Video Probabilistic Diffusion Models in Projected Latent Space

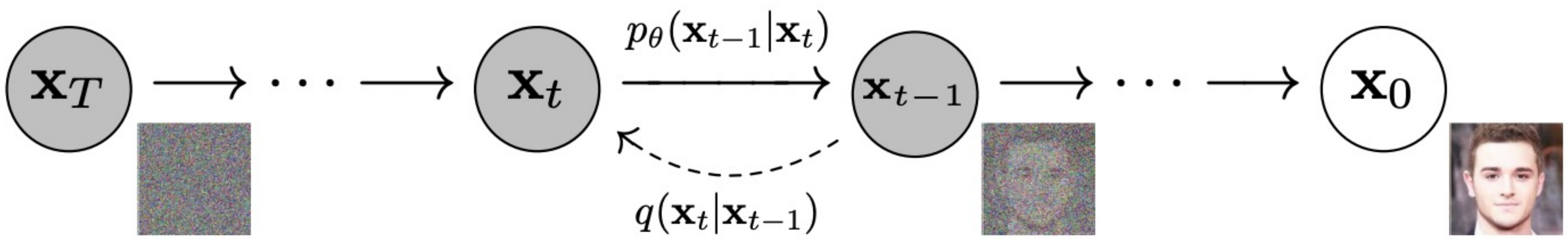
Sihyun Yu, Kihyuk Sohn, Subin Kim, Jinwoo Shin

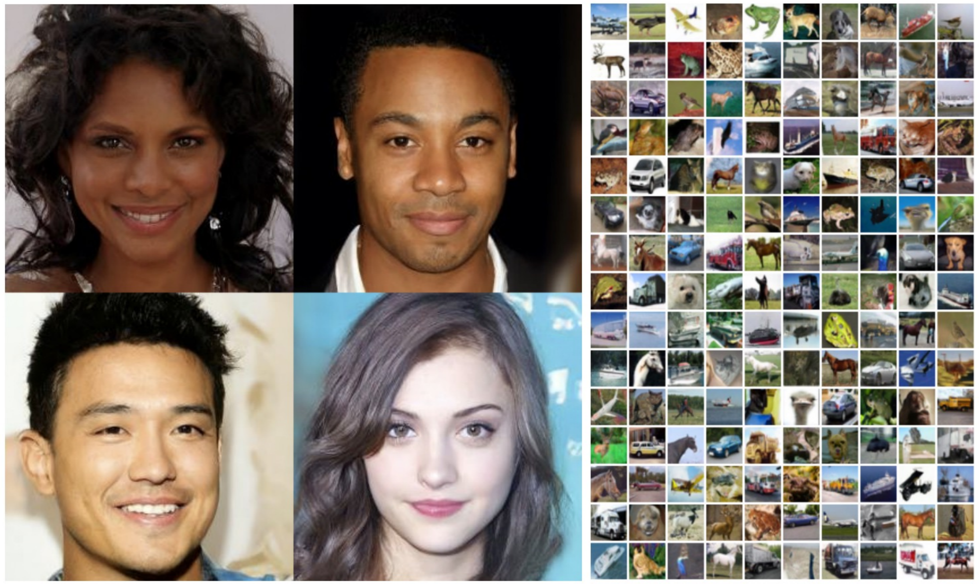
Presenter: Siyi Chen

10/03/2023

[1] Video Probabilistic Diffusion Models in Projected Latent Space. Sihyun Yu, Kihyuk Sohn, Subin Kim, Jinwoo Shin. 2023.

Diffusion Models – Image Generation

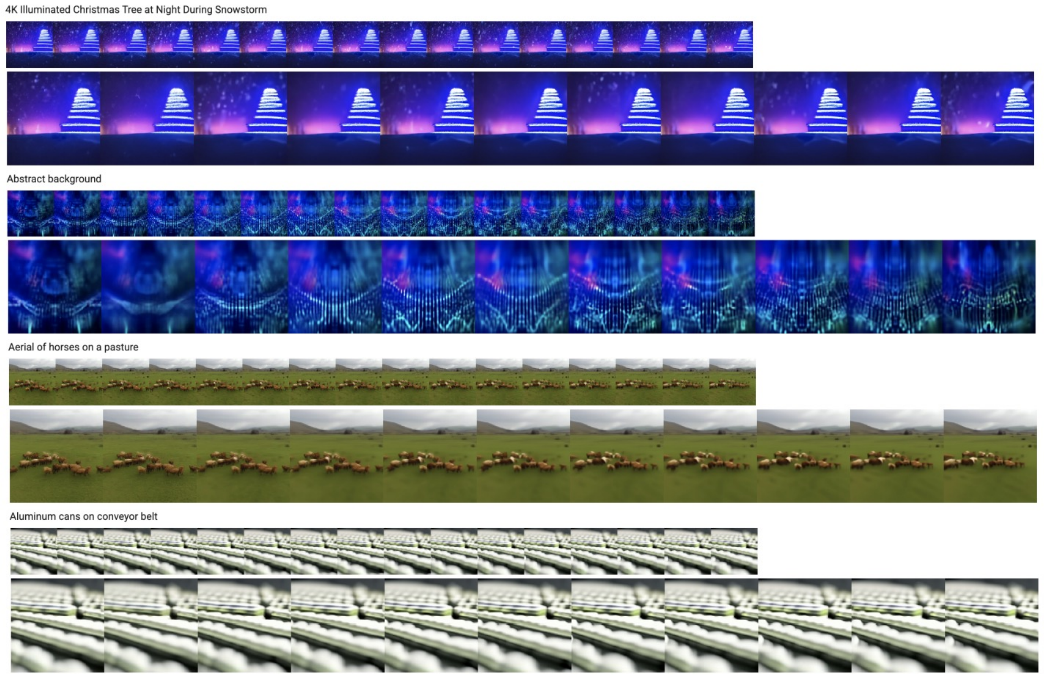




[2] Denoising Diffusion Probabilistic Models. Jonathan Ho, Ajay Jain, Pieter Abbee. 2020.

