





# REUSE AND DIFFUSE: ITERATIVE DENOISING FOR TEXT-TO-VIDEO GENERATION

REUSE AND DIFFUSE : ITERATIVE DENOISING FOR TEXT-TO-VIDEO GENERATION) 学习笔记

method  
    视频数据  
    长视频生成  
experiment

## REUSE AND DIFFUSE: ITERATIVE DENOISING FOR TEXT-TO-VIDEO GENERATION

Jiayi Gu<sup>1</sup>   Shicong Wang<sup>2</sup>   Haoyu Zhao<sup>2</sup>   Tianyi Lu<sup>2</sup>   Xing Zhang<sup>2</sup>   Zuxuan Wu<sup>2</sup>  
Songcen Xu<sup>1</sup>   Wei Zhang<sup>1</sup>   Yu-Gang Jiang<sup>2</sup>   Hang Xu<sup>1</sup>

<sup>1</sup>Huawei Noah's Ark Lab      <sup>2</sup>Fudan University  
chromexbjxh@gmail.com      zxwu@fudan.edu.cn      CSDN @kangxi11122344

### motivation:

由于计算和内存资源的限制，将LDM用于t2v任务挑战性较大  
单个LDM通常只能生成非常有限的视频帧数，且需要额外的训练成本和帧级抖动

### contribution:

Reuse and Diffuse（重用和扩散），根据LDM已经生成的帧生成更多的帧（以具有少量帧的初始视频片段为条件，通过重用原始潜在特征并遵循先前的额外的帧。）  
对autoencoder插入时间层进行finetuning实现时间一致性  
还提出了一套策略，用于组合视频文本数据

### method

stable diffusion在t2i任务中表现较好，视频合成任务会加载预训练的LDM（Variational Auto-Encoder (VAE)和U-Net）  
通过注入图中虚线框标记的temporal layer（时间层）来适应原始的U-Net，进行图像扩散到视频合成的转变。

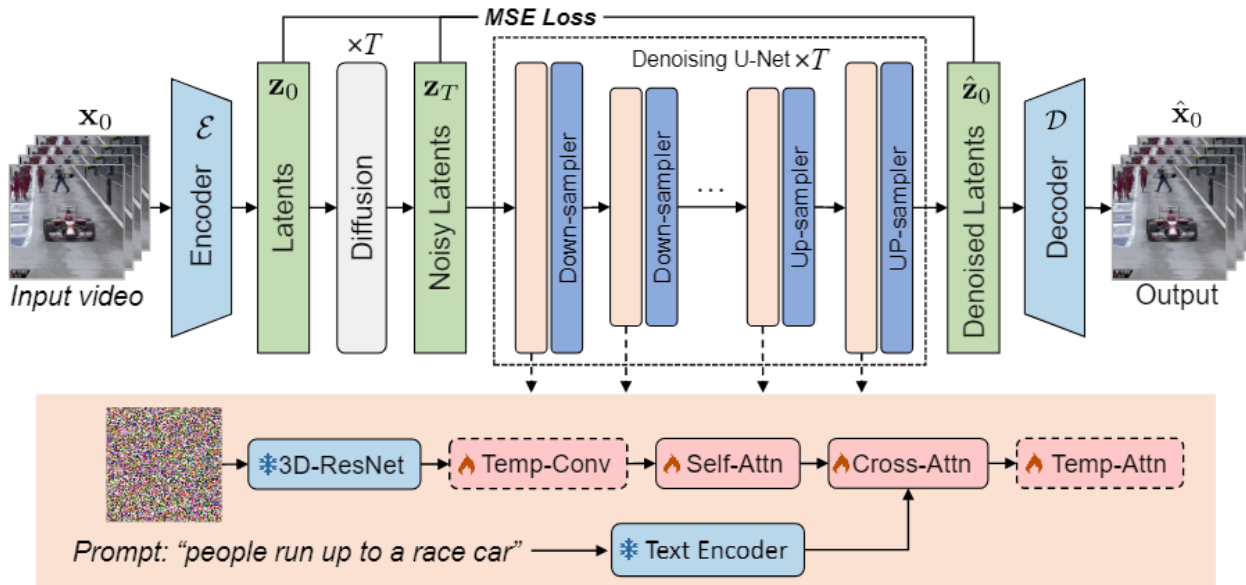


Figure 2: The architecture of VidRD is derived from an LDM for image synthesis. Modules with snowflake marks are frozen while those with flame marks are trainable. Modules with dashed boxes are added in addition to the original LDM for image synthesis.

