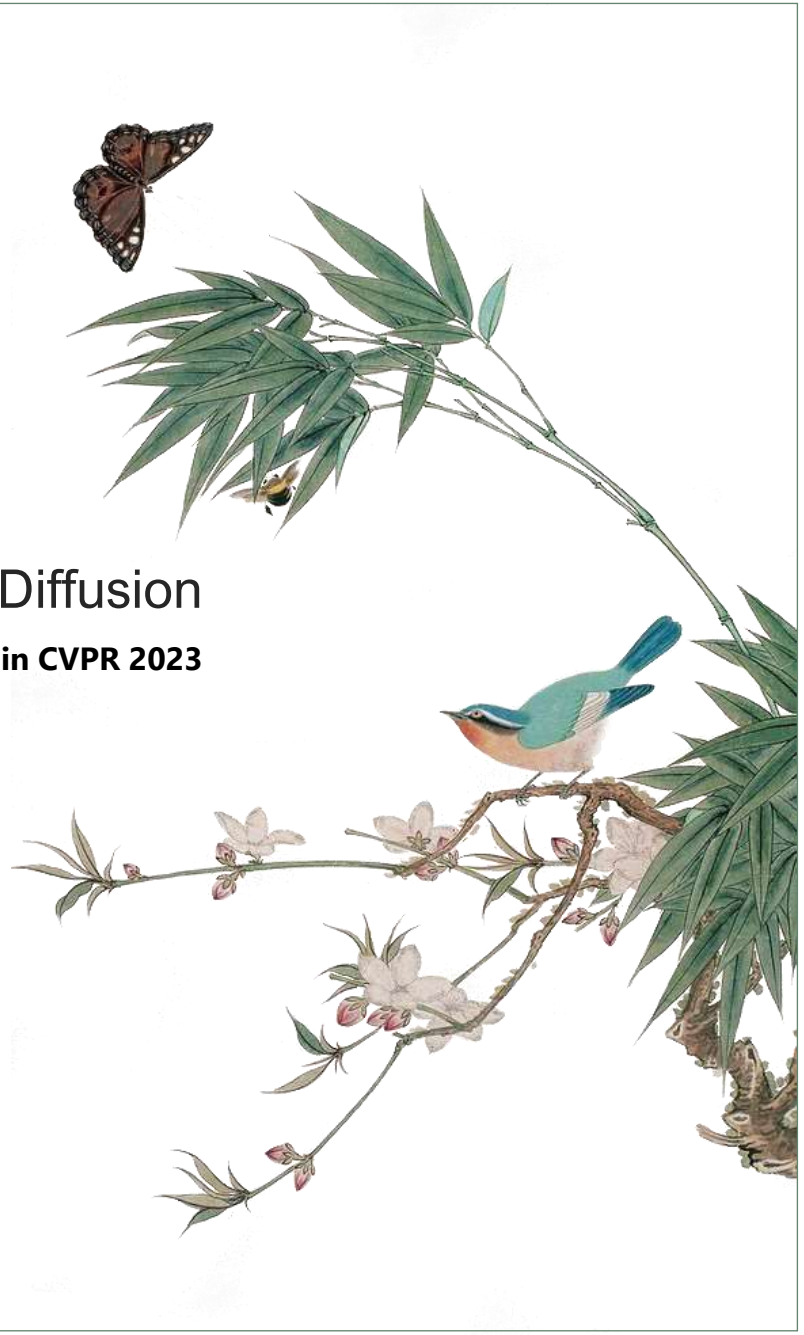
Conditional Image-to-Video Generation with Latent Flow Diffusion

included in CVPR 2023

Haomiao Ni[1],Changhao Shi[2],Kai Li[3],Sharon X[1]. Huang,Martin Renqiang Min[3]
[1]The Pennsylvania State University, University Park, PA, USA
[2]University of California, San Diego, CA, USA
[3]NEC Laboratories America, Princeton, NJ, USA



