VQ-VDM:
Video Diffusion Models with 3D VQGAN

The University of Electro-Communications, Tokyo, Japan

Ryota Kaji and Keiji Yanai

Background

- **Image generation area**
  - Stable Diffusion, Imagen, DALL-E2…
- **Video generation area**
  - Video Diffusion Models…
- **Difficulty in video generation**
  - Computationally demanding
  - Consistency of time sequence

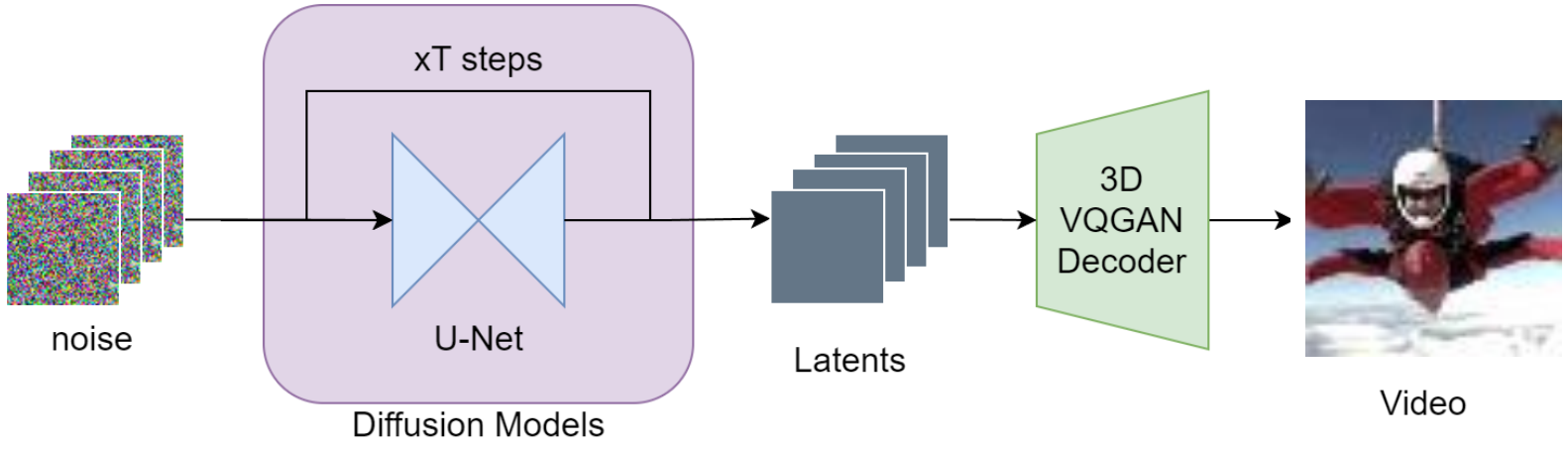
Sample examples of Stable Diffusion

Objective

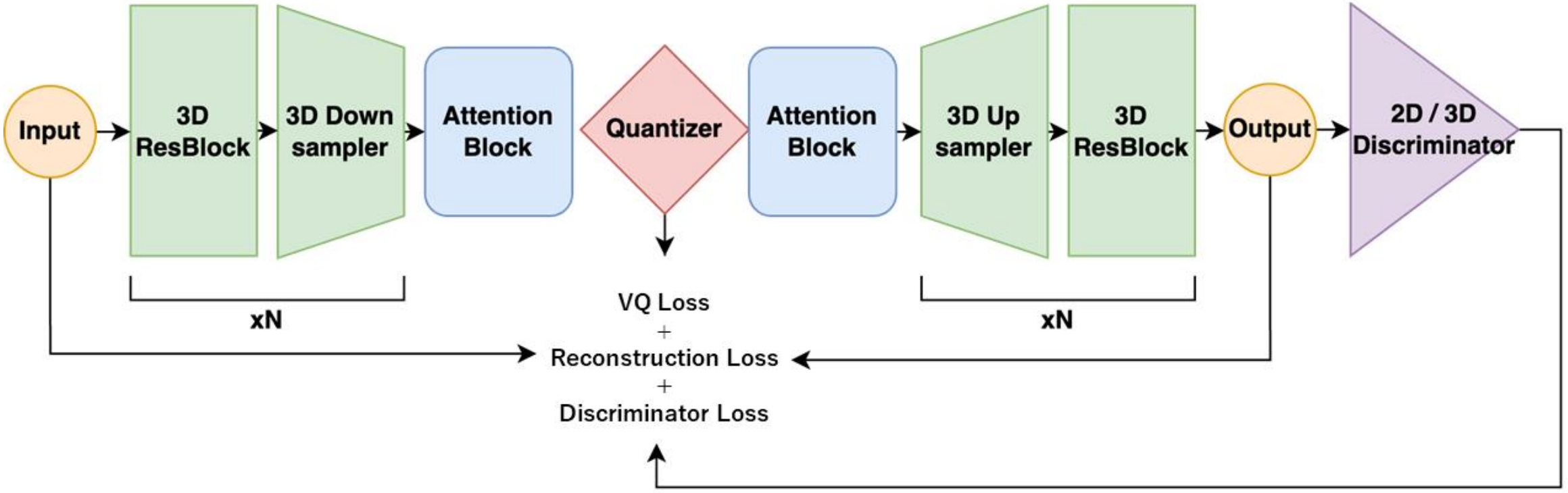
- **Performance of three typical video generation methods**
  - GAN:
    - ✓ Fast sampling, High quality generation
    - × Mode collapse
  - Autoregressively (AR):
    - ✓ High quality generation, training stability
    - × Autoregressive error
  - Diffusion Models (DM):
    - ✓ High quality generation, training stability
    - × Slow sampling
- **This study uses a 2-stage method with 3D VQGAN and Diffusion Models to reduce computational complexity.**

