2023 Summer Seminar

Integrating Image Encoder into LLM

2023. 08. 18

이정섭

1. MAGMA – Multimodal Augmentation of Generative Models through Adapter-based Finetuning, EMNLP-findings 2022

2. **Linearly Mapping from Image to Text Space, ICLR 2023**

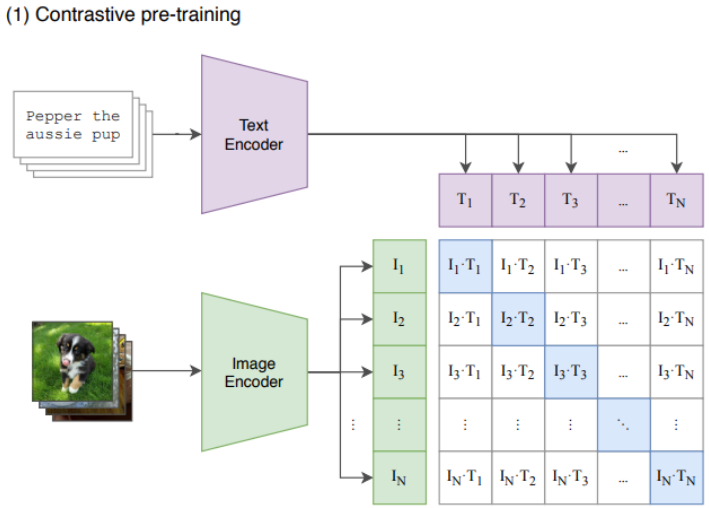
3. **Visual Instruction Tuning, arXiv 2023**

MAGMA - Multimodal Augmentation of Generative Models through Adapter-based Finetuning

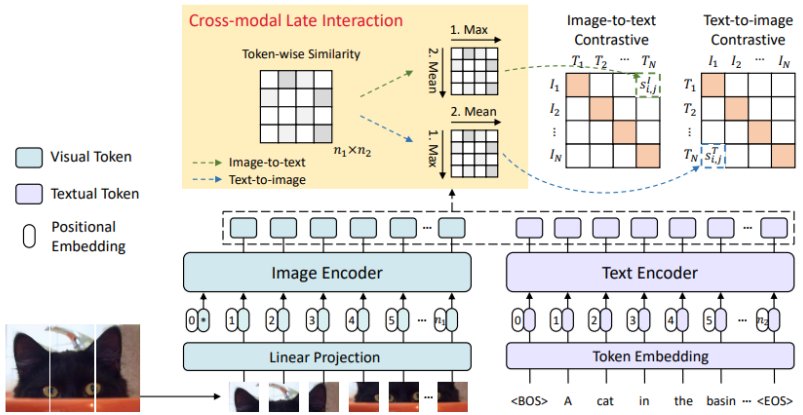
EMNLP-findings 2022

Introduction

◆ CLIP (Contrastive Language-Image Pre-training, arXiv 2021)



◆ FILIP (Fine-grained Interactive Language-Image Pre-Training, ICLR 2022)



◆ DeCLIP (Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm, CVPR 2021)

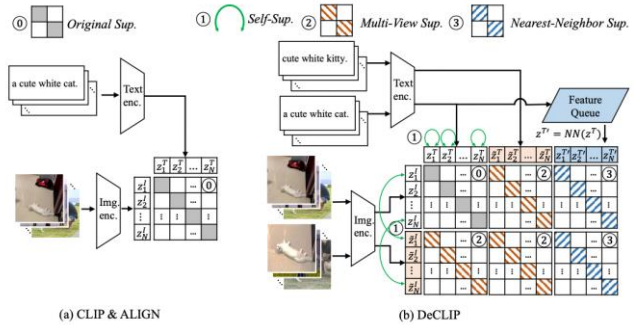
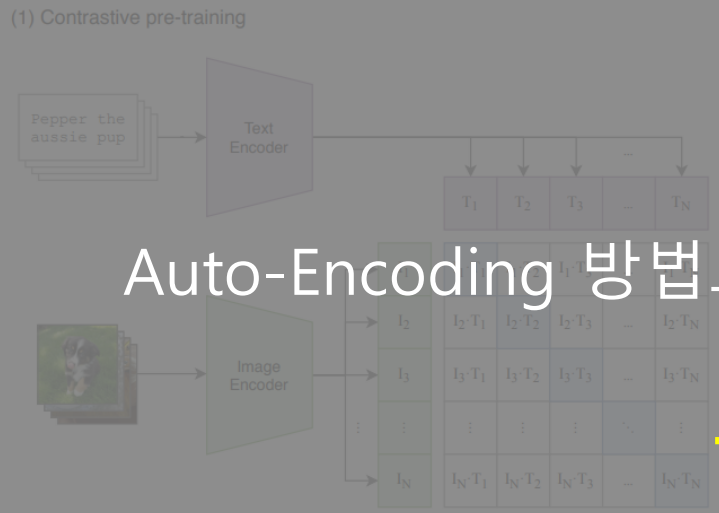


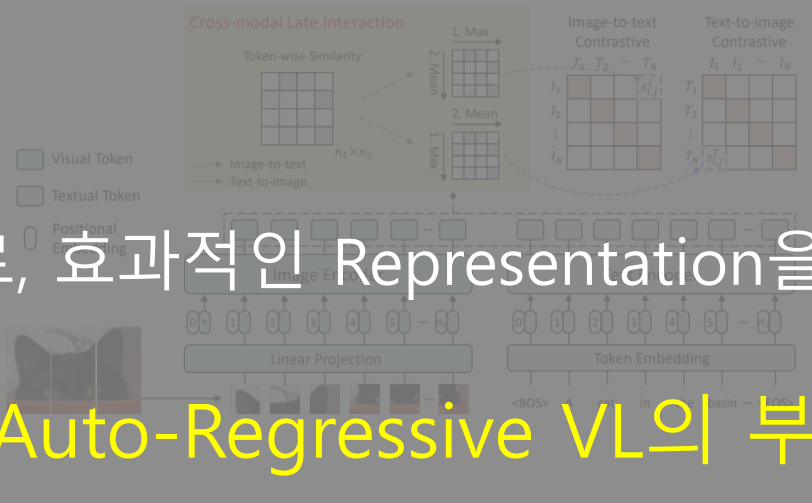
Figure 4: (a) CLIP and ALIGN jointly train an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. (b) Our DeCLIP overview. ① means Self-Supervision(SS). For image SS, we maximize the similarity between two augmented views of the same instance. For text SS, we leverage Masked Language Modeling(MLM) within a text sentence. ② represents cross-modal Multi-View Supervision(MVS). We first have two augmented views of both image and text, then contrast the 2 × 2 image-text pairs. ③ indicates Nearest-Neighbor Supervision(NNS). We sample text NN in the embedding space to serve as additional supervision. The combination of the three supervision leads to efficient multi-modal learning.

