

LetsTalk: Latent Diffusion Transformer for Talking Video Synthesis



Figure 1. We propose **LetsTalk**, a diffusion-based transformer for audio-driven portrait image animation. **Left:** Given a single reference image and audio, LetsTalk can produce a realistic and vivid video aligned with the input audio. Note that each column corresponds to the same audio. The results show that LetsTalk can drive consistent and reasonable mouth motions for the input audio. **Right:** Generation quality *vs.* inference time on the HDTF dataset, the circle area reflects the number of parameters of the method. Our LetsTalk achieves the best quality while also being highly efficient in inference, compared to current mainstream diffusion-based methods such as Haloo and AniPortrait. In addition, our base version (*LetsTalk-B*) achieves performance similar to Haloo with only 8 × fewer parameters.

Introduction

在音频驱动的头像动画生成 (Talking Head Generation) 中，任务是通过给定的音频输入，从单张图像生成一个对应音频内容的动态说话头像视频。这项任务的核心挑战在于如何解决以下问题：

1. 时间一致性：确保嘴唇与音频对齐，同时动作自然；

2. **人物一致性**: 生成视频需要与参考图像中的人物面部特征保持一致;
3. **多样性**: 生成的视频应包含自然的无意识动作（如眨眼、头部移动等）。

为了解决这些问题，研究主要分为两种方法：

1. 基于3D模型的方法：

- 使用3D变形模型（3DMM）来表示参考人物，通过音频驱动模型生成参数（如姿态、表情等），再通过神经渲染器生成视频。
- 缺点：生成的视频通常缺乏动态表现（如眨眼、头发晃动等），导致动画缺乏多样性。

2. 基于2D对应关系的方法：

- 直接学习音频和图像运动（如嘴唇和面部肌肉）的对应关系，通过潜在空间中的特征对齐来生成视频。
- 缺点：难以区分面部的静态部分（如背景、头发）与动态部分（如嘴唇），导致时间一致性和多样性不足。

近年来，扩散模型在生成高质量2D图像和视频方面表现出色。尤其是AnimateAnyone方法展示了基于扩散模型可以有效捕捉条件信号（如姿态、音频）与人类动画之间的对应关系。然而，现有方法大多基于U-Net结构，其在视频生成任务中存在以下问题：

- **多模态融合能力有限**: 难以充分利用音频和图像的特性;
- **计算复杂度高**: 在扩展至高分辨率和长视频时, 计算开销巨大。

为此, 我们提出了一种新的方法 **LetsTalk**, 它通过引入 **Transformer** 框架来解决这些问题。具体特点如下:

1. **低分辨率潜在空间生成**: 利用变分自编码器 (VAE) 将任务映射到低分辨率的潜在空间。
2. **时空注意力模块**: 在Transformer中集成空间和时间注意力机制, 增强时空一致性。
3. **多模态融合策略**:
 - 图像: 采用深度融合 (Symbiotic Fusion), 以确保生成视频与参考图像的一致性。
 - 音频: 采用浅层融合 (Direct Fusion), 以保证嘴唇动作与音频对齐, 同时保留多样性。

Related Work

音频驱动的人像动画生成任务在多个研究方向上得到了广泛探索, 包括扩散模型在视频生成中的应用、音频驱动的人像动画以及条件控制扩散模型等。

1. 扩散模型在视频生成中的应用

扩散模型近年来在图像和视频生成领域中表现卓越，其基本原理是逐步向输入添加噪声，并通过学习逆扩散过程生成目标数据。以下是相关研究的进展：

1. 图像生成：

- **Stable Diffusion** [34] 利用U-Net架构结合大规模文本-图像数据集，实现了从文本生成高质量图像。

2. 视频生成：

- **Video Diffusion Models (VDM)** [20] 在时间和空间维度上进行因子分解训练，有效提升视频生成性能。
- **ImagenVideo** [19] 基于级联扩散模型生成高分辨率视频。

3. 扩展工作：

- **VideoCrafter** [6] 和 **AnimateDiff** [17] 通过对扩散过程进行条件控制，实现了更强的生成效果。
- **DiT (Diffusion Transformer)** [31]: 引入了Transformer架构，通过图像块序列的处理机制，有效提升了扩散模型的多模态融合能力。

