EURON 6기 고급팀 6주차 발표

AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning

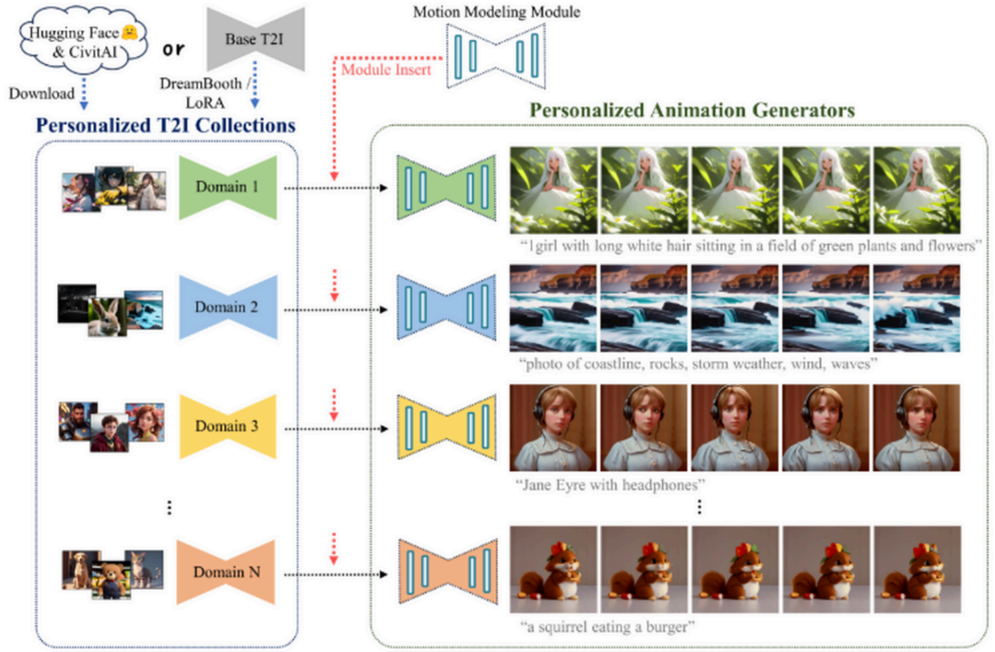
컴퓨터공학과 2070064 이현

목차

배경

〈 기존 text-to-image (T2I) 생성 모델의 발전 〉

- DreamBooth와 LoRA와 같은 몇 가지 경량의 개인화 방법이 제안됨
- 소규모 데이터셋에서 맞춤형 fine-tuning이 가능하고, 품질이 크게 향상된 맞춤형 콘텐츠를 생성할 수 있음
- 결과적으로 CivitAI나 Huggingface와 같은 모델 공유 플랫폼에서 수많은 개인화된 모델이 생성되었음



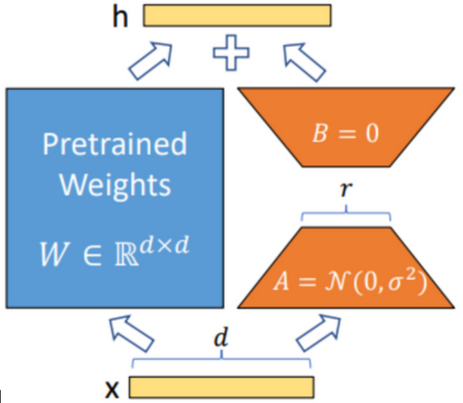
배경

〈 Personalizing T2I models 〉

- DreamBooth
  - 특정 객체나 스타일에 특화된 이미지 생성을 위해 사전 훈련된 생성 모델을 개인화하는 기술
  - 소수의 이미지를 사용하여 모델을 조정하면서도 원래의 특성을 유지

