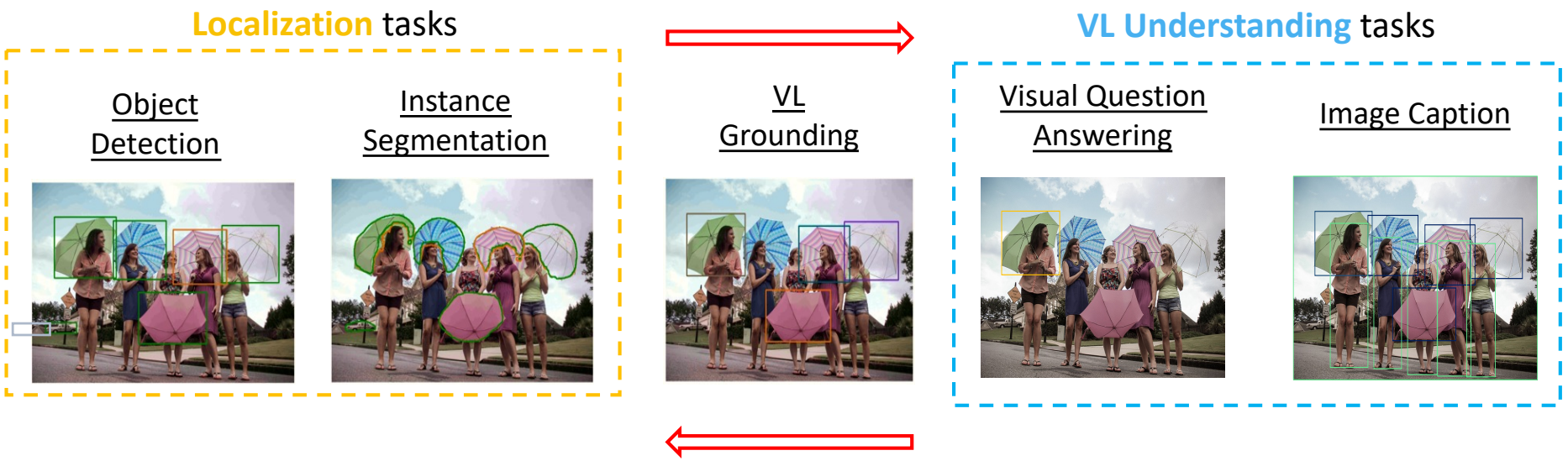
GLIPv2: Unifying Localization and Vision-Language Understanding

Haotian Zhang*[1], Pengchuan Zhang*[2], Xiaowei Hu[3], Yen-Chun Chen[3], Liunian Harold Li[4], Xiyang Dai[3], Lijuan Wang[3], Lu Yuan[3], Jenq-Neng Hwang[1], Jianfeng Gao[3]

