# Latte: Latent Diffusion Transformer for Video Generation
# 用于视频生成的潜在扩散 Transformer


## Abstract

We propose a novel Latent Diffusion Transformer, namely Latte, for video generation. Latte first extracts spatio-temporal tokens from input videos and then adopts a series of Transformer blocks to model video distribution in the latent space. In order to model a substantial number of tokens extracted from videos, four efficient variants are introduced from the perspective of decomposing the spatial and temporal dimensions of input videos. To improve the quality of generated videos, we determine the best practices of Latte through rigorous experimental analysis, including video clip patch embedding, model variants, timestep-class information injection, temporal positional embedding, and learning strategies. Our comprehensive evaluation demonstrates that Latte achieves state-of-the-art performance across four standard video generation datasets, i.e., FaceForensics, SkyTimelapse, UCF101, and Taichi-HD. In addition, we extend Latte to text-to-video generation (T2V) task, where Latte achieves comparable results compared to recent T2V models. We strongly believe that Latte provides valuable insights for future research on incorporating Transformers into diffusion models for video generation. Project page: https://maxin-cn.github.io/latte project.

**（本文贡献）：**我们提出了一种新颖的视频生成模型——潜在扩散变换器，即 Latte。**（模型细节）：**Latte 首先从输入视频中提取时空标记，然后采用一系列变换器模块（Transformer blocks）来在潜在空间中建模视频分布。**（目的）：为了处理从视频中提取的大量标记，（方法）：**我们从分解输入视频的空间和时间维度的角度，引入了四种高效的变体。**（目的）：为了提高生成视频的质量，（方法）：**我们通过严谨的实验分析确定了 Latte 的最佳实践，包括视频剪辑补丁嵌入、模型变体、时间步长类信息注入、时间位置嵌入和学习策略。**（结果 1）：我们的全面评估表明，Latte 在四个标准视频生成数据集上实现了最先进的性能，即 FaceForensics、SkyTimelapse、UCF101 和 Taichi-HD。（结果 2）：此外，我们将 Latte 扩展到文本到视频生成（T2V）任务，在这一任务中，Latte 与最近的 T2V 模型相比，取得了可比的结果。（意义）：**我们坚信，Latte 为未来研究将变换器融入扩散模型用于视频生成提供了宝贵的洞见。项目页面：https://maxin-cn.github.io/latte project。


关键词：视频生成，扩散模型，transformers


## 1 Introduction

**（研究背景）：扩散模型**（Ho 等人，2020 年；Song 等人，2021b、a）是内容创建各种任务中强大的深度生成模型，包括图像到图像生成（Meng 等人，2022 年；Zhao 等人，2022 年；Saharia 等人，2022a；Parmar 等人，2023 年）、文本到图像生成（Zhou 等人，2023 年；Rombach 等人，2022 年；Zhou 等人，2022 年；Ruiz 等人，2023 年；Zhang 等人，2023 年）和 3D 对象生成（Wang 等人，2023 年；Chen 等人，2023b；Zhou 等人，2021 年；Shue 等人，2023 年）等。**（挑战-科学问题）：**与在图像中的这些成功应用相比，生成高质量的视频仍然面临着重大挑战，这主要归因于视频的复杂和高维性质，在高分辨率帧中包含复杂的时空信息。

**（过去研究 1）：**同时，研究者们揭示了在扩散模型成功中革命性地**改进基础架构**的重要性（Nichol 和 Dhariwal, 2021; Peebles 和 Xie, 2023; Bao 等人, 2023）。**（过去研究 2）：**依赖于卷积神经网络（CNNs）的 U-Net（Ronneberger 等人, 2015）在图像和视频生成领域占据了显著地位（Ho 等人, 2022; Dhariwal 和 Nichol, 2021）。**（过去研究 3）：**相反，一方面，DiT（Peebles 和 Xie, 2023）和 U-ViT（Bao 等人, 2023）将 ViT（Dosovitskiy 等人, 2021）的架构适应于扩散模型用于图像生成，并取得了巨大的性能成就。**（过去研究 4）：**此外，DiT 证明了 U-Net 的归纳偏见对潜在扩散模型的性能并非至关重要。**（过去研究 5）：**另一方面，基于注意力的架构（Vaswani 等人, 2017）为捕获视频中的长范围上下文关系提供了一个直观的选择。**（科学问题）：**因此，一个非常自然的问题随之产生：基于 Transformer 的潜在扩散模型是否能增强真实视频的生成？

**（本文研究）：**在本文中，我们提出了一种新颖的潜在扩散 Transformer 用于视频生成，**（模型细节）：**即 Latte，它采用**视频 Transformer** 作为骨干。Latte 利用预训练的**变分自动编码器**将输入视频编码为潜在空间中的特征，其中从编码特征中提取出标记。然后，一系列 Transformer 块被应用于编码这些标记。**（出发点）：考虑到输入视频的空间和时间信息之间的固有差异以及从输入视频中提取出大量标记**，如**图 2** 所示，**（方法）：**我们从分解输入视频的空间和时间维度的角度设计了四种高效的基于 Transformer 的模型变体。

**（卷积模型过去研究）：**对于卷积模型而言，存在众多最佳实践，包括用于**问题分类**的文本表示（Pota 等人, 2020）以及用于**图像分类**的网络架构设

计（何等人, 2016）等。**（出发点）：然而，用于视频生成的基于 Transformer 的潜在扩散模型可能展现出不同的特点，这要求确定这种架构的最优设计选择**。**（方法）：**因此，我们进行了一项全面的**消融分析**，涵盖了**视频剪辑补丁嵌入**、**模型变体**、**时间步长类别信息注入**、**时间位置嵌入**和**学习策略**。**（结果）：**我们的分析**使 Latte 能够生成具有时间连贯内容的逼真视频**（见**图 1**）并**在四个标准视频生成基准测试中实现最先进的性能**，包括 FaceForensics（Rössler 等人, 2018）、SkyTimelapse（熊等人, 2018）、UCF101（Soomro 等人, 2012）和 Taichi-HD（Siarohin 等人, 2019）。**（结果）：**值得注意的是，**Latte 显著优于当前最先进的技术**，实现了最佳的 Fréchet 视频距离（FVD）（Unterthiner 等人, 2018）、Fréchet 创新距离（FID）（Parmar 等人, 2021）和创新得分（IS）。**（结果）：**此外，我们还将 Latte 扩展到**文本到视频生成任务**，**在这一任务上也取得了与当前 T2V 模型相媲美的结果**。

