

# Autoeic

: Auto-making exams of TOEIC part 1



Hairy Potatoes (23-1 모델링 H조)  
8기 엄소은, 8기 송규원, 8기 이재우, 9기 이성균



## **LISTENING TEST**

In the Listening test, you will be asked to demonstrate how well you understand spoken English. The entire Listening test will last approximately 45 minutes. There are four parts, and directions are given for each part. You must mark your answers on the separate answer sheet. Do not write your answers in your test book.

### **PART 1**

Directions: For each question in this part, you will hear four statements about a picture in your test book. When you hear the statements, you must select the one statement that best describes what you see in the picture. Then find the number of the question on your answer sheet and mark your answer. The statements will not be printed in your test book and will be spoken only one time.

1.




# Contents

? Why AuToeic?

 Dataset

 Pipeline

 Trial and Errors

 Value of Project



# AuToeic

토익 Part 1 (사진 묘사) 문항을 **자동으로 생성**해주는 알고리즘

- 사진 촬영 인력/비용 증대
- 자주 진행되는 시험(월 2회)
- 많은 역량을 들일 영역이 아님(Part 3,4,7에서 변별)



- ✓ 원하는 단어만 입력하면, 시험에 필요한 **사진을 자동 생성**
- ✓ **매력적인 오답 선지 생성**
- ✓ 많은 모의고사를 빠르고 적은 비용으로 만들어내야 하는 사설 교육업체에 유용



# 데이터 소개

## Common dataset 1) Facebook/Winoground

Dataset Preview Size: 364 MB </> API Go to dataset viewer

id (int32)	image_0 (image)	image_1 (image)	caption_0 (string)	caption_1 (string)	tag (string)	secondary_tag (string)	num_main_preds (int32)	coll (str)
0			"an old person kiss..."	"a young person kiss..."	"Adjective-Age"	""	1	"Rel"
1			"the taller person hugs..."	"the shorter person hugs..."	"Adjective-Size"	""	1	"Rel"
2			"the masked wrestler hi..."	"the unmasked wrestler hi..."	"Adjective-Manner"	"Series"	1	"Rel"
3			"a person watches an..."	"an animal watches a..."	"Noun"	""	1	"Obj"
4			"the person without..."	"the person with earring..."	"Negation, Scope"	"Morpheme-Level"	1	"Rel"

크기: 400 rows

Image-Text pair

- image\_0 (image)
- image\_1 (image)
- caption\_0 (string)
- caption\_1 (string)

## Common dataset 2) Facebook/pmd

image_url (string)	image (image)	text (string)	source (string)	meta (string)
null		"A woman wearing a net on her head..."	"coco"	"{ 'annotation': 'A woman wearing a net on her head cutting a cake. ', 'image_path':..."
null		"A woman cutting a large white sheet..."	"coco"	"{ 'annotation': 'A woman cutting a large white sheet cake.', 'image_path':..."
null		"A woman wearing a hair net cutting a..."	"coco"	"{ 'annotation': 'A woman wearing a hair net cutting a large sheet cake.', 'image_path':..."
null		"there is a woman that is cutting a..."	"coco"	"{ 'annotation': 'there is a woman that is cutting a white cake', 'image_path':..."
null		"A woman marking a cake with the back..."	"coco"	"{ 'annotation': 'A woman marking a cake with the back of a chef's knife. ', 'image_path':..."

크기: 70,000,000 rows

Image-Text pair


- image (image)
- text (string)





# 데이터 소개

## Common dataset 3) Michelecafagna26/hl

image (image)	scene (sequence)	action (sequence)	rationale (sequence)	object (sequence)	confidence (dict)	purity (dict)
	[ "in a beach", "the picture was taken on the beach.", "it is taken by a lake" ]	[ "holding an umbrella", "the lady is posing with he sun umbrella.", "she us holding a parasol" ]	[ "so they won't get a sun burn", "she is enjoying and getting pictures of her vacation.", "she is vacationing and it is sunny" ]	[ "Woman in swim suit holding parasol on sunny day.", "A woman posing for the camera, holding a pink, open umbrella and wearing a bright, floral, ruched bathing suit, by a life guard stand with lake, green trees, and a blue sky with a few clouds behind.", "A	{ "scene": [ 5, 4, 5 ], "action": [ 5, 2, 5 ], "rationale": [ 5, 5, 4 ] }	{ "scene": [ -1.216016292572, -0.861034095287, -1.416510105133 ], "action": [ -1.041449785232, -1.126106739044, -1.407878875732 ], "rationale": [ -1.726574897766, -1.012382745742, -0.973266899585 ] }

- 크기: 149997 images & 134973 high-level captions

### Image-Text pair

- image (image)
- object (sequence)

## Common dataset 4) Action-Effect

verb noun (string)	effect_sentence_list (sequence)	effect_phrases_list (sequence)	positive_image_list (images list)	negative_image_list (images list)
"arrange chairs"	[ "chairs are moved around in order", "the chairs are..." ]	[ [ "are moved", "are moved around" ], [ "are put" ], ... ]		
"arrange flowers"	[ "the flowers are in a pretty design", "flowers are..." ]	[ [ "in", "in a pretty design" ], [ "are..." ] ]		
"bake potato"	[ "i put a potato in the oven to bake it", "the..." ]	[ [ "in the oven" ], [ "safe" ], [ "is heated", ... ] ]		
"beat eggs"	[ "the eggs are stirred", "the eggs are scrambled", ... ]	[ [ "are stirred" ], [ "are scrambled" ], [ "will be..." ] ]		

- 크기: 140 verb-noun pairs & 10 images and captions

### Image-Text pair

- positive\_image\_list (image)
- negative\_image\_list (image)
- effect\_sentence\_list (string)



# 데이터 소개

## Train Dataset

- ✔ 토익 image-text pair 500개 중 480개  
Text는 정답 선지만 사용

Ex)



A cloth has been draped over a table.

