ToonCrafter: Generative Cartoon Interpolation

2024.10.15

Introduction

- 최근 live-action video frame interpolation methods들도 어느 정도 수준이 올라왔지만 여전히 Cartoon animation의 결과는 좋지 못 함. 이는 **sparsity** 와 **texture richness** 때문이라고 볼 수 있음
    - **Sparsity** : cartoon은 frame과 frame 사이를 사람이 직접 그리기 때문에, 움직임의 연결이 크고 따라서 두 프레임 사이의 움직임을 표현하는 정보가 부족함
    - **Texture richness** : 비슷한 이유로 cartoon은 텍스처가 없는 영역이 많을 확률이 높음

Video Frame Interpolation

- 이전의 연구들은 주로 움직임이 단순하다고 가정하는 **linear interpolation**에 중점
  - **phase-based methods** (위상 기반) : 이미지의 픽셀 정보를 주파수 영역에서 분석하여 움직임을 보간하는 방식
  - **kernel-based methods** (커널 기반) : 각 픽셀 주변의 정보를 바탕으로 커널(즉, 필터)을 사용해 보간을 수행하는 방식
  - **optical/feature flow-based methods** : 가장 많이 사용되는 방법으로 두 frames 간의 correspondence를 flow를 사용해서 식별한 후, warping과 fusion을 수행

dis-occlusion

complex non-linear motions

ToonCrafter!

- large dataset 의 사용으로 **generative cartoon interpolation**도 많은 연구들과 발전이 진행되고 있는데, **여전히 좋은 성능은 아니며 아래 세 가지를 그 이유**로 볼 수 있음

Issues

1. Domain gap
2. Highly compressed latent spaces
3. Lack of control

Suggestions

1. Toon rectification learning
2. Dual reference-based 3D decoder
3. Sketch encoder

Method 1 : Toon rectification learning

- **DynamiCrafter** interpolation model을 **fine tuning** 진행
- 이때 cartoon data를 바로 fine tuning 하게 되면, **catastrophic forgetting**으로 인해 기존 정보들을 많이 잊게 되는 문제가 존재하여 **Image-Context projector, Spatial Layers, Temporal layers**사용

