X:AI 5th Seminar 2024

DynamiCrafter: Animating Open-domain Images with Video Diffusion Priors

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu,
Hanyuan Liu, Xintao Wang, Tien-Tsin Wong, Ying Shan

Arxiv 2023 **출판**

2024.07.23 3**주차** ADV Session **발표**

국민대학교 AI빅데이터융합경영학과
김서령 tjfud1025@gmail.com

국민대학교
인공지능학회 X:AI

CONTENTS

국민대학교
인공지능학회 X:AI

Introduction

기존 Image Animation 기법의 한계

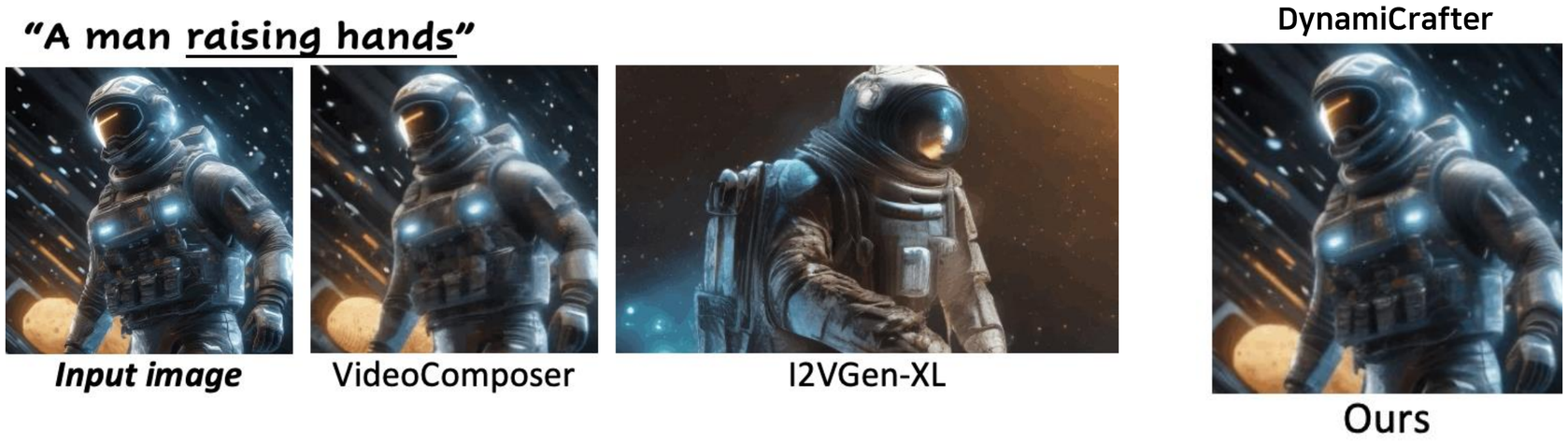
- 예측할 수 없는 변화나 움직임 (예: 구름, 유체) 애니메이션화 중점
- 특정 도메인 특화 동작 (예: 머리카락, 신체 동작) 애니메이션화 중점
- Image Animation은 visual context 이해와 detail 보존을 모두 만족해야 함
- 기존 Video Diffusion 모델 (VideoComposer, I2VGen-XL)의 문제점
  - 갑작스러운 시간적 변화 (Abrupt temporal change) : 애니메이션이 중간에 갑작스럽게 변화함
  - 낮은 시각적 일치성 (Low visual conformity) : 생성된 비디오가 입력 이미지와 시각적으로 일치하지 않음

Introduction

Image Animation은 visual context 이해와 detail 보존을 모두 만족해야 함

DynamiCrafter 제안

- 핵심 아이디어: Text2Video Diffusion에 조건부 이미지를 사용하여 Video Generation 과정을 조절함
- 문제 해결 방법: Dual-Stream Image Injection Paradigm를 도입함
  - Text-aligned Context Representation: Visual context를 이해하고 동적 콘텐츠를 생성함
  - Visual Detail Guidance: Image detail preservation 문제를 해결함



Video Diffusion Models

1. VDM의 등장

- 저해상도 비디오를 모델링하는 VDM
- Imagen-Video: cascaded DMs로 고해상도 비디오 생성

2. Text2Video Diffusion (T2V)

- 저비용으로 T2I를 T2V로 전환하는 연구
- 문제점: 텍스트 프롬프트만으로는 사용자의 의도를 정확히 반영하기 어려움

3. 추가 제어 신호 도입

- structure, pose, Canny edge 등 신호를 T2V에 도입
- 문제점: VDM에서 RGB 이미지 같은 visual condition이 충분히 탐구되지 않음

