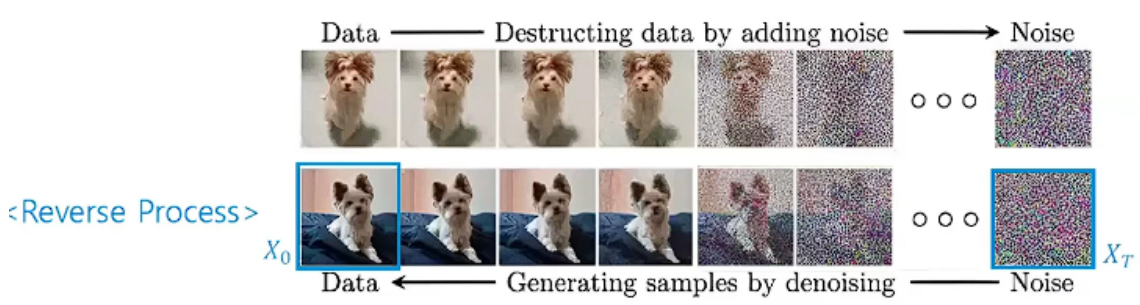
08_Explaining Diffusion based Generative Models

디퓨전 모델에 대한 분석과, 그 분석을 통한 이미지 생성을 위한 모델의 수정 방법들에 대해서 알아보자.

1. How does “Diffusion model” work?

디퓨전 모델은 이미지 데이터가 주어졌을 때, 이미지에 노이즈를 점차 더해가는 forward process를 역으로 뒤집은 Reverse Process를 학습하여 노이즈로부터 이미지를 생성하는 모델이다.

- 위쪽 과정 (Destructing data by adding noise)
  1. 원본 이미지에서 점진적으로 노이즈 추가
  2. 최종적으로 완전한 노이즈가 된 상태 X_T 도달
- 아래쪽 과정 (Generating samples by denosing / Reverse Process)
  1. 완전한 노이즈 X_T에서 점진적으로 노이즈 제거
  2. 원래의 깨끗한 데이터 X_0로 복원

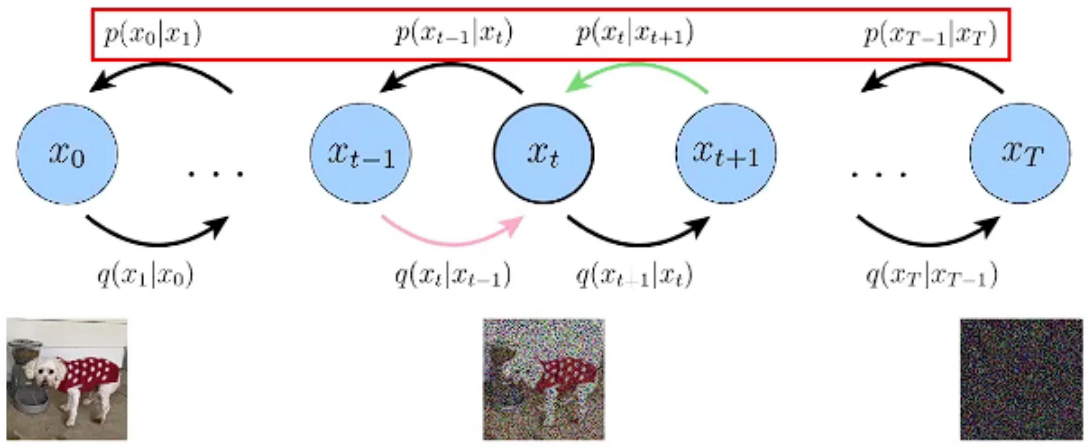


Markov Chain Process

디퓨전 모델은 마르코프 체인을 기반으로 작동하며, 확률적 과정으로 표현된다.

아래의 그림을 살펴보자.

- $q(x_t|x_{t-1})$: 이전 상태 x_{t-1}에서 현재 상태 x_t로 변환하는 과정 (노이즈 추가 과정)
- $p(x_{t-1}|x_t)$: 현재 상태 x_t에서 이전 상태 x_{t-1}를 예측하는 과정 (노이즈 제거 과정)
- **각 단계 x_t 마다 확률적으로 변환**되며, 전체적으로 확률 분포를 학습하는 과정을 가짐.