目录 | CONTENTS

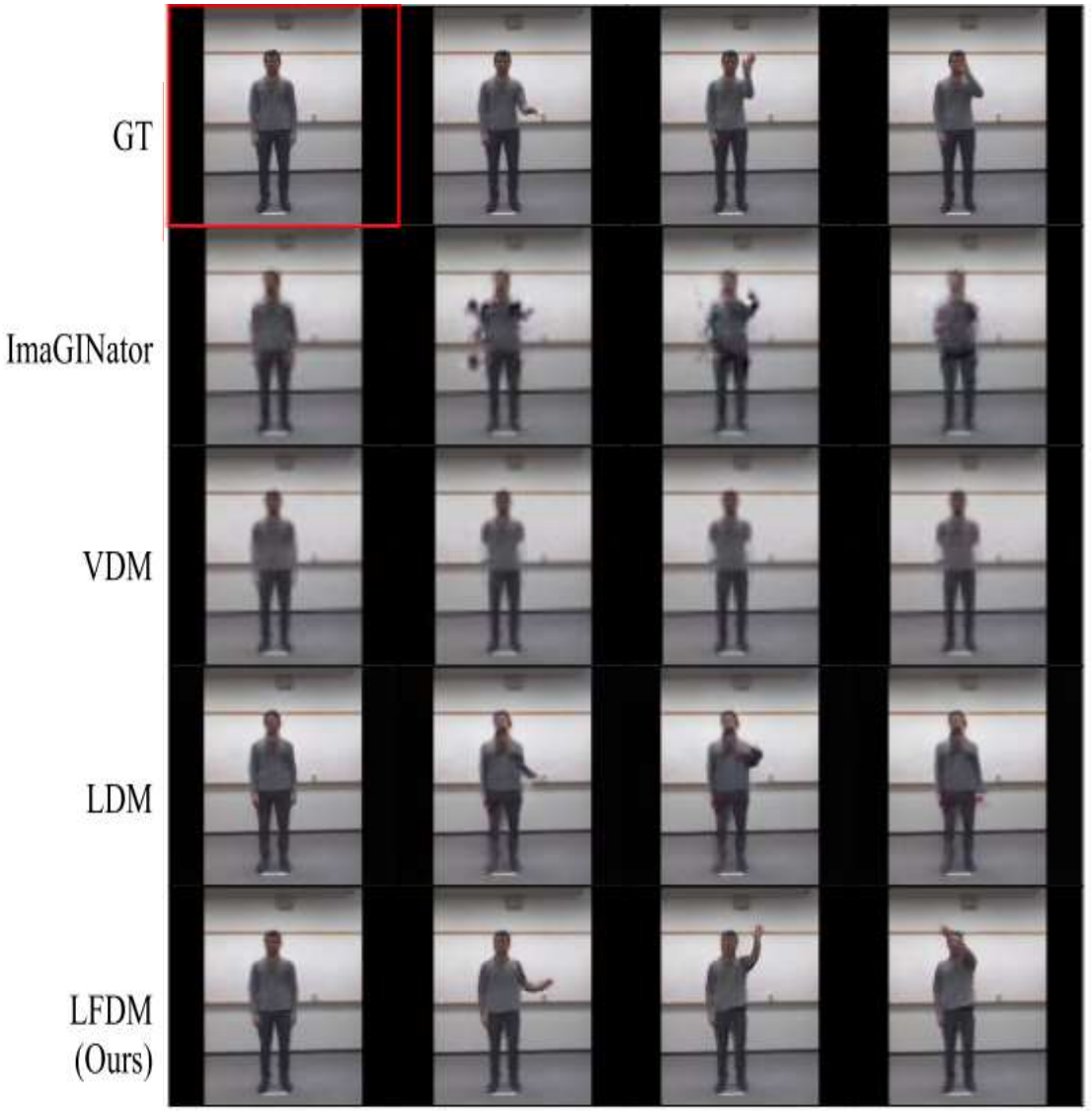
研 究 简 要 概 述

研究领域概述

本文研究的应用领域属于条件图生视频(Conditional image-to-video)。该领域的研究目标是基于一张图像（如一个人的脸）和一个文字描述的条件（如惊讶）合成一个新的可信的视频，它可以应用于艺术创作、娱乐以及用于机器学习的数据增强。其难点在于同时生成与给定图像和条件对应的空间外观和时间性动态

此前研究的不足

此前该领域绝大部分的研究工作是基于直接合成(direct-synthesis-based works), 即通过初始图片x_0和条件y分别合成每一帧的图片来生成视频。这样会导致很难同时保持空间细节和时间的一致。



"Right Hand Wave" (MHAD)

本 文 的 工 作

本文提出了一种新的潜流扩散模型(latent flow diffusion model)，官方命名为LFDM，与之前的直接合成的基于无翘曲的方法不同，LFDM更好利用给定图像的空间内容，在潜空间中进行变形来合成细节和运动

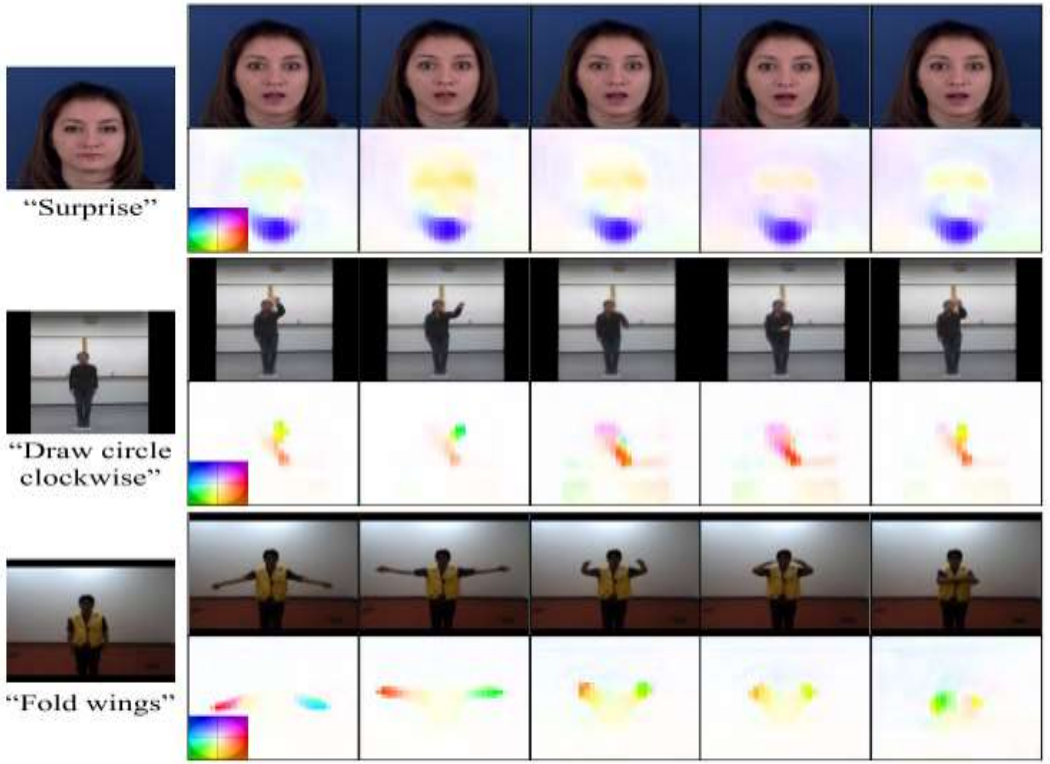
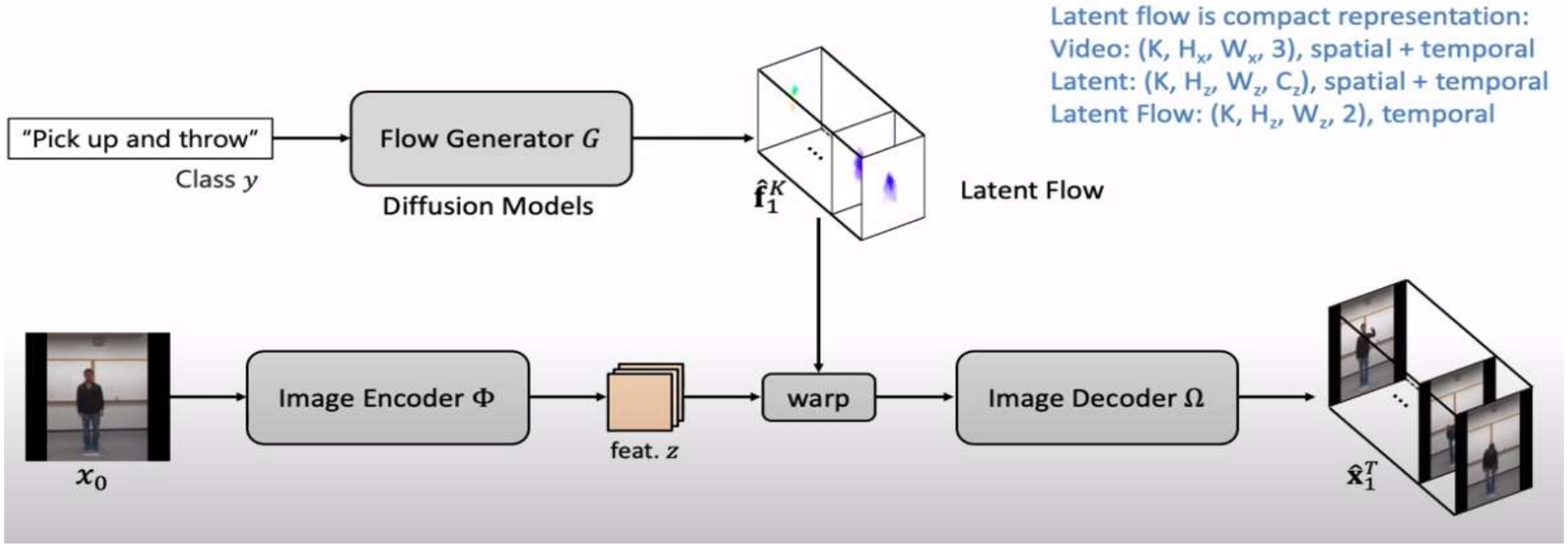


Figure 1. Examples of generated video frames and latent flow sequences using our proposed LFDM. The first column shows the given images x_0 and conditions y. The latent flow maps are *backward* optical flow to x_0 in the *latent* space. We use the color coding scheme in [4] to visualize flow, where the color indicates the direction and magnitude of the flow.