CSDN @kangxi11122344

temporal layer分为Temp-Conv（3D卷积层）、Temp-Attn（temporal attention layers）

除了这两个层，其他大多数网络层都使用stable diffusion的预训练模型 权重进行初始化

Temp-Conv 和 Temp-Attn 的参数随机初始化

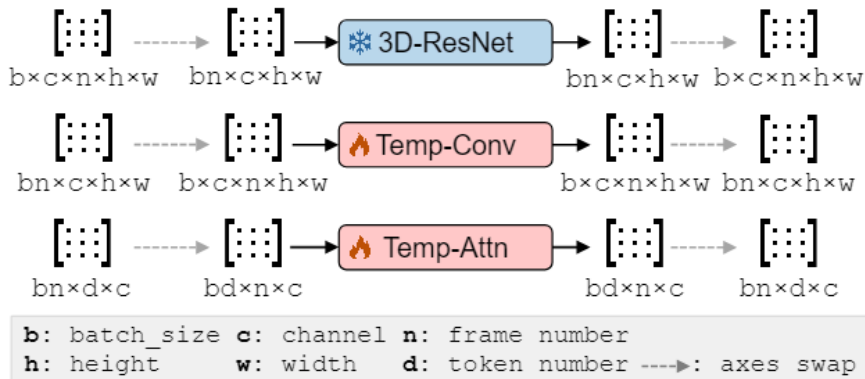


Figure 3: These are three essential network layers in VidRD. *3D-ResNet*, inherited from Stable Diffusion, treats the number of frames  $n$  as a part of batch size. This is equivalent to applying the original *2D-ResNet* frame by frame so this layer is frozen in model training. *Temp-Conv*, implemented with 3D convolutions, processes video inputs in a tube manner while *Temp-Attn* applies attention layer along temporal axis.

CSDN @kangxi11122344

2D ResNet膨胀为3D ResNet

网络层中只有一部分是可训练的，以实现有效的训练，之前的工作，使用图像数据分别微调 空间层和视频数据来训练时间层  
本工作，以端到端的方式使用纯视频数据进行训练，因为图像数据被转换为伪视频，显示出与原始视频数据相似的时间一致性

视频数据

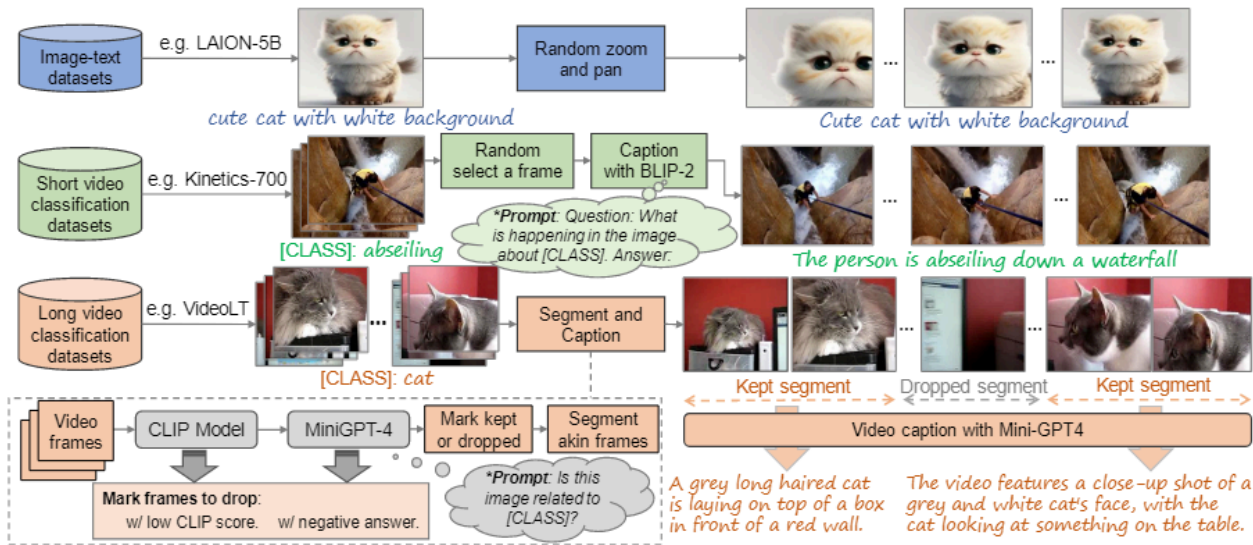


Figure 4: A set of strategies is devised for processing different types of datasets including image-text datasets, short video classification datasets, and long video classification datasets. CSDN @kangxi11122344