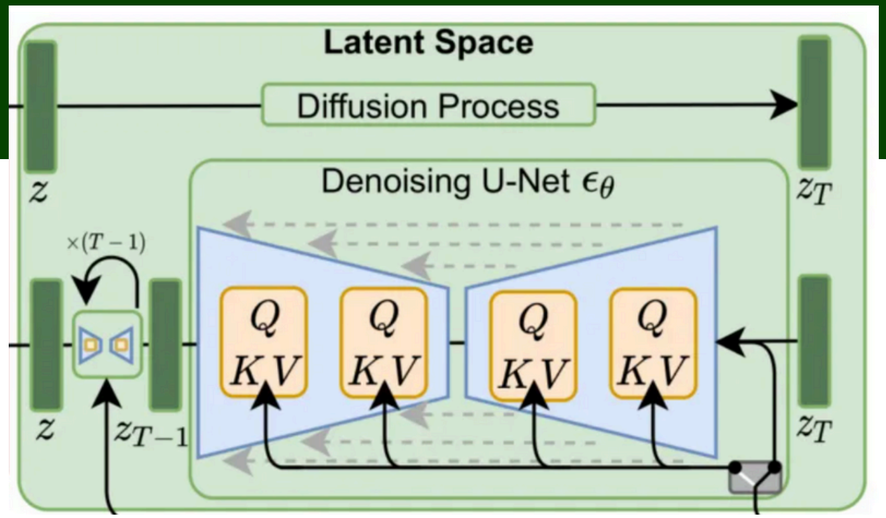
- Low-Rank Adaptation (LoRA)
  - 대규모 신경망 모델의 매개변수를 효율적으로 미세 조정하는 기술
  - 모델의 모든 매개변수를 재훈련하는 대신, 특정 층에 rank-decomposition 행렬 쌍을 추가

  (Pretrain Model weight 를 Freeze 한 상태로 (LoRA_b x LoRA_A layers) 행렬을 단순히 더함)
  - 이 새로운 가중치만을 최적화하여 기존의 학습된 특성을 유지하면서 새로운 데이터에 대해 미세 조정을 수행



배경

< Stable Diffusion >



- 텍스트로부터 이미지를 생성하는 오픈 소스 모델
- 사전 훈련된 오토인코더의 latent space에서 작동 (이미지를 인코딩하여 latent space의 표현(z0)을 얻음)
- 확산 과정: 훈련 동안, 잠재 이미지 z0는 여러 시간 단계 t=1,…,T에 걸쳐 점진적으로 교란되어

  zt라는 더 많은 노이즈가 포함된 일련의 버전을 생성 $z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon,\ \epsilon \sim \mathcal{N}(0, I)$
- 노이즈 제거: 이 구성 요소는 ϵθ(⋅)로 표현되며, 확산 과정을 역으로 수행하는 것을 학습. 이는 각 단계에서

  추가된 잡음 ϵ을 예측하여 최소화하는 평균 제곱 오차(MSE) 손실을 최소화하도록 훈련

$$\mathcal{L} = \mathbb{E}_{\mathcal{E}(x_0), y, \epsilon\sim\mathcal{N}(0,I), t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right]$$