Introduction

◆ CLIP (Contrastive Language-Image Pre-training, arXiv 2021)



◆ FILIP (Fine-grained Interactive Language-Image Pre-Training, ICLR 2022)



◆ DeCLIP (Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm, CVPR 2021)

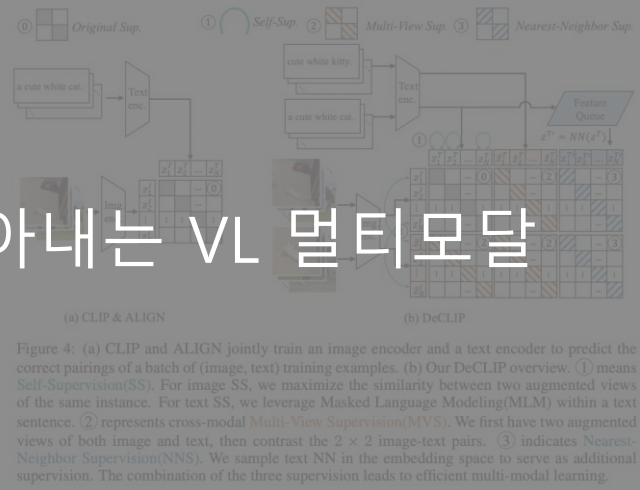


Figure 4: (a) CLIP and ALIGN jointly train an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. (b) Our DeCLIP overview. ① means Self-Supervision(SS). For image SS, we maximize the similarity between two augmented views of the same instance. For text SS, we leverage Masked Language Modeling(MLM) within a text sentence. ② represents cross-modal Multi-View Supervision(MVS). We first have two augmented views of both image and text, then contrast the 2 × 2 image-text pairs. ③ indicates Nearest-Neighbor Supervision(NNS). We sample text NN in the embedding space to serve as additional supervision. The combination of the three supervision leads to efficient multi-modal learning.

Auto-Encoding 방법으로, 효과적인 Representation을 뽑아내는 VL 멀티모달

→ Auto-Regressive VL의 부재

Introduction

i) **이전의 VL 연구에서는 이미지-텍스트에 대해 생성적 방식을 적용 하지 않음**

→ CLIP, DeCLIP 등의 연구는 인코딩을 초점으로 함

ii) **이전 VL 연구는 과도하게 많은 양의 사전 훈련 데이터와 언어 및 비전 구성 요소의 동시 훈련이 필요**

→ CLIP은 4억개의 Image-Text Pair 데이터셋을 사용

→ CLIP 및 DeCLIP은 이미지 인코더와 텍스트 인코더 모두 학습

공개된 pre-trained LM와 Vision 모델을 결합하여

강력한 multimodal 모델을 만드는 Auto-regressive VL 모델 MAGMA 소개

Examples

A picture of *an apple on a table.*

A picture of *an apple with a library sign on it*

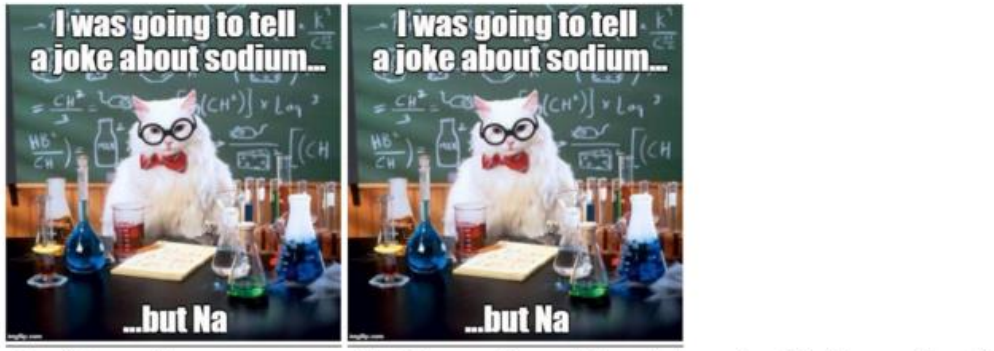
A picture of *an apple with a label on it that says iPod*

Q: What does the yellow street sign mean? A: pedestrian crossing

Q: Are the zebras walking? A: no

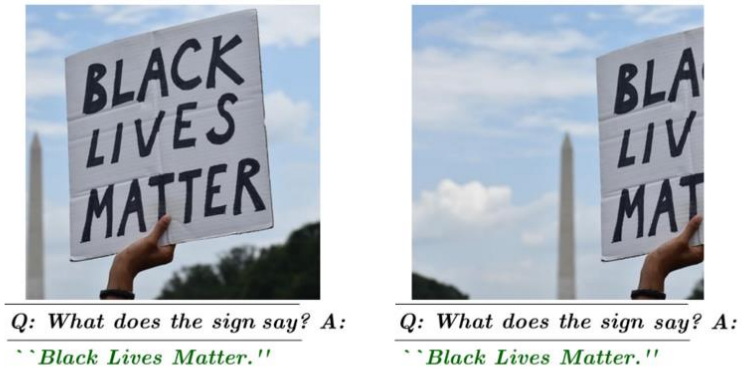
Q: What sound do these animals make? *A: bleating*

Figure 3: An example of a *2-shot* prompt for OKVQA.



A picture of *a cat in a lab coat.*

A picture of a cat in a lab coat, with the caption "I was going to tell a joke about sodium, but Na"

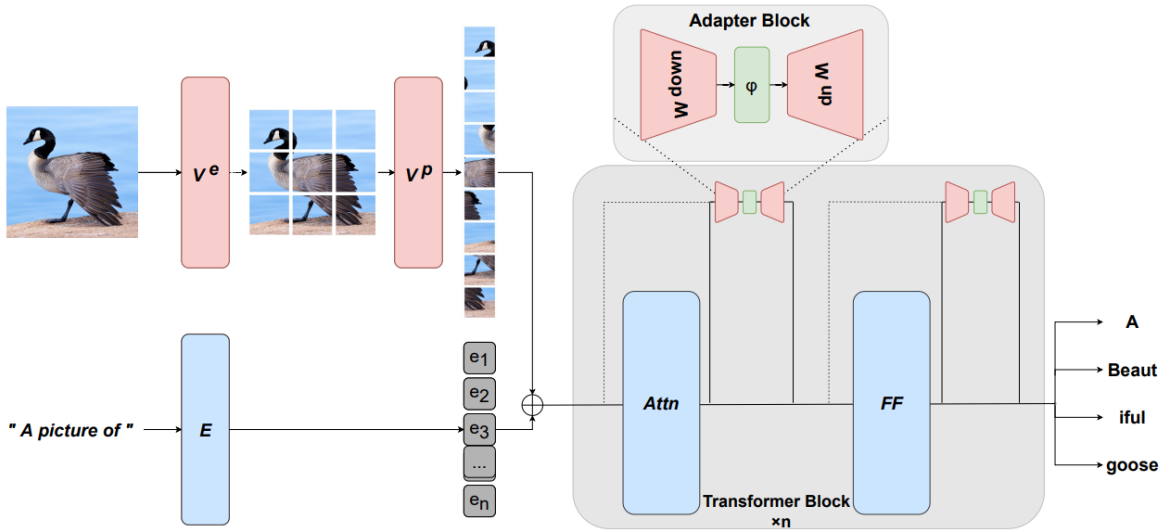


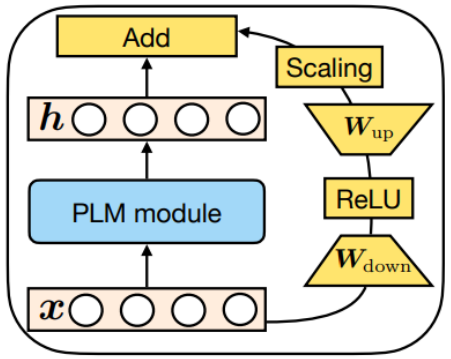
Q: What does the sign say? A: *``Black Lives Matter.''*

