以确保数据集具有较高的视觉质量和音频同步性。

3.1.1. 视频预处理

视频分割。我们首先进行编解码标准化（H.264）并去除黑边。为便于唇形同步筛选，所有视频均转换为25帧/秒。我们使用 PySceneDetect [10]进行场景切分，将视频分割为连贯的镜头，随后保留时长为 5–50 秒的视频片段，以平衡语义完整性与训练效率。值得注意的是，许多互联网视频的场景边界可能无法通过 PySceneDetect [10]检测到，因为这些剪辑视频是由固定摄像机拍摄的，背景连续，但前景可能突然发生姿态变化。这会导致在以此类视频训练音频驱动视频生成方法时，长视频生成中的视频生成窗口（例如 5 秒）内出现意料之外的人体动作。对于这类场景边界，我们使用 AutoShot [88]切分视频片段，以确保时间平滑性。经验表明，在音频驱动视频生成训练中，去除生成结果中的突然姿态变化是有效的。

视频质量筛选器。 我们从多个维度进行质量筛查，考量亮度、模糊度、美感和图像质量 [70]。阈值经过选择，以便在初筛时实现较高的召回率，并在筛选后达到较高的精确率。

3.1.2. 人物中心标注

文本描述。 为在音频注入训练中保留强有力的文本条件控制，我们还使用经过微调的 Qwen2.5-VL [2]，根据片段中均匀间隔的七帧生成结构化视频描述，涵盖背景外观、摄像机运动、主体外观和主体运动；同时使用 Qwen3-Omni [72]，根据整个音轨生成音频风格描述。音频描述着重刻画说话人相关声学属性，如年龄、性别、韵律、情绪和语速，并包含自动语音识别结果。音频风格描述有助于保持长视频生成中的情绪一致性。需要注意的是，对于音频驱动的视频生成，我们不会在训练中纳入自动语音识别信息，但这些信息未来可能有助于联合音视频生成。此外，我们将利用视觉描述中的光学字符识别信息，移除带有广播节目或电视节目等动态字幕的视频。只要静态文字叠加在整个视频中保持位置不变，就会予以保留。

骨骼序列。 我们使用 DW- Pose [76] 提取人体骨架，将其作为人体视频生成的姿态控制信号。此外，我们还利用检测到的人物边界框尺寸以及头部/手部关键点置信度来筛选

表1. 以人为中心的视频理解与生成代表性数据集的比较。我们整合了从 DWPose 派生的骨架序列 [76]，并通过 SyncNet 确保语音音频同步 [15]。在音频驱动的视频生成领域，我们是拥有同步语音音频的最大数据集，并保留全身人体动作，而非仅包含说话头部视频。

数据集名称	领域	# 视频	时长（小时）	动作类型	分辨率
WebVid-10M[3]	Open	10M	52K	Text	360P
Panda-70M[12]	Open	70M	167K	Text	720P
OpenVid-1M[48]	Open	1M	2K	Text	512P
UCF-101[54]	人体动作	13.3K	26.7	Text	240P
NTU RGB+D[49]	人体动作	114K	37	三维姿态，深度	1080P
ActivityNet[9]	人体动作	100K	849	Text	480P
TikTok[34]	单人舞蹈	350	1	骨架	600P
UBC-Fashion[79]	单人舞蹈	500	2	骨架	720P
HumanVid[67]	单人舞蹈	20K	115	骨架、摄像机	1080P
OpenHumanVid[39]	多人表演	1320万	1.67万	文本、骨架姿态、不同步音频	720P
VoxCeleb2[16]	说话头部视频	1M	2.4K	语音音频	240P
HDTF[84]	说话头部视频	368	15.8	语音音频	512P
TalkingHead-1KH[65]	说话头部视频	80K	160	语音音频	512P
CelebV-HQ[87]	说话头部视频	35K	68	文本、语音音频	512P
CelebV-Text[78]	说话头部视频	70K	279	文本，语音音频	512P
DH-FaceVid-1K[20]	说话头部视频	270K	1.2K	文本，语音音频	512P
TalkCuts [11]	单人说话	164k	507	文本、骨架姿态、语音音频	1080P
Ours	单人说话	2.3M	6.3K	文本、骨架姿态、语音音频	720P/1080P

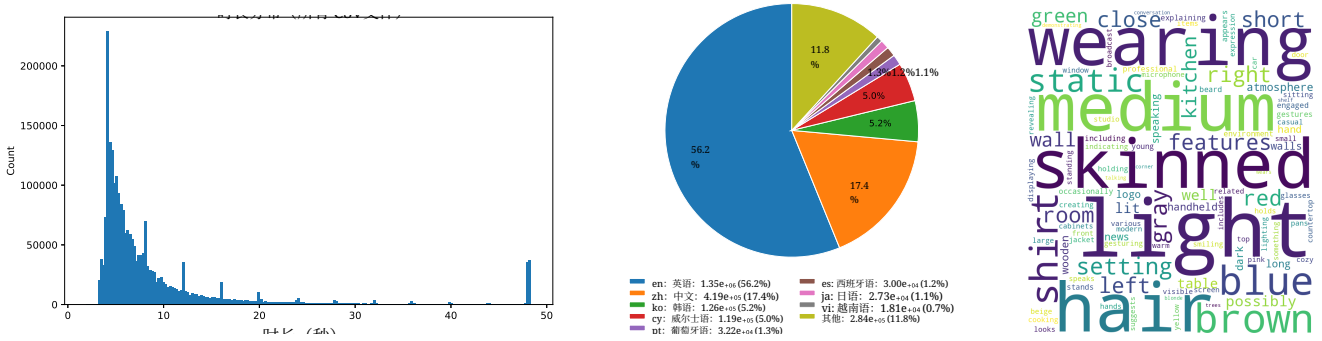


图 3. 数据集统计和一个　　分析：（左）片段时长分布；（中）来自自动检测的语言分布基于语音轨道；（右）关键词概括字幕内容。

数据：我们保留仅包含单个人物且头部关键点持续存在的视频片段，要求边界框面积 $>0.2\times$整幅图像面积，头部置信度达到 >0.9，并且在手部出现时，手部关键点置信度达到 >0.8。因此，我们的数据集能够利用百万级视频数据支持姿态驱动视频生成。

语音音频。 我们使用 SyncNet [15] 检查唇音同步并移除不同步的视频，从而强化音频条件训练。具体而言，SyncNet 学习联合音频-视觉嵌入，并输出标量同步置信度以及时间偏移（以帧为单位）。我们在数百万个片段上大规模运行该模型，并保留满足|偏移| ≤ 3 和置信度 >1.6的片段。我们观察到，即便少量不同步数据也会降低音频驱动的视频生成中的唇同步学习效果，因此我们采用这些严格阈值，以尽量减少错位。

