Language Understandings in the Real-world

RIKEN AIP
Shuhei Kurita

IJCNLP-AACL 2023 Tutorial @ Bali

Content

1. **Connecting Images and Texts**
    1. IC, VQA, SceneGraph, DenceCaptioning and beyond
    2. Referring Expression Comprehension (REC)
    3. Contrastive Learning for Images and Texts
    4. Open Vocaburaly Object Detection (OVOD) and REC
    5. REC on First-Person Vision: RefEgo
2. **Language Understandings in Navigation**

Connecting languages and vision

- There are two classical tasks:
  - Image captioning (IC)
  - visual question answering (VQA)

Captions:

- man opening the faucets to a fire hydrant letting water out onto the lawn
- a city worker turning on the water at a fire hydrant.
- worker opening valve of fire hydrant in residential neighborhood.
- a man adjusting the water flow of a fire hydrant.
- the official in a yellow vest is using a wrench on a fire hydrant.

Question: What color is the hydrant?

red

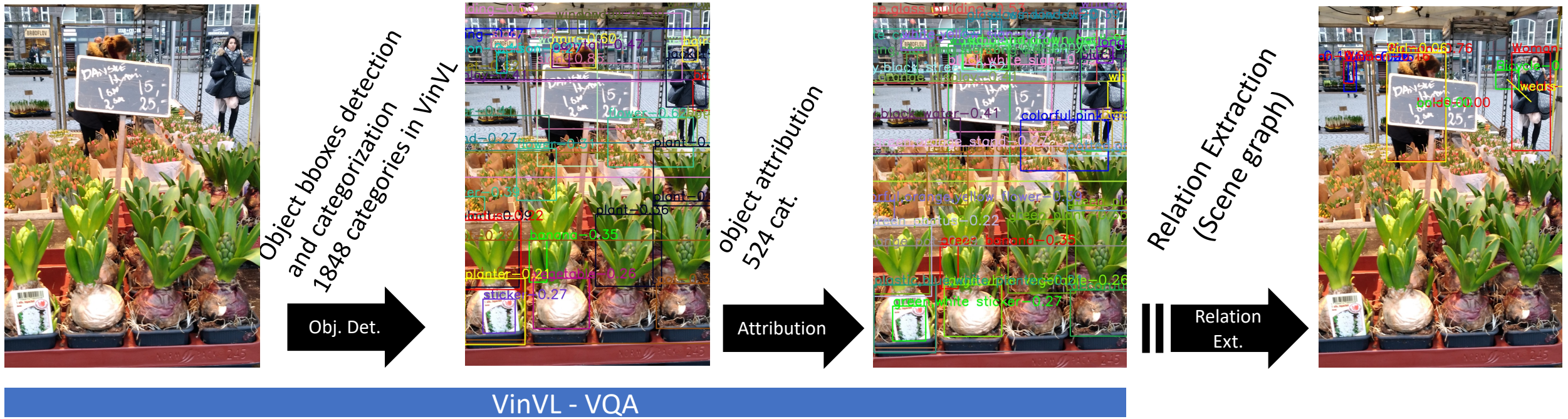
black and yellow

- IC and VQA are two major tasks that are often targeted also in modern multi-modal language models
- Can we deeply connect languages and images beyond the ”shallow” understandings of these tasks?

Connecting Images and Texts *in detail…*

Wait, what is the meaning of *connecting images and objects*? **(Open Question!)**

Example: Object detection, Attribution, Relation Extraction from an image (with VinVL)

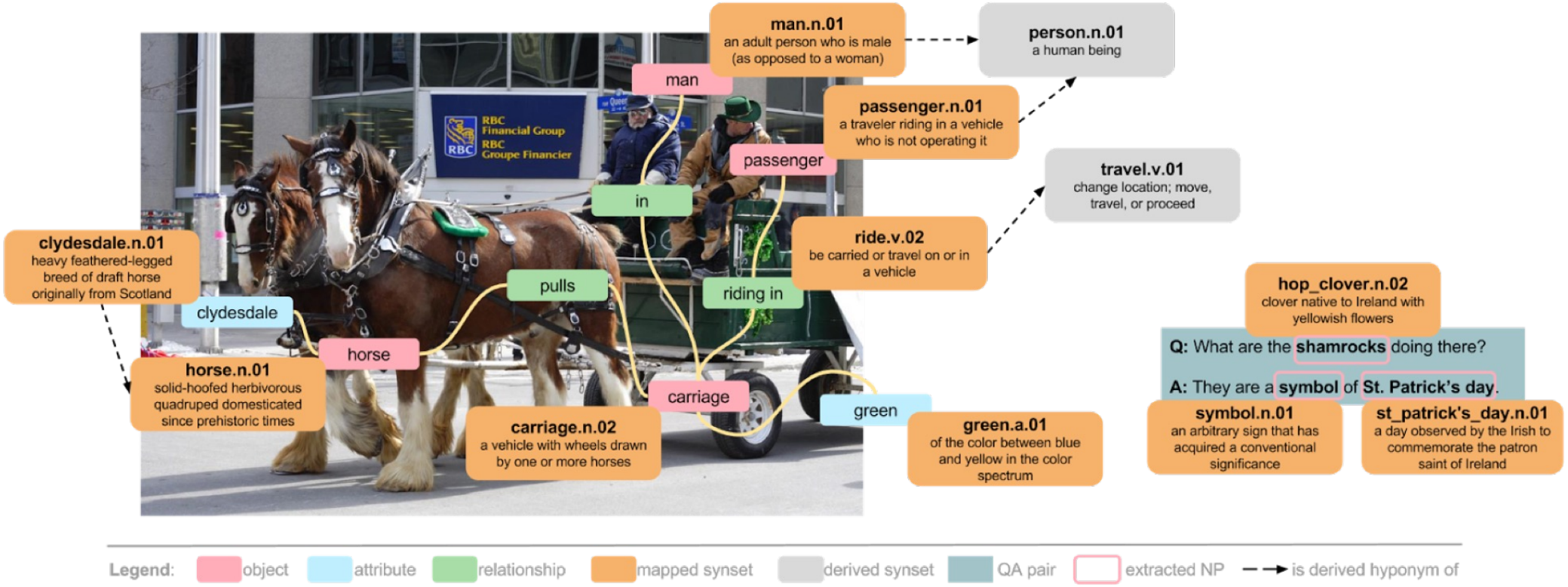


Similar to the tranditional NLP pipelines of word segmentation, POS tagging and parsing, we can extract detailed information respectively…

VinVL: Making Visual Representations Matter in Vision-Language Models, Zhang *et al*. (CVPR2021).

