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

ECCV 2024 (Oral).

날짜 : 2024-10-09

발표자 : 홍진욱

9th-together-RL

Contents

1 Introduction

2 Method

3 Training Paradigm

4 Experiments

5 Conclusion

6 Application

Introduction

Background

Limits in traditional image animation techniques:

- Animating scenes with stochastic dynamics(ex. Clouds & fluid)
- Domain-specific motions(ex. Human hair/body motions)

Non-trivial: requires both visual context understanding and detail preservation.

Recent studies(VideoComposer, I2Vgen-XL) – Lack of comprehensive image injection mechanism.

- Abrupt temporal changes, Low visual conformity

"A woman walking through a forest with a lightning bolt"

Input image

VideoComposer

I2VGen-XL

"Car driving down a road with smoke coming out of it"

Input image

VideoComposer

I2VGen-XL

Introduction

Background

DynamiCrafter

- Key Idea: To govern the video generation process of T2V diffusion models by incorporating a conditional image.
- Addressing the challenge: dual-stream image injection paradigm
  - Text-aligned context representation: Understand visual context and generate dynamic contents
  - Visual detail guidance: Image detail preservation

Input image

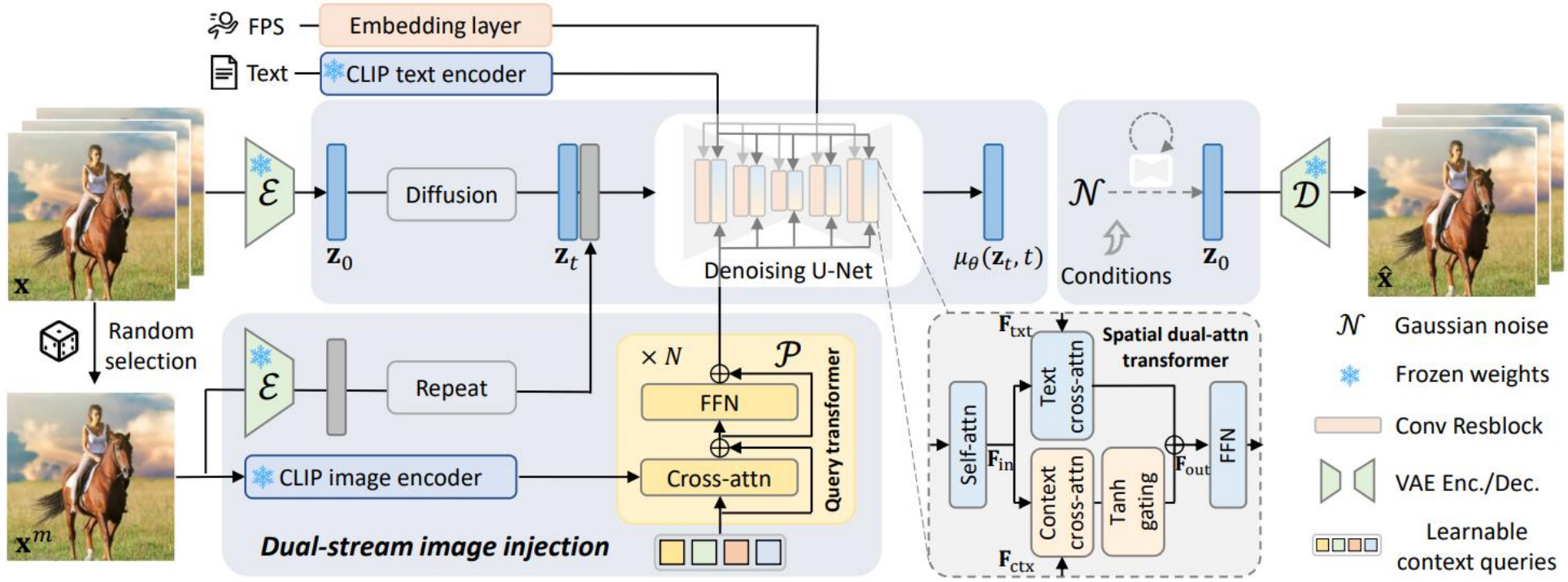
Ours

Input image

Ours

Method

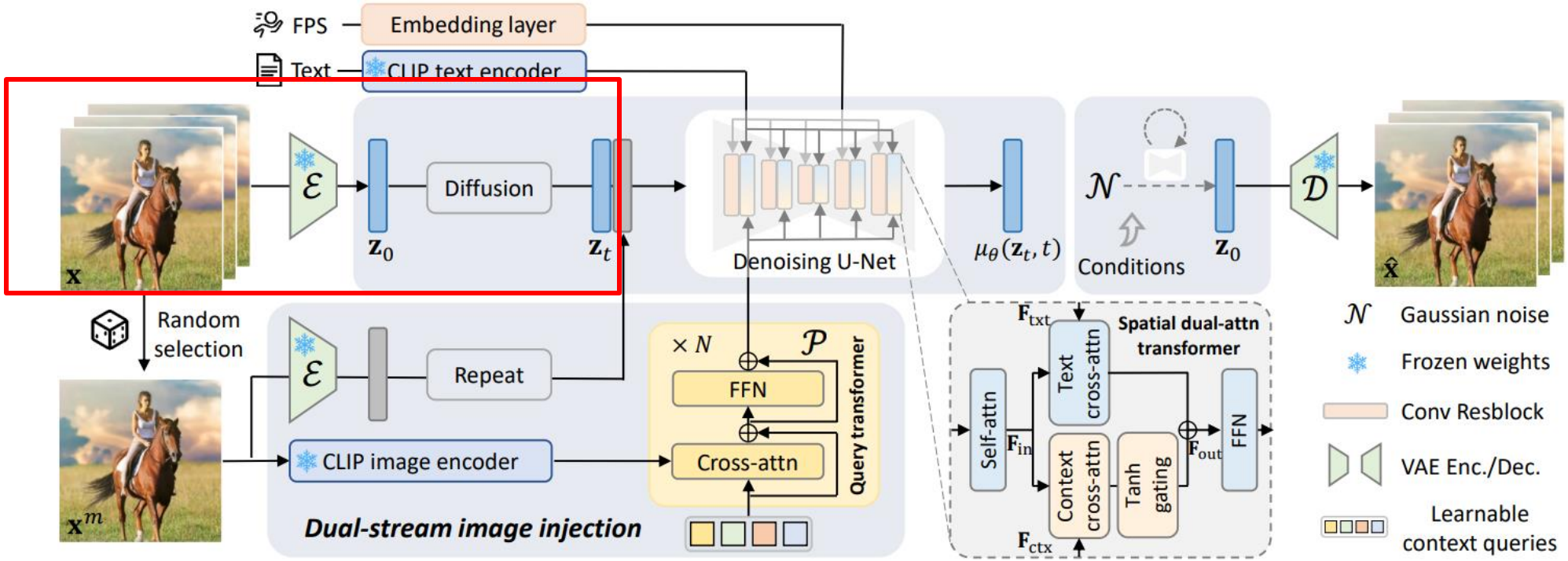
Video diffusion models



- Given image -> Context Projection-> Dual-Stream Injection -> Training -> Inference

Method

Forward Diffusion Process

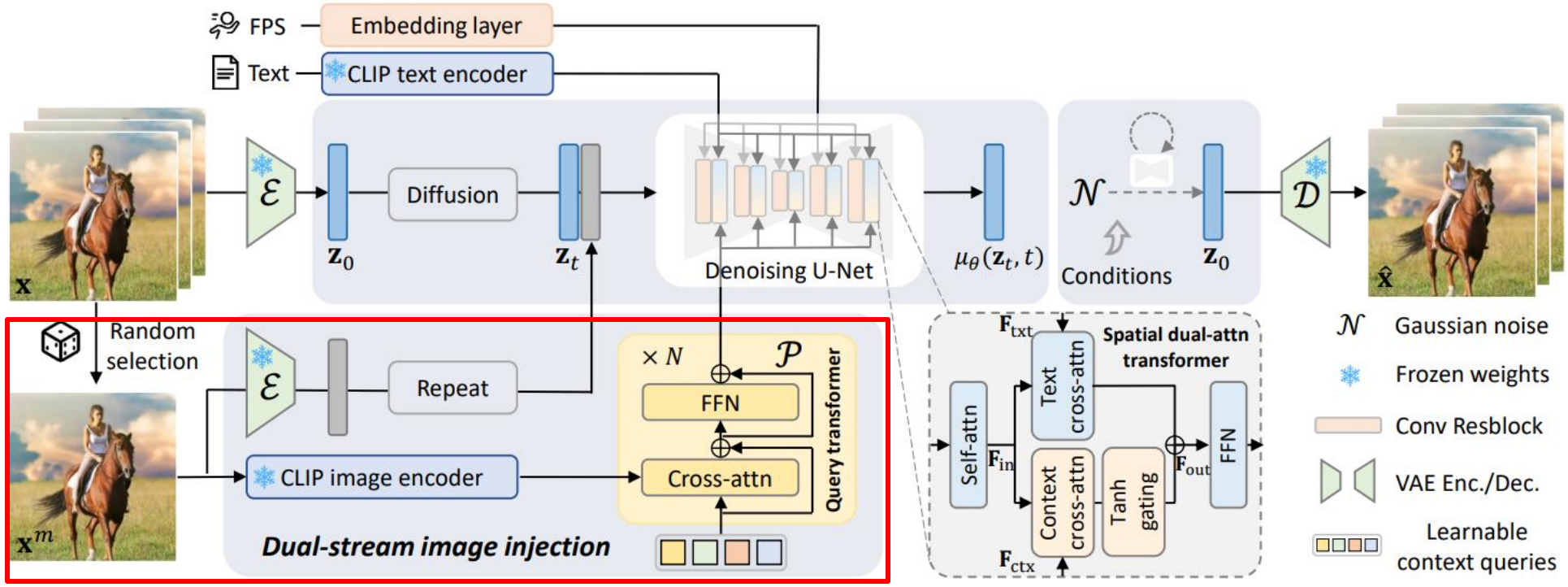


Forward Diffusion Process

- 1. 입력 단계 : 여러 개의 정적 이미지 x를 입력으로 받아 Encoder를 통해 인코딩
- 2. Diffusion : 인코딩된 이미지 feature z0에 노이즈를 추가함 -> 잠재 표현 zt가 생성됨