Q: What does the sign say? A: *``Black Lives Matter.''*

Figure 4: MAGMA's OCR capabilities. Even when text is obscured, MAGMA imputes the missing values.

Methodology





(e) Scaled PA

Figure 2: MAGMA's architecture. The layers in red are trained, and the layers in blue remain frozen.

i) **images are fed into a Visual Encoder (CLIP-visual backbone)**, which processes the raw image input and outputs a sequence of feature vectors

ii) Then an **Image Prefix module maps image features into a sequence of embedding vectors** that are input to the third model component, an auto-regressive Language Model (GPT-J 6B).

iii) Adapter: scaled residual bottleneck MLP

Evaluation Results

	VQA	OKVQA	GQA	VizWiz	SNLI-VE	NoCaps		Coco	
						CIDEr	B@4	CIDEr	B@4
MAGMA	68.0	**49.2**	54.5	35.4	79.0	93.6	27.8	91.2	31.4
SOTA	**75.5**	48.0	**72.1**	**54.7**	**86.3**	**112.2**	**33.1**	**143.3**	**41.7**
SOTA model	*SimVLM*	*PICa*	*CFR*	*Pythia*	*SimVLM*	*SimVLM*	*VIVO*	*SimVLM*	*OSCAR*

Table 2: MAGMA finetuned performance. **B@4**: NoCaps-all score. SOTA scores are to the best of our knowledge at the time of writing. If available/applicable, we compare to the SOTA score of models solving the task in an open-ended generative fashion like MAGMA (notably *SimVLM* on VQA), otherwise we compare to the general SOTA (classification setting). Models: *SimVLM* (Wang et al., 2021), *PICa* (Yang et al., 2021), *CFR* (Nguyen et al., 2021), *Pythia* (Singh et al., 2019), *VIVO* (Hu et al., 2020), *OSCAR* (Li et al., 2020).

Linearly Mapping from Image to Text Space

ICLR2023

Objective

- **인간은 비언어적 현상(non-linguistic phenomena)을 일반화하고 추론할 수 있는 반면, 정해진 form의 text data로 학습된 vision 모델은 인간처럼 풍부한 conceptual knowledge를 반영할 수 없다**

- **vision 모델과 language 모델의 conceptual representations은 기능적으로 동일한가?**

→ 혹은 linguistic supervision에 따라 그 정도가 다른가?

- **이를 위해 pre-training에서 linguistic supervision의 강도가 다른 3가지의 이미지 인코더를 사용**

- **BEiT:** 언어에 노출되지 않았고, 마스크된 이미지 섹션의 contents를 예측하는 방법으로 학습 **(w/o linguistic supervision)**
- **NFRN50 (Normalizer Free Resnet50):** WordNet hypo/hypernym 구조에 따라 labeled 데이터에 대한 이미지 분류 작업에서 사전 훈련 후, 300M 이미지 클래스 데이터로 학습 **(약간의 linguistic supervision)**
- **CLIP:** shared image-text representation space에서, full natural language captions으로 이미지를 align 하도록 사전 훈련 **(w/ linguistic supervision)**

Linguistic Supervision의 강도

- **CLIP > NFRN50 > BEiT**

Methodology

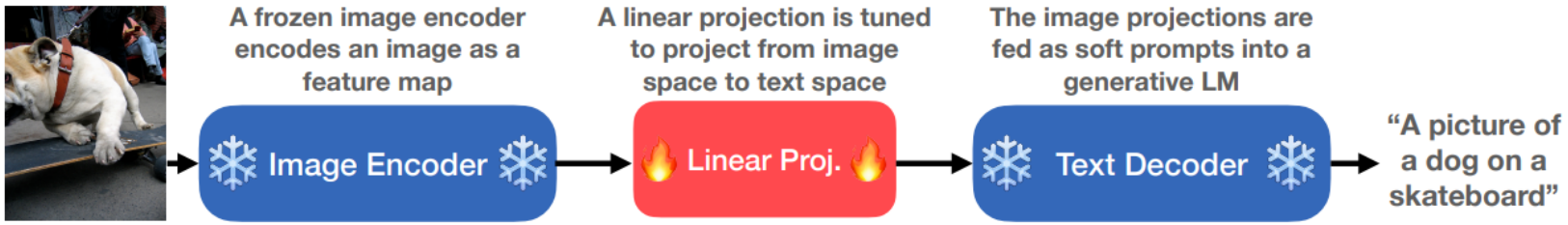


Figure 1: We train linear projections from image representations into the input space of a language model to produce captions describing images. We find that LMs can describe the contents of most image representations, but performance varies based on the type of image encoder used.

Linear Projection

→ **"Image encoder의 representation과 LM의 representation은 동일한 space에 위치?"**

Experimental Results

Image Captioning	NoCaps - CIDEr-D				NoCaps (All)		CoCo	CoCo	
	In	Out	Near	All	CLIP-S	Ref-S	CIDEr-D	CLIP-S	Ref-S
🔥NFRN50 Tuned	20.9	30.8	25.3	27.3	66.5	72.5	35.3	69.7	74.8
🔥MAGMA (released)	18.0	12.7	18.4	16.9	63.2	68.8	**52.1**	76.7	79.4
🔥MAGMA (ours)	**30.4**	**43.4**	**36.7**	**38.7**	74.3	78.7	47.5	75.3	**79.6**
❄BEIT Random	5.5	3.6	4.1	4.4	46.8	55.1	5.2	48.8	56.2
❄NFRN50 Random	5.4	4.0	4.9	5.0	47.5	55.7	4.8	49.5	57.1
❄BEIT	20.3	16.3	18.9	18.9	62.0	69.1	22.3	63.6	70.0
❄NFRN50	21.3	31.2	26.9	28.5	65.6	71.8	36.2	68.9	74.1
❄BEIT FT.	**38.5**	**48.8**	**43.1**	**45.3**	73.0	78.1	51.0	74.2	78.9
❄CLIP	34.3	48.4	41.6	43.9	**74.7**	**79.4**	**54.9**	**76.2**	**80.4**
VQA n-shots	0				1		**2**	4	
Blind	20.60				35.11		36.17	36.99	
🔥NFRN50 Tuned	27.15				37.47		38.48	39.18	
🔥MAGMA (ours)	24.62				39.27		40.58	41.51	
🔥MAGMA (reported)	32.7				40.2		**42.5**	43.8	
❄NFRN50 Random	25.34				36.15		36.79	37.43	
❄BEIT	24.92				34.35		34.70	31.72	
❄NFRN50	27.63				37.51		38.58	39.17	
❄CLIP	33.33				39.93		**40.82**	40.34	