Scene Graph Generation on Visual Genome

- Visual Genome [Krishna 2016]

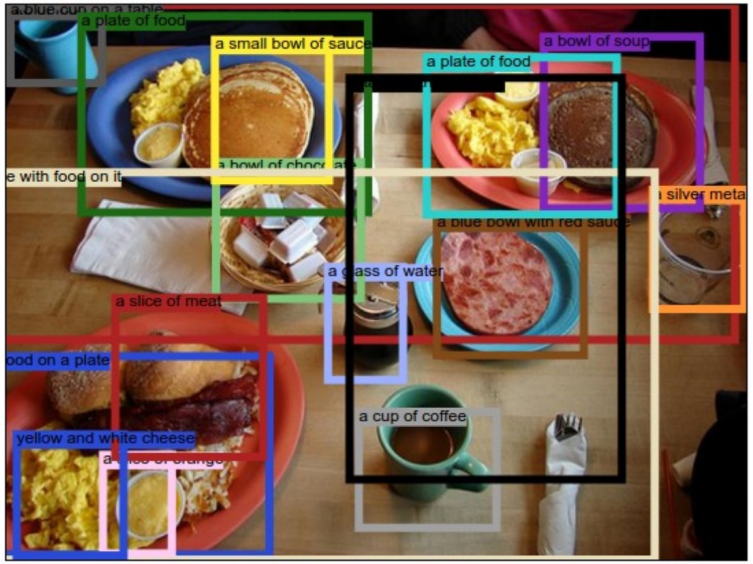


- Attribute and Relation extractions are familiar tasks also in NLP.
- However, SceneGraph Generation is a super difficult task so far.
- *First of all, we really need to extract all object attributes and relations in images*?
  - (cf. well-know *frame-problem)*

Dense Captioning

- Dense Captioning
  - Describing objects in images in detailed texts
  - A derivation from the Visual Genome Dataset

bus parked on the street. a city street scene. front windshield of a bus. man walking on sidewalk. a silver car parked on the street. a city scene. a green traffic light. a building in the background. the bus has a number. a large building. a brick building. red brick building with windows. a blue sign with a white arrow. white lines on the road.



a plate of food. food on a plate. a blue cup on a table. a plate of food. a blue bowl with red sauce. a bowl of soup. a cup of coffee. a bowl of chocolate. a glass of water. a plate of food. a silver metal container. a small bowl of sauce. table with food on it. a slice of orange. a table with food on it. a slice of meat. yellow and white cheese.

https://cs.stanford.edu/people/karpathy/densecap/

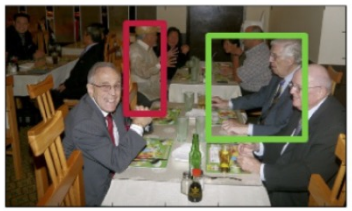
Referring expression comprehension (Visual Grounding)

- Searching and predicting bounding boxes of objects referred in texts from images

During **a gay pride parade** in **an Asian city**, **some people** hold up **rainbow flags** to show their **support**.
A group of youths march down **a street** waving **flags** showing **a color spectrum**.
Oriental people with **rainbow flags** walking down **a city street**.
A group of people walk down **a street** waving **rainbow flags**.
People are **outside** waving **flags** .

Flickr 30k entities

RefCOCO testA

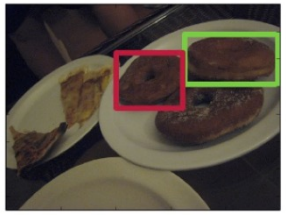


second back old guy on right

old man

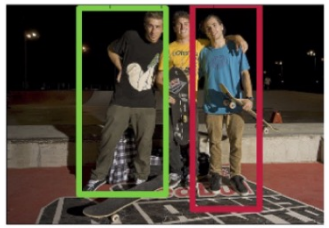
RefCOCO testB

white cup at top center left



donut on top of the other middle

RefCOCO+ testA



man not holding a skateboard

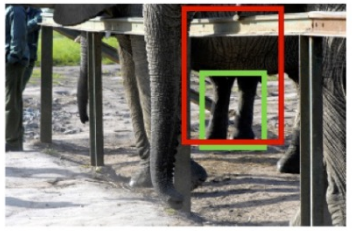
girl looking at pizza

RefCOCO+ testB

screen with no people

two pieces of broccoli forming a reverse left

RefCOCOg validation



furthest away animal legs

the wine glass in the middle right

a short ski leaning up against the table

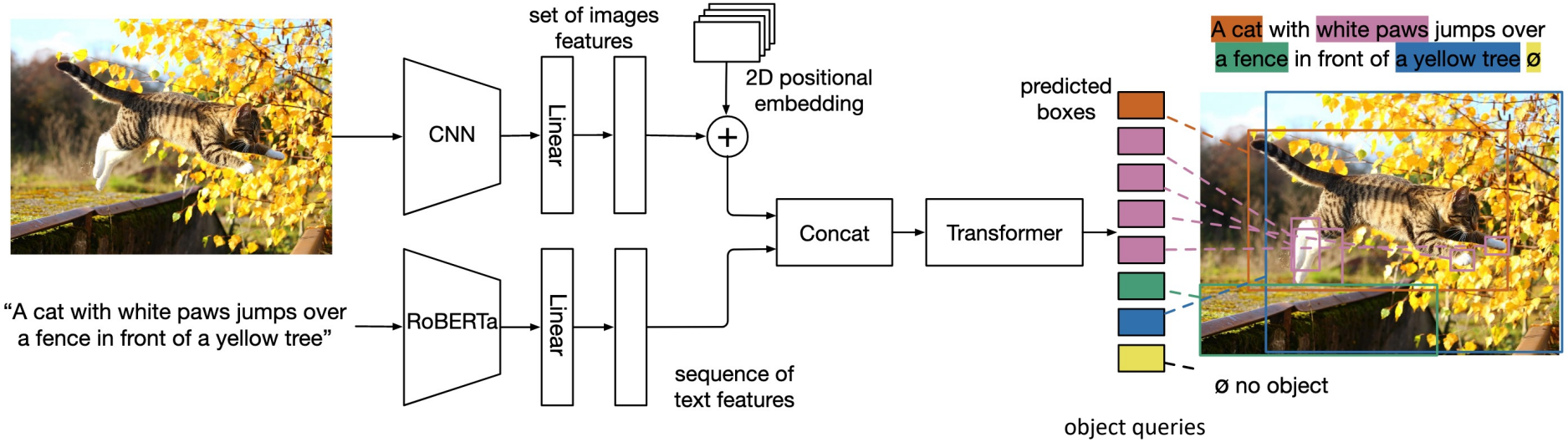
a woman in a striped sweater talking on a phone

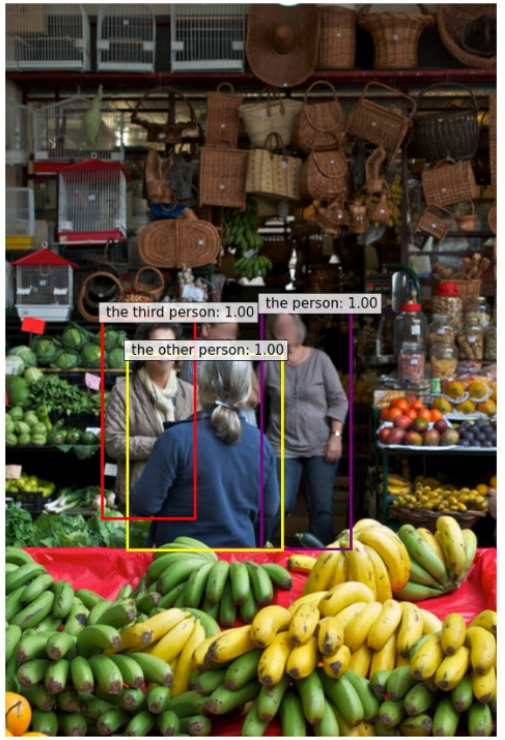
RefCOCO / RefCOCO+ / RefCOCOg

Green: Annotated, Red: Pred. Err.