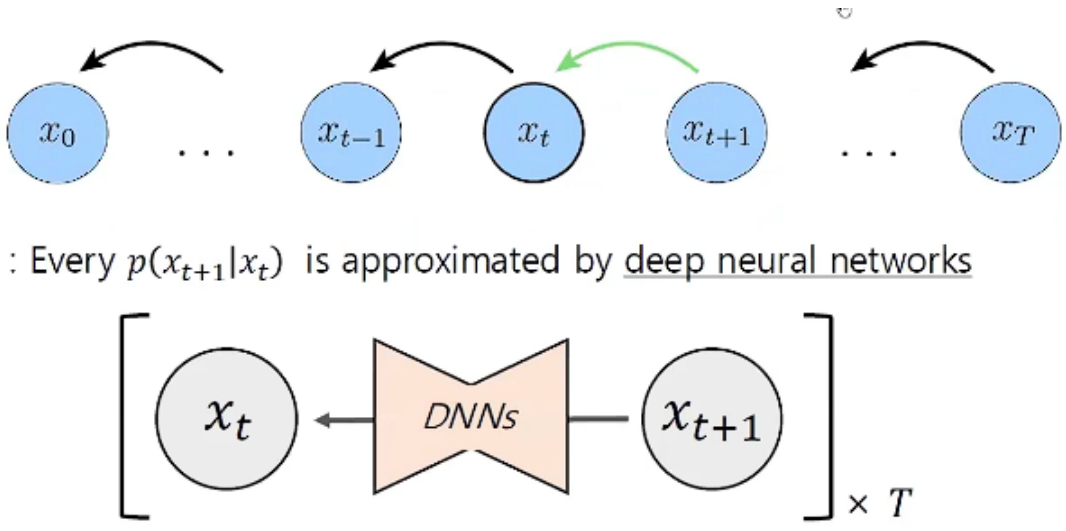


Neural Network Approximation

DNNs를 사용하여 $p(x_{t-1}|x_t)$를 근사

확산 과정은 수학적으로 정의되지만, 역방향 과정(노이즈 제거)은 딥러닝 모델이 학습하여 실행

즉, 신경망은 노이즈 제거를 통해 점진적으로 현실적인 데이터를 복원하는 방법을 학습함



Stable Diffusion

stable diffusion 모델은 Latent Diffusion Model(LDM)을 기반으로 하는 텍스트-투-이미지 생성 모델이다. 이 모델은 이미지의 픽셀 공간에서 직접 작업하지 않고, 잠재 공간에서 확산 과정을 수행하여 계산량을 줄이고 효율성을 높인다.

1. **모델의 주요 구조**

(1) Pixel Space (입력/출력)

- x : 원본 이미지 (픽셀 공간)
- $\tilde{x}$: 모델이 생성한 이미지
- ε (Encoder) : 픽셀 공간에서 잠재 공간으로 변환하는 인코더
- D (Decoder) : 잠재 공간에서 다시 픽셀 공간으로 변환하는 디코더

Stable Diffusion은 **픽셀 공간에서 직접 노이즈를 추가하는 것이 아니라, 잠재 공간에서 작업**하여 더 효율적으로 학습한다.

(2) Latent Space (잠재 공간)

- Diffusion Process
  - 주어진 잠재 변수 z에 점진적으로 노이즈 z_T를 추가
  - 완전한 노이즈에서 시작하여 점차 깨끗한 z_0으로 복원하는 과정을 학습
- Denoising U-Net ϵ_θ
  - 노이즈 제거를 담당하는 U-Net 구조의 신경망

- Skip Connection 사용 : 각 레이어의 정보를 다음 레이어로 전달하여 중요한 정보 보존
- Cross-Attention 적용 : 텍스트 및 이미지 조건을 반영하여 이미지 생성 성능 향상

(3) Conditioning (조건부 정보 입력)

Stable Diffusion은 텍스트-투-이미지 변환이 가능한 모델로, 다양한 조건을 활용한다.