总之，我们的主要贡献如下：

我们提出了 Latte，一种新颖的潜在扩散 Transformer，它采用视频 Transformer 作为骨干。此外，我们还引入了**四种模型变体**，以有效捕获视频中的**时空分布**。

- **（目的）：为了提高生成视频的质量**，**（过程）：**我们全面探讨了**视频剪辑补丁嵌入**、**模型变体**、**时间步类信息注入**、**时间位置嵌入**和**学习策略**等方面，**（目的）：以确定基于 Transformer 的扩散模型在视频生成中的最佳实践**。
- **（结果）：**在四个标准视频生成基准上的实验结果显示，与最先进的方法相比，**Latte 能够生成具有时间一致内容的逼真视频**。此外，当应用于**文本到视频生成任务**时，**Latte 显示出可比较的结果**。

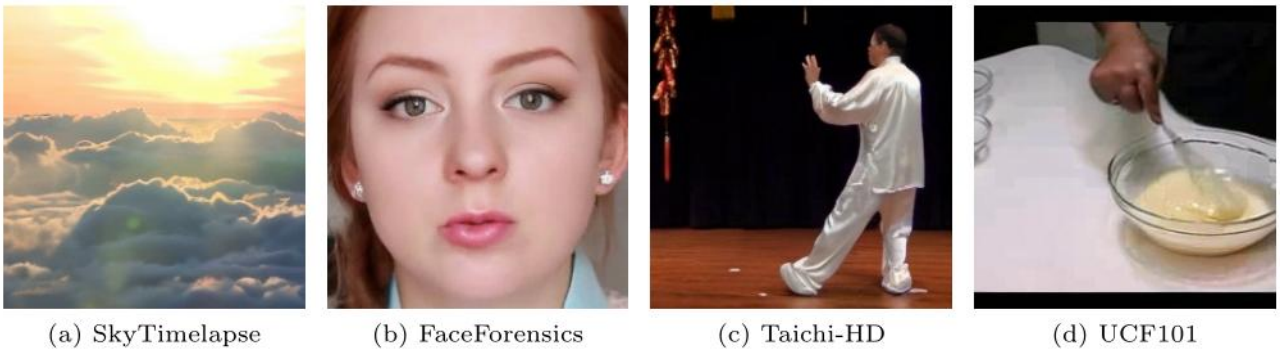



**图 1**：在四个数据集上的样本视频（256×256）。Latte 生成了具有时间连贯内容的逼真视频。请点击图片播放视频片段。

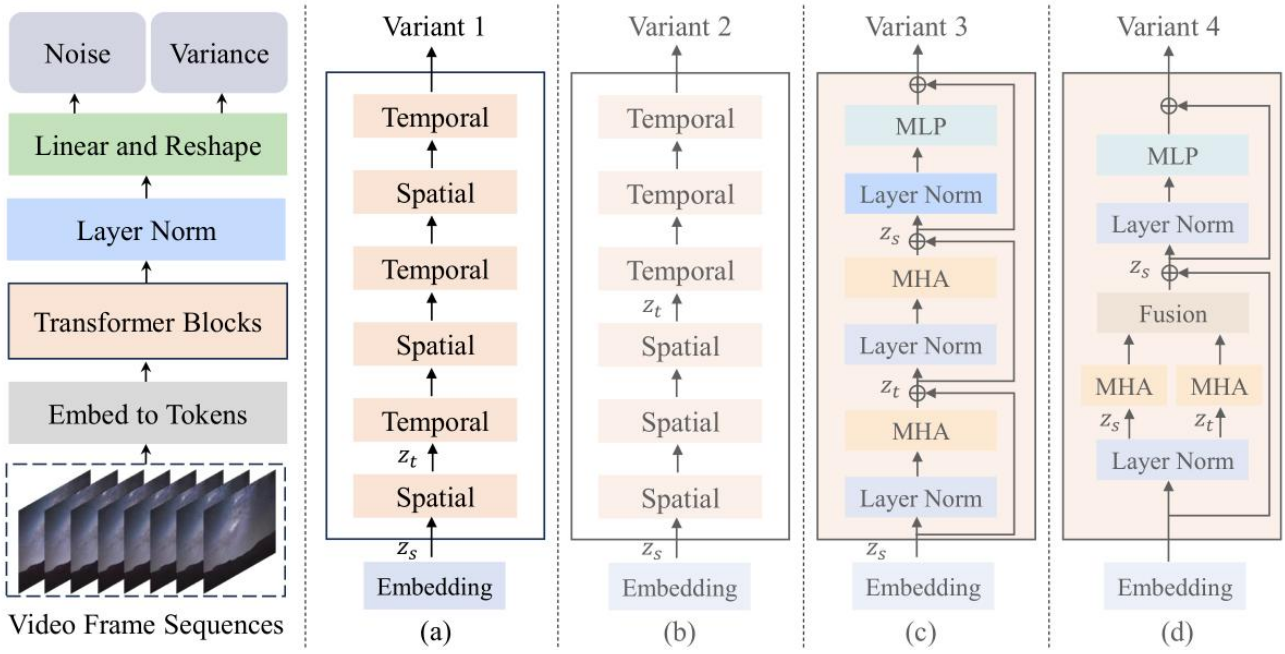



**图 2**：Latte 用于视频生成的流程图。提出了四种 Latte 模型变体，用以高效捕获视频中的时空信息。图中浅橙色表示的每个块代表一个 Transformer 块。标准 Transformer 块（如**图 4b** 所述）被应用于（a）和（b）。同时，（c）和（d）使用了我们各自的 Transformer 块变体。为了简化，图中未显