LFDM的算法过程

设 $n \sim \mathcal{N}(0, I)$是体积和形状为 $K_n \times H_n \times W_n \times C_n$ 的高斯噪声体积，其中 K_n、H_n、W_n 和C_n分别为长度、高度、宽度和通道数。给定一个初始图片 x_0 和一个条件 y，假设 $\mathbf{x}_0^K = \{x_0, x_1, \ldots, x_K$ 是基于条件y 的真实视频，cI2V生成的目标就是学习得出将噪声体积n转化为合成视频 $\tilde{\mathbf{x}}_1^K = \{\hat{x}_1, \ldots, \hat{x}_K\}$ 的映射。



本 研 究 的 优 势

由于LFDM是通过合成以y为条件的潜在光流序列从而在潜在空间中扭曲图像x_0以生成新的视频。所以LFDM可以更好地保存受试者的外观，保证运动的连续性，也可以推广到看不见的图像。

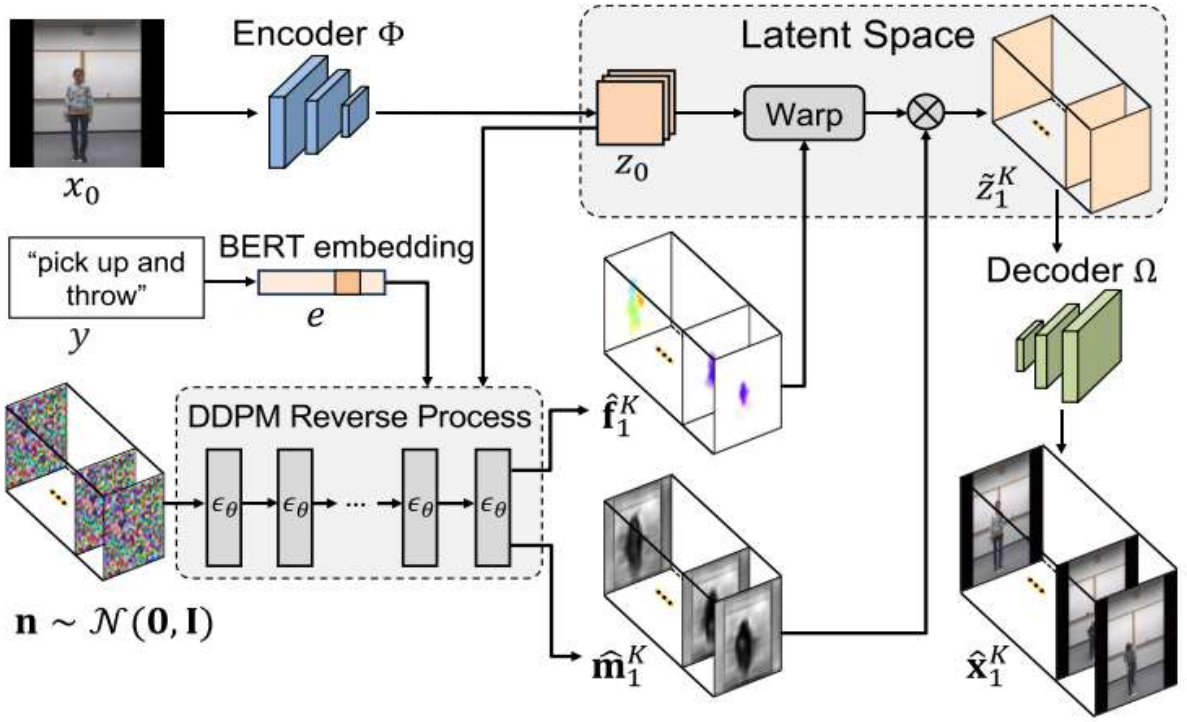
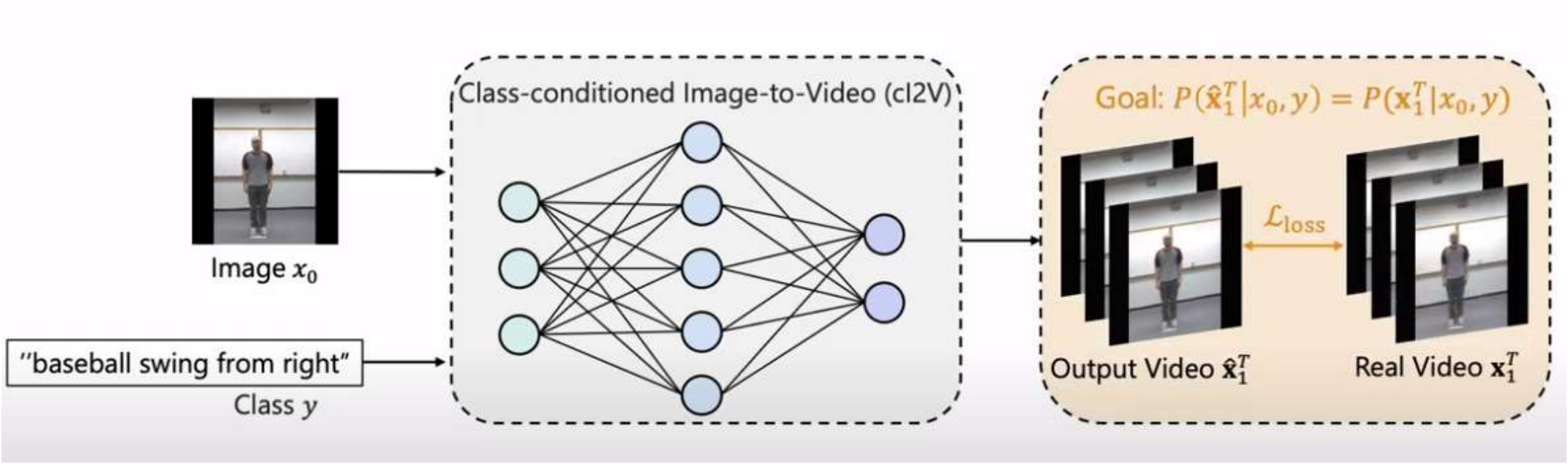


Figure 2. The video generation (*i.e.*, inference) process of LFDM. The generated flow sequence $\hat{\mathbf{f}}_1^K$ and occlusion map sequence $\hat{\mathbf{m}}_1^K$ have the same *spatial* size as image latent map z_0. The brighter regions in $\hat{\mathbf{m}}_1^K$ mean those are regions less likely to be occluded.