MDETR

- Referring expression comprehension, VQA, etc.
- Introducing “object queries” that are learnable vectors and paired to both bounding boxes and textual spans in pretraining
- Pretrain is done in MS COCO, Visual Genome, Flicker30k
  - DETR loss (L1 & GIoU of bboxes)
  - Soft token prediction loss (for text spans)
  - Contractive alignment loss (InfoNCE-based object-text matching loss)

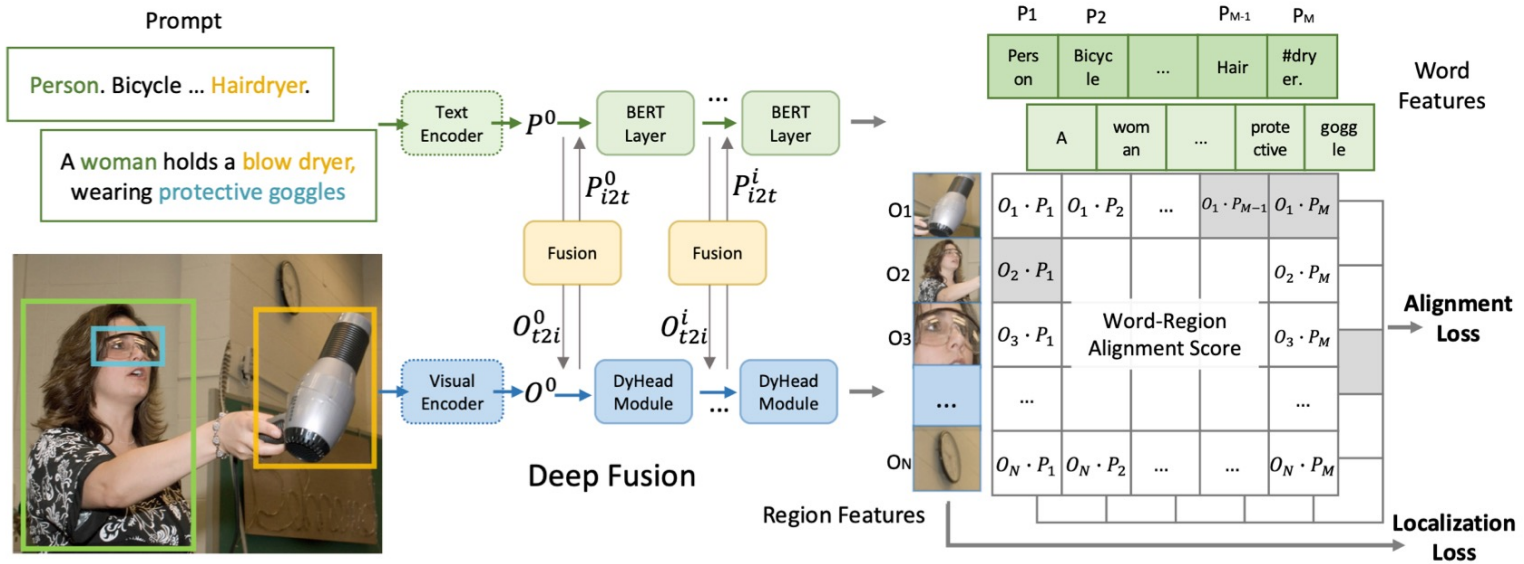


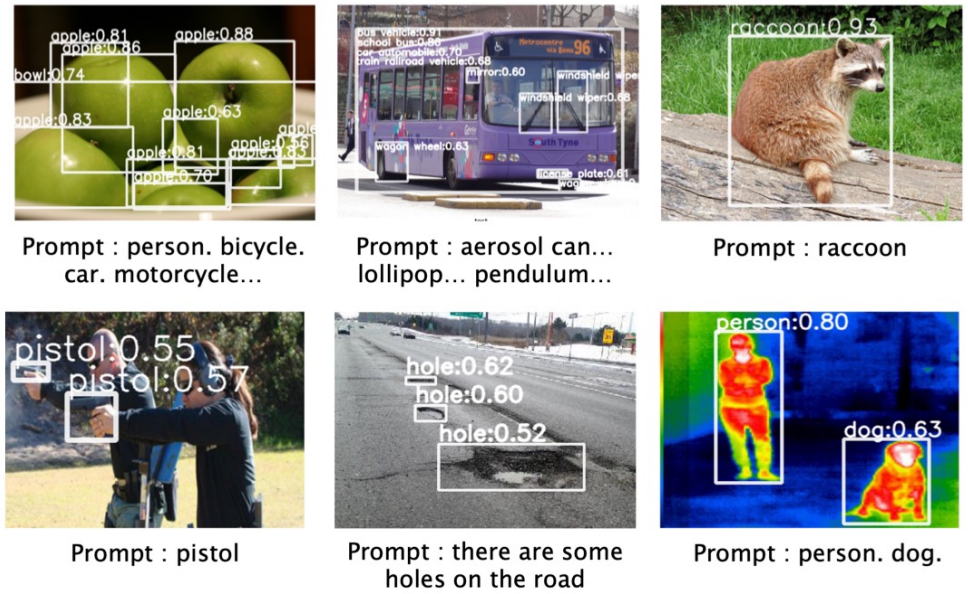


MDETR -- Modulated Detection for End-to-End Multi-Modal Understanding, Kamath *et al.*, (ICCV2021).

Improvement of MDETR

- GLIPv2 (Zhang, NeurIPS2022)
  - Deep fusion
  - Contrastive learning inter and intra-images





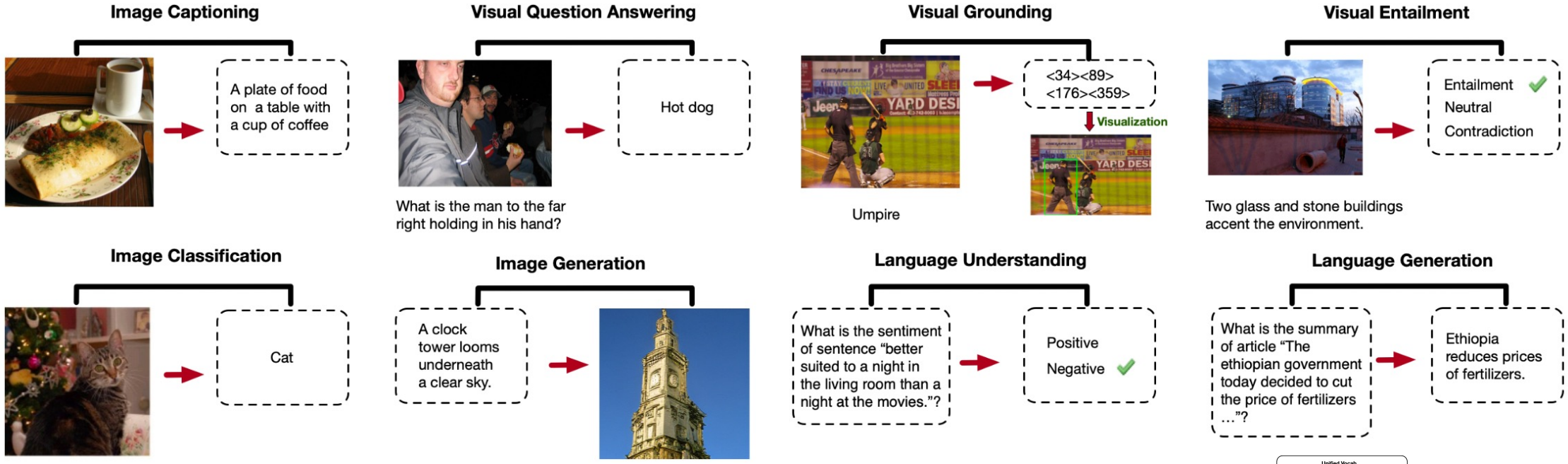
➢Contrastive learning of intra-images (similar to MDETR).

➢Contrastive learning of inter-images (similar to CLIP).

GLIPv2: Unifying Localization and Vision-Language Understanding, Zhang *et al*. (NeurIPS2023).

REC with Language Model

- OFA [Wang 2022]



- Joint multi-task multi-dataset in one language modeling.