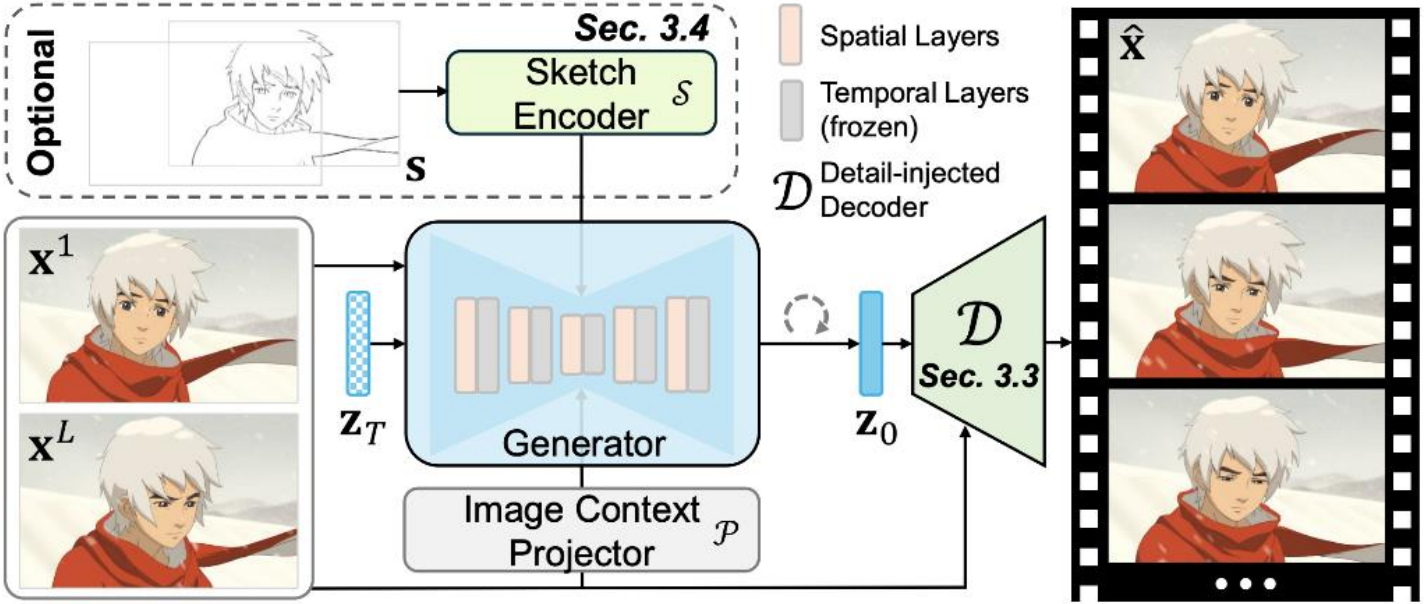
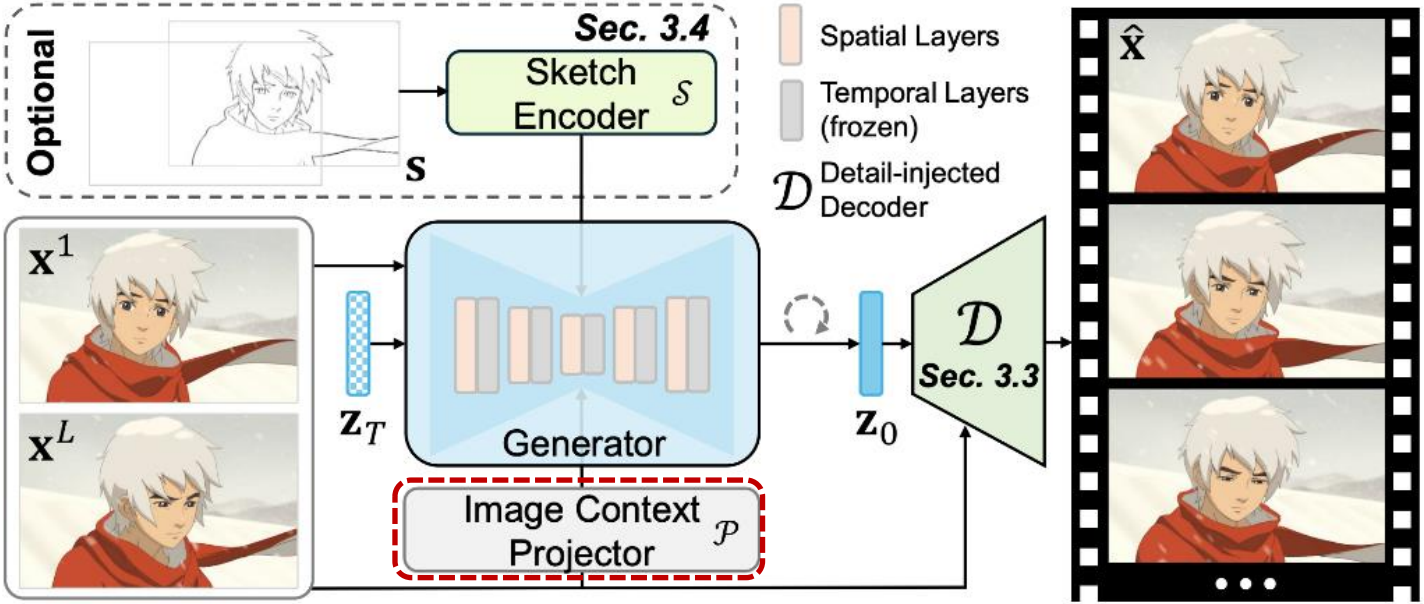