## Test Dataset

- ✔ 토익 image-text pair 500개 중 20개

Ex)

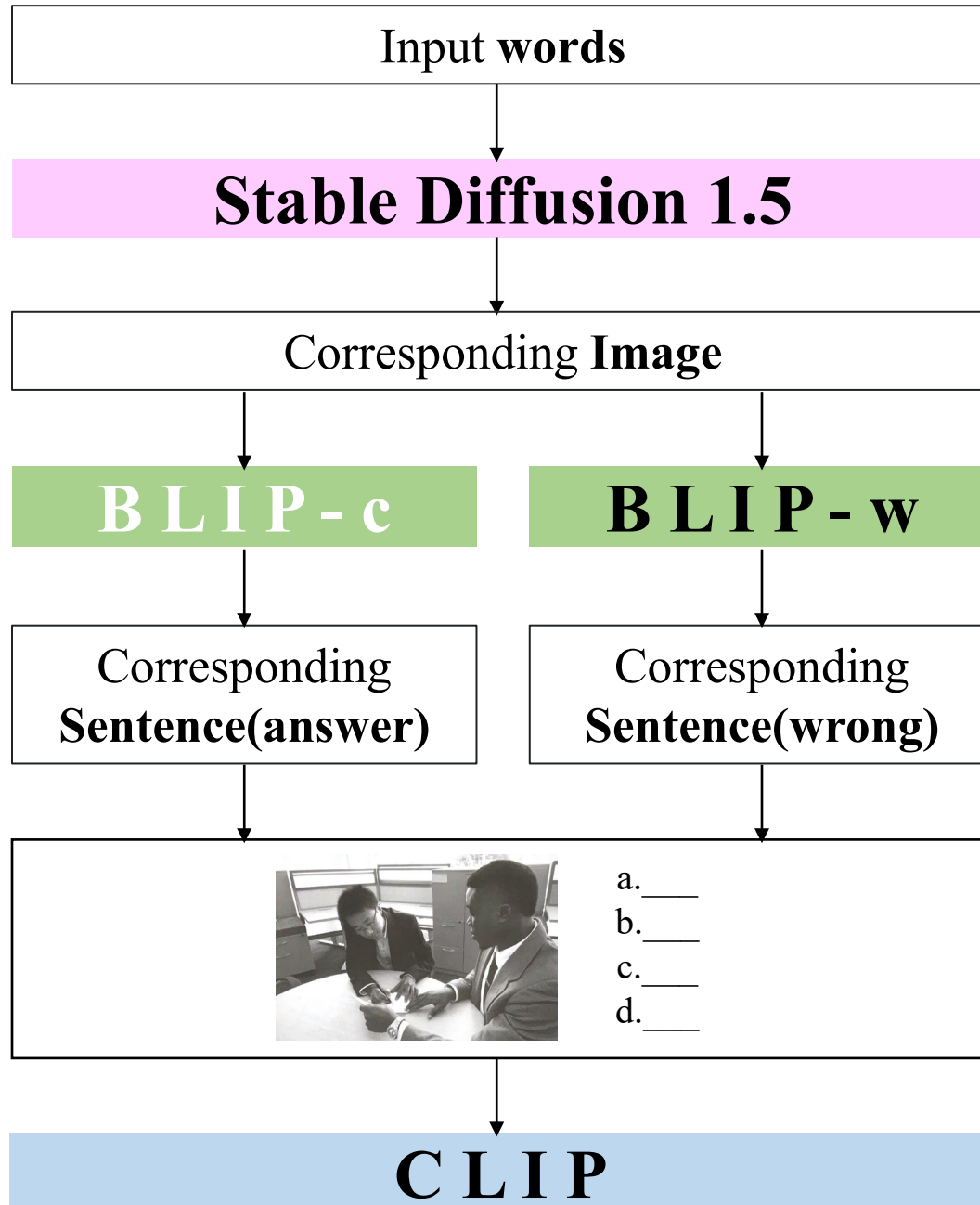


Some bicycles are resting against  
a barrier.