[1]University of Washington, [2]Meta AI, [3]Microsoft, [4]UCLA

General Purpose Vision Models

- Grounded VL Understanding

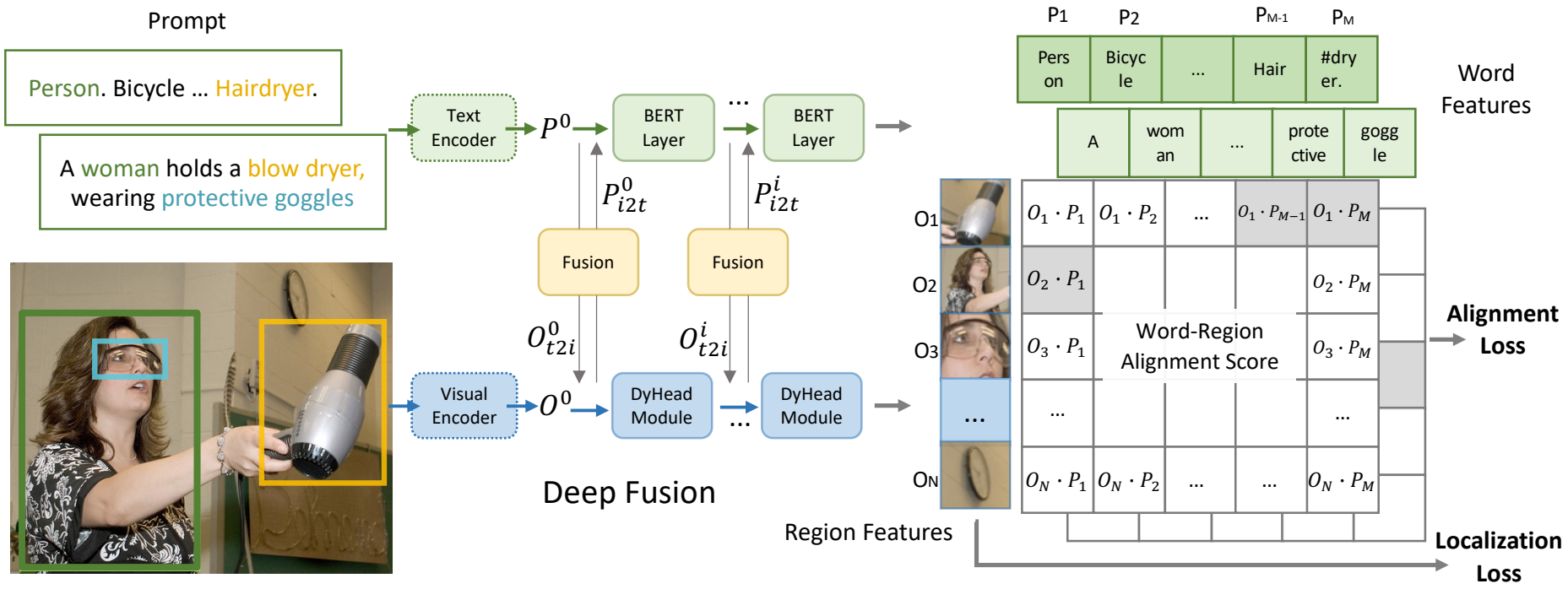


- General Purpose Vision Model requires **Unification**: Localization & VL Understanding
  1) Localization Tasks: Vision-only, fine-grained outputs.
  2) VL Understanding Tasks: Both modalities, high-level semantic outputs.

- To achieve **mutual benefits**; **simplify pre-training procedure; reduce pre-training cost** – "Grounded VL Understanding"

GLIP: Grounded Language-Image Pre-training

- GLIP[2] : A Unified Framework for Detection and Grounding



- Reformulate the **Object Detection** task as **Phrase Grounding** task.
- Pre-train the model with **scalable** and **semantic-rich** grounded data.

[2] Li Liunian*, Pengchuan Zhang*, Haotian Zhang*, et al. "Grounded Language-Image Pre-training." , CVPR2022 (**Best Paper Finalist**)

GLIPv2: Unifying Localization and Vision-Language Understanding

NEURAL INFORMATION PROCESSING SYSTEMS

- Potential Limitations during Pre-training in GLIP

Query: Santa Claus climbing stairs

- Limitation: ***Intra***-image region-word Contrastive Loss

['Santa', 'Claus']