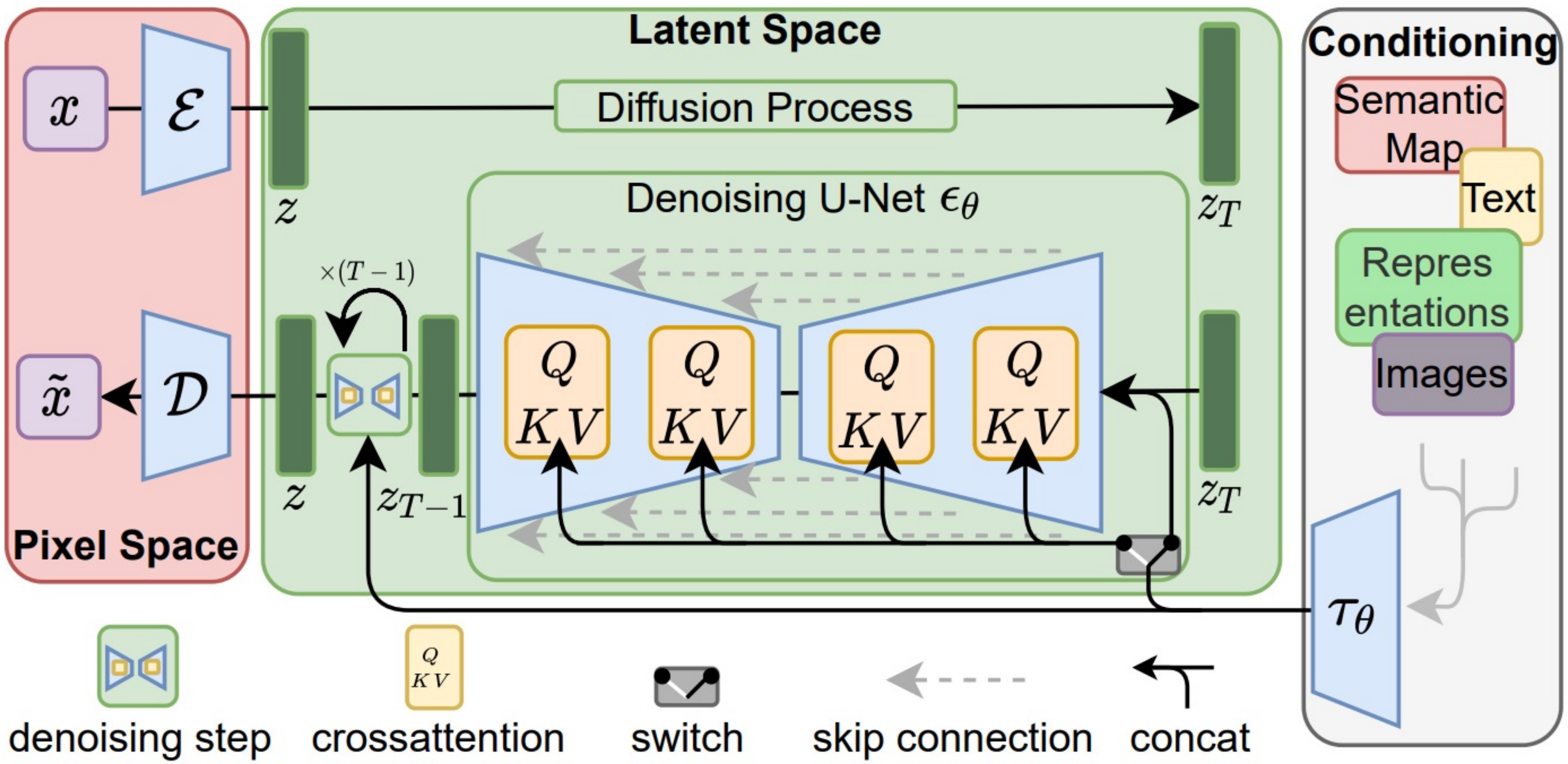
Video Diffusion Models - Video Generation

- Previous works extended from images in frame-space
  - Suffer from computation & memory inefficiency
    - Solved by latent diffusion + special autoencoder
  - Not flexible enough to support high-quality long video generation
    - Solved by special diffusion model design



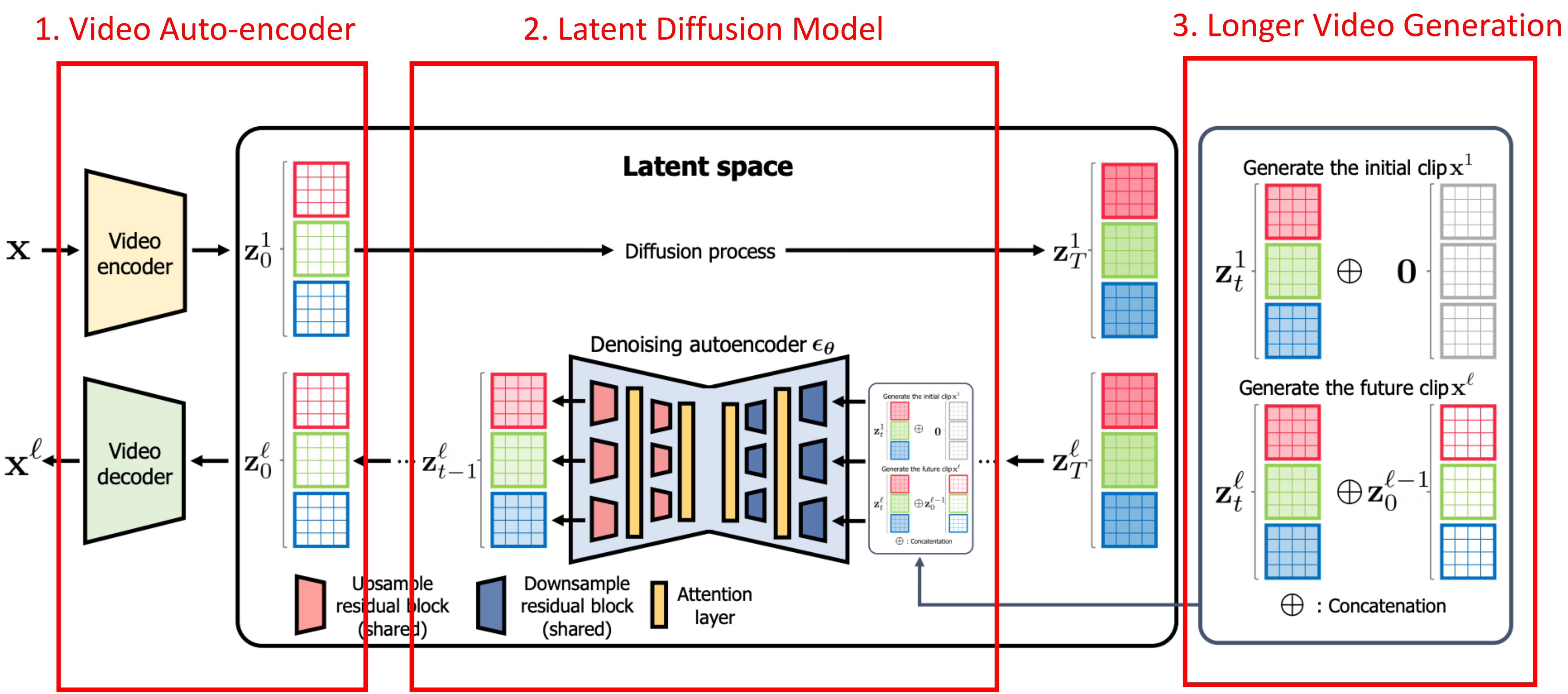
[3] Video Diffusion Models. Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, David J. Fleet. 2022.

