RE-ALigning Language to Visual Objects with an Agentic Workflow

Yuming Chen, Jiangyan Feng, Haodong Zhang, Lijun Gong, Feng Zhu, Rui Zhao

Qibin Hou#, Ming-Ming Cheng, Yibing Song#

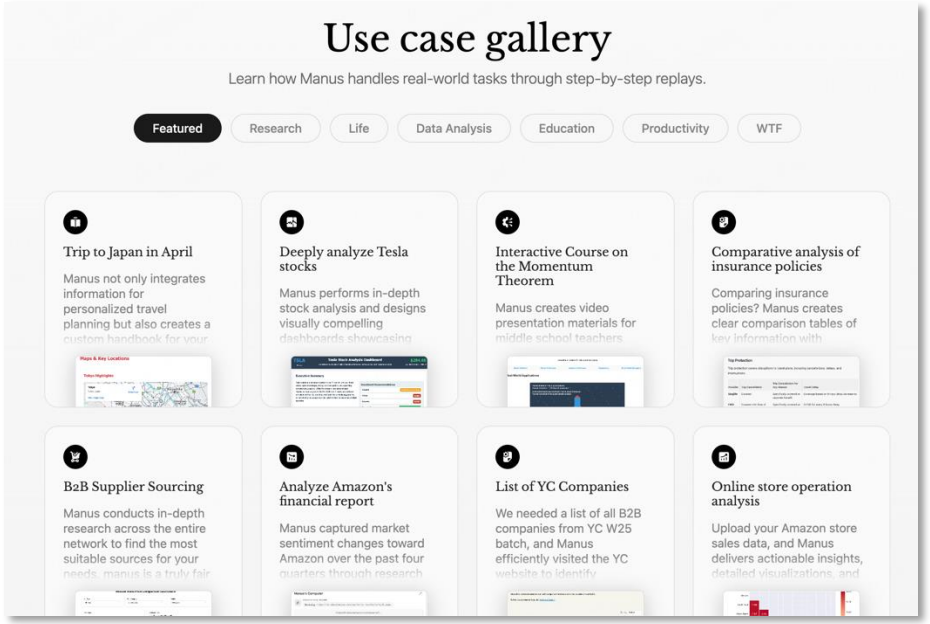
ICLR

Key Point

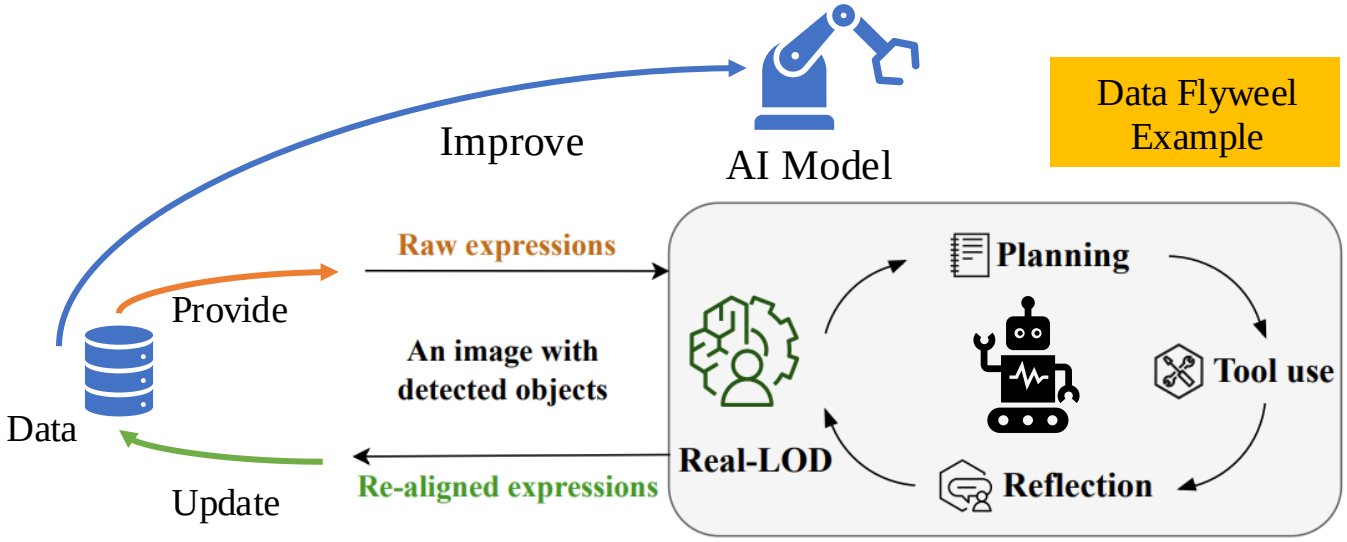
Real-LOD

Agentic Workflow: from Assistant to Data Flywheel

- Agent is not only a *simple assistant* but also can establish workflows that serve as **flywheel** to sustaining **high-value data assets** across AI industries
- In this paper, we show an application of agentic workflow to demonstrate its potential…



Manus[1]: Recent Viral Agent Assistant



Our agentic workflow proposed to provide data with **high Visual-Language alignment quality**

[1] Manus: https://manus.im/

Content

Real-LOD

- **Background**
- Method