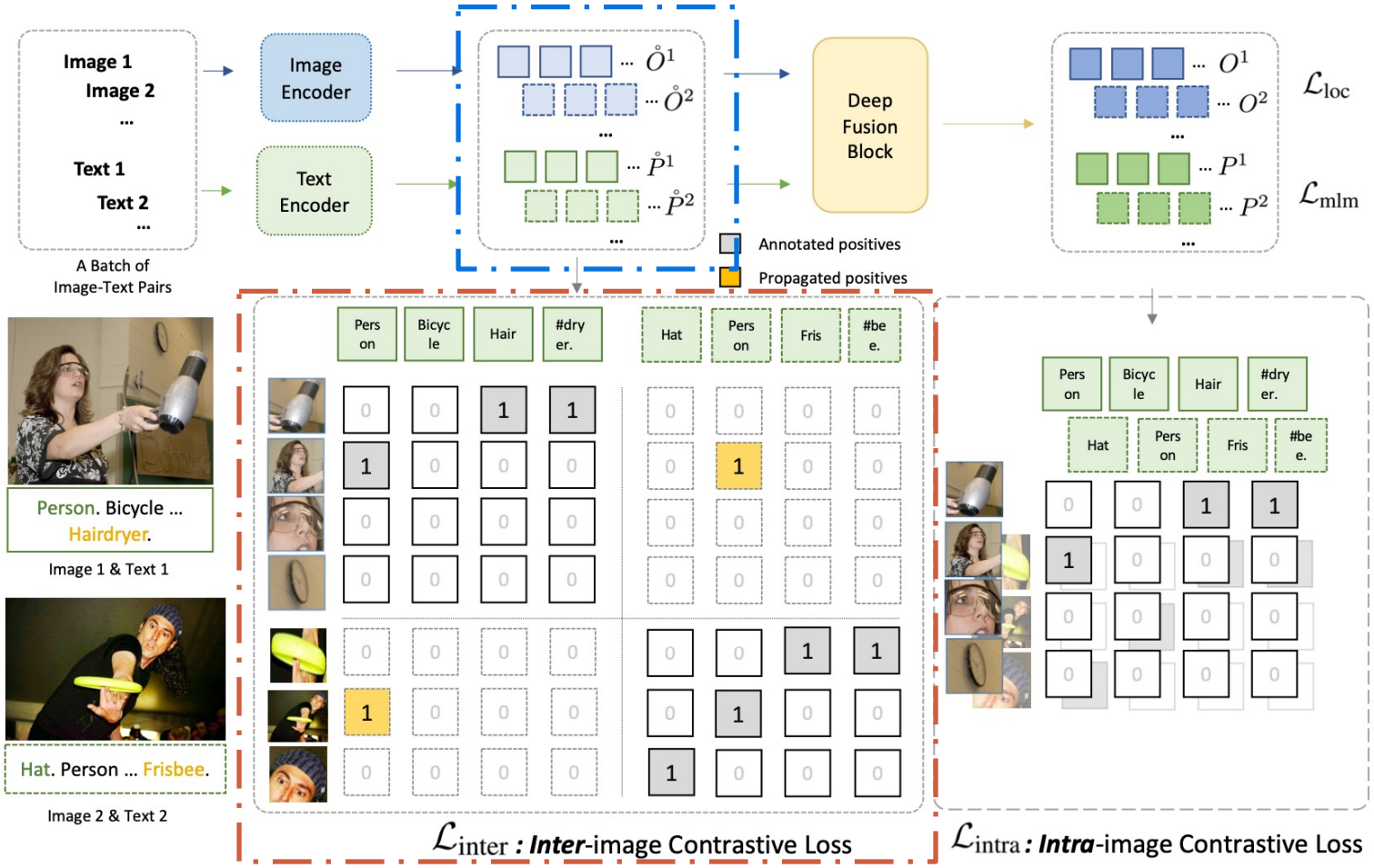
['Santa', 'Claus', 'climbing', 'stairs']

- The supervision signal becomes weak, when the caption is not long, and entities are not a lot.

- To make the pre-training tasks become harder and let the model better learn the information from self-trained image-text pair data…