示 VAE 的编码和解码过程。

## 2 RelatedWork

**（研究背景）：**视频生成旨在同时产生具有高质量视觉外观和一致运动的逼真视频。**（研究划分）：**在这一领域的先前研究可以分为三大类。**（领域 1）：**首先，一些研究致力于扩展**基于 GAN** 的强大图像生成器的能力，以创建视频（Vondrick 等人，2016 年；Saito 等人，2017 年；Wang 等人，2020b、a；Kahembwe 和 Ramamoorthy，2020 年）。**（局限性）：**然而，这些方法通常面临着与模态崩溃相关的挑战，限制了它们的有效性。**（领域 2）：**其次，一些方法提出使用**自回归模型**学习数据分布（Ge 等人，2022 年；Rakhimov 等人，2021 年；Weissenborn 等人，2020 年；Yan 等人，2021 年）。**（优势）：**尽管这些方法通常提供良好的视频质量并且展现出**更稳定的收敛性**，**（局限性）：**但它们的缺点是需要**大量的计算资源**。**（领域 3）：**最后，最近视频生成方面的进展集中在构建**基于扩散模型**的系统上（Ho 等人，2020 年；Harvey 等人，2022 年；Ho 等人，2022 年；Singer 等人，2022 年；Mei 和 Patel，2023 年；Blattmann 等人，2023b 年；Wang 等人，2023b；Chen 等人，2023c；Wang 等人，2023c），取得了令人期待的结果。**（局限性）：**然而，基于 Transformer 的扩散模型尚未得到深入探索。**（最近研究）：**最近的同时进行的工作 VDT（Lu 等人，2023 年）探索了类似的想法。**（本文研究的开创性）：**与 VDT 的不同之处在于，我们**在视频生成方面对不同的 Transformer 骨干进行了系统分析**，并**讨论了第 3.2 节和第 3.3 节中的相对最佳实践**。VDT 与我们的变体 3 类似。我们在**图 6d** 中展示了这些模型变体之间的性能差异，**（测试结果）：结果显示变体 1 优于变体 3**。

**（Transformer 地位）：**Transformer 已成为主流的模型架构，并在许多领域取得了显著的成功，例如图像修复（Ma 等人，2022 年，2021 年，2023 年）、图像超分辨率（Luo 等人，2022 年；Huang 等人，2017 年）、图像裁剪（Jia 等人，2022 年）、伪造检测（Jia 等人，2021 年）、人脸识别（Luo 等人，2021a，b 年）和自然语言处理（Devlin 等人，2019 年）等领域。**（Transformer 发展历史）：**Transformer 最初出现在语言领域（Vaswani 等人，2017 年；Kaplan 等人，2020 年），很快就因其出色的能力而建立了声誉。随着时间的推移，这些模型已经灵活地适应了预测图像的任务，并在图像空间和离散码本中自回归地执行这一功能（Chen 等人，2020 年；Parmar 等人，2018 年）。在最新的发展中，Transformer 已被整合到扩散模型中，将其范围扩展到了非空间数据和图像的生成。这包括文本编码和解码任务（Rombach 等人，2022 年；Saharia 等人，2022b 年）、生成 CLIP 嵌入（Ramesh 等人，2022 年），以及逼真的图像生成（Bao 等人，2023 年；Peebles 和 Xie，2023 年）。

## 3 方法论

我们从第 3.1 节简要介绍**潜在扩散模型**开始。随后，在第 3.2 节中介绍了 Latte 的模型变体。最后，在第 3.3 节中讨论了 Latte 的实证分析。

### 3.1 潜在扩散模型的基础

**（概念）：**潜在扩散模型（Latent Diffusion Models, LDMs）（Rombach 等人，2022 年）。LDMs 是一种高效的扩散模型（Ho 等人，2020 年；Song 等人，2021b 年），通过在**潜在空间**而不是像素空间中进行扩散过程。LDMs 首先利用预训练的**变分自动编码器**的编码器 E 将输入数据样本 x ∈ $p_{data}$(x) 压缩成**较低维度的潜在编码** z = E(x)。随后，它通过两个关键过程学习数据分布：**扩散**和**去噪**。

扩散过程逐渐将**高斯噪声**引入潜在编码 z，生成扰动样本 $z_t = \sqrt{\alpha_t}z + \sqrt{1-\alpha_t}\epsilon$，其中 $\epsilon \sim N(0,1)$，遵循跨越 T 个阶段的马尔可夫链。在这个上下文中，$\alpha_t$ 作为噪声调度器，其中 t 表示扩散的时间步。

去噪过程被训练**（目的）：用于理解逆扩散过程**，**（目的）：以预测一个更少噪声的 $z_{t-1}$ ：** $p_\theta(z_{t-1}|z_t) = N(\mu_\theta(z_t), \Sigma_\theta(z_t))$，其中变分下界的对数似然缩减为 $L_\theta = -logp_\theta(z_{t-1}|z_t) + \sum_t D_{KL}(q(z_{t-1}|z_t, z_0)||p_\theta(z_{t-1}|z_t))$。这里，$\mu_\theta$ 使用去噪模型 $\epsilon_\theta$ 实现，并通过简单的目标进行训练。

$$\mathcal{L}_{simple} = \mathbb{E}_{\mathbf{z}\sim p(z),\ \epsilon\sim\mathcal{N}(0,1),\ t}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t)\|_2^2\right]. \quad (1)$$

根据（Nichol 和 Dhariwal，2021 年），**（目的）：为了使用学习的逆过程协方差 $\Sigma_\theta$ 训练扩散模型**，需要优化完整的 $D_{KL}$ 项，因此使用完整的 $\mathcal{L}$ 进行训练，表示为 $\mathcal{L}_{vlb}$。此外，$\Sigma_\theta$ 是使用 $\epsilon_\theta$ 实现的。

**（扩展方法）：**我们**将 LDMs 扩展到视频生成**，具体为：1）使用编码器 E 将每个视频帧压缩到潜在空间中；2）扩散过程在视频的潜在空间中进行，

（目的）：以建模潜在的空间和时间信息。在这项工作中，$\epsilon_\theta$ 是用 Transformer 实现的。我们通过同时使用 $\mathcal{L}_{simple}$ 和 $\mathcal{L}_{vlb}$ 训练所有模型。

### 3.2 Latte 的模型变体

如图 2 所示，提出了 Latte 的四个模型变体，（目的）：以有效捕获视频中的时空信息。

变体 1：如图 2（a）所示，这种变体的 Transformer 骨干由两种不同类型的 Transformer 块组成：空间 Transformer 块和时间 Transformer 块。前者专注于在具有相同时间索引的标记之间仅（目的）：捕获空间信息，而后者以“交错融合”的方式跨时间维度（目的）：捕获时间信息。