- 이 과정을 통해, 모델은 노이즈를 첨가하고 이를 다시 예측하여 제거하는 방법을 배움

배경

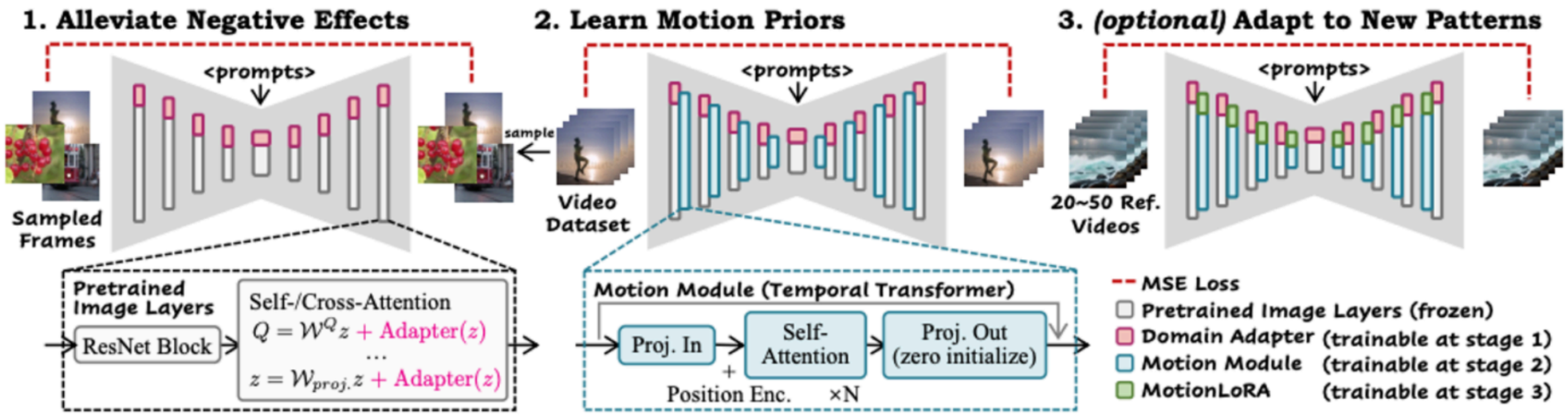
〈 한계점 〉

- 개인화된 T2I 모델의 출력은 정적인 이미지이다. 즉, 시간적 자유도가 부족하다.
- 사용자는 민감한 hyperparameter 튜닝, 개인화된 동영상 수집, 엄청난 계산 리소스를 감당할 수 없기 때문에 개인화된 T2I 모델을 사용하는 것은 어렵다.

〈 AnimateDiff 〉

- 기존의 개인화된 T2I 모델의 시각적 품질을 유지하면서 애니메이션 이미지를 생성하는 모델로 전환한다.
- 모델별 튜닝이 필요하지 않고 시간이 지남에 따라 매력적인 콘텐츠 일관성을 달성할 수 있다.

2. AnimateDiff 소개

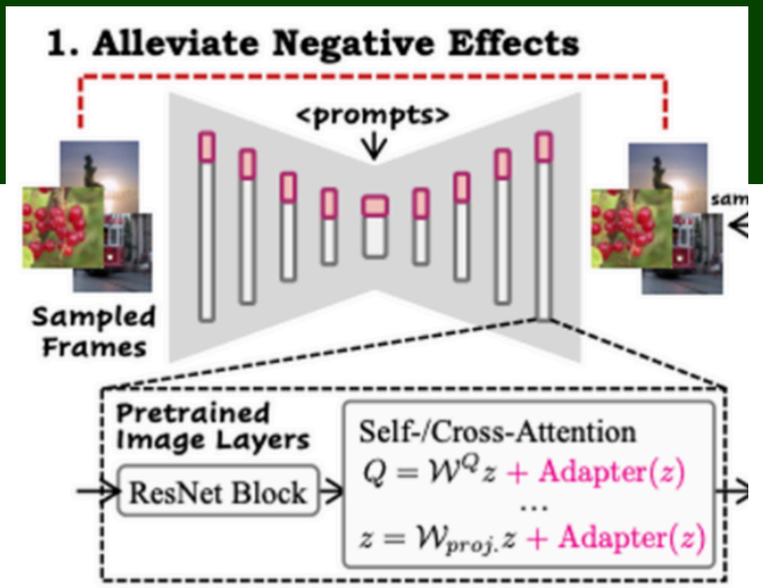


1. 도메인 어댑터: 훈련 중에만 사용되며, 기본 T2I 사전 훈련 데이터와 비디오 훈련 데이터 간의 차이에 의해 발생하는 부정적인 영향을 완화하는 역할