GLIPv2: Unifying Localization and Vision-Language Understanding

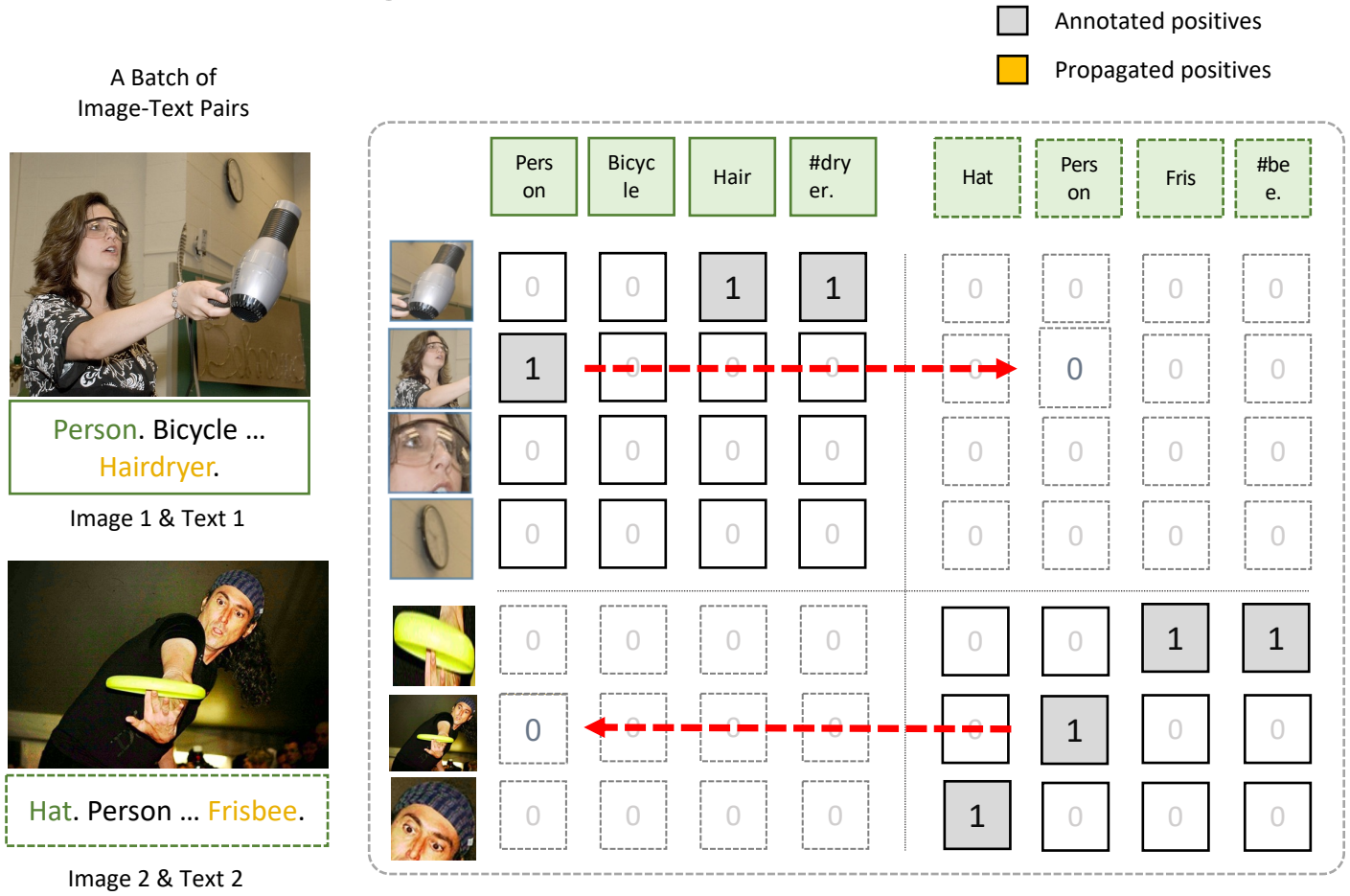
- A stronger VL pre-training task: ***Inter***-Image region-word Contrastive Loss



- Further increase the number of negative samples for across instances.
- 'Shallow' features directly from image and text encoder. Not after Fusion.
- Goal: Learn more **discriminative** region-word features.

GLIPv2: Unifying Localization and Vision-Language Understanding

- Label Propagation



Standard Contrastive Loss (No label propagation)