- Semantic Map : 의미론적 지도
- Text : 텍스트 입력
- Representations : 다양한 조건 표현
- Images : 추가 이미지 조건

이러한 정보는 Croess-Attention을 통해 Denoising U-Net에 반영되어 이미지를 더 정밀하게 제어할 수 있다.

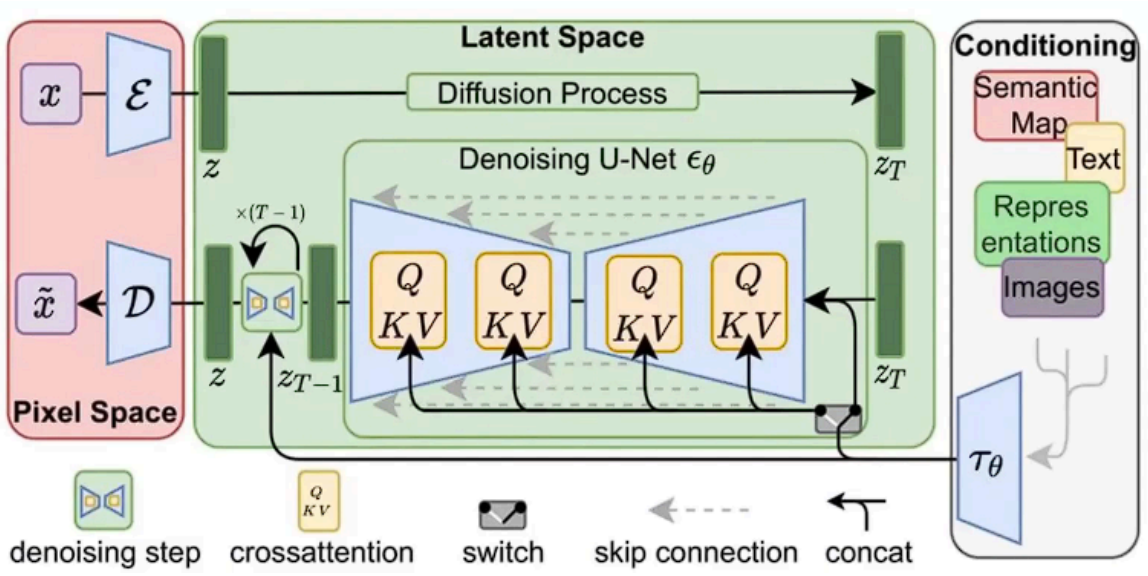
2. 모델의 주요 연산 요소

그림에 표시된 여러 기호들은 모델의 핵심 연산을 나타낸다.

- Denoising Step (노이즈 제거 단계): zT→zT−1→...→z0 점진적으로 노이즈 제거
  zT→zT−1→...→z0z_T \to z_{T-1} \to ... \to z_0
- Cross-Attention: 텍스트 또는 외부 조건을 이미지 생성에 반영하는 메커니즘
- Switch: 네트워크 내에서 정보 흐름을 제어하는 메커니즘
- Skip Connection: 깊은 네트워크에서도 중요한 정보를 유지하기 위한 연결
- Concat: 여러 정보를 병합하여 더 풍부한 표현을 생성

3. Stable Diffusion이 동작하는 방식

1. 이미지 또는 텍스트 입력을 받아 잠재 공간에서 확산 모델을 수행
2. Diffusion Process를 통해 노이즈를 점진적으로 추가한 후, Denoising U-Net을 사용하여 노이즈를 제거
3. Cross-Attention을 이용하여 텍스트 조건을 반영
4. 최종적으로 Decoder가 잠재 공간에서 복원된 이미지를 픽셀 공간으로 변환하여 최종 결과 생성