Method

Dual-Stream Image Injection Paradigm

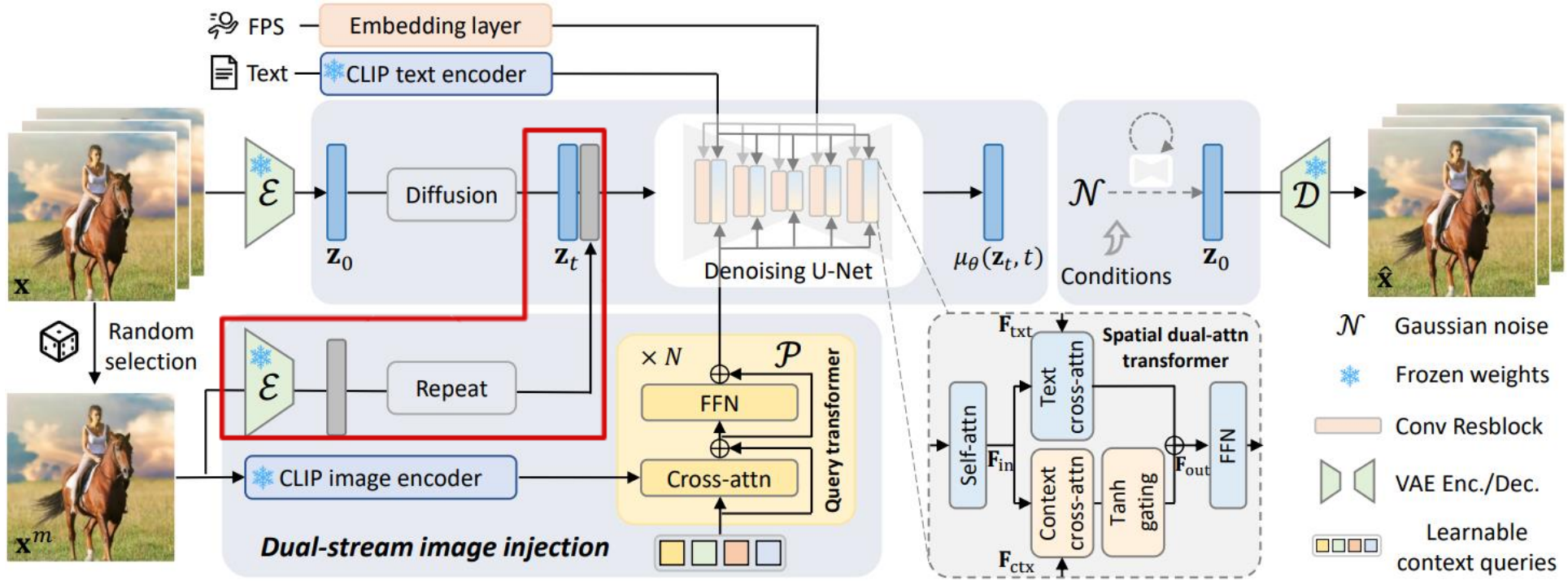


Dual-Stream Image Injection Paradigm

- 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

Method

Visual detail guidance (VDG)



Visual detail guidance (VDG)