假设我们有一个视频剪辑在潜在空间中 $V_L \in R^{F\times H\times W\times C}$ 。我们首先将 $V_L$ 转换为一个标记序列，表示为 $\hat{z} \in R^{n_f\times n_h\times n_w\times d}$ 。这里的 F、H、W 和 C 分别表示潜在空间中视频帧的数量、视频帧的高度、宽度和通道数。视频剪辑在潜在空间中的标记总数是 $n_f \times n_h \times n_w$，而 d 表示每个标记的维度。我们将时空位置嵌入 $p$ 加入到 $\hat{z}$ 中。最后，我们将得到的 $z = \hat{z} + p$ 作为 Transformer 骨干的输入。

我们将 z 重新整形成 $z_s \in R^{n_f\times t\times d}$ （目的）：作为空间 Transformer 块的输入，（目的）：用于捕获空间信息。这里，$t = n_h \times n_w$ 表示每个时间索引的标记数量。随后，包含空间信息的 $z_s$ 被重新整形成 $z_t \in R^{t\times n_f\times d}$，作为（目的）：用于捕获时间信息的时间 Transformer 块的输入。

变体 2：与变体 1 中的时间“交错融合”设计相反，这种变体利用“后期融合”方法（目的）：来结合时空信息（Neimark 等人，2021 年；Simonyan 和 Zisserman，2014 年）。如图 2（b）所示，这种变体包含与变体 1 中相同数量的 Transformer 块。与变体 1 类似，空间 Transformer 块和时间 Transformer 块的输入形状分别为 $z_s \in R^{n_f\times t\times d}$ 和 $z_t \in R^{t\times n_f\times d}$。

变体 4：在这种变体中，我们将多头注意力（MHA）分解为两个组件，每个组件利用一半的注意力头，如图 2（d）所示。我们使用不同的组件分别处理空间和时间维度中的标记。这些不同组件的输入形状分别为 $z_s \in R^{n_f\times t\times d}$ 和 $z_t \in R^{t\times n_f\times d}$。一旦计算了两个不同的注意力操作，我们将 $z_t \in R^{t\times n_f\times d}$ 重新整形为 $z_t' \in R^{n_f\times t\times d}$。然后将 $z_t'$ 加到 $z_s$ 中，作为 Transformer 块中下一个模块的输入。

在 Transformer 的主干结构之后，一个关键的步骤涉及解码视频令牌序列，（目的）：以推导出预测的噪声和预测的协方差。这两个输出的形状与输入 $V_L$ 的形状相同，其中 $V_L \in R^{F\times H\times W\times C}$。根据以往的研究（Peebles 和 Xie，2023 年；Bao 等，2023 年），我们通过使用标准的线性解码器以及重塑操作来完成这一步骤。

### 3.3 Latte 的实证分析

我们对 Latte 中关键组件进行了全面的实证分析，（目的）：旨在发现将 Transformer 作为潜在扩散模型中的主干结构集成到视频生成中的最佳实践。

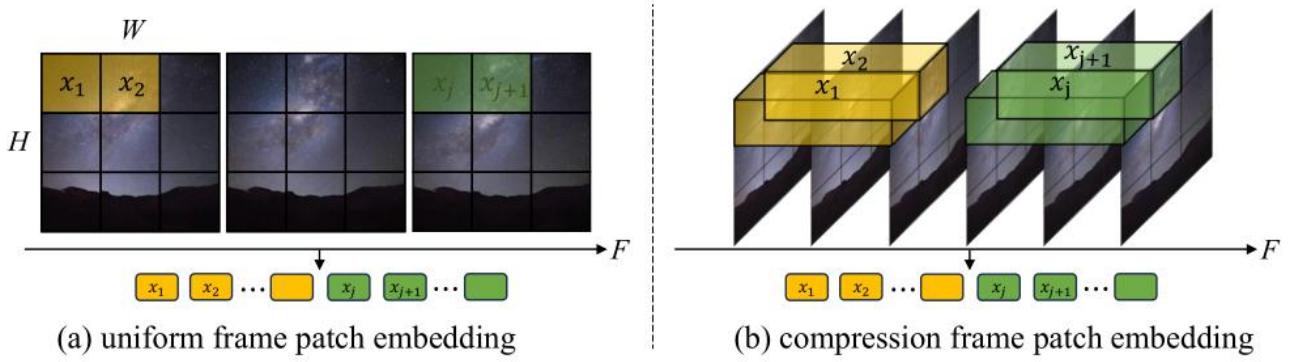



图 3：视频剪辑补丁嵌入。 （a）我们采样 F 帧，并使用 ViT 中描述的方法将每个单独的视频帧嵌入到令牌中。 （b）我们考虑捕获时间信息，然后将 ViT 补丁嵌入方法从 2D 扩展到 3D，并随后沿时间维度提取管道。（目的）：为了便于理解，我们在这里使用原始视频剪辑来演示补丁嵌入方法。视频潜在空间中的补丁嵌入遵循相同的处理方法。

#### 3.3.1 潜在视频片段补丁嵌入

（目的）：为了嵌入一个视频片段，我们探索了以下两种方法，以分析在令牌中集成时间信息的必要性，即 1）均匀帧补丁嵌入和 2）压缩帧补丁嵌入。

**均匀帧补丁嵌入**。如**图 3（a）**所示，我们将 ViT（Dosovitskiy 等，2021 年）中概述的补丁嵌入技术应用到每个视频帧中。具体而言，当从每个视频帧中提取非重叠的图像补丁时，$n_f$，$n_h$ 和 $n_w$ 分别等同于 F、$\frac{H}{h}$ 和 $\frac{W}{w}$。这里，h 和 w 分别表示图像补丁的高度和宽度。

**压缩帧补丁嵌入**。第二种方法是通过将 ViT 补丁嵌入扩展到时间维度（目的）：来建模潜在视频片段中的时间信息，如**图 3（b）**所示。我们沿着时间维度提取管道，并以步幅 s 映射它们到令牌。在这里，与**非重叠均匀帧补丁嵌入**相比，$n_f$ 相当于 $\frac{F}{s}$。与前者相比，这种方法在补丁嵌入阶段内在地包含了时空信息。需要注意的是，在使用**压缩帧补丁嵌入**方法的情况下，一个额外的步骤是在标准线性解码器和重塑操作之后，通过 **3D 转置卷积**对输出的潜在视频进行**时间上采样**的集成。