2. Key Components in Diffusion Models

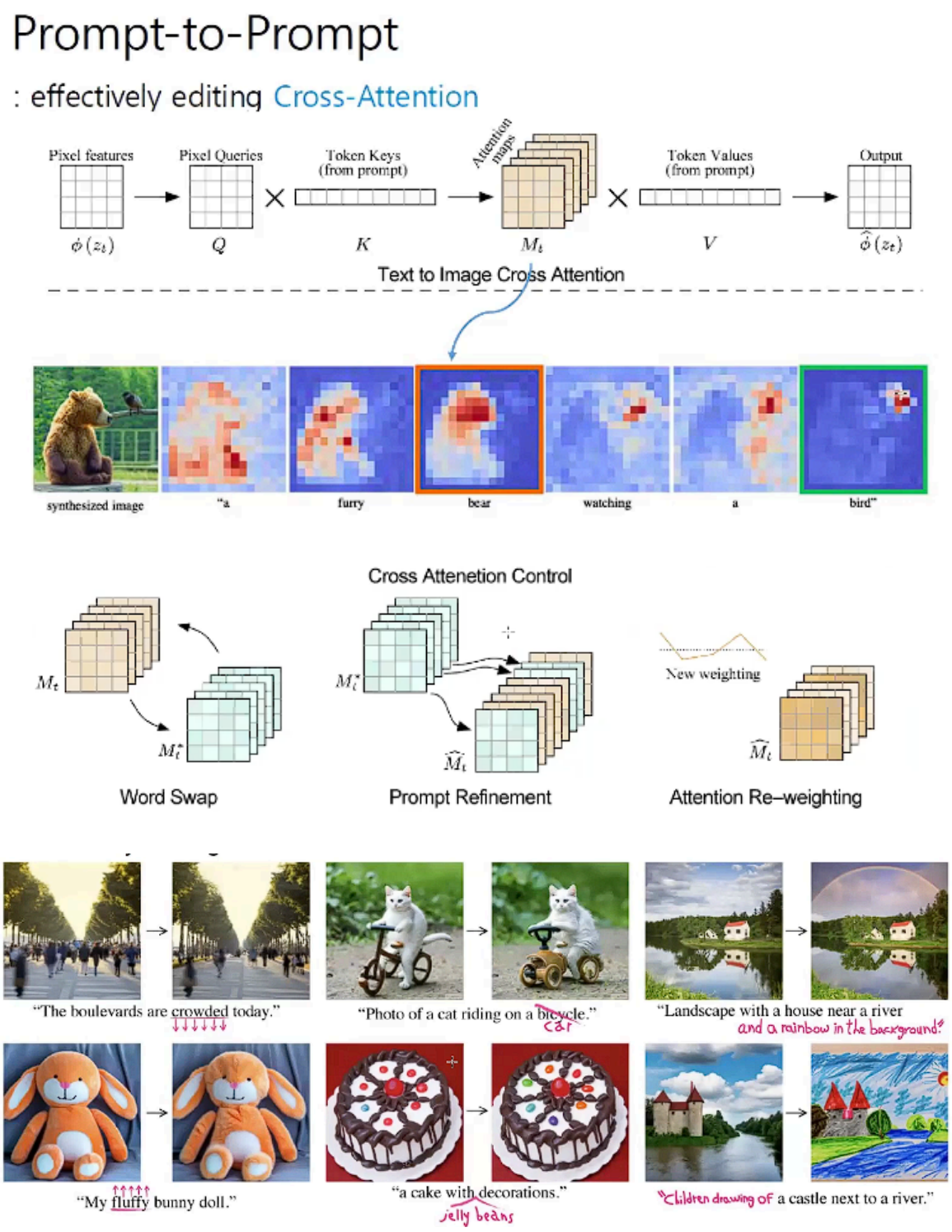
2-1. Attention Modules

- Self-Attention : 모델이 자체적으로 학습한 이미지 패턴 간의 관계를 고려
- Cross-Attention : 텍스트 프롬프트와 이미지 특징 간의 연관성을 학습

Prompt-to-Prompt

Prompt-toprompt기법은 **Cross-Attention 조정**을 통해 입력 텍스트 기반으로 생성된 이미지의 특정 부분을 변경하는 방식이다.

