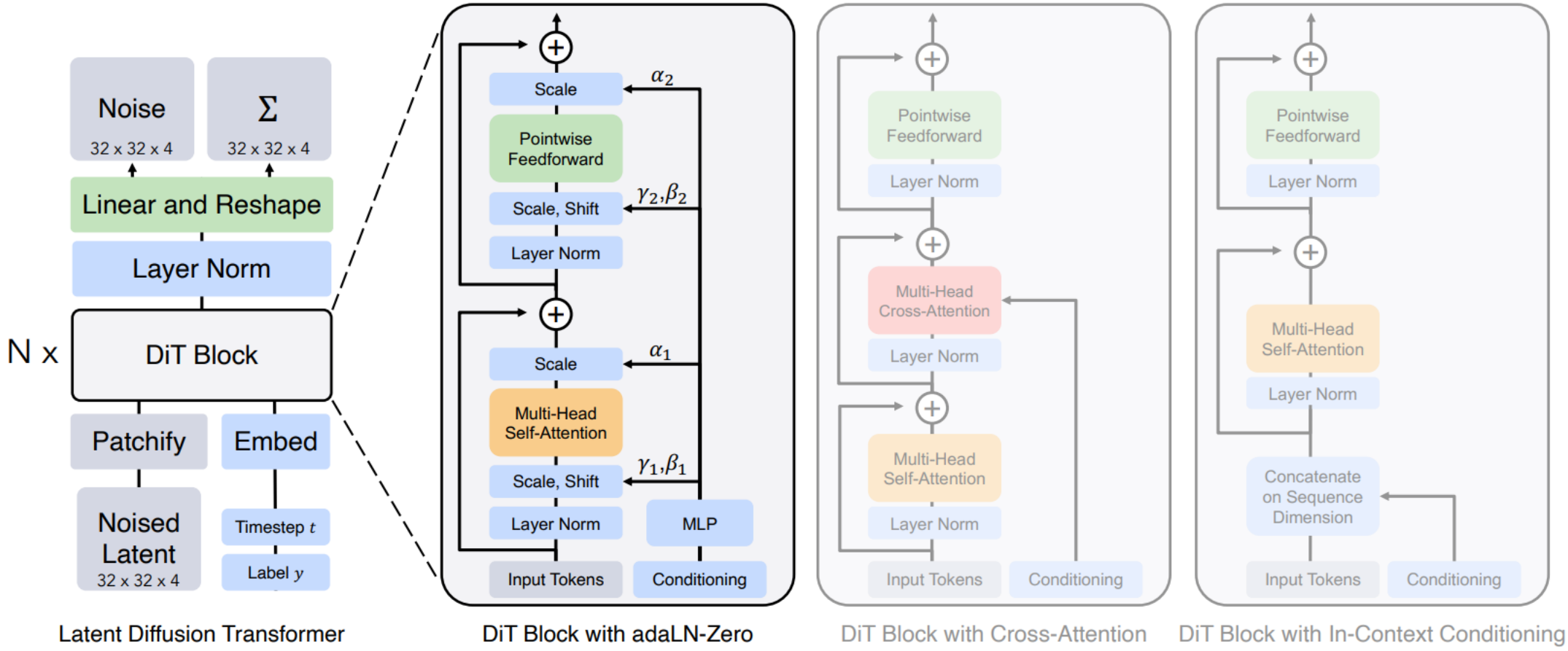
AI 실전 12주차

Computer Vision - Text to Video

DiT (Diffusion Transformer)



Latent Diffusion Transformer | DiT Block with adaLN-Zero | DiT Block with Cross-Attention | DiT Block with In-Context Conditioning

CogVideoX

Text Prompt: Push upward at a low angle, slowly look up, an evil dragon suddenly appears on the iceberg, and then the dragon spots you and rushes towards you. Hollywood movie style

