



# 1. AnimateDiff: Animate Your Personalized Text-To Image Diffusion Models without Specific Tuning



## AnimateDiff

- 별다른

모델별 튜닝 없이 개인화된 Text to Image(T2I) 모델을 애니메이션화하는 것을 목표로 하는 프레임워크

- 실제 비디오에서 이동 가능한 모션 우선순위를 학습  
→ 모든 개인화된 Text to Image 모델에 적용

- 또한,

**MotionLoRA**라는 경량 파인 튜닝 기술을 제안

⇒ 새로운 모션 패턴에 쉽게 적응할 수 있도록 함

## 1. Introduction

- 고품질의 개인화된 T2I 모델을 애니메이션 생성기로 직접 변환하는 **AnimateDiff**를 제안
  - 이를 위해 훈련된 모션 모듈을 사용하여 개인화된 T2I에 부드럽고 시각적으로 매력적인 애니메이션을 생성
- **AnimateDiff**의 핵심
  - 비디오 데이터셋에서 합리적인 **모션 우선순위**를 학습하는 플러그 앤 플레이 모션 모듈
  - 또한, **MotionLoRA**라는 경량 파인 튜닝 기술을 사용하여 사전 훈련된 모션 모듈을 새로운 모션 패턴에 적응시킴
- 실험 결과

- AnimateDiff 및 MotionLoRA가 다양한 개인화된 T2I 모델에서 유망한 결과를 제공
- Transformer 아키텍처가 모션 우선순위를 모델링하는 데 적절하다는 것을 입증  
⇒ 특정 상황에 맞춘 fine-tuning 없이도 모든 개인화된 T2I 모델의 애니메이션 생성 능력을 활성화(범용적이다~)

## 2. Related Work

### Text-to-Image Diffusion Models

- GLIDE: 텍스트 조건을 도입하고 분류기 안내를 통합하는 것이 더 만족스러운 결과를 낳는다는 것을 입증
- DALL-E 2: CLIP의 공동 특성 공간을 활용하여 텍스트-이미지 정렬을 개선
- Imagen: 대규모 언어 모델과 카스케이드 아키텍처를 결합하여 사실적인 결과를 달성
- 잠재 확산 모델(Latent Diffusion Model): 확산 프로세스를 오토인코더의 잠재 공간으로 이동하여 효율성을 향상
- eDiff-I: 다양한 생성 단계에 특화된 확산 모델 양상을 사용

### Personalizing T2I models

- 사전 훈련된 T2I를 사용한 창작을 용이하게 하기 위해 많은 연구가 효율적인 모델 개인화에 집중하고 있음
  - 기본 T2I에 개념이나 스타일을 참조 이미지를 사용하여 도입하는 것을 의미
- 이를 달성하기 위한 가장 직접적인 접근 방법은 모델의 완전한 fine-tuning임
  - 전반적인 품질을 크게 향상시킬 수 있는 잠재력이 있지만, 참조 이미지 세트가 작을 때 catastrophic forgetting을 초래할 수 있다는 문제가 존재
    - ▼ catastrophic forgetting
      - 새로운 데이터를 학습할 때 이전에 학습한 정보를 완전히 잊어버리는 것을 의미
      - 주로 신경망이나 딥러닝 모델에서 나타나며, 이전 작업에 대한 정보가 새로운 작업을 수행하는 동안 손실되어 이전 작업의 성능이 저하될 수 있음
      - 이를 해결하기 위한 방법으로는 고정된 가중치를 사용하여 이전 정보를 보존하거나, 경험적 손실을 최소화하여 새로운 정보를 효율적으로 통합하는

방법이 있음

- 대신에, DreamBooth는 보존 손실로 전체 네트워크를 미세 조정하고 소수의 이미지만 사용
- 텍스트 역변환은 각 새로운 개념에 대해 토큰 임베딩을 최적화함
  - 저랭크 적응(Low-Rank Adaptation, LoRA)는 기본 T2I에 추가적인 LoRA 레이어를 도입하여 가중치 잔차만 최적화함으로써 위의 미세 조정 과정을 용이하게 함
- 또한, 개인화 문제를 해결하는 인코더 기반 접근 방법도 있음