Latent Diffusion Models



[4] High-Resolution Image Synthesis with Latent Diffusion Models. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Bjorn Ommer. 2022.

Projected Latent Video Diffusion Model



[1] Video Probabilistic Diffusion Models in Projected Latent Space. Sihyun Yu, Kihyuk Sohn, Subin Kim, Jinwoo Shin. 2023.

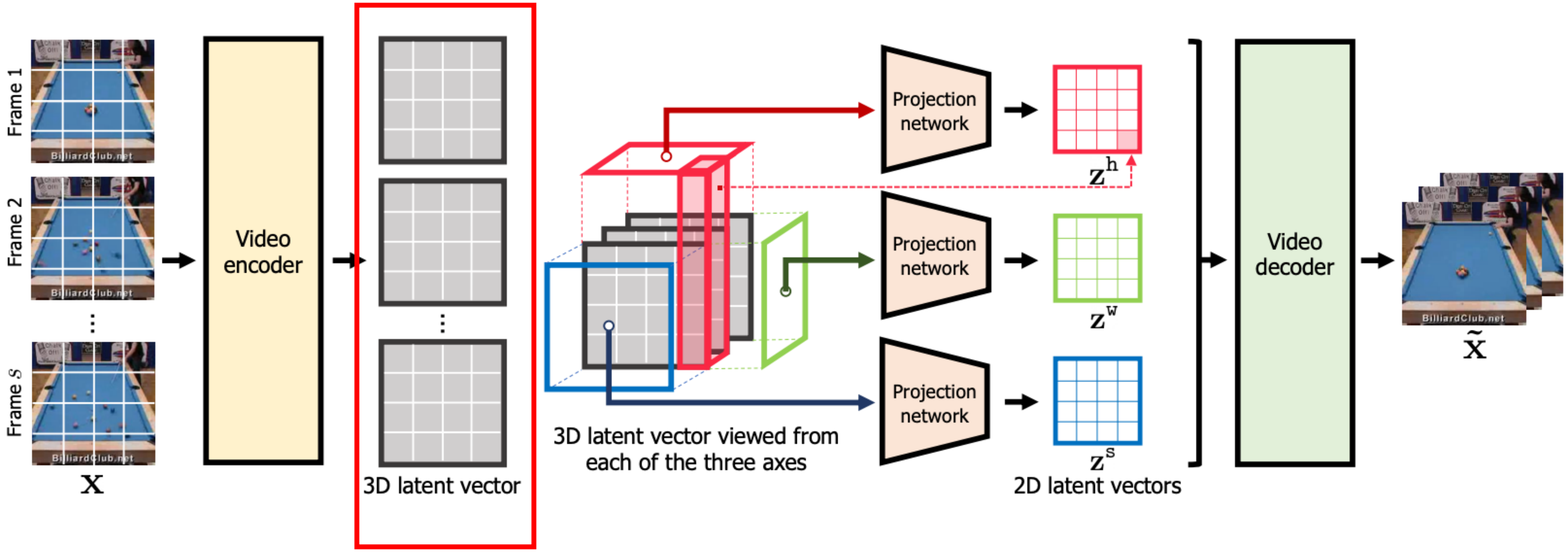
Video Auto-encoder

$$\boxed{\mathbf{u} := f^{\mathtt{shw}}_{\phi_{\mathtt{shw}}}(\mathbf{x}), \quad \text{where } \mathbf{u} = [u_{shw}] \in \mathbb{R}^{C\times S\times H'\times W'},}$$

$$z^{\mathtt{s}}_{hw} := f^{\mathtt{s}}_{\phi_{\mathtt{s}}}(u_{1hw}, \ldots, u_{Shw}),\ 1 \le h \le H',\ 1 \le w \le W',$$

$$z^{\mathtt{h}}_{sw} := f^{\mathtt{h}}_{\phi_{\mathtt{h}}}(u_{s1w}, \ldots, u_{sH'w}),\ 1 \le s \le S,\ 1 \le w \le W',$$

$$z^{\mathtt{w}}_{sh} := f^{\mathtt{w}}_{\phi_{\mathtt{w}}}(u_{sh1}, \ldots, u_{shW'}),\ 1 \le s \le S,\ 1 \le h \le H'.$$



[1] Video Probabilistic Diffusion Models in Projected Latent Space. Sihyun Yu, Kihyuk Sohn, Subin Kim, Jinwoo Shin. 2023.

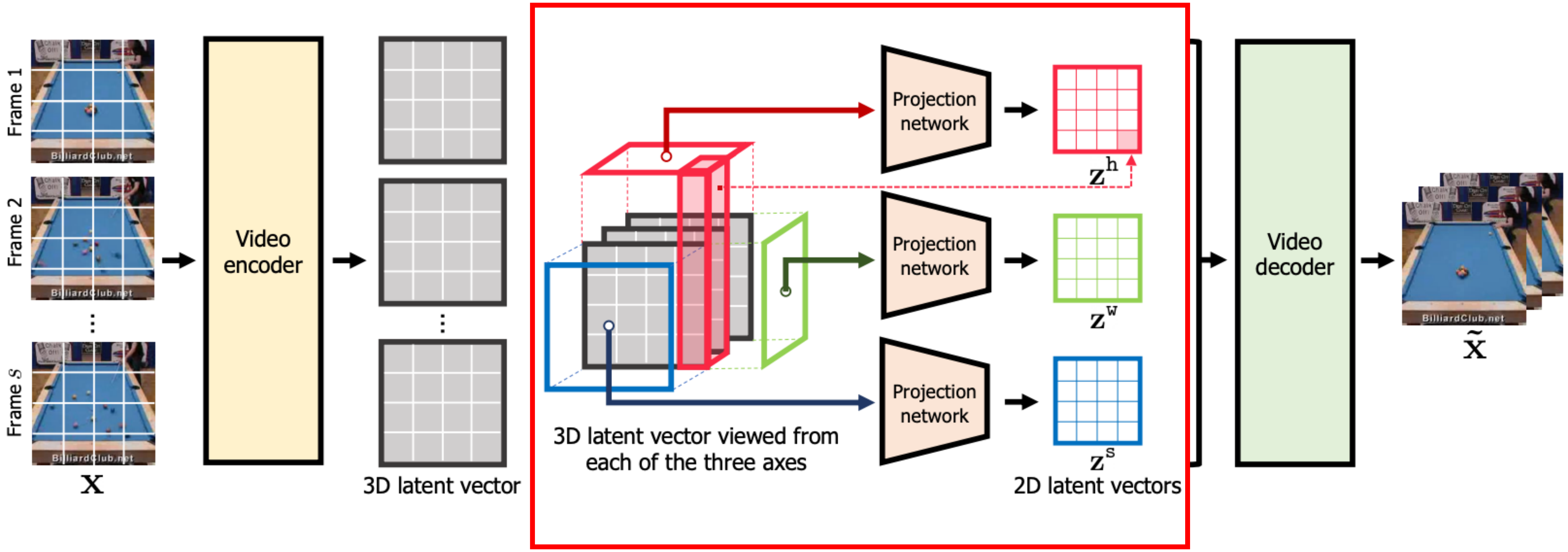
Video Auto-encoder

$$\mathbf{u} := f^{\mathtt{shw}}_{\phi_{\mathtt{shw}}}(\mathbf{x}), \quad \text{where } \mathbf{u} = [u_{shw}] \in \mathbb{R}^{C\times S\times H'\times W'},$$