Text Prompt: An old-fashioned automobile drives through the streets of the Republic. While driving right in the middle of it, bombs suddenly fall from the sky, the car is blown up, the people in the car are blown up, the screen shakes, the movie winds up

CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer

CogVideoX 구조

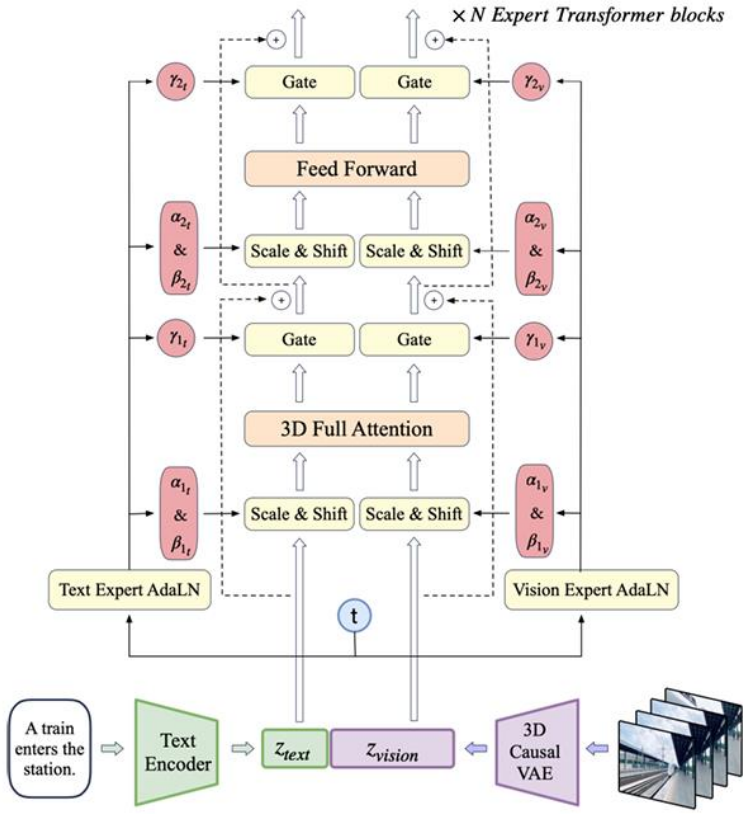


Figure 3: **The overall architecture of CogVideoX.**

- DiT 구조를 이용한 이미지 생성 + 시간
  - 시간 개념을 구하기 위해 하나의 이미지를 **3D VAE Encoder 처리를 함.**

3D Causal VAE

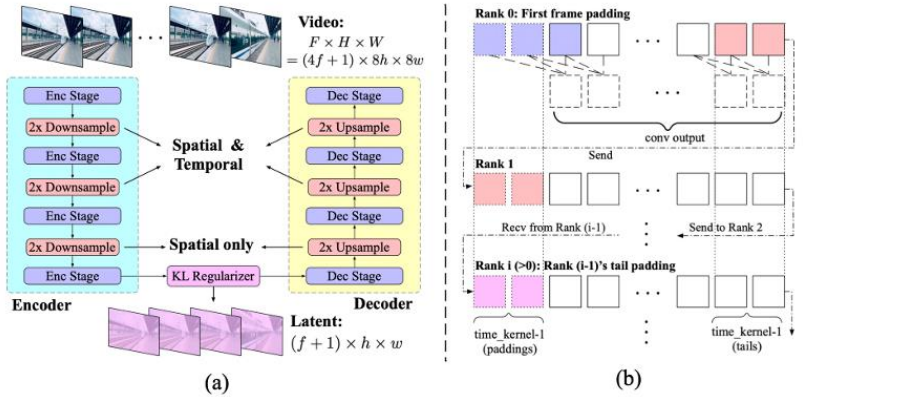
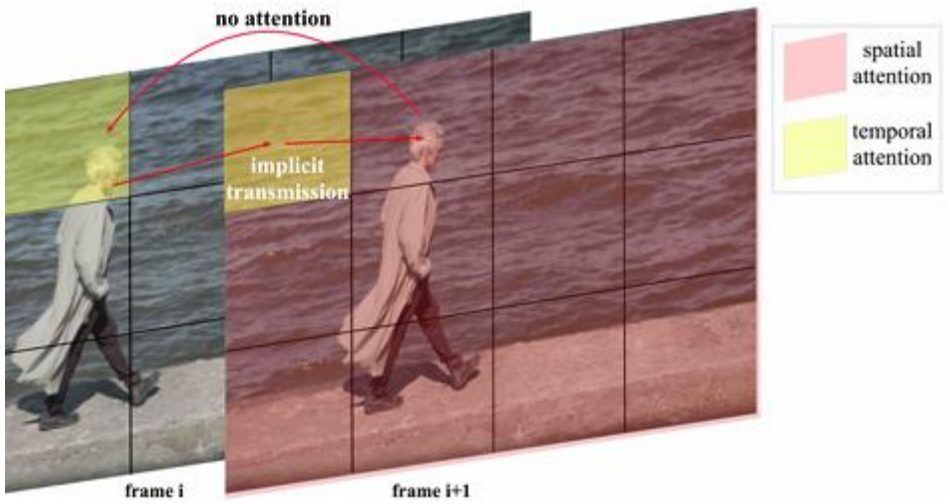