### 3.3.2 时间步类信息注入

从简单和直接的集成到复杂和微妙的集成视角，我们探索了两种方法（目的）：来将时间步或类别信息 c 集成到我们的模型中。（方法 1）：第一种方法是将其**视为令牌**，并且我们将这种方法称为所有令牌。（方法 2）：第二种方法类似于**自适应层归一化**（AdaLN）（Perez 等，2018 年；Peebles 和 Xie，2023 年）。我们使用**线性回归**根据输入 c 计算 $\gamma_c$和 $\beta_c$，得到方程 $AdaLN(h,\ c) = \gamma_c\, LayerNorm(h) + \beta_c$，其中 $h$ 表示 Transformer 块内的隐藏嵌入。此外，我们还对 $\alpha_c$ 进行**回归**。该方法直接应用于 Transformer 块内的**任何残差连接（RCs）之前**，导致 RCs（h, c） $= \alpha_c$h + AdaLN（h, c）。我们将其称为**可伸缩自适应层归一化**（S-AdaLN）。S-AdaLN 的架构如**图 4a** 所示。

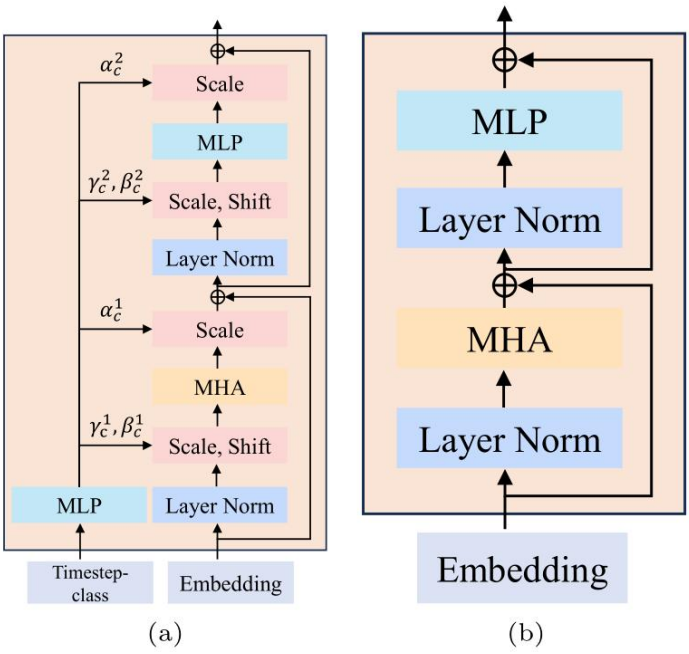



**图 4**：（a）第 3.3.2 节中描述的 S-AdaLN 架构。（b）图 2 中（a）和（b）使用的标准 Transformer 块架构。MLP 和 MHA 分别代表多层感知层和多头注意力。

### 3.3.3 时间位置嵌入

时间位置嵌入（目的）：使模型能够理解时间信号。我们探索了以下两种方法，将时间位置嵌入注入到模型中：

1）绝对位置编码方法：该方法利用**正弦和余弦函数的不同频率**（Vaswani 等，2017 年）（目的）：来使模型能够识别视频序列中每帧的精确位置；

2）相对位置编码方法：该方法采用了**旋转位置嵌入**（RoPE）（Su 等，2021 年），（目的）：使模型能够把握相邻帧之间的时间关系。

### 3.3.4 利用学习策略增强视频生成

我们的目标是（目的）：确保生成的视频在保持时间一致性的同时呈现最佳的视觉质量。（探索的科学问题）：我们探索是否可以通过引入两种额外的学习策略来增强生成视频的质量，即利用预训练模型进行学习和利用图像-视频联合训练进行学习。

**利用预训练模型进行学习**。**（观点）：**预训练的图像生成模型已经学习到了世界的外观。因此，有许多视频生成工作将它们的模型**建立在预训练的图像生成模型之上**，**（目的）：**以学习世界的运动方式（Wang 等，2023b 年；Blattmann 等，2023a 年）。**（局限性）：**然而，这些工作主要是基于潜在扩散模型内的 U-Net。**（探索方向）：**值得探索的是基于 Transformer 的潜在扩散模型的必要性。

我们从在 ImageNet 上预训练的 DiT 模型（Peebles 和 Xie，2023 年；Deng 等，2009 年）初始化 Latte。**（缺陷）：**直接从预训练的 DiT 模型初始化可能会遇到缺失或不兼容的参数问题。**（目的）：**为了解决这些问题，我们实施了以下策略。**（方法）：**在预训练的 DiT 模型中，对每个令牌应用了一个**位置嵌入** $p \in R^{n_h \times n_w \times d}$。**（数据量）：**然而，在我们的视频生成模型中，我们的令牌数量是预训练的 DiT 模型的 $n_f$ 倍。**（方法）：**因此，我们将时间位置嵌入从 $p \in R^{n_h \times n_w \times d}$ 复制 $n_f$ 次 **（目的）：**以适应视频生成模型的令牌数目，从 $p \in R^{n_h \times n_w \times d}$ 扩展到 $p \in R^{n_f \times n_h \times n_h \times d}$。**（数据量）：**此外，预训练的 DiT 模型包含一个**标签嵌入层**，类别数量为 1000。**（缺陷）：**然而，所使用的视频数据集要么缺少标签信息，要么包含的类别数量明显少于 ImageNet。**（出发地）：**由于我们既针对**无条件的视频生成**，又针对**有条件的视频生成**，DiT 中的原始标签嵌入层不适用于我们的任务，**（方法）：**因此我们选择直接丢弃 DiT 中的标签嵌入并应用零初始化。

**使用图像-视频联合训练进行学习**。基于 CNN 的视频扩散模型的先前工作提出了一种**联合图像-视频训练策略**，极大地提高了生成视频的质量（Ho 等人，2022 年）。**（科学问题）：**我们探讨这种训练策略是否也能提高基于 Transformer 的视频扩散模型的性能。**（目的）：**为了实现视频和图像生成的同时训练，**（方法）：**我们将同一数据集中随机选择的视频帧追加到所选视频的末尾，每个帧都是独立采样的。**（目的）：**为了确保我们的模型能够生成连续的视频，**（方法）：**与视频内容相关的标记被用于用于建模时间信息的时间模块中，而**帧标记**被排除在外。

## 4 实验

本节首先概述了实验设置，包括**数据集**、**评估指标**、**基准**、**Latte 配置**以及实现细节。随后，我们针对 Latte 的最佳实践选择和模型规模进行了**消融实验**。最后，我们将实验结果与最新技术进行了比较，并展示了文本到视频生成的结果。