2. 音频驱动的人像动画

该方向主要关注如何通过音频驱动生成高保真和表情丰富的动态人像：

1. 早期方法：

- 主要关注嘴唇同步，例如 **LipSyncExpert** [32] 针对特定身份实现了高精度嘴唇同步。

2. 多模态表示学习：

- **ATVGnet** [7] 和 **MakeItTalk** [55] 通过引入面部标志点和解耦表示，提高了视频帧生成质量。

3. 融合3D信息：

- **SadTalker** [52] 和 **DiffTalk** [38] 引入了3D信息来增强头部动作和表情生成。

4. 扩散模型：

- **Diffused Heads** [39] 和 **DreamTalk** [30] 使用扩散模型生成自然且高保真的偶像动画。

3. 条件控制扩散模型

扩散模型生成能力的提高促使研究者探索在生成过程中引入更多控制信息：

1. 文本引导：

- **ControlNet** [51] 在预训练的扩散模型中加入了空间定位任务的条件控制。

2. 视频生成：

- **AnimateAnyone** [21] 通过使用 **ReferenceNet** 确保参考图像细节的一致性。

3. 运动控制：

- **MCDiff** [8] 和 **LaMD** [22] 专注于运动控制下的视频生成。

总结：现有方法多采用U-Net架构处理多模态条件。然而，扩散Transformer（如DiT）在这一领域仍未得到充分研究。为填补这一空白，本文探讨了基于Transformer架构的扩散条件控制策略，提升多模态条件下的生成质量。

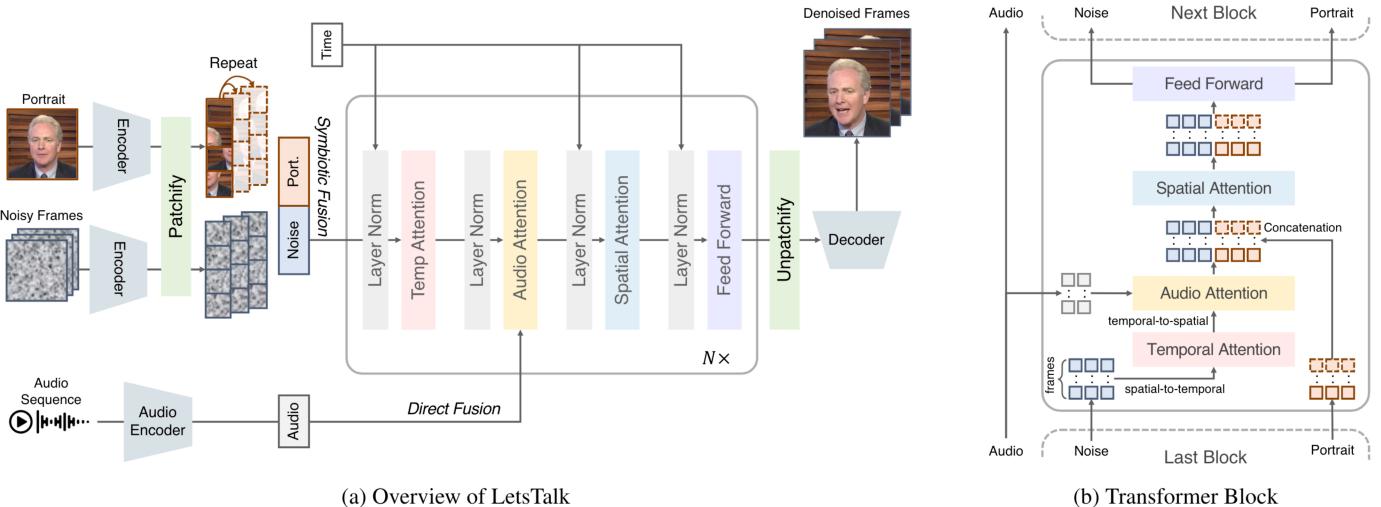


Figure 2. The overview of our method (a) and the illustration of our designed transformer block (b). For better illustration, we omit the timestep encoder and Layer Norm in (b). LetsTalk integrates transformer blocks equipped with both temporal and spatial attention modules, designed to capture intra-frame spatial details and establish temporal correspondence across time steps. After obtaining portrait and audio embeddings, *Symbiotic Fusion* is used for fusing the portrait embedding and *Direct Fusion* is for fusing the audio embedding. Notably, we repeat the portrait embedding along the frame axis to make it have the same shape as the noise embedding.