# 모델 구조



## Stage 1.

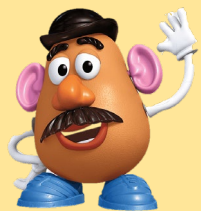
출제자가 원하는 텍스트(상황)를 입력하면  
Diffusion 모델을 통해  
상응하는 사진을 생성

## Stage 2.

사진에 맞는 정답 선지와  
매력적인 오답 선지를  
BLIP 모델을 이용하여 생성

## Stage 3.

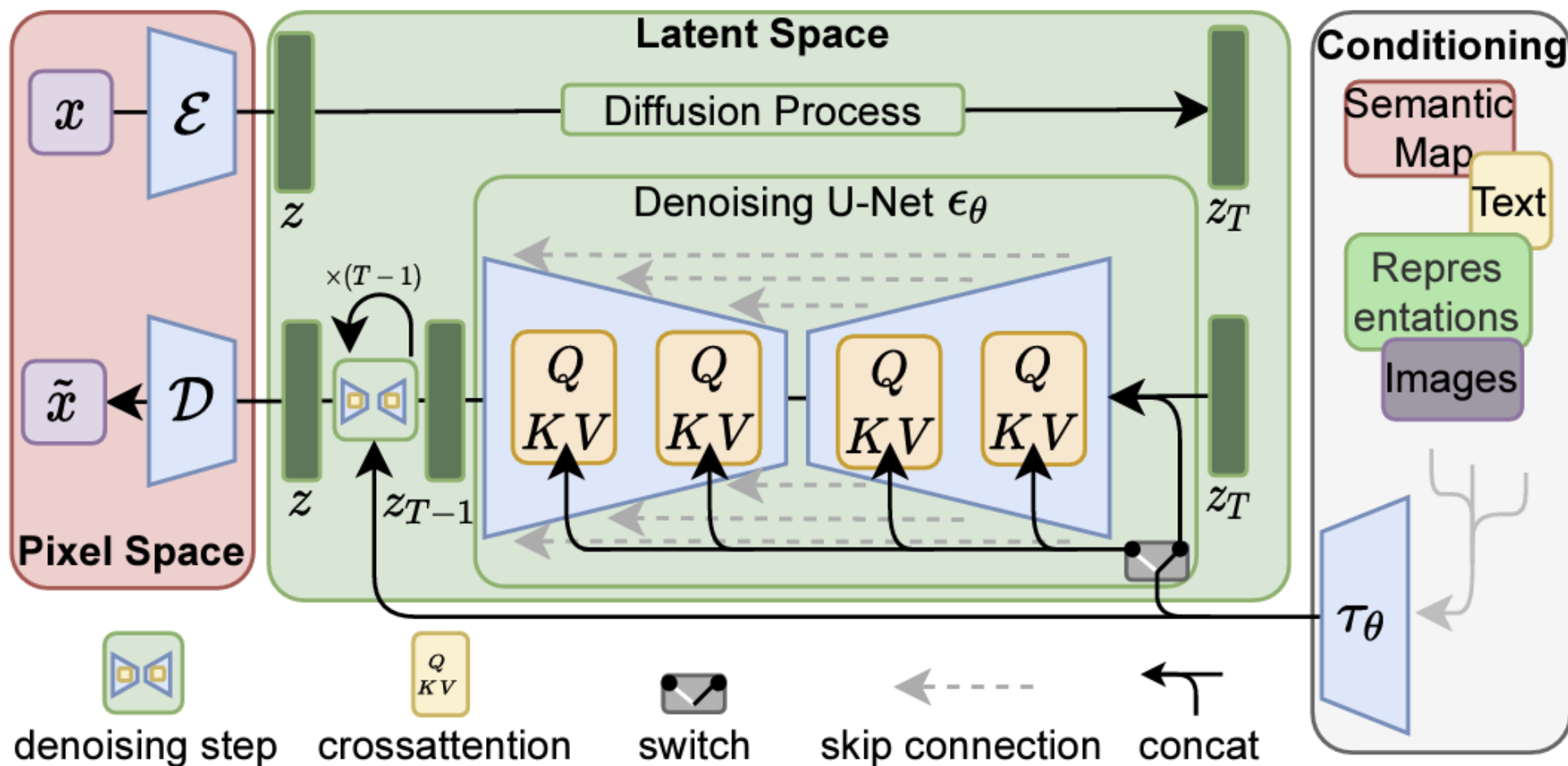
CLIP 모델을 이용하여  
출제가 잘 되었는지 평가

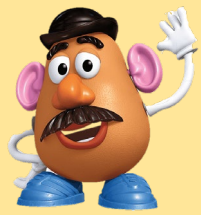


# 모델 설명

## Stable Diffusion 1.5

- Text-to-image 모델
- 텍스트 설명에 기반하여 상세한 이미지를 생성
- 이미지에서 이미지 변환과 같은 작업 수행

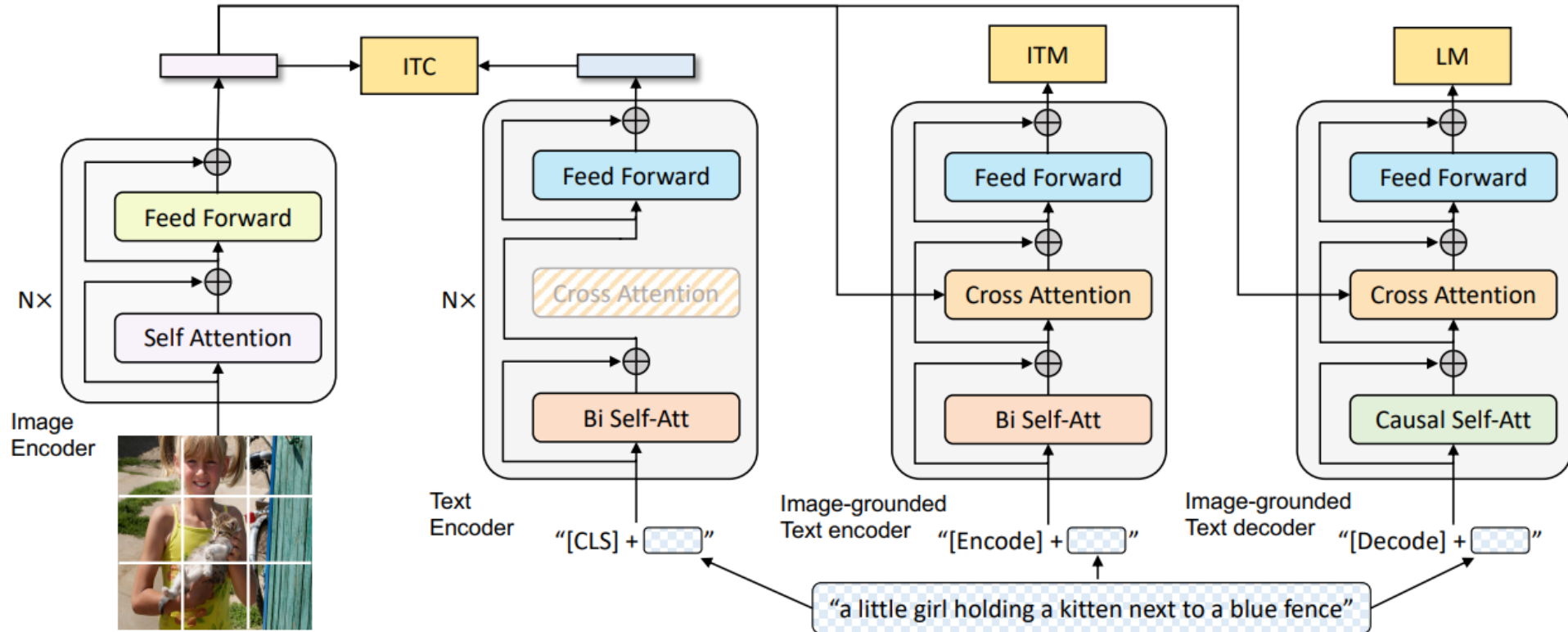




# 모델 설명

## BLIP

- Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation
- Multimodal mixture of Encoder-Decoder (MED)