Video Diffusion Models

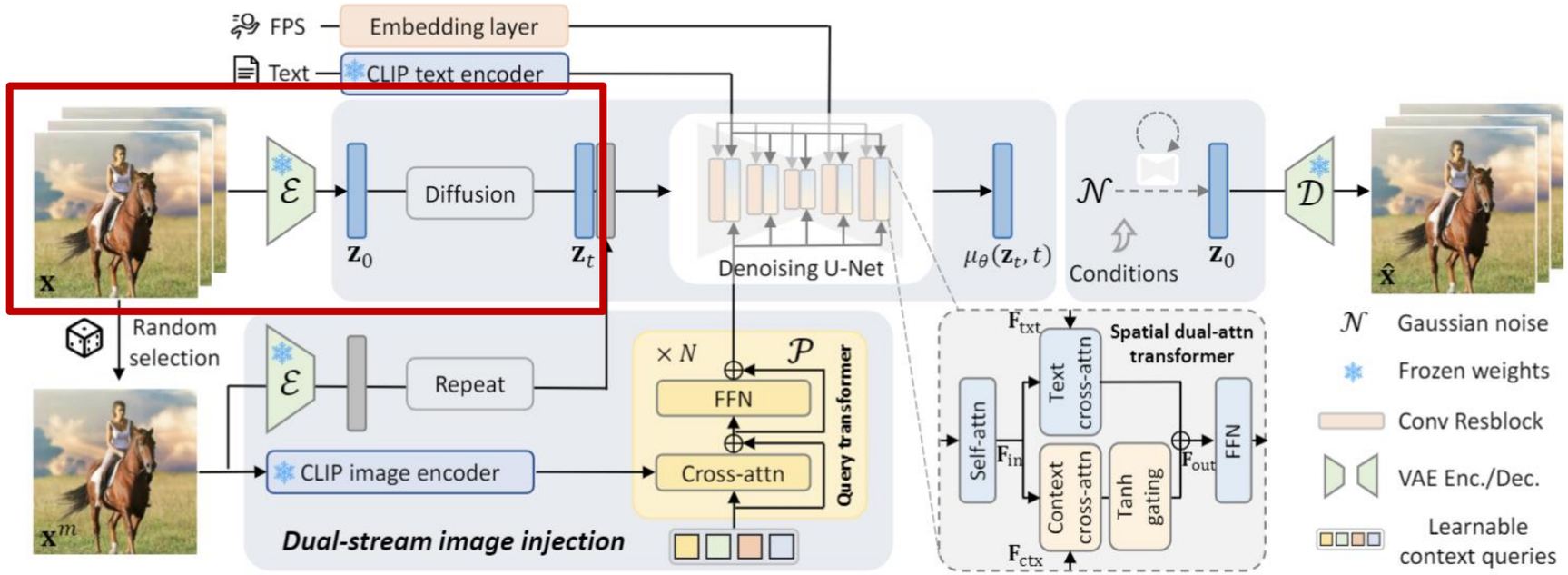
4. Visual Condition 관련 연구

- Seer, VideoComposer, I2VGen-XL 등의 연구
- 문제점: 특정 도메인에 집중(Seer), 시간적으로 일관된 프레임과 현실적인 움직임 생성 불가(VideoComposer),
  입력 이미지의 visual detail 보존 실패(I2VGen-XL)