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

Method

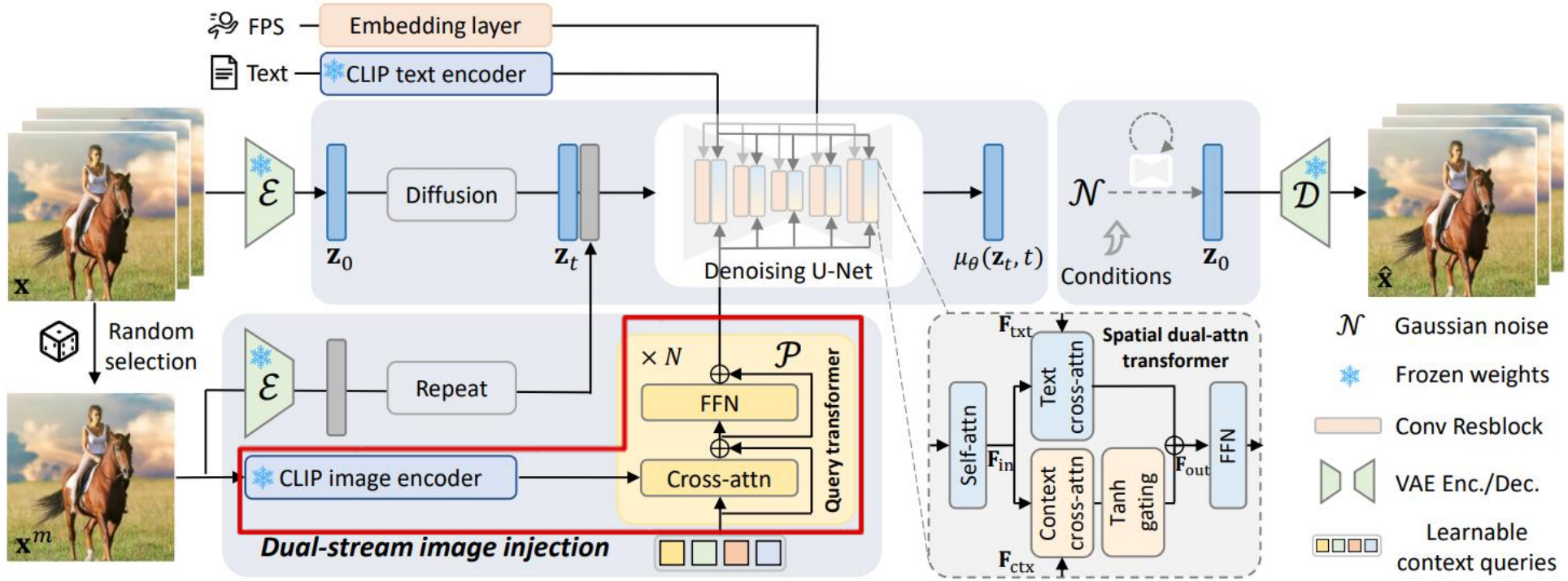
Visual detail guidance (VDG)

Visual detail guidance

- VDM은 입력 이미지와 유사한 동영상을 생성하지만, 일부 시각적 불일치가 발생(CLIP이 시각적 디테일을 완전히 보존x).
- 해결책: 추가 시각적 디테일을 제공하기 위해 이미지와 초기 노이즈를 결합하여 U-Net에 주입.

Method

Text-aligned context representation

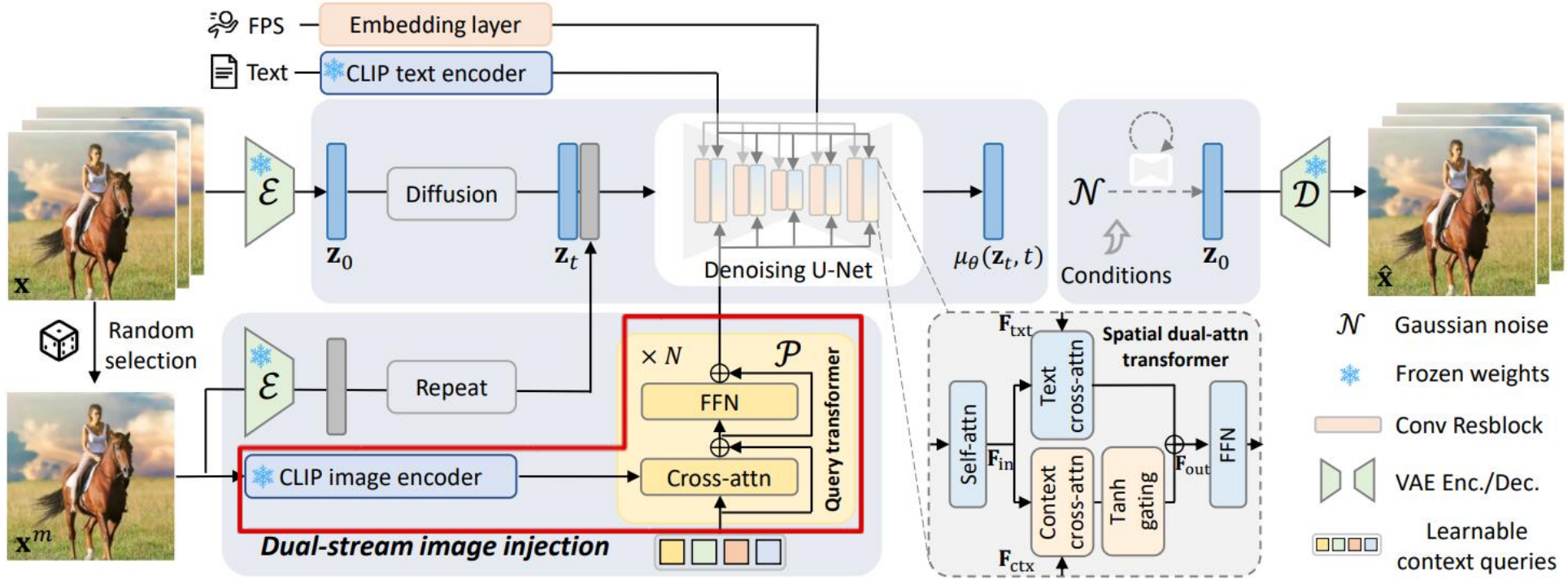


Text-aligned context representation

1. 여러 개의 정적 이미지 x에서 랜덤으로 임의의 프레임 x^t를 선택하고, CLIP image encoder로 인코딩하여 시각 피처 f_{vis}를 추출함