4. 方法

4.1. 音频驱动的视频生成

给定一张参考图像、一段音频输入，以及一条描述视频内容和音频风格的文本提示，我们可以生成与音频同步、同时保留参考图像内容的视频，如图4所示。我们的基础模型（Wan2.2-5B）是一个经过预训练、用于文本/图像生成视频的扩散Transformer（扩散变换器），并搭配视频VAE，其下采样倍率为$(t,\ h,\ w)=(4,\ 16,\ 16)$，压缩比高于常用模型（如 Wan2.1-14B）。因此，在生成相同分辨率的视频时，这一潜空间的速度4×比以往更快。与 Wan-S2V [24], 的发现一致，我们发现，使用一组 {运动帧上下文、去噪视频和参考图像}，并配以恰当的位置嵌入，通过滑动窗口即可生成流畅连贯、时长达一分钟的视频。其视觉质量、主体一致性以及

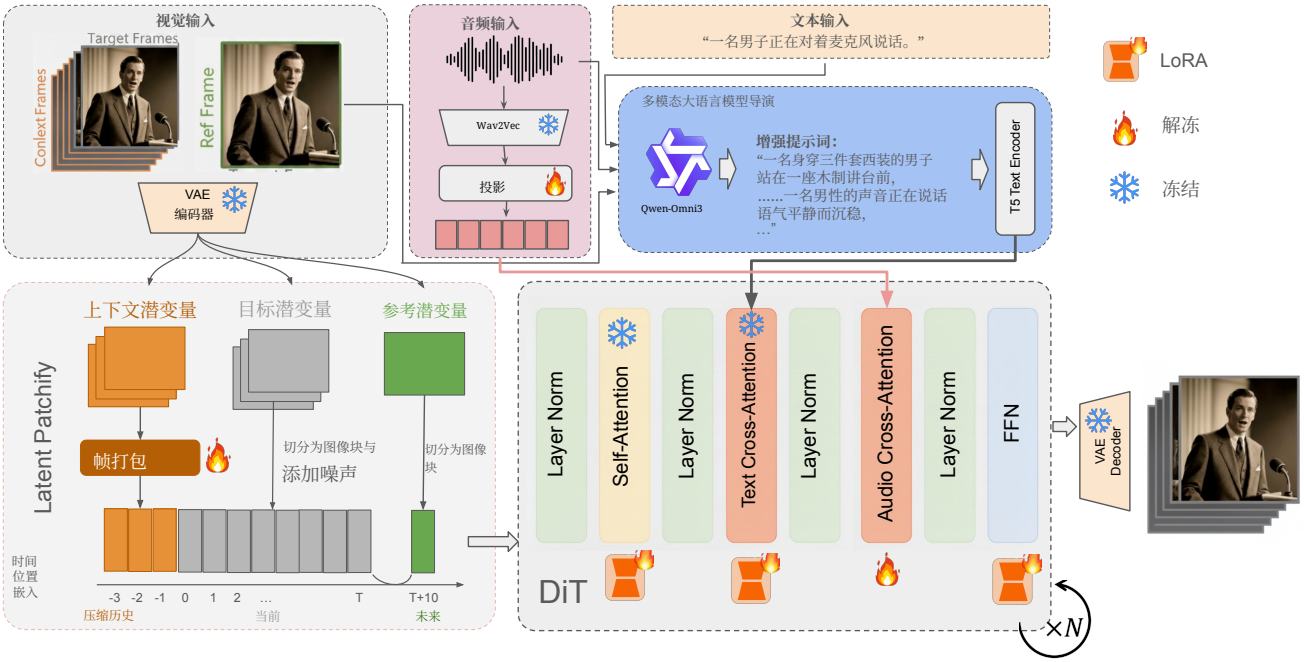


图 4。TalkVerse-5B 模型架构。Wav2Vec特征经过投影，并通过稀疏音频交叉注意力注入。FramePack 压缩上下文；参考图像位置嵌入可防止滑动窗口之间发生漂移。多模态大语言模型导演根据音频和参考帧重写提示词。变分自编码器采用 (t,h,w) =(4,16,16) 下采样，相较于典型设置实现 4× 词元缩减。

这种设计的抗漂移能力总体优于采用首帧图像条件的方法 [23, 38]。训练时，视频时长只需达到运动帧上下文与去噪视频时长之和，例如在我们的设置中为 3 秒 + 4，远短于一分钟的目标时长。

4.1.1. 音频条件注入

我们使用 Wav2Vec [1], 对多层 Wav2Vec特征进行聚合，并通过投影层融合来自不同语义层级的音频线索；随后使用因果一维卷积对其进行压缩，使每个潜在帧都能接收时间对齐的音频片段。音频特征通过从视觉词元到音频词元的逐潜变量交叉注意力注入 DiT 模块，从而确保精确的唇音频同步。音频交叉注意力以稀疏方式应用于选定的 DiT 模块，以降低计算开销，并采用端到端方式进行训练。

4.1.2. 长视频生成

长视频中的图像条件与防漂移。 我们发现，图像条件对于保持长视频中视觉外观的一致性至关重要。我们采用略长于去噪视频最后一个潜变量的 位置嵌入（例如，10 个潜变量）来引导生成。这种分配方式使生成过程在未来“追随”参考图像，而不是在当前步骤复制它，从而有效地将参考图像中的人体姿势与训练视频中的姿势解耦，并避免复制粘贴伪影。因此，在推理时，我们可以在不同滑动窗口间共享同一张参考图像，而不会引发循环的身体动作；相比之下，首帧条件往往会将每个窗口的起始潜变量重置为与参考姿势完全一致。最后，这种方案还支持为不同窗口注入不同的参考图像，从而逐步改变背景，同时保持主体外观一致，生成更具电影感的长视频。

上下文历史帧。 遵循 Wan-S2V [24], ，我们将一段此前运动帧窗口作为上下文，并使用 FramePack 模块 [81] 对其进行打包，以保持各个短视频片段（例如 5 秒）之间的平滑过渡。FramePack 能有效压缩运动帧，减少开销巨大的三维自注意力（时间、宽度和高度）中的令牌总数。分块嵌入通过数据驱动的方式与音频交叉注意力训练联合训练。为使 {运动帧上下文、去噪视频、参考图像} 的位置嵌入保持一致，我们为运动帧分配负位置嵌入，以表明上下文是历史信息。