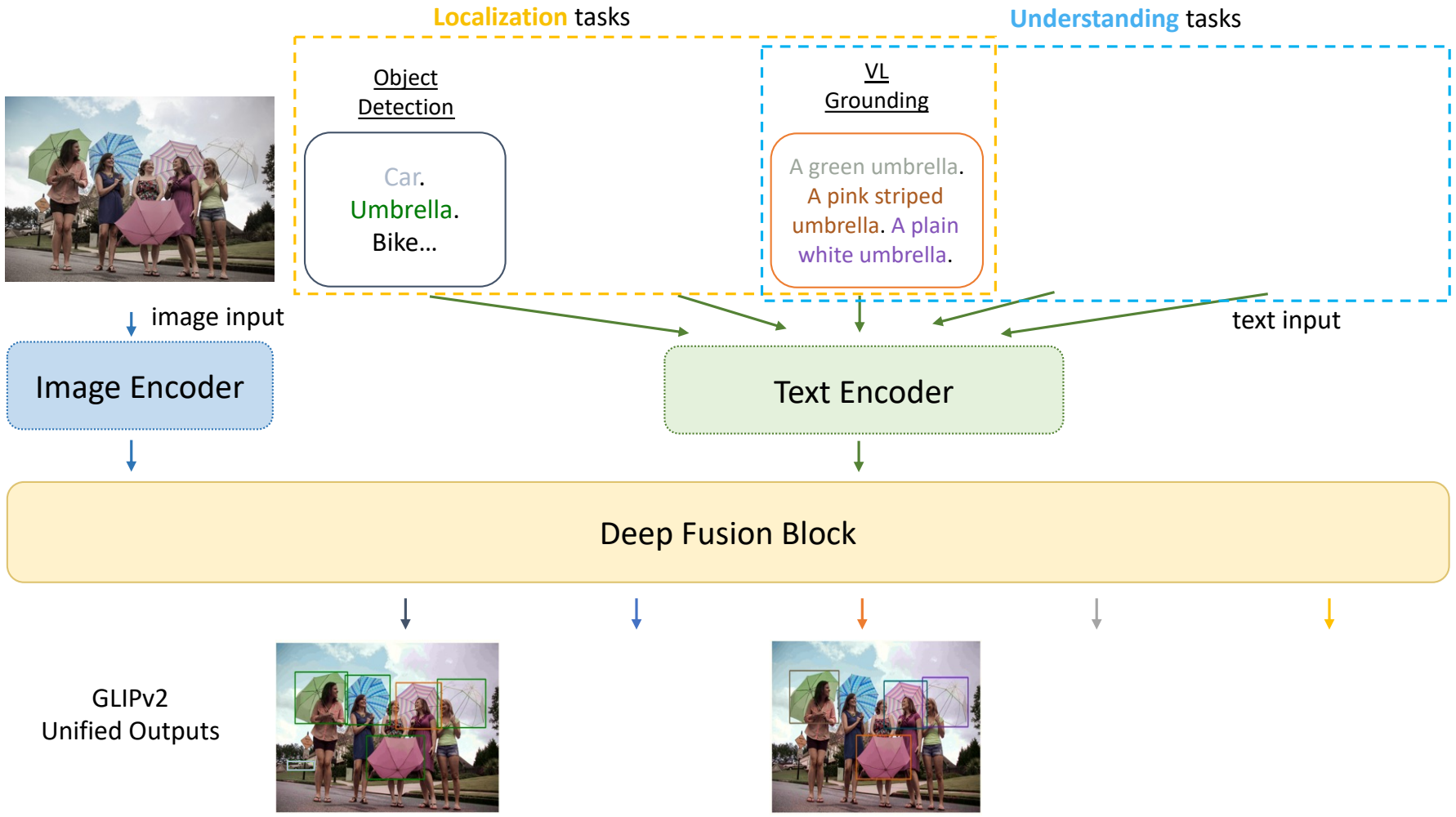
- Different from the standard Contrastive Loss, e.g., CLIP[3].
- Only propagate positives to detection-type texts
- Not propagate positives to grounding-type texts

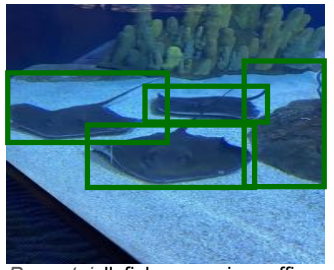
[3] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." arXiv preprint arXiv:2103.00020 (2021).