Method

Text-aligned context representation



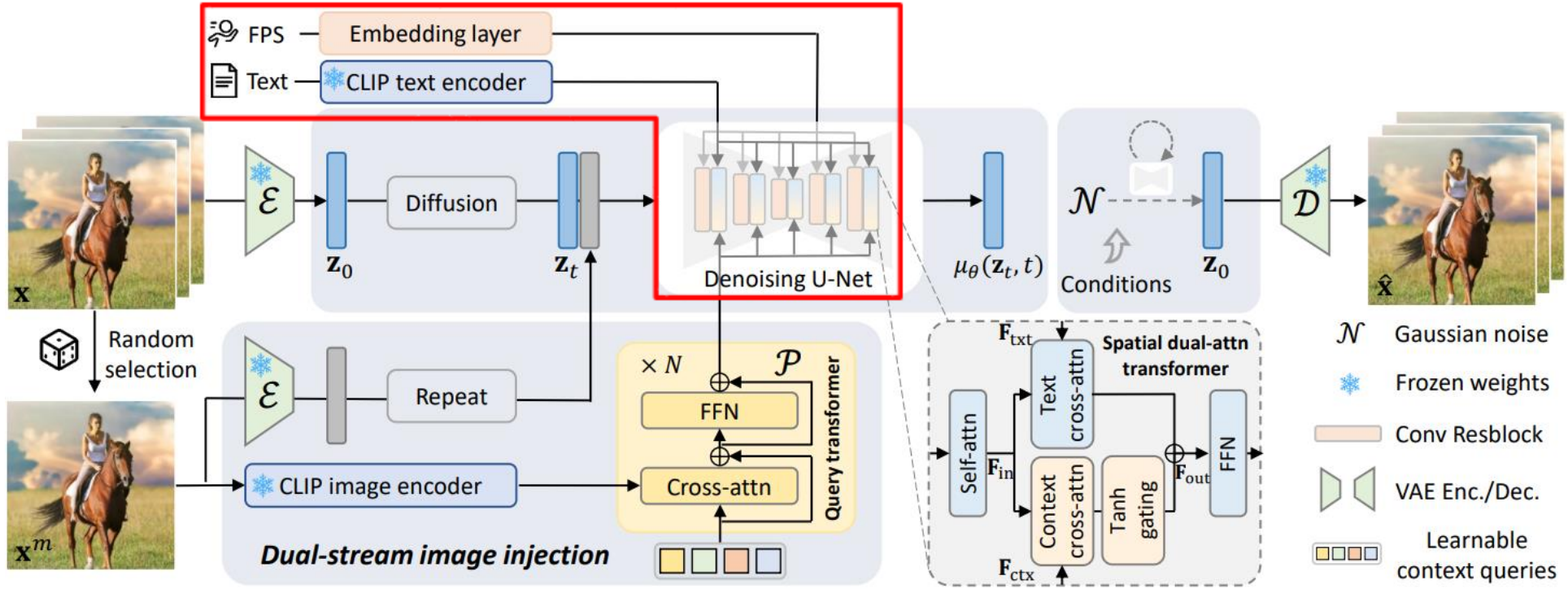
Text-aligned context representation

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

Method

Video diffusion models

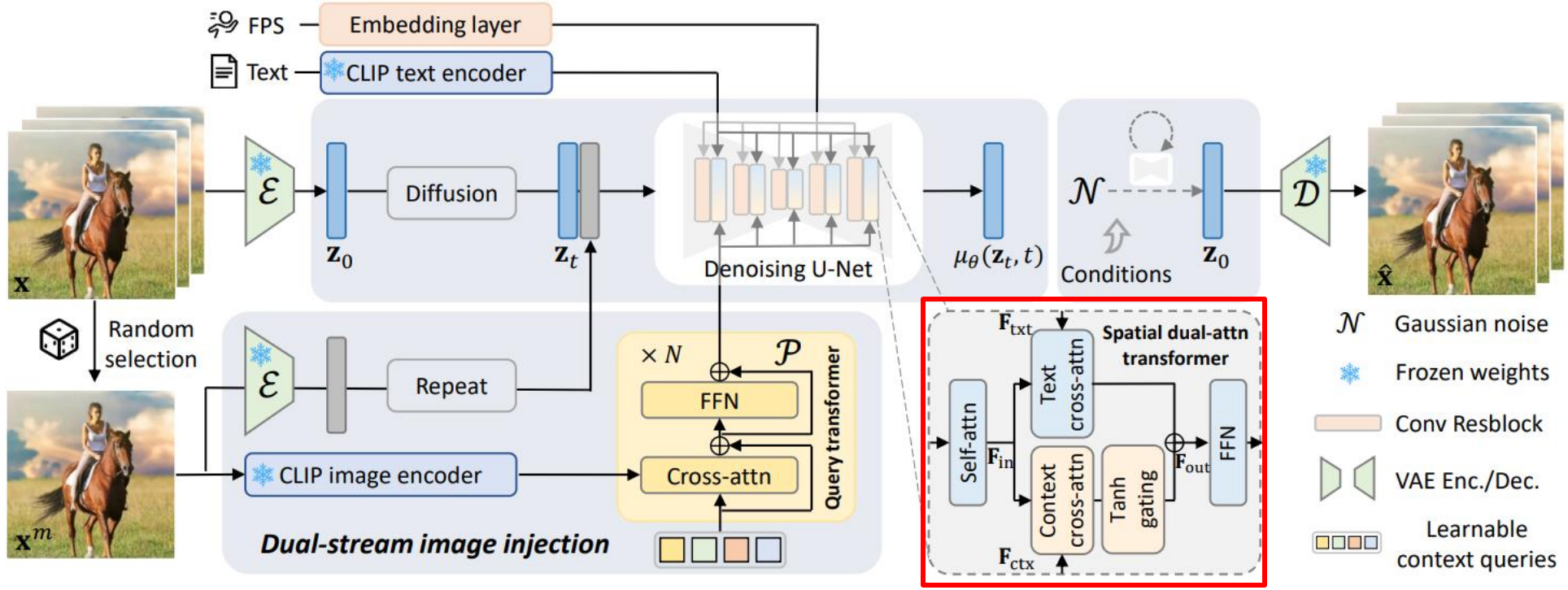


Denoising U-Net

1. 노이즈가 추가된 잠재 표현 z_t 가 Denoising U-Net에 입력됨
2. Denoising U-Net은 텍스트 임베딩을 조건으로 하여 z_t 에서 노이즈를 제거함
3. 이 과정에서 z_t 는 다시 노이즈가 제거된 z_0 로 변환됨

Method

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 mab F_{out} 를 생성하여 다음 단계로 전달함

Method

Spatial dual-attention transformer

$$\mathbf{F}_{\text{out}} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}_{\text{txt}}^{\top}}{\sqrt{d}}\right)\mathbf{V}_{\text{txt}} + \lambda \cdot \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}_{\text{ctx}}^{\top}}{\sqrt{d}}\right)\mathbf{V}_{\text{ctx}}$$

$$\text{where } \mathbf{Q} = \mathbf{F}_{\text{in}}\mathbf{W}_{\mathbf{Q}},\ \mathbf{K}_{\text{txt}} = \mathbf{F}_{\text{txt}}\mathbf{W}_{\mathbf{K}}\ \mathbf{V}_{\text{txt}} = \mathbf{F}_{\text{txt}}\mathbf{W}_{\mathbf{V}}$$

$$\mathbf{K}_{\text{ctx}} = \mathbf{F}_{\text{ctx}}\mathbf{W}'_{\mathbf{K}},\ \mathbf{V}_{\text{ctx}} = \mathbf{F}_{\text{ctx}}\mathbf{W}'_{\mathbf{V}}$$