"**[image] Q: [q] A:**" 의 프롬프트로 VQA 진행

VQA는 Accuracy 기반 평가

Linguistic Supervision의 강도:

- **CLIP < NFRN50 < BEiT**

CIDEr-D (Vedantam et al., 2015) **:** 시각적으로 유용할 정확한 단어를 생성할때 점수 ↑

CLIP-S (CLIPscore): reference 없이 이미지와 캡션 간의 유사성 평가

RefCLIPScore: reference를 포함하여 이미지와 캡션 간의 유사성 평가

Experimental Results

Image Captioning	NoCaps - CIDEr-D				NoCaps (All)		CoCo	CoCo	
	In	Out	Near	All	CLIP-S	Ref-S	CIDEr-D	CLIP-S	Ref-S
🔥NFRN50 Tuned	20.9	30.8	25.3	27.3	66.5	72.5	35.3	69.7	74.8
🔥MAGMA (released)	18.0	12.7	18.4	16.9	63.2	68.8	**52.1**	76.7	79.4
🔥MAGMA (ours)	**30.4**	**43.4**	**36.7**	**38.7**	74.3	78.7	47.5	75.3	**79.6**
❄BEIT Random	5.5	3.6	4.1	4.4	46.8	55.1	5.2	48.8	56.2
❄NFRN50 Random	5.4	4.0	4.9	5.0	47.5	55.7	4.8	49.5	57.1
❄BEIT	20.3	16.3	18.9	18.9	62.0	69.1	22.3	63.6	70.0
❄NFRN50	21.3	31.2	26.9	28.5	65.6	71.8	36.2	68.9	74.1
❄BEIT FT.	**38.5**	**48.8**	**43.1**	**45.3**	73.0	78.1	51.0	74.2	78.9
❄CLIP	34.3	48.4	41.6	43.9	**74.7**	**79.4**	**54.9**	**76.2**	**80.4**

NFRN50 Tuned: frozen LM, Visual Encoder 학습 → LM space에 Visual Encoder를 **fit**

BEiT FT.: 기존 BEiT은 MIM으로 Linguistic Supervision이 없었는데,

BEiT ckpt에 ImageNet-22k로 fine-tuning하여 Linguistic Supervision 주입

Experimental Results

Image Captioning	NoCaps - CIDEr-D				NoCaps (All)		CoCo	CoCo	
	In	Out	Near	All	CLIP-S	Ref-S	CIDEr-D	CLIP-S	Ref-S
🔥NFRN50 Tuned	20.9	30.8	25.3	27.3	66.5	72.5	35.3	69.7	74.8
🔥MAGMA (released)	18.0	12.7	18.4	16.9	63.2	68.8	**52.1**	76.7	79.4
🔥MAGMA (ours)	**30.4**	**43.4**	**36.7**	**38.7**	74.3	78.7	47.5	75.3	**79.6**
❄BEIT Random	5.5	3.6	4.1	4.4	46.8	55.1	5.2	48.8	56.2
❄NFRN50 Random	5.4	4.0	4.9	5.0	47.5	55.7	4.8	49.5	57.1
❄BEIT	20.3	16.3	18.9	18.9	62.0	69.1	22.3	63.6	70.0
❄NFRN50	21.3	31.2	26.9	28.5	65.6	71.8	36.2	68.9	74.1
❄BEIT FT.	**38.5**	**48.8**	**43.1**	**45.3**	73.0	78.1	51.0	74.2	78.9
❄CLIP	34.3	48.4	41.6	43.9	**74.7**	**79.4**	**54.9**	**76.2**	**80.4**

VQA n-shots	0	1	2	4
Blind	20.60	35.11	36.17	36.99
🔥NFRN50 Tuned	27.15	37.47	38.48	39.18
🔥MAGMA (ours)	24.62	39.27	40.58	41.51
🔥MAGMA (reported)	32.7	40.2	**42.5**	43.8
❄NFRN50 Random	25.34	36.15	36.79	37.43
❄BEIT	24.92	34.35	34.70	31.72
❄NFRN50	27.63	37.51	38.58	39.17
❄CLIP	33.33	39.93	**40.82**	40.34

Jointly-tuned Models vs **Linearly Mapping Models**

Jointly-tuned < Linearly Mapping

→ LM 디코더를 학습하는 것이 큰 의미는 없다

Experimental Results

Image Captioning	NoCaps - CIDEr-D				NoCaps (All)		CoCo	CoCo	
	In	Out	Near	All	CLIP-S	Ref-S	CIDEr-D	CLIP-S	Ref-S
NFRN50 Tuned	20.9	30.8	25.3	27.3	66.5	72.5	35.3	69.7	74.8
MAGMA (released)	18.0	12.7	18.4	16.9	63.2	68.8	**52.1**	76.7	79.4
MAGMA (ours)	**30.4**	**43.4**	**36.7**	**38.7**	74.3	78.7	47.5	75.3	**79.6**
BEIT Random	5.5	3.6	4.1	4.4	46.8	55.1	5.2	48.8	56.2
NFRN50 Random	5.4	4.0	4.9	5.0	47.5	55.7	4.8	49.5	57.1
BEIT	20.3	16.3	18.9	18.9	62.0	69.1	22.3	63.6	70.0
NFRN50	21.3	31.2	26.9	28.5	65.6	71.8	36.2	68.9	74.1
BEIT FT	**38.5**	**48.8**	**43.1**	**45.3**	73.0	78.1	51.0	74.2	78.9
CLIP	34.3	48.4	41.6	43.9	**74.7**	**79.4**	**54.9**	**76.2**	**80.4**

VQA n-shots	0	1	**2**	4
Blind	20.60	35.11	36.17	36.99
NFRN50 Tuned	27.15	37.47	38.48	39.18
MAGMA (ours)	24.62	39.27	40.58	41.51
MAGMA (reported)	32.7	40.2	**42.5**	43.8
NFRN50 Random	25.34	36.15	36.79	37.43
BEIT	24.92	34.35	34.70	31.72
NFRN50	27.63	37.51	38.58	39.17
CLIP	33.33	39.93	**40.82**	40.34