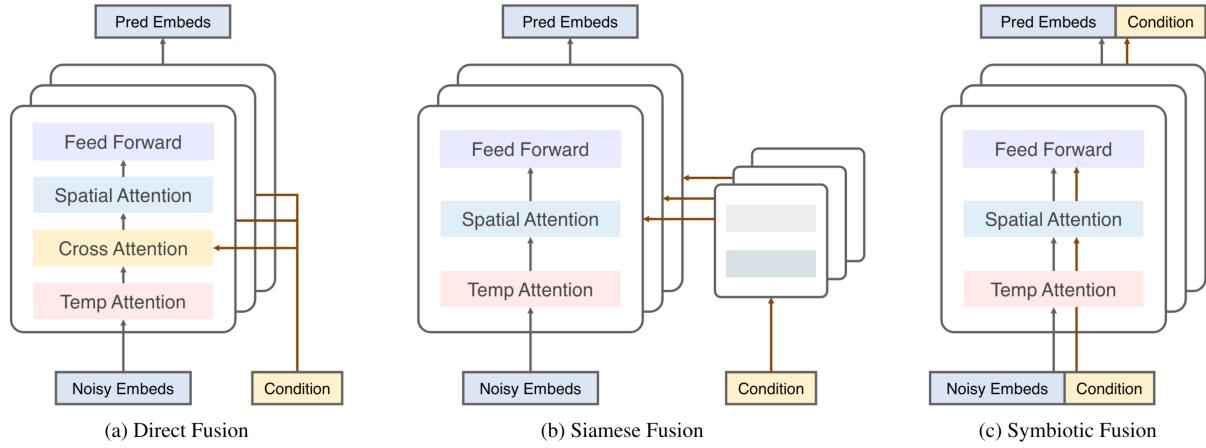


Figure 3. Illustration of three multimodal fusion schemes, our transformer backbone is formed by the left-side blocks. (a) **Direct Fusion**. Directly feeding condition into each block's cross-attention module; (b) **Siamese Fusion**. Maintaining a similar transformer and feeding the condition into it, extracting the corresponding features to guide the features in the backbone; (c) **Symbiotic Fusion**. Concatenating modality with the input at the beginning, then feeding it into the backbone, achieving fusion via the inherent self-attention mechanisms.

Method

本文提出了一种名为 **LetsTalk** 的扩散Transformer框架，用于音频驱动的人像视频生成。

3.1 Preliminaries

扩散模型的基本原理

扩散模型通过逐步添加噪声来破坏数据 x_0 ，生成 x_t ，其过程定义为：

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbf{I}), \quad (1)$$

其中：

- $\bar{\alpha}_t$ 是超参数，表示每一步的噪声权重。
- x_t 是第 t 步生成的数据。
- \mathcal{N} 表示高斯分布。

通过重参数化，得：

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, 1). \quad (2)$$

扩散模型的目标是学习逆过程 $p_\theta(x_{t-1}|x_t)$ ，用以生成无噪声的数据。模型训练通过优化以下目标函数实现：

$$L_{\text{simple}}(\theta) = \mathbb{E}_{x,\epsilon,t} [\|\epsilon - \epsilon_\theta(x_t, t)\|^2]. \quad (3)$$

扩散Transformer简介

与 U-Net 不同，DiT 在潜在空间中操作，大幅降低高分辨率生成的计算复杂度。生成视频的目标被简化为在潜在空间中生成特征 $F \in \mathbb{R}^{F \times H \times W \times C}$ ，其中 F 表示帧数。

3.2 Efficient Spatial-Temporal-Aware Transformer Backbone

问题：时空融合的挑战

直接在 Transformer 中对 $F \times P$ 的空间-时间特征进行自注意力计算会导致极高的计算开销，其复杂度为 $\mathcal{O}((F \times P)^2)$ 。

解决方案：交替时空融合

本文提出了“交替融合”的方法，分开计算空间注意力和时间注意力：

1. **空间注意力**：仅在同一帧内计算特征间的关系，注意力矩阵大小为 (F, P, P) 。
2. **时间注意力**：在同一位置跨帧计算关系，注意力矩阵大小为 (P, F, F) 。

最终的计算复杂度显著降低为：

$$\mathcal{O}(F \times P^2 + P \times F^2). \quad (4)$$

Transformer结构设计

交替堆叠以下模块实现时空融合：

1. **空间注意力模块**：聚焦同一帧中的空间信息。
2. **时间注意力模块**：处理不同帧之间的时序关系。

公式表示如下：

- 空间注意力计算：

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \quad (5)$$

其中 $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ 是空间特征的线性投影。

3.3 Multimodal Fusion Schemes for Portrait and Speech Audio

多模态融合的三种方案

1. Direct Fusion:

- 条件信息直接通过交叉注意力模块传入骨干网络的每一层。
- 优点：实现简单，计算效率高。
- 缺点：浅层融合可能导致深层特征对齐问题。

2. Siamese Fusion:

- 设计与骨干网络相似的 Siamese Transformer，用于提取条件信息，并逐层融合到主网络。
- 优点：显著提升对齐精度。
- 缺点：引入了额外的模型参数。