4.1.3. 多模态大语言模型导演

在分钟级视频生成中，文本提示中可能包含引导生成的故事线。因此，能够接收用户提示和音频/图像特征，并对其进行改写以生成多样且准确的视频字幕的多模态大语言模型导演，对生成过程至关重要。在这一阶段，改写后的视频字幕还能——



图 5. 推理过程中使用视频编辑技巧，借助预训练的音频驱动参考图像生成视频模型进行视频配音。

通过匹配训练阶段的文本提示风格来缩小训练-推理差距。为将多模态指令融入连贯的计划，我们使用 Qwen3-Omni [72] 作为多模态大语言模型导演。它从音频中提取情绪，从参考帧中提取场景/外观描述符，然后将其与用户提示词结合，生成简洁的故事线。故事线优先考虑用户意图，其次是音频线索，再次是图像参考，并逐一列出角色特征、背景布局、动作、情绪变化、视觉风格和摄像机规划。我们采用少样本上下文模板来稳定输出。最终确定的故事线通过交叉注意力作为文本条件注入，从而在整个生成过程中提供一致的引导。因此，经过训练的音频驱动的DiT模型能够确保长视频中的唇形同步、视觉外观一致性和平滑的窗口过渡，同时多模态大语言模型导演能够引导生成有趣且符合用户提示和音频/情绪风格的内容。

4.1.4. 训练技术。

由于采用更高压缩比变分自编码器的 Wan2.2-5B 在微调期间更容易产生伪影，我们发现全参数调优并不适用于 5B模型，这与 14B 模型的结果不同 [24]。我们通过低秩适配（LoRA） [32]对所有扩散变换器权重进行音频驱动微调，并对音频投影、交叉注意力以及与 FramePack 相关的权重进行全参数微调。具体而言，对于预训练权重矩阵 W，我们引入低秩残差ΔW，并得到适配后的权重 $W' = W + \Delta W$，其中ΔW 分解为 $\Delta W = AB^\top$，$A \in \mathbb{R}^{n\times d}$，$B \in \mathbb{R}^{m\times d}$，$d \ll \min(n, m)$。通过仅训练扩散变换器模块内部注意力模块和 MLP块中的小矩阵 A 和 B （并保持 W 冻结），我们大幅降低了内存和计算资源开销，同时保留基础模型的先验知识。这使模型能够更稳定地适配同步语音线索，并将视觉伪影降至最低。为了加快音频条件注入的收敛速度，我们将

表 2. 在 EMTD 测试集上的定量比较。 † 我们的实验结果来自 Wan-S2V-14B 官方检查点。

方法	FID↓	FVD↓	Sync-C↑	HKC↑	HKV↑	CSIM↑
EchoMimicV2 [45]	33.42	217.71	4.44	0.425	0.150	0.519
EMO2 [56]	27.28	129.41	4.58	0.553	0.198	0.650
FantasyTalking [64]	22.60	178.12	3.00	0.281	0.087	0.626
混元数字人 [13]	18.07	145.77	4.71	0.379	0.145	0.583
Wan-S2V-14B [24]	15.66	129.57	6.41†	0.435	0.142	0.677
TalkVerse-5B	18.66	147.21	5.47	0.412	0.151	0.658

音频交叉注意力层的权重初始化为文本交叉注意力权重，而非采用随机初始化。模型训练的主要难点在于加入音频条件控制能力，并平滑上下文帧过渡，同时不损害文本条件控制能力。为避免新层干扰预训练模型的知识，我们使用一个零初始化的线性层，将音频交叉注意力层连接到主扩散变换器模块，类似于 ControlNet [80]。LoRA 训练也有助于保留文本条件控制先验。

4.2. 应用于视频配音

我们发现，一旦音频驱动的图像转视频生成模型训练完成，它便能以零样本方式泛化到音频驱动的视频转视频生成。如图5所示，给定输入视频后，我们调整生成流程，以便灵活控制内容保留和运动合成。我们不再以纯高斯噪声初始化扩散过程，而是在指定噪声水平下，向编码后的视频潜变量中注入受控噪声 $\alpha \in [0, 1]$。具体而言，我们首先使用 VAE 编码器将输入视频编码到潜空间中，然后按以下方式初始化去噪过程：

$$\mathbf{z}_t = (1-\alpha)\mathbf{z}_0 + \alpha\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \tag{1}$$

其中，$\mathbf{z}_0$ 表示干净的视频潜变量，$\mathbf{z}_t$ 则是带噪声的初始化结果。随后，去噪过程仅从时间步 $t = \alpha T$ 进行到 $t = 0$，其中 T 为时间步总数。

我们通过实证将 $\alpha = 0.95$ 设为默认值，因为观察到，在基于流的扩散模型中，运动动态主要是在高噪声去噪步骤中合成的。较高的噪声水平使模型能够生成新的音频同步运动和表情，而其余低噪声步骤则用于细化细节，并保留原视频的场景结构。噪声强度 α 可直观地控制配音效果：较低的值（$\alpha \to 0$）能保留更多原视频内容，而较高的值（$\alpha \to 1$）则能实现更灵活的运动生成。此外，为了在多片段生成过程中保持各片段间的时间一致性，我们会动态更新参考图像，方法是从每个视频片段中随机采样一帧，确保参考



图 6. 与 EchoMimicV 的定性比较3 [44] （案例 1 和 2）以及 OmniAvatar [23] （案例 3），按从上到下的顺序排列。

条件输入在整个视频过程中始终与当前场景上下文保持一致。

5. 实验

5.1. 实现细节

我们遵循 Wan2.2-5B，并在训练期间使用与 720p 面积匹配的分辨率分桶。我们以预训练 Wan2.2-5B 为起点，在 TalkVerse 上从头训练新引入的参数（即音频投影、交叉注意力和与 FramePack 相关的权重），同时冻结 DiT，并对 DiT 应用低秩适配微调。音频层稀疏注入 DiT 模块 $3k$ 的 $k = 0$、.....、9 以及最后一个模块中，共计 58 亿个参数。由于 GPU 显存限制，训练采用分布式数据并行，每次最多使用 73 个上下文 + 80 视频帧，在 64 个图形处理器上每个图形处理器的批量大小为 1。低秩适配的秩和 alpha 均为 128。我们使用 AdamW，学习率分别为 1e-5（完整微调参数）和 1e-4（低秩适配）。对于 FramePack 的训练，我们采用混合长度采样：如果视频 ≥73+80 帧，则随机采样一个 73+80 窗口（前 73 帧作为上下文，后 80 帧作为视频）；否则，将上下文潜变量置零，并随机采样一个 80 帧的视频窗口。为与我们方法中的位置嵌入保持一致，我们在视频窗口之后按时间顺序采样参考图像。对于文本和音频，我们均使用 CFG=6.5。由于计算资源限制，我们仅训练一周（∼1个 epoch），并在本节报告这些结果。我们认为，通过更多的计算资源，性能还可以进一步优化