5. DynamiCrafter 제안

- 텍스트 조건이 있는 VDM을 통해 open domain 이미지를 애니메이션화

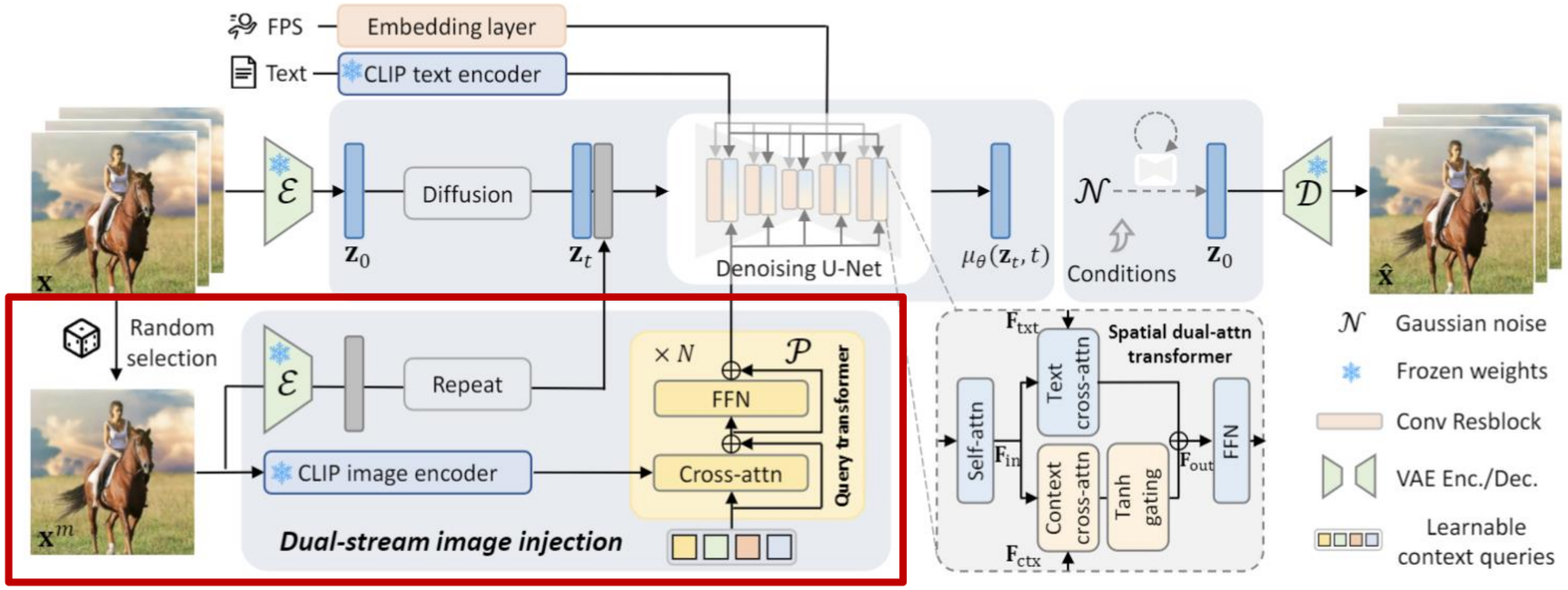
Forward Diffusion Process



Forward Diffusion Process

1. 입력 단계 : 여러 개의 정적 이미지 x를 입력으로 받아 Encoder를 통해 인코딩됨
2. Diffusion : 인코딩된 이미지 feature z_0에 노이즈를 추가함 → 잠재 표현 z_t가 생성됨