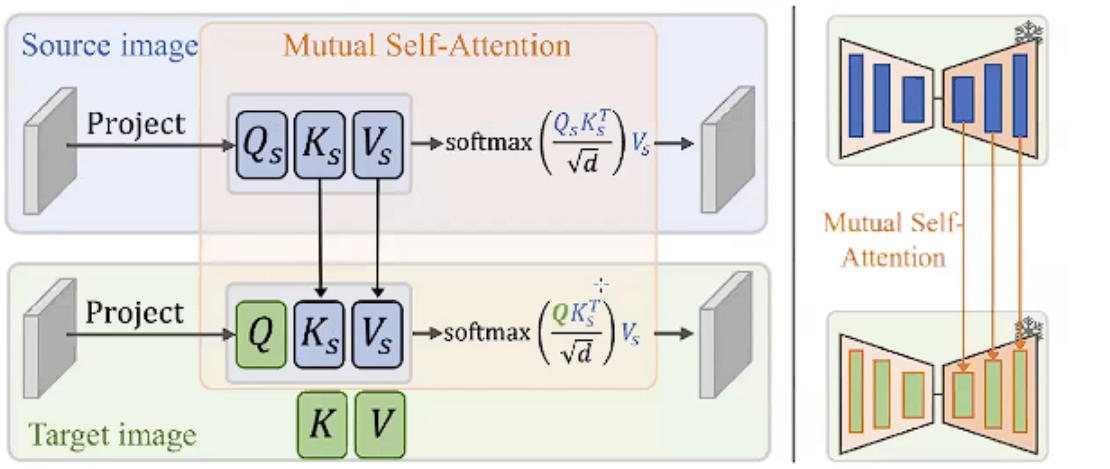
- Cross-Attention Control : 입력 프롬프트의 키(Key), 쿼리(Query), 값(Value) 행렬을 수정하여 특정 단어가 강조되도록 수정
- Word Swap : 텍스트 일부를 변경하여 이미지의 특정 요소를 바꾸는 방식
- Prompt Refinement : 기존 프롬프트를 조정하여 더 정교한 이미지 생성 가능
- Attention Re-weighting : 특정 단어의 중요도를 조정하여 이미지의 스타일이나 구성에 영향을 줌



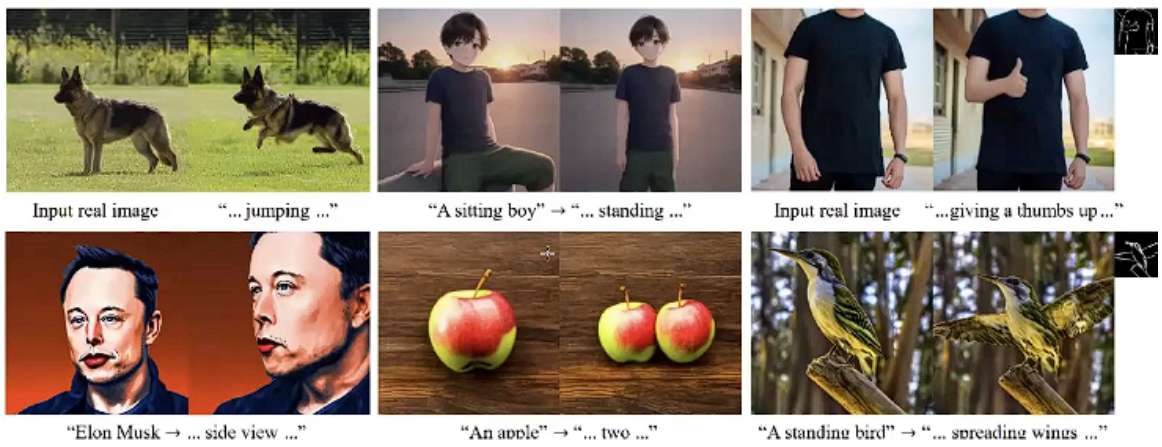
MasaCtrl

MasaCtrl은 **Mutual Self-Attention**을 활용하여 **원본 이미지와 타겟 이미지 간의 관계를 조정**하는 방식이다.

- 특정 이미지를 유지하면서 일부 속성만 변화시키는 것이 가능하다.
- Self-Attention을 제어하여, 기존 이미지의 핵심 특성을 유지하면서 변형을 수행한다.