- Example
- Experiments

Real-LOD

Language-based Object Detection

- The essence of *Object Detection* is to **Align** Language Data to Visual Objects

Align

person
single noun word

- person with helmet
- person on motorcycle

short phrases

- person wearing black boots and jeans in the middle of image.
- person sat on the back of a black motorcycle
- person sat behind someone wearing a skirt.

long description

Early

Coarse-grained
Categories

Peoples usually describe objects in various forms

Longer!

More Detailed!

More Aspects!

More Challenging!

Fine-grained
Expressions

Recent

LOD
Language-based Object Detection

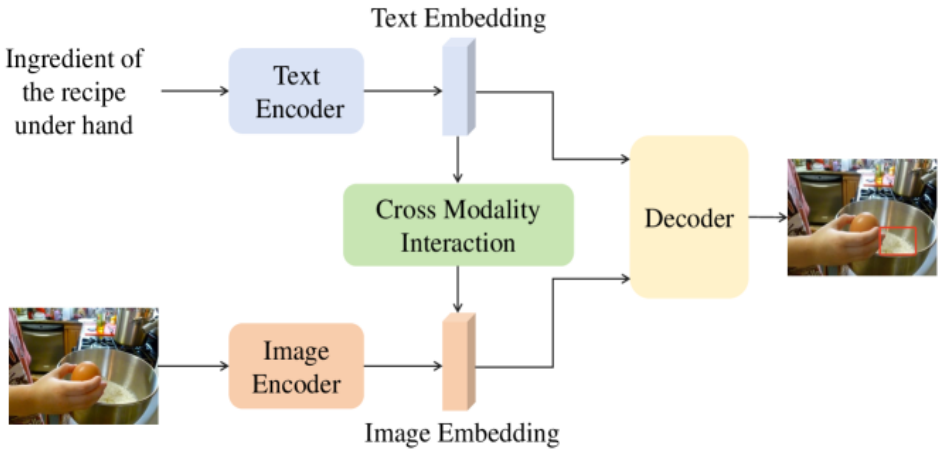
Real-LOD

Visual-Language Alignment

- **Quantity** & **Diversity** of **Visual-Language data** for alignment is **important**!

