KUBIG 24-WINTER

COMPUTER VISION STUDY

BLIP-Diffusion

KUBIG 17기 임청수

01. Abstract

■ 전체 요약

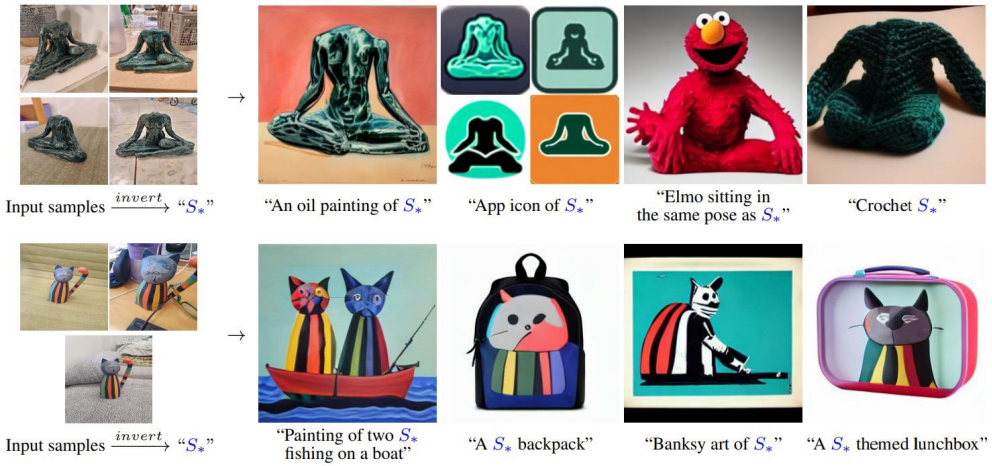
- subject driven text-to-image generation model이 빠르게 성장
- 이미지 내의 특정 객체(피사체)에 대한 프롬프트 기반 이미지 생성 모델을 의미

ex) dream booth, textual-inversion 등

- 기존 fine tuning 방법론의 단점

1) 피사체를 파인튜닝하는데 많은 시간 소요

2) 피사체를 높은 충성도(fidelity)로 구현하기에 여려움



01. Abstract

■ 전체 요약

blip-diffusion model

- few shot으로 원하는 피사체를 학습하고 높은 성능으로 피사체 기반 이미지를 생성
- 두 단계로 파인튜닝되는데 1단계는 blip-2 기반으로 피사체를 학습하며 2단계는 stable diffusion 기반으로 프롬프트 임베딩에 피사체 표현을 삽입함
- 학습 시 데이터는 동일한 피사체를 가진 여러 장의 이미지와 피사체 키워드를 사용
- 특정 피사체를 활용하기 위해 40~120 step의 fine tuning만 사용하므로 기존 dream booth보다 20배 속도 향상을 이룸.

02. Introduction

■ Introduction

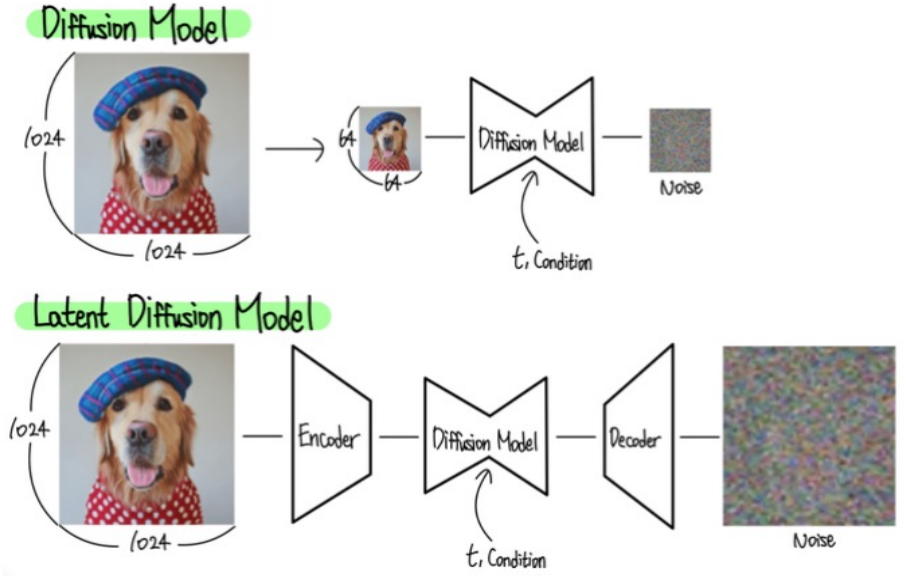
- subject driven text-to-image generation을 위한 fine-tuning은 멀티 모달 제어를 활용하지 않았기 때문에 비효율적이었음(textual inversion, dreambooth)
- 따라서 멀티모달 표현을 효율적으로 학습할 수 있는 blip-2을 활용하여 피사체를 텍스트 형태로 삽입하는데 최적화 진행
- 학습 시 데이터는 동일한 피사체를 가진 여러 장의 이미지와 피사체 키워드를 사용하여 멀티모달 임베딩 출력.
- 임베딩과 텍스트 프롬프트를 결합하여 새로운 이미지 생성을 가이드함
- zero-shot 및 few shot 피사체 중심 생성에서 높은 효율성을 가져옴