计算。

指标。 由于大多数开源基准都侧重于说话头像，我们在 EMTD 数据集 [45] 上进行评估，并按照 Wan-S2V-14B [24] 的设置，与 EchoMimicV2 [45], 、EMO2 [56], 、FantasyTalking [64], 和 Hunyuan-Avatar [13]进行比较。我们报告 FID [29], 、FVD [59], 、唇形同步 Sync-C [15], 、ArcFace [19]相似度（CSIM）、手部关键点置信度（HKC）和手部关键点方差（HKV）。由于 PSNR 和 SSIM 在音频条件下效果不佳，因此我们不予报告。

5.2. 主要结果

唇形同步与身份一致性。 我们的 5B模型在 FID/FVD/唇形同步/手部质量/身份一致性方面的表现可与最先进的方法相媲美。与基于 Wan2.1-14B模型的 Wan-S2V-14B 和 FantasyTalking 相比，这一表现所需的推理 $11.2\times$ 计算资源更低，原因是 $4\times$ 我们采用了更高的 VAE压缩，减少了潜在令牌数量，此外还带来 $\approx 2.8\times$ 由于使用 5B 而非 14B 参数而实现的加速（14/5）。Hunyuan-Avatar 还使用了大型的 13B Hunyuan-Video 视频模型，其速度明显慢于我们的方法。在图 6中，我们展示了与 EchoMimicV3 [44] 和 OmniAvatar [23]的定性比较。与开源学术方法相比，我们的模型能够生成更自然的身体动作，并保持更好的主体一致性。

长视频生成。 我们还以 25 帧每秒对模型进行 40 秒长视频生成评估，并与 Sta-



图 7. 使用 StableAvatar 进行 40 秒长视频生成（25 帧每秒）的定性比较 [58]。



图 8. 扩散变换器权重的低秩适配训练与全量微调的定性比较。

bleAvatar [58]见图 7。可以看到，经过 1000 帧后，我们的模型在长视频外观保持能力方面可与 StableAvatar 媲美，足以满足大多数实际应用场景。

5.3. 消融实验

LoRA 训练。 我们在图 8 中比较了对 DiT权重进行 LoRA 训练与全面微调的效果。我们可以看到，与对 DiT权重进行全面微调相比，LoRA 训练能够更好地保持视觉质量，尤其是在关节手部和物体方面。我们的观察是，小型预训练扩散变换器模型（例如 5B模型或 1.3B模型）在微调过程中可能比强大的 14B模型更容易出现视觉伪影。然而，笨重的 14B模型在应用中进行推理时可能慢得难以接受，并阻碍社区的广泛参与。我们认为，5B模型在性能和推理速度之间实现了良好平衡，配合一种合适的训练策略。此外，通过适当的蒸馏和工程优化，它还有望实现实时运行。

6. 结论

我们推出了 TalkVerse，这是一个大规模音频同步视频语料库，面向以音频或姿势为条件的单人视频生成。该数据集包含 230 万个片段（6300 小时），分辨率为 720p/1080P，每个片段都配有时间对齐的语音音频、DWPose 骨骼序列以及视频/音频字幕。基于这些数据，我们训练了一个 Wan2.2-5B 扩散变换器，其潜空间速度显著快于广泛使用的设置。尽管模型规模较小，其感知质量和唇形同步效果仍可与更大的 14B 系统相媲美，且 10×推理成本更低。这得益于通过稀疏交叉注意力注入由聚合 Wav2Vec 特征构成的音频条件、基于 FramePack 的历史压缩以实现平滑的窗口过渡，以及参考图像位置嵌入策略，使持续数分钟的生成漂移更少。多模态大语言模型导演（Qwen3- Omni）通过将音频和图像线索作为依据来改写用户提示词，从而生成连贯且风格一致的故事线。通过受控噪声注入，我们的模型还能以零样本方式泛化至视频配音任务。

局限性与未来工作。 5B模型在复杂、极端的手部关节动作和全身动作方面可能表现不佳。尽管该数据集可支持姿态驱动视频生成和音视频联合生成等其他任务，但由于计算资源限制，本文未开展这些实验。未来，我们将调整该数据集以适用于这些任务。

TalkVerse：让人人都能进行分钟级音频驱动的视频生成

补充材料

7. 关于实验的更多细节

7.1. 训练细节

我们使用 AdamW（10^{-5} 学习率）对 Wan S2V 5B 模型和新增的音频交叉注意力进行微调，并选择性地解冻时间嵌入、投影层、输出头以及可选的 Transformer 模块。DiT 中的低秩适配参数缩放至 $10\times$，使其达到 10^{-4}。训练时，每个图形处理器的批次大小为 1，并通过梯度累积将总批次大小设为 256，同时进行梯度裁剪（范数为 1.0）。为进行无分类器引导，我们对文本、图像和音频条件应用 10% 的丢弃率。启用 FramePack 模块时，我们会将运动帧和目标帧联合编码到变分自编码器空间中进行训练，以维持长视频生成中的时间连续性。否则，变分自编码器空间中的运动上下文帧为全零张量。

7.2. 空间损失加权

我们使用 ROI 损失，通过 DWPose 提高人脸和身体区域重建误差的权重 [76]，以估计身体和人脸的边界框及关键点，并将其归一化后栅格化为与变分自编码器空间分辨率对齐的二值掩码，随后在扩散去噪目标上构建空间损失加权。具体而言，令 L_{full}、L_{body}和 L_{face} 分别表示整帧、身体边界框和人脸边界框上的逐像素损失。最终训练损失为

$$L = \frac{1 \cdot L_{\text{full}} + 1 \cdot L_{\text{body}} + 1 \cdot L_{\text{face}}}{Z}, \qquad (2)$$

其中，Z 是归一化因子（例如权重之和），因此身体框和人脸框内的区域相比背景具有更高的相对重要性。

7.3. 数据预处理流程

我们整合多个来源的元数据，构建视频-音频-文本数据集，并确保所有片段都包含有效的 25帧/秒视频流和音频流。为控制分辨率，我们采用分辨率分桶策略（例如 720p），将每个片段调整大小并居中裁剪至最接近的目标分桶。我们采样一段连续帧序列用于去噪，可选择在序列前添加运动上下文帧，并从附近的窗口中提取条件参考图像。

音频会重采样至 16千赫，并与视频片段严格对齐，必要时进行填充。最后，我们加载缓存的文本嵌入（由 T5文本编码器预先提取）以及用于无分类器引导的空嵌入，从而为训练提供一套全面同步的输入。