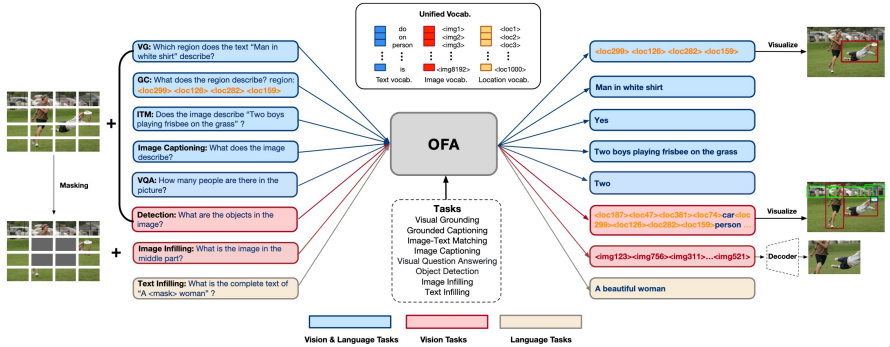


Figure 2: A demonstration of the pretraining tasks, including visual grounding, grounded captioning, image-text matching, image captioning, VQA, object detection, image infilling as well as text infilling.

OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework, Wang *et al*. (ICML2022).

REC with Language Model

- OFA [Wang 2022]

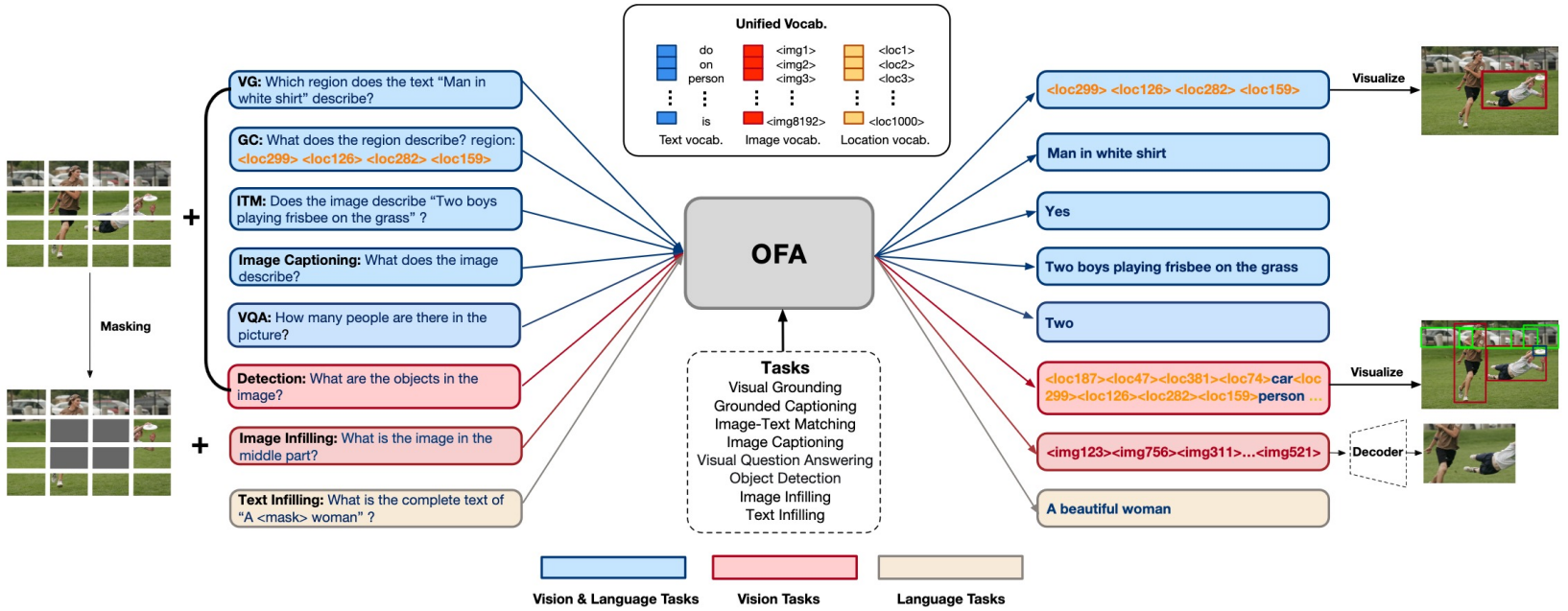
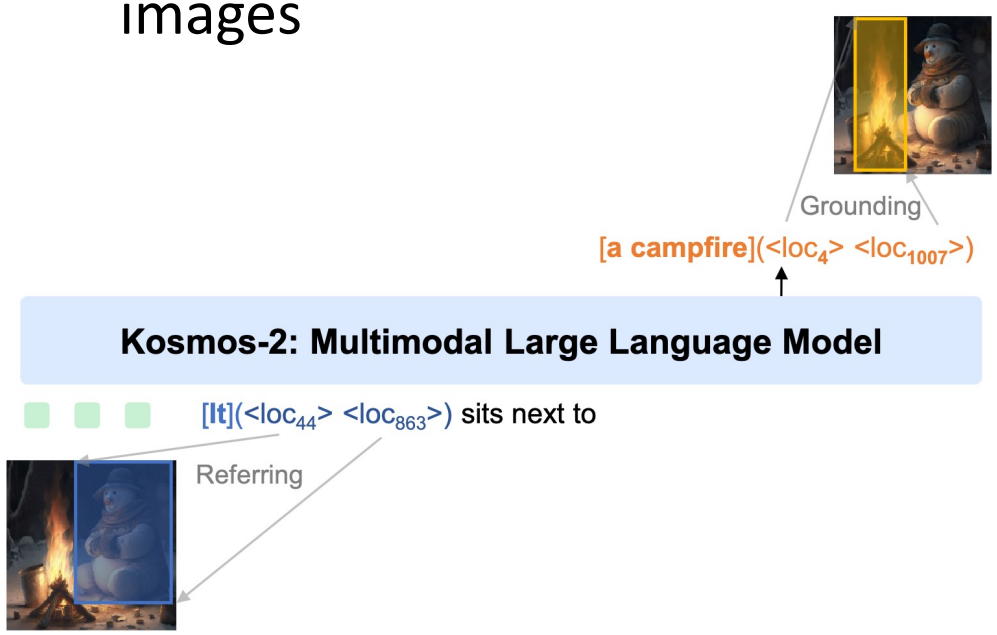


Figure 2: A demonstration of the pretraining tasks, including visual grounding, grounded captioning, image-text matching, image captioning, VQA, object detection, image infilling as well as text infilling.