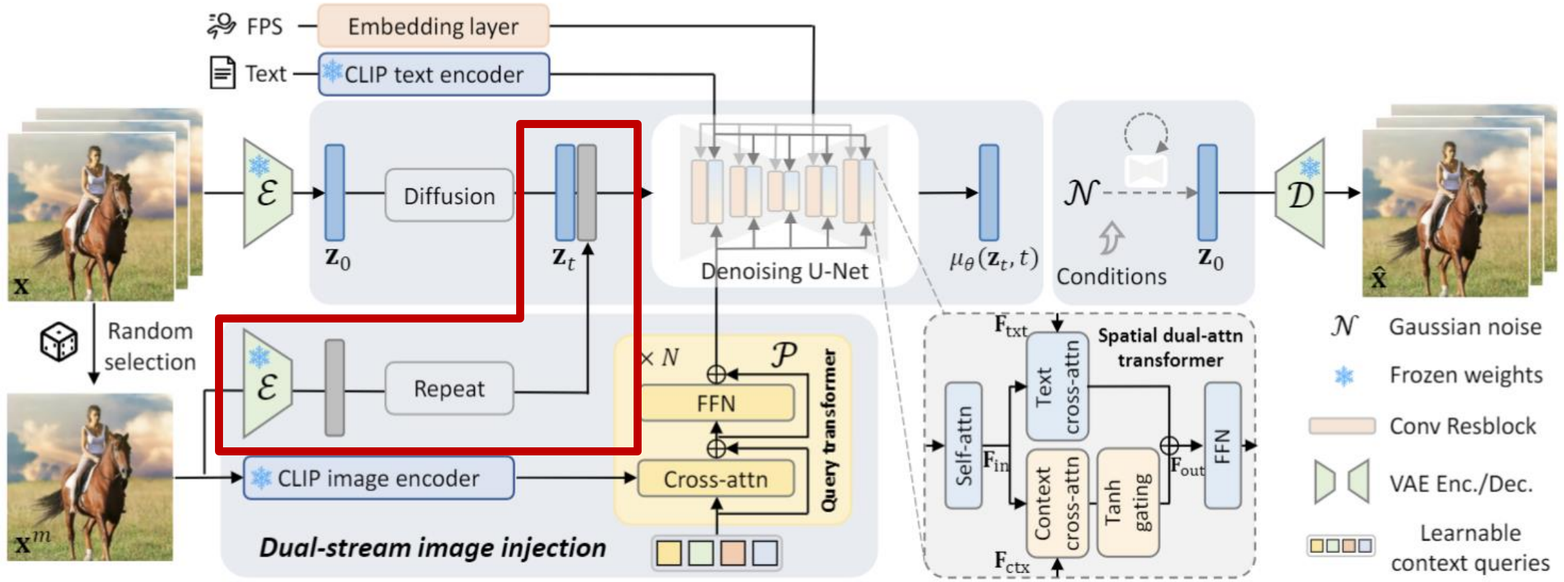
Dual-Stream Image Injection Paradigm



Dual-Stream Image Injection Paradigm

- Visual Detail Guidance : 비디오 생성 시 이미지의 시각적 세부 사항을 유지하도록 함
- Text-aligned Context Representation : 이미지 정보를 텍스트 임베딩 공간에 투영하여 비디오 생성에 필요한 문맥을 이해하도록 함

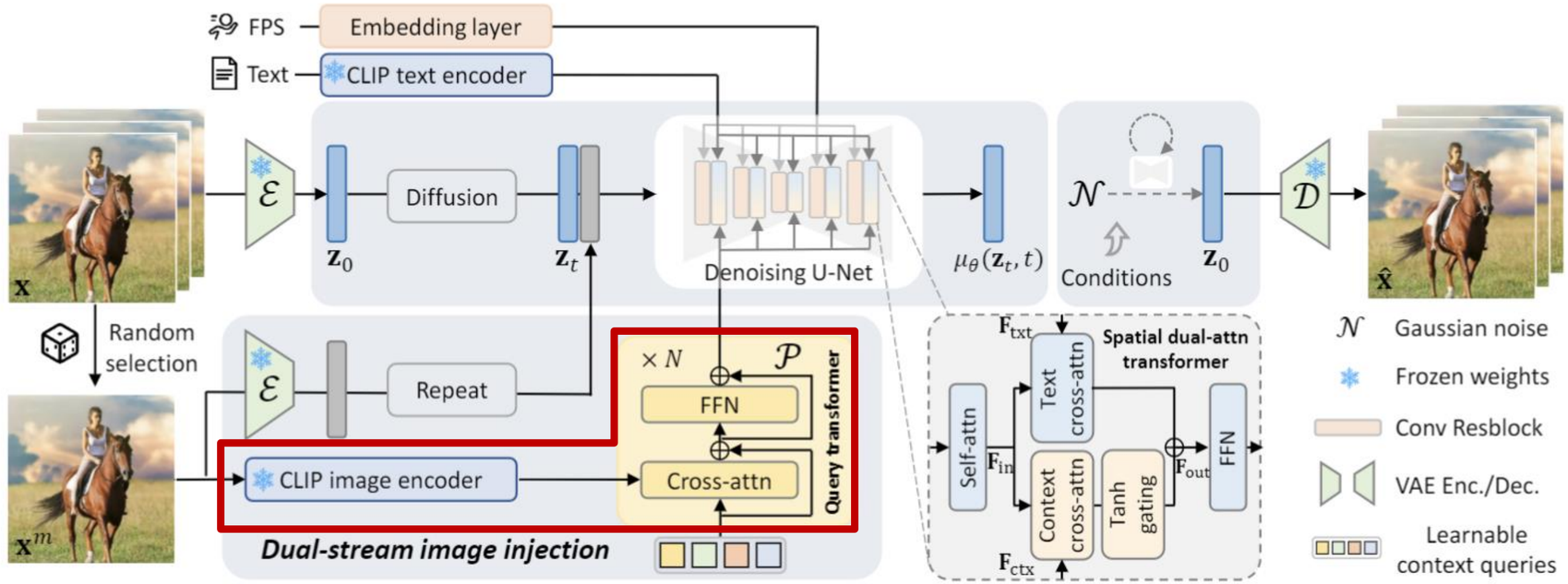
Visual detail guidance (VDG)



Visual detail guidance (VDG)

1. 여러 개의 정적 이미지 x에서 랜덤으로 임의의 프레임 x^m를 선택하고, Encoder를 통해 인코딩함
2. 인코딩된 이미지 feature를 반복하여 Sequence를 생성함
3. 이 Sequence를 조건부 이미지로 사용하고, 초기 노이즈와 결합하여 z_t를 생성함