$$z^{\mathtt{s}}_{hw} := f^{\mathtt{s}}_{\phi_{\mathtt{s}}}(u_{1hw},\ldots,u_{Shw}),\ 1 \le h \le H',\ 1 \le w \le W',$$

$$z^{\mathtt{h}}_{sw} := f^{\mathtt{h}}_{\phi_{\mathtt{h}}}(u_{s1w},\ldots,u_{sH'w}),\ 1 \le s \le S,\ 1 \le w \le W',$$

$$z^{\mathtt{w}}_{sh} := f^{\mathtt{w}}_{\phi_{\mathtt{w}}}(u_{sh1},\ldots,u_{shW'}),\ 1 \le s \le S,\ 1 \le h \le H'.$$



[1] Video Probabilistic Diffusion Models in Projected Latent Space. Sihyun Yu, Kihyuk Sohn, Subin Kim, Jinwoo Shin. 2023.

Video Auto-encoder

$$\mathbf{u} := f^{\mathtt{shw}}_{\phi_{\mathtt{shw}}}(\mathbf{x}), \quad \text{where } \mathbf{u} = [u_{shw}] \in \mathbb{R}^{C\times S\times H'\times W'},$$

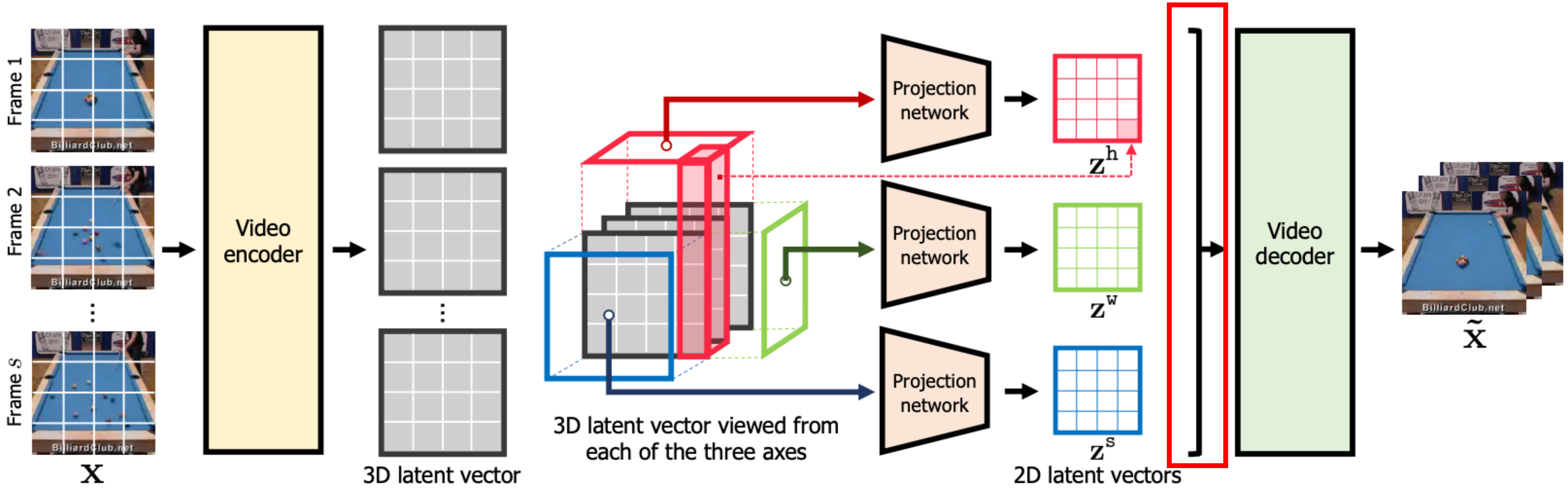
$$z^{\mathtt{s}}_{hw} := f^{\mathtt{s}}_{\phi_{\mathtt{s}}}(u_{1hw},\dots,u_{Shw}),\ 1\le h\le H',\ 1\le w\le W',$$

$$z^{\mathtt{h}}_{sw} := f^{\mathtt{h}}_{\phi_{\mathtt{h}}}(u_{s1w},\dots,u_{sH'w}),\ 1\le s\le S,\ 1\le w\le W',$$

$$z^{\mathtt{w}}_{sh} := f^{\mathtt{w}}_{\phi_{\mathtt{w}}}(u_{sh1},\dots,u_{shW'}),\ 1\le s\le S,\ 1\le h\le H'.$$