GLIPv2: Unifying Localization and Vision-Language Understanding

NEURAL INFORMATION PROCESSING SYSTEMS

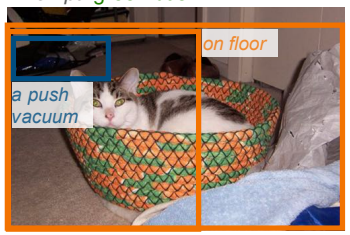
- Wide range of downstream tasks



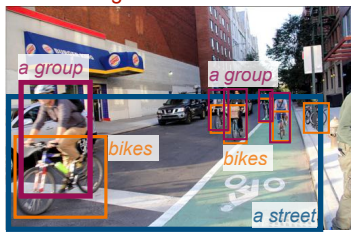


Prompt: jellyfish. penguin. puffin. shark. starfish. Stingray.

Prompt: green bush.



Input: Where is a push vacuum?
Prediction: on floor
Gold: background



Generated Caption: a group of people riding bikes down a street.

GLIPv2: Unifying Localization and Vision-Language Understanding

NEURAL INFORMATION PROCESSING SYSTEMS

- Performance
  - **One model architecture**

Model	Model Type	COCO-Det (test-dev)	ODinW (test)	LVIS (minival)	COCO-Mask (test-dev)	Flickr30K (test)	PhraseCut (test)	VQA (test-dev / test-std)	Captioning (Karpathy-test)
Mask R-CNN (23)	Localization	39.8	-	33.3 / -	- / 37.1	-	-	-	-
DETR (9)		42.0	-	17.8 / -	-	-	-	-	-
DyHead-T (15)		49.7	60.8	-	-	-	-	-	-
DyHead-L (15)		60.3*	-	-	-	-	-	-	-
VisualBERT (34)	Understanding	-	-	-	-	71.33	-	70.8 / 71.0	-
UNITER (12)		-	-	-	-	-	-	73.8 / 74.0	-
VinVL (58)		-	-	-	-	-	-	**76.5 / 76.6**	130.8
GPV (21)	Localization & Understanding	-	-	-	-	-	-	62.5 / -	102.3
UniT (24)		42.3	-	-	-	-	-	67.6 / -	-
MDETR (25)		-	-	24.2 / -	-	84.3	53.7	70.6 / 70.6	-
Unicorn (55)		-	-	-	-	80.4	-	69.2 / 69.4	119.1
GLIP-T (36)	Localization & Understanding	55.2	64.9	-	-	85.7	-	-	-
GLIP-L (36)		61.5*	68.9	-	-	87.1	-	-	-
GLIPv2-T (**Ours**)	Localization & Understanding	55.5	66.5	50.6 / 41.4	53.5 / 42.0	86.5	59.4	71.6 / 71.8	122.1
GLIPv2-B (**Ours**)		58.8	69.4	57.3 / 46.2	59.0 / 45.8	87.5	**61.3**	73.1 / 73.3	128.5
GLIPv2-H (**Ours**)		**60.6 (62.4*)**	**70.4**	**59.8 / 48.8**	**59.8 / 48.9**	**87.7**	**61.3**	74.6 / 74.8	**131.0**

Table 1: One model architecture results. For COCO-Det test-dev, * indicates multi-scale evaluation. For LVIS, we report the numbers for both `bbox` and `segm` on minival to avoid data contamination due to the pre-training. For Flickr30K test, we report the metric under `R@1`. For COCO-Mask, we also report both `bbox` and `segm` on test-dev.

