Attention to Detail: Fine-Grained Vision-Language Alignment

Image generated by DALL-E 3

Kai-Wei Chang

CS @ UCLA

kw@kwchang.net

UCLA ENGINEERING Computer Science

See http://kwchang.net for more information

Children can recognize objects through language descriptions

UCLA ENGINEERING
Computer Science

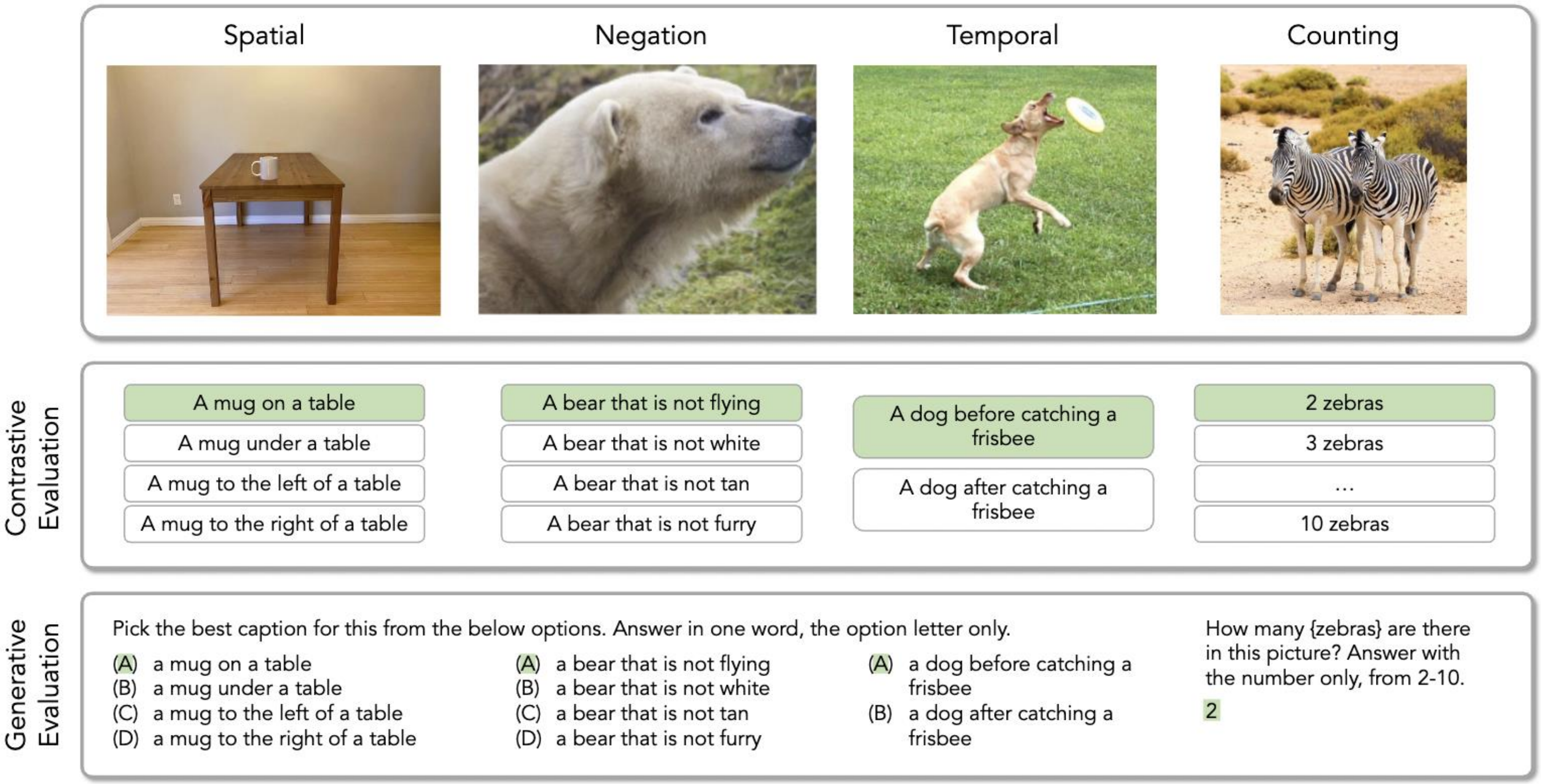
How to teach a model to locate “paramedics”

Visual Recognition with Language Descriptions

From ChatGPT:

Paramedic typically wears a uniform with a patch or logo identifying them as a member of an emergency medical services (EMS) team. The uniform may include a shirt, pants, and jacket with reflective strips for visibility. They may also wear protective gear such as gloves, goggles, and a mask. They often carry equipment such as a backpack with medical supplies, a radio, and a defibrillator. They may also wear a duty belt with a flashlight, scissors, and other tools. The appearance can vary depending on the agency.

Scale Can't Overcome Pragmatics: The Impact of Reporting Bias on Vision-Language Reasoning.



Amita Kamath

Scale Can't Overcome Pragmatics: The Impact of Reporting Bias on Vision-Language Reasoning. Amita Kamath, Jack Hessel, Khyathi Chandu, Jena Hwang, Kai-Wei Chang, Ranjay Krishna

	Model	Spatial	Negation	Counting	Temporal
(a)	CLIP ViT-B/32	30.6	11.5	43.4	58.5
	+ ML Div.	27.4	15.5	23.3	51.5
	CLIP ViT-B/16	27.7	12.7	48.1	55.0
	CLIP ViT-L/14	28.4	12.3	64.1	52.0
	CLIP ViT-g/14	28.4	12.7	59.0	52.0
	CLIP ViT-H/14	26.0	13.2	60.0	59.0
(b)	LLAVA-1.5-7B	37.6	33.4	47.3	72.5
	LLAVA-1.5-13B	61.7	28.4	48.9	74.5
	Molmo 7B-O	75.5	38.4	77.5	78.0
	Molmo 7B-D	87.6	41.3	83.8	80.5
(c)	LLAVA-1.6-m7B	60.0	40.6	52.9	70.0
	QwenVL 7B-Chat	47.1	24.2	84.6	67.5
	Qwen2VL 7B-Inst.	98.3	56.1	85.8	84.0
	GPT4o	91.5	22.2	90.9	95.0
	GPT o1	97.6	64.7	88.2	97.0
	Gemini 1.5-Flash	98.5	46.4	84.6	81.5
	Gemini 1.5-Pro	92.0	49.0	87.8	85.0
	Claude-3 Haiku	65.5	28.9	83.4	70.0
	Claude-3.5 Sonnet	95.4	42.0	92.3	83.5
	Random Chance	25.0	25.0	11.1	50.0
	Human Estimate	100	100	100	100

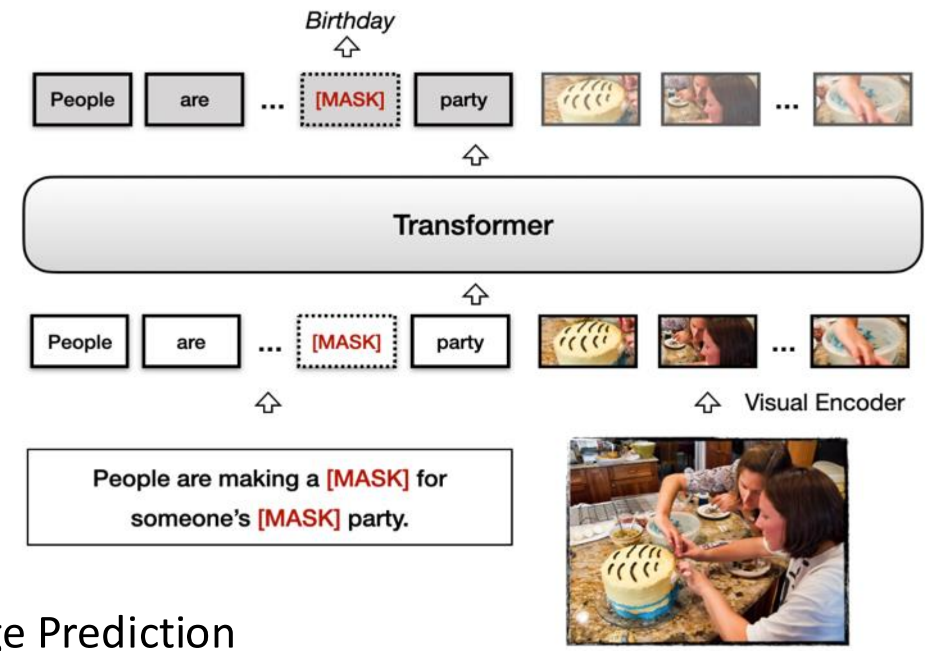
Outline

- Teach Machines using Language descriptions
- Describe Visual Features in Language (Verbalized Representation Learning)
- Refine Large Vision Language Model’s Alignment for Reasoning

How we can teach machines using language description?

Pre-train VisualBERT

Masked language modeling with the image



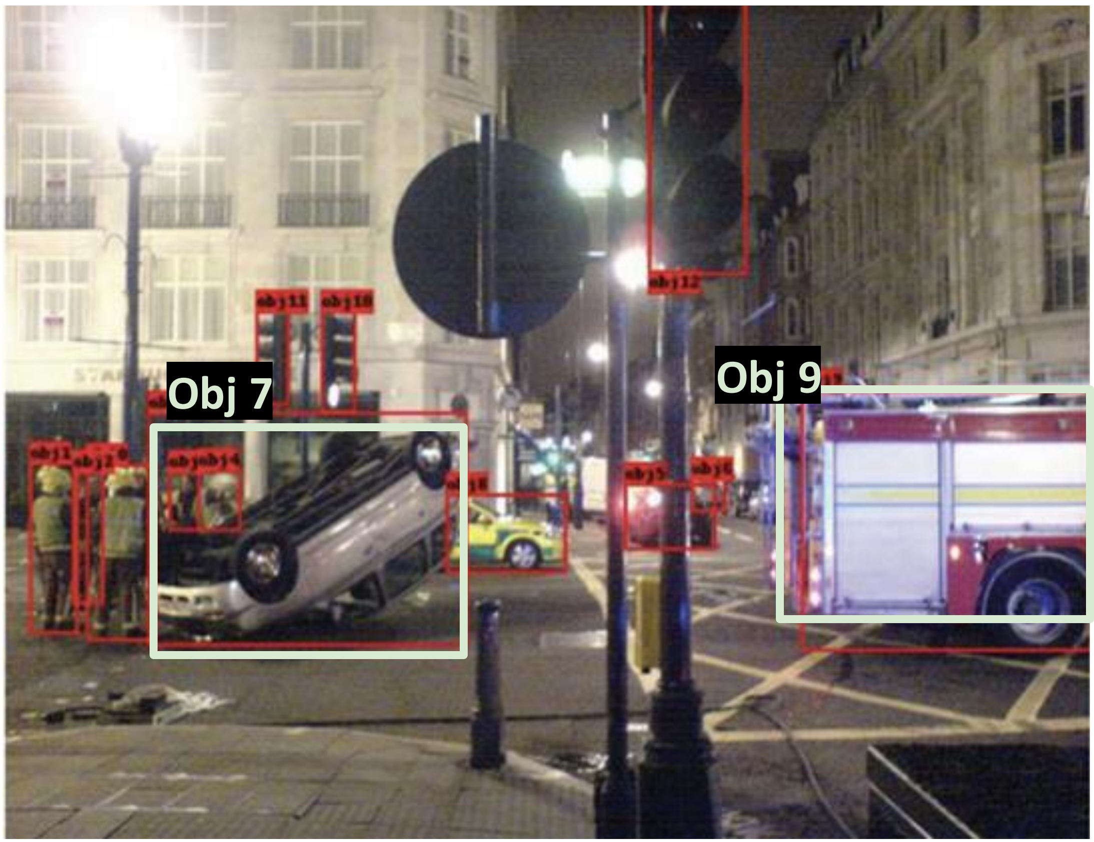
Sentence-Image Prediction

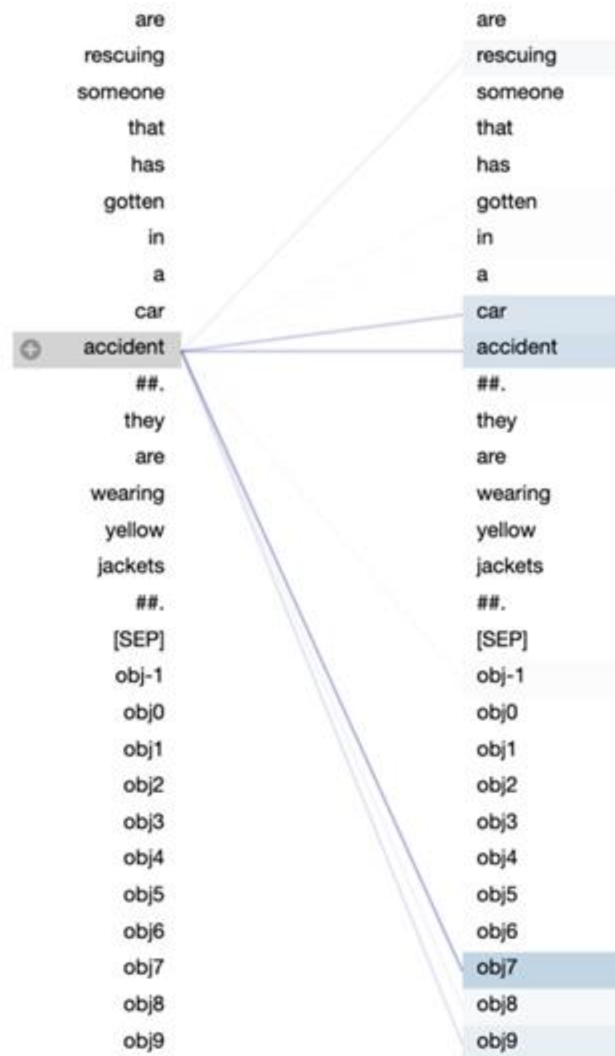
People are making a cake for someone's birthday party.	positive
People are playing a ball in the park.	negative