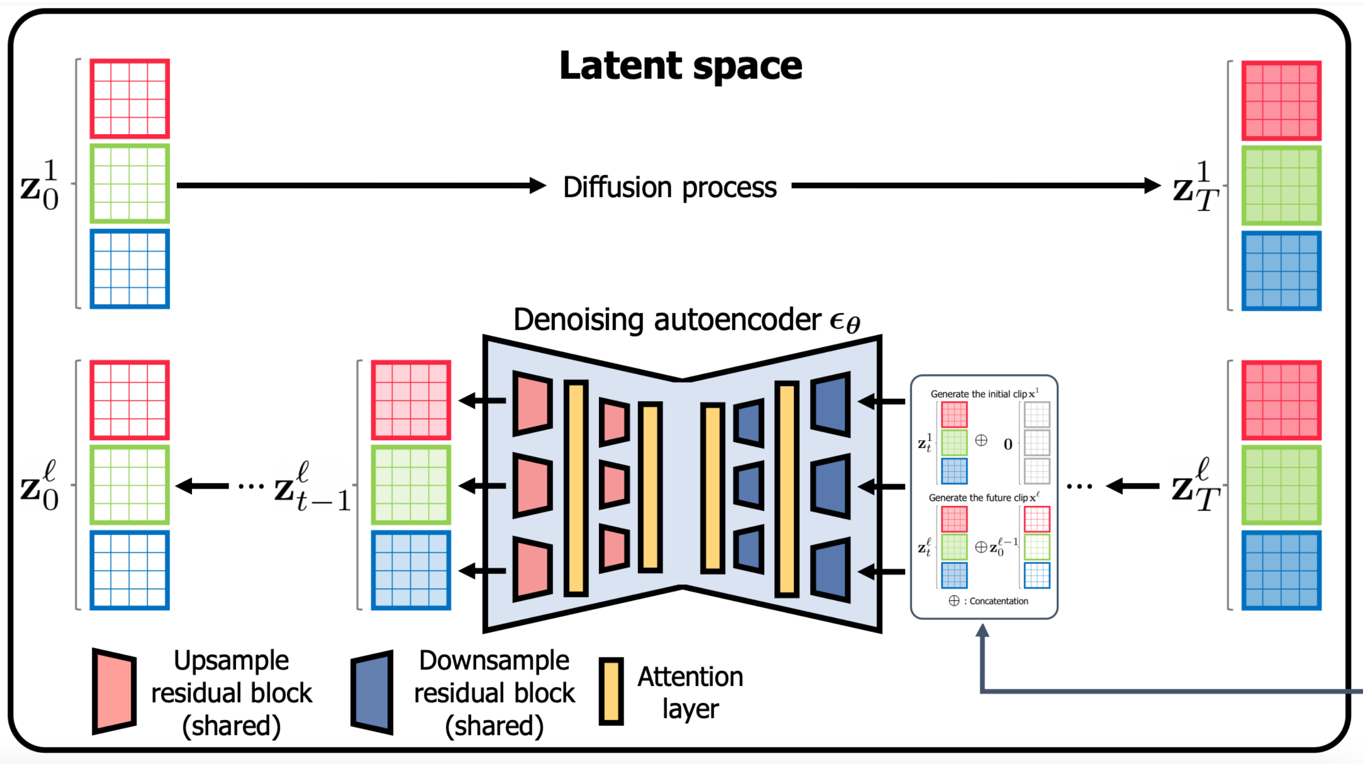
$$\mathbf{v} = (v_{shw}) \in \mathbb{R}^{3C\times S\times H'\times W'}$$

$$v_{shw} := [z_{hw}, z_{sw}, z_{sh}].$$



[1] Video Probabilistic Diffusion Models in Projected Latent Space. Sihyun Yu, Kihyuk Sohn, Subin Kim, Jinwoo Shin. 2023.