Figure 4: (a) The structure of the 3D VAE in CogVideoX. It comprises an encoder, a decoder and a latent space regularizer, achieving a 8×8×4 compression from pixels to the latents. (b) The context parallel implementation on the temporally causal convolution.

- 기본적으로, Encoder와 Decoder에서 보면 **2x Downsampling**
  - Spatial은 8배
  - temporal은 4배 줄어드는 구조
- **<u>현재와 과거에 영향을 적게 함.(많이 변화할 필요x)</u>**

Expert Transformer - **3D Full Attention**



- 현재, 다음 프레임 각 영역별로 동시 계산 적용

Mixed Training

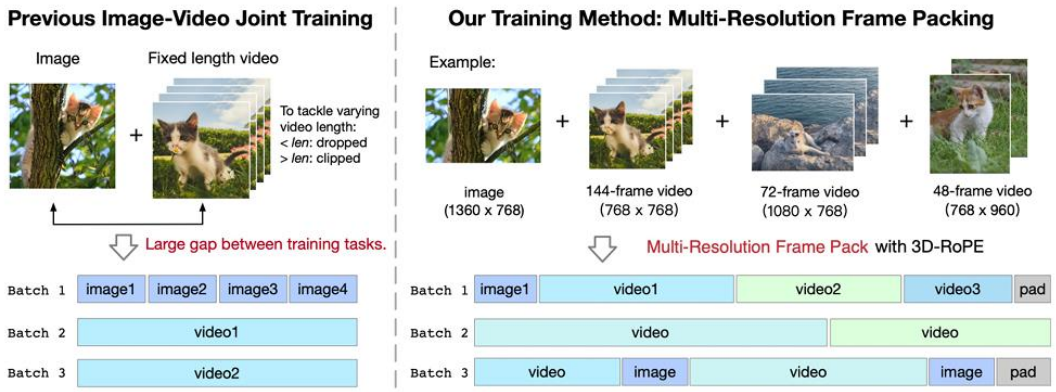
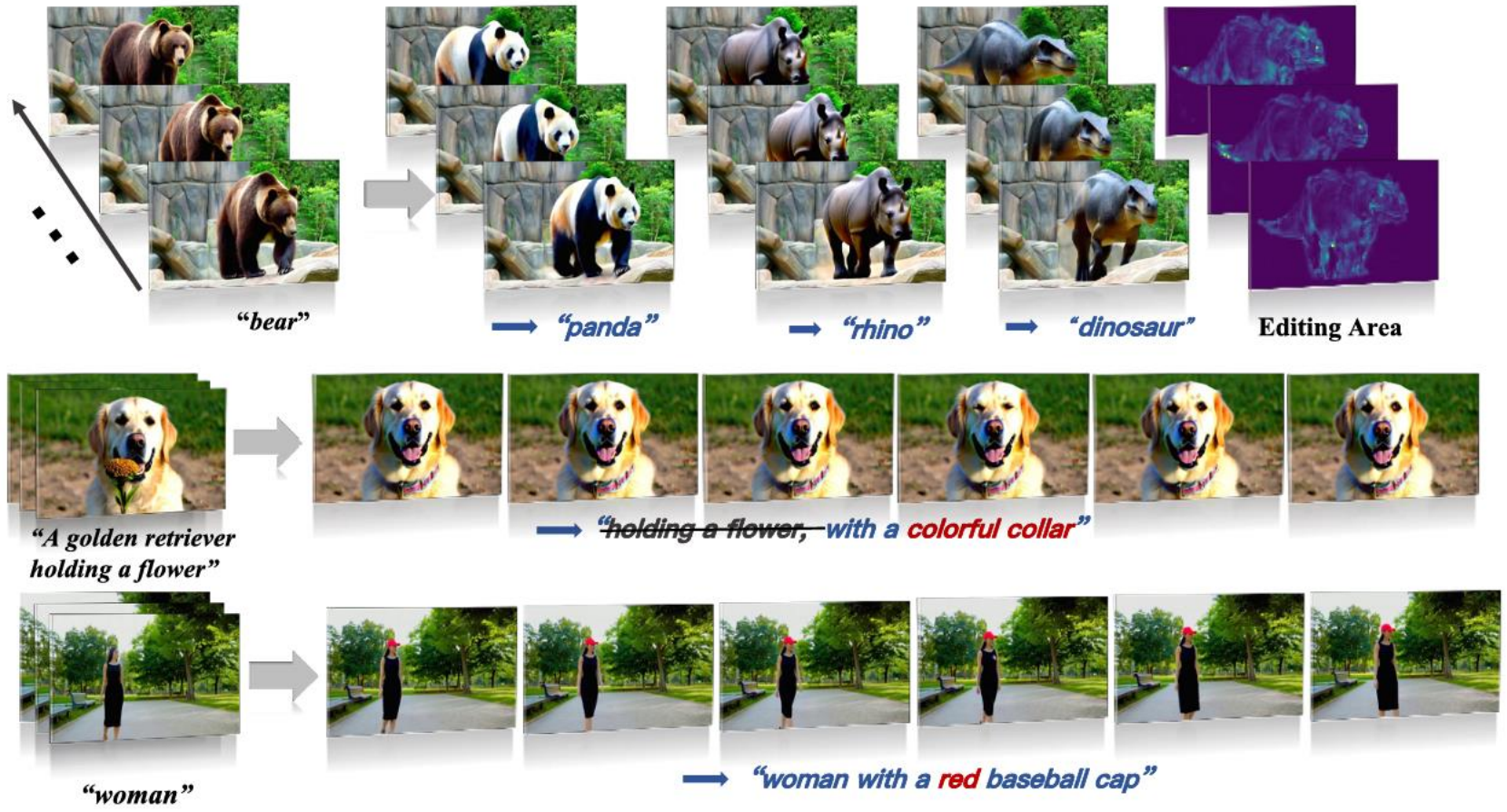


Figure 6: The diagram of mixed-duration training and Frame Pack. To fully utilize the data and enhance the model's generalization capability, we train on videos of different duration within the same batch.