VisualBERT: A Simple and Performant Baseline for Vision and Language

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, Kai-Wei Chang

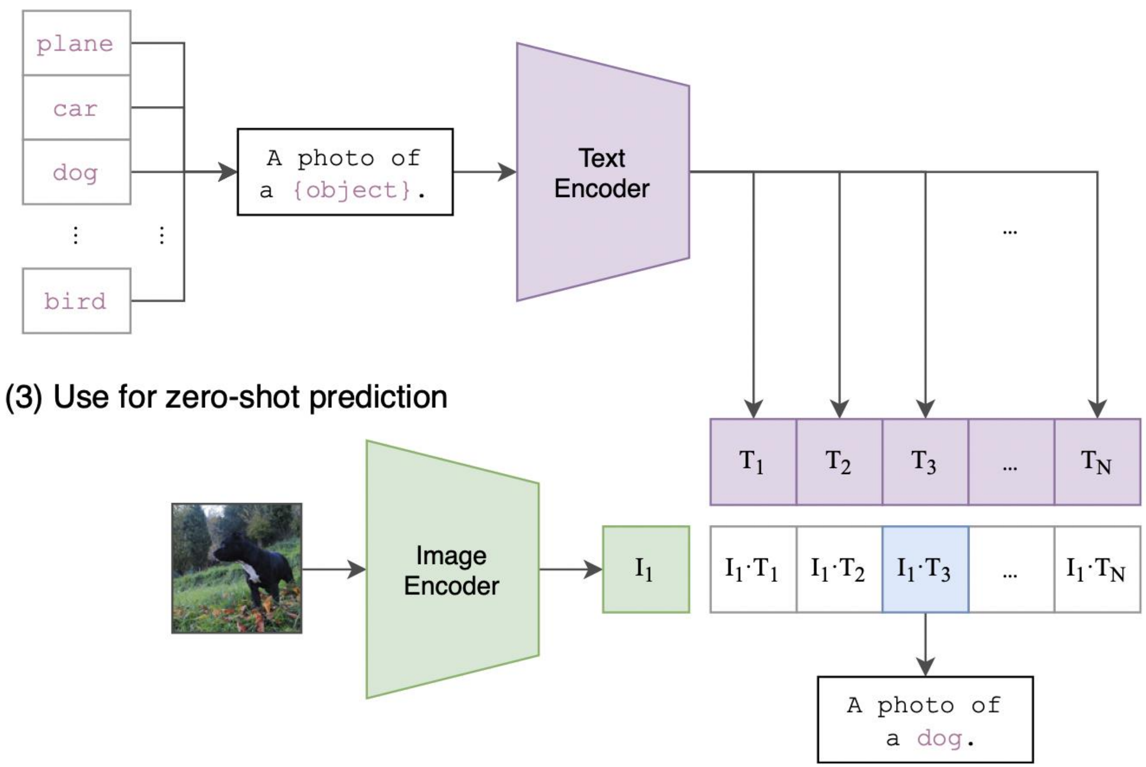
Learning High-level Concepts





Contrastive Language-Image Pretraining (CLIP)

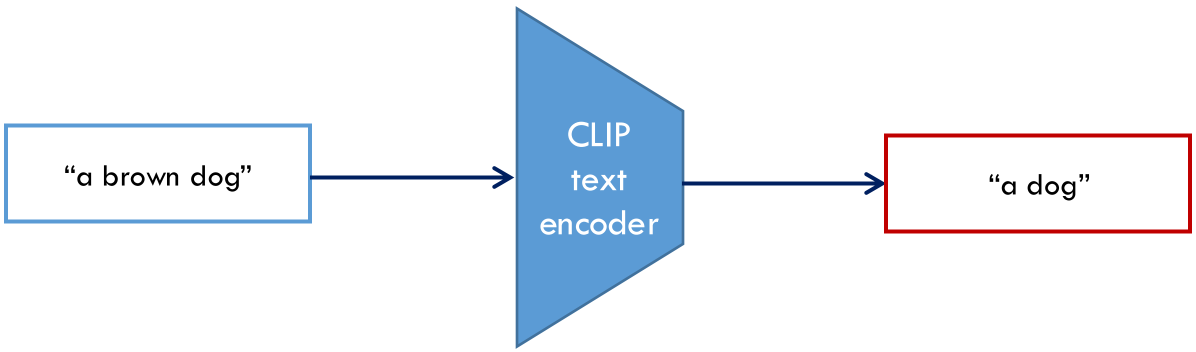
- CLIP (Radford et al., 2021): image classification as image-text matching
- Leverage millions of image-caption data



Text encoders are performance bottlenecks

- What if the text encoder isn't perfect

Amita Kamath



You can't cram the meaning of a single $&!#*
sentence into a single $!#&* vector!

Professor Raymond J. Mooney

Text Encoders are Performance Bottlenecks in Contrastive Vision-Language Models

Amita Kamath, Jack Hessel, and Kai-Wei Chang, in EMNLP, 2023.

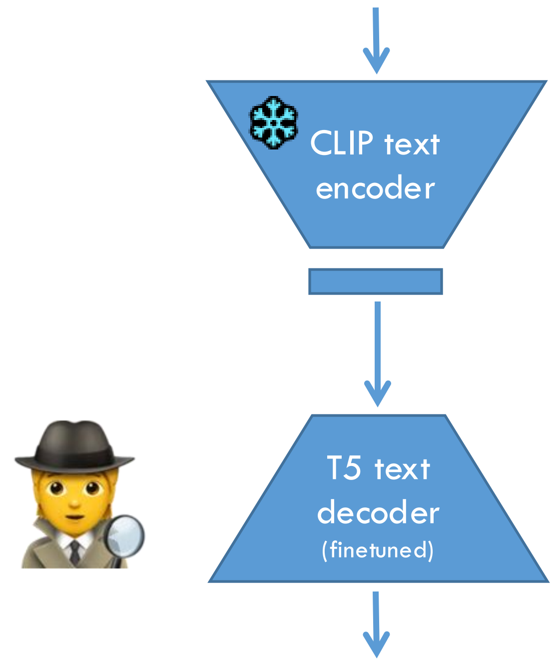
Probing text encoder in CLIP

1. Create increasingly compositional text prompts

2. Feed them into CLIP's text encoder

3. Try to decode out the original prompt

an iguana
a happy dinosaur
a surfer carrying a lifeguard
an orangutan eating and an officer flying



an iguana ✓
an amusing dinosaur ✓
a lifeguard carrying a surfer ✗
an orangutan and an officer eating an orangutan ✗

UCLA ENGINEERING
Computer Science

Text-Encoder Reconstruction Performance

$\mathcal{T}(x)$	Embed. size	Avg. EM (%)
CLIP ViT-B/32	512	13.2
CLIP ViT-L/14	768	28.5
negCLIP ViT-B/32	512	28.6
RoBERTaCLIP ViT-B/32	512	28.9
Proof-of-concepT5	1024	<u>*98.9*</u>

Lesson 1:

Align Vision-Language at a fine-grained level of granularity

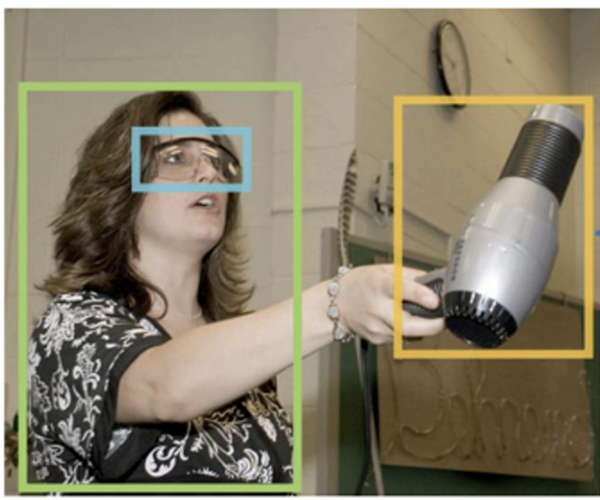
GLIP: Object Detection as Phrase Grounding

Harold (Liunian) Li

Prompt

Person. Bicycle … Hairdryer.

A woman holds a blow dryer, wearing protective goggles



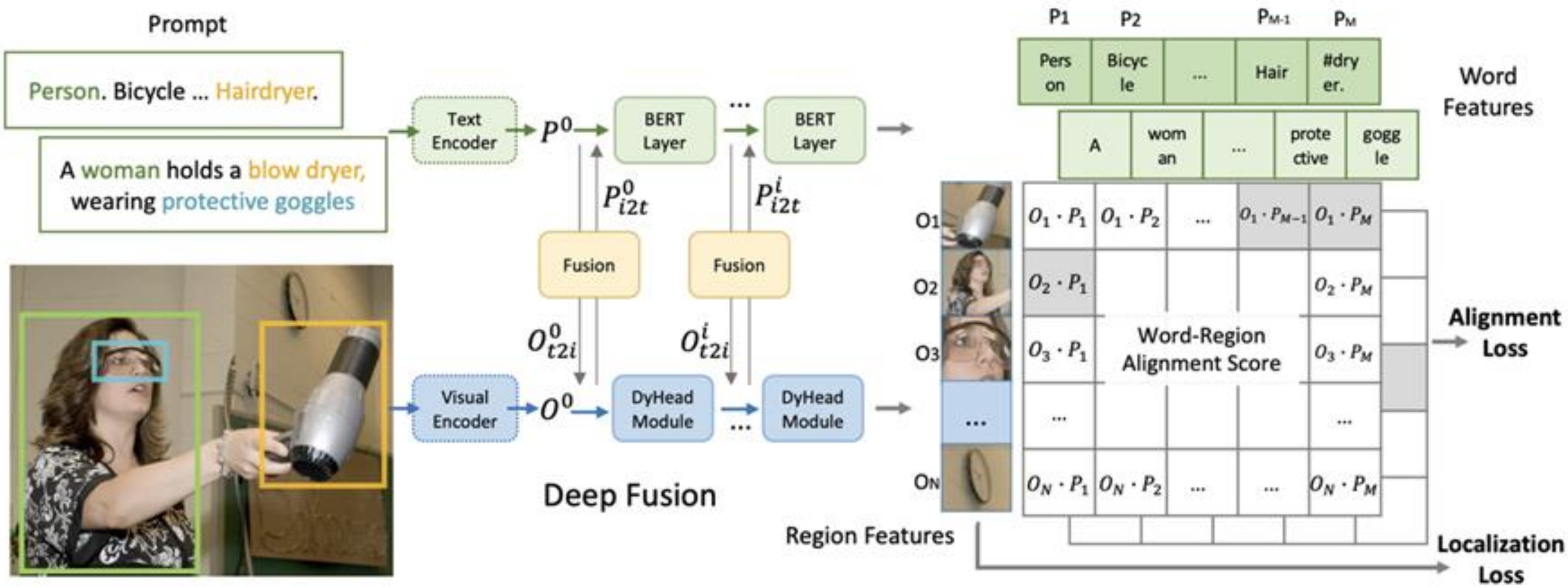
Phrase Grounding : Given a sentence and an image, locate the entities in the image

Grounded Language-Image Pre-training

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao, in CVPR, 2022. Best Paper Finalist

GLIP: Overview

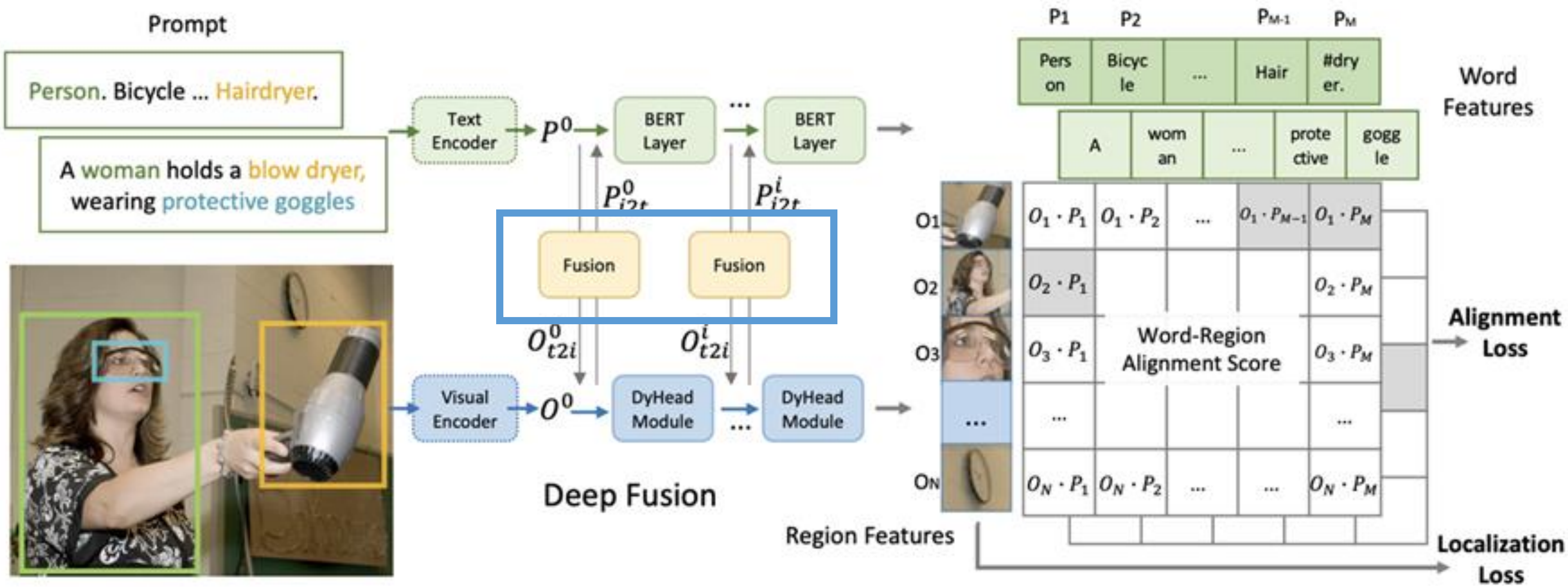
- Align objects to phrases in text



$$O = \mathrm{Enc}_I(\mathrm{Img}), \ P = \mathrm{Enc}_L(\mathrm{Prompt}), \ S_{\mathrm{ground}} = OP^\top,$$