本　文　贡　献

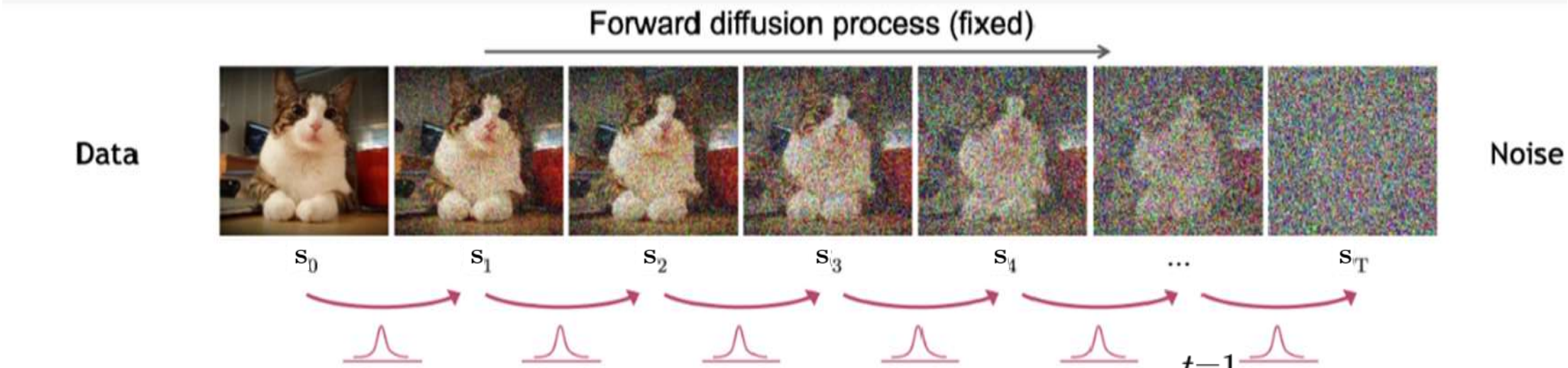
本 文 创 新 点

本文提出了一种新的潜流扩散模型(latent flow diffusion model, LFDM)，通过在给定条件下合成潜空间内的时间相干流序列来实现图像到视频的生成。本文是第一个应用扩散模型来产生潜在流来实现条件图像到视频任务的研究工作。



本 文 创 新 点

潜流扩散模型(latent flow diffusion model, LFDM)是基于去噪扩散概率模型(denoising diffusion probabilistic model, DDPM)。从数据分布$\mathbf{s}_0 \sim q(\mathbf{s}_0)$中给定一个样本，$DDPM$的前向过程通过根据方差表$\beta_1, \ldots, \beta_T$逐步向$\mathbf{s}_0$加入噪声从而产生一个马尔科夫链$\mathbf{s}_1, \ldots, \mathbf{s}_T$。



$$q(\mathbf{s}_t \mid \mathbf{s}_{t-1}) = \mathcal{N}\left(\mathbf{s}_t; \sqrt{1-\beta_t}\mathbf{s}_{t-1}, \beta_t \mathbf{I}\right) \longrightarrow q(S_{1:T}|S_0) = \prod_{T}^{t=1} q(S_t|S_{t-1})$$

本 文 创 新 点

其中方差β_t保持不变。当β_t较小时，后验概率$q(st_1|st)$可以用对角高斯逼近。

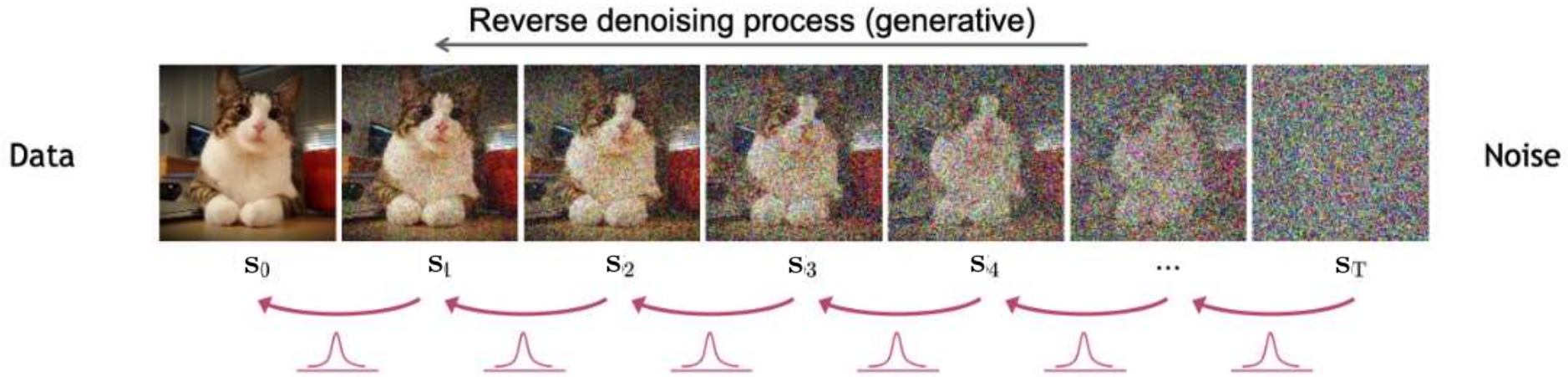
$$q(\mathbf{s}_t|\mathbf{s}_{t-1}) = \mathcal{N}(\mathbf{s}_t; \sqrt{1-\beta_t}\mathbf{s}_{t-1}, \beta_t\mathbf{I})$$

此外，如果链的 T 足够大，s_T可以很好地用标准高斯分布$\mathcal{N}(\mathbf{0}, \mathbf{I})$逼近，这就表示真正的后验概率$q(st_1|st)$。可以用$p\theta(s_t_1|st)$来估算

$$p_\theta(\mathbf{s}_{t-1}|\mathbf{s}_t) = \mathcal{N}(\mathbf{s}_{t-1}; \mu_\theta(\mathbf{s}_t), \sigma_t^2\mathbf{I})$$

本 文 创 新 点

DDPM的逆向过程(也称作采样)则是通过对一个高斯噪声$\mathbf{s}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$在一个马尔科夫链$\mathbf{s}_{T-1}, \mathbf{s}_{T-2}, \ldots, \mathbf{s}_0$中通过学习到的$p_\theta(\mathbf{s}_{t-1}|\mathbf{s}_t)$不断进行去噪的过程，最终产生一个样本$\mathbf{s}_0 \sim p_\theta(\mathbf{s}_0)$