Method

- **Consisting of Stage 1 and Stage 2**
  - Stage1: Compressing video into latent representation using 3D VQGAN
  - Stage2: Training latent video representations with Diffusion Models
- **Combine both learned stages to generate a video**



Stage1：3D VQGAN for compressing videos



Stage1：Loss of 3D VQGAN

- **Reconstruction, VQ, Discriminator loss is used for training**
  - Reconstruction loss is MSE Loss and Perceptual Loss (VGG)

$$\mathcal{L}_{recon} = ||x - \hat{x}||^2 + P_{loss}(x, \hat{x})$$

  - VQ loss for Codebook and Encoder outputs closer together

$$\mathcal{L}_{vq} = ||sg[z_e(x)] - e||_2^2 + \beta||z_e(x) - sg[e]||_2^2$$

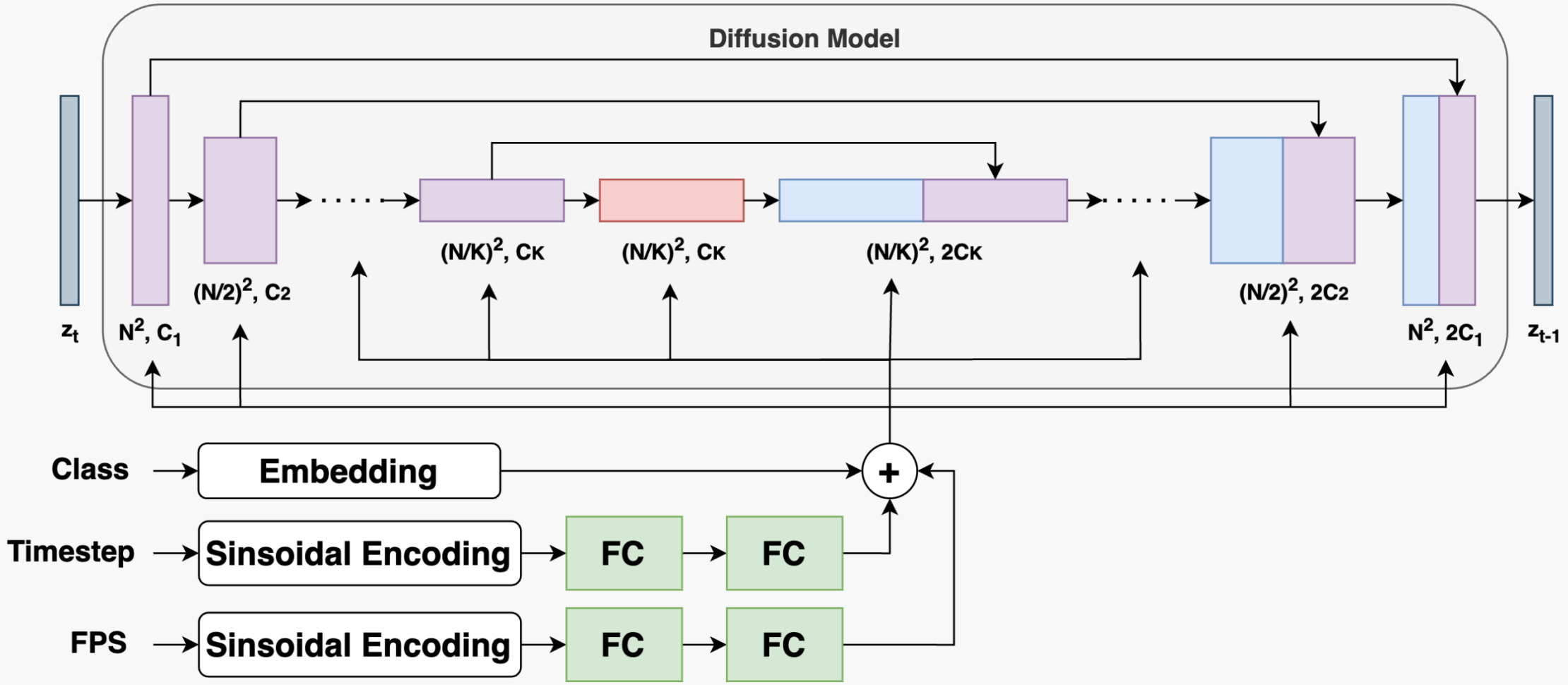
  - Discriminator Loss based on DCGAN

$$\mathcal{L}_{disc} = logD(x) + log(1 - D(\hat{x}))$$

  - Internal features of the Discriminator are also used as auxiliary loss

$$\mathcal{L}_{disc\_aux} = \sum_i ||D^{(i)}(x) - D^{(i)}(\hat{x})||^2$$

Stage2： Video Diffusion Models



Stage2： Video Diffusion Models

- Training by DDPM standard formulation

$$\mathcal{L}_{uncondition}(\theta) := \quad \mathbb{E}_{t,z_0,\epsilon}[||\epsilon - \epsilon_\theta(z_t, t, f)||^2]$$

$$\mathcal{L}_{class\_condition}(\theta) := \quad \mathbb{E}_{t,z_0,\epsilon}[||\epsilon - \epsilon_\theta(z_t, t, f, c)||^2]$$

- Generate video with Classifier-free guidance

$$\hat{\epsilon}_\theta(z_t, t, f, c) = w \cdot (\epsilon_\theta(z_t, t, f, c) - \epsilon_\theta(z_t, t, f)) + \epsilon_\theta(z_t, t, f)$$

Quantitative evaluation on UCF-101