表 3. 我们的 5B模型的超参数。

超参数	值
语言编码器	T5-XXL
VAE	Wan2.2-VAE
变分自编码器步幅	(4，16，16)
音频编码器	wav2vec2-large-xlsr-53-english
补丁大小	(1，2，2)
模型维度	3072
前馈网络维度	14336
频率嵌入维度	256
注意力头	24
Transformer 层	30
QK 归一化 / 交叉注意力归一化	是 / 是
音频维度	1024
音频注入层	$\{0, 3, 6, 9, 12, 15, 18, 21, 24, 27, 29\}$
运动帧	73
潜在维度	48
ν-尺度	1.0
采样帧率	25帧/秒
采样帧数	120
扩散偏移	5
默认采样步数	50

7.4. 推理过程

我们的推理流程支持多种分辨率下的图像转视频生成和视频转视频“配音”。我们将文本提示（T5）、音频（wav2vec2）和参考图像（变分自编码器）编码，用于对扩散模型进行条件控制。对于配音，我们不从纯噪声开始，而是将输入视频编码为潜变量，在目标时间步 t_{dub} 添加噪声，然后从该状态开始去噪，以保留粗略结构，同时同步唇部运动。潜变量序列通过流匹配求解器（例如 UniPC，50 步）迭代去噪，再由变分自编码器解码。关于 CFG 引导尺度，我们发现取 3 到 6 能获得良好效果，这是在（1）视觉质量和运动强度，以及（2）唇形同步质量之间的权衡。CFG 权重越大，唇形同步质量越好，但视觉质量和运动强度越差。

7.5. Wan S2V 5B 超参数设置

我们在表格中总结了实验所用的主要配置3。

参考文献

[1] Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli。《wav2vec 2.0：语音表征自监督学习框架》。神经信息处理系统进展，33：12449–12460，2020。5[2] Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang 等。《Qwen2.5-vl 技术报告》。arXiv 预印本 arXiv:2502.13923，2025。2，3[3] Max Bain、Arsha Nagrani、Gul Varol 和 Andrew Zisserman。《Frozen in Time：用于端到端检索的视频与图像联合编码器》。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第1728–1738页，2021年。2，4 – 。[4]Omer Bar-Tal、Hila Chefer、Omer Tov、Charles Herrmann、Roni Paiss、Shiran Zada、Ariel Ephrat、Junhwa Hur、Yuanzhen Li、Tomer Michaeli 等。《Lumiere：用于视频生成的时空扩散模型》。arXiv 预印本 arXiv:2401.12945，2024年。1[5] Michael J. Black、Priyanka Patel、Joachim Tesch 和 Jinlong Yang。《BEDLAM：一个呈现细致逼真动画运动的人体合成数据集》。载于 CVPR，第8726–8737页。IEEE，2023年。2[6] Andreas Blattmann、Tim Dockhorn、Sumith Kulal、Daniel Mendelevitch、Maciej Kilian、Dominik Lorenz、Yam Levi、Zion English、Vikram Voleti、Adam Letts 等。《Stable Video Diffusion：将潜在视频扩散模型扩展至大规模数据集》。arXiv 预印本 arXiv:2311.15127，2023年。1[7] Andreas Blattmann、Robin Rombach、Huan Ling、Tim Dockhorn、Seung Wook Kim、Sanja Fidler 和 Karsten Kreis。《对齐你的潜变量：利用潜变量扩散模型进行高分辨率视频合成》。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第22563–22575页，2023年。[8]Tim Brooks、Janne Hellsten、Miika Aittala、Ting-Chun Wang、Timo Aila、Jaakko Lehtinen、Ming-Yu Liu、Alexei Efros 和 Tero Karras。《生成动态场景的长视频》。神经信息处理系统进展，35：31769–31781，2022年。1[9] Fabian Caba Heilbron、Victor Escorcia、Bernard Ghanem 和 Juan Carlos Niebles。《ActivityNet：用于人类活动理解的大规模视频基准》。载于 IEEE计算机视觉与模式识别会议论文集，第961–970页，2015年。2，4[10] Brandon Castellano。PySceneDetect。3[11] Jiaben Chen、Zixin Wang、Ailing Zeng、Yang Fu、Xueyang Yu、Siyuan Cen、Julian Tanke、Yihang Chen、Koichi Saito、Yuki Mitsufuji 等。《TalkCuts：用于多镜头人类语音视频生成的大规模数据集》。arXiv 预印本 arXiv:2510.07249，2025年。2，4[12]Tsai-Shien Chen、Aliaksandr Siarohin、Willi Menapace、Ekaterina Deyneka、Hsiang-wei Chao、Byung Eun Jeon、Yuwei Fang、Hsin-Ying Lee、Jian Ren、Ming-Hsuan Yang 等。《Panda-70M：利用多个跨模态教师为7000万条视频生成描述》。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第13320–13331页，2024年。1，2，3，4

[13] Yi Chen、Sen Liang、Zixiang Zhou、Ziyao Huang、Yifeng Ma、Junshu Tang、Qin Lin、Yuan Zhou 和 Qinglin Lu。混元视频头像：面向多个角色的高保真音频驱动人体动画。arXiv预印本 arXiv:2505.20156，2025年。1、2、6、7[14] Zhiyuan Chen、Jiajiong Cao、Zhiquan Chen、Yuming Li 和 Chenguang Ma。Echomimic：通过可编辑的关键点条件实现逼真的音频驱动肖像动画。arXiv预印本 arXiv:2407.08136，2024年。1[15] Joon Son Chung 和 Andrew Zisserman。时光之外：野外环境下的自动唇形同步。载于《计算机视觉—ACCV 2016研讨会：ACCV 2016国际研讨会》，中国台湾台北，2016年11月20日至24日，修订版精选论文，第二部分