### Model Architecture

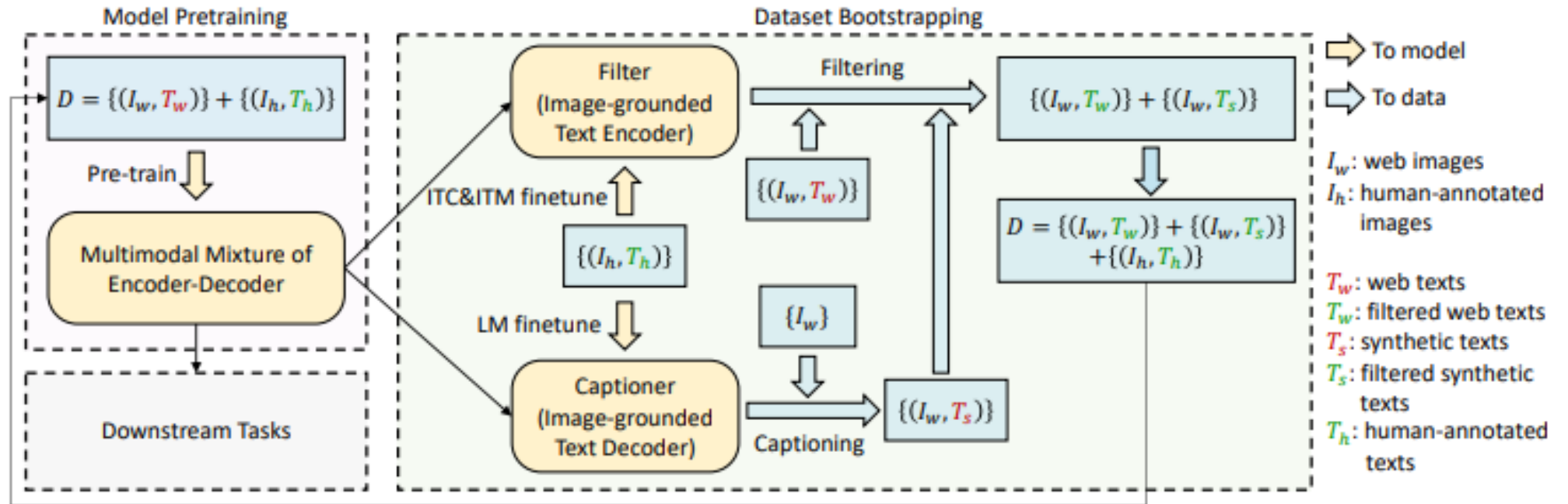
1. Unimodal encoder: Text는 BERT로 & Image는 ViT로 encoding
2. Image-grounded text encoder: Visual information이 주입됨
3. Image-grounded text decoder: Bi-self attention 제거, casual self-attention 추가



# 모델 설명

## BLIP

- Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation
- Multimodal mixture of Encoder-Decoder (MED)



## CapFilt

1. Image-Text Contrastive Loss (ITC) : 같은 {image, text} pair에 있으면 코사인 유사도가 높게, 반대면 유사도가 낮게 나오도록 학습함
2. Image-Text Matching Loss (ITM) : {image, text} 쌍이 match 됐는지 예측하도록 학습
3. Language Modeling Loss (LM) : 생성한 text가 original text와 얼마나 유사한지 학습