GLIPv2: Unifying Localization and Vision-Language Understanding

NEURAL INFORMATION PROCESSING SYSTEMS

- Performance
  - **One set of weights for localization tasks**

Model	Direct Evaluation				Prompt Tuning				
	COCO-Mask (minival)	ODinW (test)	LVIS-Det (minival)	Flickr30K (minival)	COCO-Det (test-dev)	ODinW (test)	LVIS (minival)	COCO-Mask (test-dev)	PhraseCut (test)
GLIP-T	46.6/–	46.5	26.0	85.7	–	46.5	-	-	-
GLIP-L	49.8/–	52.1	37.3	87.1	58.8	67.9	-	-	-
GLIPv2-T	**47.3**/35.7	48.5	**29.0**	86.0	53.4 (-2.1)	64.8 (-1.7)	49.3 / 34.8 (-1.3 / -6.6)	53.2 / 41.2 (-0.3 / -0.8)	49.4
GLIPv2-B	61.9†/43.4	54.2	48.5	87.2	59.0 (+0.2)	67.3 (-2.1)	56.8 / 41.7 (-0.5 / -4.5)	58.8 / 44.9 (-0.2 / -0.9)	55.9
GLIPv2-H	64.1†/47.4	**55.5**	50.1	87.7	**60.2 / 61.9*** (-0.4 / -0.5)	**69.1** (-1.3)	**59.2 / 43.2** (-0.6 / -5.7)	**59.8 / 47.2** (-0.0 / -1.7)	**56.1**

Table 2: One set of weights results v.s. Original GLIP. * indicates multi-scale evaluation. Numbers in red clearly points out the difference between the prompt tuning and full fine-tuning results (see Table 1). Numbers in gray mean that they are not in *zero-shot* manner. †: these two numbers are artificially high due to some overlap between COCO-minival and VisualGenome-train.

"Object Detection in the Wild"

GLIPv2: Unifying Localization and Vision-Language Understanding

NEURAL INFORMATION PROCESSING SYSTEMS

- Explainable grounded VQA & Image Captioning

Input: Where is the strainer? [MASK]
Prediction: counter
Gold: counter

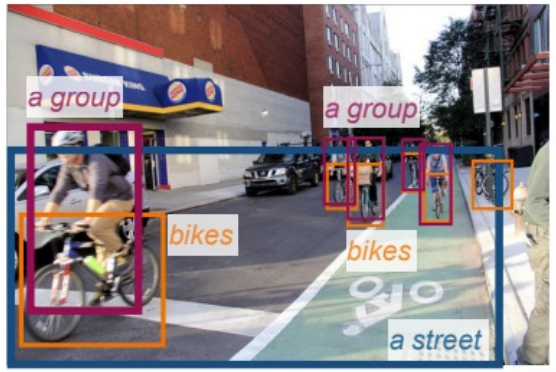
Input: What is the man wearing? [MASK]
Prediction: jacket
Gold: ski suit

Input: Where is a push vacuum? [MASK]
Prediction: on floor
Gold: background

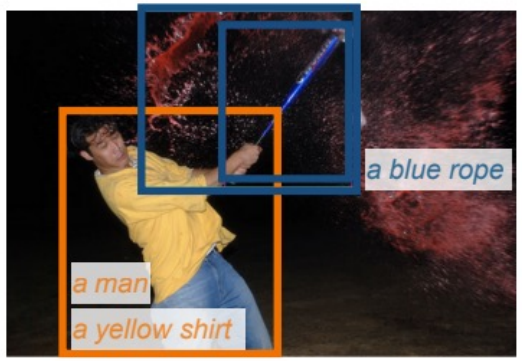
Visual Question Answering (VQA)

Image Captioning

Generated Caption: a man riding a motorcycle on a dirt road.



Generated Caption: a group of people riding bikes down a street.

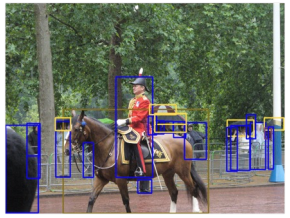


Generated Caption: a man in a yellow shirt is holding a blue rope.

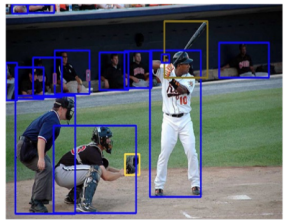
GLIPv2: Unifying Localization and Vision-Language Understanding

NEURAL INFORMATION PROCESSING SYSTEMS

COCO



Prompt: person. dog ... backpack. umbrella. horse. toothbrush.



Prompt: person. hairdryer ... baseball bat. baseball glove. bottle. toothbrush.