- In REC or object detection, OFA predicts bounding box of (x1 y1 x2 y2) points as a language model outputs

REC with Language Model

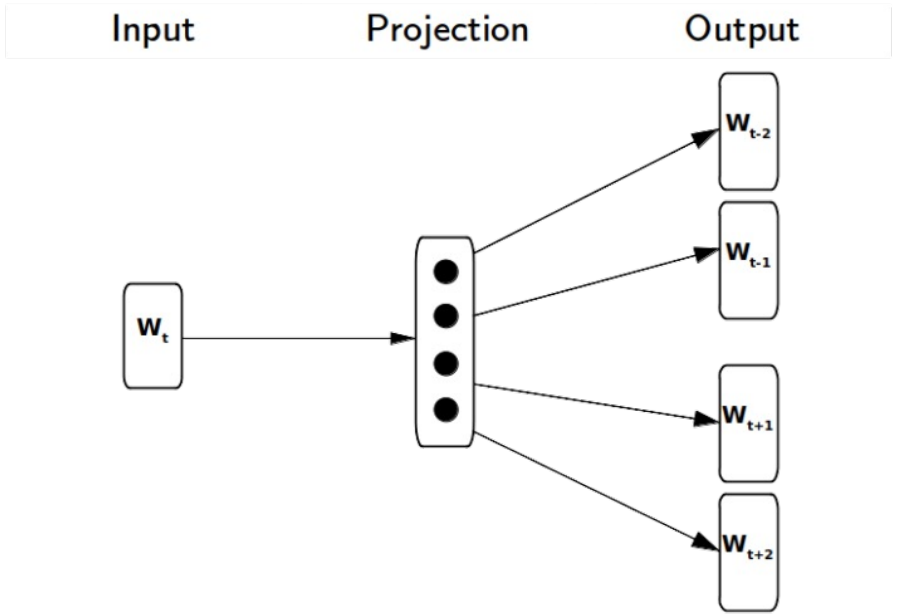
- Microsoft KOSMOS-2
  - Inserting references to image regions (bounding boxes) in input/output of language models
  - Trained with GRIT dataset of 90M images



Kosmos-2: Grounding Multimodal Large Language Models to the World, Peng *et al*. (2023). arXiv.

Contrastive learning of Images and Texts

Contrastive learning

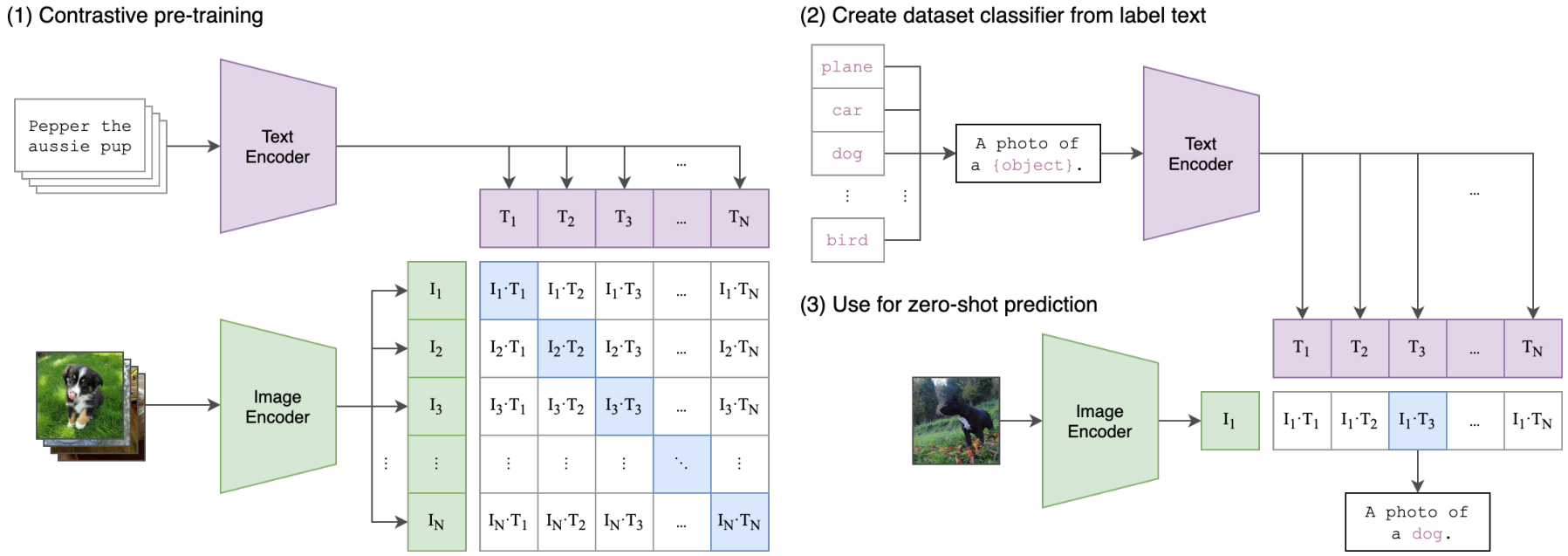


Skip-gram learning in word2vec [Mikolov 2013]

SimCLR [Chen 2020]

Contrastive learning of Images and Texts

- CLIP [Radford, 2021]



- (1) In pretraining, align captions and images in the batch as contrastive learning
- (2) In inference, with extraporated text labels, models can predict zero-shot object detection (classification) for the new text label classes.
- No use of existing dense annotations such as MS-COCO or VisualGenomein pretraining.
  - Suitable for matching the entire image and text
  - Probably not suitable for matching parts of images and parts of texts
- Contrastive learned models of V&L perform like ”bag-of-words” text understanings (Yuksekgonul, ICLR2023).

Object Detection and Object Class Labels

- In conventional object detections, the number of object categories are fixed parameters due to the training dataset:

Dataset	# images	# boxes	# categories
Pascal VOC	11.5 k	27 k	20
MS COCO	159 k	896 k	80
Objects 365	1,800 k	29,000 k	365
LVIS v1.0	159 k	1,514 k	1,203

➢ When we'd like to pair image objects to texts, fixed-number object categories are not suitable?

REC and Open Vocaburaly Object Detection

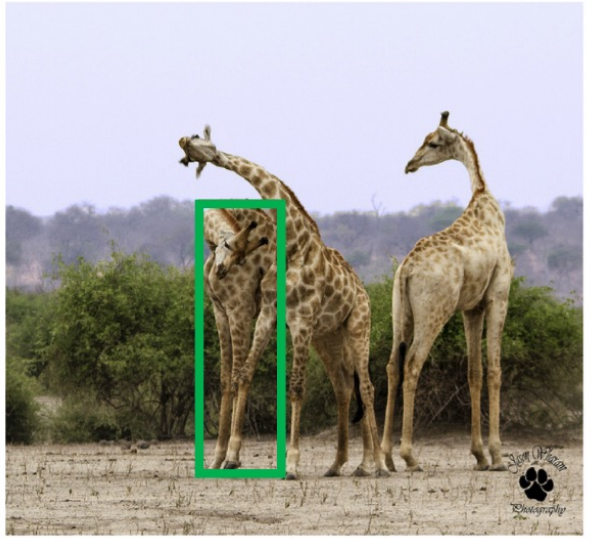
- **Referring Expression Comprehension**
  - *Detection of objects that are referred by some texts from images. Also simply called as visual grounding.*
  - Go back to around 2014 (ReferItGame, EMNLP2014).
  - Trained and evaluated in human-annotated sets of RefCOCO, RefCOCO+, RefCOCOg.
  - Good at longer texts (Especially in case of the RefCOCOg dataset)
    "A car outside to the right of the red box.", "a horse being ridden by number 6 jockey"
  - Representative papers
    - MDETR -- Modulated Detection for End-to-End Multi-Modal Understanding, Kamath *et al*., ICCV2021.
    - OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework, Wang *et al*. ICML2022.
- **Open-Vocabulary Object Detection**
  - *Detecting objects of new object classes that are not seen in training from the given textual labels in the inference. Some kind of out-of-domain object-class object detection.*
  - Becoming a major topic from around 2021.
  - Matching the detection results (image regions) with textual phrases with external pretrained models such as CLIP.
  - Good at short phrases, attributes of object synonyms.
    "red car", "white monitor", "black display"
  - Representative papers
    - Open-vocabulary Object Detection via Vision and Language Knowledge Distillation (ViLD), Gu *et al*., ICLR2022.
    - Detecting Twenty-thousand Classes using Image-level Supervision (Detic), Zhou *et al*.,, ECCV2022.