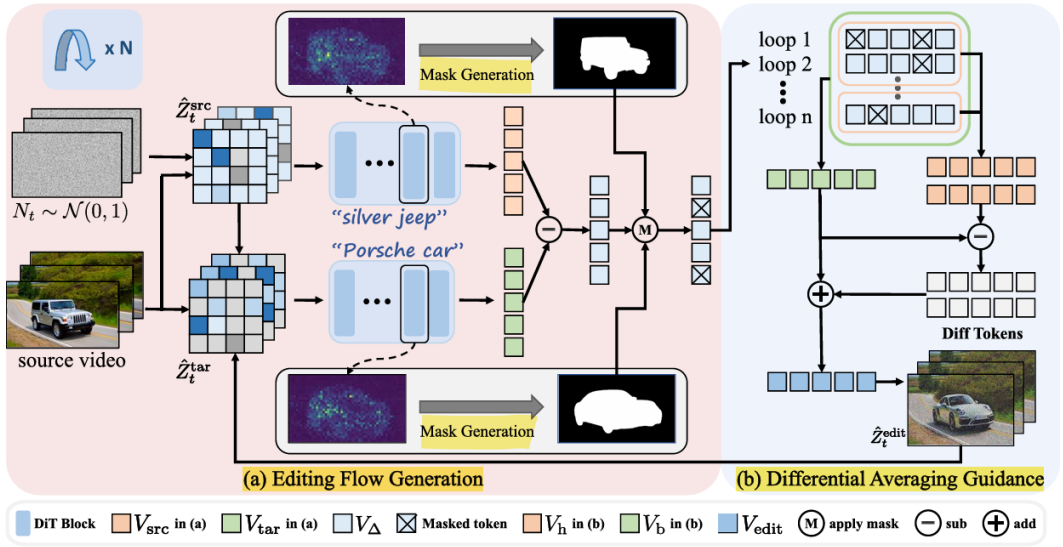
- 이미지, 영상 생성에 차이를 줄이기 위한 다른 영상 이미지들을 연결하여 학습 방안 제시
  - 이전 방식으로 공간을 동일하게 하며, 큰 이미지는 자르기 처리
  - 공간이 작은 이미지는 패딩 처리
  - 또한 이미지 크기 점차 크게 적용