GLIP: Overview

- Align objects to phrases in text



$$O^i_{\text{t2i}}, P^i_{\text{i2t}} = \text{X-MHA}(O^i, P^i), \quad i \in \{0, 1, .., L-1\}$$

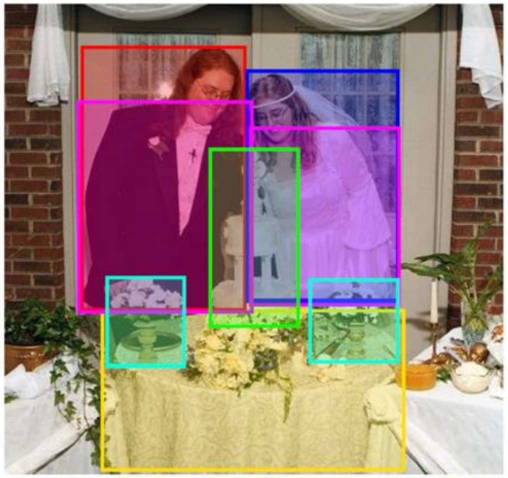
$$O^{i+1} = \text{DyHeadModule}(O^i + O^i_{\text{t2i}}),$$

$$P^{i+1} = \text{BERTLayer}(P^i + P^i_{\text{i2t}}),$$

UCLA ENGINEERING Computer Science

Pre-training with Scalable Semantic-Rich Data

- Detection data have **limited categories**
  - Objects365: 365 categories
  - LVIS: ~1,200 categories
  - Visual Genome: ~1,600 categories
- **Gold grounding data**:
  - Flickr30K: **44,518** unique phrases
  - VG Caption:**110,689** unique phrases



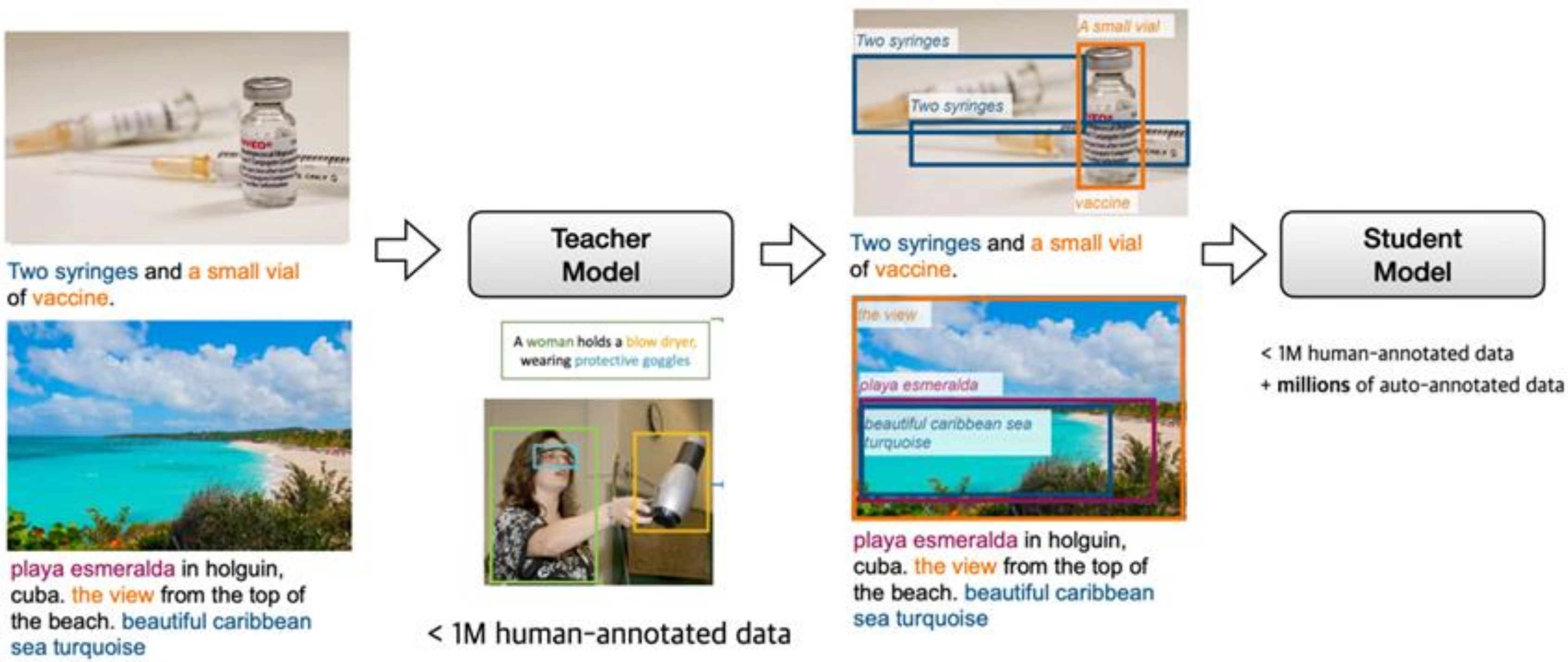
A couple in their wedding attire stand behind a table with a wedding cake and flowers.
A bride and groom are standing in front of their wedding cake at their reception.
A bride and groom smile as they view their wedding cake at a reception.
A couple stands behind their wedding cake.
Man and woman cutting wedding cake.

Can further generate pseudo-data from image captions

Scaling up with image-caption data

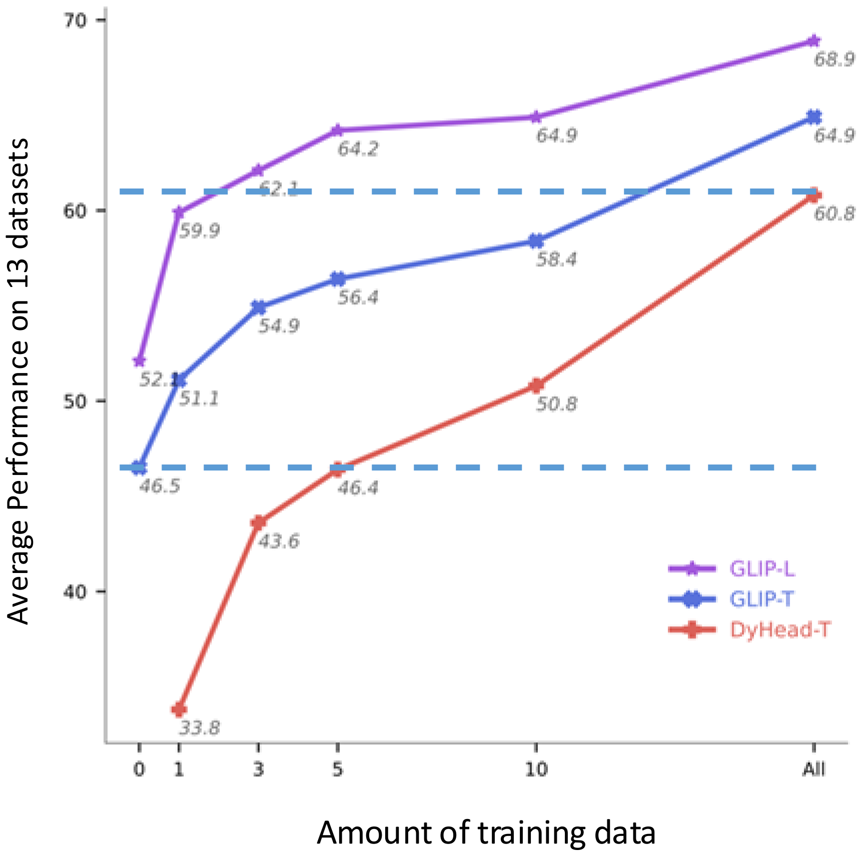
Constructed by **distant supervision** (Craven et al., 1998)

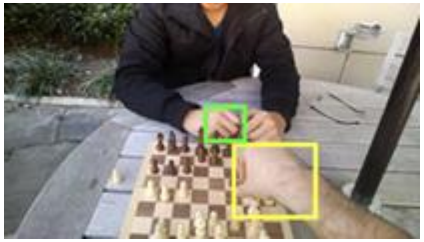
24M image-caption data, with **78.1M** boxes and **58.4M** unique phrases



Object Detection in the Wild : Data Efficiency

13 downstream tasks from https://public.roboflow.com/object-detection





0-shot GLIP-T ~= 5-shot DyHead-T

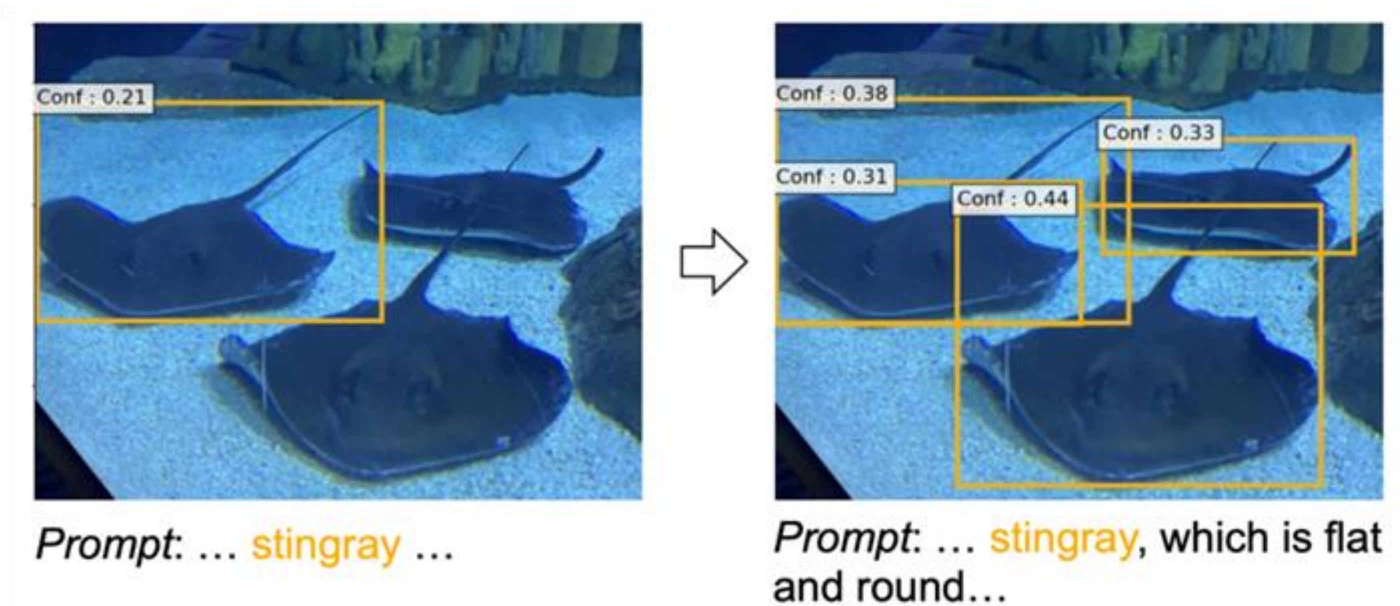
1-shot GLIP-T ~= 10-shot DyHead-T

1-shot GLIP-L ~= Fully-supervised DyHead-T

UCLA ENGINEERING
Computer Science

Object Detection with Instructions

- Learn from human instructions on the fly



Understand properties and attributes

Users can change the model behaviour by changing the instruction

Lesson 2:

Attention to detail: Description fine-tuning

Desco: Learning through Descriptions

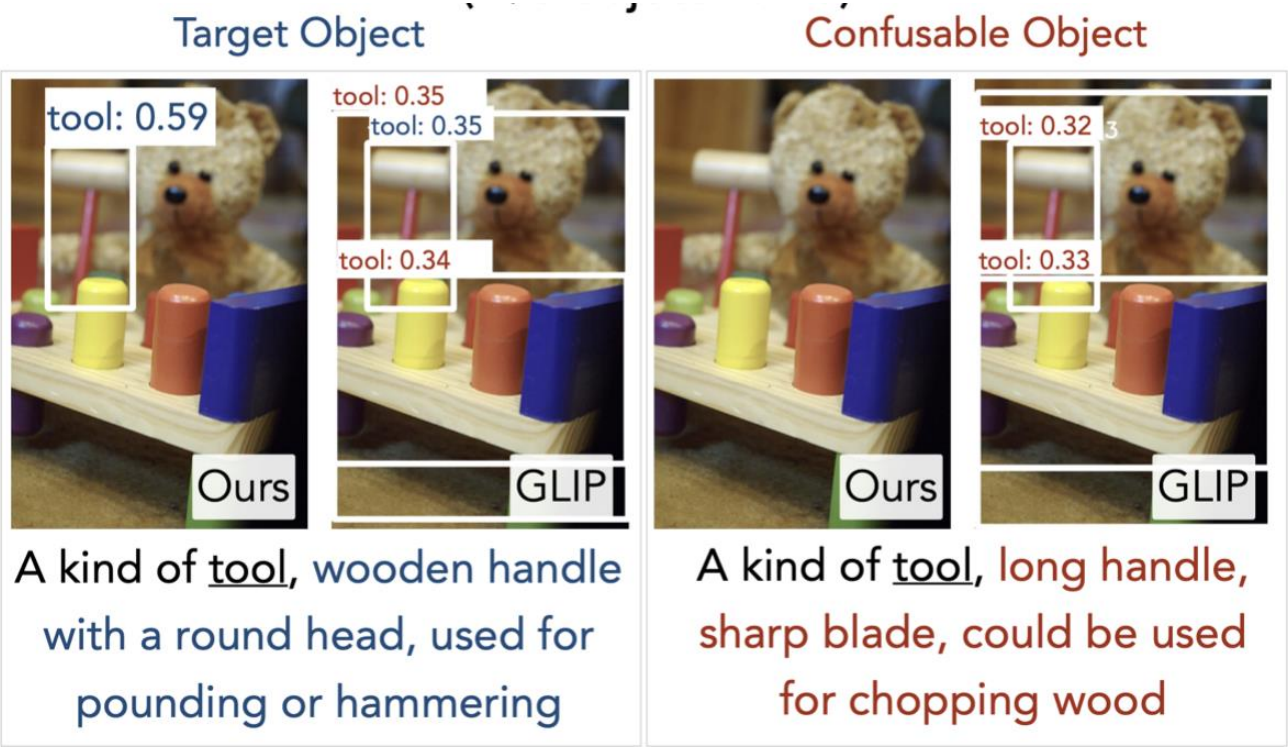
Harold (Liunian) Li

DesCo: Learning Object Recognition with Rich Language Descriptions

Liunian Harold Li, Zi-Yi Dou, Nanyun Peng, and Kai-Wei Chang, in *NeurIPS*, 2023.

Top-1 System of OmniLabel Challenge at CVPR workshop 2023

Desco: Learning through Descriptions



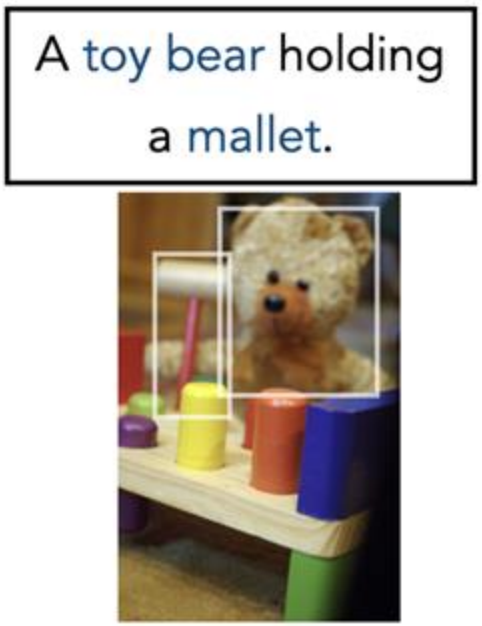
Harold (Liunian) Li

DesCo: Learning Object Recognition with Rich Language Descriptions

Liunian Harold Li, Zi-Yi Dou, Nanyun Peng, and Kai-Wei Chang, in *NeurIPS*, 2023.

Winning System for Omnilabel Challenge at CVPR workshop 2023

Challenge #1: Fine-Grained Descriptions Rare in Data



Reporting bias: humans do not write obvious things

When writing captions, we tend to directly use entity names rather than descriptions for subparts, shapes, textures, etc.

The World of an Octopus: How Reporting Bias Influences a Language Model's Perception of Color

Cory Paik, Stéphane Aroca-Ouellette, Alessandro Roncone, Katharina Kann

UCLA ENGINEERING
Computer Science

Solution: Generating Descriptions from LLMs

User:

What are the useful features for identifying **mallet**?

GPT-3:

Mallet is a kind of tool, wooden handle,

…

Build a vocabulary of 10K noun phrases on Conceptual Captions and VG

Sample descriptions for each noun phrase (<1 day via API)

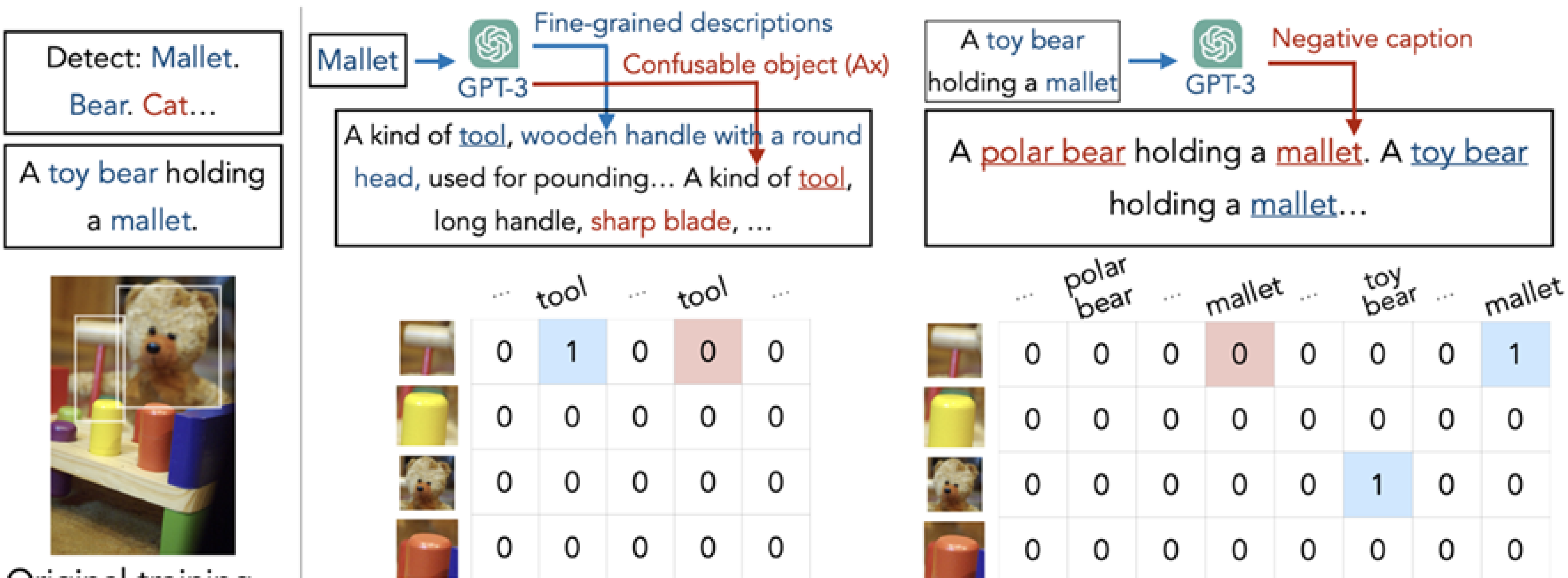
UCLA ENGINEERING Computer Science

Challenge #2: Model Might Ignore Description

- The model is not incentivized to "read" the descriptions

UCLA ENGINEERING
Computer Science

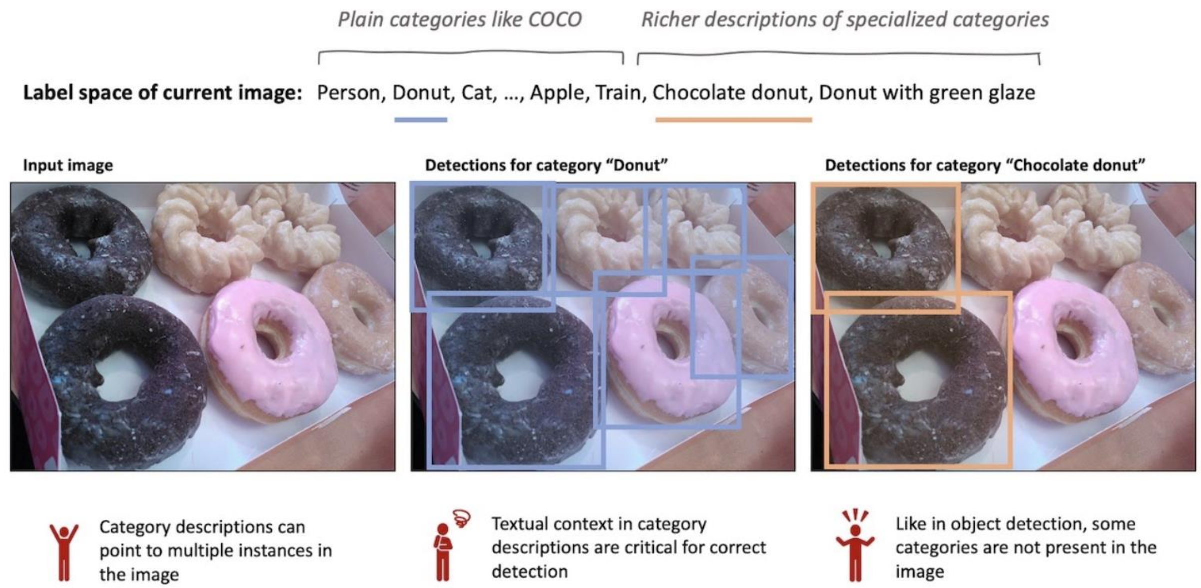
Solution: Context-Sensitive Query



Original training data for GLIP

Description-rich and context-sensitive data for DESCO-GLIP

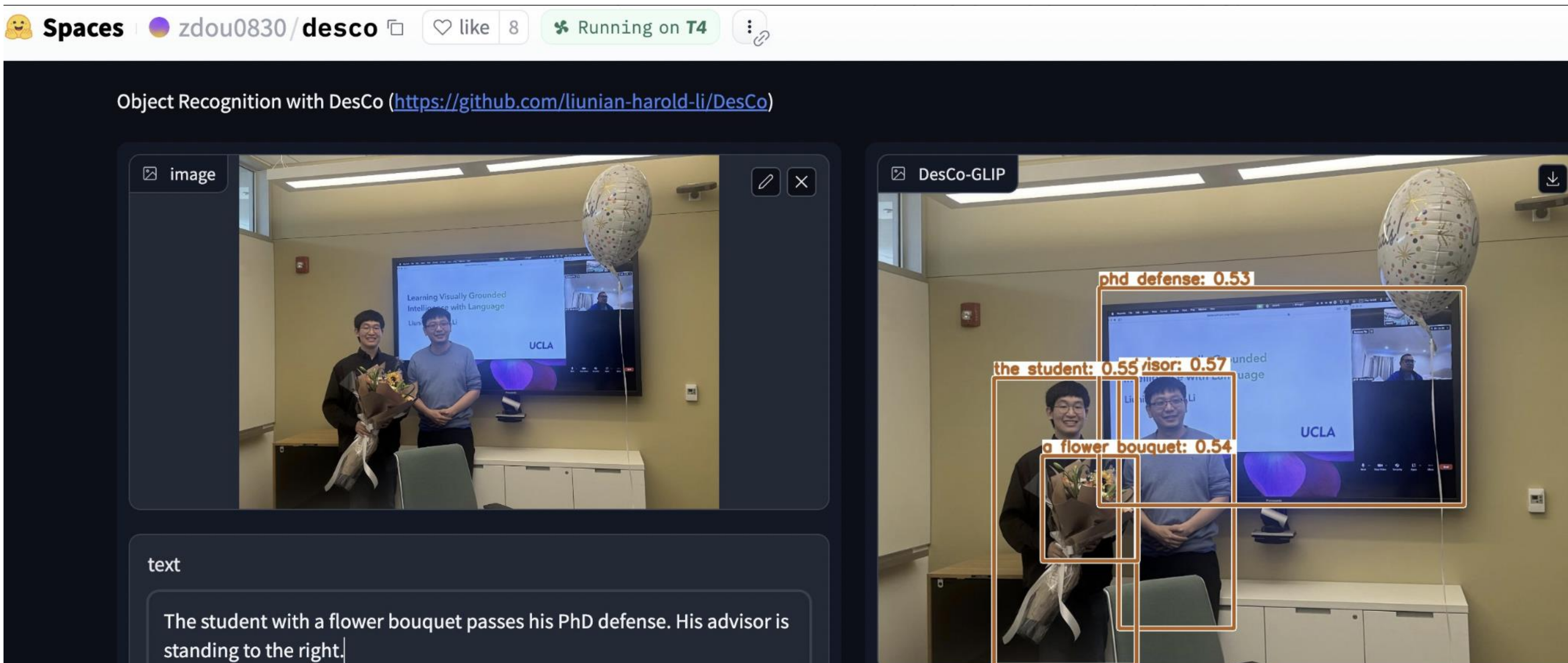
Results



Model	Backbone	LVIS MiniVal [15]				OmniLabel [30]			
		APr	APc	APf	AP	AP	APc	APd	APd-P
MDETR [15]	RN101	20.9	24.9	24.3	24.2	-	-	4.7	9.1
MaskRCNN [15]	RN101	26.3	34.0	33.9	33.3	-	-	-	-
RegionCLIP [44]	ResNet-50	-	-	-	-	2.7	2.7	2.6	3.2
Detic [46]	Swin-B	-	-	-	-	8.0	15.6	5.4	8.0
K-LITE [33]	Swin-T	14.8	18.6	24.8	21.3	-	-	-	-
GroundingDINO-T [21]	Swin-T	18.1	23.3	32.7	27.4	-	-	-	-
GroundingDINO-L [21]	Swin-L	22.2	30.7	38.8	33.9	-	-	-	-
GLIP-L [19]	Swin-L	28.2	34.3	41.5	37.3	25.8	32.9	21.2	33.2
GLIP-T [19]	Swin-T	20.8	21.4	31.0	26.0	19.3	23.6	16.4	25.8
DESCO-GLIP	Swin-T	**30.8**	**30.5**	**39.0**	**34.6**	**23.8**	**27.4**	**21.0**	**30.4**
FIBER-B [7]	Swin-B	25.7	29.0	39.5	33.8	25.7	30.3	22.3	34.8
DESCO-FIBER	Swin-B	**34.8**	**35.5**	**43.9**	**39.5**	**29.3**	**31.6**	**27.3**	**37.7**

Demo

https://huggingface.co/spaces/zdou0830/desco



UCLA ENGINEERING Computer Science

Describe Visual Features in Language
--Verbalized Representation Learning

How human recognizes objects?