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

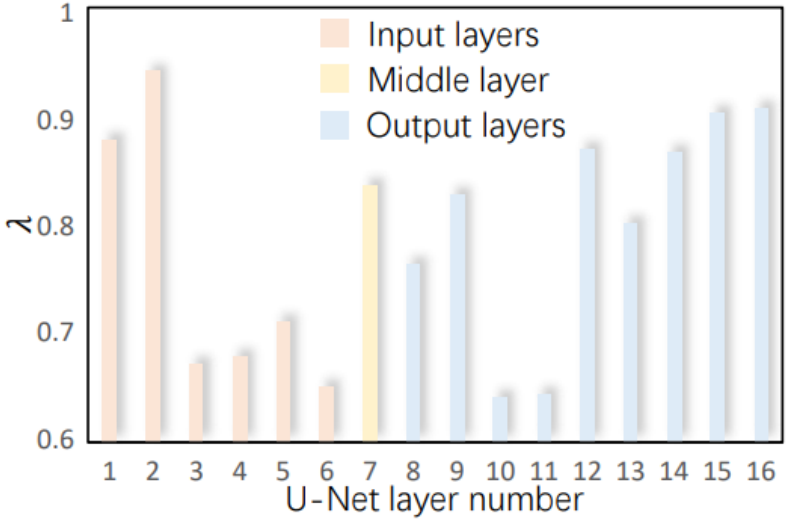
Context Cross-attention

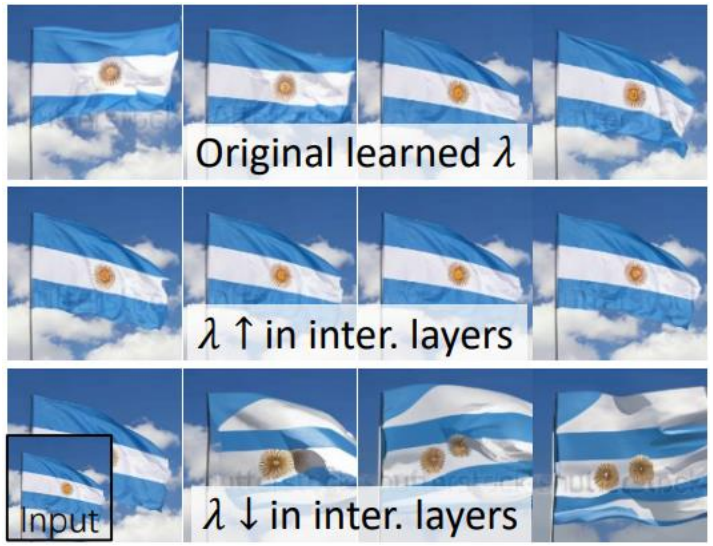
- 쿼리 Q와 컨텍스트 키 K_{cxt}의 내적을 $\sqrt{d}$로 나누어 스케일링함
- Softmax 함수로 유사도를 계산한 후, 컨텍스트 값 V_{cxt}와 결합함
- 가중치 λ로 조정함

-> Text Cross-attention와 가중치로 조정된 Context Cross-attention의 결과를 더하여 최종 출력 Fout 을

Method

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

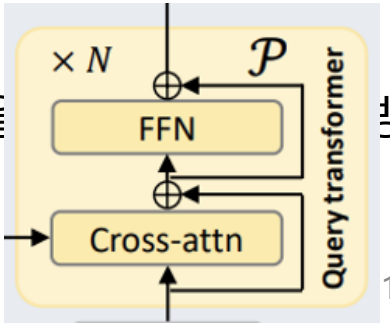
중간 레이어에서 λ 조절 실험

- λ 를 증가시키면 프레임 간 움직임이 줄어듦
- λ 를 감소시키면 객체의 형태를 유지하기 어려움
- 최적의 값을 찾아 텍스트 정보와 컨텍스트 정보를 균형 있게 반영하여, 프레임 간의 움직임과 객체의 형태를 일관되게 유지하도록 함

Training Paradigm

Training Paradigm

- 듀얼 스트림 통합: 이미지 정보는 컨텍스트 제어와 VDG(Visual Detail Guidance) 두 개의 스트림을 통해 결합됨.
- 세 단계 학습 전략:
    1. 이미지 컨텍스트 표현 네트워크 P 학습.
    2. P를 T2V 모델에 적응.
    3. VDG와 함께 fine-tuning 진행.
- 처음에는 가벼운 T2I 모델로 P를 학습하고, 이후 T2V 모델의 공간적 레이어와 공동으로 학습해 호환성을 맞춤.
- 프레임 무작위 선택 이유:
    1. 네트워크가 concat된 이미지를 특정 위치의 프레임에 매핑하는 shortcut을 방지하기 위함
    2. 컨텍스트 표현을 더 유연하게 하기 위함.



Experiments

Implementation Details

- Dataset: WebVid10M (256×256)
- Details:
  - T2V 모델 VideoCrafter와 T2I 모델 Stable-Diffusion-v2.1 (SD) 기반
  - P 학습: 100만 step, learning rate 1×10^{-4}, batch size: 64
  - P, T2V 모델학습: 30만 step fine-tuning, learning rate 5×10^{-5}, batch size: 64
- Inference:
  - DDIM 샘플러 사용.
  - Classifier-free guidance로 이미지와 텍스트 조건을 조정.

$$\begin{aligned}\hat{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}_{\text{img}}, \mathbf{c}_{\text{txt}}) = \epsilon_\theta(\mathbf{z}_t, \varnothing, \varnothing) \\ + s_{\text{img}}(\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_{\text{img}}, \varnothing) - \epsilon_\theta(\mathbf{z}_t, \varnothing, \varnothing)) \\ + s_{\text{txt}}(\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_{\text{img}}, \mathbf{c}_{\text{txt}}) - \epsilon_\theta(\mathbf{z}_t, \mathbf{c}_{\text{img}}, \varnothing))\end{aligned}$$

- Evaluation:
  - Perceptual Input Conformity (PIC): 입력 이미지와 동영상의 시각적 일치도 측정.

$$\text{PIC} = \frac{1}{L}\sum_{l}(1 - D(\mathbf{x}^{\text{in}}, \mathbf{x}^{l}))$$