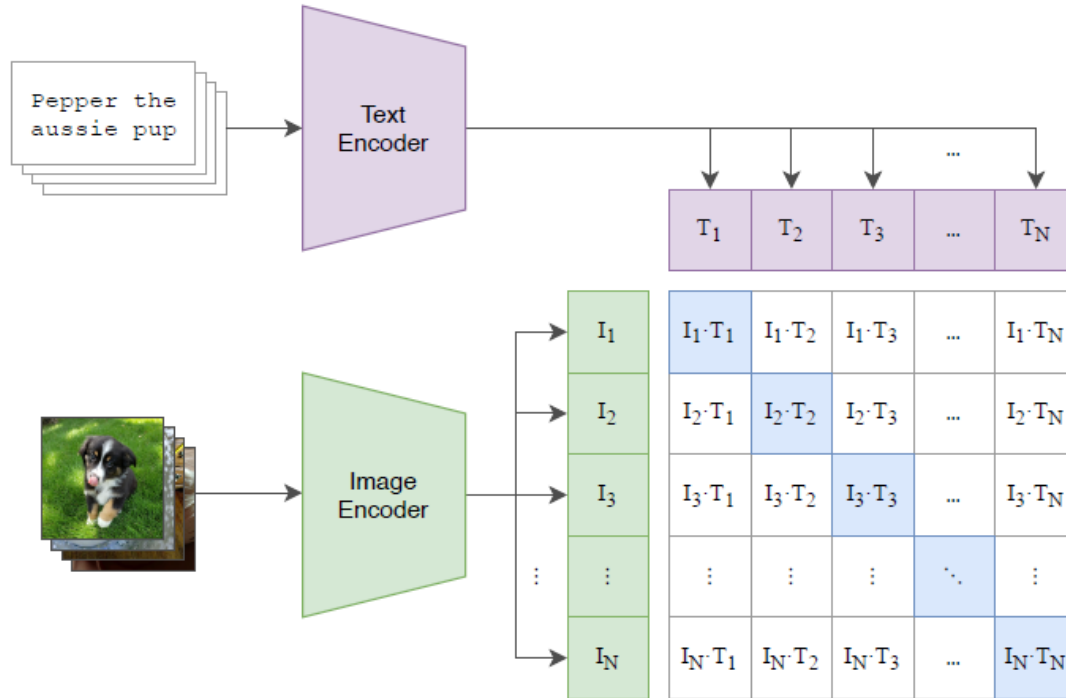


# 모델 설명

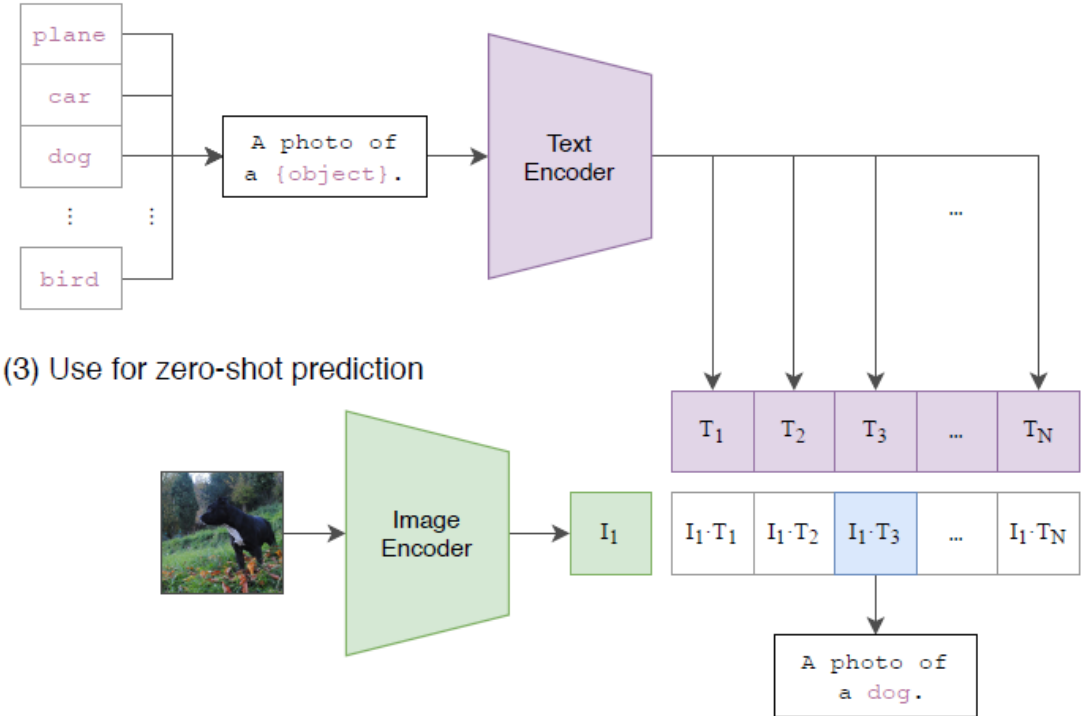
## CLIP

CLIP(Contrastive Language–Image Pre-training) is an **open source, multi-modal, zero-shot model**. Given an image and text descriptions, the model can predict the most relevant text description for that image, without optimizing for a particular task.

