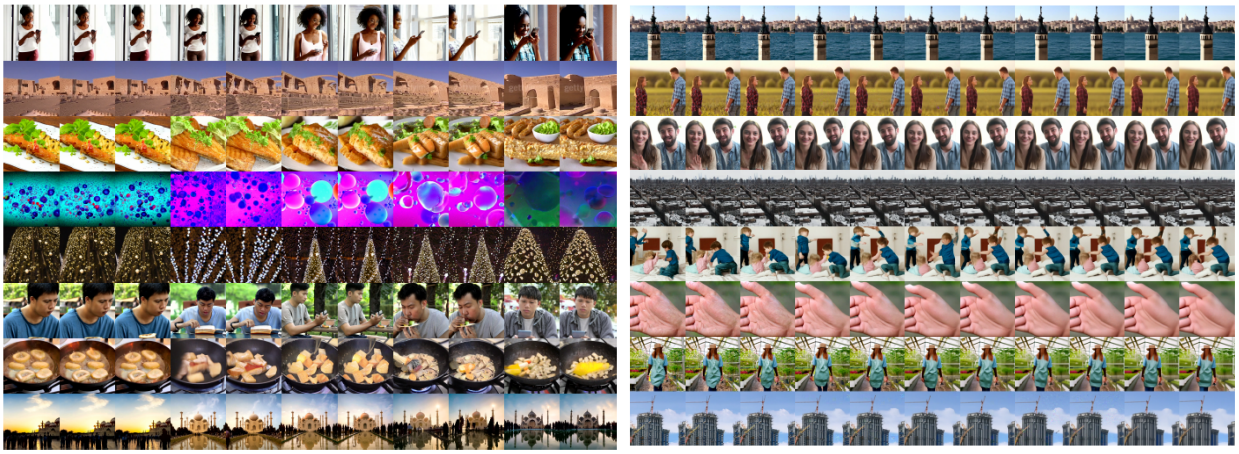
Video Diffusion Models (ICLR 2022)

1. Introduction and Background
   - Application: Diffusion model 은 최근 이미지 및 오디오 생성에 큰 두각을 보이고 있다. 하지만, video generation 부분에서는 문제점이 많았다. 이번 논문을 통해 diffusion model 의 video generation 부분에서의 가능성을 입증하고 발전 가능성을 제안한다.
   - Related Works: Name 2-3 directly related works and summarize their approaches.
     - Development and validation of the 3D U-Net algorithm for segmentation of pelvic lymph nodes on diffusion-weighted images : 3D U-Net 을 diffusion 에 처음을 도입
     - RePaint: Inpainting using Denoising Diffusion Probabilistic Models : 위 논문과 유사한 solution 제시 (Conventional replacement method -> Gradient method)
   - Problem:
     - diffusion model 의 경우 메모리의 한계로 인해 한번에 16 개의 frame 밖에 생성해내지 못한다. 따라서 추가적인 frame 을 생성하는 방법들이 연구되었다. 그 중 autoregressive, imputation 의 방법이 소개되었는데 그 방법으로 replacement method 가 사용되었다.
     - Replacement method 의 한계 : $E_q[x^b|z_t,x^a] = E_q[x^b|z_t] + (\sigma_t^2/\alpha_t)\nabla_{z^b_t} \log q(x^a|z_t)$ 에서 $\nabla_{z^b_t} \log q(x^a|z_t)$ 부분이 소실되는 문제
     - x^a 항 소실로 인해 x^b 의 연관성 부족
2. Innovation
   - Reconstruction-guided sampling(Reconstruction sampling) : $q(x^a|z_t) \approx N[\hat{x}^a_\theta(z_t),(\sigma_t^2/\alpha_t^2)I]$ 형식의 gaussian distribution 을 근사하여 사용
   - $$\tilde{\mathbf{x}}_\theta(\mathbf{z}_t) = \hat{\mathbf{x}}_\theta(\mathbf{z}_t) - \frac{w_r \alpha_t}{2} \nabla_{\mathbf{z}_t} \|\mathbf{x}^a - \hat{\mathbf{x}}^a_\theta(\mathbf{z}_t)\|_2^2$$
   - 이때 다음 식 중 The additional gradient term 을 guidance 의 형태로도 해석할 수 있다. 더 큰 가중치 계수(wr > 1)를 선택하면 샘플 품질이 향상되는 경향이 있다.
   - $$\nabla_{\mathbf{z}_\lambda} \log p^i(\mathbf{c}|\mathbf{z}_\lambda) = -\frac{1}{\sigma_\lambda}[\boldsymbol{\epsilon}^*(\mathbf{z}_\lambda, \mathbf{c}) - \boldsymbol{\epsilon}^*(\mathbf{z}_\lambda)]$$
3. Claim



Comparing the replacement method (left) vs the reconstruction guidance method (right)

- frame 간의 연관성을 향상시킴으로써 기존 16 frame 보다 더 긴 영상을 추출할 수 있게 되었고 diffusion model 의 video generation 가능성을 입증했다.