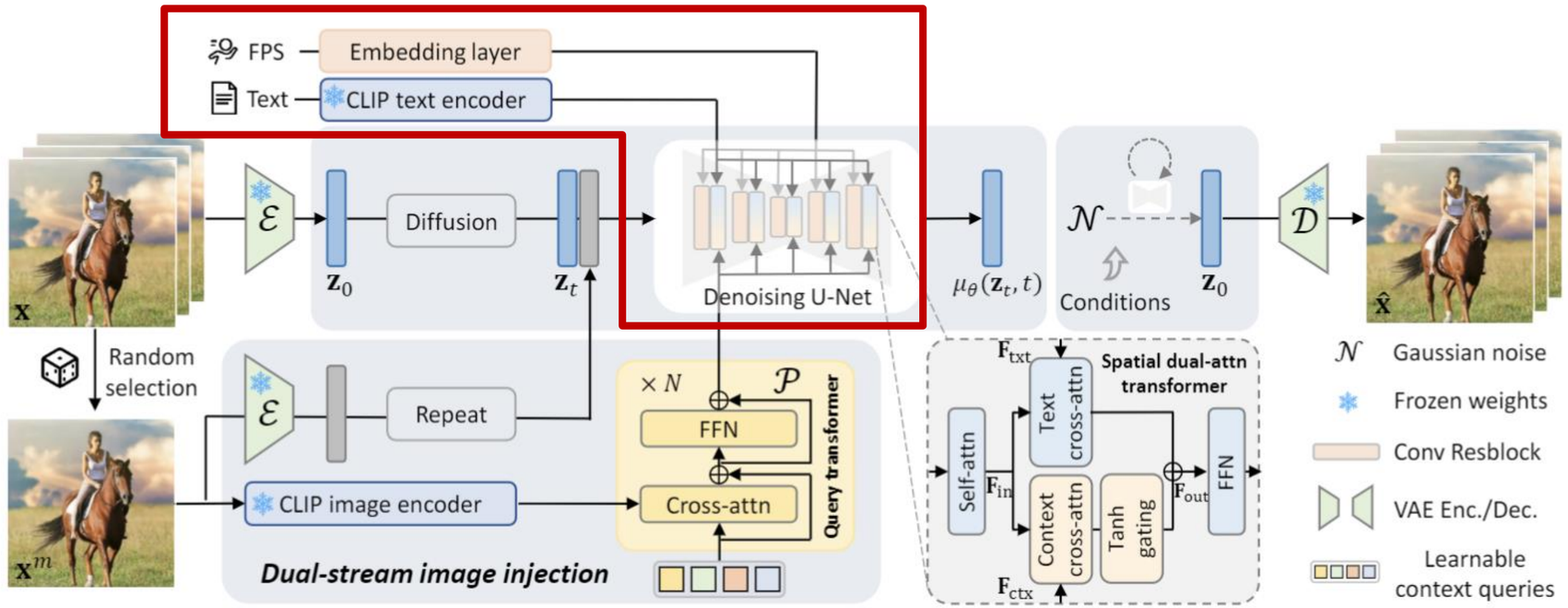
Text-aligned context representation



Text-aligned context representation

1. 여러 개의 정적 이미지 x에서 랜덤으로 임의의 프레임 x^m를 선택하고, CLIP image encoder로 인코딩하여 시각 피처 F_{vis}를 추출함
2. Query Transformer : 시각 피처를 text-aligned context representation으로 변환함
   - Cross-attention 메커니즘을 통해 시각 피처 F_{vis} 와 Learnable context queries Q 간의 상호작용을 학습함
   - FFN을 통해 최종 컨텍스트 표현 F_{ctx}를 생성