## Animating personalized T2Is

- 기존 연구 대부분은 모든 매개변수를 업데이트하고 원래 T2I의 특징 공간을 수정하는 것에 초점 → 개인화된 모델과는 거리가 좀 있음
  - Text2Cinemagraph: 흐름 예측을 통해 영화를 생성하는 것을 제안
  - Align-Your Latents: 일반적인 비디오 생성기에서 고정된 이미지 레이어를 개인화 할 수 있다는 것을 입증
- 최근에는 개인화된 T2I 모델을 애니메이션화하는 데 유망한 결과를 보여주는 비디오 생성 접근 방법이 존재
  - Tune-a-Video: 단일 비디오에서 소수의 매개변수를 미세 조정
  - Text2Video-Zero: 사전 정의된 affine 행렬을 기반으로 사전 훈련된 T2I를 애니메이션화하는 데 사용되는 잠재 래핑을 통한 훈련 방법을 소개

## 3. Preliminary

keywords) Stable Diffusion, Low-Rank Adaptation (LoRA)

### Stable Diffusion

- 사전 훈련된 오토인코더  $\epsilon(\cdot)$ 와  $D(\cdot)$ 의 잠재 공간 내에서 확산 프로세스를 수행
  - 훈련 중에 인코딩된 이미지  $z_0 = \epsilon(x_0)$ 는 forward diffusion을 통해  $z_t$ 로 변형

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, I),$$

encoded image:  $\epsilon(x_0)$   
~~~~~  
~~~~~  
 $\bar{\alpha}_t$  noise strength

- 이 과정을 반전시키기 위해 노이즈를 추가하는 방법을 학습하는 노이즈 제거 네트워크  $\epsilon_\theta(\cdot)$ 는 추가된 노이즈를 예측하도록 장려되며, MSE 손실에 의해 촉진됨

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

- $\epsilon_\theta(\cdot)$ 이 네트워크 블록으로 구현되어 있음
- 네트워크 블록은 ResNet, 공간 자기 주의 계층, 그리고 텍스트 조건을 도입하는 교차 주의 계층으로 구성된 U-Net을 포함

## Low-rank Adaptation(LoRA)

- 대규모 모델의 미세 조정을 가속화하는 방법
  - 모든 모델 매개변수를 재훈련하는 대신 랭크 분해 행렬의 쌍을 추가하고, 새롭게 도입된 가중치만 최적화
  - 훈련 가능한 매개변수를 제한하고 원래의 가중치를 고정시킴으로써, LoRA는 재앙적인 잊혀짐 문제의 발생 가능성성이 낮음
- 구체적으로, 랭크 분해 행렬은 사전 훈련된 모델 가중치  $W \in \mathbb{R}^{m \times n}$ 의 잔차로 작용
  - 새로운 모델 가중치

$$\mathcal{W}' = \mathcal{W} + \Delta\mathcal{W} = \mathcal{W} + AB^T,$$

$\mathcal{W}$ : pre-trained model weight,  $\mathcal{W} \in \mathbb{R}^{m \times n}$   
 $A$ : rank decomposition

- $A \in \mathbb{R}^{m \times r}$ ,  $B \in \mathbb{R}^{n \times r}$
- $r$ : 하이퍼 파라미터, LoRA 레이어의 rank
- LoRA는 어텐션 계층에만 적용되어 모델 미세 조정의 비용과 저장 공간을 더욱 줄임

## 4. AnimateDiff

★ 비디오 데이터로부터 전이 가능한 모션 우선순위를 학습하자.

→ 특정 조정 없이 개인화된 T2I에 적용될 수 있음

## 4-1. Alleviate Negative Effects From Training Data with Domain Adapter

- 기본 T2I를 훈련하는 데 사용된 고품질 이미지 데이터셋과 우리가 모션 우선순위를 학습하는 데 사용하는 대상 비디오 데이터셋 간에는 무시할 수 없는 품질 도메인 간격이 존재
  - 불가피한 성능 저하가 발생한다는 의미
- 품질 도메인 간격을 피하기 위해 별도의 도메인 어댑터를 제안
  - LoRA를 사용하여 구현, 기본 T2I의 셀프/교차 어텐션 레이어에 삽입  
⇒ 기존 T2I 모델의 지식을 보존
  - 품질 차이를 학습하지 않고도 모션 모듈을 개선 가능
    - ex) 쿼리(Q) 프로젝션

$$Q = \mathcal{W}^Q z \xleftarrow{\text{feature}} + \text{AdapterLayer}(z) = \mathcal{W}^Q z + \underbrace{\alpha}_{\text{solar}} \cdot AB^T z,$$