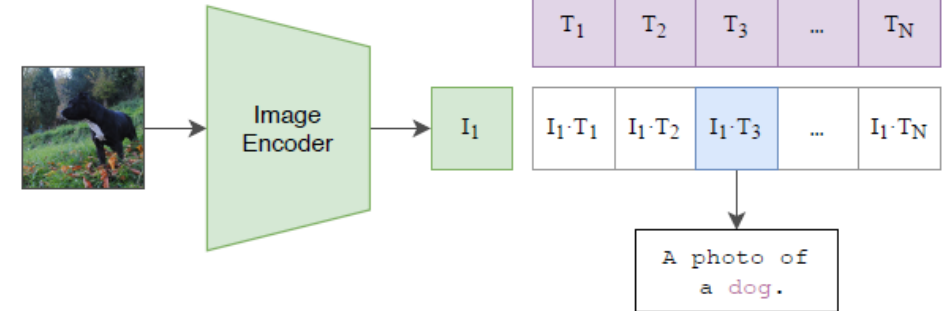
(1) Contrastive pre-training



(2) Create dataset classifier from label text



(3) Use for zero-shot prediction



- Contrastive Language: 비슷한 representation 은 latent space 내에서 가까이 있어야 하고, 비슷하지 않은 representation은 멀리 떨어져 있어야 한다.
- Zero-shot model**: 이전에 한번도 보지 않았던 label에 대해서도 분류할 수 있다.
- Text & image pair를 학습하고 zero-shot 분류기로 image에 맞는 text를 반환