13， 251 263 2017 4, 7 页码 – 。

[16] Springer，。Joon Son Chung、Arsha Nagrani 和 Andrew Zisserman。Voxceleb2：深度说话人识别。载于 INTERSPEECH，第1086–1090页。ISCA，2018年。4[17] Enric Corona、Andrei Zanfir、Eduard Gabriel Bazavan、Nikos Kolotouros、Thiemo Alldieck 和 Cristian Sminchisescu。Vlogger：用于具身虚拟形象合成的多模态扩散。arXiv预印本 arXiv:2403.08764，2024年。2[18] Jiahao Cui、Hui Li、Yun Zhan、Hanlin Shang、Kaihui Cheng、Yuqi Ma、Shan Mu、Hang Zhou、Jingdong Wang 和 Siyu Zhu。Hallo3：基于扩散变换器网络的高动态、逼真肖像图像动画。arXiv预印本 arXiv:2412.00733，2024年。1、2[19] Jiankang Deng、Jia Guo、Niannan Xue 和 Stefanos Zafeiriou。Arcface：用于深度人脸识别的加性角度间隔损失。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第4690–4699页，2019年。7[20] Donglin Di、He Feng、Wenzhang Sun、Yongjia Ma、Hao Li、Wei Chen、Lei Fan、Tonghua Su 和 Xun Yang。Dh-facevid-1k：用于人脸视频生成的大规模高质量数据集。载于《IEEE/CVF国际计算机视觉会议论文集》，第12124–12134页，2025年。4[21] Yikang Ding、Jiwen Liu、Wenyuan Zhang、Zekun Wang、Wentao Hu、Liyuan Cui、Mingming Lao、Yingchao Shao、Hui Liu、Xiaohan Li 等。Kling-avatar：为级联式长时长虚拟形象动画合成奠定基础的多模态指令。arXiv预印本 arXiv:2509.09595，2025年。1、2[22] Ariel Ephrat、Inbar Mosseri、Oran Lang、Tali Dekel、Kevin Wilson、Avinatan Hassidim、William T. Freeman 和 Michael Rubinstein。鸡尾酒会上的聆听：用于语音分离的说话人无关音视频模型。《ACM图形学汇刊》，37(4):112，2018年。1[23] Qijun Gan、Ruizi Yang、Jianke Zhu、Shaofei Xue 和 Steven Hoi。Omniavatar：通过自适应身体动画实现高效的音频驱动虚拟形象视频生成。arXiv预印本 arXiv:2506.18866，2025年。1、2、5、7[24] Xin Gao、Li Hu、Siqi Hu、Mingyang Huang、Chaonan Ji、Dechao Meng、Jinwei Qi、Penchong Qiao、Zhen Shen、Yafei Song 等。Wan-s2v：音频驱动的电影级视频生成。arXiv预印本 arXiv:2508.18621，2025年。1、2、4、5、6、7[25] Chunhui Gu、Chen Sun、David A Ross、Carl Vondrick、Caroline Pantofaru、Yeqing Li、Sudheendra Vijayanarasimhan、

George Toderici、Susanna Ricco、Rahul Sukthankar 等。Ava：时空定位的原子视觉动作视频数据集。载于《IEEE计算机视觉与模式识别会议论文集》，第6047–6056页，2018年。2 – 。[26] Yuwei Guo、Ceyuan Yang、Anyi Rao、Zhengyang Liang、Yaohui Wang、Yu Qiao、Maneesh Agrawala、Dahua Lin 和 Bo Dai。AnimateDiff：无需特定调优即可动画化个性化文本到图像扩散模型。载于《国际学习表征会议》（ICLR），2023年。1[27] Agrim Gupta、Lijun Yu、Kihyuk Sohn、Xiuye Gu、Meera Hahn、李飞飞、Irfan Essa、Lu Jiang 和 Jos´e Lezama。基于扩散模型的照片级真实感视频生成。arXiv预印本 arXiv:2312.06662，2023年。1[28] Tianyu He、Junliang Guo、Runyi Yu、Yuchi Wang、Jialiang Zhu、Kaikai An、Leyi Li、Xu Tan、Chunyu Wang、Han Hu 等。Gaia：零样本说话数字人生成。arXiv预印本 arXiv:2311.15230，2023年。1[29]Martin Heusel、Hubert Ramsauer、Thomas Unterthiner、Bernhard Nessler 和 Sepp Hochreiter。采用双时间尺度更新规则训练的生成对抗网络会收敛至局部纳什均衡。神经信息处理系统进展，第30卷，2017年。7[30] Jonathan Ho、Tim Salimans、Alexey Gritsenko、William Chan、Mohammad Norouzi 和 David J Fleet。视频扩散模型。神经信息处理系统进展（NeurIPS），第35卷：8633–8646，2022年。1[31] Steven Hogue、Chenxu Zhang、Hamza Daruger、Yapeng Tian 和 Xiaohu Guo。DiffTED：通过基于扩散的伴随语音手势实现单次音频驱动的TED演讲视频生成。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第1922–1931页，2024年。2 – 。[32] Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。低秩适配：大语言模型的低秩适配。arXiv预印本 arXiv:2106.09685，2021年。6[33] Li Hu。Animate Anyone：用于角色动画的连贯且可控的图像到视频合成。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第8153–8163页，2024年。2[34] Yasamin Jafarian 和 Hyun Soo Park。通过观看社交媒体舞蹈视频，学习穿着衣服的人体的高保真深度。载于 CVPR，第12753–12762页。计算机视觉基金会 / IEEE，2021年。4[35] Yasamin Jafarian 和 Hyun Soo Park。通过观看社交媒体舞蹈视频，学习穿着衣服的人体的高保真深度。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第12753–12762页，2021年。2[36] Jianwen Jiang、Chao Liang、Jiaqi Yang、Gaojie Lin、Tianyun Zhong 和 Yanbo Zheng。Loopy：通过长期运动依赖驾驭音频驱动的肖像虚拟形象。arXiv预印本 arXiv:2409.02634，2024年。1、2[37] Jianwen Jiang、Weihong Zeng、Zerong Zheng、Jiaqi Yang、Chao Liang、Wang Liao、Han Liang、Yuan Zhang 和 Mingyuan Gao。Omnihuman-1.5：通过认知模拟赋予虚拟形象主动思维。arXiv预印本 arXiv:2508.19209，2025年。1[38] Zhe Kong、Feng Gao、Yong Zhang、Zhuoliang Kang、Xiaoming Wei、Xunliang Cai、Guanying Chen 和 Wenhan Luo。让他们交谈：音频驱动的多人对话视频生成。arXiv预印本 arXiv:2505 22647，2025年 5 . . [39] Hui Li、Mingwang Xu、Yun Zhan、Shan Mu、Jiaye Li、Kaihui Cheng、Yuxuan Chen、Tan Chen、Mao Ye、Jingdong Wang 等。Openhumanvid：用于增强以人为中心的视频生成的大规模高质量数据集。arXiv预印本 arXiv:2412.00115，2024年。1、2、3、4[40] Yitong Li、Martin Min、Dinghan Shen、David Carlson 和 Lawrence Carin。文本生成视频。载于《AAAI人工智能会议论文集》，2018年。1[41] Gaojie Lin、Jianwen Jiang、Chao Liang、Tianyun Zhong、Jiaqi Yang 和 Yanbo Zheng。Cyberhost：通过区域码本注意力驾驭音频驱动的虚拟形象扩散模型。arXiv预印本 arXiv:2409.01876，2024年。1、2[42] Gaojie Lin、Jianwen Jiang、Jiaqi Yang、Zerong Zheng 和 Chao Liang。Omnihuman-1：重新思考单阶段条件人类动画模型的规模扩展。arXiv预印本 arXiv:2502.01061，2025年。1、2[43] Shanchuan Lin、Xin Xia、Yuxi Ren、Ceyuan Yang、Xuefeng Xiao 和 Lu Jiang。用于单步视频生成的扩散对抗后训练。arXiv预印本 arXiv:2501 08316，2025年 1 . . [44] Rang Meng、Yan Wang、Weipeng Wu、Ruobing Zheng、Yuming Li 和 Chenguang Ma。Echomimicv3：统一多模态、多任务人体动画只需13亿个参数。arXiv预印本 arXiv:2507.03905，2025年。7[45] Rang Meng、Xingyu Zhang、Yuming Li 和 Chenguang Ma。Echomimicv2：迈向引人注目、简化且半身的人体动画。载于 CVPR，第5489–5498页。计算机视觉基金会 / IEEE，2025年。2、6、7[46] Antoine Miech、Dimitri Zhukov、Jean-Baptiste Alayrac、Makarand Tapaswi、Ivan Laptev 和 Josef Sivic。HowTo100M：通过观看一亿个带旁白的视频片段学习文本-视频嵌入。载于 IEEE/CVF国际计算机视觉大会论文集，第2630 2640, 2019 2 – 页。A Nagrani、J Chung 和 A [47] Zisserman。Voxceleb：大型说话人识别数据集。INTERSPEECH 2017，2017年。2[48] Kepan Nan、Rui Xie、Penghao Zhou、Tiehan Fan、Zhenheng Yang、Zhijie Chen、Xiang Li、Jian Yang 和 Ying Tai。OpenVid-1M：用于文本生成视频的大规模高质量数据集。arXiv预印本 arXiv:2407.02371，2024年。2、4[49] Amir Shahroudy、Jun Liu、Tian-Tsong Ng 和 Gang Wang。NTU RGB+ D：用于3D人体活动分析的大型数据集。载于 IEEE计算机视觉与模式识别会议论文集，第1010–1019页，2016年。2、4[50] Ruizhi Shao、Youxin Pang、Zerong Zheng、Jingxiang Sun 和 Yebin Liu。Human4dit：采用四维扩散变换器的自由视角人体视频生成。arXiv预印本 arXiv:2405.17405，2024年。2