- Humans excel in recognizing objects with few examples using their *inherent language understanding*

Thalassoma Pavo

Thalassoma Bifasciatum

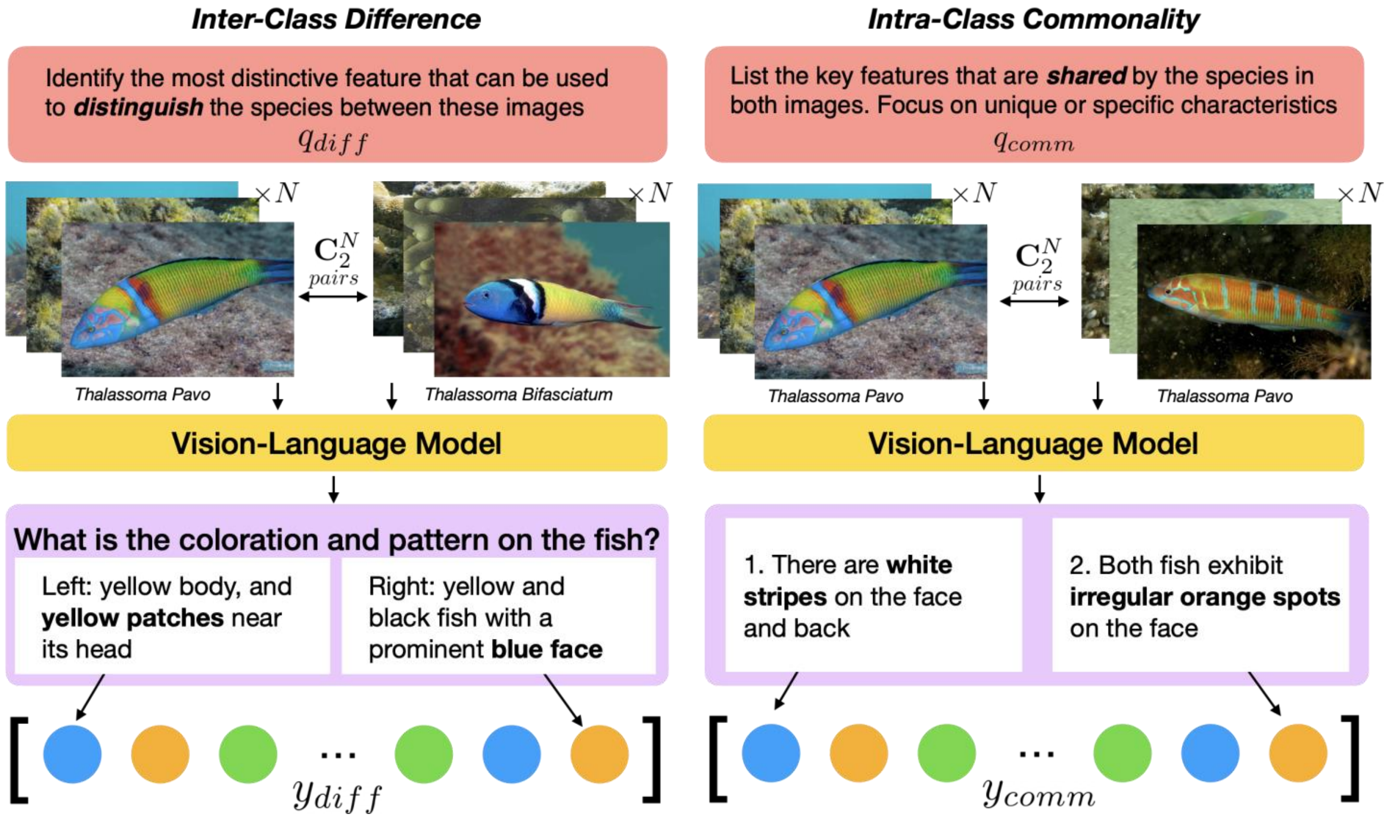
Thalassoma Pavo

Thalassoma Pavo

Both fish feature a white stripe around their neck and a yellow body. However, fish on the left has an ***irregular white and orange pattern*** *on its head. The right one has a* ***solid blue head****.*

The right one has ***more stripe*** *on its body. However, both fish* ***display irregular, mosaic-like patterns*** *on their heads, predominantly in white and orange hues.*

Verbalized Representation Learning



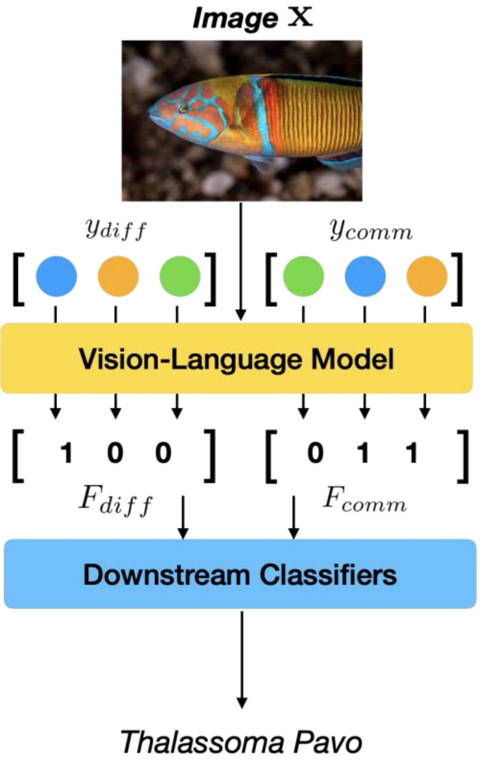
Verbalized Representation Learning for Interpretable Few-Shot Generalization

Cheng-Fu Yang, Da Yin, Wenbo Hu, Nanyun Peng, Bolei Zhou, Kai-Wei Chang

UCLA ENGINEERING Computer Science

Automatic Feature Engineering

- These verbalized features are mapped to numeric values via a Vision-Language Model
- The resulting representations can be used in any downstream model
- Language as an information bottleneck



UCLA ENGINEERING
Computer Science

Experiment

- Significantly outperforms LLM + Attribute Bottleneck (LLM-Mutate), In-context Learning (LLaVA-ICL), Lora-Finetuning (LLaVA-SFT)

Method	Lichen	Wrasse	Wild Rye	Manzanita	Bulrush	Average
Zero-Shot Methods						
CLIP Class Name [10]	23.3	32.0	32.0	26.0	26.0	27.86
Full Dataset Methods (200+ Images per Species)						
CLIP Prompt Tuning [10]	23.3	20.0	40.0	20.0	20.0	24.66
Classification by Description [27]	30.0	34.0	36.0	28.0	20.0	29.60
LLM-Mutate-70B [10] (1-prompt)	31.6	24.0	44.0	40.0	22.0	32.32
LLM-Mutate-70B [10] (10-prompt)	48.3	44.0	58.0	58.0	42.0	50.06
Few-Shot Methods (10 Images per Species)						
LLM-Mutate-7B[†] (10-prompt)	35.0	48.0	38.0	44.0	26.0	38.20
LLM-Mutate-70B[†] (10-prompt)	46.6	44.0	46.0	44.0	40.0	44.13
LLaVA-ICL-7B	16.6	28.0	22.0	18.0	30.0	22.92
LLaVA-SFT-7B	41.6	50.0	58.0	42.0	28.0	43.92
LLaVA-VRL-7B (**Ours**)	58.3	48.0	74.0	**66.0**	46.0	58.46
LLaVA-VRL-72B (**Ours**)	**71.6**	**72.0**	**74.0**	56.0	**66.0**	**67.92**

UCLA ENGINEERING Computer Science

Attention to detail is Required for Visual Reasoning

Context-Sensitive Text-Rich Visual Reasoning

ESTVQA

ConTextual

	ESTVQA	ConTextual
Instruction	What can we eat here?	Get the number of the boat with three yellow and one red round items hanging from it.
OCR	Angelo's Car Hop Service Hamburgers Laundromat	SS273 WH97 SS266 SS681 SS138
GPT4 w/ OCR Response	You can eat hamburgers at Angelo's Car Hop Service.	SS681

ConTextual: Evaluating Context-Sensitive Text-Rich Visual Reasoning in Large Multimodal Models

Rohan Wadhawan, Hritik Bansal, Kai-Wei Chang, and Nanyun Peng, in ICML, 2024.

ConTextual Dataset - Visual Scenarios

https://con-textual.github.io/

Navigation

Group gates by direction.

Shopping

Calculate the total cost of assembling an outfit consisting of a white dress, heeled sandals, and sunglasses.

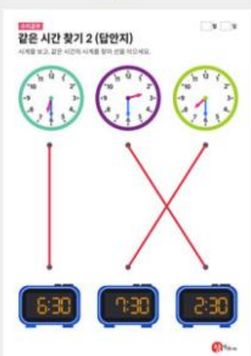
Miscellaneous Natural Scenes

Determine the word obfuscated by the outfielder.

Abstract

Identify a common proverb portrayed in this image.

Time

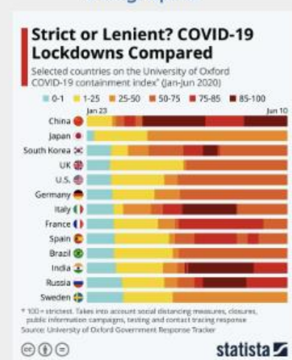


Comment whether the analog clocks are correctly matched to the digital clocks.

Web Usage

Which items have experienced a price drop?

Infographic



Which countries have had a non-zero containment index that lies between 85-100?

Application Usage

List the exercises where the corresponding illustration showcases a single movement.

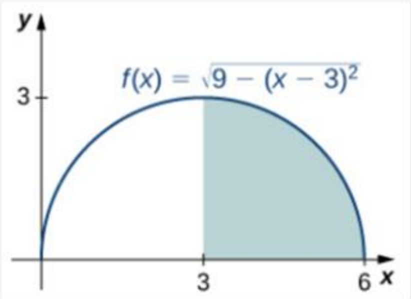
ConTextual: Evaluating Context-Sensitive Text-Rich Visual Reasoning in Large Multimodal Models

Rohan Wadhawan, Hritik Bansal, Kai-Wei Chang, and Nanyun Peng, in ICML, 2024.

#	Model	Method	Source	Date	ALL	Time	Shop.	Nav.	Abs.	App.	Web.	Info.	Misc. NS.
-	Human Performance	-	Link	2024-01-24	**69.6**	64.0	64.0	73.5	75.5	64.0	58.0	72.0	78.0
1	GPT-4o 🥇	LMM 🖼️	Link	2024-05-18	**62.8**	32.0	70.0	60.0	98.0	72.0	62.0	48.0	64.7
2	GPT-4o-mini-2024-07-18 🥈	LMM 🖼️	Link	2024-07-18	**61.7**	22.0	62.0	62.0	98.0	72.0	64.0	42.0	67.3
3	Claude-3.5-Sonnet-2024-06-20 🥉	LMM 🖼️	Link	2024-07-18	**57.5**	22.0	52.0	66.0	96.0	68.0	64.0	44.0	56.7
4	Gemini-1.5-Flash-Preview-0514	LMM 🖼️	Link	2024-05-18	**56.0**	30.0	51.0	52.1	84.0	63.0	63.2	42.8	61.7
5	Gemini-1.5-Pro-Preview-0514	LMM 🖼️	Link	2024-05-18	**52.4**	24.0	46.9	39.6	84.0	45.8	59.2	43.8	64.0
6	GPT-4V(ision)	LMM 🖼️	Link	2024-01-24	**47.4**	18.0	54.0	48.0	100.0	48.0	42.0	28.0	48.0
7	Gemini-Pro-Vision	LMM 🖼️	Link	2024-01-24	**40.2**	16.0	32.7	28.6	65.3	44.9	43.8	20.0	52.8
8	Claude-3-Opus-2024-02-29	LMM 🖼️	Link	2024-03-05	**38.1**	18.0	32.0	34.0	68.0	44.0	38.0	18.0	44.7
9	LLaVA-Next-34B	LMM 🖼️	Link	2024-03-05	**36.8**	10.0	36.0	30.6	66.0	36.0	28.0	12.0	51.3

- GPT-4o is bad at time-reading and infographics but outperforms humans in abstract contexts (memes, etc.).
- Current models have limited capacity for fine-grained visual perception

Benchmarking Math Skills



ID: 926

Question: What is the size of the shaded area under the curve? Round the answer to 2 decimal places.