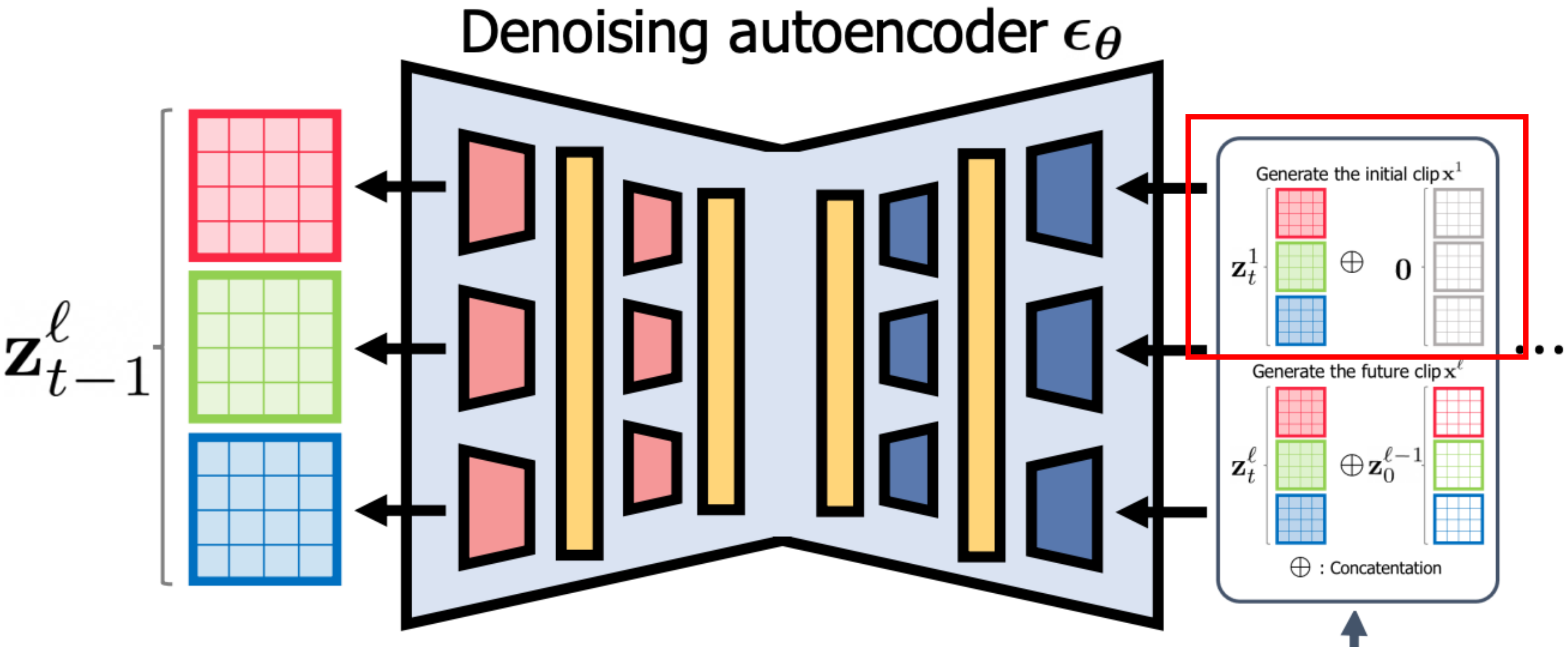
Latent Diffusion Model



- U-Net: Share parameters
- Attention layer

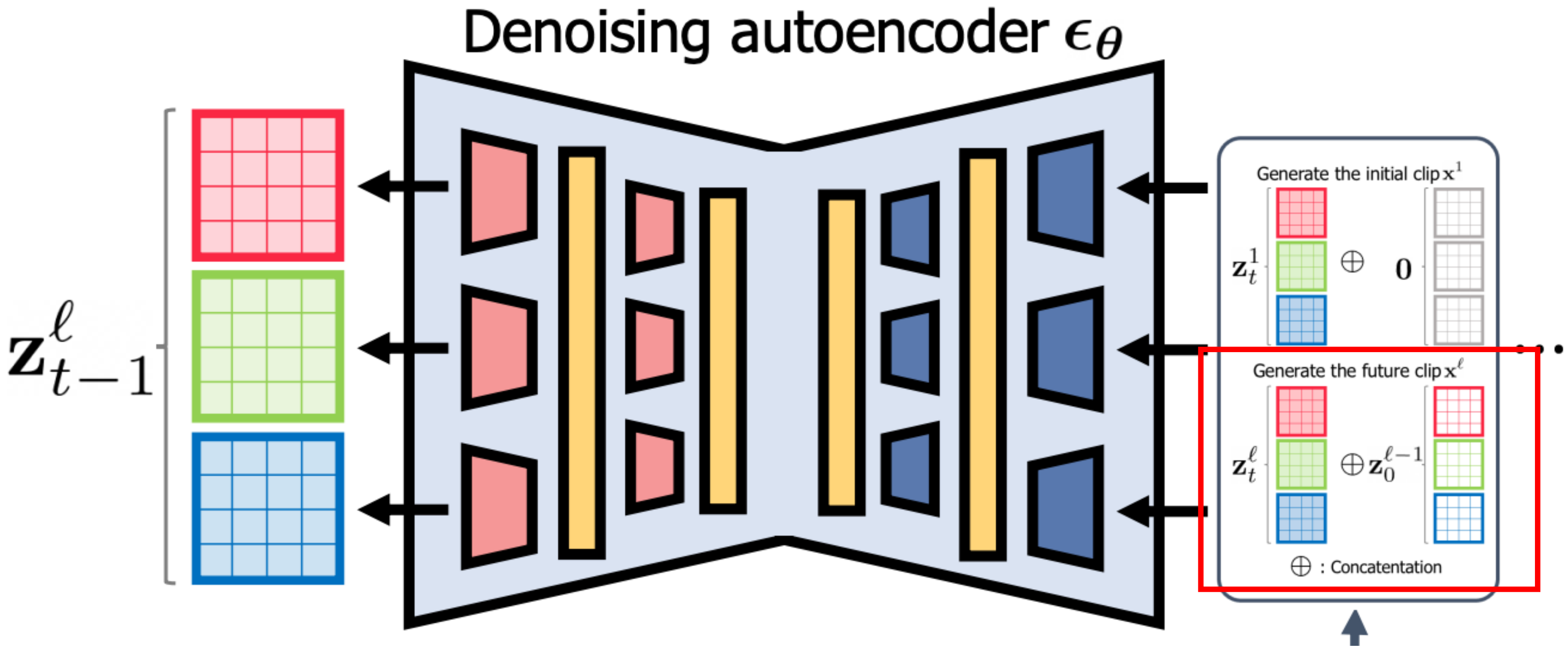
[1] Video Probabilistic Diffusion Models in Projected Latent Space. Sihyun Yu, Kihyuk Sohn, Subin Kim, Jinwoo Shin. 2023.

Longer Video Generation



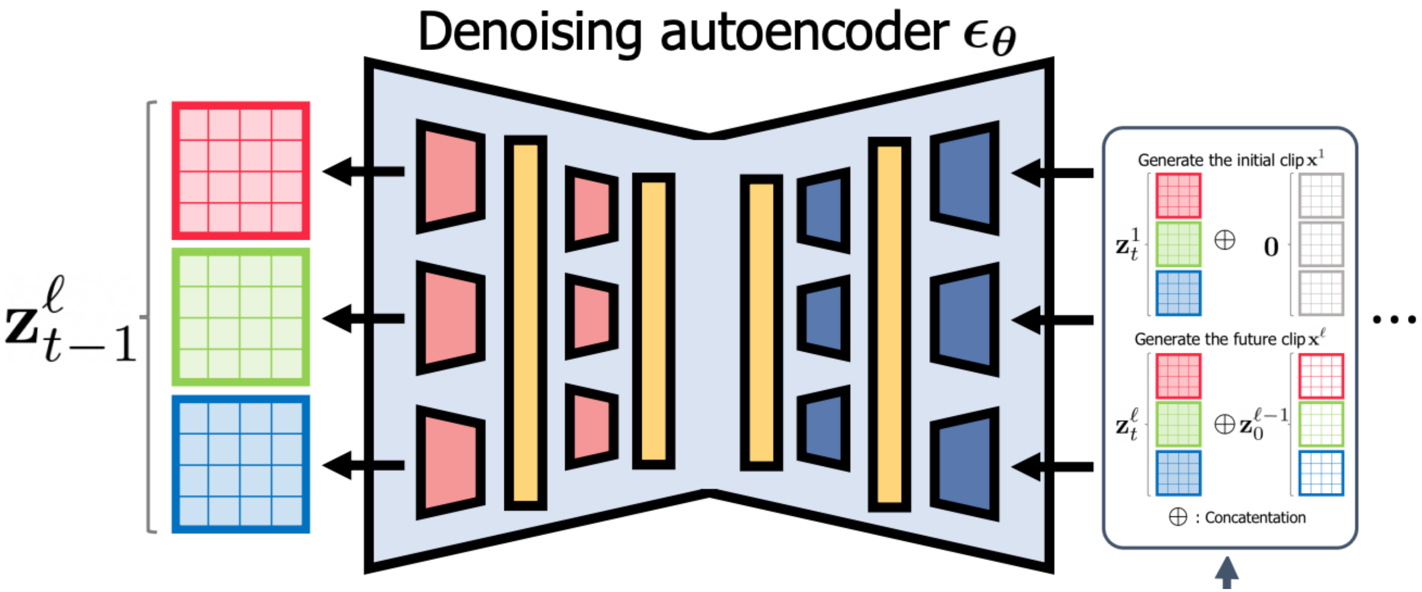
[1] Video Probabilistic Diffusion Models in Projected Latent Space. Sihyun Yu, Kihyuk Sohn, Subin Kim, Jinwoo Shin. 2023.

Longer Video Generation



[1] Video Probabilistic Diffusion Models in Projected Latent Space. Sihyun Yu, Kihyuk Sohn, Subin Kim, Jinwoo Shin. 2023.

Longer Video Generation



$$\mathbb{E}_{(\mathbf{x}_0^1,\mathbf{x}_0^2),\boldsymbol{\epsilon},t}\Big[\lambda||\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}_t^2,\mathbf{z}_0^1,t)||_2^2 + (1-\lambda)||\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}_t^2,\mathbf{0},t)||_2^2\Big]$$

[1] Video Probabilistic Diffusion Models in Projected Latent Space. Sihyun Yu, Kihyuk Sohn, Subin Kim, Jinwoo Shin. 2023.

Quiz

- What is the missing variable in the algorithm?
- A. z_{t+1}^{l}
- B. z_{t}^{l-1}
- C. z_{0}^{l-1}
- D. z_{0}^{l}

Algorithm 1 projected latent video diffusion model (PVDM)

---

for $\ell = 1$ to L **do** $\triangleright$ *Iteratively generate video clips* $\mathbf{x}^{\ell}$.