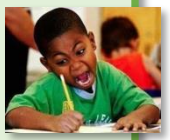
Discriminative Model → CLIP[1] → Generative Model

General Generative Framework



Training Data Quantity:

- *OWL* [2] : **10B** images
- *GDINO* [3] : **5.8M** images
- *APE-B* [4] : **2.6M** images



Collection Challenge:

- Objects are **harder** to label than whole images
- The costs for large-scale data are **extremely high**

Expression Diversity

- Training Dataset: **Simple**
- Testing Dataset: **Complex**

OmniLabel [5] — low

Images	Method	AP	AP-categ	AP-descr
All	RegionCLIP [53]	2.7	2.7	2.6
	Detic [55]	8.0	15.6	5.4
	MDETR [20]	-	-	4.7
	GLIP-T [29]	19.3	23.6	16.4
	GLIP-L [29]	25.8	32.9	21.2
	FIBER-B [12]	25.7	30.3	22.3

- Introducing *VLM* for **expression generation** is promising, but…

[1] Radford et al., Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.
[2] Minderer et al., OWL-ViT: Open-World Object Detection with Vision Transformers. ECCV 2022.
[3] Liu et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. ECCV 2024.
[4] Shen et al., Aligning and Prompting Everything All at Once for Universal Visual Perception. CVPR 2024.
[5] Schulter et al., OmniLabel: A Challenging Benchmark for Language-Based Object Detection. ECCV 2024.

Real-LOD

Hallucination of VLM

- **Small object** and **General Prompt** cause *VLM* to misperceive objects, damaging **Quality**

<Prompt>: What is the object in the red box?
<Response>: The object is a blue and red ***hat worn by a young girl***.

(a) For Small Object

<Prompt>: What is the object in the red box?
<Response>: The object is ***a flag*** with a design that resembles ***the Union Jack***.

(b) With Extended Crop

<Prompt>: Describe the person in the red box.
<Response>: The person is a woman with long hair wearing a ***black top***.

(c) With General Prompt

<Prompt>: What is the color of person's clothes.
<Response>: The person is wearing ***a black and white striped shirt, dark pants, and black shoes.***