### 4.1 实验设置

**数据集**。我们主要在**四个公开数据集**上进行了全面的实验：FaceForensics（Rössler 等人，2018）、SkyTimelapse（Xiong 等人，2018）、UCF101（Soomro 等人，2012）和 Taichi-HD（Siarohin 等人，2019）。在（Skorokhodov 等人，2022 年）的实验设置中，除了 UCF101 外，我们使用所有可用的数据集的训练集。对于 UCF101，我们**同时使用训练集和测试集**。我们从这些数据集中提取 **16 帧**的视频片段，使用**特定的采样间隔**，并将每一帧调整大小为 **256×256 分辨率**进行训练。

**评估指标**。在定量比较评估中，我们采用了**三个评估指标**：Fréchet Video Distance（FVD）（Unterthiner 等人，2018）、Fréchet Inception Distance（FID）（Parmar 等人，2021）和 Inception Score（IS）（Saito 等人，2017）。我们的主要关注点是 FVD，**（理论基础）：**因为其基于图像的对应物 FID 与人类主观判断更为接近。遵循 StyleGAN-V 引入的评估指南，我们通过分析包含 16 帧的 2,048 个视频片段来计算 FVD 分数。我们仅在评估 UCF101 上的生成质量时使用 IS，因为它利用了 UCF101 精调的 C3D 模型（Saito 等人，2017）。

**基准方法**。我们与最近的方法进行比较，**（目的）：**以定量评估结果，包括 MoCoGAN（Tulyakov 等人，2018）、VideoGPT（Yan 等人，2021）、MoCoGAN-HD（Tian 等人，2021）、DIGAN（Yu 等人，2022）、StyleGAN-V（Skorokhodov 等人，2022）、PVDM（Yu 等人，2023）、MoStGAN-V（Shen 等人，2023）和 LVDM（He 等人，2023）。此外，我们在 UCF101 数据集上对我们提出的方法与先前方法之间的 IS 进行**额外比较**。

**Latte 配置**。我们使用一系列 N 个 Transformer 块来构建我们的 Latte 模型，每个 Transformer 块的隐藏维度为 D，具有 N 个多头注意力。与 ViT 相似，我们确定了**四种不同参数数量**的 Latte 配置，如**表 4** 所示。

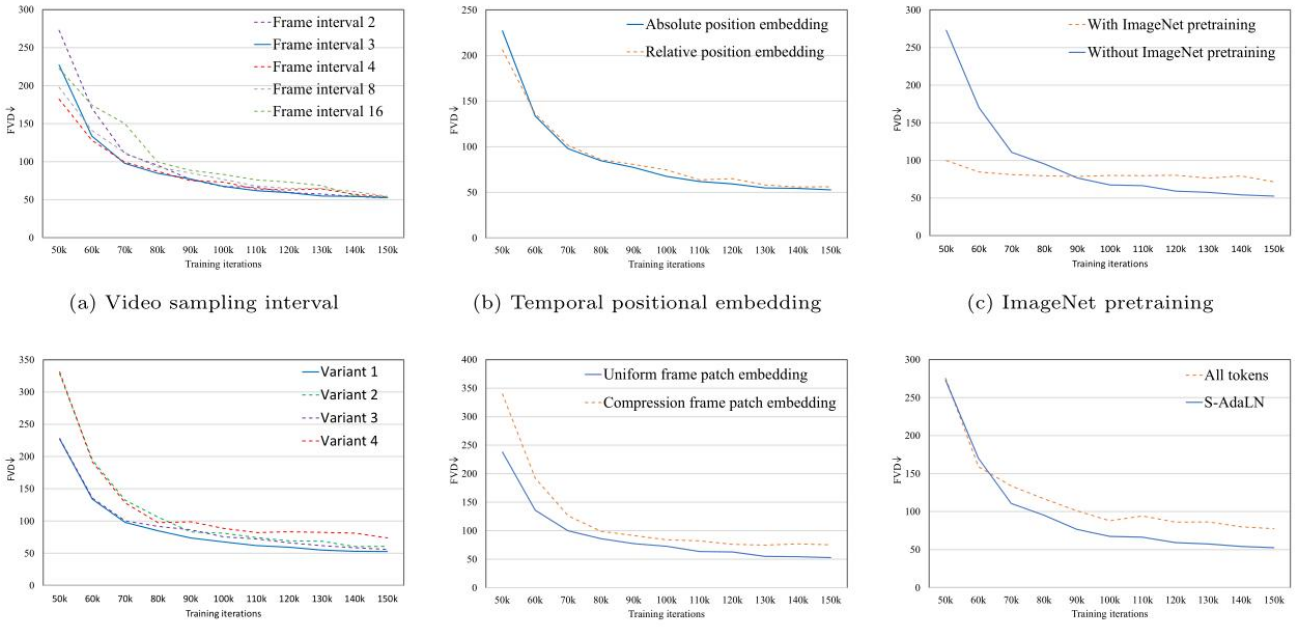
| Model | Layer numbers N | Hidden size D | Heads H | Param |
|---|---|---|---|---|
| Latte-S | 12 | 384 | 6 | 32.48M |
| Latte-B | 12 | 768 | 12 | 129.54M |
| Latte-L | 24 | 1024 | 16 | 456.81M |
| Latte-XL | 28 | 1152 | 16 | 673.68M |

**表 4**：Latte 模型的详细信息。我们遵循不同模型大小的 ViT 和 DiT 模型配置。

**实现细节**。我们使用 AdamW 优化器，并使用恒定学习率 1×10^(-4)来训练所有模型。**水平翻转**是唯一使用的数据增强技术。遵循生成建模作品中的常见实践（Peebles 和 Xie，2023；Bao 等，2023），在整个训练过程中保持 Latte 权重的指数移动平均（EMA），采用衰减率为 0.9999。所有报告的结果直接来自 EMA 模型。我们借用了稳定扩散 1.4 中的预训练变分自动编码器。

### 4.2 消融研究

在本节中，我们在 FaceForensics 数据集上进行实验，以检验第 3.3 节中描述的不同设计、第 3.2 节中描述的模型变体、视频采样间隔以及模型规模对模型性能的影响。




(a) Video sampling interval (b) Temporal positional embedding (c) ImageNet pretraining

(d) Model variants (e) Video clip patch embedding (f) Timestep-class information injection