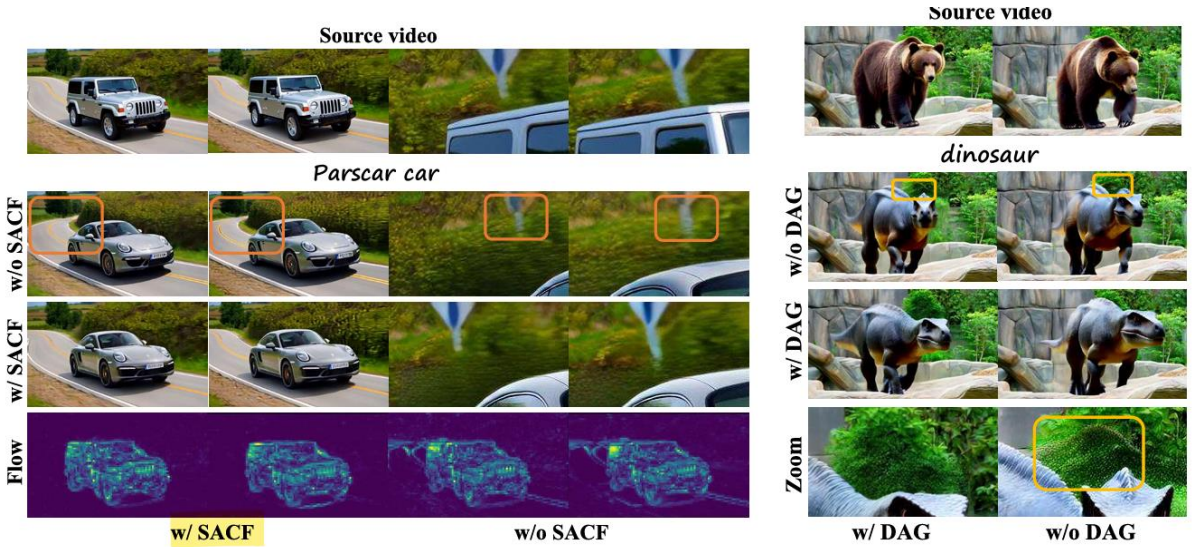
FlowDirector



FlowDirector: Training-Free Flow Steering for Precise Text-to-Video Editing

FlowDirector 구조



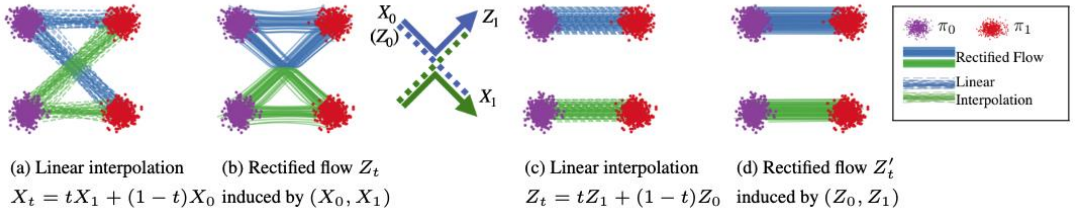


- **attention masking**
  - 해당 객체의 물체를 마스킹 처리함. Attention(Query, Key, Value) 내용들의 계산으로 구해짐.
  - Object Text, Attention shape를 동일하게

differential averaging guidance

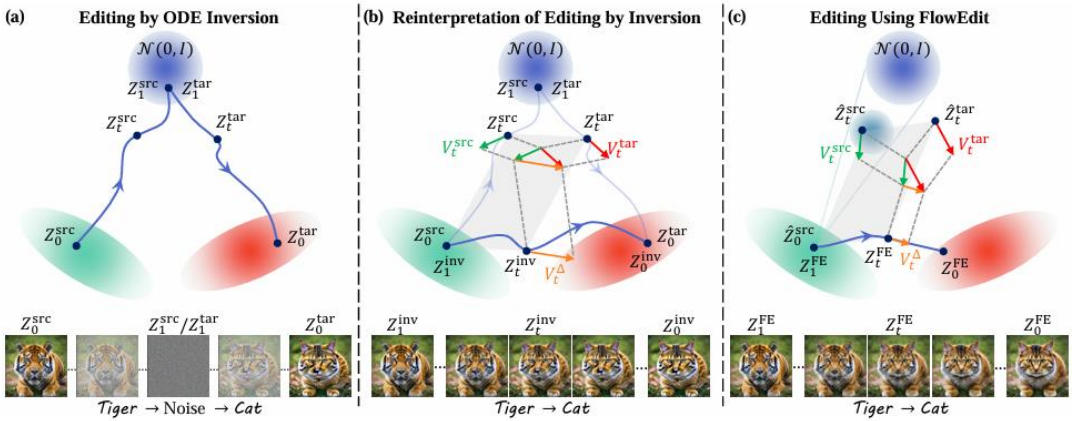
- 이미지 생성하고 나온 과정(Denoise처리 유사, 속도)를 평균화 처리(쉽게 말해 모션블러)

이미지 생성 잘하기 위한 방안



- 기존의 Diffusion 모델에서 Denoise 하는 과정들을 구하는 것을 목표로 함.
- FlowDirector에서 적용한 것으로 **Rectified flow 기법으로, Denoise 하는 과정(velocity, 속도)를 학습하는 방안 제시**
- 이 기법을 FLUX 모델에서 채택

이미지를 생성를 잘 하기 위한 방안



- **Random noise start**
  - 노이즈를 초기화 할 때 대상의 이미지(source image)를 대상으로 노이즈 초기화
  - 연관성 없이 이미지 생성
- **Noise Inversion**
  - 이미지 노이즈를 연관성 있게 학습 과정으로 적용한 것 정확도 높으나 속도 느림.
- **Inversion-Free**
  - 이미지 연관성 학습 없이도 성능 좋게 나옴!