(d) With Specified Prompt

- By adopting **appropriate operations**, hallucination can be alleviated

Content

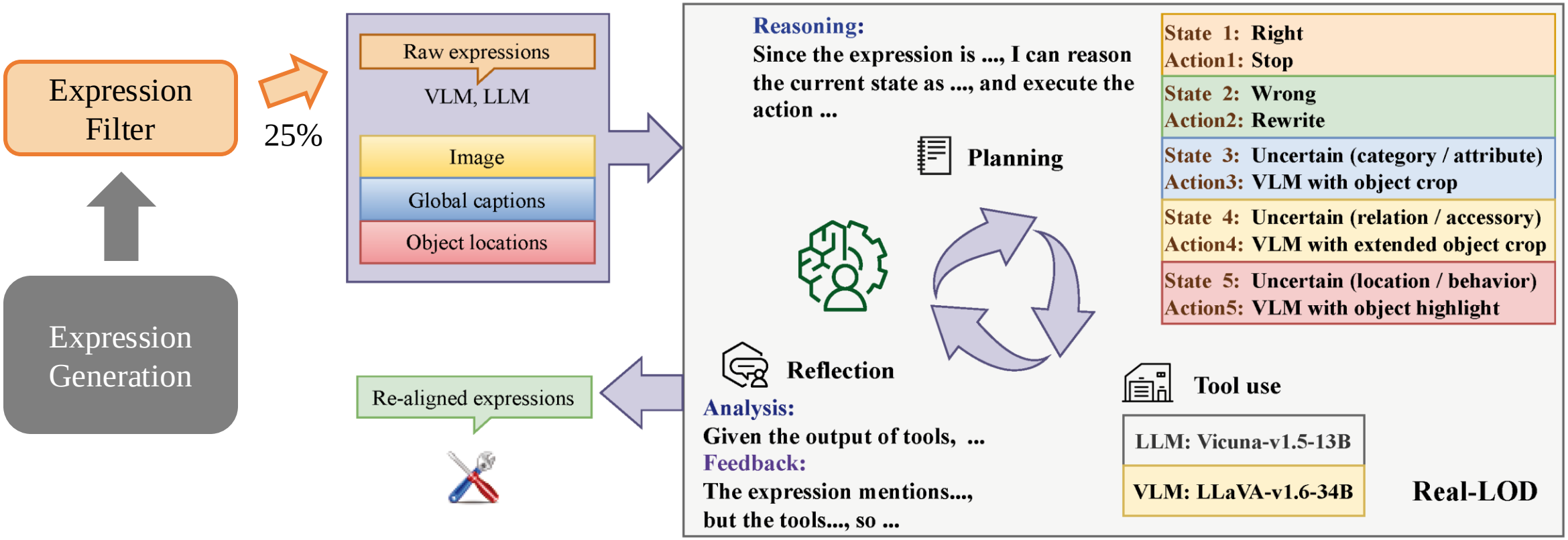
Real-LOD

- Background
- **Method**
- Example
- Experiments

Real-LOD

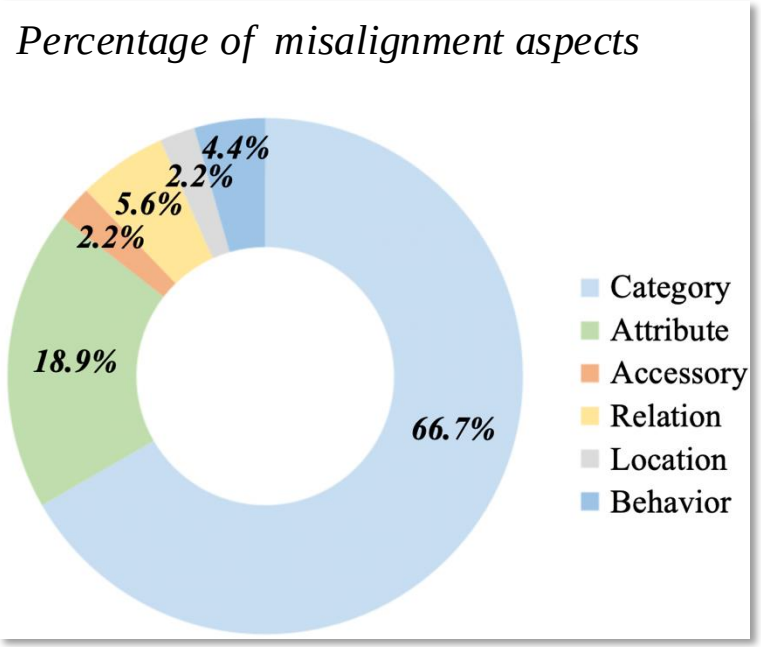
Agentic Workflow

- To improve the **alignment quality** of large-scale VL data, **Real-LOD** involves cycles with **3 steps**: ***Planning***, ***Tool use***, and ***Reflection***



Real-LOD

Agentic Workflow



Analyze

State 1: ***Correct***

State 2: ***error***

State 3: ***Uncertain*** Category/Attribute → **Action 3:** Object crop

State 4: ***Uncertain*** Relation/Accessory → **Action 4:** Extended object crop

State 5: ***Uncertain*** Location/Behavior → **Action 5:** Object Highlight

Language prompt

Here are some prior knowledge about this image. The object in the image is a {*object category*}. There is a possible description of this image, it may not be precise enough:{*current expression*}. Answer the following questions. {*Questions generated by Real-LOD*}.

Visual prompt

Visual prompts of three actions for VLM:

After analyzing the **misalignment** caused by **hallucination**, **5 states/actions** are specifically designed for **Real-LOD**

Content

Real-LOD

- Background
- Method
- **Example**
- Experiments

Example

Real-LOD

Input

Object Category horse

Bbox Coordinates [223.07, 267.79, 101.47, 59.39]

Image Caption

The image depicts a group of people on horseback in a forested area. The person riding the horse in the red box is wearing a blue shirt and dark pants, and appears to be leading the group. The horse is brown and is walking on a dirt path. The riders are wearing backpacks, suggesting they might be on a trail ride or a guided tour. The forest is lush with green trees, and the sky is visible through the canopy, indicating it's a sunny day. The overall atmosphere of the image is one of outdoor adventure and exploration.