图像-文本数据通过随机缩放和平移来生成多个图像，并进一步组合成伪视频。

短视频的数据集（Kinetics-700），根据每个视频的分类标签给出合适的文本caption。

包含多个场景的长视频（VideoLT），segment-then-caption strategy，先分段在加字幕（Mini-GPT4）。

## 长视频生成

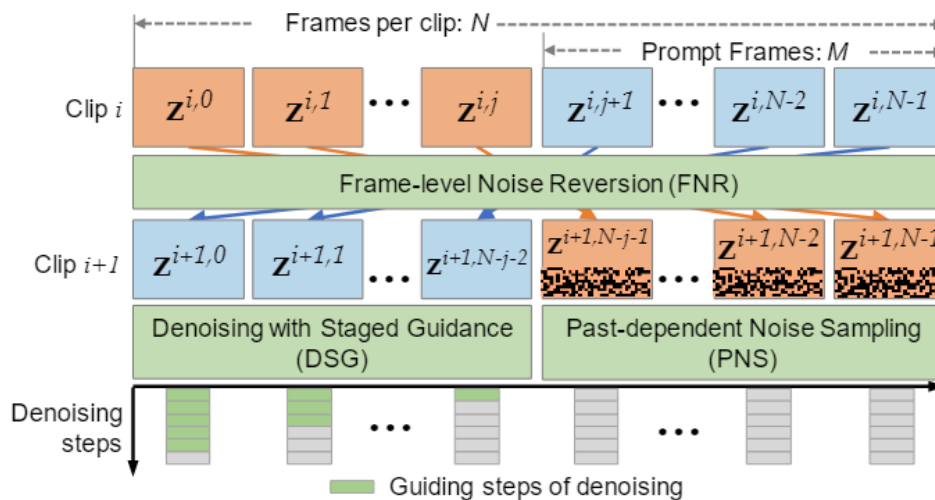


Figure 5: Videos can be generated clip by clip iteratively with a single LDM. After each iteration,  $N$  frames are generated and the last  $M$  frames are used as prompt frames for the next iteration. Three key strategies are proposed for generating natural and smooth videos. Frame-level Noise Reversion (FNR) is used as a basic module for re-using the initial noise in a reversed order from the last video clip. Past-dependent Noise Sampling (PNS) brings new random noise for the last several video frames. Temporal consistencies between video clips are refined by Denoising with Staged Guidance (DSG). CSDN @kangxi11122344

FNR：为了生成平滑的视频，迭代地重用初始噪声，但每次以相反的顺序。

PNS：为了减轻视频内容循环程度，

$$\begin{cases} z_T^{i-1,N-j-1} & \text{if } j < M \\ \frac{\alpha}{\sqrt{1+\alpha^2}} z_T^{i-1,N-j-1} + \epsilon^{i,j} & \text{otherwise} \end{cases}$$

$$\epsilon_{i,j} \sim \mathcal{N}(0, \frac{1}{1+\alpha^2})$$
$$\alpha_{i,j} \geq 0$$
$$\mathbf{z}_{i,j} = \{\mathbf{z}_{i-1,N-j-1} \quad 1+\alpha^2$$

√

$$\mathbf{z}_{i-1,N-j-1} + \epsilon_{i,j} \text{ if } j < M$$
$$\text{otherwise } \mathbf{z}_{i,j} \sim \mathcal{N}(0, 1 + \alpha^2)$$
$$\alpha \geq 0$$

M帧为参考帧，在参考帧之外加额外的随机噪声，α越小，随机噪声占的比率越大

DSG：提高帧之间的连续性，主要是视频clip 之间的连续性， $\mathbf{z}_0^{i,N-1}$  和  $\mathbf{z}_0^{i+1,0}$ ，

$$\mathbf{z}_{t-1}^{i,j} = \begin{cases} \mathbf{z}_{t-1}^{i-1,N-j-1} & \text{if } t > (1-\beta)T + \frac{\beta T j}{M}, \beta \in [0, 1] \\ \text{DDIM}(\mathbf{z}_t^{i,j}, t) & \text{otherwise} \end{cases}$$

前几帧重用上一视频clip的latent futures,β越小，重用程度越小

experiment

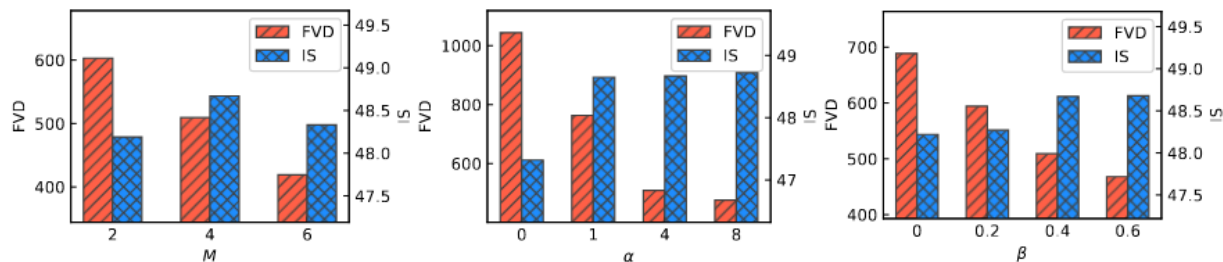
由静态图像的随机缩放和平移产生的伪视频有助于提高时间一致性但损害视觉外观

Strategy	IS ↑	FVD ↓
VidRD w/o pseudo-videos	42.00	451.16
VidRD w/ pseudo-videos	40.87	433.22

Table 3: This is a comparison between fine-tuning VidRD with and without pseudo-videos by random zooming and panning static images. FVD and IS are evaluated on UCF-101 and all experiments here are in a zero-shot manner.

使用伪视频微调和不使用伪视频微调

与使用静态图像仅训练空间层相比，由静态图像的随机缩放和平移产生的伪视频有助于提高时间一致性但损害视觉外观。



(a) Ablation studies on M with α = 4 and β = 0.4. (b) Ablation studies on α in PNS with M = 4 and β = 0.4. (c) Ablation studies on β in DSG with M = 4 and α = 0.4.

对M（参考帧数量）、α、β消融实验

越大，重用的越多，量化指标较高，视频循环严重