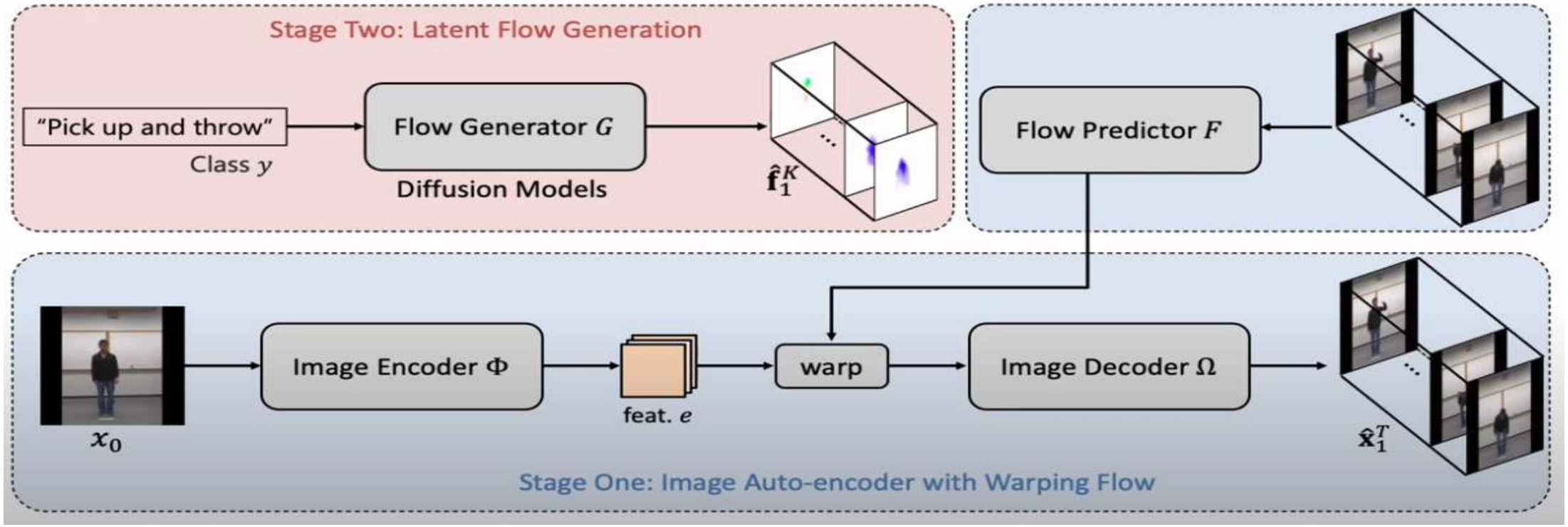


$$p(S_T) = \mathcal{N}(S_T; 0, I) \longrightarrow p_\theta(S_{0:T}) = p(S_T) \prod_T^{t=1} p_\theta(S_{t-1}|S_t)$$

$$p_\theta(\mathbf{S}_{t-1}|\mathbf{S}_t) = \mathcal{N}(\mathbf{S}_{t-1}; \mu_\theta(\mathbf{S}_t, t), \sigma_t^2 \mathbf{I})$$

本 文 创 新 点

针对$LFDM$，提出了一种新的两阶段训练策略，将空间内容生成和时间动态解耦，其中第一阶段训练潜流自编码器，第二阶段训练基于条件的3D $U-Ne$扩散模型。这种分离的训练过程也使$LFDM$很容易扩展到新的领域。



本 文 创 新 点

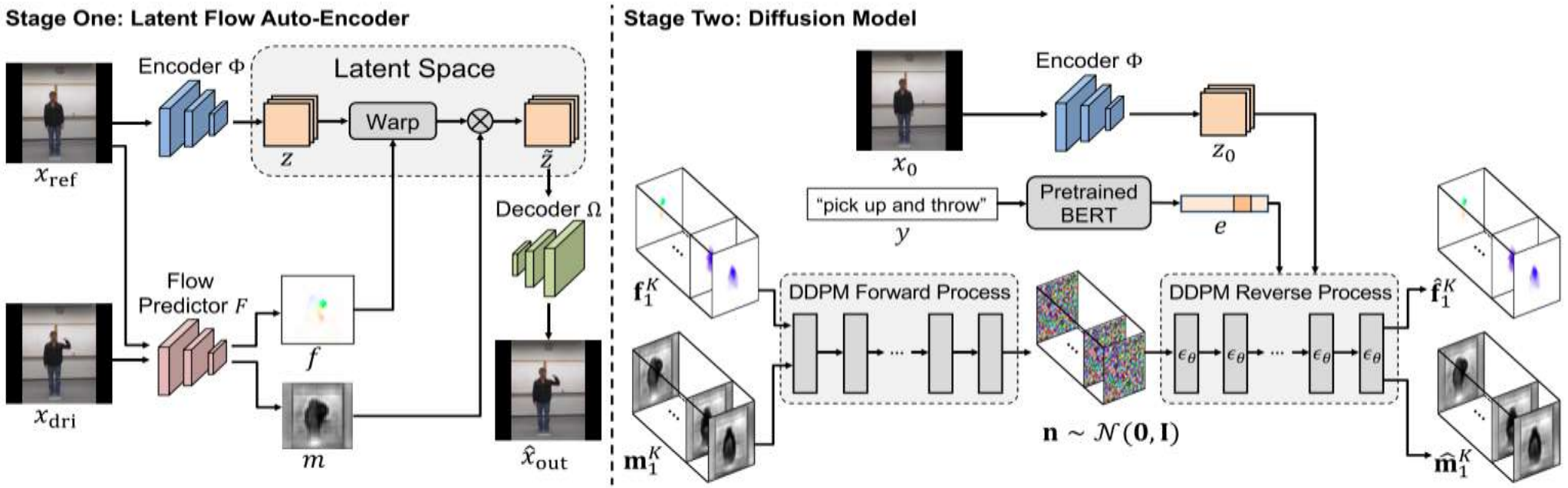


Figure 3. The training framework of LFDM. On the left is stage one for training latent flow auto-encoder while on the right is stage two for training diffusion model. In stage two, the encoder Φ is the one already trained in stage one, and the latent flow sequence $\mathbf{f}_1^K$ and occlusion map sequence $\mathbf{m}_1^K$ are estimated between x_0 and each frame in ground truth video $\mathbf{x}_1^K$ using the trained flow predictor F from stage one.