Ground truth: 7.07

GPT-4V:
To find the size of the shaded area under the curve $f(x) = \sqrt{9-(x-3)^2}$, we need to integrate the function over the given interval.

The given function represents the upper half of a circle with a radius of 3 and centered at the point (3,0). The shaded region corresponds to the area of a quarter circle with a radius of 3. The formula for the area of a circle is πr^2. For a quarter of the circle, the area is $\frac{1}{4}\pi r^2$.

Given: $r = 3$

$$\text{Area} = \frac{1}{4}\pi(3^2)$$

$$\text{Area} = \frac{1}{4}\pi(9)$$

$$\text{Area} = \frac{9}{4}\pi$$

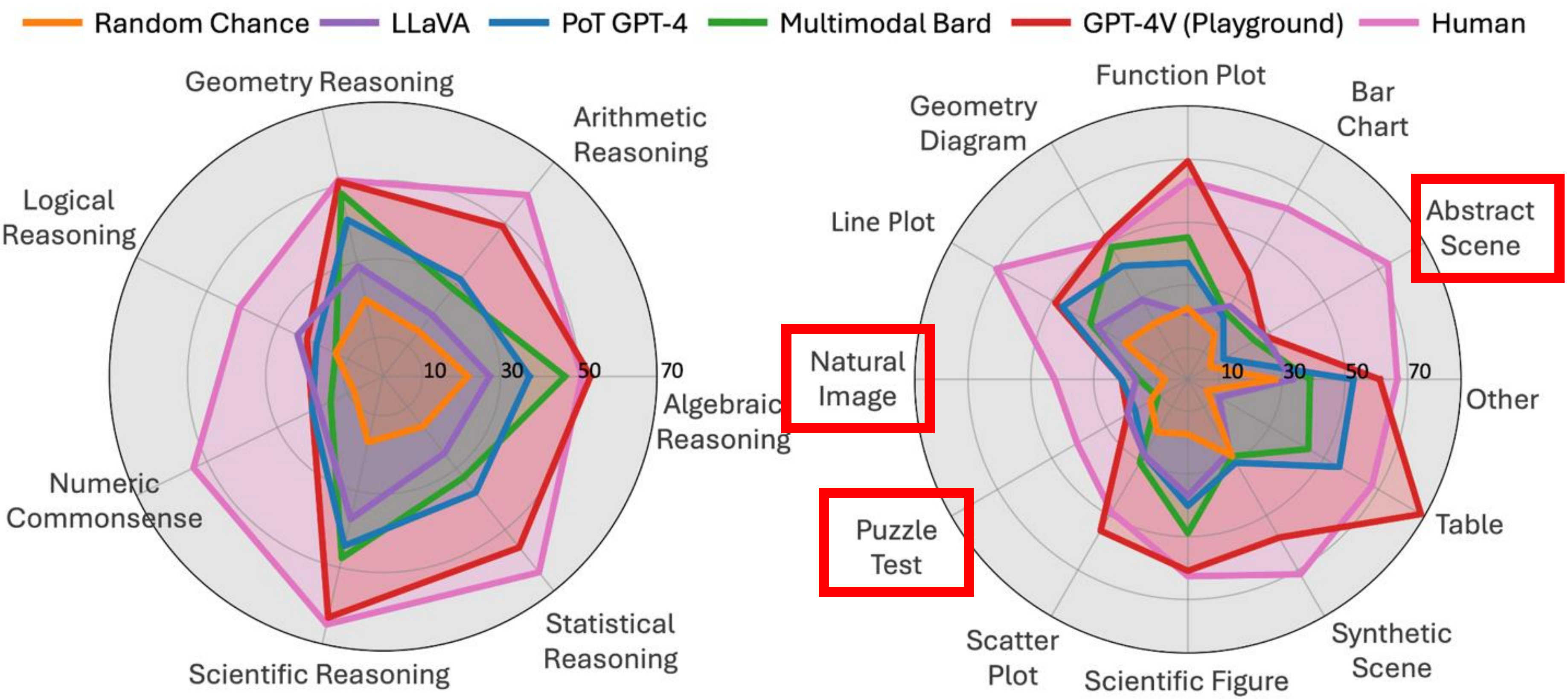
$$\text{Area} = 7.07 \text{ (rounded to 2 decimal places)}$$

So, the size of the shaded area under the curve is approximately 7.07 square units.
Correct output: 7.07

MathVista

Evaluating Math Reasoning in Visual Contexts

https://mathvista.github.io/



(a) Mathematical reasoning (b) Visual context

https://mathvista.github.io/

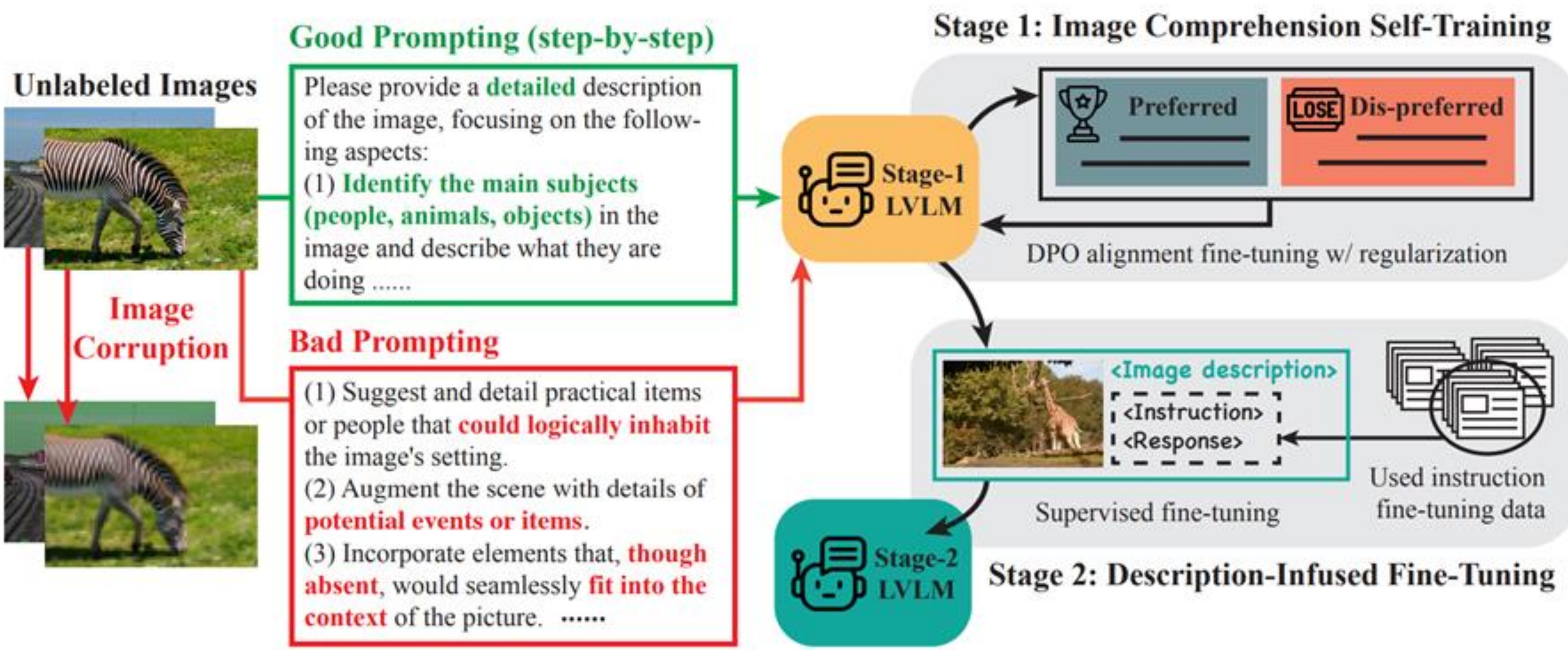
[MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts]

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao, in ICLR, 2024.

Attention to detail:

Refine Large Vision Language Models with Description Fine-tuning

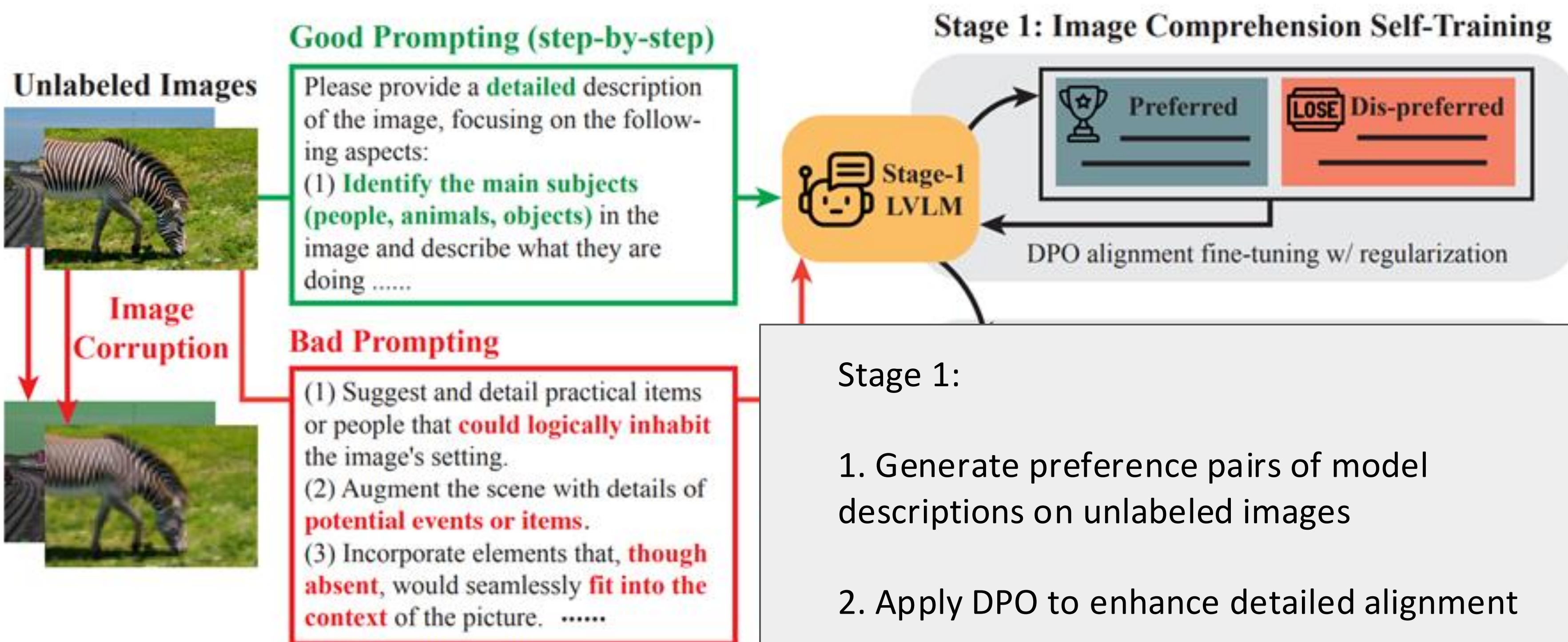
Refine LVLM with Self-Training



Enhancing Large Vision Language Models with Self-Training on Image Comprehension

Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Zou, Kai-Wei Chang, and Wei Wang, in *NeurIPS*, 2024.

How can we improve fine-grained visual perception in reasoning?



Stage 1:

1. Generate preference pairs of model descriptions on unlabeled images

2. Apply DPO to enhance detailed alignment

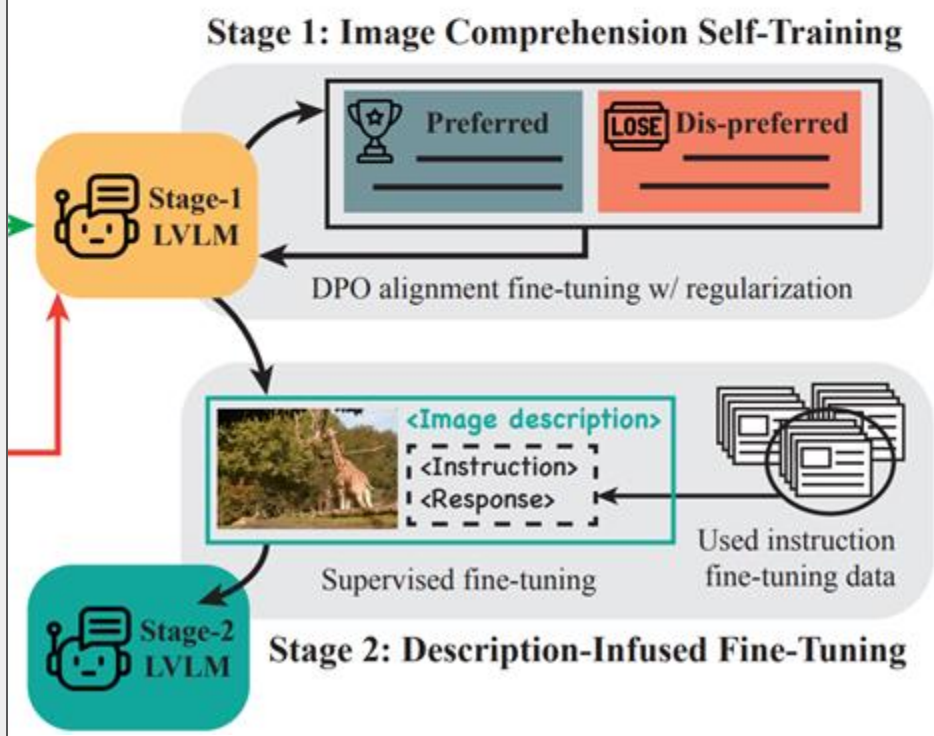
Enhancing Large Vision Language Models with Self-Training on Image Comprehension

Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Zou, Kai-Wei Chang, and Wei Wang, in *NeurIPS*, 2024.