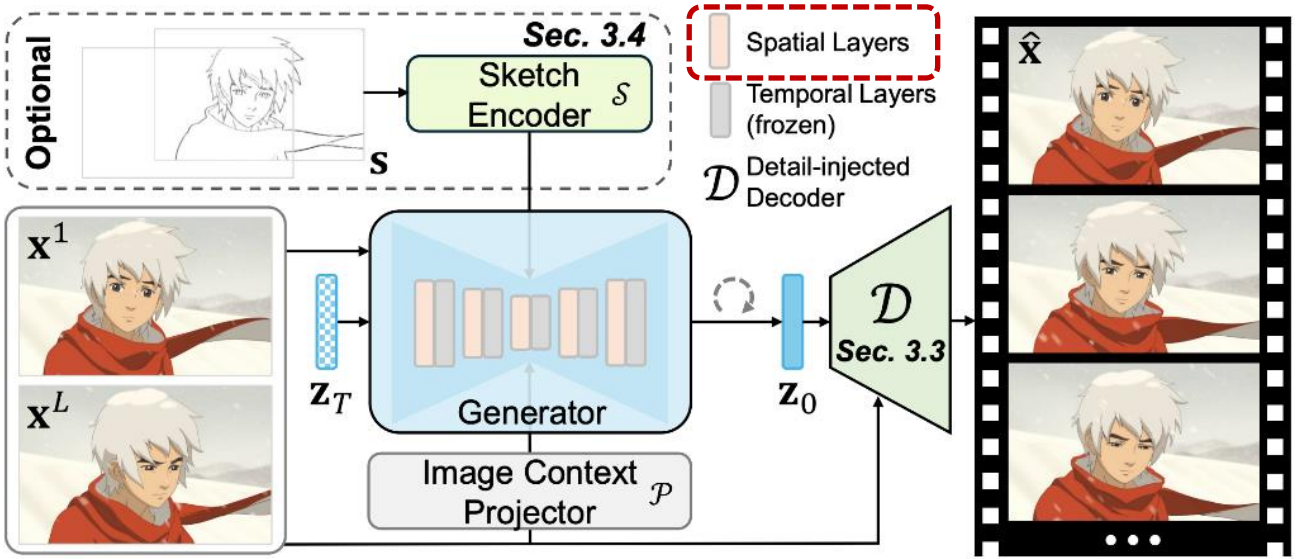
Image-Context projector (Toon rectification learning)

- 모델이 input frames의 **context를 이해**하도록 돕는 부분
- **cartoon 장면의 context**를 더 잘 이해할 수 있게 fine tuning 되어야 함



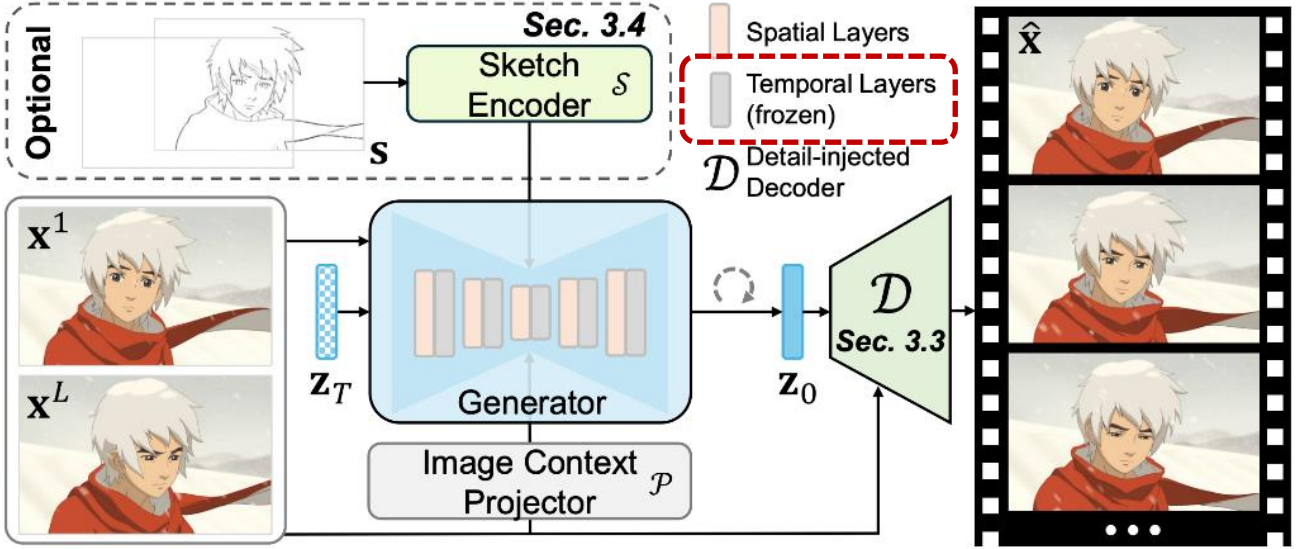
Spatial layers (Toon rectification learning)