- 도메인 어댑터는 추론 시에 삭제됨

## 4-2. Learn Motion Priors with Motion Module

- 사전 훈련된 T2I 모델 위에 모션 다이내믹스를 모델링하기 위해 필요한 두 가지 과정이 있음
  - 1) 2차원 확산 모델을 3차원 비디오 데이터로 확장
  - 2) 시간 축을 따라 효율적인 정보 교환을 가능하게 하는 하위 모듈 설계
- 네트워크 확장
  - 이미지 레이어가 독립적으로 각 비디오 프레임을 처리할 수 있도록 모델을 수정하는 방법을 소개
  - 이를 위해 비디오 텐서를 5차원 형식으로 변환하고, 이미지 레이어를 통과할 때 시간 축을 무시하여 각 프레임을 독립적으로 처리하도록 하였음
- 모듈 설계
  - 시간적 모델링을 위해 Transformer 아키텍처를 채택하고 시간 축에 맞게 조정하는 것이 적합함을 확인

$$z_{out} = \text{Attention}(Q, K, V) = \text{Softmax}(QK^T / \sqrt{c}) \cdot V,$$

$\underbrace{\quad\quad}_{\text{---}} \quad \underbrace{\quad\quad}_{\text{---}}$  (분류된) projection

- 이러한 변화로 인해 T2I 모델은 애니메이션 클립의 모션 다이내믹스를 캡처할 수 있게 되었음
- 또한, 모션 모듈이 애니메이션의 프레임 순서를 파악할 수 있도록 사인 함수 위치 인코딩이 필수적이며, 추가된 모듈이 무해한 효과를 일으키지 않도록 출력 투사 레이어를 영점으로 초기화하고 잔차 연결을 추가하는 것이 중요

## 4-3. Adapt to New Motion Patterns with MotionLoRA

- MotionLoRA는 사전 훈련된 모션 모듈을 새로운 모션 패턴에 효과적으로 조정하기 위한 효율적인 미세 조정 방법임
  - 적은 참조 비디오와 훈련 반복으로도 우수한 결과를 얻는 것이 가능
  - 이를 통해 모델 조정 및 공유가 가능하며, 합성 능력도 갖추고 있음

## 4-4. AnimateDiff In Practice

### Training

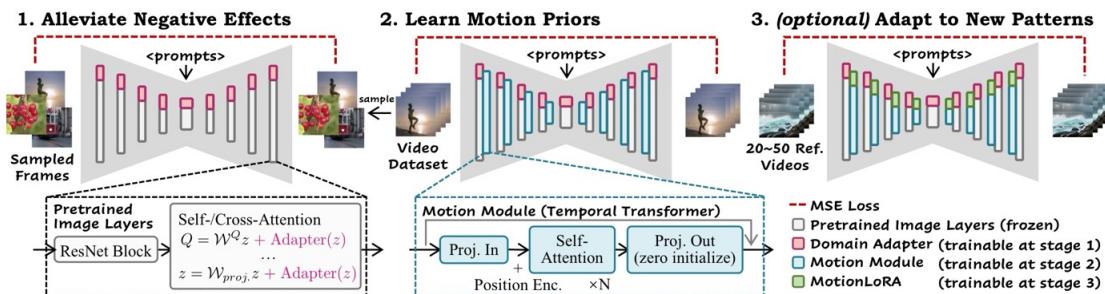


Figure 3: **Training pipeline of AnimateDiff.** AnimateDiff consists of three training stages for the corresponding component modules. Firstly, a domain adapter (Sec. 4.1) is trained to alleviate the negative effects caused by training videos. Secondly, a motion module (Sec. 4.2) is inserted and trained on videos to learn general motion priors. Lastly, MotionLoRA (Sec. 4.3) is trained on a few reference videos to adapt the pre-trained motion module to new motion patterns.