03. Related work

■ Stable Diffusion : High-Resolution Image Synthesis with Latent Diffusion Models

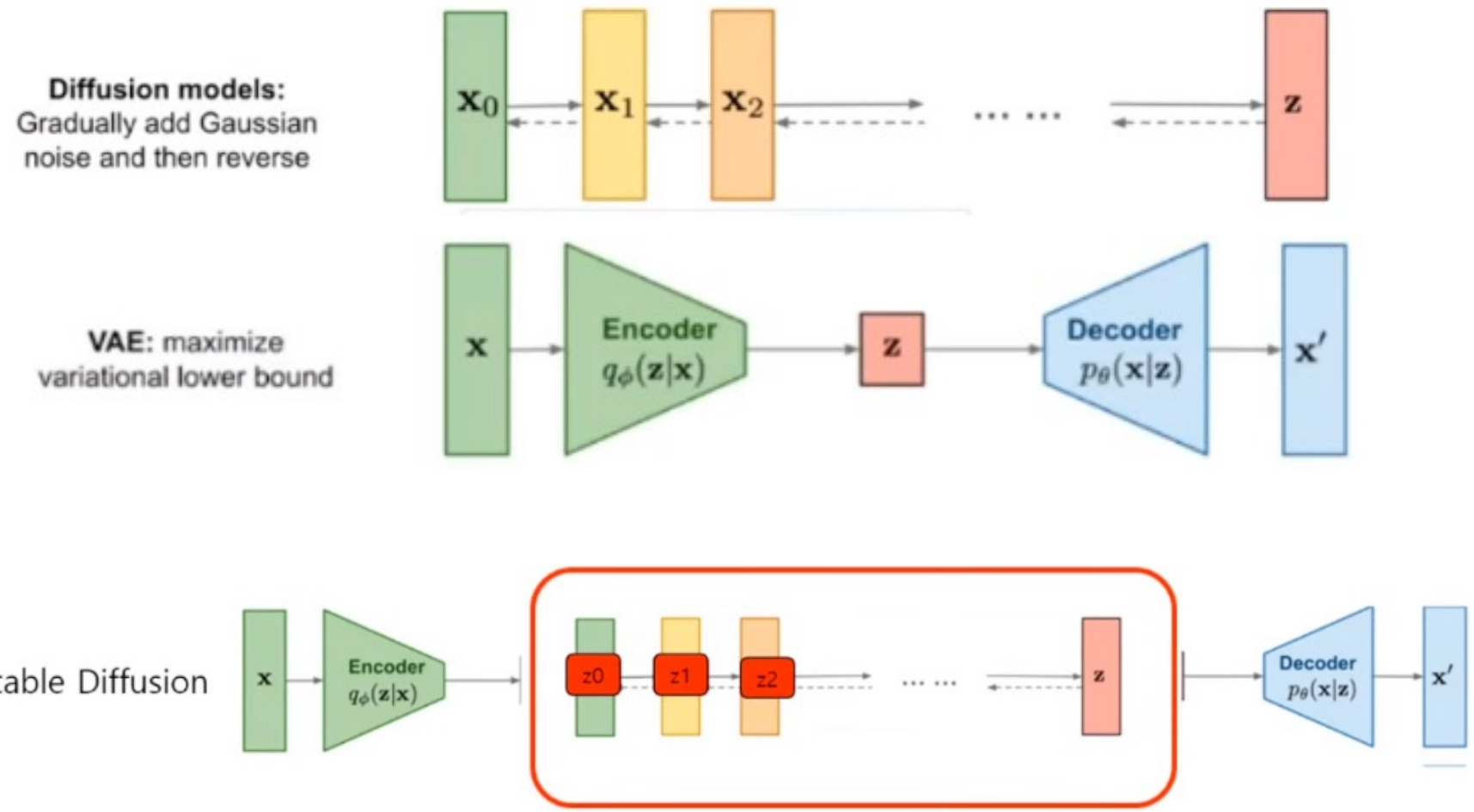
Imagen, DALLE-2 등 엄청난 수준의 이미지를 생성하는 Diffusion Model 등장