- Sharing the same architecture as StableDiffusion v2.1
- Video frames의 appearance distribution을 학습
- **Real-world frame 이 생성되는 것을 방지**하고, cartoon frame 을 생성할 수 있도록 fine tuning



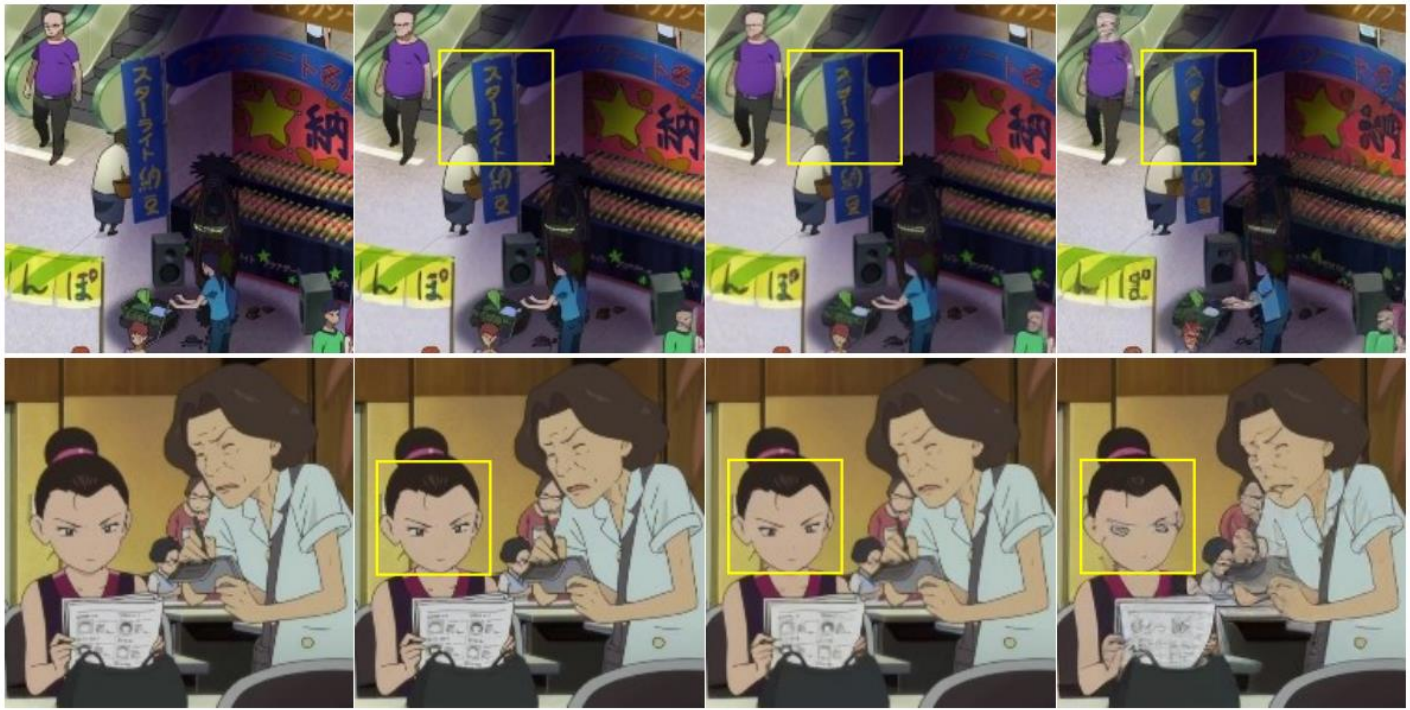
Temporal layers (Toon rectification learning)