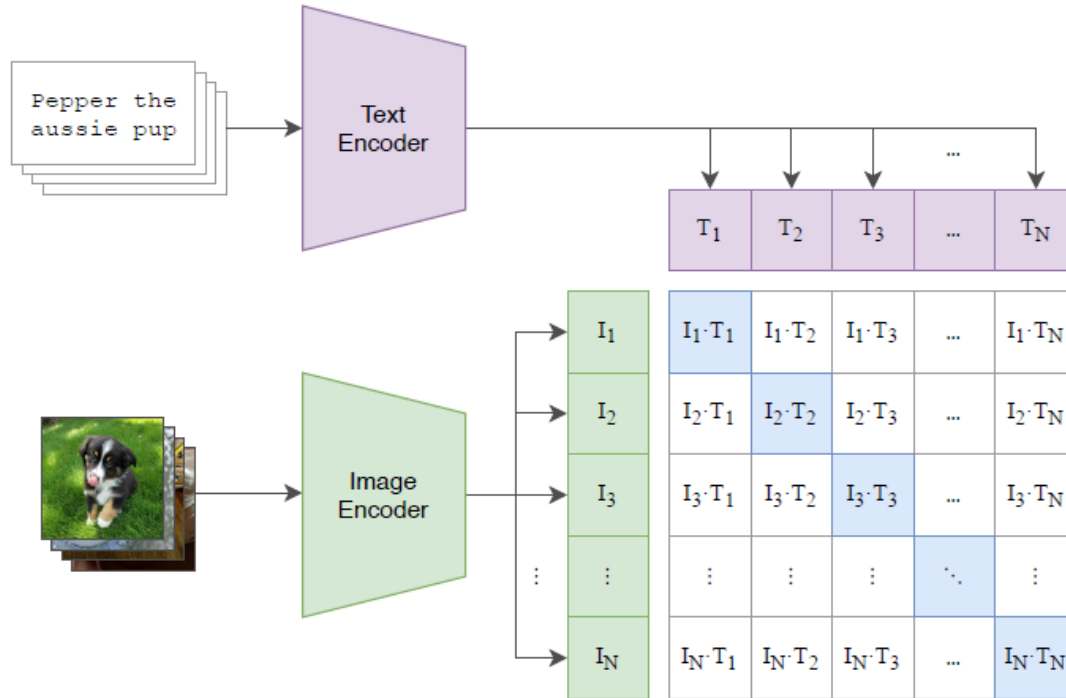


# 모델 설명

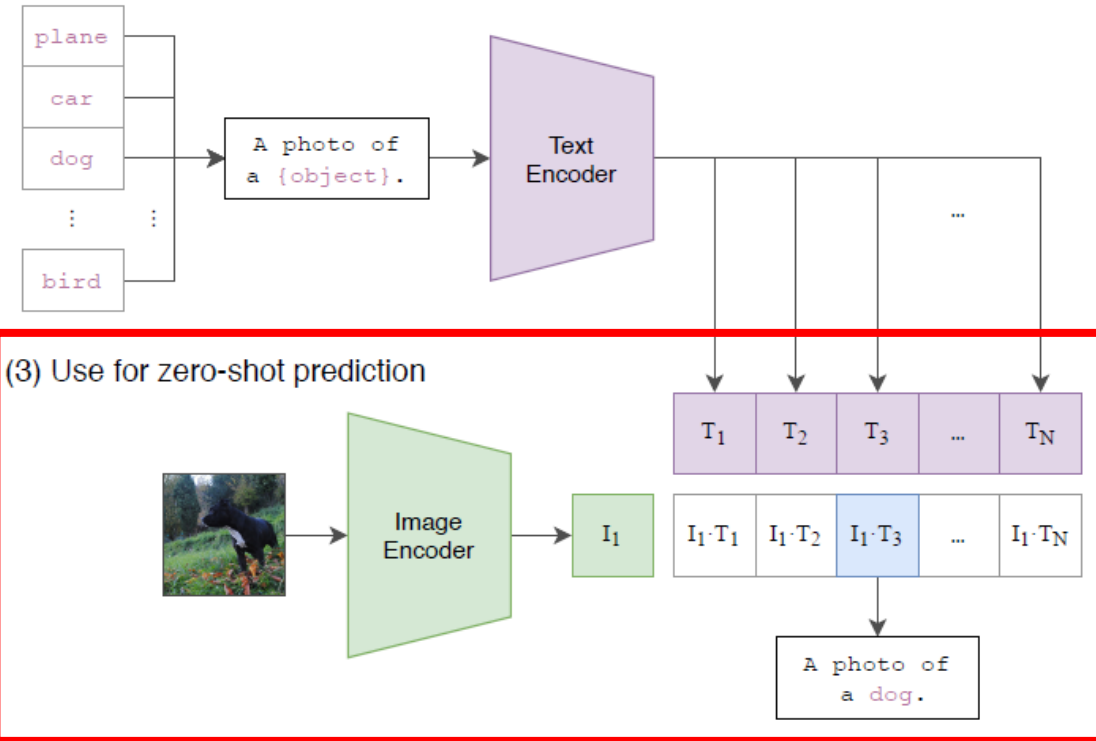
## CLIP

CLIP(Contrastive Language–Image Pre-training) is an **open source, multi-modal, zero-shot model**. Given an image and text descriptions, the model can predict the most relevant text description for that image, without optimizing for a particular task.

(1) Contrastive pre-training



(2) Create dataset classifier from label text



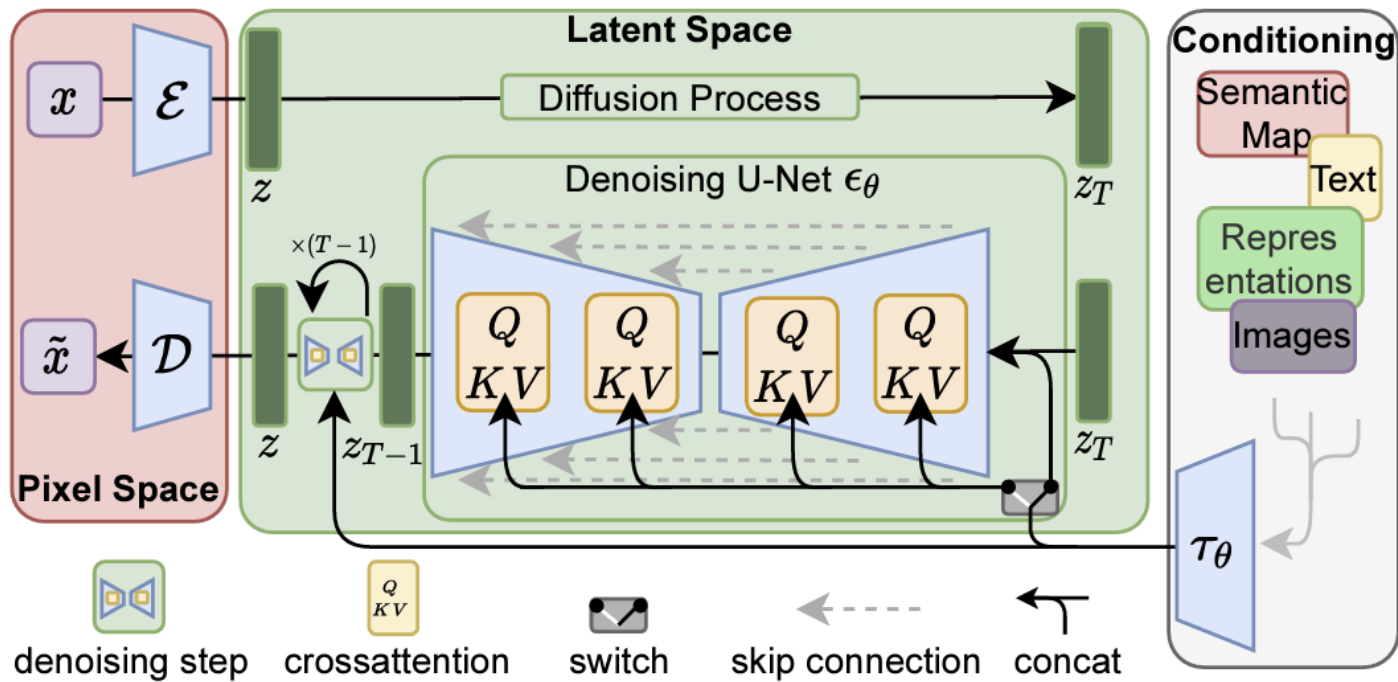
- 과정 : 1) 'A photo of a 라벨' 의 문장 만든 후 토큰화  
2) Text encoder(transformer) 이용  
3) 텍스트/이미지 인코딩 생성  
4) 한 이미지와 라벨의 정보가 담긴 텍스트 인코딩 사이 코사인 유사도 구하기  
5) 가장 유사도가 큰 값을 라벨로 반환



# Stage 1. prompt-to-image

Black and white photo,  
Some bicycles have been left unattended

## Stable Diffusion 1.5





# Stage 2. image-to-prompt

Image & **Correct** answer  
Paired Dataset



One of the men is writing  
on a document.