- 하지만 저화질 이미지만 생성할 수 있다는 한계로 super resolution 모델을 따로 학습하여 붙여줘야 함
- diffusion model은 이미지 픽셀 단위로 생성하기 때문에 사람 눈의 인지가 잘 되지 않는 Non perceptual한 부분을 학습하는데 초점을 맞추고 있음
- stable diffusion은 이러한 부분을 개선하여 perceptual한 부분에 초점을 맞춰 학습하는 diffusion model을 제안



03. Related work

■ Stable Diffusion : High-Resolution Image Synthesis with Latent Diffusion Models



03. Related work

■ Stable Diffusion : High-Resolution Image Synthesis with Latent Diffusion Models

빨간색 음영 부분은 auto encoder, 초록색 음영 부분은 latent diffusion model,오른쪽 회색 음영 부분 condition

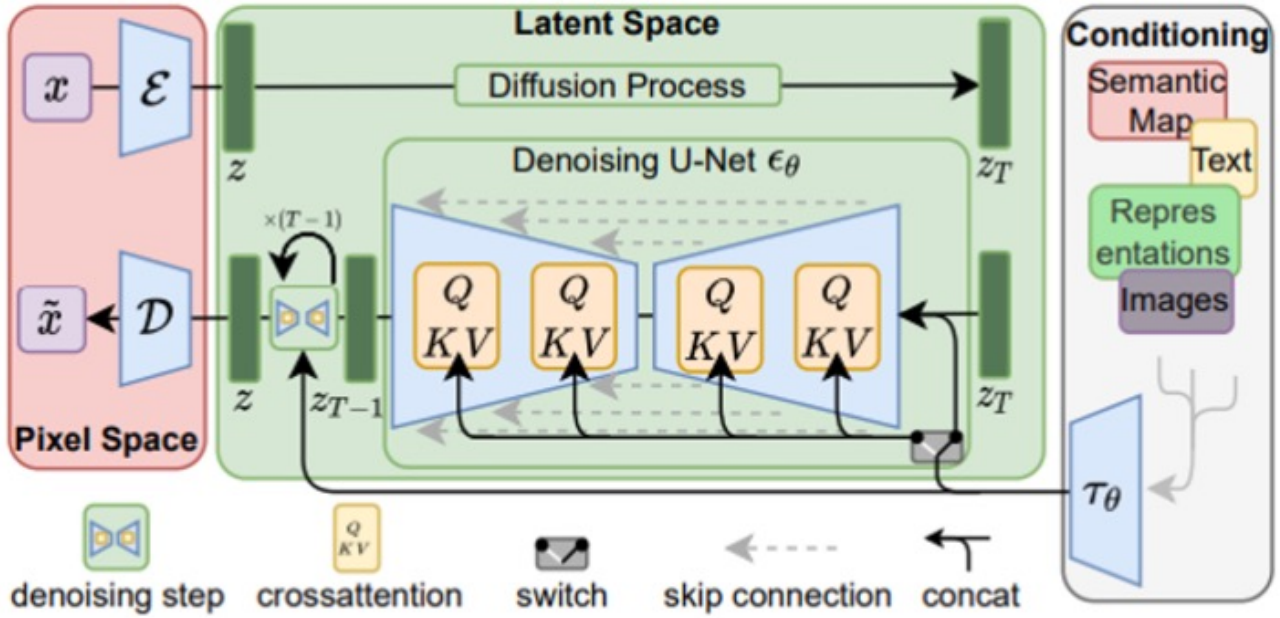


Figure 3. We condition LDMs either via concatenation or by a more general cross-attention mechanism. See Sec. 3.3

03. Related work

■ Stable Diffusion : text to image

- Text를 Condition으로 받아 이미지를 생성
- pretrained LLM의 text encoder를 통해 embedding을 생성한 후 cross attention을 통해 이미지 정보인 z와 연산하여 상관관계 반영
- query는 image z_t, key, value는 condition(Text)로 cross attention 연산 진행

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon-\epsilon_\theta(z_t,t,\tau_\theta(y))\|_2^2\right]$$