- Video frames 간의 **motion dynamics를 포착**
- video frames간의 motion dynamics를 잘 포착하고, real-world motion prior를 유지하기 위해 **frozen 시킴**

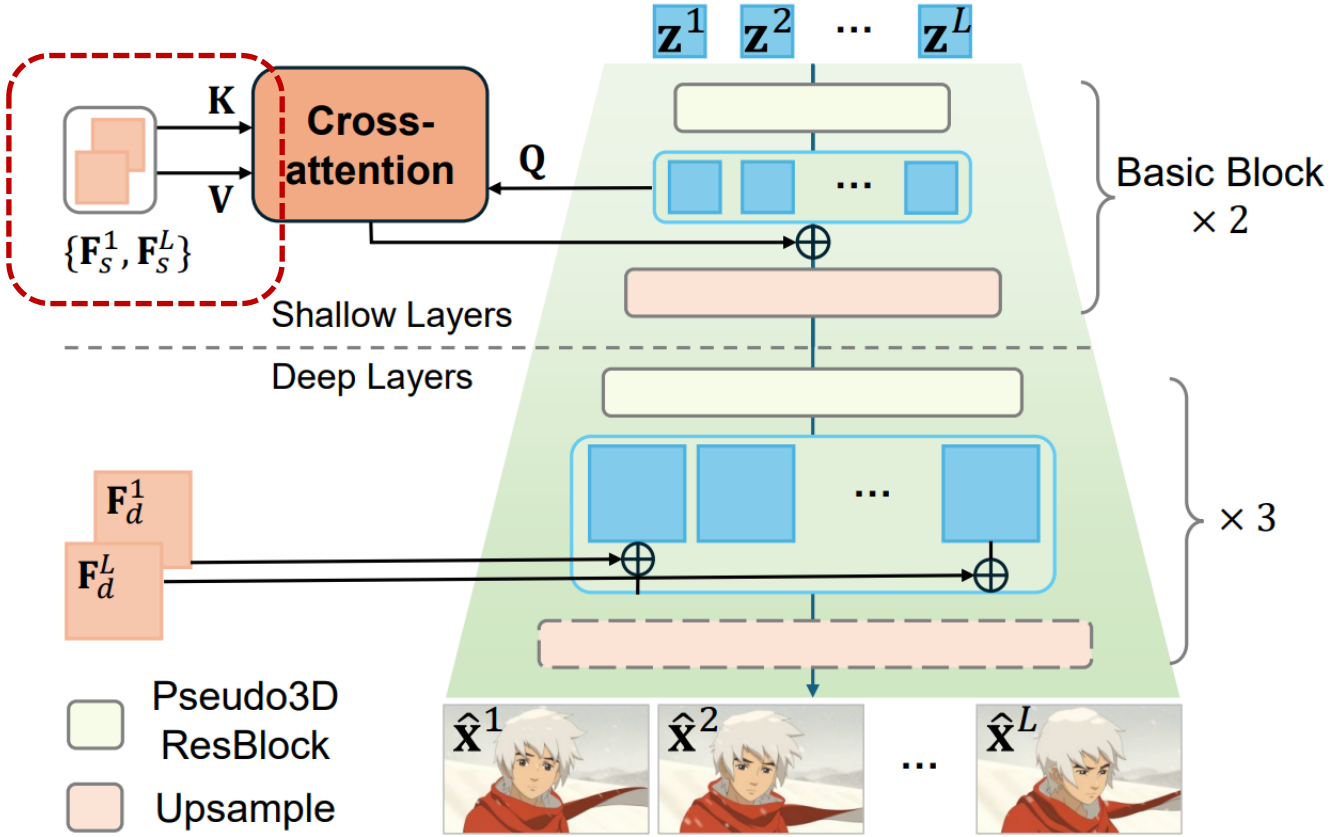


Method 2 : Dual reference-based 3D decoder (1)

- 대부분의 diffusion models은 **highly compressed latent spaces**를 사용하고, 이때문에 **디테일한 그림의 생성이 어려움**