Can you specify if the raw expression that ***brown horse walking on sand*** is referring to the target object? If not, please modify it to a correct expression.

Output

The expression ***brown horse walking on sand*** is a wrong one, and a correct expression could be ***brown horse walking on dirt path with rider in blue shirt and hat, leading the way***.

Example

Real-LOD

First Cycle

Planning

Reasoning: The expression describes a brown horse that is walking on sand, but the image caption mentions the horse is walking on a dirt path, and there is no mention of sand in the caption. Based on this information, I reason the state is 4 where I am uncertain of object relation with surroundings. So I plan for Action 4 where I can crop a local object region in this image for VLM to further re-perceive. The description of VLM supplements the current information for me to conduct a more thorough expression examination.
Action: [{'Tool_name': 'VLM', 'Tool_parameters': {'Prompt': 'Questions:1. Is the horse walking on sand or on a dirt path?', 'Image editing': 'extended object crop.'}}]
Response: I will consult with a VLM to acquire supplement information that will enable me to evaluate the current expression more thoroughly.

Tool use

VLM output: 1. The horse is walking on a dirt path, not on sand.

Reflection

Feedback: The expression {***brown horse walking on sand***} describes a wrong relation (on sand) of the horse.

Example

Real-LOD

Second Cycle

Planning

Reasoning: Based on the feedback, I reason the state is 2 where I am certain the expression {***brown horse walking on sand***} is wrong. So I plan for Action 2 where I use an LLM to rewrite the expression, and then check the correctness of the new expression.
Actions: [{'Tool_name': 'LLM', 'Tool_parameters': {'Prompt: Image Caption, ..., Response, ...'}}]
Response: The expression {***brown horse walking on sand***} contains flaws, now I will first leverage the LLM to create an alternative expression for you.

Tool use

LLM output: New Description: {***brown horse walking on dirt path with rider in blue shirt and hat, leading the way***}

Reflection

Feedback: The expression {***brown horse walking on dirt path with rider in blue shirt and hat, leading the way***} describes the object category 'horse' and provides similar context to the caption.

Final Cycle

Planning

Reasoning: Based on the feedback, I reason the state is 1 where I am certain the new expression {***brown horse walking on dirt path with rider in blue shirt and hat, leading the way***} is correct. So I plan for Action 1 to stop.
Actions: []
Response: The expression {***brown horse walking on sand***} is a wrong one, and a correct one could be {***brown horse walking on dirt path with rider in blue shirt and hat, leading the way***}.