그림8. Stable Diffusion Model Loss Function

03. Related work

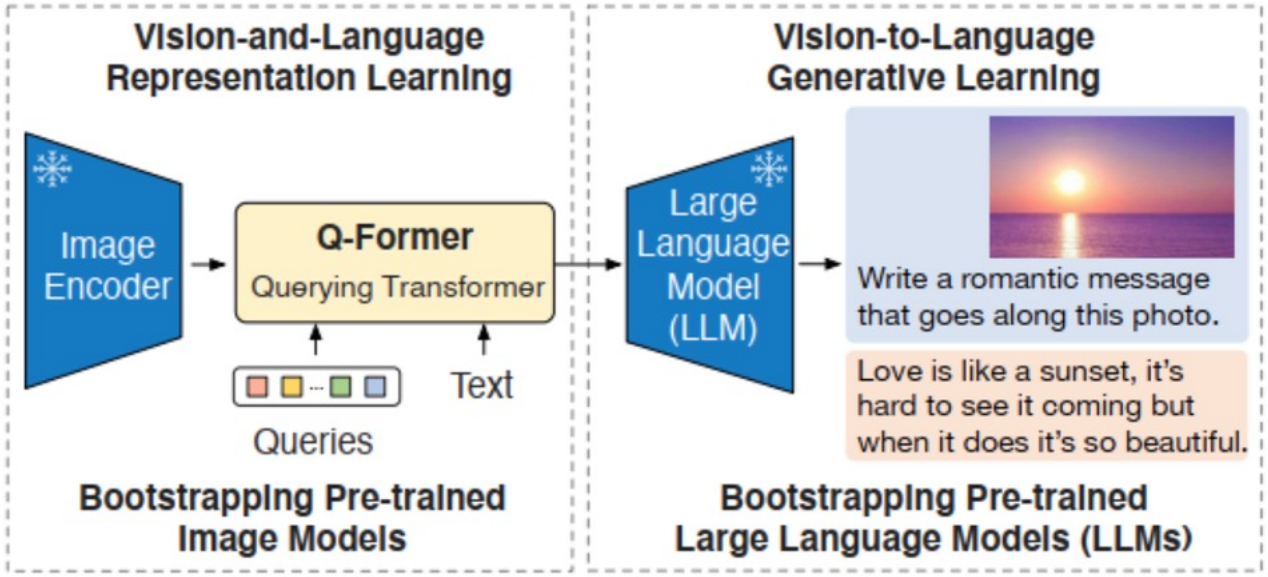
■ Stable Diffusion : High-Resolution Image Synthesis with Latent Diffusion Models

- Denoising 과정에서 AutoEncoder 사용
- pixel 공간이 아닌 latent space에서 Denoising을 진행하면서 수백 개의 GPU가 필요한 컴퓨팅 코스트를 줄임
- 아키텍처 상의 Cross-Attention을 사용함으로써, 다른 도메인(text, audio, image 등)을 함께 모델상에서 사용 가능

03. Related work

■ blip-2

- frozen image encoder와 LLM 모델을 연결해주는 새로운 방법(Q-former) 제시
- frozen model을 통해 발생한 modality gap은 Q-former를 통해 해결

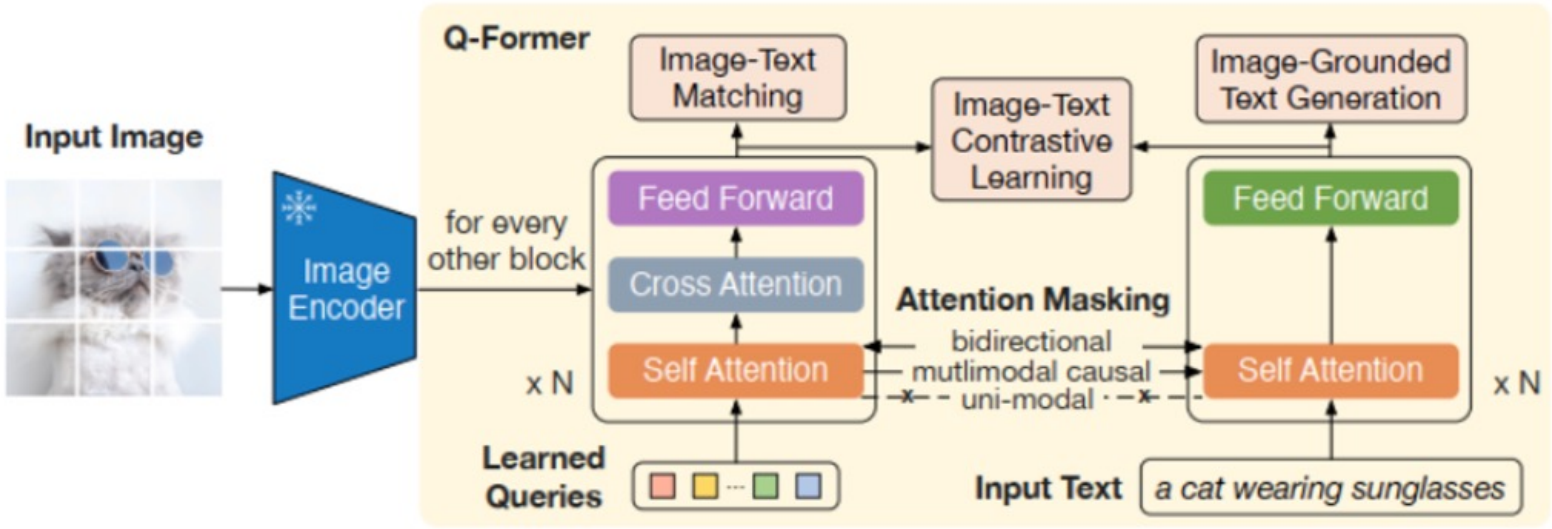


03. Related work

■ blip-2

- blip-2는 두 단계 학습 진행
  - frozen image encoder를 통해 representation을 얻고 Q-former를 활용하여 frozen LLM에 넘겨줌.