Method 2 : Dual reference-based 3D decoder



Method 3 : Sketch encoder

- 생성되는 frames를 더 **control 하기 위해 sketch-based control** 이용
- **ControlNet**과 동일한 전략으로 학습

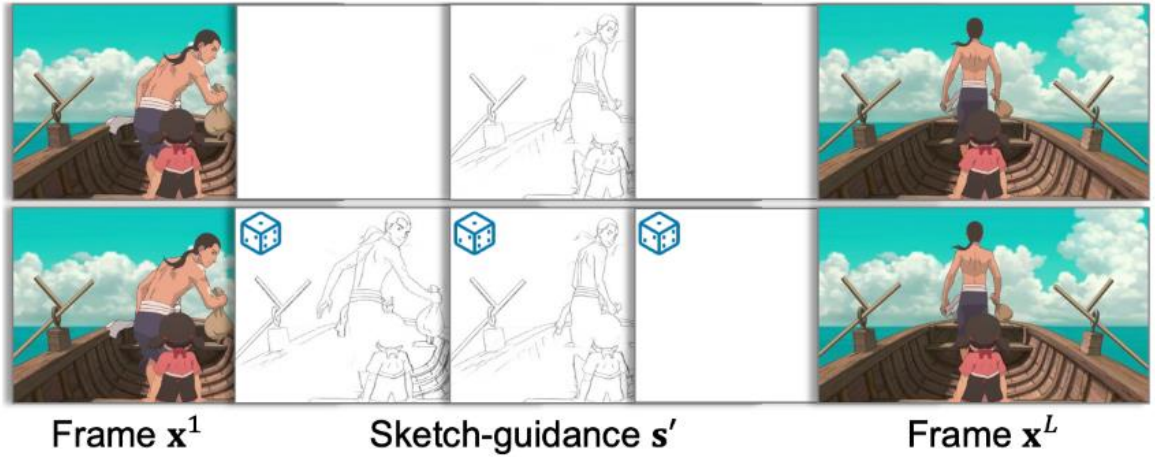
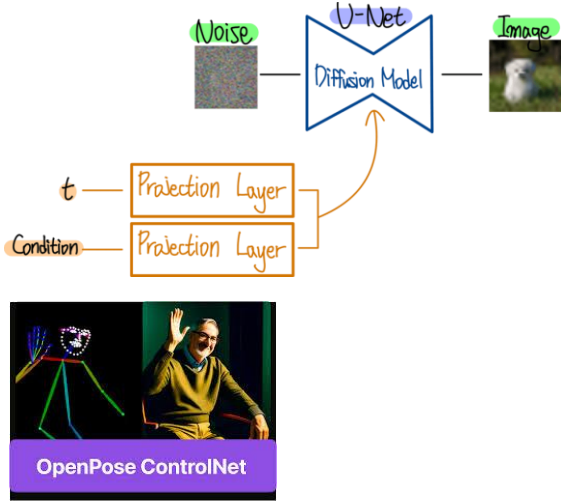


Figure 4. Examples of different patterns of sketch-guidance: (top) bisection (n=1) and (bottom) random position.