  Sample the random noise $\mathbf{z}_T^{\ell} \sim p(\mathbf{z}_T)$.

  for $t = T$ to 1 **do**

    if $\ell = 1$ **then**

      Unconditional score $\boldsymbol{\epsilon}_t = \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}_t^{\ell}, \mathbf{0}, t)$.

    else

      Conditional score $\boldsymbol{\epsilon}_t = \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}_t^{\ell}, \boxed{}, t)$.

    end if

    Sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}_{\mathbf{z}}, \mathbf{I}_{\mathbf{z}})$.

    Compute $\mathbf{z}_{t-1}^{\ell} = \frac{1}{\sqrt{1-\beta_t}}\left(\mathbf{z}_t^{\ell} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_t\right) + \sigma_t \boldsymbol{\epsilon}$.

  end for

  Decode the ℓ-th clip $\mathbf{x}^{\ell} = g_{\boldsymbol{\psi}}(\mathbf{z}_0^{\ell})$.

end for

Output the generated video $[\mathbf{x}^1, \ldots, \mathbf{x}^L]$.

---

[1] Video Probabilistic Diffusion Models in Projected Latent Space. Sihyun Yu, Kihyuk Sohn, Subin Kim, Jinwoo Shin. 2023.

Quiz

- What is the missing variable in the algorithm?
- A. z_{t+1}^{l}
- B. z_{t}^{l-1}
- C. z_{0}^{l-1}
- D. z_{0}^{l}

Algorithm 1 projected latent video diffusion model (PVDM)

for $\ell = 1$ to L **do** $\triangleright$ *Iteratively generate video clips* $\mathbf{x}^{\ell}$.
  Sample the random noise $\mathbf{z}_T^{\ell} \sim p(\mathbf{z}_T)$.
  for $t = T$ to 1 **do**
    if $\ell = 1$ **then**
      Unconditional score $\boldsymbol{\epsilon}_t = \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}_t^{\ell}, \mathbf{0}, t)$.
    else
      Conditional score $\boldsymbol{\epsilon}_t = \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}_t^{\ell}, \mathbf{z}_0^{\ell-1}, t)$.
    end if
    Sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}_{\mathbf{z}}, \mathbf{I}_{\mathbf{z}})$.
    Compute $\mathbf{z}_{t-1}^{\ell} = \frac{1}{\sqrt{1-\beta_t}}\Big(\mathbf{z}_t^{\ell} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_t\Big) + \sigma_t \boldsymbol{\epsilon}$.
  end for
  Decode the ℓ-th clip $\mathbf{x}^{\ell} = g_{\boldsymbol{\psi}}(\mathbf{z}_0^{\ell})$.
end for
Output the generated video $[\mathbf{x}^1, \ldots, \mathbf{x}^L]$.

[1] Video Probabilistic Diffusion Models in Projected Latent Space. Sihyun Yu, Kihyuk Sohn, Subin Kim, Jinwoo Shin. 2023.

Experiments

- Dataset: UCF101 & SkyTimelapse
- Model:
  - Autoencoder: TimeSformer
  - 3D-to-2D projection mapping: Transformer architecture
- Baselines:
  - GPT/GAN/Diffusion based SOTA models
- Metrics:
  - Inception score (IS)
  - Fréchet video distance (FVD)

Results

Table 1. FVD_{16} and FVD_{128} values (lower values are better) of video generation models on UCF-101 and SkyTimelapse. Bolds indicate the best results, and we mark our method as blue. We report FVD values of other baselines obtained by the reference (StyleGAN-V [47]). N/M-s denotes the model is evaluated with the DDIM sampler [51] with N steps (for the initial clip) and M steps (for future clips).

	UCF-101		SkyTimelapse	
Method	$\text{FVD}_{16}\downarrow$	$\text{FVD}_{128}\downarrow$	$\text{FVD}_{16}\downarrow$	$\text{FVD}_{128}\downarrow$
VideoGPT [65]	2880.6	N/A	222.7	N/A
MoCoGAN [57]	2886.8	3679.0	206.6	575.9
+ StyleGAN2 [28]	1821.4	2311.3	85.88	272.8
MoCoGAN-HD [55]	1729.6	2606.5	164.1	878.1
DIGAN [67]	1630.2	2293.7	83.11	196.7
StyleGAN-V [47]	1431.0	1773.4	79.52	197.0
PVDM-S (ours); 100/20-s	457.4	902.2	71.46	159.9
PVDM-L (ours); 200/200-s	398.9	**639.7**	61.70	137.2
PVDM-L (ours); 400/400-s	**343.6**	648.4	**55.41**	**125.2**

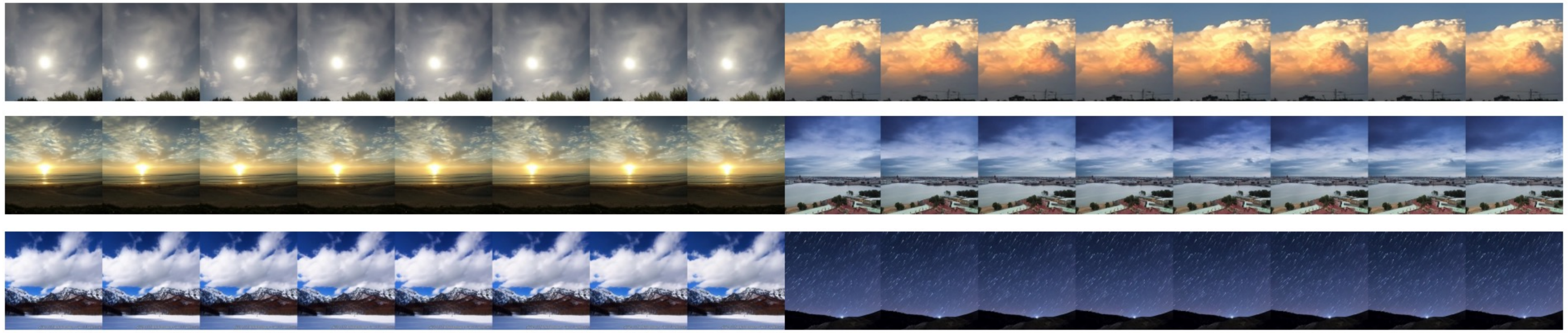
Table 2. IS values (higher values are better) of video generation models on UCF-101. Bolds indicate the best results and subscripts denote the standard deviations. * denotes the model is trained on train+test split, otherwise the method uses only the train split for training.

Method	IS $\uparrow$
MoCoGAN [57]	$12.42_{\pm 0.07}$
ProgressiveVGAN [1]	$14.56_{\pm 0.05}$
LDVD-GAN [23]	$22.91_{\pm 0.19}$
VideoGPT [65]	$24.69_{\pm 0.30}$
TGANv2 [43]	$28.87_{\pm 0.67}$
StyleGAN-V* [47]	$23.94_{\pm 0.73}$
DIGAN [67]	$29.71_{\pm 0.53}$
VDM* [21]	$57.00_{\pm 0.62}$
TATS [12]	$57.63_{\pm 0.24}$
PVDM-L (ours)	$\mathbf{74.40}_{\pm 1.25}$

[1] Video Probabilistic Diffusion Models in Projected Latent Space. Sihyun Yu, Kihyuk Sohn, Subin Kim, Jinwoo Shin. 2023.