How can we improve fine-grained visual perception in reasoning?

Stage 2:

Fine-tune LVLM with generated detailed image description



Enhancing Large Vision Language Models with Self-Training on Image Comprehension

Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Zou, Kai-Wei Chang, and Wei Wang, in *NeurIPS*, 2024.

Stage 1: Image Comprehension Self-Training

Task: Explain what is depicted in the photograph.

Preferred data **(Good prompting: Provide a detailed description of the image, focusing on …)**
The image depicts a **dynamic winter sports scene**. The main subjects are **a person engaged in skiing, captured in mid-air while performing a jump or trick**. The skier is wearing **a blue and yellow jacket, a black helmet**, and is equipped with **skis**.....

Preferred response generation:

- Image: Original image
- Prompt: GPT-4 to generate image descriptions.
  - We test these prompts on MSCOCO samples.
- SFT on the preferred data alone can be similar to system-2 distillation.

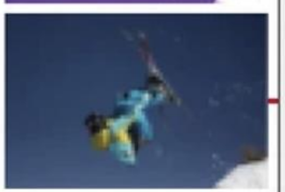
UCLA ENGINEERING
Computer Science

Stage 1: Image Comprehension Self-Training

Task: Explain what is depicted in the photograph.

Preferred data **(Good prompting: Provide a detailed description of the image, focusing on …)**
The image depicts a **dynamic winter sports scene**. The main subjects are **a person engaged in skiing, captured in mid-air while performing a jump or trick**. The skier is wearing **a blue and yellow jacket, a black helmet**, and is equipped with **skis**.....

(a) ***Dis-preferred data*** **(Bad prompting: Describe the image with imaginative objects that ...)**
...... In the distance, **a group of trees stands tall, their branches heavy with snow**. ...... Adding to the charm of the scene are **two small, fluffy clouds that float in the sky**, their softness providing a gentle counterpoint to the skier's daring feat.



Dispreferred response generation:

1. Bad prompting: designed to elicit inaccurate descriptions by setting up a slightly different task (describe objects that would logically exist in the image) for the model.
2. Prompts are similarly generated using GPT-4 with human filtering.

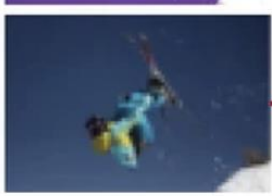
Stage 1: Image Comprehension Self-Training

Task: Explain what is depicted in the photograph.

***Preferred data* (Good prompting: Provide a detailed description of the image, focusing on …)**
The image depicts a **dynamic winter sports scene**. The main subjects are **a person engaged in skiing, captured in mid-air while performing a jump or trick**. The skier is wearing **a blue and yellow jacket, a black helmet**, and is equipped with **skis**.....

(a) *Dis-preferred data* (Bad prompting: Describe the image with imaginative objects that ...)
...... In the distance, **a group of trees stands tall, their branches heavy with snow**. ...... Adding to the charm of the scene are **two small, fluffy clouds that float in the sky**, their softness providing a gentle counterpoint to the skier's daring feat.

(b) *Dis-preferred data* (Normal prompting + Image corruption: Color jitter.)
...... The skier, clad in a **vibrant purple jacket** and blue pants, ... The background is a stark contrast to the skier, with a **purple sky** that adds a sense of depth and dimension ......



(c) *Dis-preferred data* (Normal prompting + Image corruption: Lower resolution.)
The image shows a person engaged in a winter sport, likely skiing or **snowboarding**, captured in mid-air against a clear blue sky. The individual is wearing a blue and yellow suit, ... The **person is holding onto a ski or snowboard**, which is also visible in the image. The **motion blur effect** ......

UCLA ENGINEERING
Computer Science

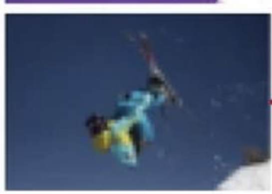
Stage 1: Image Comprehension Self-Training

Task: Explain what is depicted in the photograph.

Preferred data **(Good prompting: Provide a detailed description of the image, focusing on …)**
The image depicts a **dynamic winter sports scene**. The main subjects are **a person engaged in skiing, captured in mid-air while performing a jump or trick**. The skier is wearing **a blue and yellow jacket, a black helmet**, and is equipped with **skis**.....

(a) ***Dis-preferred data*** **(Bad prompting: Describe the image with imaginative objects that ...)**
...... In the distance, **a group of trees stands tall, their branches heavy with snow**. ...... Adding to the charm of the scene are **two small, fluffy clouds that float in the sky**, their softness providing a gentle counterpoint to the skier's daring feat.

(b) ***Dis-preferred data*** **(Normal prompting + Image corruption: Color jitter.)**
...... The skier, clad in a **vibrant purple jacket** and blue pants, ... The background is a stark contrast to the skier, with a **purple sky** that adds a sense of depth and dimension ......



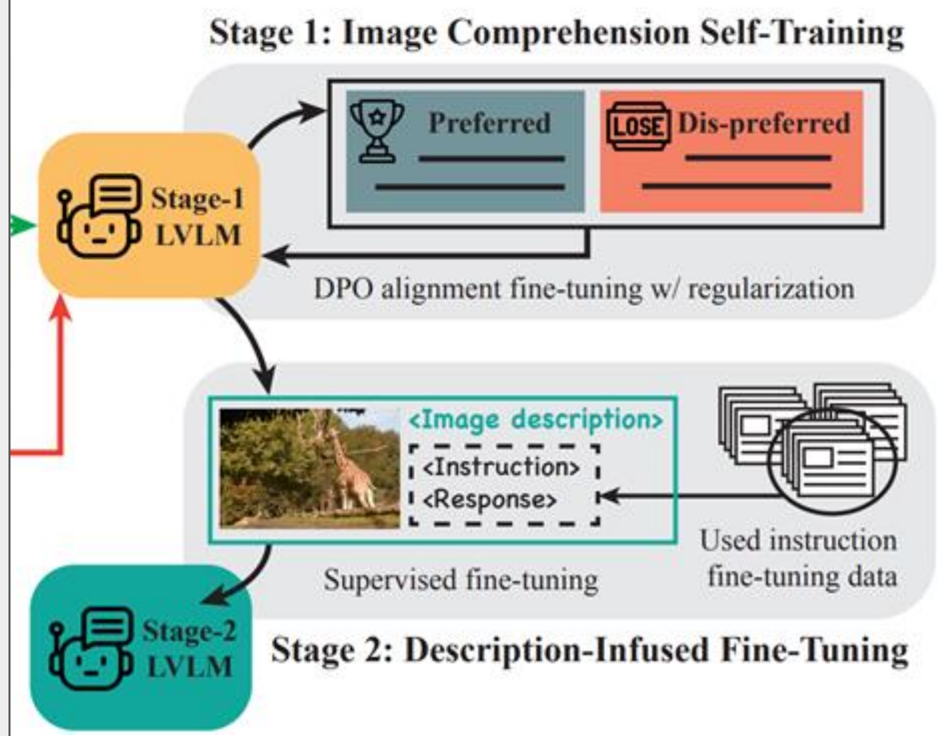
(c) ***Dis-preferred data*** **(Normal prompting + Image corruption: Lower resolution.)**
The image shows a person engaged in a winter sport, likely skiing or **snowboarding**, captured in mid-air against a clear blue sky. The individual is wearing a blue and yellow suit, ... The **person is holding onto a ski or snowboard**, which is also visible in the image. The **motion blur effect** ......

$$\text{Update } \boldsymbol{\theta}_1 = \operatorname{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{(\mathbf{x},\mathbf{y}_g,\mathbf{y}_b)\in D} \Big[\underbrace{\ell\Big(\lambda \log \frac{p_{\boldsymbol{\theta}}(\mathbf{y}_g|\mathbf{x})}{p_{\boldsymbol{\theta}_0}(\mathbf{y}_g|\mathbf{x})} - \lambda \log \frac{p_{\boldsymbol{\theta}}(\mathbf{y}_b|\mathbf{x})}{p_{\boldsymbol{\theta}_0}(\mathbf{y}_b|\mathbf{x})}\Big)}_{\text{DPO}} - \underbrace{\alpha \log p_{\boldsymbol{\theta}}(\mathbf{y}_g|\mathbf{x})}_{\text{Regularizer}}\Big].$$

How can we improve fine-grained visual perception in reasoning?

Stage 2:

Fine-tune LVLM with generated detailed image description



Enhancing Large Vision Language Models with Self-Training on Image Comprehension

Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Zou, Kai-Wei Chang, and Wei Wang, in *NeurIPS*, 2024.

Stage 2: Description-Infused Fine-Tuning

- Randomly select a small set (50k) data
- Infuse instructions with image description

```
Image description:  {model description}
<original instruction>
```

for $i = 1, \ldots m$ **do**

Randomly sample $\mathbf{x}_{\text{des}} \sim \{\mathbf{x}_{\text{des}}^{(i)}\}_{i \in [M]}$.

Generate model image description $\mathbf{y}_{\text{des}} \sim p_{\boldsymbol{\theta}_t}(\cdot|\mathbf{v}^{(i)}, \mathbf{x}_{\text{des}})$.

Add $\left([\mathbf{y}_{\text{des}}, \mathbf{x}^{(i)}], \mathbf{y}^{(i)}\right)$ to D_{des}.

end for

Update $\widehat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{(\mathbf{x},\mathbf{y}) \in D_{\text{des}}} \ell\Big(\log p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})\Big)$.

UCLA ENGINEERING
Computer Science

Performance

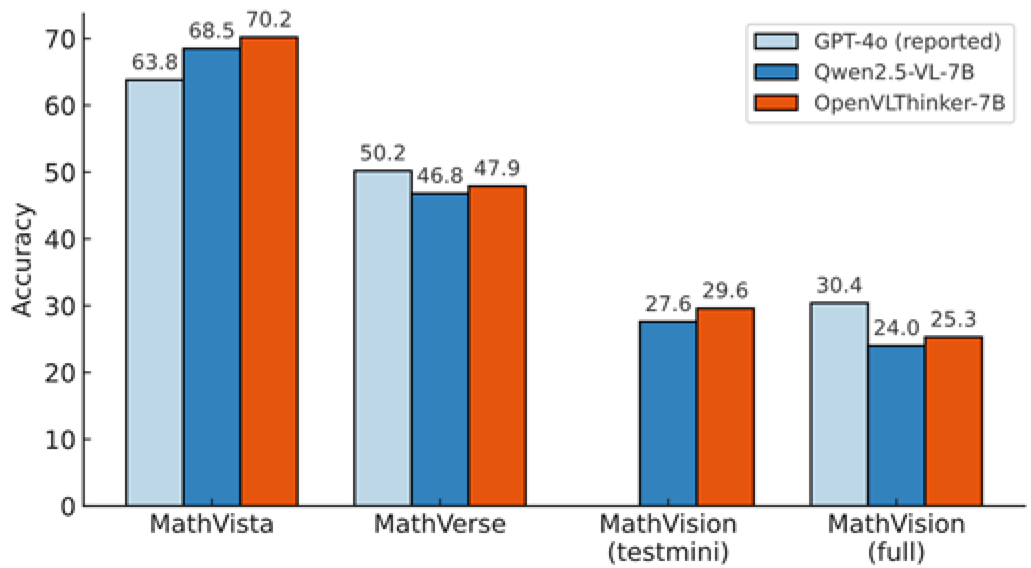
Model	ScienceQA	TextVQA	ChartQA	LLaVA-Bench	MMBench	MM-Vet	MathVista
InstructBLIP (7B)	60.5	50.1	–	60.9	36.0	26.2	25.3
mPLUG-OWL2 (7B)	64.5	54.3	–	59.9	64.5	36.2	22.2
LLaVA-v1.5 (7B)	66.8	58.2	6.3	65.4	64.3	31.1	25.1
w/ POVID	68.8	–	–	68.7	64.9	31.8	–
w/ STIC	**69.5**	**61.4**	**6.6**	**68.9**	**65.3**	**32.6**	**27.2**
LLaVA-v1.6 (7B)	68.9	60.3	36.4	77.3	63.7	42.2	34.6
w/ STIC	**75.3**	**65.2**	**41.5**	**79.2**	**67.8**	**45.0**	**37.0**

Achieve 4% improvement over 7 benchmarks

OpenVLThinker: R1-like Reasoning for Vision-Language Models

Can we distillate reasoning capability into vision-language LLMs?