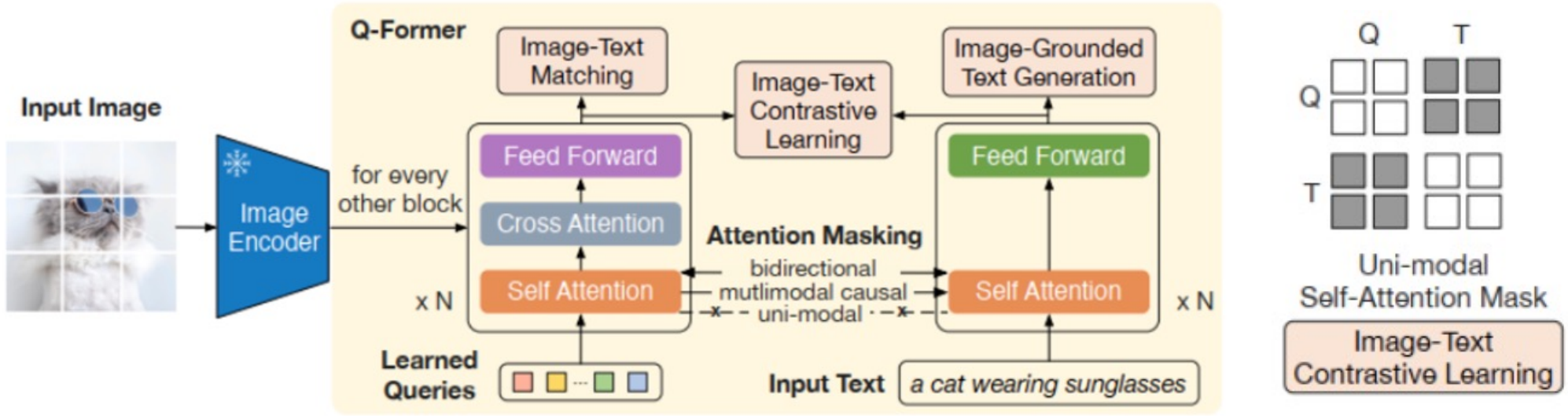
LLM은 VL generative learning 수행



03. Related work

■ blip-2 - ITC(Image-Text Contrastive Learning)

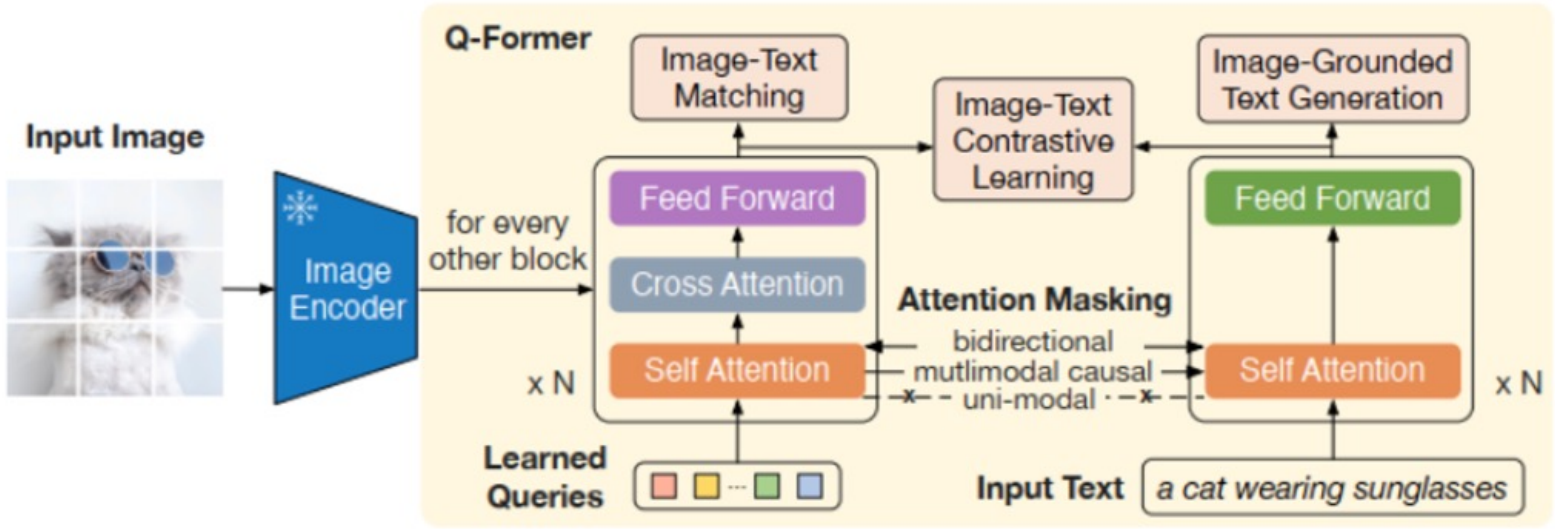
- image transformer에서 나온 query output과 text transformer에서 나온 output간의 pairwise 유사도를 계산하고, 가장 값이 높은 pair를 query-text pair로 선정
  - image와 text가 서로의 정보를 공유하면 안되므로 Uni-modal Self-Attention Mask를 사용