二阶段的编码器Φ是在第一阶段已经训练的编码器，利用第一阶段训练的流预测器F估计出x_0和基准视频$\mathbf{x}_1^K$每一帧之间的潜流序列$\mathbf{f}_1^K$和遮挡贴图序列m_1^K

本 文 创 新 点

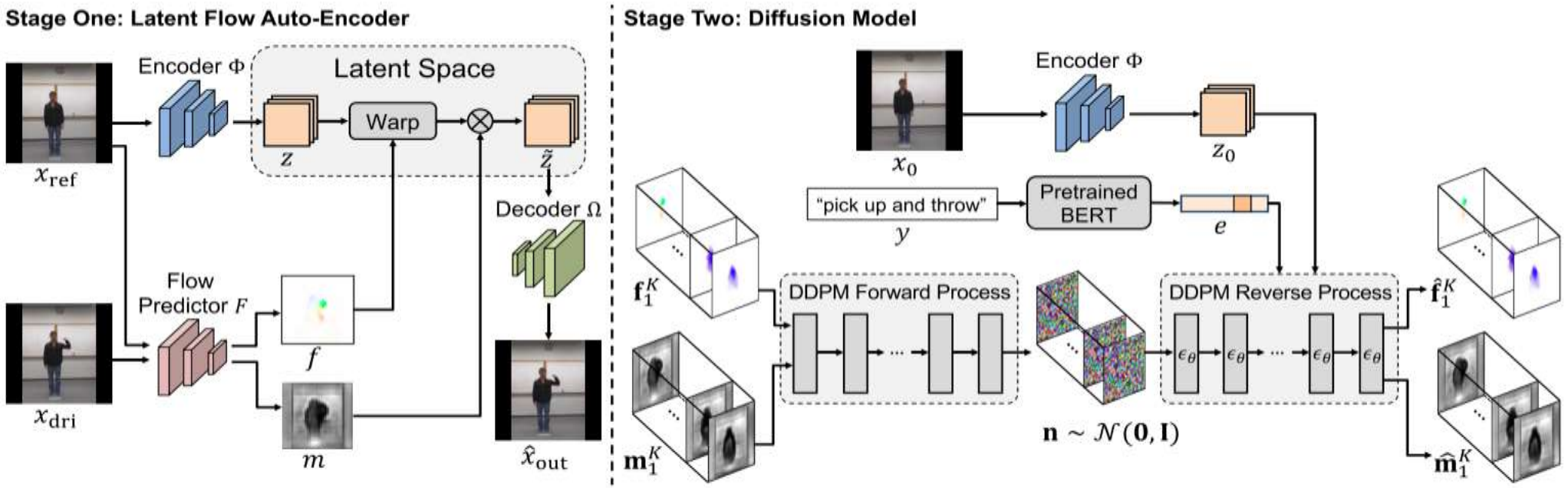


Figure 3. The training framework of LFDM. On the left is stage one for training latent flow auto-encoder while on the right is stage two for training diffusion model. In stage two, the encoder Φ is the one already trained in stage one, and the latent flow sequence $\mathbf{f}_1^K$ and occlusion map sequence $\mathbf{m}_1^K$ are estimated between x_0 and each frame in ground truth video $\mathbf{x}_1^K$ using the trained flow predictor F from stage one.

第一阶段，以无监督方式训练潜在流自动编码器($LFAE$)来估计视频帧、参考帧和驱动帧之间的潜在流，并将参考帧与潜在流一起扭曲重构驱动帧

第二阶段，我们训练一个基于三维$unet$的扩散模型(DM)来产生以图像x_0和类y为条件的时间相干潜流序列

本 文 创 新 点

本文在多个数据集上进行了广泛的实验，包括面部表情、人类动作和手势的视频，结果表明，LFDM方法始终优于以前的最先进的方法。

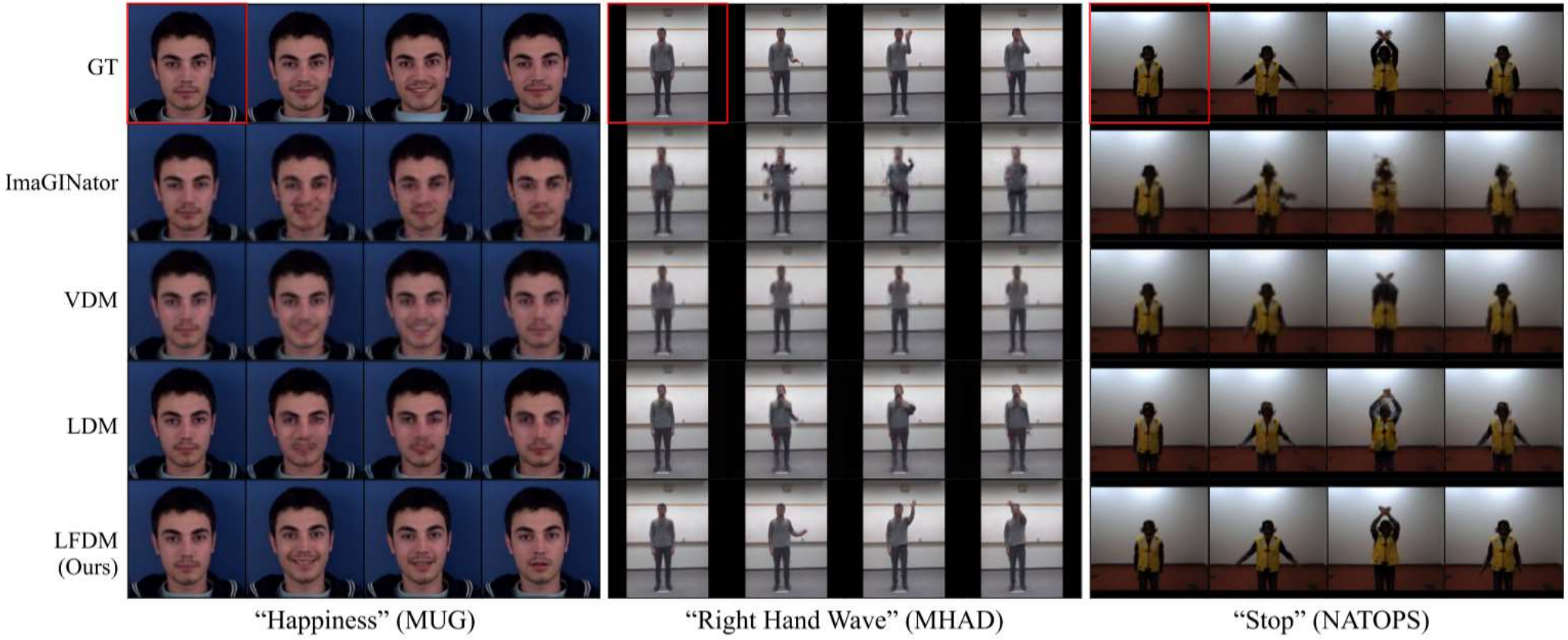


Figure 4. Qualitative comparison among different methods on multiple datasets for cI2V generation. First image frame x_0 is highlighted with red box and condition y is shown under each block. To simplify coding, all the models are designed to also generate starting frame $\hat{x}_0$. The video frames of GT (ground truth), LDM and LFDM have 128×128 resolution while results of ImaGINator and VDM are 64×64.