Linguistic Supervision의 관점

Linguistic Supervision의 세기

CLIP > NFRN50 > BEIT이 성능과 일치

Experimental Results

Image Captioning	NoCaps - CIDEr-D				NoCaps (All)		CoCo	CoCo	
	In	Out	Near	All	CLIP-S	Ref-S	CIDEr-D	CLIP-S	Ref-S
🔥NFRN50 Tuned	20.9	30.8	25.3	27.3	66.5	72.5	35.3	69.7	74.8
🔥MAGMA (released)	18.0	12.7	18.4	16.9	63.2	68.8	**52.1**	76.7	79.4
🔥MAGMA (ours)	**30.4**	**43.4**	**36.7**	**38.7**	74.3	78.7	47.5	75.3	**79.6**
❄BEIT Random	5.5	3.6	4.1	4.4	46.8	55.1	5.2	48.8	56.2
❄NFRN50 Random	5.4	4.0	4.9	5.0	47.5	55.7	4.8	49.5	57.1
❄BEIT	20.3	16.3	18.9	18.9	62.0	69.1	22.3	63.6	70.0
❄NFRN50	21.3	31.2	26.9	28.5	65.6	71.8	36.2	68.9	74.1
❄BEIT FT.	**38.5**	**48.8**	**43.1**	**45.3**	73.0	78.1	51.0	74.2	78.9
❄CLIP	34.3	48.4	41.6	43.9	**74.7**	**79.4**	**54.9**	**76.2**	**80.4**

VQA n-shots	0	1	**2**	4
Blind	20.60	35.11	36.17	36.99
🔥NFRN50 Tuned	27.15	37.47	38.48	39.18
🔥MAGMA (ours)	24.62	39.27	40.58	41.51
🔥MAGMA (reported)	32.7	40.2	**42.5**	43.8
❄NFRN50 Random	25.34	36.15	36.79	37.43
❄BEIT	24.92	34.35	34.70	31.72
❄NFRN50	27.63	37.51	38.58	39.17
❄CLIP	33.33	39.93	**40.82**	40.34

BEIT vs **BEIT FT**

Linguistic Supervision이 전혀 없는 BEiT 모델에, ImageNet-22k 데이터를 fine-tuning하면

Linguistic information을 넣어주면 Image와 Text Space가 유사해짐을 발견

Experimental Results

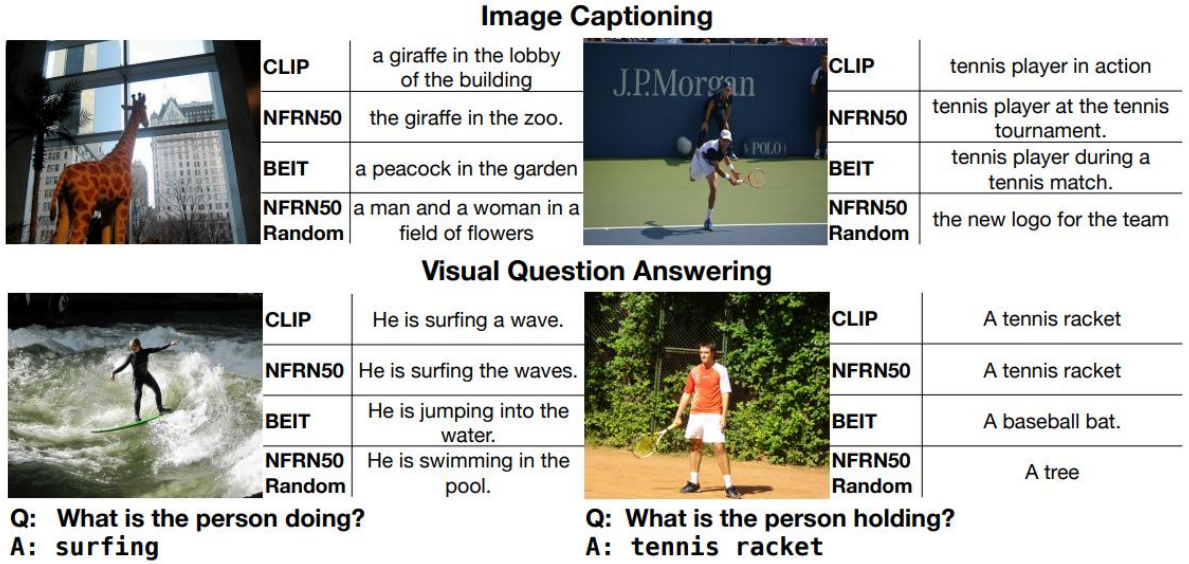
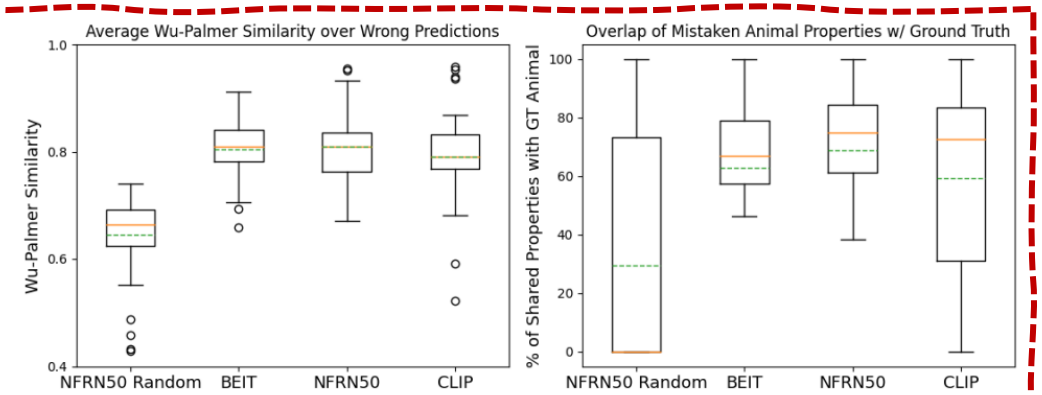


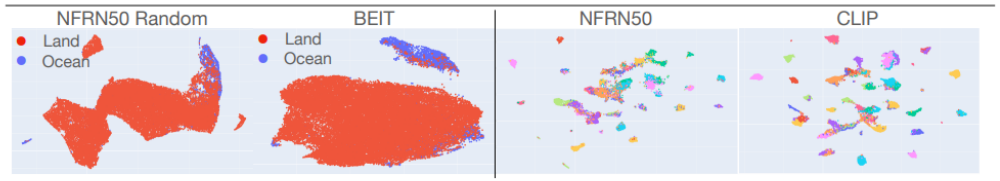
Figure 2: Curated examples of captioning and zero-shot VQA illustrating the ability of each model to transfer information to the LM without tuning either model. We use these examples to also illustrate common failure modes for BEIT prompts of sometimes generating incorrect but conceptually related captions/answers.