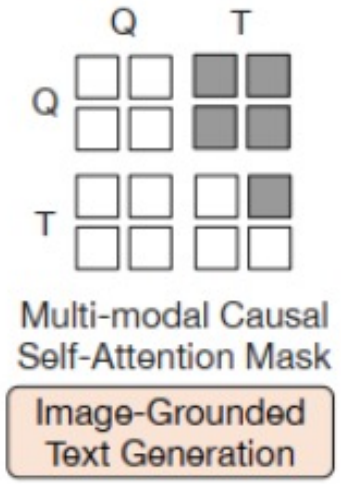


03. Related work

■ blip-2 - ITG(Image-grounded Text Generation)

- Encoder에서 뽑아낸 이미지 정보는 공유된 self-attention layer를 통해 text tokens로 전달되어 text generation 진행
- query는 text와 관련된 이미지 정보들을 전달하도록 학습됨
- query가 text 정보를 미리 보면 안되므로 Multi-modal Causal Self-Attention Mask를 활용하여 query가 text 정보를 참고하지 못하도록 설정

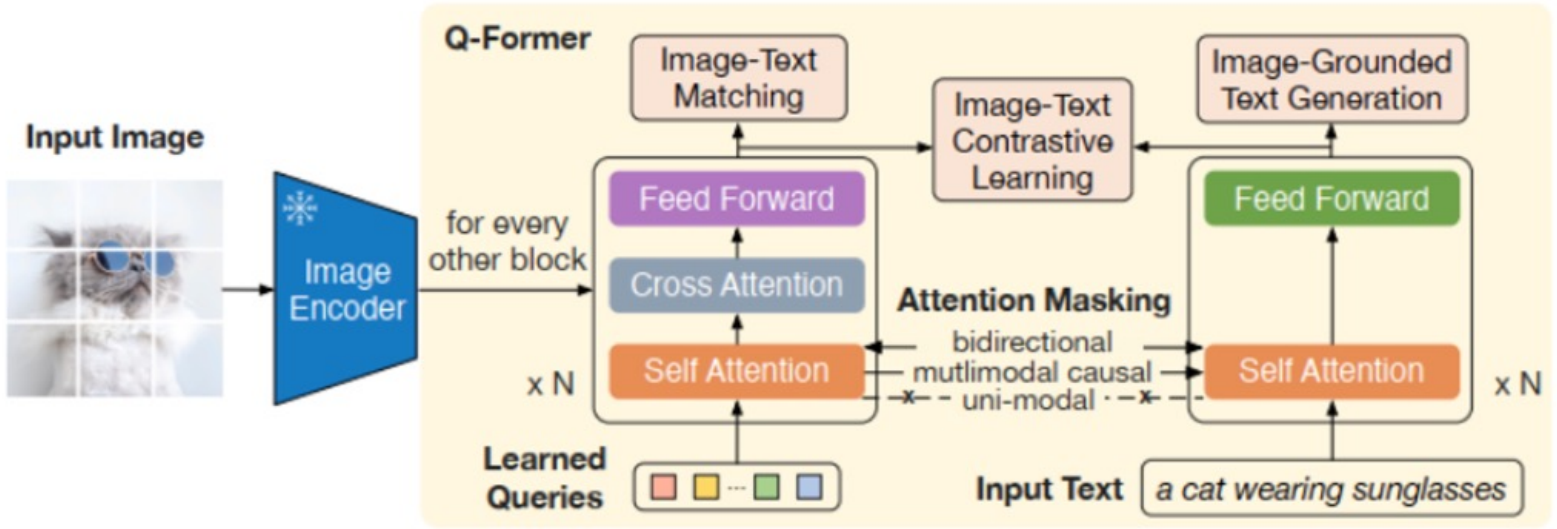


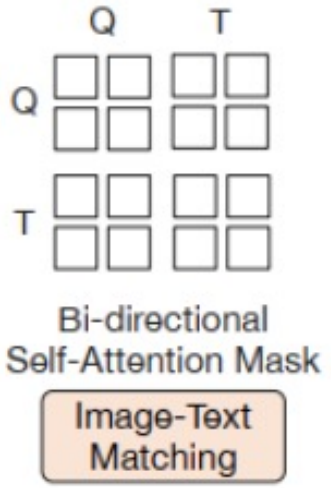


03. Related work

■ blip-2 - ITM(Image-Text Matching)