3. Symbiotic Fusion:

- 在输入阶段将条件特征与主特征拼接，通过自注意力实现隐式融合。
- 优点：高效地利用条件信息，减少参数开销。

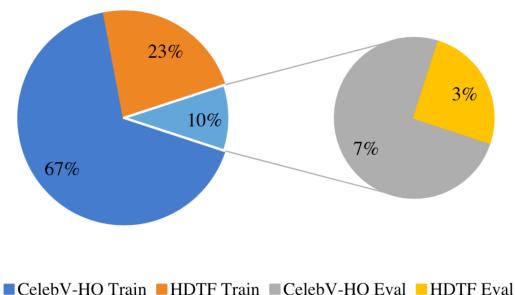
不同模态的融合策略

1. 图像融合：

- 采用 Symbiotic Fusion，参考图像特征与潜在空间特征拼接，增强生成视频与参考图像的一致性。

2. 音频融合：

- 采用 Direct Fusion，将音频特征通过交叉注意力逐层引入，确保嘴唇动作与音频对齐，同时保留多样性。



Datasets	Raw		Filtered	
	#IDs	#Hours	#IDs	#Hours
HDTF	411	15.75	356	13.95
CelebV-HQ	35,666	68	22,423	41.27
Train	-	-	20,501	49.69
Eval	-	-	2,278	5.522

Figure 4. The dataset splits for HDTF and CelebV-HQ, including raw/filtered data and training/evaluation distributions.

Method	FID ↓	FVD ↓	SSIM ↑	E-FID ↓
DreamTalk	105.434	529.510	0.406	3.466
SadTalker	102.371	362.634	0.414	2.168
AniPortrait	78.284	331.117	0.623	2.903
Hallo	45.754	285.988	0.698	2.235
Ours (Base)	46.240	289.385	0.666	2.587
Ours (Large)	37.134	219.119	0.632	1.962

Table 1. The quantitative comparisons with the existed portrait image animation approaches on the HDTF dataset.

Method	FID ↓	FVD ↓	SSIM ↑	E-FID ↓
DreamTalk	64.297	293.905	0.448	5.915
SadTalker	42.017	77.802	0.469	3.372
AniPortrait	33.408	107.236	0.578	3.722
Hallo	8.504	49.431	0.580	3.280
Ours (Base)	9.585	36.159	0.598	3.349
Ours (Large)	7.610	33.175	0.584	3.038

Table 2. The quantitative comparisons with the existed portrait image animation approaches on the CelebV-HQ dataset.

Reference Portrait	FID ↓	FVD ↓	SSIM ↑	E-FID ↓
Direct Fusion	52.999	389.01	0.497	3.149
Siamese Fusion	47.779	286.692	0.651	2.597
Symbiotic Fusion	46.240	289.385	0.666	2.587

Table 3. Experimental results of three multimodal fusion schemes, guided by reference portraits, on the HDTF validation dataset, employing the Letstalk-B model.

Driven Audio	FID ↓	FVD ↓	SSIM ↑	E-FID ↓
Symbiotic Fusion	51.880	403.628	0.471	3.398
Siamese Fusion	48.041	318.637	0.585	2.706
Direct Fusion	46.240	289.385	0.666	2.587

Table 4. Experimental results of three multimodal fusion schemes, driven by audio inputs, on the HDTF validation dataset, employing the Letstalk-B model.

Experiment

为了验证 **LetsTalk** 的性能，我们设计了一系列实验，涵盖数据集设置、定量结果、定性结果以及消融实验。

4.1 Experimental Setups

数据集

我们主要在两个公开数据集上进行实验：

1. **HDTF** [53]:

- 包含多种年龄、性别和背景的视频，主要为近景或半身镜头。
- 数据过滤后保留了 356 个视频剪辑，约 13.95 小时。

2. **CelebV-HQ** [56]:

- 包含超过 35,000 段视频，数据清理后保留了 22,423 个视频，约 41.27 小时。

训练数据提取

- 每个视频剪辑提取 16 帧，每帧分辨率为 256×256 。
- 训练时，采用间隔采样方法，从每个视频中提取帧序列。

评估指标

实验使用以下指标评估生成质量：

1. **Frechet Inception Distance (FID)**:

- 测量生成数据和真实数据的分布差异，值越低表示质量越高。

2. Frechet Video Distance (FVD):

- 扩展到视频的版本，用于衡量生成视频与真实视频之间的相似性。

3. Structural Similarity Index Measure (SSIM):

- 衡量生成视频与真实视频的结构相似性，值越高越好。

4. Expression-FID (E-FID):

- 使用面部表情参数，通过 FID 评估生成视频的表情真实性。
-

4.2 Quantitative Results

在 HDTF 数据集上的定量结果：

Method	FID ↓	FVD ↓	SSIM ↑	E-FID ↓
DreamTalk	105.434	529.510	0.406	3.466
SadTalker	102.371	362.634	0.414	2.168
AniPortrait	78.284	331.117	0.623	2.903
Hallo	45.754	285.988	0.698	2.235
Ours (Base)	46.240	289.385	0.666	2.587
Ours (Large)	37.134	219.119	0.632	1.962

在 CelebV-HQ 数据集上的定量结果：

Method	FID ↓	FVD ↓	SSIM ↑	E-FID ↓
DreamTalk	64.297	293.905	0.448	5.915
SadTalker	42.017	77.802	0.469	3.372
AniPortrait	33.408	107.236	0.578	3.722
Hallo	8.504	49.431	0.580	3.280
Ours (Base)	9.585	36.159	0.598	3.349
Ours (Large)	7.610	33.175	0.584	3.038

4.3 Qualitative Results

实验可视化

在 HDTF 和 CelebV-HQ 数据集上，与其他方法的对比结果如下：

1. 生成的嘴唇动作与音频对齐：

- LetsTalk 能更好地生成自然的嘴唇运动，并与输入音频保持一致。

2. 人物一致性：

- 得益于 Symbiotic Fusion 融合策略，生成视频中的人物特征保持高度一致。

3. 多样性：

- Direct Fusion 的浅层融合使生成的视频包含更多自然的无意识动作（如眨眼和头部移动）。
-

4.4 Ablation Study

1. 参考图像融合的影响

Fusion Scheme	FID ↓	FVD ↓	SSIM ↑	E-FID ↓
Direct Fusion	52.999	389.01	0.497	3.149
Siamese Fusion	47.779	286.692	0.651	2.597
Symbiotic Fusion	46.240	289.385	0.666	2.587

2. 音频融合的影响

Fusion Scheme	FID ↓	FVD ↓	SSIM ↑	E-FID ↓
Symbiotic Fusion	51.880	403.628	0.471	3.398
Siamese Fusion	48.041	318.637	0.585	2.706
Direct Fusion	46.240	289.385	0.666	2.587

3. 多样性的实验

通过让 LetsTalk 作为无条件视频生成器（去除音频输入），我们比较了三种融合方案对生成多样性的影响：

- **Symbiotic Fusion**: 生成的视频几乎无运动。
- **Siamese Fusion**: 生成了少量运动。
- **Direct Fusion**: 生成了自然且丰富的动作。

Conclusion

我们提出了 **LetsTalk**，一种基于扩散Transformer的多模态音频驱动头像动画生成方法。本文的方法通过引入模块化的时空注意力机制和合理的多模态融合策略，在保持高质量生成的同时，显著提升了生成视频的时间一致性和多样性。

关键贡献

1. 创新性架构：

- 首次在音频驱动的头像动画生成任务中引入扩散 Transformer，结合模块化的时空注意力机制，增强了视频生成的空间一致性和时间一致性。

2. 多模态融合方案：

- 深入分析并总结了三种多模态融合策略（Direct Fusion、Siamese Fusion 和 Symbiotic Fusion），并根据不同模态（图像和音频）的特点分别采用最佳融合方案：
 - 图像：Symbiotic Fusion，保证生成视频与参考图像高度一致。
 - 音频：Direct Fusion，确保嘴唇动作与音频对齐，并保留多样性。

3. 实验验证：

- 在 HDTF 和 CelebV-HQ 数据集上的大量实验表明，**LetsTalk** 在生成保真度、时间一致性和音频动画对齐方面均超越了现有方法。

局限性与未来方向

尽管 **LetsTalk** 在多个指标上取得了领先，但仍存在一些挑战和改进空间：

1. 资源和数据集限制：

- 使用的公开数据集较小，未来研究可探索更大规模的数据集和多样化的场景。

2. 长时动画生成：

- 当前方法在生成长时间视频时，帧之间的过渡存在累积误差。未来可通过改进时间建模或帧生成机制来解决此问题。

3. 高分辨率生成：

- 尽管引入了 VAE 降低计算开销，但高分辨率视频生成仍是一个挑战，未来可探索更高效的生成策略。
-

总结

LetsTalk 提供了一种新的视角，将扩散Transformer应用于音频驱动的头像动画生成任务，推进了该领域在生成质量和多样性方面的技术边界。