Yes, via iterative SFT+RL



70.2% accuracy on MathVista, exceeding GPT-4o and its base model.
The performance is comparable to Qwen2-VL-72B (70.5%)

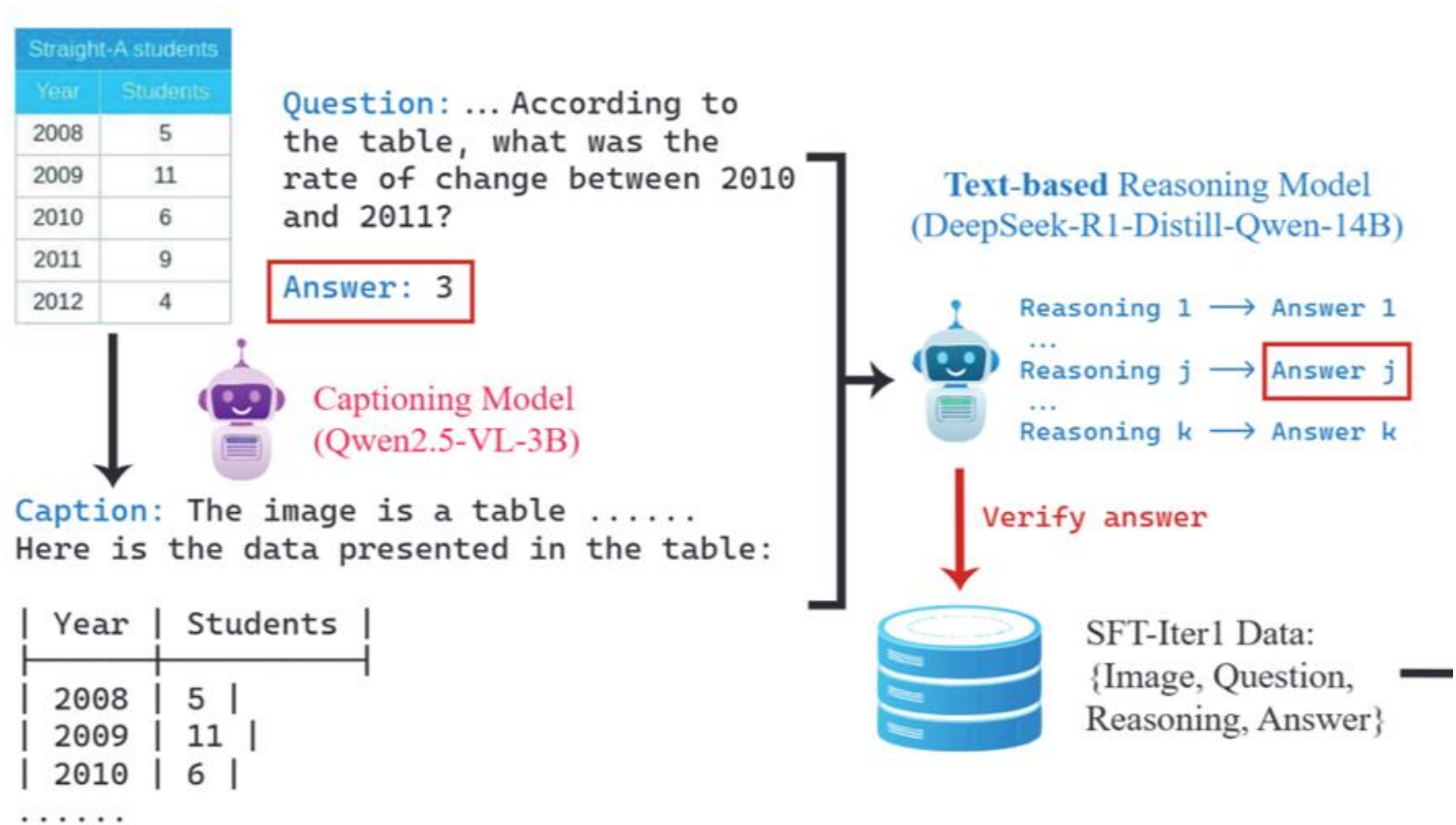
OpenVLThinker: An Early Exploration to Complex Vision–Language Reasoning via Iterative Self–Improvement

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, Kai–Wei Chang

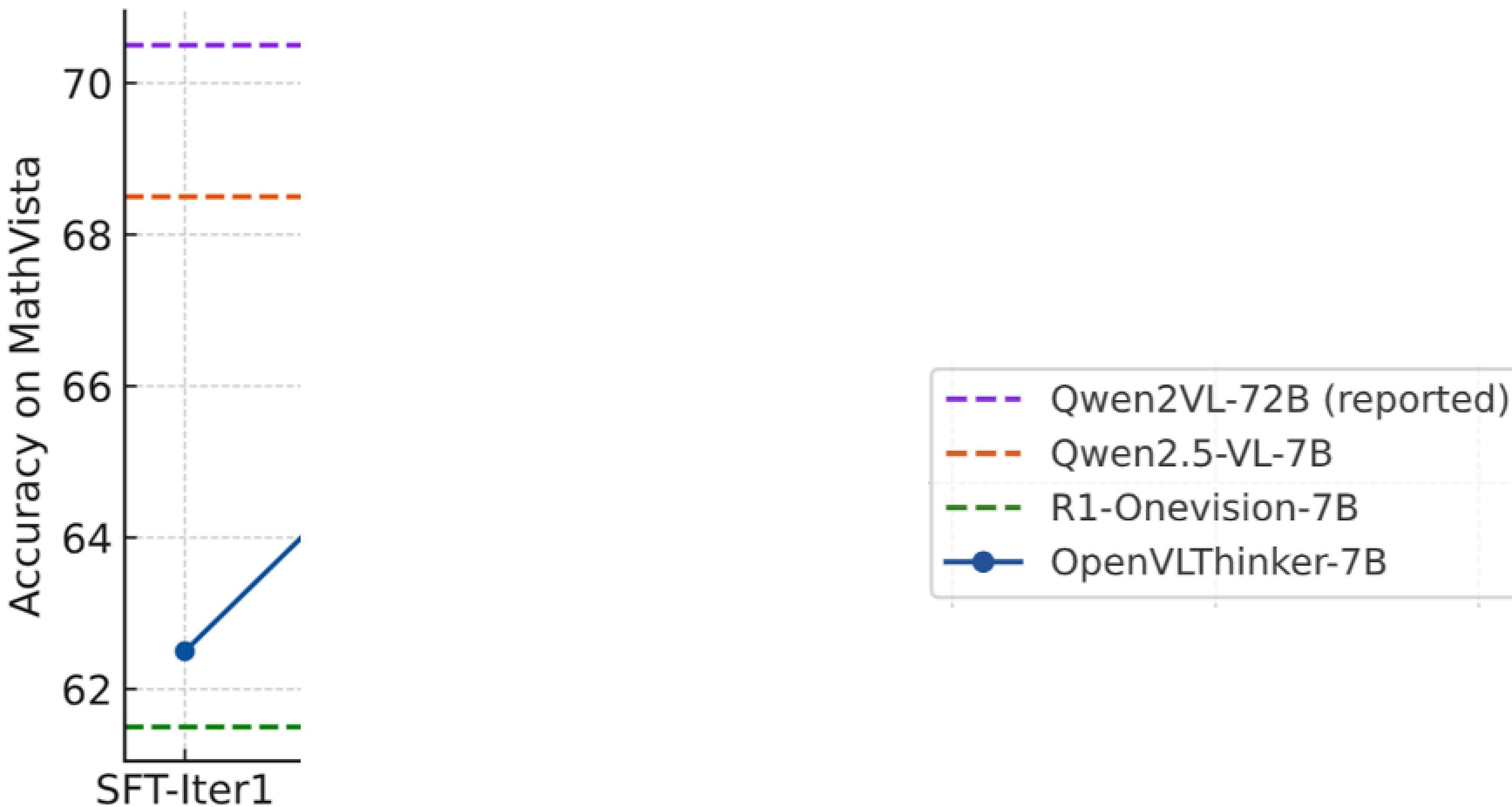
```
Assistant:
<think>
Okay, so I have this triangle XYZ with sides labeled as 9x, 4x + 5, and 6x + 3.
The problem is asking me to find the length of side XY, which is 4x + 5.

First, I remember that in a triangle, all sides are equal because it’s an
equilateral triangle. That means all sides are the same length. So, I set the
expressions for the sides equal to each other.

I start with the sides: 9x = 4x + 5 and 9x = 6x + 3.

Let me solve the first equation, 9x = 4x + 5. I subtract 4x from both sides:
9x - 4x = 5, which simplifies to 5x = 5. Then, I divide both sides by 5: x = 1.

Now, I substitute x = 1 into the expression for side XY, which is 4x + 5. So,
4(1) + 5 = 4 + 5 = 9.

I double-check my work. The other sides are 9x and 6x + 3. Substituting x = 1,
9x becomes 9(1) = 9, and 6x + 3 becomes 6(1) + 3 = 9.

Both sides are equal to 9, so the triangle is equilateral with all sides
equal to 9.
</think>

The length of side XY is 9.

<answer>D</answer>
```

UCLA ENGINEERING
Computer Science

Collect Reasoning Paths



UCLA ENGINEERING
Computer Science

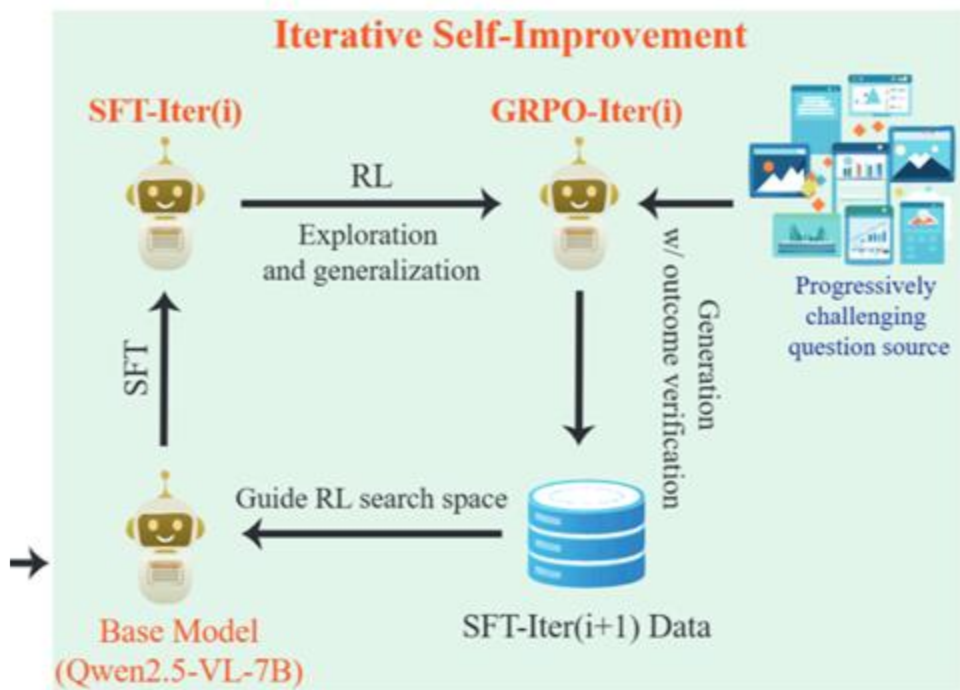

Accuracy on MathVista
70
68
66
64
62
SFT-Iter1
Qwen2VL-72B (reported)
Qwen2.5-VL-7B
R1-Onevision-7B
OpenVLThinker-7B

Iterative SFT & RL

Iterative Self-Improvement

GRPO-Iter: Train LVLM using RL (GRPO)

SFT-Iter: Train on reasoning paths generated by previous GRPO…



UCLA ENGINEERING
Computer Science

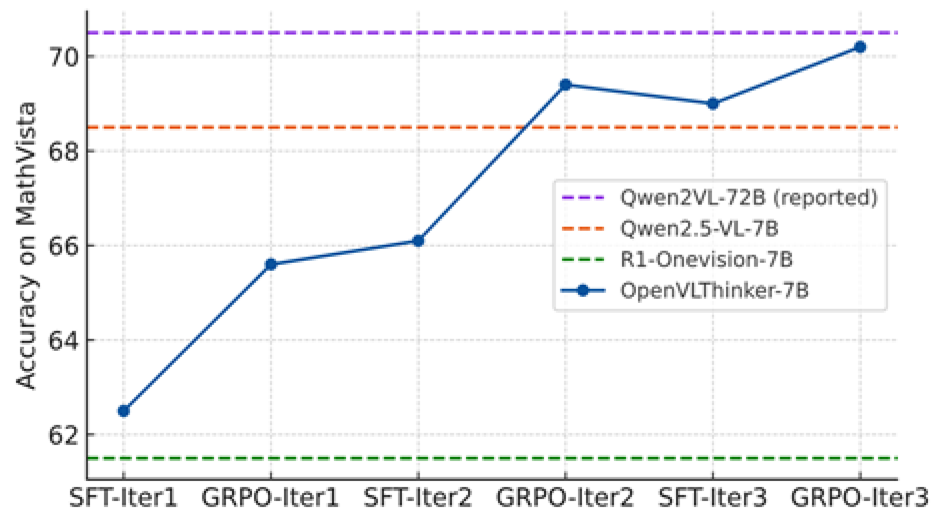
OpenVLThinker: Iterative SFT & RL

Role of SFT and RL

We hypothesized that

- SFT plays a role in setting up the model's reasoning frameworks.

- RL plays as a more significant contributor to generalization.



Conclusion

- If we aim for AI to behave like humans, we need the model to attention to details
- Instruction fine-tuning guides models to focus on details

- Thank you for your attention.
- Question?