- Image와 text의 fine-grained alignment를 학습
- 가지고 있는 모든 정보를 참고해도 문제가 없기 때문에, Bi-directional Self-Attention Mask를 사용

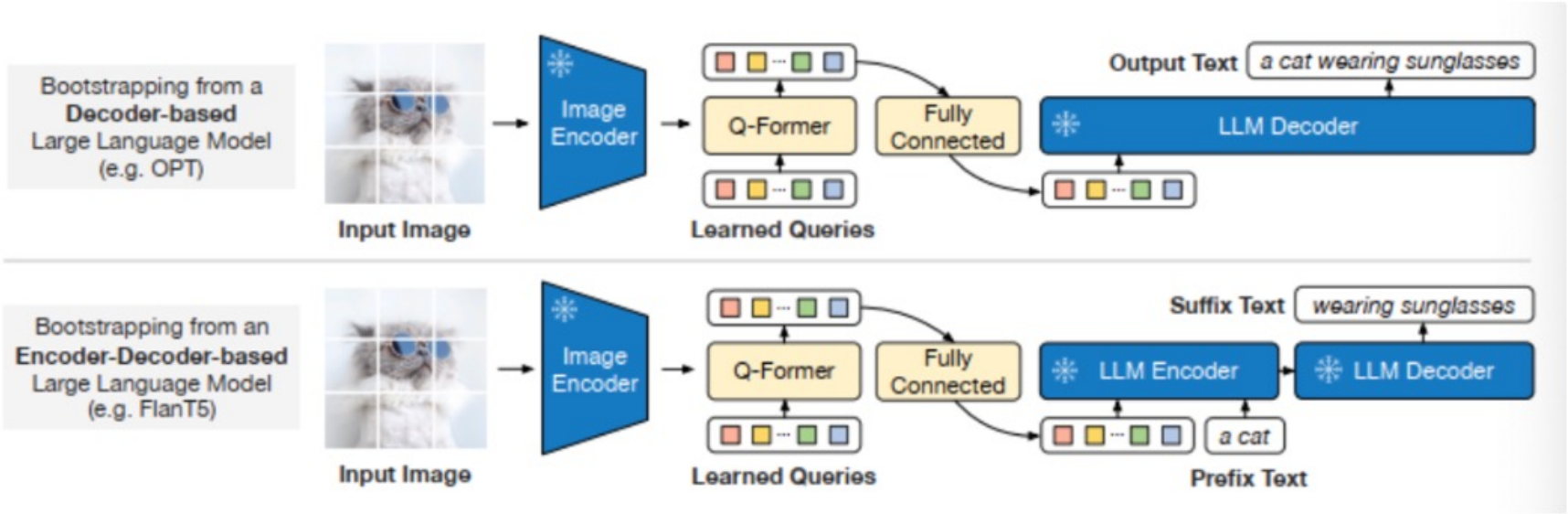




03. Related work

■ blip-2 - ITM(Image-Text Matching)

- Q-Former의 output query는 완전연결 계층(Fully Connected Layer)를 통해 LLM로 전달
- FC layer를 통해 query의 Dimension을 text embedding과 동일하게 낮춤
- text embedding 앞에 붙여서 visual prompt로 사용