**图 6**：设计选择的消融分析。我们设计了几项消融研究，以探索基于 Transformer 的视频扩散模型在 FaceForensics 上的 FVD 方面的最佳实践。请放大以获得更好的视图。

**视频片段补丁嵌入**。我们检验了第 3.3.1 节详细描述的**两种**视频片段补丁嵌入方法的影响。**（发现）**：在**图 6e** 中，压缩帧补丁嵌入方法的性能**明显落后**于均匀帧补丁嵌入方法。这一发现与视频理解方法 ViViT 所得到的结果**相矛盾**。**（推测）**：我们推测，使用压缩帧补丁嵌入方法导致了**时空信号的丢失**，使得 Transformer 主干难以学习视频的分布。

**时间步类信息注入**。**（发现）**：如**图 6f** 所示，S-AdaLN 的性能**显著优于**所有标记。**（观点）**：我们认为这种差异可能源自于**所有标记**仅将时间步或标签信息引入模型的输入层，**（难以有效传播至整个模型）**：这可能面临有效传播至整个模型的挑战。**（观点）**：相比之下，S-AdaLN 以**更适应性的方式**将时间步或标签信息编码到每个 Transformer 块中。**（观点）**：这种信息传递方法似乎更有效，可能有助于提高性能并加快模型收敛速度。

**时间位置嵌入**。**图 6b** 展示了**两种不同**的时间位置嵌入方法对模型性能的影响。**（发现）**：采用绝对位置嵌入方法往往会比备选方法产生稍微**更好**的结果。

**学习策略增强视频生成**。如**图 6c**所示，我们观察到训练的初始阶段**受益于**在 ImageNet 上对模型进行**预训练**，**（目的）：使其能够快速实现在视频数据集上的高质量性能**。**（局限性）：**然而，随着迭代次数的增加，使用预训练模型初始化的模型性能往往会在某个水平附近**稳定下来**，远远低于使用随机初始化的模型。

**（解释）：**这一现象可以通过两个因素来解释：1）在 ImageNet 上预训练的模型提供了**良好的表示**，这可能有助于模型在早期阶段**快速收敛**；2）ImageNet 和 FaceForensics 之间的**数据分布存在显著差异**，这使得模型难以将在 ImageNet 上学到的知识适应到 FaceForensics 上。

正如**表 2** 和**表 1** 所示，**（发现）：**我们发现图像-视频联合训练（“Latte+IMG”）**显著提高了** FID 和 FVD。沿时间轴连接额外的随机采样帧到视频中**（好处）：**使得模型能够容纳更多的示例在每个批次中，这可以增加训练模型的多样性。

**视频采样间隔**。我们探索了构建每个训练视频的 16 帧剪辑的**各种采样率**。**（发现）：**如**图 6a** 所示，在训练过程中，使用**不同采样率**的模型在**早期阶段存在显著的性能差距**。然而，随着训练迭代次数的增加，性能逐渐变得一致，**（结论）：这表明不同的采样率对模型性能影响不大**。我们选择了采样间隔为 3，**（目的）：以确保生成的视频具有合理的连续性，以进行与最先进方法的比较实验**。

**模型变体**。我们评估了 Latte 的模型变体，如第 3.2 节所述。我们努力使所有不同模型的参数**数量相等**，**（目的）：以确保公平比较**。我们从头开始训练所有模型。**（发现）：**如**图 6d** 所示，**变体 1 随着迭代次数的增加表现最佳**。值得注意的是，与其他三种模型变体相比，**变体 4 的浮点运算操作（FLOPs）大约只有四分之一**，如**表 3** 所详细说明的那样。因此，**变体 4 在四个变体中表现最不理想**并不足为奇。

| | Variant 1 | Variant 2 | Variant 3 | Variant 4 |
|---|---|---|---|---|
| Params (M) | 673.68 | 673.68 | 676.33 | 676.44 |
| FLOPs (G) | 5572.69 | 5572.69 | 6153.15 | 1545.15 |

**表 3**：不同模型变体的参数数量和 FLOPs（浮点运算）。

在变体 2 中，一半的 Transformer 块最初**（目的）：**用于空间建模，然后剩余的一半**（目的）：**用于时间建模。**（缺陷）：**这种划分可能会导致在后续的时间建模过程中丧失空间建模能力，最终影响性能。**（观点）：**因此，我们认为使用完整的 Transformer 块（包括多头注意力、层归一化和多线性投影）可能比仅使用多头注意力（变体 3）更有效地建模时间信息。

| Model | Layer numbers N | Hidden size D | Heads H | Param |
|---|---|---|---|---|
| Latte-S | 12 | 384 | 6 | 32.48M |
| Latte-B | 12 | 768 | 12 | 129.54M |
| Latte-L | 24 | 1024 | 16 | 456.81M |
| Latte-XL | 28 | 1152 | 16 | 673.68M |

**表 4**：Latte 模型的详细信息。我们根据不同的模型尺寸遵循 ViT 和 DiT 模型的配置。

**模型大小**。根据**表 4**，我们在 FaceForensics 数据集上训练了**四个不同尺寸**的 Latte 模型（XL、L、B 和 S）。**图 8** 清晰地展示了随着训练迭代次数增加，相应的 FVD 的变化趋势。**（发现）：**可以清楚地观察到，**增加模型大小通常与显著的性能改善相关**，这也在图像生成工作中得到了指出（Peebles 和 Xie，2023）。

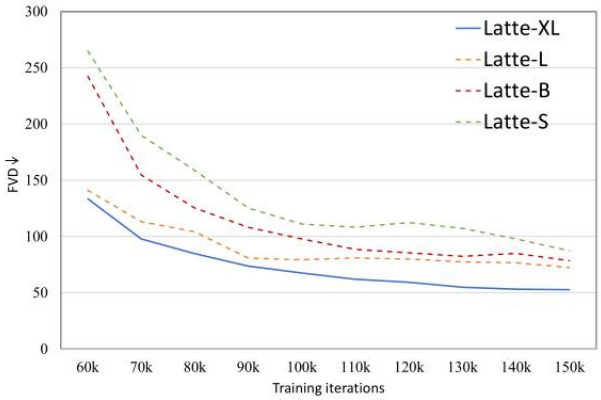



**图 8**：不同 Latte 模型尺寸的性能。通常，增加模型的尺寸可以显著提高其性能。

#### 4.3 与最先进技术的比较

根据第 4.2 节的消融研究，我们可以得到基于 Transformer 的潜在视频扩散模型的最佳实践（即，**模型变体 1**、**统一帧块嵌入**、**S-AdaLN** 和**绝对位置嵌入方法**、**图像-视频联合训练**）。我们使用这些最佳实践下提出的 Latte 对当前最先进技术进行比较。