4. Experiment
   - Unconditional video modeling : We use the data loader provided by TensorFlow Datasets without further processing, and we train on all 13,320 videos. we use the C3D network [51]2 for calculating FID and IS, using 10,000 samples generated from our model.

Method	Resolution	FID↓	IS↑
MoCoGAN [52]	16x64x64	26998 ± 33	12.42
TGAN-F [26]	16x64x64	8942.63 ± 3.72	13.62
TGAN-ODE [18]	16x64x64	26512 ± 27	15.2
TGAN-F [26]	16x128x128	7817 ± 10	22.91 ± .19
VideoGPT [62]	16x128x128		24.69 ± 0.30
TGAN-v2 [41]	16x64x64	3431 ± 19	26.60 ± 0.47
TGAN-v2 [41]	16x128x128	3497 ± 26	28.87 ± 0.47
DVD-GAN [14]	16x128x128		32.97 ± 1.7
Video Diffusion (ours)	16x64x64	**295 ± 3**	**57 ± 0.62**
real data	16x64x64		60.2

Table 1: Unconditional video modeling results on UCF101.

Table 2: Video prediction on BAIR Robot Pushing.

Method	FVD↓
DVD-GAN [14]	109.8
VideoGPT [62]	103.3
TrIVD-GAN-FP [33]	103.3
Transframer [35]	100
CCVS [31]	99
VideoTransformer [59]	94
FitVid [4]	93.6
NUWA [61]	86.9
Video Diffusion (ours)	
ancestral sampler, 512 steps	68.19
Langevin sampler, 256 steps	**66.92**

Table 3: Video prediction on Kinetics-600.

Method	FVD↓	IS↑
Video Transformer [59]	170 ± 5	
DVD-GAN-FP [14]	69.1 ± 0.78	
Video VQ-VAE [57]	64.3 ± 2.04	
CCVS [31]	55 ± 1	
TrIVD-GAN-FP [33]	25.74 ± 0.66	12.54
Transframer [35]	25.4	
Video Diffusion (ours)		
ancestral, 256 steps	18.6	15.39
Langevin, 128 steps	**16.2 ± 0.34**	**15.64**

   - 다른 모델들에 비해 FID, IS, FVD(video prediction) 평가지표가 압도적으로 뛰어남을 확일 할 수 있다.
   - Classifier free guidance : weight 에 따른 평가지표 차이

Table 5: Effect of classifier-free guidance on text-to-video generation (large models). Sample quality is reported for 16x64x64 models trained on frameskip 1 and 4 data. The model was jointly trained on 8 independent image frames per 16-frame video.

Frameskip	Guidance weight	FVD↓	FID-avg↓	IS-avg↑	FID-first↓	IS-first↑
1	1.0	41.65/43.70	12.49/12.39	10.80/10.07	16.42/16.19	12.17/11.22
	2.0	50.19/48.79	10.53/10.47	13.22/12.10	13.91/13.75	14.81/13.46
	5.0	163.74/160.21	13.54/13.52	14.80/13.46	17.07/16.95	16.40/14.75
4	1.0	56.71/60.30	11.03/10.93	9.40/8.90	16.21/15.96	11.39/10.61
	2.0	54.28/51.95	9.39/9.36	11.53/10.75	14.21/14.04	13.81/12.63
	5.0	185.89/176.82	11.82/11.78	13.73/12.59	16.59/16.44	16.24/14.62

   -

5. Substantiation:
   - 이번 논문을 통해 diffusion model 보다 transformer, VAE 에 치중되어있던 video generation 에서 diffusion model 에서도 좋은 품질과, 긴 영상 시간을 얻을 수 있다는 가능성을 보여주었다.
   - 하지만 diffusion model 의 특성상 계산량이 압도적으로 많고 frame 을 생성하기에 상당히 오랜 시간이 걸린다는 단점은 해결하지 못했다.
   - 이후 latent space 를 활용함과 동시에 성능이 뛰어난 기존 image generation model 을 활용하는 등 diffusion model 에서 다양한 시도들이 이루어진다.

6. Rating:
   - 9 points
   - diffusion model 이 왜 video generation 에 약한지를 분석하고 이를 해결하기 위한 방안을 합리적으로 제시.
   - Background 에서 이론적인 부분을 자세히 설명
   - Replacement Method 의 한계와 이를 보안안 Reconstruction-guided Sampling 을 합리적으로 유도
   - 해결책이 기존에 classifier free guidance 와 연결됨을 보임
   - 연구 결과에서 평가지표 뿐만 아니라 실제 시각적으로도 뛰어난 성능임을 보임