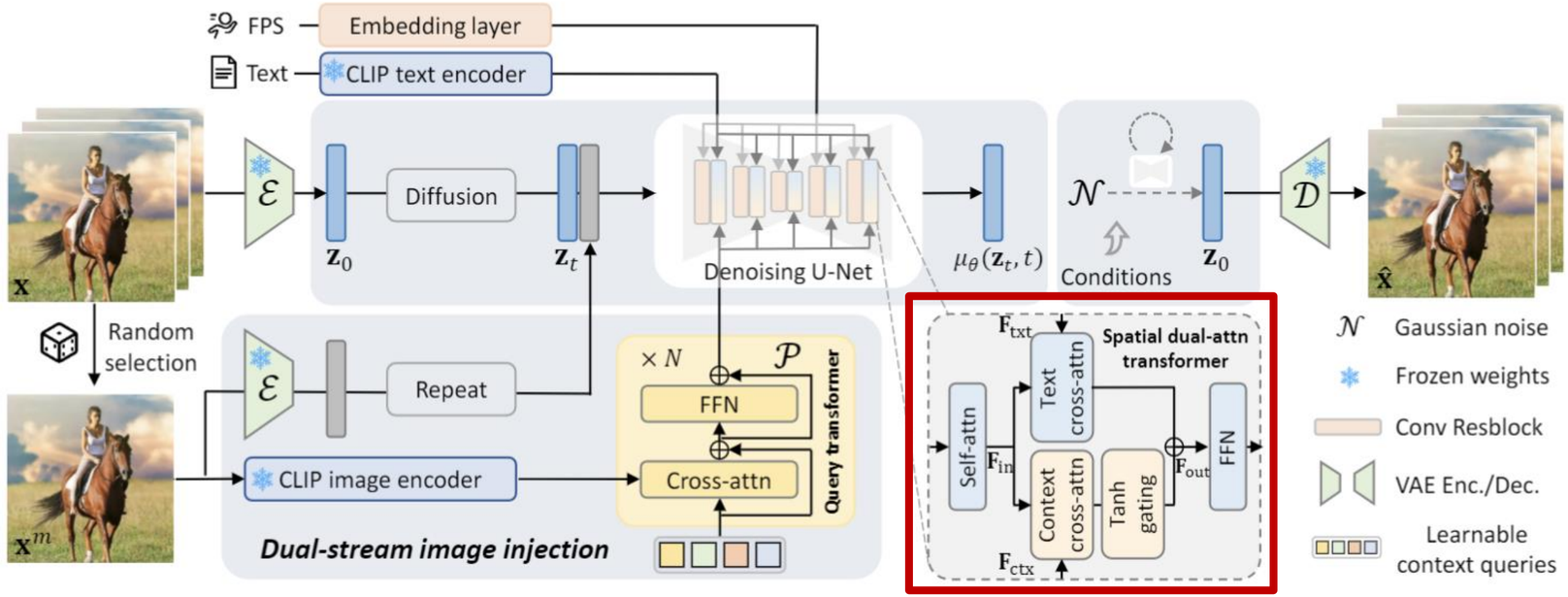
Reverse Diffusion Process



Denoising U-Net

1. 노이즈가 추가된 잠재 표현 z_t가 Denoising U-Net에 입력됨
2. Denoising U-Net은 텍스트 임베딩을 조건으로 하여 z_t에서 노이즈를 제거함
3. 이 과정에서 z_t는 다시 노이즈가 제거된 z_0로 변환됨

Spatial dual-attention transformer



Spatial dual-attention transformer

1. Self-attention : Denoising U-Net의 입력 feature map F_{in}에서 각 위치의 정보를 다른 위치와 비교하여 학습함
2. Text Cross-attention : 텍스트 임베딩 F_{txt}와 이미지 입력 feature map 간의 cross-attention을 수행함
3. Context Cross-attention : context 정보를 feature map에 반영함
4. Tanh Gating : feature map 출력을 gating하여 특정 정보가 더 강조되거나 억제되도록 조절함
5. FFN : 최종 출력 feature map F_{out}를 생성하여 다음 단계로 전달함

Spatial dual-attention transformer

$$F_{\text{out}} = \text{Softmax}\left(\frac{QK_{\text{txt}}^{\top}}{\sqrt{d}}\right)V_{\text{txt}} + \lambda \cdot \text{Softmax}\left(\frac{QK_{\text{ctx}}^{\top}}{\sqrt{d}}\right)V_{\text{ctx}}$$

Text Cross-attention

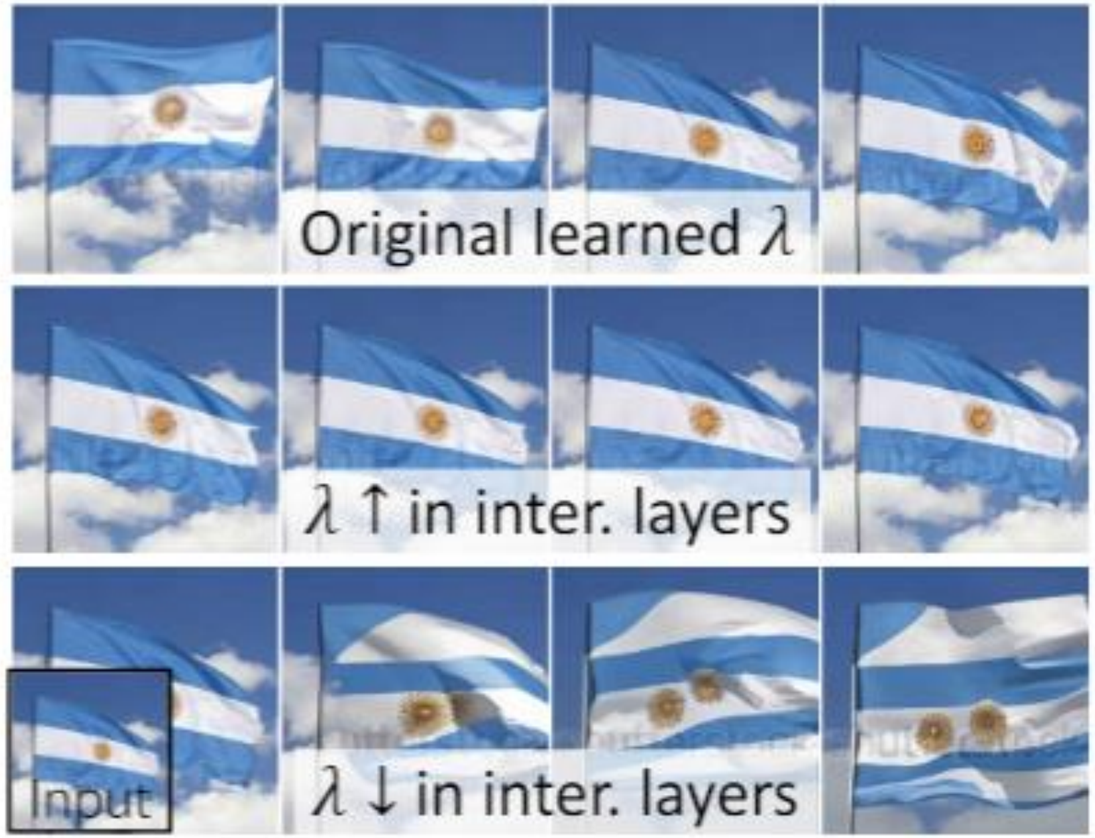
- 쿼리 Q와 텍스트 키 K_{txt}의 내적을 $\sqrt{d}$로 나누어 스케일링함
- softmax 함수로 유사도를 계산한 후, 텍스트 값 V_{txt}와 결합함