2. Motion 모듈: 비디오 데이터로부터 동작의 일반적인 패턴을 파악하여 학습.
훈련된 모션 모듈은 개인 맞춤형 T2I 모델에 통합됨

3. MotionLoRA : 선택적으로 사용되며, 동작 모듈을 더욱 특정한 동작 패턴에 맞추도록 미세 조정하는 과정.
적은 수의 참조 비디오만을 사용하여 새로운 동작 패턴에 적응시킬 수 있음.

3. AnimateDiff의 훈련 과정

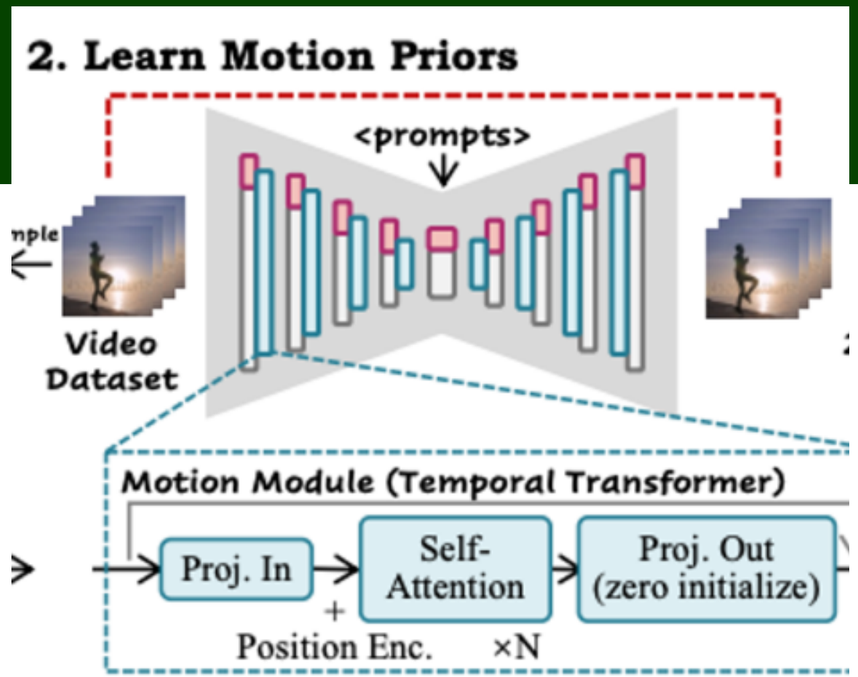


〈 도메인 어댑터를 사용하여 훈련 데이터의 부정적인 영향 완화 〉

- 훈련 데이터의 문제:
  - 비디오 데이터셋은 수집이 어렵고, 시각적 품질이 이미지 데이터셋에 비해 낮음.
  - 비디오 프레임 처리 시 모션 블러, 압축 아티팩트, 워터마크 등이 문제가 될 수 있음.
- 도메인 어댑터 :
  - 기본 T2I 모델의 지식을 보존, 도메인 간의 품질 차이로 인한 부정적인 영향을 줄이기 위해

    => 도메인 정보를 별도의 네트워크인 '도메인 어댑터'에 적합하도록 제안
  - 도메인 어댑터 층은 LoRA를 이용하여 구현되며, 기본 T2I의 self-/cross-attention layers 에 삽입