### ★ Latent Diffusion Model과 유사

- sampling 된 동영상 데이터  $x_0^{1:N}$ 은 먼저 사전 학습된 autoencoder를 통해 프레임 별로 latent code  $z_0^{1:N}$ 으로 인코딩 됨
- 이후 forward diffusion schedule을 사용하여 latent code에 noise 추가
  - 모션 모듈로 확장된 diffusion network는 noisy한 latent code와 텍스트 프롬프트를 입력으로 사용하고, L2 loss 항에 의해 latent code에 추가되는 noise 강도

를 예측

- 최종 목적 함수

$$\mathcal{L} = \mathbb{E}_{\substack{\mathcal{E}(x_0^{1:f}), y, \epsilon^{1:f} \sim \mathcal{N}(0, I), t \\ \text{sampling 된 동영상 데이터}}} \left[ \|\epsilon - \epsilon_\theta(z_t^{1:f}, t, \tau_\theta(y))\|_2^2 \right].$$

by 사전 학습된 AutoEncoder

latent code

## Inference

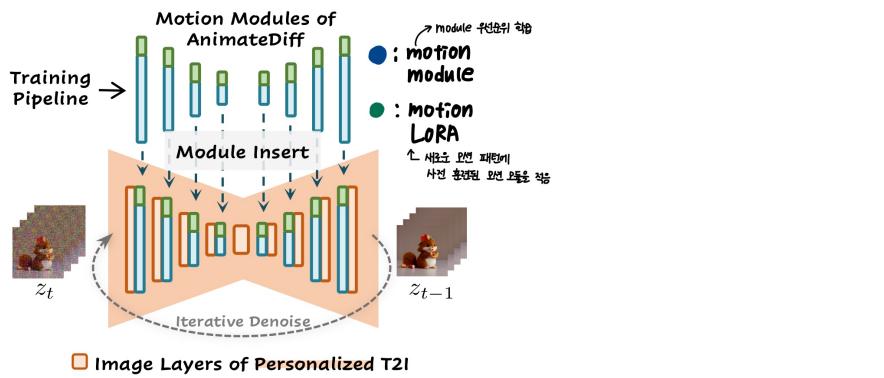


Figure 2: Inference pipeline.

- 추론 시 개인화된 T2I 모델은 확장을 거친 다음, 일반 애니메이션 생성을 위해 모션 모듈로 주입되며, 선택적으로 MotionLoRA가 개인화된 모션을 갖는 애니메이션 생성을 위해 사용됨
- 단순히 추론 시 도메인 어댑터를 삭제하는 대신, 실제로는 개인화된 T2I 모델에 주입하고 방정식 (4)에서 스케일러  $\alpha$ 를 변경하여 기여하는 정도를 조정 가능

$$Q = \mathcal{W}^Q \overset{\text{feature}}{\check{z}} + \text{AdapterLayer}(z) = \mathcal{W}^Q z + \overset{\text{scalar}}{\check{\alpha}} \cdot AB^T z,$$

Equation (4)

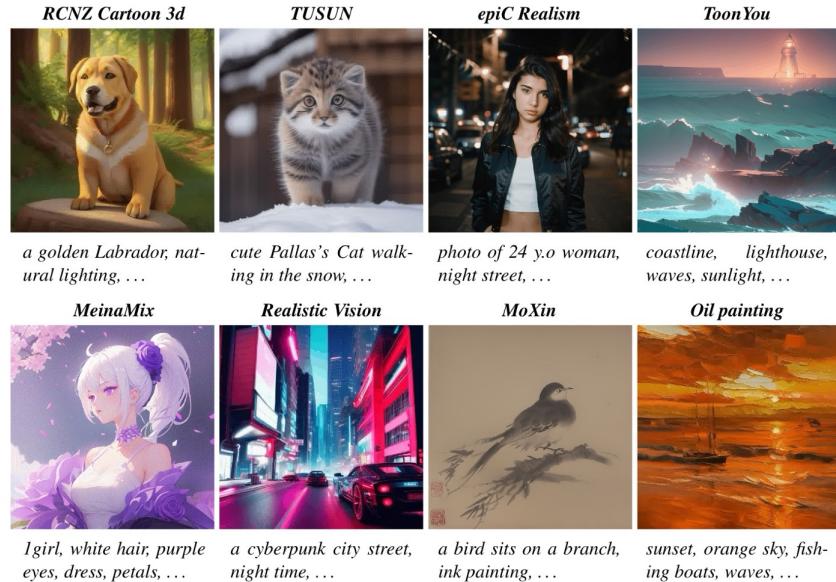
- 마지막으로, 애니메이션 프레임은 역(reverse) 확산 프로세스를 수행하고 잠재 코드를 디코딩함으로써 얻을 수 있음