Experiments

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**

- Frechet 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

Experiments

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

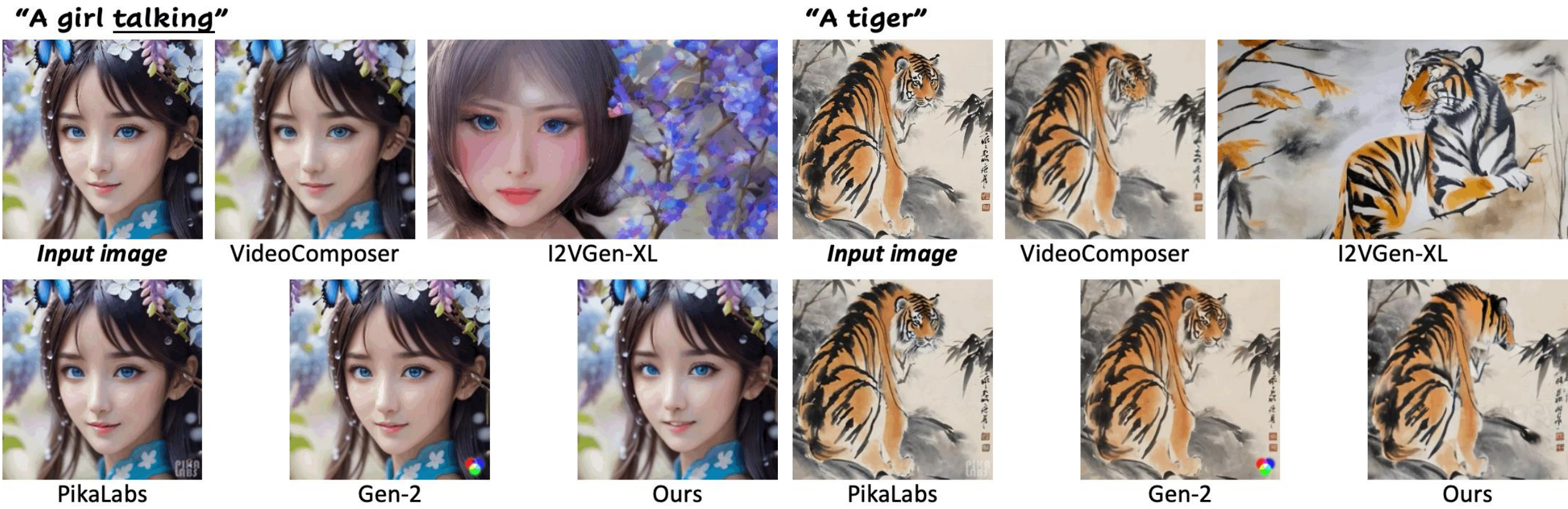
PikaLabs

Gen-2

Ours

Experiments

Qualitative Evaluation



Experiments

Qualitative Evaluation

Property	Proprietary		Open-source		
	PikaLabs	Gen-2	VideoComposer	I2VGen-XL	Ours
M.Q. ↑	28.60%	22.91%	2.09%	7.56%	**38.84%**
T.C. ↑	32.09%	26.05%	2.21%	6.51%	**33.14%**
I.C. ↑	79.07%	64.77%	18.14%	15.00%	**79.88%**

- Motion Quality, Temporal Coherence, Input Conformity에 대한 user study 결과

Experiments

Ablation Studies

- 듀얼 스트림 이미지 주입과 학습 패러다임에 대한 ablation 결과

Metric	Ours	Dual-stream image injection				Training paradigm	
		w/o ctx	w/o VDG	w/o λ	Ours$_G$	Ft. ent.	1st frame
FVD ↓	234.66	372.80	159.24	241.38	286.84	364.11	309.23
PIC ↑	0.5803	0.4916	0.6945	0.5708	0.5717	0.5564	0.5673

Experiments

Ablation Studies

- Stage 적용 비교 결과

Experiments

Ablation Studies

- 여러 학습 패러다임에 대하여 시각적으로 비교한 결과

Experiments

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

Conclusion

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

Applications

Applications

- 1. 스토리텔링: ChatGPT를 사용하여 스토리 스크립트와 해당 이미지를 생성한다. 그런 다음 DynamiCrafter를 사용하여 스토리 스크립트로 이미지 를 애니메이션화하여 스토리텔링 동영상을 생성할 수 있다.
- 2. 반복되는 동영상 생성: 약간의 수정을 통해 반복되는 동영상 생성을 용이하게 하도록 조정될 수 있다. 학습 중에는 x^1과 x^L 을 모두 VDG로 제공하고 다른 프레임은 비워둔다. Inference 시에는 둘 다 입력 이미지로 설정한다.
- 3. 프레임 보간: 입력 이미지 x^l과 x^L 을 다르게 설정하여 중간 프레임을 생성할 수 있다.

Fin

Thank you!