[51] Aliaksandr Siarohin、St´ephane Lathuili`ere、Sergey Tulyakov、Elisa Ricci 和 Nicu Sebe。一阶运动模型用于图像动画。神经信息处理系统进展，32，2019年。2[52] Aliaksandr Siarohin、Oliver J Woodford、Jian Ren、Menglei Chai 和 Sergey Tulyakov。关节式动画的运动表示。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第13653–13662页，2021年。2[53] Uriel Singer、Adam Polyak、Thomas Hayes、Xi Yin、Jie An、Songyang Zhang、Qiyuan Hu、Harry Yang、Oron Ashual、Oran Gafni 等。Make-a-video：无需文本-视频数据的文本生成视频。arXiv预印本 arXiv:2209 14792，2022年 1 . . [54] K Soo mro。UCF101：野外视频中包含101类人体动作的数据集。arXiv 预印本 arXiv:1212 0402，2012年 2、4 . . [55] Michal Stypulkowski、Konstantinos Vougioukas、Sen He、Maciej Zieba、Stavros Petridis 和 Maja Pantic。扩散人头：扩散模型在说话人脸生成方面胜过生成对抗网络。载于 IEEE/CVF冬季计算机视觉应用会议论文集，第5091–5100页，2024年。1[56] Linrui Tian、Siqi Hu、Qi Wang、Bang Zhang 和 Liefeng Bo。E MO2：末端执行器引导的音频驱动虚拟形象视频生成。arXiv预印本 arXiv:

2501 10687, 2025 2, 6, 7 . . [57] Linrui Tian、Qi Wang、Bang Zhang 和 Liefeng Bo。Emo：在弱条件下，利用音频到视频扩散模型生成富有表现力的肖像视频，让肖像栩栩如生。载于欧洲计算机视觉会议，第244–260页。Springer，2025年。1、2[58] Shuyuan Tu、Yueming Pan、Yinming Huang、Xintong Han、Zhen Xing、Qi Dai、Chong Luo、Zuxuan Wu 和 Yu-Gang Jiang。StableAvatar：无限长度音频驱动虚拟形象视频生成。arXiv预印本 arXiv:2508.08248，2025年。1、2、8[59] Thomas Unterthiner、Sjoerd van Steenkiste、Karol Kurach、Rapha¨el Marinier、Marcin Michalski 和 Sylvain Gelly。FVD：一种用于视频生成的新指标。7[60] Ruben Villegas、Mohammad Babaeizadeh、Pieter-Jan Kindermans、Hernan Moraldo、Han Zhang、Mohammad Taghi Saffar、Santiago Castro、Julius Kunze 和 Dumitru Erhan。Phenaki：根据开放领域文本描述进行可变长度视频生成。载于国际学习表征会议，2022年。1[61] Ang Wang、Baole Ai、Bin Wen、Chaojie Mao、Chen-Wei Xie、Di Chen、Feiwu Yu、Haiming Zhao、Jianxiao Yang、Jianyuan Zeng 等。Wan：开放且先进的大规模视频生成模型。arXiv预印本 ar Xiv:2503.20314，2025年。2[62] Cong Wang、Kuan Tian、Jun Zhang、Yonghang Guan、Feng Luo、Fei Shen、Zhiwei Jia ng、Qing Gu、Xiao Han 和 Wei Yang。V-Express：用于渐进式肖像视频生成训练的条件丢弃。arXiv预印本 arXiv:2406.02511，2024年。1[63] Jiuniu Wang、Hangjie Yuan、Dayou Chen、Yingya Zhang、Xiang Wang 和 Shiwei Zhang。ModelScope文本到视频技术报告。arXiv预印本 arXiv:2308.06571，2023年。1 [64] Mengchao Wang、Qiang Wang、Fan Jiang、Yaqi Fan、Yunpeng Zhang、Yonggang Qi、Kun Zhao 和 Mu Xu。FantasyTalking：通过连贯性实现逼真说话肖像生成