REC and Open Vocaburaly Object Detection

- **Referring Expression Comprehension**
  - *Detection of objects that are referred by some texts from images. Also simply called as visual grounding.*
  - Good at longer texts (Especially in case of the RefCOCOg dataset)

    "A car outside to the right of the red box.", "a horse being ridden by number 6 jockey"
- **Open-Vocabulary Object Detection**
  - *Detecting objects of new object classes that are not seen in training from the given textual labels in the inference. Some kind of out-of-domain object-class object detection.*
  - *Matching the detection results (image regions) with textual phrases with external pretrained models such as CLIP.*

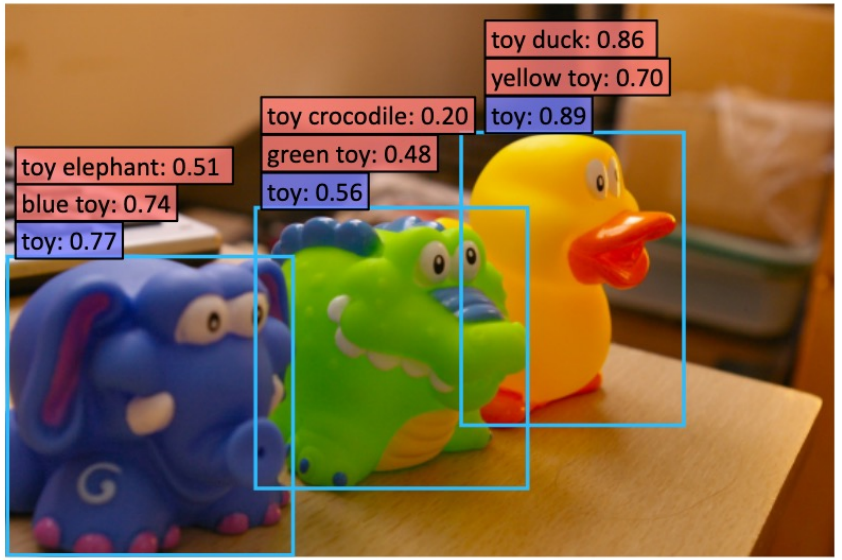
Although they look similar, they are indeed different tasks so far?



RefCOCO:
1. giraffe on left
2. first giraffe on left

RefCOCO+:
1. giraffe with lowered head
2. giraffe head down

RefCOCOg:
1. an adult giraffe scratching its back with its horn
2. giraffe hugging another giraffe

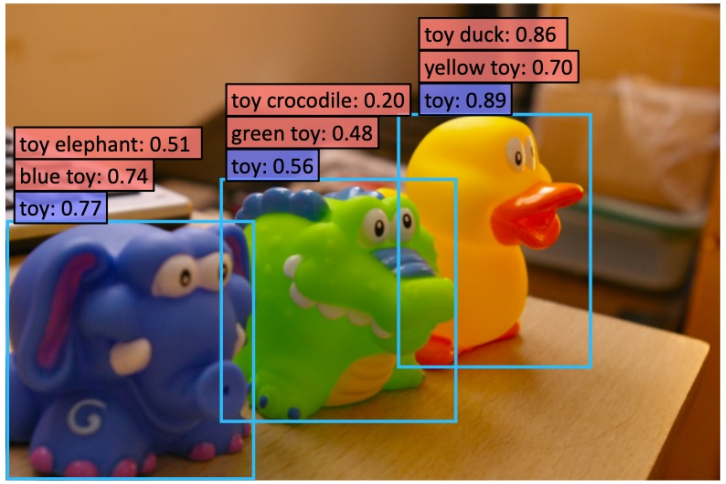


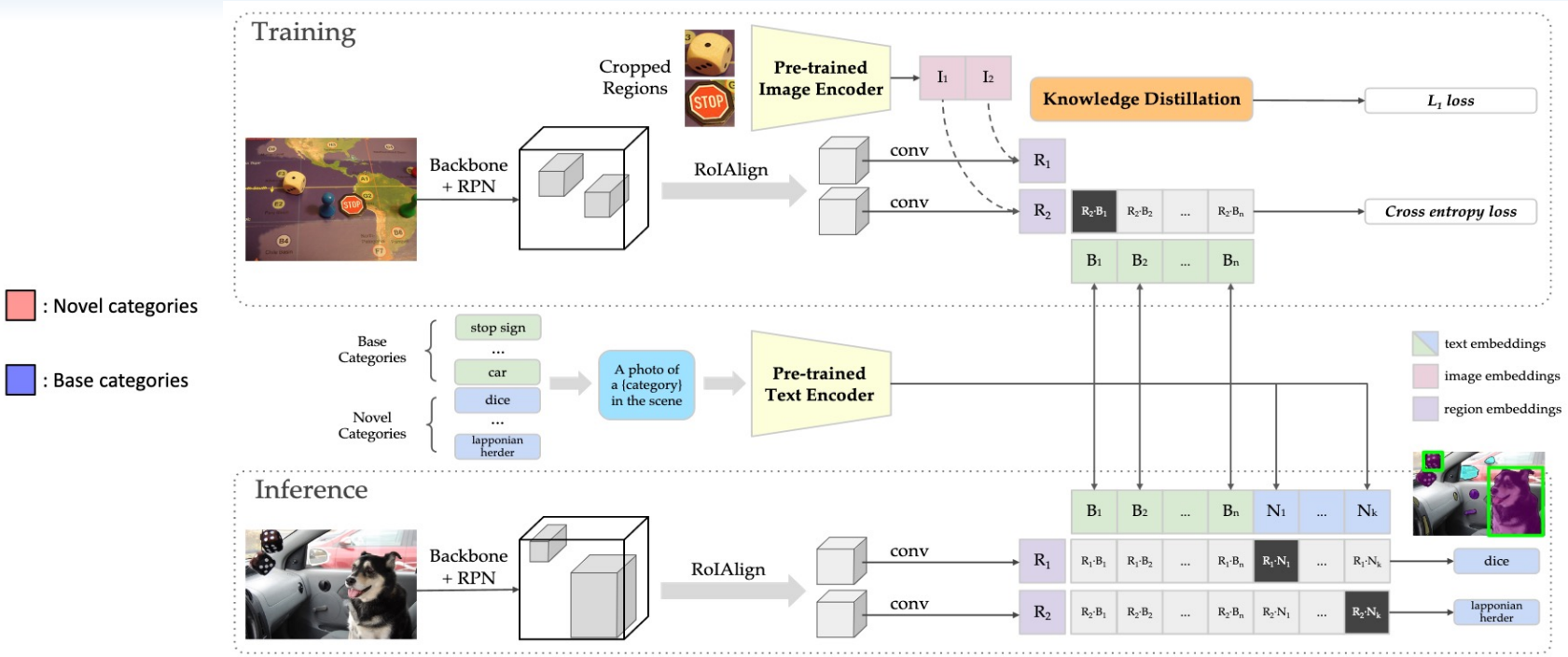
: Novel categories

: Base categories

(ViLDより)

ViLD





- Open Vocabulary Object Detection: Detecting objects of new object classes that are not seen in training from the given textual labels in the inference.
  - Open Vocabulary Object Detection is differect from zero-shot object detection (ZOD).
  - Classifying image regions (bounding boxes) into a new unseen classes that are obtained through CLIP.
- Good for simple phrase and paraphrasing, such as “toy”/ ”toy duck” or “display”/ ”monitor”.