Prompt: person. cup. sink ... microwave. refrigerator. bear.

Prompt: person. chair. dining table ... potted plant. vase.

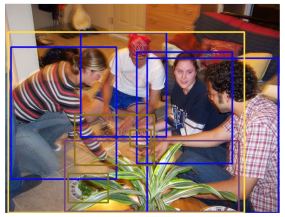
Prompt: person. hairdryer ... baseball bat. baseball glove. bottle. toothbrush.

Prompt: person. cup. sink ... microwave. refrigerator. bear.

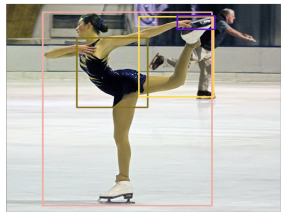
COCO-Mask

Flick30k

Prompt: Mounted officers in bright green jackets sit on their horses wearing helmets.



Prompt: 2 couples are eating dinner on the floor behind a large plant.



Prompt: A woman figure skater in a blue costume holds her leg by the blade of her skate

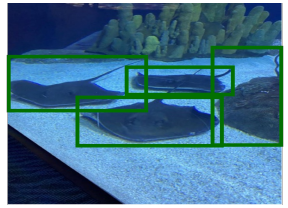
Prompt: tissue. jacket. ... fork. pineapple. dinning table.

Prompt: donut. wineglass ... banana. pineapple.

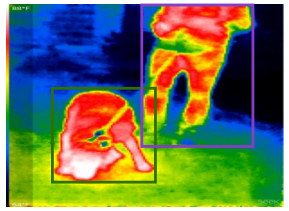
Prompt: person. teddy bear ... lollipop. flower.

LVIS

ODinW



Prompt: fish. jellyfish. penguin. puffin. shark. starfish. stingray

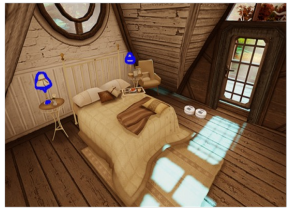


Prompt: dog. person.

Prompt: smoke.

Prompt: green bush

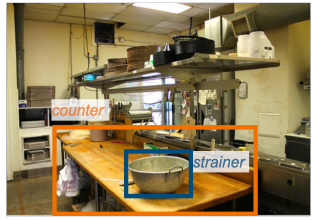
Prompt: window has a frame



Prompt: brown lampshade

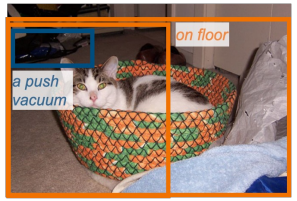
PhraseCut

VQA

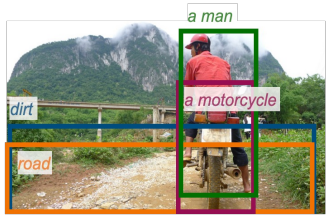


Input: Where is the strainer? [MASK]
Prediction: counter
Gold: counter

Input: What is the man wearing? [MASK]
Prediction: jacket
Gold: ski suit

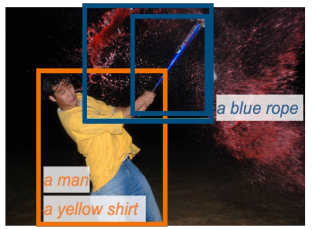


Input: Where is a push vacuum? [MASK]
Prediction: on floor
Gold: background



Generated Caption: a man riding a motorcycle on a dirt road.

Generated Caption: a group of people riding bikes down a street.



Generated Caption: a man in a yellow shirt is holding a blue rope.

COCO-Caption

Follow Ups

- 2nd 'Computer Vision in the Wild' Workshop @ CVPR 2023 (in Preparing)

https://computer-vision-in-the-wild.github.io/eccv-2022/

- For more details about our paper, please refer to the following links:

GLIPv2 Paper

GLIPv2 Code

Hugging Face Demo


NEURAL INFORMATION
PROCESSING SYSTEMS
Q&A
Thank you!