3. AnimateDiff의 훈련 과정



〈 MOTION PRIORS 학습과 MOTION MODULE 〉

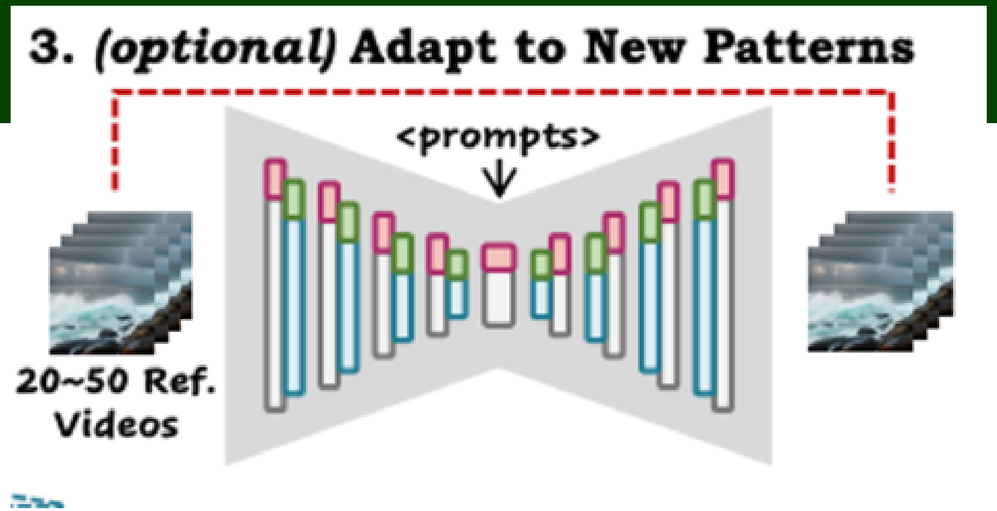
1. Network Inflation 네트워크 확장

- 2차원 확산 모델을 3차원 비디오 데이터를 다룰 수 있게 확장

2. 모션 모듈 설계

- 시간 축을 따라 효율적인 정보 교환을 가능하게 하는 서브 모듈을 설계
- Transformer 구조를 시간 축을 따라 작동하도록 수정하여 temporal Transformer로 만듦
- 애니메이션의 각 프레임 위치를 인코딩하기 위해 사인 위치 인코딩을 사용
- 인코딩된 벡터들은 self-attention 메커니즘을 사용하여 서로의 정보를 참조
- 초기에 모듈이 안정적으로 작동하도록 출력 레이어를 0으로 초기화하고, residual connection을 추가

3. AnimateDiff의 훈련 과정



〈 MotionLoRA를 이용하여 새로운 동작 패턴에 적응하기 〉

- MotionLoRA: 비디오에서 새로운 움직임 패턴을 학습하고 조정하는 기술
- 카메라의 줌, 이동, 회전 등을 처리하는 데 특히 유용
- 적은 수의 참조 비디오와 짧은 훈련 시간으로도 효과적으로 모델을 조정
- 기존 self-attention 레이어에 LoRA 레이어를 추가하여 프레임을 개선
- 이 레이어들은 새로운 움직임 패턴을 특별히 훈련하여 다양한 비디오 효과를 통합하고 조정
- 다양한 훈련된 모델을 결합하여 새로운 움직임 효과를 생성할 수 있는 능력도 가짐

4. AnimateDiff의 추론 과정

〈 Inference 추론 〉

학습된 모델을 사용해서 새로운 데이터에 대한 예측을 하는 과정

- 모델 확장: 개인화된 T2I 모델에 모션 모듈이 주입됨

  선택적으로 MotionLoRA가 개인화된 동작 생성을 위해 추가됨
- 도메인 어댑터 조정: 추론 시에도 도메인 어댑터를 모델에 주입하고,

  그 기여도를 조정하기 위해 스케일러 α를 변경
- 애니메이션 프레임 생성: 역 확산 과정을 통해 최종적으로 애니메이션 프레임이 생성됨

5. EXPERIMENTS

- Setting: Stable Diffusion V1.5를 기반으로 AnimateDiff를 구현 + WebVid10M 데이터셋으로 동작 모듈을 훈련

- 질적 양적 비교 실험을 진행
  - AnimateDiff는 Text2Video-Zero, Tune-a-Video, Gen-2, 그리고 Pika Labs와 같은 최신 도구들과 비교 분석
  - 양적 평가는 사용자 연구와 CLIP 지표를 통해 텍스트 정렬, 도메인 유사성, 동작의 부드러움을 중점적으로 비교
- ABLATIVE STUDY
  - 도메인 어댑터의 스케일러 조정이 시각적 품질 개선에 기여함을 확인
  - 동작 모듈 디자인 비교에서는 트랜스포머 기반 설계가 더 효과적임을 보였음
  - MotionLoRA의 효율성 검토 결과, 적은 데이터로도 효과적인 학습이 가능하나 참조 비디오가 제한적일 때는 성능 저하가 발생함을 확인하였음

6. 결론 및 ETHICS STATEMENT

〈 결론 〉

- 자연스러운 움직임과 높은 일관성을 가진 애니메이션 이미지를 성공적으로 생성하였음
- AnimateDiff는 추가 훈련 비용 없이 제어 가능한 생성을 지원하며,

  다양한 분야에서 사용될 수 있는 잠재력을 가지고 있음

〈 윤리적 고려 〉

- 생성 AI를 사용한 부정적인 콘텐츠 생성 및 허위 정보의 확산, 인간 관련 콘텐츠 생성 남용
- 위험 최소화 전략: 법적 틀을 준수, 개인 정보 보호를 존중, 긍정적인 콘텐츠 생성을 장려, 콘텐츠 안전 검사기 도입

감사합니다 :)