OVOD and Segmentation

Detic

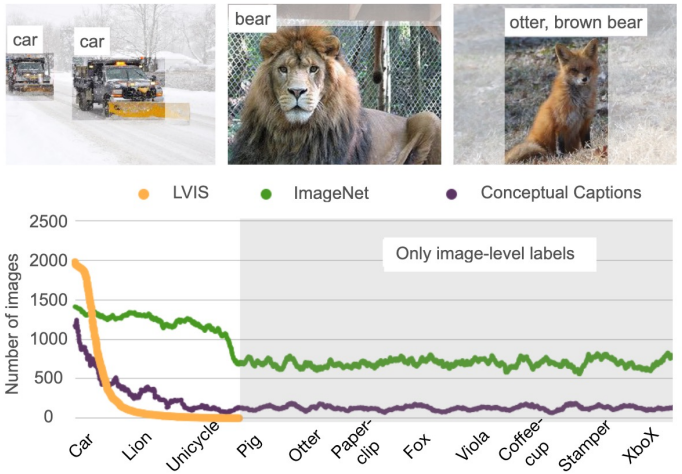
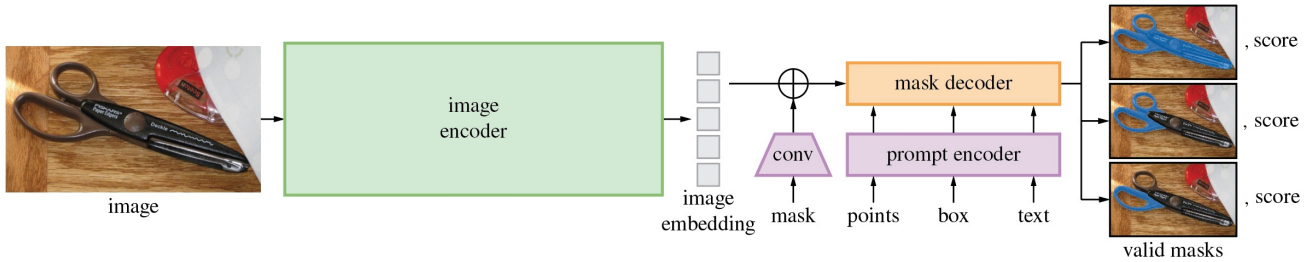


Fig. 1: Top: Typical detection results from a strong open-vocabulary LVIS detector. The detector misses objects of "common" classes. **Bottom:** Number of images in LVIS, ImageNet, and Conceptual Captions per class (smoothed by averaging 100 neighboring classes). Classification datasets have a much larger vocabulary than detection datasets.

Segment Anything



(Meta AI.)

Grounding DINO

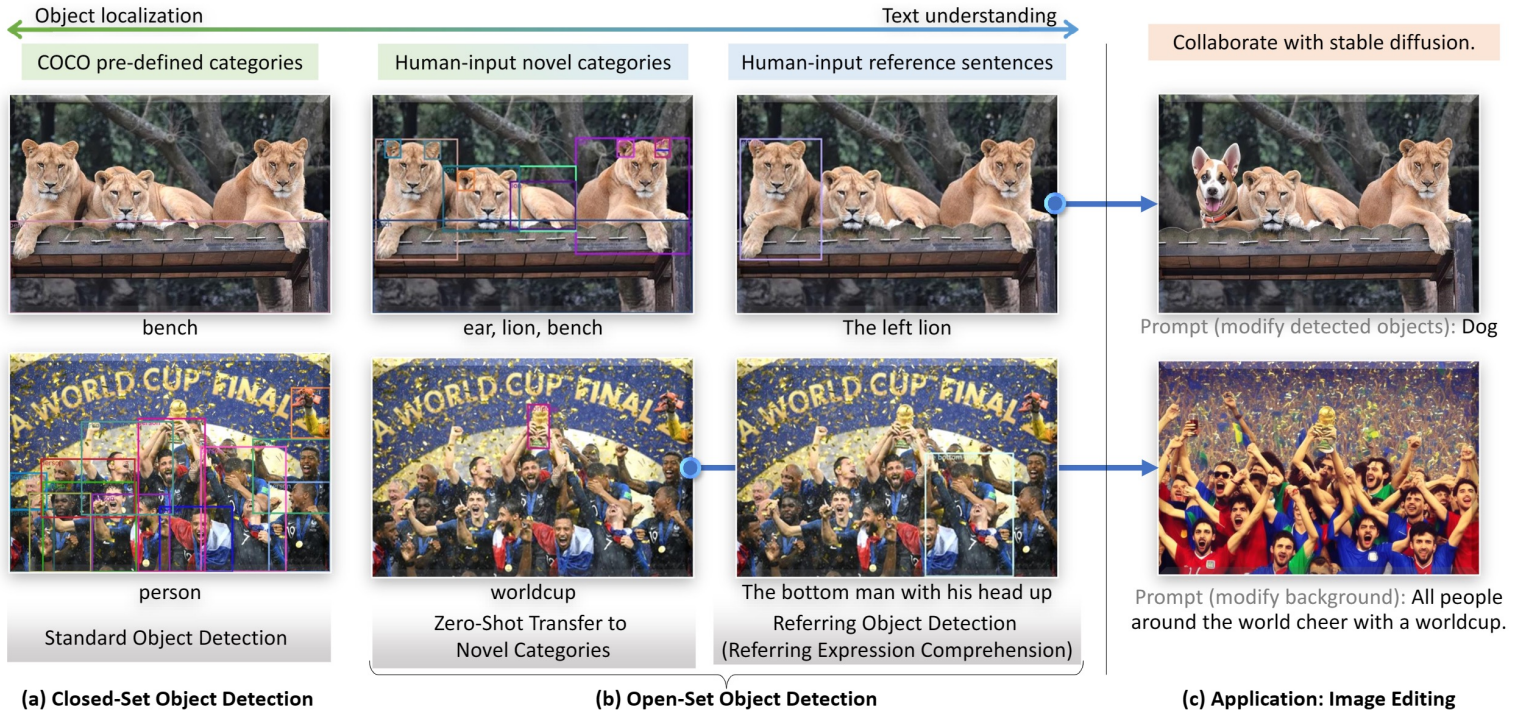
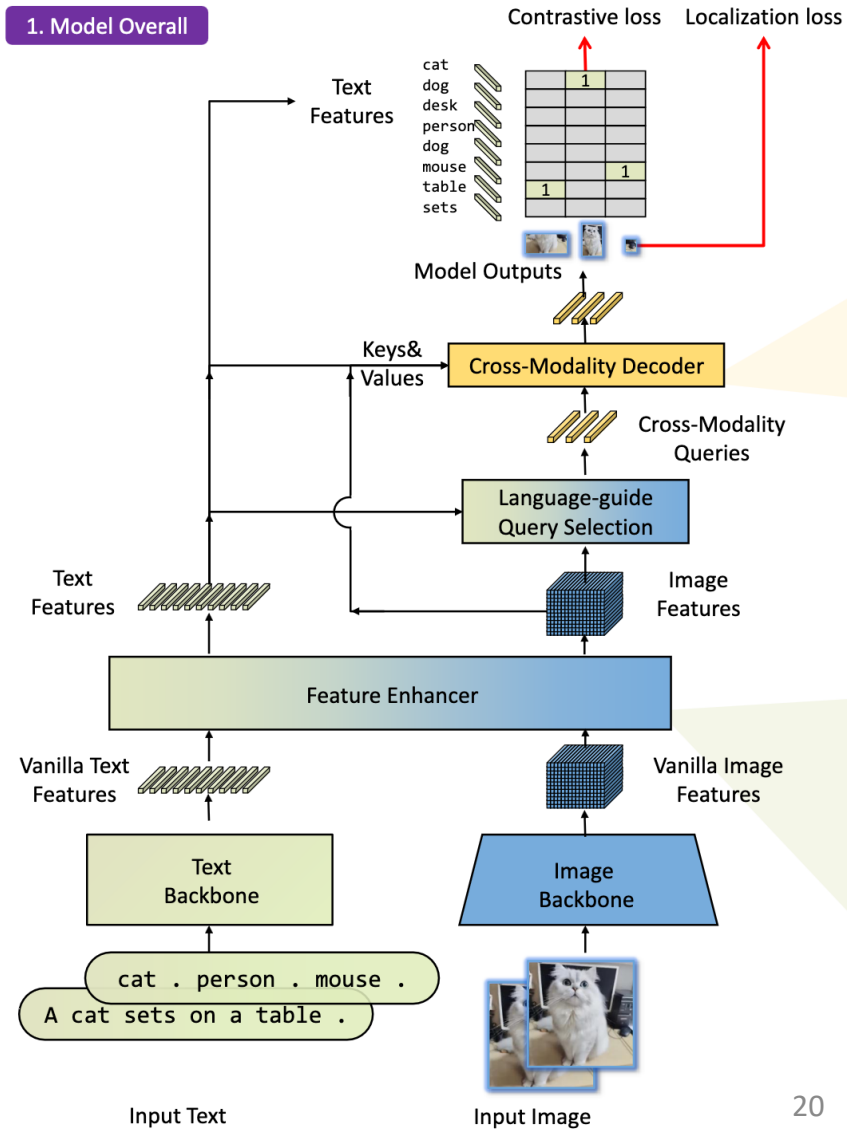