Context Cross-attention

- 쿼리 Q와 컨텍스트 키 K_{ctx}의 내적을 $\sqrt{d}$로 나누어 스케일링함
- Softmax 함수로 유사도를 계산한 후, 컨텍스트 값 V_{ctx}와 결합함
- 가중치 λ로 조정함

Text Cross-attention와 가중치로 조정된 Context Cross-attention의 결과를 더하여 최종 출력 F_{out}을 생성함

Observations and analysis of λ



U-Net Layer 특성

- 중간 레이어 (Intermediate Layers) : 주로 객체의 모양과 포즈와 관련됨
- 양쪽 끝 레이어 (End Layers) : 주로 비디오의 외관(appearance)과 관련됨

중간 레이어에서 λ 조절 실험

- λ를 증가시키면 프레임 간 움직임이 줄어듭
- λ를 감소시키면 객체의 형태를 유지하기 어려움

→ 최적의 λ 값을 찾아 텍스트 정보와 컨텍스트 정보를 균형 있게 반영하여, 프레임 간의 움직임과 객체의 형태를 일관되게 유지하도록 함

Quantitative Evaluation

Method	UCF-101			MSR-VTT		
	FVD ↓	KVD ↓	PIC ↑	FVD ↓	KVD ↓	PIC ↑
VideoComposer	576.81	65.56	0.5269	377.29	26.34	0.4460
I2VGen-XL	571.11	**58.59**	0.5313	289.10	14.70	0.5352
Ours	**429.23**	62.47	**0.6078**	**234.66**	**13.74**	**0.5803**

- Fréchet Video Distance (FVD): 감성 품질과 시간적 일관성을 평가하는 지표로, 낮을수록 품질이 좋음
- Kernel Video Distance (KVD): 시간적 일관성을 중점으로 평가하는 지표로, 낮을수록 좋음
- Perceptual Input Conformity (PIC): 입력 이미지와 애니메이션 결과의 일치도를 측정함. 높을수록 입력 이미지와 일치함

→ **UCF-101에서 KVD지표를 제외하고, 모든 평가 지표에서 기존 방법보다 우수한 성과를 냄**

Qualitative Evaluation

"An anime scene with windmills standing tall ..."

Input image

VideoComposer

I2VGen-XL

PikaLabs

Gen-2

Ours

"Some people walking on a road with pedestrian crossing"

Input image

VideoComposer

I2VGen-XL

PikaLabs

Gen-2

Ours

Qualitative Evaluation

"A girl talking"

Input image

VideoComposer

I2VGen-XL

PikaLabs

Gen-2

Ours

"A tiger"