Short Videos

(a) UCF-101



(b) SkyTimelapse

Figure 4. Illustrations of random 16 frames, 256×256 resolution video synthesis results of PVDM trained on UCF-101 and SkyTimelapse datasets. We visualize the frames of each video with stride 2.

[1] Video Probabilistic Diffusion Models in Projected Latent Space. Sihyun Yu, Kihyuk Sohn, Subin Kim, Jinwoo Shin. 2023.

Longer Videos

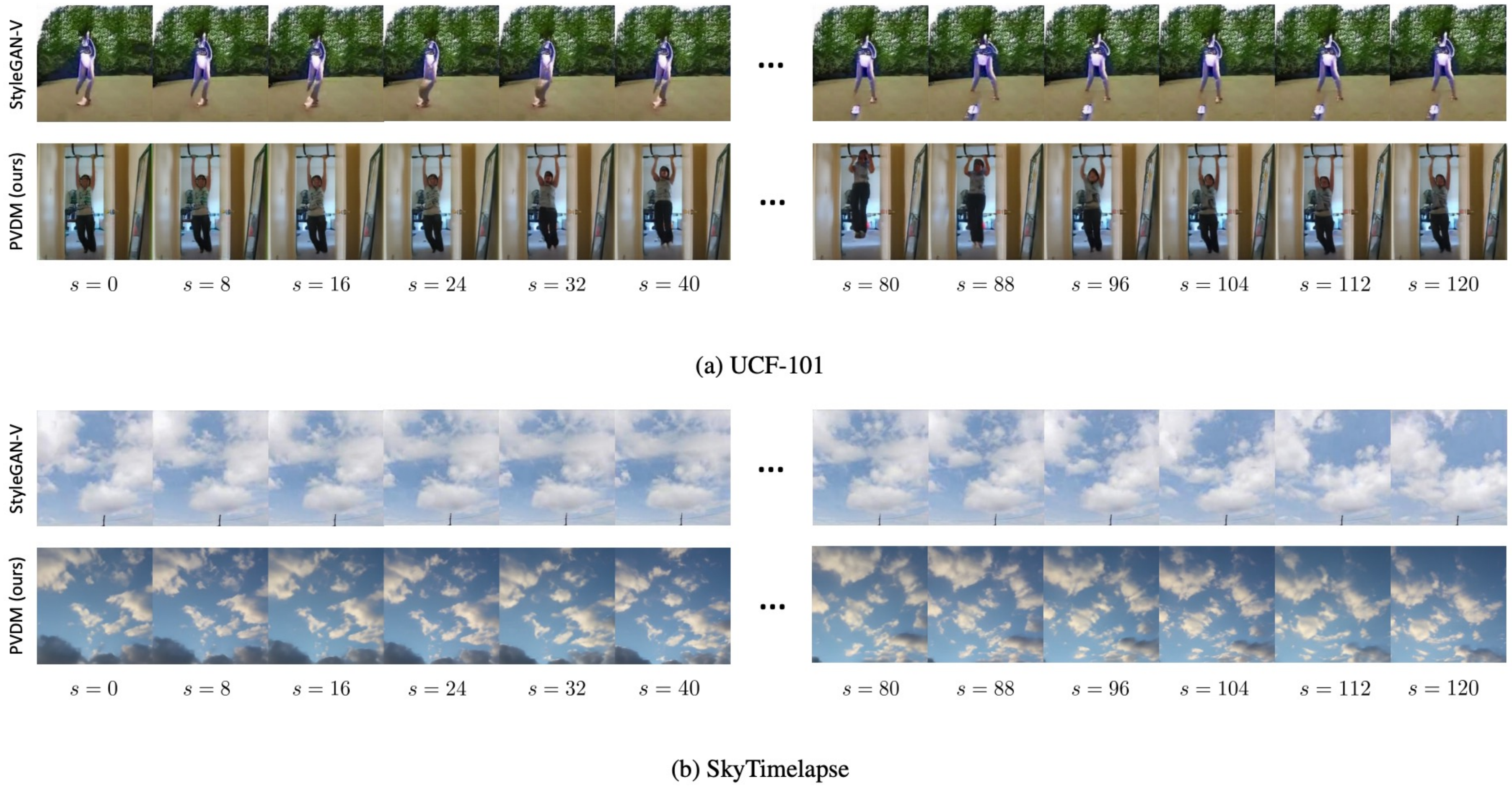


Figure 3. 256×256 resolution, 128 frame video synthesis results of StyleGAN-V and PVDM, trained on (a) UCF-101 and (b) SkyTimelapse.[1]

[1] Video Probabilistic Diffusion Models in Projected Latent Space. Sihyun Yu, Kihyuk Sohn, Subin Kim, Jinwoo Shin. 2023.

Time & Memory Efficiency

Table 5. Maximum batch size for training and time (s), memory (GB) for synthesizing a 256×256 resolution video measured with a single NVIDIA 3090Ti 24GB GPU. N/A denotes the values cannot be measured due to the out-of-memory problem. N/M-s denotes the model is evaluated with the DDIM sampler [51] with N steps (for the initial clip) and M steps (for future clips).

	Train	Inference (time/memory)	
Length →	16	16	128
TATS [12]	0	84.8/18.7	434/19.2
VideoGPT [65]	0	139/15.2	N/A
VDM [21]; 100/20-s	0	113/11.1	N/A
PVDM-L (ours); 200/200-s	2	20.4/5.22	166/5.22
PVDM-L (ours); 400/400-s	2	40.9/5.22	328/5.22
PVDM-S (ours); 100/20-s	**7**	**7.88/4.33**	**31.3/4.33**

[1] Video Probabilistic Diffusion Models in Projected Latent Space. Sihyun Yu, Kihyuk Sohn, Subin Kim, Jinwoo Shin. 2023.

Conclusion & Discussion

- Latent diffusion model for video generation
- Image-like 2D latent space
- Longer video generation

- Future Direction
  - Text-to-video latent diffusion models

Reference

[1] Video Probabilistic Diffusion Models in Projected Latent Space. Sihyun Yu, Kihyuk Sohn, Subin Kim, Jinwoo Shin. 2023.

[2] Denoising Diffusion Probabilistic Models. Jonathan Ho, Ajay Jain, Pieter Abbee. 2020.

[3] Video Diffusion Models. Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, David J. Fleet. 2020.

[4] High-Resolution Image Synthesis with Latent Diffusion Models. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Bjorn Ommer. 2022.

Thank you!