Train

BLIP - c

Some bicycles are  
parked near a wall  
[Correct]



BLIP - w

A bike is leaning  
against a column  
[Wrong]

Image & **Wrong** answer  
Paired Dataset



One of the men is checking  
his watch.

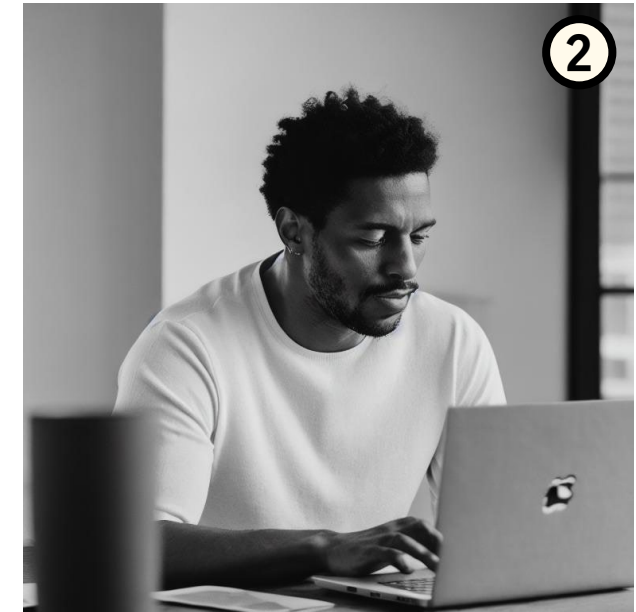
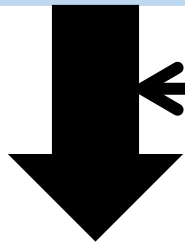
Train



# Stage 3. Validation

CLIP

Dot Product  $\rightarrow$  Softmax



some bicycles are  
parked near a wall

a bike is leaning  
against a column.

the door is no  
longer on the  
hinges

a person sits on a  
bicycle and makes it  
move

he's typing on  
a computer

the man is  
picking up a  
laptop

the man is  
typing on his  
laptop

the shirt is being  
fastened securely

	some bicycles are parked near a wall	a bike is leaning against a column.	the door is no longer on the hinges	a person sits on a bicycle and makes it move	he's typing on a computer	the man is picking up a laptop	the man is typing on his laptop	the shirt is being fastened securely
image 1	99.16	0.09	0.00	0.75	0.00	0.00	0.00	0.00
image 2	0.00	0.00	0.00	0.06	20.48	15.45	63.97	0.04





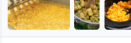
# Trial and Error

Dataset Preview

Size: 364 MB

API

Go to dataset view

id (int32)	image_0 (image)	image_1 (image)	caption_0 (string)	caption_1 (string)	tag (string)	secondary_tag (string)	num_main_preds (int32)	coll (str)
0	image_url (string)	image (image)	text (string)	source (string)	meta (string)			
1	image (image)	scene (sequence)	action (sequence)	rationale (sequence)	object (sequence)	confidence (dict)	purity (dict)	
2		[ "in a beach", "the picture was"	[ "holding an umbrella", "the"	[ "so they won't get a"	[ "Woman in swim suit"	{ "scene": [ 5, 4, 5 ],	{ "scene": [ -1.216816292572	
3	image (image)	verb noun (string)	effect_sentence_list (sequence)	effect_phrases_list (sequence)	positive_image_list (images list)	negative_image_list (images list)		
4	image (image)	"arrange chairs"	[ [ "chairs are moved around in order", "the chairs are..." ] ]	[ [ "are moved", "are moved around" ], [ "are put" ], ... ]				
	image (image)	"arrange flowers"	[ [ "the flowers are in a pretty design", "flowers are..." ] ]	[ [ "in", "in a pretty design" ], [ "are..." ] ]				
	image (image)	"bake potato"	[ [ "i put a potato in the oven to bake it", "the..." ] ]	[ [ "in the oven" ], [ "safe" ], [ "is heated", ... ] ]				
	image (image)	"beat eggs"	[ [ "the eggs are stirred", "the eggs are scrambled", ... ] ]	[ [ "are stirred" ], [ "are scrambled" ], [ "will be..." ] ]				

## Multiple Datasets

- 데이터셋의 종류와 조합(비례배분)에 따라 모델의 성능의 차이가 많이 남(not robust)



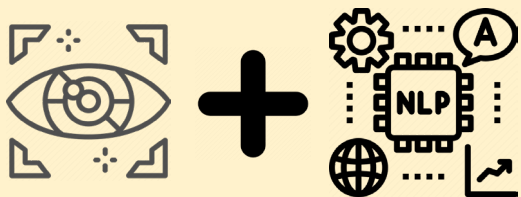
## GPT 2

- 단어 1개 / 단어 뭉치 / 문장 입력하면 이어서 문장 작성 (동일 주제 문장 생성)
- BUT 유사한 의미의 문장은 생성 X  
→ BLIP으로 오답 선지 학습시켜 문장 생성





# 프로젝트의 가치



CV + NLP를 종합적으로  
구현해볼 수 있는 기회



Prompt – image – prompt & deep learning



TOEIC 문제 LC Part 1 자동 출제



우리의 목적에 맞게  
fine-tuning 해볼 수 있는 좋은 기회



**그럼 토익 문제를 만들어봅시다 !**