运动合成。arXiv 预印本 arXiv:2504.04842，2025年。6, 7[65] Ting-Chun Wang、Arun Mallya 和 Ming-Yu Liu。用于视频会议的单次自由视角神经说话头合成。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 10039–10049 页，2021年。2, 4 – 。[66] Yaohui Wang、Piotr Bilinski、Francois Bremond 和 Antitza Dantcheva。Imaginator：用于视频生成的条件时空生成对抗网络。载于 IEEE/CVF 冬季应用计算机视觉会议论文集，第 1160–1169 页，2020年。1[67] Zhenzhi Wang、Yixuan Li、Yanhong Zeng、Youqing Fang、Yuwei Guo、Wenran Liu、Jing Tan、Kai Chen、Tianfan Xue、Bo Dai 和 Dahua Lin。Humanvid：揭示相机可控的人体图像动画训练数据。载于神经信息处理系统进展，2024年。2, 4[68] Zhenzhi Wang、Yixuan Li、Yanhong Zeng、Yuwei Guo、Dahua Lin、Tianfan Xue 和 Bo Dai。基于结构化视频扩散的多身份人体图像动画。arXiv 预印本 arXiv:2504.04126，2025年。2[69] Zhenzhi Wang、Jiaqi Yang、Jianwen Jiang、Chao Liang、Gaojie Lin、Zerong Zheng、Ceyuan Yang 和 Dahua Lin。Interacthuman：布局对齐音频条件下的多概念人体动画。arXiv 预印本 arXiv:2506 09984，2025年 2 。[70] Haoning Wu、Zicheng Zhang、Weixia Zhang、Chaofeng Chen、Liang Liao、Chunyi Li、Yixuan Gao、Annan Wang、Erli Zhang、Wenxiu Sun 等。Q-align：通过离散文本定义等级训练大型多模态模型进行视觉评分。arXiv 预印本 arXiv:2312.17090，2023年。3[71] Liangbin Xie、Xintao Wang、Honglun Zhang、Chao Dong 和 Ying Shan。Vfhq：用于视频人脸超分辨率的高质量数据集和基准。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 657–666 页，2022年。2[72] Jin Xu、Zhifang Guo、Hangrui Hu、Yunfei Chu、Xiong Wang、Jinzheng He、Yuxuan Wang、Xian Shi、Ting He、Xinfa Zhu 等。Qwen3-omni 技术报告。arXiv 预印本 arXiv:2509.17765，2025年。2, 3, 6[73] Mingwang Xu、Hui Li、Qingkun Su、Hanlin Shang、Li-wei Zhang、Ce Liu、Jingdong Wang、Luc Van Gool、Yao Yao 和 Siyu Zhu。Hallo：用于肖像图像动画的分层音频驱动视觉合成。arXiv 预印本 arXiv:2406.08801，2024年。1[74] Sicheng Xu、Guojun Chen、Yu-Xiao Guo、Jiaolong Yang、Chong Li、Zhenyu Zang、Yizhong Zhang、Xin Tong 和 Baining Guo。Vasa-1：实时生成逼真的音频驱动说话人脸。arXiv 预印本 arXiv:2404 10667，2024年 1 。[75] Hongwei Xue、Tiankai Hang、Yanhong Zeng、Yuchong Sun、Bei Liu、Huan Yang、Jianlong Fu 和 Baining Guo。借助大规模视频转录推进高分辨率视频语言表征。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 5036–5045 页，2022年。2

[76] Zhendong Yang、Ailing Zeng、Chun Yuan 和 Yu Li。采用两阶段蒸馏的有效全身姿态估计。载于 IEEE/CVF国际计算机视觉大会论文集，第 4210–4220 页，2023年。2, 3, 4, 1[77] Zhenhui Ye、Ziyue Jiang、Yi Ren、Jinglin Liu、Jinzheng He 和 Zhou Zhao。Geneface：通用且高保真的音频驱动 3D 说话人脸合成。载于第十一届国际学习表征会议，2022年。2[78] Jianhui Yu、Hao Zhu、Liming Jiang、Chen Change Loy、Weidong Cai 和 Wayne Wu。Celebv-text：大规模人脸文本-视频数据集。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 14805–14814 页，2023年。4[79]Polina Zablotskaia、Aliaksandr Siarohin、Bo Zhao 和 Leonid Sigal。Dwnet：用于姿态引导的人体视频生成的密集变形网络。arXiv 预印本 arXiv:1910.09139，2019年。2, 4[80] Lvmin Zhang、Anyi Rao 和 Maneesh Agrawala。为文生图扩散模型添加条件控制。载于 ICCV，第 3813–3824 页。IEEE，2023年。6[81] Lvmin Zhang、Shengqu Cai、Muyang Li、Gordon Wetzstein 和 Maneesh Agrawala。下一帧预测视频扩散模型中的帧上下文打包与漂移抑制。载于第三十九届神经信息处理系统年会，2025年。5[82] Wenxuan Zhang、Xiaodong Cun、Xuan Wang、Yong Zhang、Xi Shen、Yu Guo、Ying Shan 和 Fei Wang。Sadtalker：学习逼真的 3D 运动系数，以实现风格化的音频驱动单图说话人脸动画。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 8652–8661 页，2023年。2[83] Yuang Zhang、Jiaxi Gu、Li-Wen Wang、Han Wang、Junqi Cheng、Yuefeng Zhu 和 Fangyuan Zou。Mimicmotion：通过置信度感知姿态引导实现高质量人体运动视频生成。arXiv 预印本 arXiv:2406.19680，2024年。2[84] Zhimeng Zhang、Lincheng Li、Yu Ding 和 Changjie Fan。利用高分辨率音视频数据集进行光流引导的单次说话人脸生成。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 3661–3670 页，2021年。1, 2, 4[85] Jian Zhao 和 Hui Zhang。用于图像动画的薄板样条运动模型。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 3657–3666 页，2022年。2[86] Daquan Zhou、Weimin Wang、Hanshu Yan、Weiwei Lv、Yizhe Zhu 和 Jiashi Feng。Magicvideo：利用潜在扩散模型高效生成视频。arXiv 预印本 arXiv:2211.11018，2022年。1[87] Hao Zhu、Wayne Wu、Wentao Zhu、Liming Jiang、Siwei Tang、Li Zhang、Ziwei Liu 和 Chen Change Loy。Celebv-hq：大规模视频人脸属性数据集。载于欧洲计算机视觉会议论文集，第 650–667 页。Springer，2022年。1, 2, 4[88] Wentao Zhu、Yufang Huang、Xiufeng Xie、Wenxian Liu、Jincan Deng、Debing Zhang、Zhangyang Wang 和 Ji Liu。Autoshot：短视频数据集与最先进的镜头边界检测。在 *IEEE/CVF*计算机视觉与模式识别会议论文集，第2238–2247页，2023年。3