**定性结果**。**图 5** 展示了 Latte 在 UCF101、Taichi-HD、FaceForensics 和 SkyTimelapse 上的视频合成结果。我们的方法在所有场景中始终提供逼真、高分辨率（256x256 像素）的视频生成结果。这包括捕捉人脸的运动和处理运动员的显著转换。值得注意的是，我们的方法在挑战性较大的 UCF101 数据集中合成高质量视频方面表现出色，而其他比较方法在这个任务上通常会出现问题。更多结果可在项目网站上查看。

| Method | IS ↑ | FID ↓ |
|---|---|---|
| MoCoGAN | 10.09 | 23.97 |
| VideoGPT | 12.61 | 22.7 |
| MoCoGAN-HD | 23.39 | 7.12 |
| DIGAN | 23.16 | 19.1 |
| StyleGAN-V | 23.94 | 9.445 |
| PVDM | 60.55 | 29.76 |
| Latte (ours) | 68.53 | 5.02 |
| Latte+IMG (ours) | **73.31** | **3.87** |

**表 1**：在 UCF101 和 FaceForensics 数据集上，Latte 与其他最先进技术的创新得分和 FID（Fréchet 创新距离）比较。我们使用 PVDM 提供的预训练模型生成相应的视频，并报告它们的对应值。此处，“IMG”表示视频-图像联合训练。

| Method | FaceForensics | SkyTimelapse | UCF101 | Taichi-HD |
|---|---|---|---|---|
| MoCoGAN | 124.7 | 206.6 | 2886.9 | - |
| VideoGPT | 185.9 | 222.7 | 2880.6 | - |
| MoCoGAN-HD | 111.8 | 164.1 | 1729.6 | 128.1 |
| DIGAN | 62.5 | 83.11 | 1630.2 | 156.7 |
| StyleGAN-V | 47.41 | 79.52 | 1431.0 | - |
| PVDM | 355.92 | 75.48 | 1141.9 | 540.2 |
| MoStGAN-V | 39.70 | 65.30 | 1380.3 | - |
| LVDM | - | 95.20 | 372.0 | 99.0 |
| Latte (ours) | 34.00 | 59.82 | 477.97 | 159.60 |
| Latte+IMG (ours) | **27.08** | **42.67** | **333.61** | **97.09** |

**表 2**：不同数据集上视频生成模型的 FVD（Fréchet 视频距离）值。其他基线模型的 FVD 值根据参考文献 StyleGAN-V 或原始论文报告。此外，我们使用 PVDM 的官方代码，严格遵循训练方法，在 FaceForensics 和 TaichiHD 上重新训练，并报告它们的 FVD 结果。同时，我们使用 PVDM 提供的 UCF101 和 SkyTimelapse 的预训练模型生成相应的视频，并报告它们的 FVD 值。此处，“IMG”表示视频-图像联合训练。

**定量结果**。在**表 2** 中，我们分别提供了 Latte 和其他比较方法的定量结果。**（结果）**：我们的方法在所有数据集上**都显著优于先前的工作**，**（结论）**：**表明了我们方法在视频生成方面的优越性**。在**表 1** 中，我们报告了在 FaceForensics 上的 FID 和在 UCF101 上的 IS **（目的）**：**以评估视频帧质量**。**（结果）**：我们的方法表现出色，具有 3.87 的 FID 值和 73.31 的 IS 值，**明显超过其他方法的能力**。

#### 4.4 扩展到文本到视频生成

（目的）：为了探索我们提出的方法的潜力能力，我们将 Latte 扩展到文本到视频生成。我们采用图 2（a）中所示的方法构建我们的 Latte T2V 模型。第 4.2 节提到利用预训练模型可以促进模型训练。因此，我们利用预训练的 PixArt-α（512×512 分辨率）（Chen 等人，2023a）的权重（目的）：来初始化 Latte T2V 模型中的空间变换器块的参数。（出发点）：由于常用视频数据集 WebVid-10M（Bain 等人，2021）的分辨率低于 512×512，我们在高分辨率视频数据集 Vimeo25M 上训练我们的模型，该数据集由（Wang 等人，2023b）提出。我们在这两个数据集的子集上训练我们的 T2V 模型，其中包含大约 330,000 个文本-视频对。我们在图 7 中根据视觉质量与最近的 T2V 模型 VideoFusion（Luo 等人，2023）和 VideoLDM（Blattmann 等人，2023b）进行比较。结果显示，我们的 Latte 可以生成可比较的 T2V 结果。更多结果可以在我们的项目网站上找到。此外，我们选择了 2,048 个采样视频来计算 FVD 和 FID 分数。结果得到的 FVD 和 FID 值分别为 328.20 和 50.72。


图 5：在 UCF101、Taichi-HD、FaceForensics 和 SkyTimelapse 上，来自不同方法的样本视频。


(a) *A dog in astronaut suit and sunglasses floating in space.*


(b) *An astronaut flying in space, 4k, high resolution.*

**图 7**：文本条件视频样本。与当前领先的 VideoFusion 和 Align your Latents T2V 模型相比，Latte 取得了可比较的结果。我们利用 VideoFusion 的官方在线平台以及提供的提示生成视频。此外，我们使用 VideoLDM 官方网站上可用的视频，因为他们没有发布他们的代码和相关模型。

## 5 结论

本文提出了 Latte，一种简单而通用的**视频扩散**方法，它采用**视频 Transformer** 作为骨干来生成视频。**（目的）：为了提高生成的视频质量，（本研究贡献）：**我们确定了提出模型的最佳实践，包括剪辑块嵌入、模型变体、时间步类信息注入、时间位置嵌入和学习策略等。**（结论）：**综合实验表明，Latte 在四个标准视频生成基准上取得了最先进的结果。此外，与当前的 T2V 方法相比，我们也取得了可比较的文本到视频结果。**（展望）：**我们坚信 Latte 可以为未来关于将基于 Transformer 的骨干集成到视频生成的扩散模型以及其他模态的研究提供宝贵的见解。

## 6 数据可用性声明

本研究结果的支持数据是公开可用的。