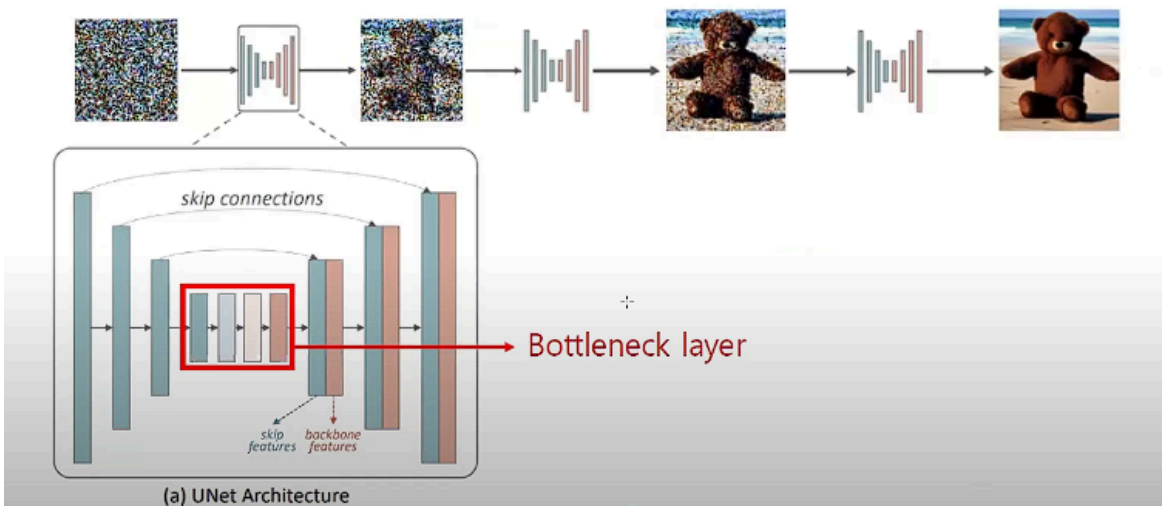
(a) Mutual self-attention control



2-2. U-Net Bottleneck

디퓨전 모델에서는 노이즈를 제거하면서 점진적으로 선명한 이미지를 복원하는 과정이 있다. 이때, U-Net이 중요한 역할을 하며, 특히 인코더와 디코더 사이 가운데 있는 Bottleneck Layer가 핵심이다.