Linguistic Supervision이 부족한 **BEiT 모델**은 다음과 같은 실수를 범함

Case Studies



(a) **Left**: Wu-Palmer Similarity for captions in which the models don't mention the animal show that BEIT, NFRN50, and CLIP are all similarly close, meaning that even if they predict the wrong animal, it is on average very taxonomically similar. **Right**: When the model mistakes one animal for another in the dataset, how similar are the AWA properties for the true animal and the one it mistakes it most for? The average number of overlapping properties show that animals predicted from BEIT are at least as similar to the real animal as NFRN50 and CLIP. Median is shown as the solid orange line while the dashed green line shows the mean.



(b) **UMAP projections of AWA images:**While NFRN50 and CLIP cluster tightly along lexical categories (color coded by animal), BEIT clusters the most distinctly along animals that live in water/the ocean; the randomly initialized NFRN50 mostly randomly overlap in one cluster.

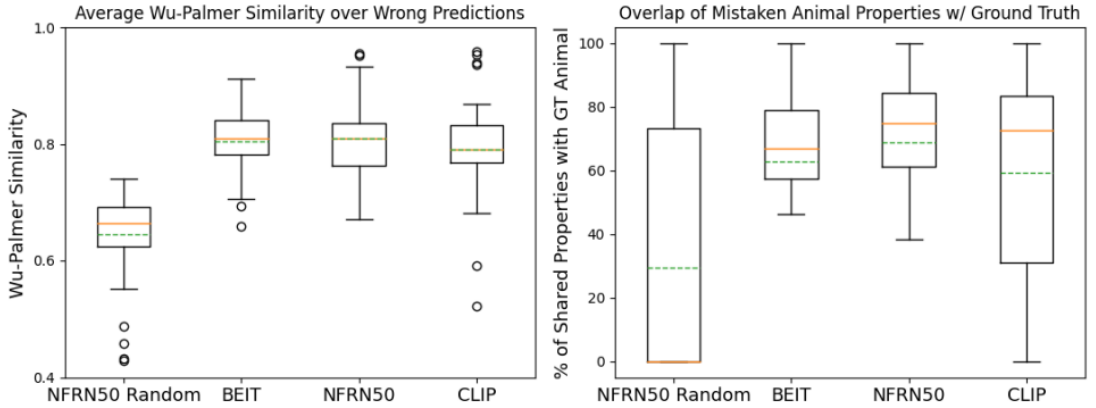
Figure 4

Wu-Palmer Similarity: WordNet 분류법에서 실제 단어와 생성된 단어 사이의 거리를 계산하여 단어가 정답에 얼마나 근접했는지 측정하는 방법

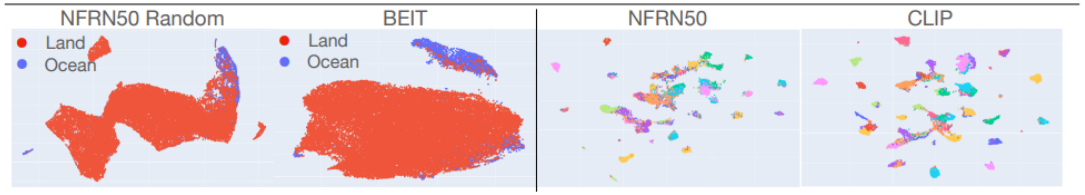
Data: 50개의 동물 클래스를 포함하는 37,000개의 총 이미지를 포함하는 AWA(Animals With Attributes 2) 데이터셋

- 각 동물 클래스는 동물을 설명하는 85가지 속성(예: '발톱', ' 줄무늬 ' , ' 정글 ')에 대한 주석을 제공.
- Caption을 생성했을 때, WordNet synset of the ground truth animal label에 언급된 동물의 similarity로 평가
- **Accuracy** → CLIP**:** 59%, NFRN50: 43%, BEIT 13%, NFRN50 Random: 0.4%

Case Studies



(a) **Left**: Wu-Palmer Similarity for captions in which the models don't mention the animal show that BEIT, NFRN50, and CLIP are all similarly close, meaning that even if they predict the wrong animal, it is on average very taxonomically similar. **Right**: When the model mistakes one animal for another in the dataset, how similar are the AWA properties for the true animal and the one it mistakes it most for? The average number of overlapping properties show that animals predicted from BEIT are at least as similar to the real animal as NFRN50 and CLIP. Median is shown as the solid orange line while the dashed green line shows the mean.



(b) **UMAP projections of AWA images:**While NFRN50 and CLIP cluster tightly along lexical categories (color coded by animal), BEIT clusters the most distinctly along animals that live in water/the ocean; the randomly initialized NFRN50 mostly randomly overlap in one cluster.

Figure 4

UMAP Projections of AWA images:

- **NFRN50, CLIP**은 (동물별로 색으로 구분)된 동물을 밀접하게 클러스터링
- **BEIT**은 물/바다에 사는 동물을 따라 가장 뚜렷하게 클러스터
- **Random NFRN50**은 대부분 하나의 클러스터에서 겹침

→ CLIP, NFRN50은 Language Supervision으로 Fine-grained conceptual representation을 생성할 수 있지만,

→ BEiT은 Language Supervision을 진행하지 않아 Coarse-grained conceptual representation을 생성

Conclusion

- 세 개의 이미지 인코더 중 하나로 **LM을 프롬프트하면, 이미지의 의미론적 내용이 LM에 효과적으로 전송됨**을 보여줌
- **성능은 또한 이미지 인코더가 가진 linguistic supervision의 강도에 비례**
- **BEiT는 대부분 Coarse-grained visual concepts를 전송**하고, LM이 정확한 어휘 범주를 생성하게 함에 있어 어려움을 겪는 것으로 보임
  - → LM이 이미지 데이터에 대해 directly trained model 의해 학습된 것과 구조적으로 매우 유사한 conceptual space를 학습하지만, 정확한 similarity는 이미지 인코더가 받는 supervision 유형에 따라 다르다는 증거로 해석
  - → linguistic supervision의 강도에 따라 Fine-grained visual concepts를 전송

Visual Instruction Tuning

arXiv 2023

Introduction

- 사용자의 Instruct로 LM과 인간이 상호작용 하는 연구가 대두되고 있지만, VLM에서 이는 고려되지 않고 있음

- 본 연구에서는 Visual Instruct Tuning을 진행하는 방법에 대해 다룸
  - Visual Instruct 데이터셋 제작
  - Visual Instruct 모델 학습

Introduction

- 사용자의 Instruct로 LM과 인간이 상호작용 하는 연구가 대두되고 있지만, VLM에서 이는 고려되지 않고 있음

MAGMA 연구 ⇔ GPT-2, GPT-3 연구

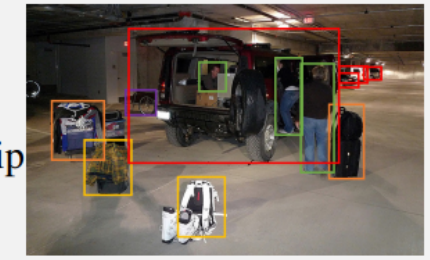
- 본 연구에서는 Visual Instruct Tuning을 진행하는 방법에 대해 다룸
  - Visual Instruct 데이터셋 제작
  - Visual Instruct 모델 학습

LLaVA 연구 ⇔ ChatGPT 연구

GPT-assisted Visual Instruction Data Generation

Context type 1: Captions

A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.



Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], person: [0.63, 0.222, 0.686, 0.516], person: [0.444, 0.233, 0.487, 0.34], backpack: [0.384, 0.696, 0.485, 0.914], backpack: [0.755, 0.413, 0.846, 0.692], suitcase: [0.758, 0.413, 0.845, 0.69], suitcase: [0.1, 0.497, 0.173, 0.579], bicycle: [0.282, 0.363, 0.327, 0.442], car: [0.786, 0.25, 0.848, 0.322], car: [0.783, 0.27, 0.827, 0.335], car: [0.86, 0.254, 0.891, 0.3], car: [0.261, 0.101, 0.787, 0.626]

3개의 Response type으로 GPT-4를 통해 Visual Instruction 데이터 생성

- **Conversation**
- **Detailed description**
- **complex reasoning**

=> 총 158k의 데이터셋 생성

Response type 1: conversation

Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV).
Question: Where is the vehicle parked?

Response type 3: complex reasoning

Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the

Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip.

GPT-assisted Visual Instruction Data Generation

Instructions for brief image description. The list of instructions used to briefly describe the image content are shown in Table 8. They present the same meaning with natural language variance.

- "Describe the image concisely."
- "Provide a brief description of the given image."
- "Offer a succinct explanation of the picture presented."
- "Summarize the visual content of the image."
- "Give a short and clear explanation of the subsequent image."
- "Share a concise interpretation of the image provided."
- "Present a compact description of the photo's key features."
- "Relay a brief, clear account of the picture shown."
- "Render a clear and concise summary of the photo."
- "Write a terse but informative summary of the picture."
- "Create a compact narrative representing the image presented."

Table 8: The list of instructions for brief image description.

Instructions for detailed image description. The list of instructions used to describe the image content in detail are shown in Table 9. They present the same meaning with natural language variance.

- "Describe the following image in detail"
- "Provide a detailed description of the given image"
- "Give an elaborate explanation of the image you see"
- "Share a comprehensive rundown of the presented image"
- "Offer a thorough analysis of the image"
- "Explain the various aspects of the image before you"
- "Clarify the contents of the displayed image with great detail"
- "Characterize the image using a well-detailed description"
- "Break down the elements of the image in a detailed manner"
- "Walk through the important details of the image"
- "Portray the image with a rich, descriptive narrative"
- "Narrate the contents of the image with precision"
- "Analyze the image in a comprehensive and detailed manner"
- "Illustrate the image through a descriptive explanation"
- "Examine the image closely and share its details"
- "Write an exhaustive depiction of the given image"

Table 9: The list of instructions for detailed image description.

```
messages = [ {"role":"system", "content": f"""You are an AI visual assistant, and you are seeing a single image. What you see are provided with five sentences, describing the same image you are looking at. Answer all questions as you are seeing the image.

Design a conversation between you and a person asking about this photo. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question. Ask diverse questions and give corresponding answers.

Include questions asking about the visual content of the image, including the object types, counting the objects, object actions, object locations, relative positions between objects, etc. Only include questions that have definite answers:
(1) one can see the content in the image that the question asks about and can answer confidently;
(2) one can determine confidently from the image that it is not in the image. Do not ask any question that cannot be answered confidently.

Also include complex questions that are relevant to the content in the image, for example, asking about background knowledge of the objects in the image, asking to discuss about events happening in the image, etc. Again, do not ask about uncertain details. Provide detailed answers when answering complex questions. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized. You can include multiple paragraphs if necessary."""}
]
for sample in fewshot_samples:
    messages.append({"role":"user", "content":sample['context']})
    messages.append({"role":"assistant", "content":sample['response']} )
messages.append({"role":"user", "content":'\n'.join(query)})
```

GPT-assisted Visual Instruction Data Generation

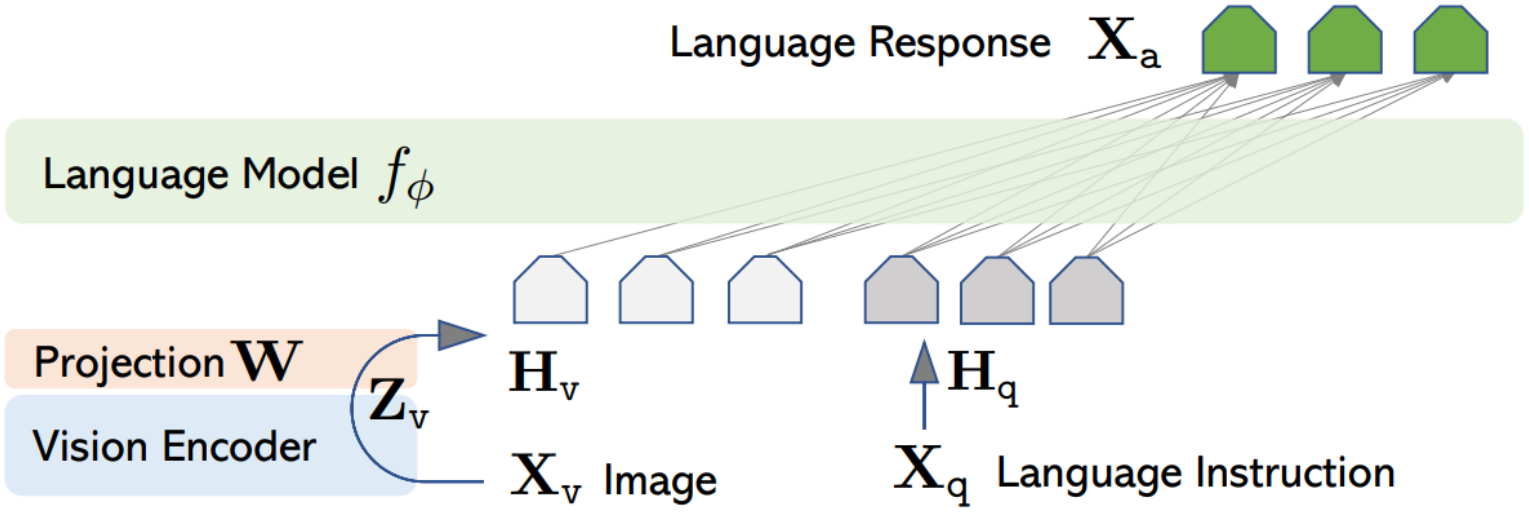
Dataset Viewer Auto-converted to Parquet API