## 5. Experiments

- [Stable Diffusion V 1.5](#)로 구현
- WebVid-10M 데이터셋을 활용하여 모션 모듈 구현

## 5-1. Qualitative Results

- Civitai(2022)에서 수집한 다양한 개인화된 T2I 모델을 사용하여 평가되었음



- 다양한 도메인을 포함, 종합적인 benchmark로 가능
- 각 sample은 고유하게 개인화된 T2I에 해당함
- MotionLoRA를 사용하여 few-shot 유형 제어를 달성한 결과도 제시
  - 성능 비교를 위한 baseline
    - Text2Video Zero, Tune-a-Video, Gen-2, Pika Labs

(pdf 아크로뱃으로 열어서 동영상 motion을 확인해 보면 AnimateDiff로 생성된 영상들이 훨씬 배경 왜곡이나 모션의 연속성 등이 더 자연스럽다.)

## 5-2. Quantitative Comparison

✓ 텍스트 정렬, 도메인 유사성, 모션 부드러움

Table 1: Quantitative comparison. A higher score indicates superior performance.

Method	User Study (↑)			CLIP Metric (↑)		
	Text.	Domain.	Smooth.	Text.	Domain.	Smooth.
Text2Video-Zero	1.620	<b>2.620</b>	1.560	32.04	84.84	96.57
Tune-a-Video	2.180	1.100	1.615	<b>35.98</b>	80.68	97.42
<b>Ours</b>	<b>2.210</b>	2.280	<b>2.825</b>	31.39	<b>87.29</b>	<b>98.00</b>

- 사용자 연구 및 CLIP 메트릭을 사용하여 AnimateDiff와 기타 두 방법 간의 정량적 비교를 실시하였음
  - 사용자 연구에서는 참가자들이 각 방법으로 생성된 애니메이션을 위 세 가지 측면을 기준으로 순위를 매기도록 요청했으며, CLIP 메트릭은 애니메이션 프레임과 개인화된 T2I를 사용하여 생성된 참조 이미지 사이의 유사성을 측정하였음

## 5-3. Ablative Study

- 도메인 어댑터
  - AnimateDiff의 도메인 어댑터를 조사하기 위해, 스케일러를 조정하여 어댑터 레이어의 영향을 추론 중에 조절하는 연구를 수행
   
→ 어댑터의 스케일러가 감소함에 따라 시각적 품질이 향상되는 것을 확인
- 모션 모듈 설계
  - 시간적 트랜스포머를 사용한 모션 모듈 디자인을 컨볼루션 상대 디자인과 비교
  - 컨볼루션 모듈은 정렬은 하지만 모션을 포함하지 않음을 확인
- MotionLoRA의 효율성
  - MotionLoRA의 효율성을 매개변수 효율성과 데이터 효율성 측면에서 조사
   
→ 작은 매개변수 스케일과 적은 참조 비디오로도 원하는 모션 패턴을 학습하는 능력을 확인
  - 그러나 참조 비디오의 수가 제한되면 품질 저하가 관찰되었음

## 5-4. Controllable Generation

- AnimateDiff를 ControlNet과 결합하여 추출된 깊이 맵 시퀀스로 생성을 제어하였음
    - 최근의 비디오 편집 기술과 대조적으로, 무작위로 샘플링된 노이즈에서 애니메이션을 생성하였음
- ⇒ 섬세한 모션 세부 사항(ex. 머리카락과 표정)과 높은 시각적 품질을 보임

## 6. Conclusion

- 해당 논문에서는 AnimateDiff를 소개
  - 개인화된 텍스트 대 이미지(T2I) 모델을 애니메이션 생성을 위한 실용적인 파이프라인으로 직접 변환

- 품질을 저하시키지 않으면서 사전에 학습된 도메인 지식을 유지하고, 의미 있는 모션 선행 지식을 학습하며, 가볍고 효과적인 파인 튜닝 기법인 MotionLoRA를 통해 모션을 개인화할 수 있음
  - AnimateDiff는 다양한 개인화된 T2I 모델을 사용하여 폭넓은 평가를 거쳐 효과적이고 일반적으로 적용 가능함을 검증
  - 또한, 기존의 콘텐츠 제어 방법과의 호환성을 보여줌으로써 추가 교육 비용을 발생시키지 않고 제어 가능한 생성이 가능
- ⇒ 개인화된 애니메이션에 대한 효과적인 기준을 제공하며 다양한 응용 분야에서 활용 가능