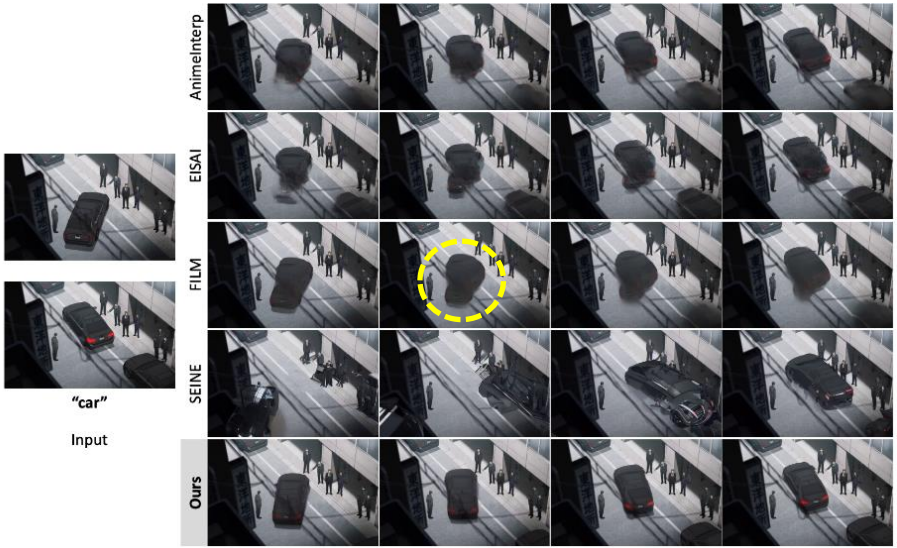
https://ffighting.net/deep-learning-paper-review/diffusion-model/controlnet/

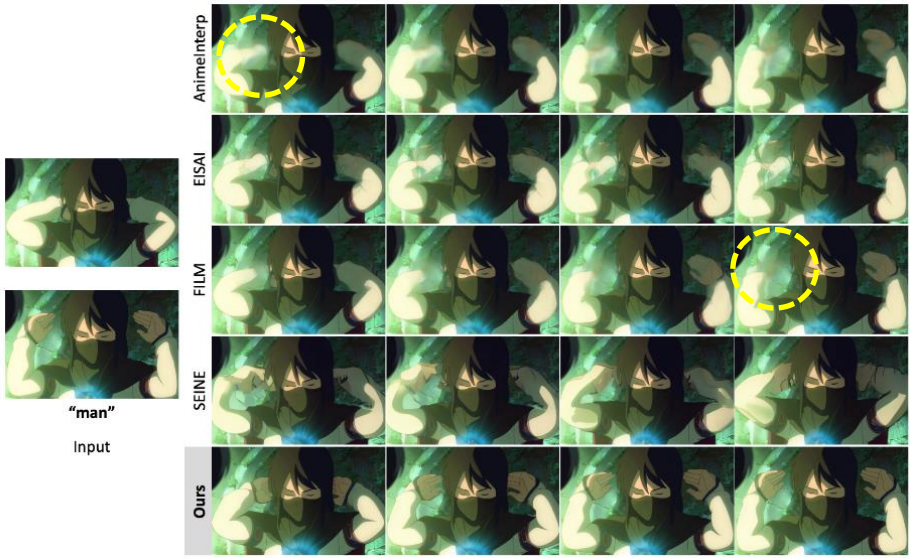
Quantitative Comparisons

- **Fre´chet Video Distance (FVD), Kernel Video Distance (KVD)** : evaluate the quality and temporal motion dynamics of generated videos
- **LPIPS** : generated video frames와 ground-truth videos 간의 perceptual similarity
- **CLIP** : text 와 generated video frames, 실제 frame과 generated video frames 간의 유사성 평가