Content

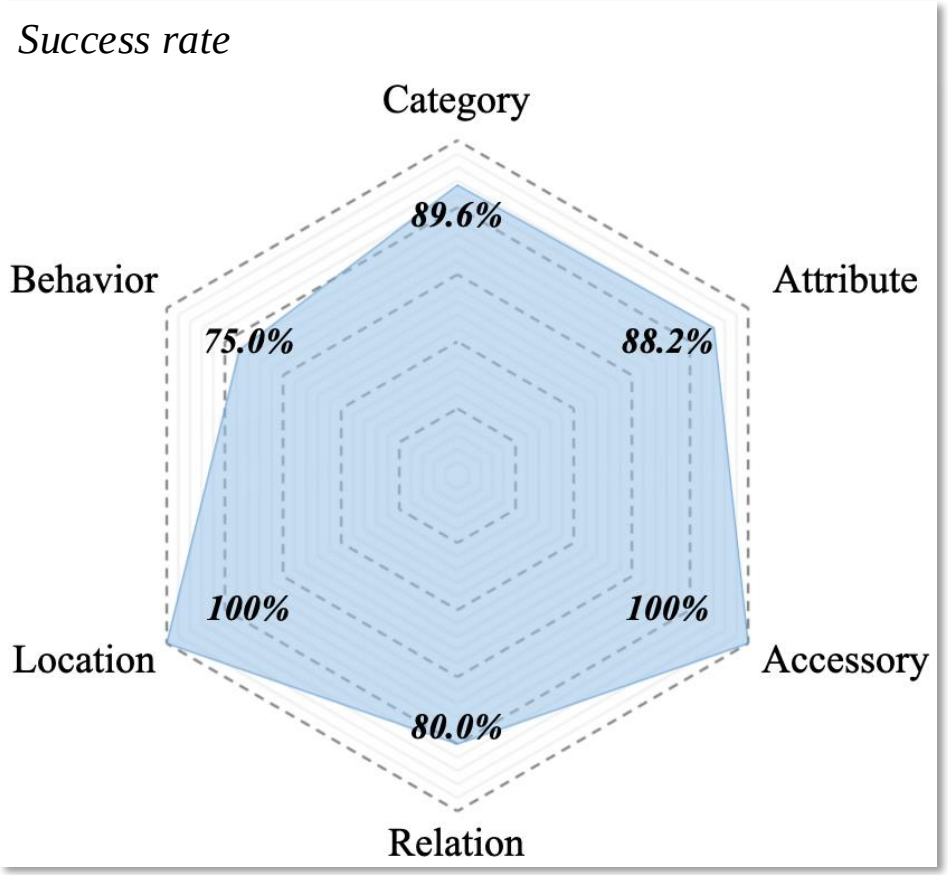
Real-LOD

- Background
- Method
- Example
- **Experiments**

Experiment

Real-LOD

Re-aligned Data Analysis



Category

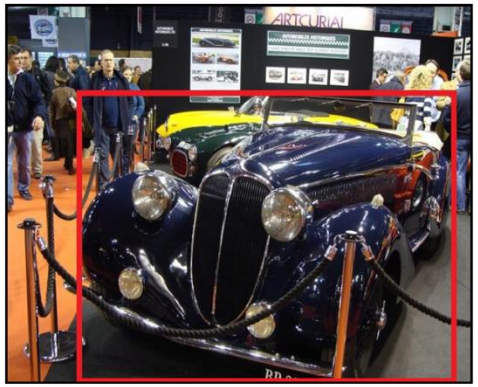
Raw: a black and silver toaster oven
Refined: black and silver coffee machine

Relation

Raw: bread on the dining table
Refined: fresh white bread held by girl

Location

Raw: cylindrical instrument being played by a person on the left side of the image
Refined: flute being played by person on the right side of the image



Attribute

Raw: yellow sports car
Refined: vintage sports car with blue finish

Accessory

Raw: person with a walking stick
Refined: person with a yellow-white mop for ice rank cleaning

Behavior

Raw: man in gray jacket is playing frisbee while a group of people watch and play instruments
Refined: man in a gray hoodie is directing the orchestra

Real-LOD

Ablation Study on OmniLabel Benchmark

Test subset	Training data type	#Img	AP-des	AP-des-pos	AP-des-S	AP-des-M	AP-des-L
COCO	A	933k	21.2	59.4	31.3	21.1	18.6
	B	695k	22.2	59.4	32.4	21.9	19.4
	C	863k	**24.2**	**59.6**	**35.2**	**24.2**	**21.1**
O365	A	933k	27.6	43.1	39.8	25.5	17.9
	B	695k	28.5	43.7	40.2	26.2	18.5
	C	863k	**32.4**	**48.5**	**47.5**	**30.0**	**21.3**
OI	A	933k	30.5	43.0	37.2	30.3	23.2
	B	695k	31.4	43.7	38.1	31.2	24.0
	C	863k	**33.5**	**44.9**	**42.2**	**32.9**	**24.8**