- 인코더-디코더 구조에서 가장 정보가 압축되는 지점
- 노이즈가 점점 사라지는 중간 단계를 효율적으로 학습할 수 있도록 설계됨

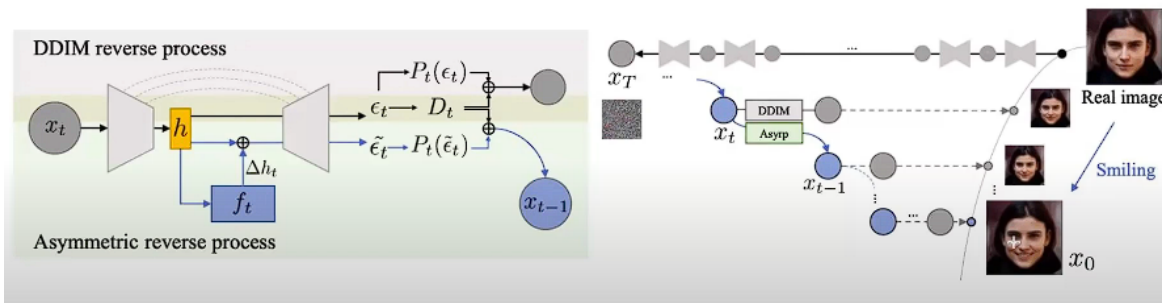


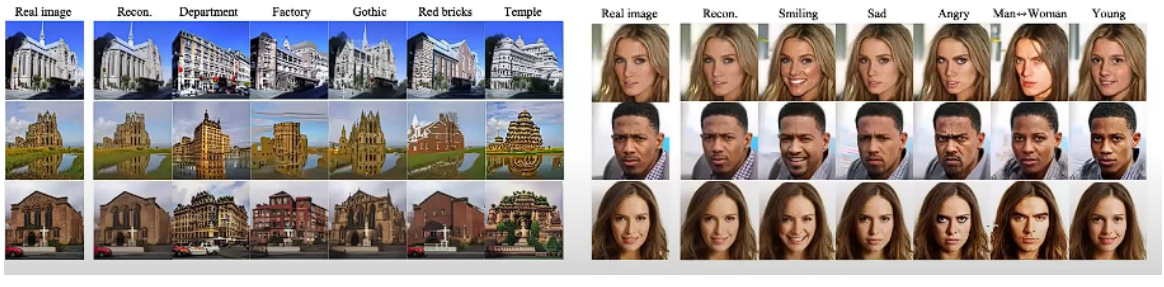
H-space

Denoising Diffusion Implicit Models의 Reverse Process에서 H-space를 활용하여 보다 효율적으로 노이즈를 줄일 수 있다.

H-space에서 연산을 수행하면 원본 이미지의 다양한 속성을 바꾸는 것이 가능하다.

- 예시:
  - 건물 스타일 변형 : 원본 건물을 다양한 스타일(고딕, 공장, 붉은 벽돌 등)로 변형 가능
    - 얼굴 감정 및 특성 변화 : 미소, 화남, 젊음 등 특정 속성을 조절 가능

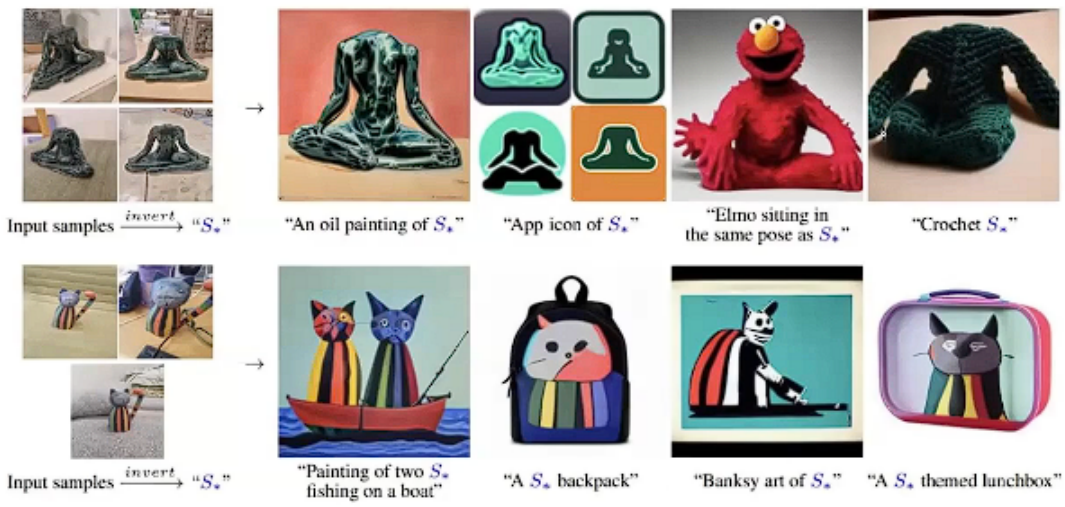




Textual Tokens & Textual Inversion

텍스트 임베딩을 이용해 특정 스타일이나 사물을 학습하고, 이를 기반으로 새로운 샘플을 생성하는 기법이다.

- Textual Tokens : 텍스트 정보를 효과적으로 반영하는 부분
- Textual Inversion : 특정 단어나 개체를 학습한 후, 이를 새로운 프롬프트에 활용하여
  원하는 스타일을 유지하면서 새로운 이미지를 생성
- 예시:
  - 특정 개체를 학습한 후, 다양한 문장 프롬프트에서 해당 개체를 활용해 새로운 이미지를 생성할 수 있음



Inspiration Tree

이미지의 개념을 나누고, 특정 스타일을 학습하는 방법이다.

- 이미지를 계층적으로 분석하여 세부적인 특징을 나누고, 이를 바탕으로 새로운 스타일을 생성
- 개념을 트리 형태로 정리해서 학습하고 활용
- 예시:
  - 특정 그림 스타일을 학습하고, 이를 활용하여 다양한 형태의 이미지를 생성
  - 특정 개체의 여러 변형된 모습을 생성할 수 있도록 학습