Metric	AnimeInterp	EISAI	FILM	SEINE	Ours
FVD ↓	196.66	146.65	189.88	98.96	**43.92**
KVD ↓	8.44	5.55	8.01	2.93	**1.52**
LPIPS ↓	0.1890	0.1729	**0.1702**	0.2519	0.1733
CLIP_{img} ↑	0.8866	0.9083	0.9006	0.8531	**0.9221**
CLIP_{txt} ↑	0.3069	0.3097	0.3083	0.2962	**0.3129**
CPBD ↑	0.5974	0.6413	0.6317	0.6630	**0.6723**

Qualitative Comparisons





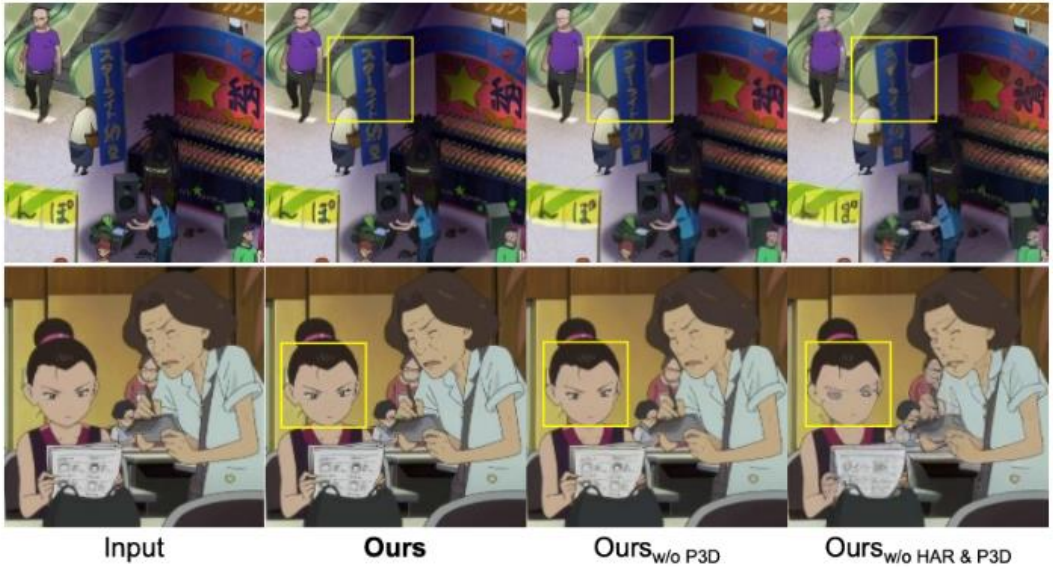
Ablation Study : **Toon rectification learning**

- I : directly using the pre-trained backbone model
- II : fine-tuning the imagecontext projector (ICP) and entire denoising U-Net (Spatial+Temporal layers)
- III : fine-tuning ICP and spatial layers while bypassing temporal layers in forwarding during training
- IV (Ours): fine-tuning ICP and spatial layers while keeping temporal layers frozen
- V : fine-tuning only ICP

Variant*	ICP	Spa.	Temp.	Bypass Temp.	FVD ↓	CLIP_{img} ↑
I (DCinterp)					86.62	0.8637
II	✓	✓	✓		70.73	0.8978
III	✓	✓		✓	291.45	0.7997
IV (Ours)	✓	✓			**52.73**	**0.9096**
V	✓				81.45	0.8875

Dual-reference-based 3D VAE decoder

Variant	Ref.	Temp.	PSNR ↑	SSIM ↑	LPIPS ↓
Ours	✓	✓	**33.83**	**0.9450**	**0.0204**
$\text{Ours}_{\text{w/o P3D}}$	✓	✗	32.51	0.9270	0.0326
$\text{Ours}_{\text{w/o HAR \& P3D}}$	✗	✗	29.49	0.8670	0.0426



Sparse sketch guidance

Thank you

Q&A