Input image

VideoComposer

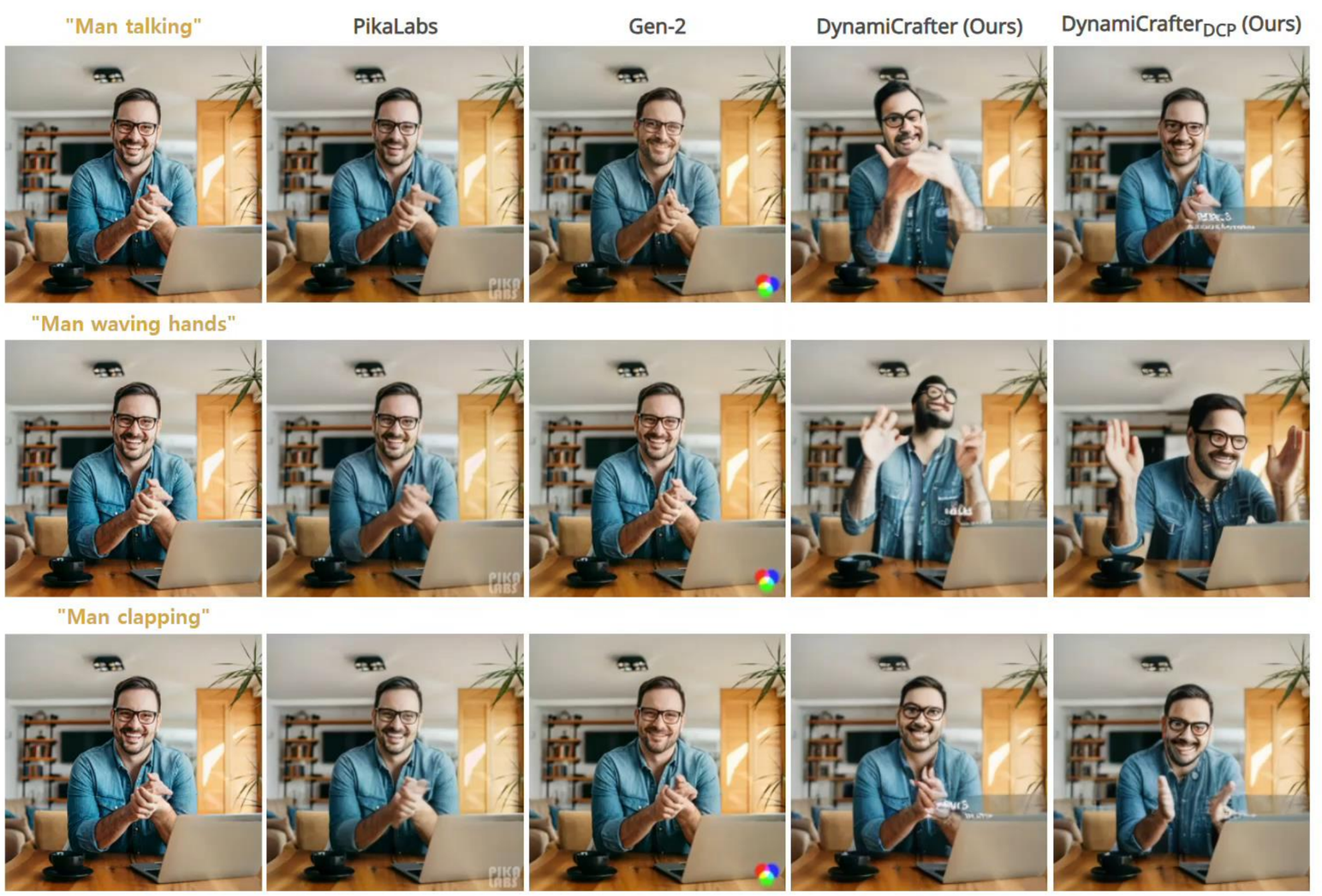
I2VGen-XL

PikaLabs

Gen-2

Ours

Discussions on Motion Control using Text



WebVid10M 데이터셋을 필터링하고 재주석해서 새로운 데이터셋을 구성함

- 기존 데이터셋의 캡션은 장면 설명에 초점이 많고 동적 설명이 적어, 모델이 동작을 학습하는 데 한계가 있었음
- 이미지 애니메이션을 위해 장면 설명은 이미지 조건으로, 동작 설명은 텍스트 조건으로 분리하여 모델을 훈련함

Conclusion

- DynamiCrafter는 pretrained video diffusion prior를 활용하여 정적 이미지를 애니메이션으로 변환하는 프레임워크임
- dual-stream image injection mechanism를 도입함
- 오픈 도메인 이미지를 애니메이션으로 변환하는 데 있어서 탁월한 성능을 보여줌
- 구축된 데이터셋을 사용하여 이미지 애니메이션을 위한 text-based dynamic control를 탐구함

Reference

https://doubiiu.github.io/projects/DynamiCrafter/#top

https://github.com/Doubiiu/DynamiCrafter

X:AI 5th Seminar 2024

ThankYou

국민대학교 AI빅데이터융합경영학과
김서령 tjfud1025@gmail.com

국민대학교
인공지능학회 X:AI