本文的实验结果

本文实验数据

本文对以下数据集采取了全面的实验，分别是MUG,MHAD,NATOPS。

Model	MUG			MHAD			NATOPS		
	FVD↓	cFVD↓	sFVD↓	FVD↓	cFVD↓	sFVD↓	FVD↓	cFVD↓	sFVD↓
ImaGINator [77]	170.73	257.46±62.88	319.37±95.23	889.48	1406.56±260.70	1175.74±327.99	721.17	1122.13±150.74	1042.69±416.16
VDM [27]	108.02	182.90±69.56	213.59±97.70	295.55	531.20±104.25	398.09±121.16	169.61	410.71±105.97	350.59±125.03
LDM_{64} [57]	123.88	196.49±66.99	236.26±76.08	280.26	515.29±125.70	427.03±112.31	251.72	506.40±125.08	491.37±231.85
$LFDM_{64}$ (Ours)	**27.57**	**77.86±20.27**	**108.36±39.60**	**152.48**	**339.63±52.88**	**242.61±28.50**	**160.84**	**376.14±106.13**	**324.45±116.21**
LDM_{128} [57]	126.28	208.03±64.86	241.49±75.18	337.43	594.34±150.31	497.50±110.16	344.81	627.84±169.52	623.13±320.85
$LFDM_{128}$ (Ours)	**32.09**	**84.52±24.81**	**114.33±42.62**	**214.39**	**426.10±63.48**	**328.76±34.42**	**195.17**	**423.42±117.06**	**369.93±159.26**

Table 1. Quantitative comparison among different methods on multiple datasets for cI2V generation. The 64 and 128 in the subscript of LDM and LFDM indicate that the resolution of synthesized video frames are 64×64 and 128×128, respectively.

FVD(Fr´echet Video Distance)用来衡量生成视频的视觉质量，时间一致性和样本多样性，FVD首先使用在Kinetics-400数据集上预训练的视频分类网络I3D，获得真实视频和合成视频的特征表示。然后计算真实视频特征分布与合成视频特征分布之间的距离。

本 文 的 启 发

本 文 的 启 发

尽管本文在多个方面都取得了突出成果但是仍然存在一些可以继续探究的地方

1. 目前的LFDM实验仅限于包含单个运动对象的视频

2. 当前的LFDM是基于类标签而不是自然文本描述

3. 与GAN模型相比，在采用1000步DDPM采样时，LFDM模型的采样速度要慢得多

本 文 的 启 发

针对本文目前的一些问题我们可以做如下工作：

1. 将LFDM拓展应用到多对象流的视频生成
2. 将LFDM与transformer结合从而将条件拓展到自然文本描述
3. 研究一些快速采样方法应用到DDPM的采样流程中

本文的代码仓库

https: //github.com/nihaomiao/CVPR23_LFDM

其 他 相 关 论 文

Denoising Diffusion Probabilistic Models, NeurIPS 2020

Jonathan Ho, Ajay Jain, Pieter Abbeel

UC Berkeley

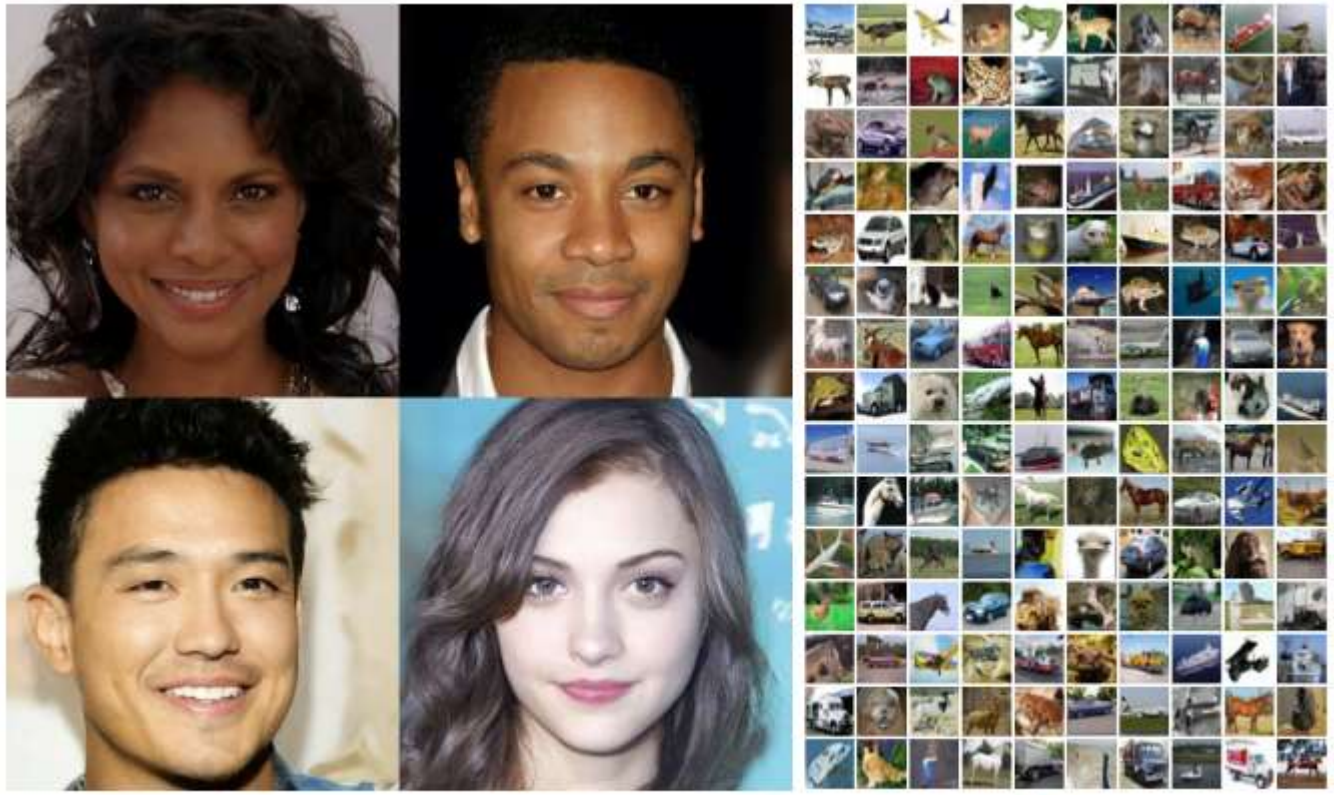


Figure 1: Generated samples on CelebA-HQ 256 × 256 (left) and unconditional CIFAR10 (right)

其 他 相 关 论 文

FVD: A new Metric for Video Generation, ICLR 2019

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, Sylvain Gelly
Google Brain

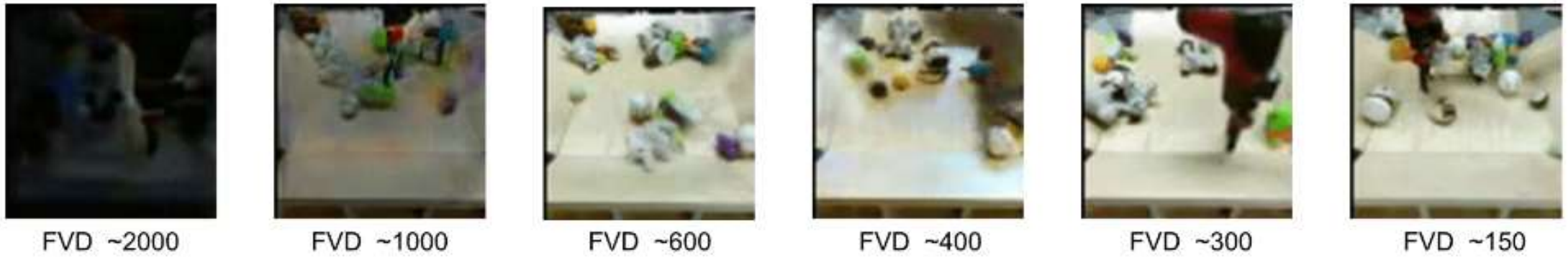


Figure 1: Generated videos by various models ranked according to FVD (lower is better).