Methods	Based	Resolution	FVD(↓)	KVD(↓)	Conference
VideoGPT	AR	128x128	24.69	-	arXiv 2021
DVD-GAN	GAN	128x128	27.38	-	arXiv 2019
TGANv2	GAN	128x128	28.87	1209	IJCV 2020
DIGAN	GAN	128x128	32.7	577	ICLR 2021
CogVideo*	AR	160x160	50.46	626	arXiv 2022
VDM	DM	64x64	57	-	NIPS 2022
TATS	AR	128x128	**79.28**	**332**	ECCV 2022
Ours	DM	128x128	64.13	425	-

* CogVideo uses pretrained text-to-image model and finetuning 5.4 million videos.

Quantitative evaluation on Sky Timelapse

- **Achieved state-of-the-art scores in FVD evaluation**

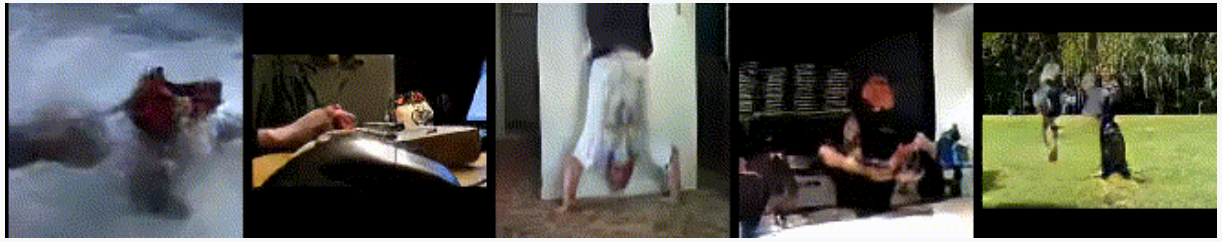
Method	Based	Resolution	FVD(↓)	KVD(↓)	Conference
MoCoGAN-HD	GAN	128x128	183.6	13.9	ICLR 2021
DIGAN	GAN	128x128	<u>114.6</u>	6.8	ICLR 2022
TATS	AR	128x128	132.6	**5.7**	ECCV 2022
Ours	DM	128x128	**109.4**	<u>5.9</u>	-

Qualitative evaluation on UCF-101

- Comparable to TATS in SOTA methods

DIGAN

TATS



Ours

Comparison of generation speed

- Speed comparison with VDM

Method	Resolution	100 step time [s]
VDM	16x64x64	35.26±2.43
Ours	16x128x128	**3.95±0.01**

- VQ-VDM generated videos that are 4 times larger in the spatial direction.
- However, the proposed method worked about 9 times faster.

Conclusion

- We introduced video diffusion models with 3D VQGAN.
- Our method outperformed all but the SOTA methods in class conditional generation.
- Our method achieved state-of-the-art FVD in unconditional generation on Sky-Timelapse dataset.
- Our method generated high-resolution video about 9 times faster than conventional VDM.