04. Method

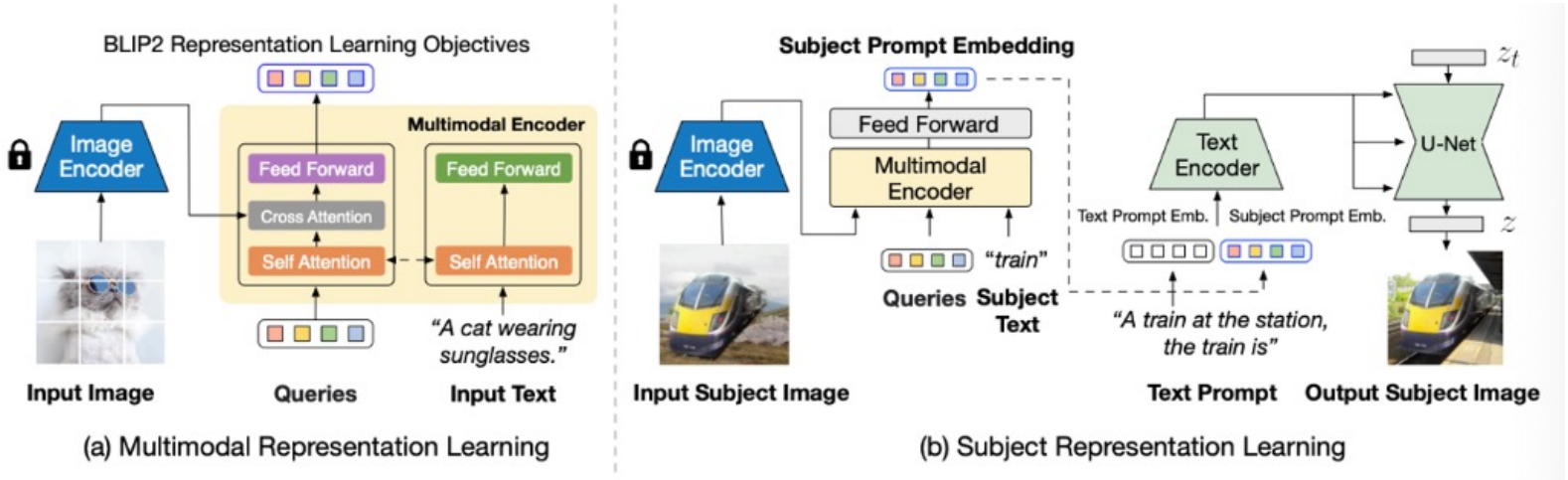
BLIP-Diffusion

Left : multi-modal representation learning stage로 blip-2의 인코더 구조를 활용하여 text-aligned image representation을 학습

Right : subject representation learning stage로 input image에 포함된 피사체에 대해 텍스트 프롬프트를 입력하여 새로운 배경으로 합성

1단계에서 출력된 subject prompt embedding은 2단계에서 text prompt embedding과 함께 입력

latent diffusion model은 subject image를 생성. 이때 image encoder는 frozen pre-training model을 사용



05. Experiment

■ Zero shot and few shot subject-driven generation

사전 학습된 모델을 사용하기 때문에 zero shot으로도 subject based generation 가능

05. Experiment

■ Zero shot and few shot subject-driven generation

fine tuning시 high-fidelity generation이 가능해짐

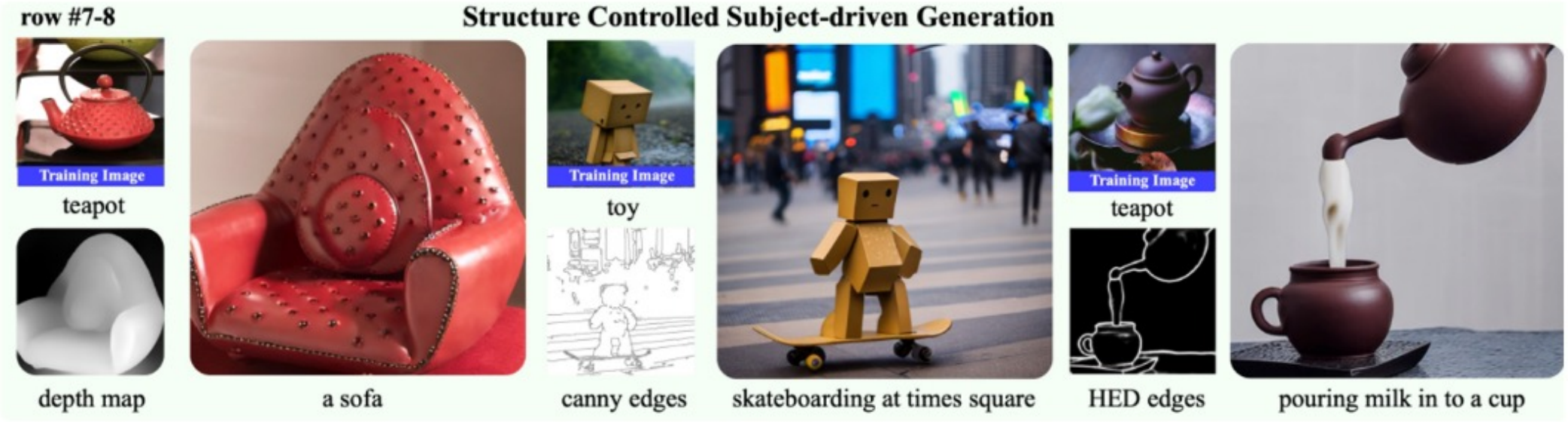
05. Experiment

■ Structure Controlled subject driven generation

controlNet 사용 시 구조와 피사체를 변경할 수 있음

Adding Conditional Control to Text-to-Image Diffusion Models

https://arxiv.org/abs/2302.05543



05. Experiment

■ Structure Controlled subject driven generation