其 他 相 关 论 文

Video Probabilistic Diffusion Models in Projected Latent Space,CVPR2023
[1]Sihyun Yu,[2]Kihyuk Sohn, [1]Subin Kim, [1]Jinwoo Shin
1KAIST 2Google Research

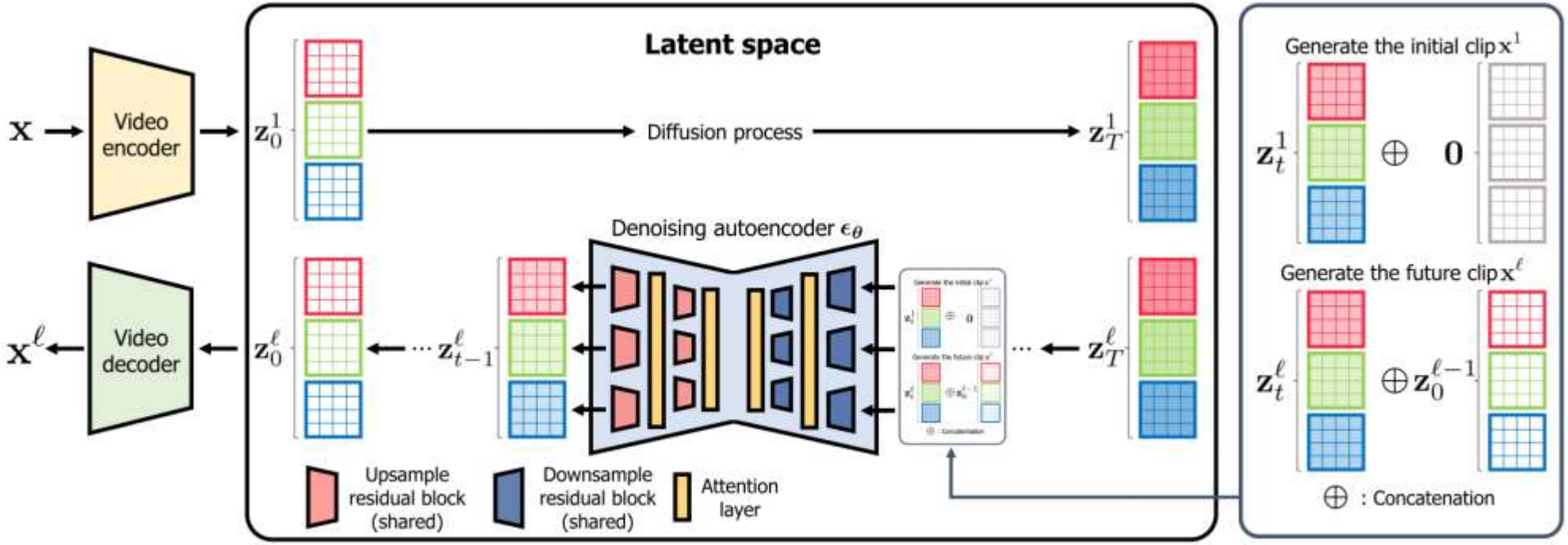


Figure 1. Overall illustration of our projected latent video diffusion model (PVDM) framework. PVDM is composed of two components: (a) (left) an autoencoder that maps a video into 2D image-like latent space (b) (right) a diffusion model operates in this latent space.

其 他 相 关 论 文

HouseDiffusion: Vector Floorplan Generation via a Diffusion Model with Discrete and Continuous Denoising,CVPR2023
Mohammad Amin Shabani, Sepidehsadat Hosseini, Yasutaka Furukawa
Simon Fraser University

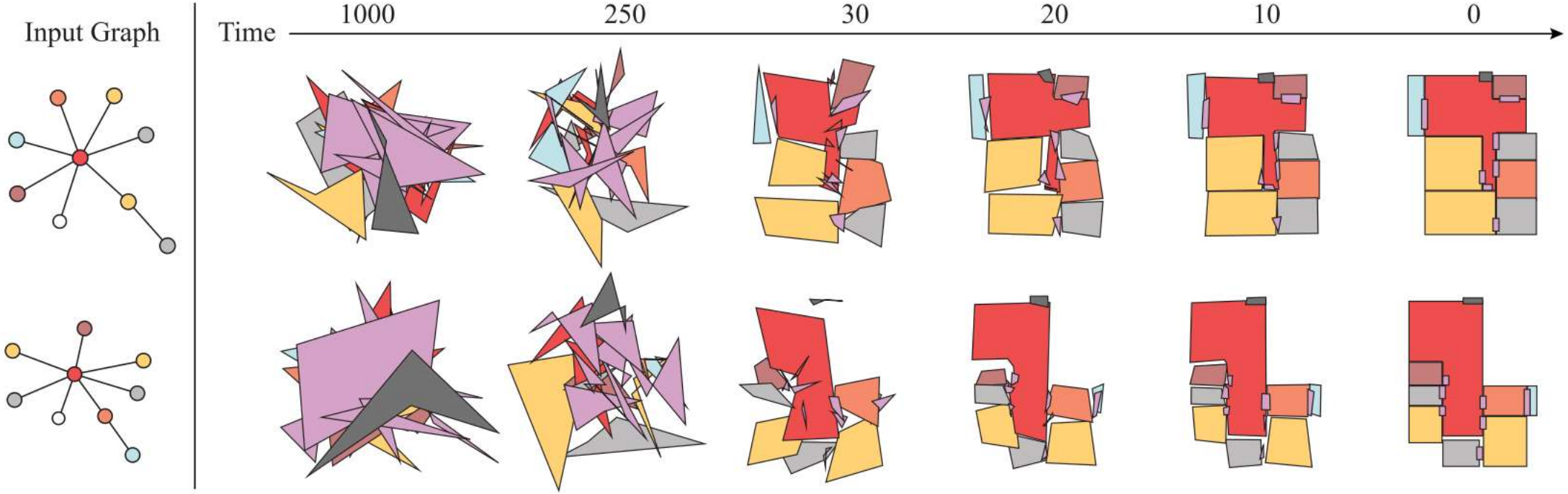


Figure 1. Given a bubble diagram as the input constraint, HouseDiffusion directly generates a vector floorplan by initializing the room/door coordinates with Gaussian noise and iteratively denoising them. Qualitative and quantitative evaluations demonstrate that HouseDiffusion significantly outperforms the current state-of-the-art with large margins.

欢迎各位指导讨论

Let ‘s Have a Brainstorm!