Split

train (394k rows)

image (string)	conversations (list)	id (string)
"000000215677.jpg"	[{ "from": "human", "value": "<image>\nWhat skill set might someone need to perform such a frisbee trick?" }, { "from": "gpt", "value": "To perform the frisbee trick shown in the image, where the man is passing a frisbee between or underneath his legs, a person would need a combination of skills. These skills include good hand-eye coordination, agility, balance, flexibility, and dexterity. Additionally, the ability to throw and catch the frisbee accurately while maintaining control of bodily movements would also be essential. To perfect the trick, practicing these skills and building up muscle memory through repetition would be beneficial." }]	"000000215677"
"000000296754.jpg"	[{ "from": "human", "value": "<image>\nWhat precautions is the woman taking while walking in this weather?" }, {…	"000000296754"
"000000543038.jpg"	[{ "from": "human", "value": "What made the photographs unique and helped to create a memorable moment?\n<image>"…	"000000543038"

Visual Instruction Tuning



LLM (fφ(·)): LLaMA

Vision Encoder: CLIP-ViT-L

Figure 1: LLaVA network architecture.

$$\mathbf{X}_{\text{instruct}}^{t} = \begin{cases} \text{Random choose } [\mathbf{X}_{\text{q}}^{1}, \mathbf{X}_{\text{v}}] \text{ or } [\mathbf{X}_{\text{v}}, \mathbf{X}_{\text{q}}^{1}], & \text{the first turn } t = 1 \\ \mathbf{X}_{\text{q}}^{t}, & \text{the remaining turns } t > 1 \end{cases}$$

$$p(\mathbf{X}_{\text{a}} | \mathbf{X}_{\text{v}}, \mathbf{X}_{\text{instruct}}) = \prod_{i=1}^{L} p_{\boldsymbol{\theta}}(\boldsymbol{x}_i | \mathbf{X}_{\text{v}}, \mathbf{X}_{\text{instruct},<i}, \mathbf{X}_{\text{a},<i}),$$

Visual Instruction Tuning

Stage 1: Pre-training for Feature Alignment.

- CC3M을 595k로 필터링 후, instruction-following data로 변환한 데이터 학습
  - 1) Image Description 중 하나를 랜덤으로 뽑아 이미지 설명 요청
  - 2) 원본 이미지의 caption을 정답으로 학습
  - 3) Mapping Layer만 업데이트 (LLM, Vision Encoder 고정)

Stage 2: Fine-tuning End-to-End.

- Multi-model Chatbot 시나리오
  - 158k 데이터에서 2개의 single turn, 1개의 multi-turn으로 샘플링하여 학습
- Science QA 시나리오
  - 과학 질문 이미지 추론 데이터 학습

Experimental Results

1) Multimodal Chatbot 시나리오

	Conversation	Detail description	Complex reasoning	**All**
Full data	83.1	75.3	96.5	85.1
Detail + Complex	81.5 (-1.6)	73.3 (-2.0)	90.8 (-5.7)	81.9 (-3.2)
Conv + 5% Detail + 10% Complex	81.0 (-2.1)	68.4 (-7.1)	91.5 (-5.0)	80.5 (-4.4)
Conversation	76.5 (-6.6)	59.8 (-16.2)	84.9 (-12.4)	73.8 (-11.3)
No Instruction Tuning	22.0 (-61.1)	24.0 (-51.3)	18.5 (-78.0)	21.5 (-63.6)

Table 3: Relative scores for different settings *w.r.t.* GPT-4 (text-only) on 30 randomly sampled images from COCO Val 2014. Each image is associated one short question, one detailed question, and one complex reasoning question, resulting in a total of 90 questions. Following the same setting as our data generation pipeline, GPT-4 uses ground truth image captions and bounding boxes as visual input. We prompt GPT-4 with the answers from our model outputs and the answers by GPT-4 (text-only), and let it compare between both responses and give a rating with an explanation.

Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

User	What is unusual about this image?
LLaVA	The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.
	[*Start a new conversation, and clear the history*]
User	What's happening in the scene?
LLaVA	The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.
User	What is unusual about this image?
GPT-4 [32]	The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.
User	Can you explain this meme in detail?
BLIP-2	a man is sitting on the back of a yellow cab
User	Can you explain this meme in detail?
OpenFlamingo	The man is drying his clothes on the hood of his car.

Experimental Results

1) ScienceQA

Method	Subject			Context Modality			Grade		Average
	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	
Representative & SoTA methods with numbers reported in the literature									
Human [30]	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
GPT-3.5 [30]	74.64	69.74	76.00	74.44	67.28	77.42	76.80	68.89	73.97
GPT-3.5 w/ CoT [30]	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
LLaMA-Adapter [55]	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
MM-CoT$_{Base}$ [57]	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37	84.91
MM-CoT$_{Large}$ [57]	95.91	82.00	90.82	95.26	88.80	92.89	92.44	90.31	91.68
Results with our own experiment runs									
GPT-4	84.06	73.45	87.36	81.87	70.75	90.73	84.69	79.10	82.69
LLaVA	90.36	95.95	88.00	89.49	88.00	90.66	90.93	90.90	90.92
LLaVA+GPT-4 (complement)	90.36	95.50	88.55	89.05	87.80	91.08	92.22	88.73	90.97
LLaVA+GPT-4 (judge)	91.56	96.74	91.09	90.62	88.99	93.52	92.73	92.16	**92.53**

Table 6: Results (accuracy %) on Science QA dataset. Question classes: NAT = natural science, SOC = social science, LAN = language science, TXT = text context, IMG = image context, NO = no context, G1-6 = grades 1-6, G7-12 = grades 7-12.

	
User LLaVA	Can you explain this meme in detail? The meme in the image is a creative and humorous take on food, with a focus on chicken nuggets, as the center of the universe. The meme begins with the phrase "Sometimes I just look at pictures of the Earth from space and I marvel at how beautiful it all is..." The main image features a tray of cooked chicken nuggets, with the chicken pieces positioned to resemble the continents and islands on a world map. The meme playfully suggests that the chicken nuggets represent the Earth, and the various locations depicted in the photo are actually chicken nugget versions of different places. The punchline of the meme appears in smaller text at the bottom, reading "I mean, it's not the real Earth, but how beautiful it is all is." This lighthearted and entertaining meme combines food and humor, using the image of chicken nuggets to create a funny and imaginative representation of the world.
User GPT-4 [32]	Can you explain this meme? This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets. The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world. The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.
User BLIP-2	Can you explain this meme in detail? sometimes i just look at pictures of the earth from space and marvel how beautiful it is
User OpenFlamingo	Can you explain this meme in detail? It's a picture of a chicken nugget on the International Space Station.

Table 5: Example prompt demonstrating LLaVA and GPT-4's visual input capability. The prompt requires image understanding.

Interaction VL

LLaVA Chatbot

What is the man doing in this picture?

The man in the picture is sleeping on a plane, wearing headphones.

The man in the picture is sleeping on a plane, wearing headphones.

What country does he seem to be from?

The man seems to be from China, as indicated by his ethnicity.

Thank you

Q&A