A: “raw expression”; **B:** “raw expression with filter”; **C:** “raw expression with filter + Real-LOD”;

Experiment

Real-LOD

OmniLabel Benchmark

LOD Method	#Img	AP-des	AP-des-pos	AP-des-S	AP-des-M	AP-des-L
GLIP (Swin-L)	17.5M	21.2	33.2	37.7	18.9	10.8
mm-GDINO (Swin-B)	12M	20.8	33.1	31.9	19.8	14.1
FIBER (Swin-B)	4M	22.3	34.8	38.6	19.5	12.4
Real-Model (Swin-B)	**0.18M**	**36.5**	**52.1**	**54.4**	**33.2**	**25.5**

DOD Benchmark

LOD Method	#Img	Full	Presence	Absence
GDINO (Swin-B)	5.8M	20.1	20.7	22.5
mm-GDINO (Swin-B)	12M	24.2	23.9	25.9
APE-B (ViT-L)	2.6M	30.0	29.9	30.3
Real-Model (Swin-B)	**0.18M**	**34.1**	**34.4**	**33.2**

Experiment

Real-LOD

RefCOCO/g/+ Benchmark

LOD Method	#Img	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val-u	test-u
APE-A (ViT-L)	2.0M	34.2	34.8	36.1	33.5	32.3	36.0	38.9	40.5
Real-Model (Swin-B)	**0.18M**	**74.0**	**79.6**	**66.0**	**76.4**	**83.1**	**68.5**	**80.8**	**81.2**
GDINO* (Swin-B)	5.8M	-	-	-	73.6	82.1	64.1	78.3	78.1
APE-B* (ViT-L)	2.6M	84.6	89.2	80.9	76.4	82.4	66.5	80.0	80.1
Real-Model* (Swin-B)	**0.24M**	**91.3**	**93.1**	**88.0**	**85.4**	**90.3**	**78.6**	**88.4**	**89.0**

* indicates that the model employs RefCOCO/g/+ for training

OVDEval Benchmark

LOD Method	#Img	color	material	position	relationship	negation	avg
GLIP (Swin-L)	17.5M	6.7	15.8	48.1	33.2	51.8	31.1
OmDet (ConvNext-B)	1.1M	24.5	22.5	47.7	51.8	55.8	40.4
FIBER (Swin-B)	4M	9.4	17.7	48.1	33.2	58.1	33.3
Real-Model (Swin-B)	**0.18M**	**25.7**	**22.5**	**59.3**	**41.9**	**68.4**	**43.6**

Experiment

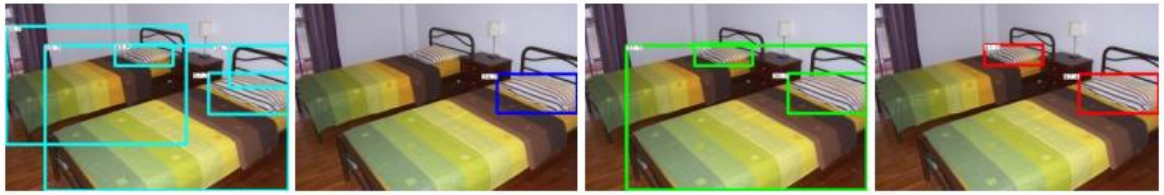
Real-LOD

Visualization

Query: "This item is used to keep warm in colder weather."

Query : "Woman in wedding dress next to a man in suit."

Query : "Pillow placed at the head of the bed."



(a) GLIP-L (b) APE-B (c) mm-GDINO (d) **Our**

Query : "These two people each have a pink surfboard."

Query : "The fire extinguisher on the left."

Query : "Cows that are laid down."

(a) GLIP-L (b) APE-B (c) mm-GDINO (d) **Our**

Thanks For Watching!

GitHub

Main Page

Wechat