Figure 1. (a) Closed-set object detection requires models to detect objects of pre-defined categories. (b) Previous work zero-shot transfer models to novel categories for model generalization. We propose to add Referring expression comprehension (REC) as another evaluation for model generalizations on novel objects with attributes. (c) We present an image editing application by combining Grounding DINO and Stable Diffusion [42]. Best view in colors.

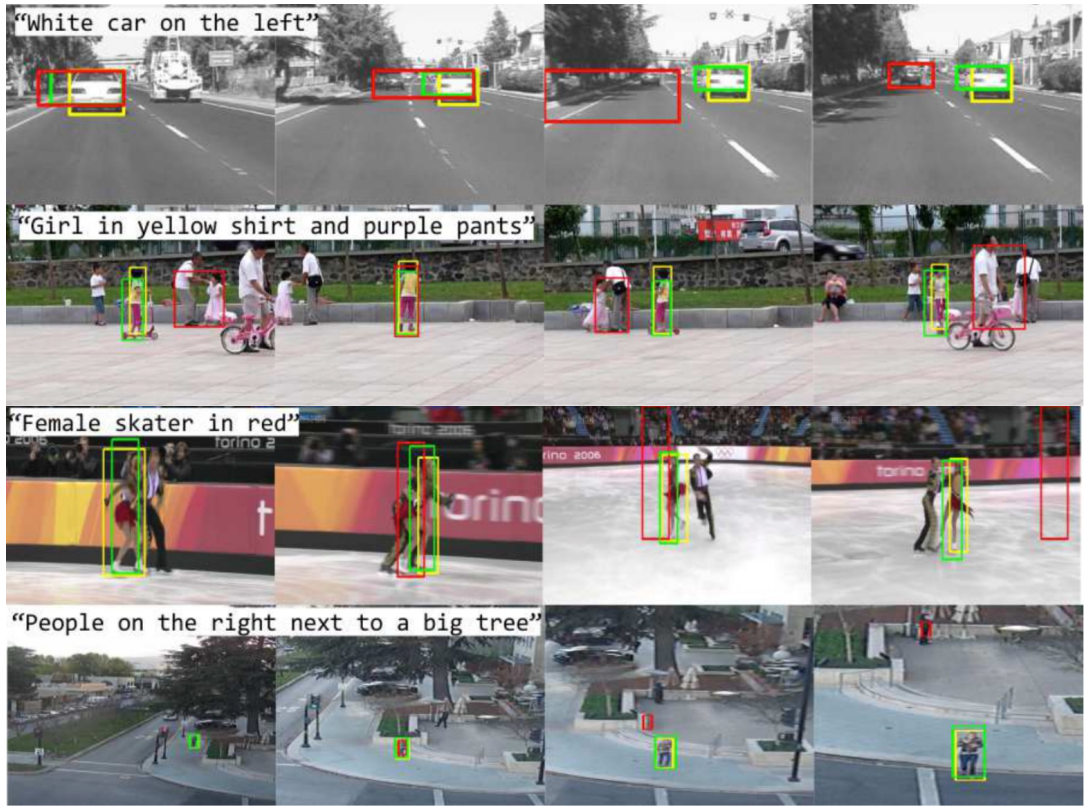


- Beyond GLIP performance in **OVOD**.
  - However, OFA still out-performs gounding DINO in **REC**?

Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection, Liu *et al*., 2023. arXiv.

Referring expression comprehension on videos

- Searching and predicting bounding boxes of objects referred in texts from videos…
- Unfortunately, existing videos are in limited domains.
- Especially no first-person vision referring expression comprehension dataset although there are so many realistic applications.



Lingual OTB99/ImageNet Videos dataset

A woman with a stroller.

A girl riding a horse.

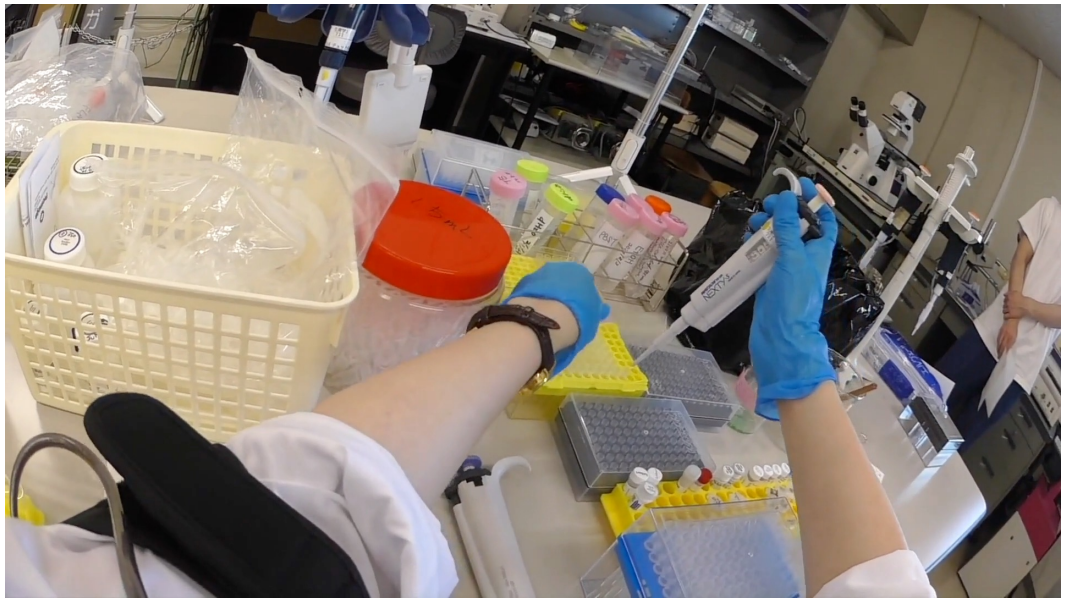
ReferDAVIS

A person on the right dressed in blue black walking while holding a white bottle.

A person showing his skate board skills on the road.

Refer-YouTube-VOS

First Person Vision vs Third Person Vision



(Ego4D)

(The VIRAT Video Dataset)

First Person Vision

- wearable camera
- car & robotic view

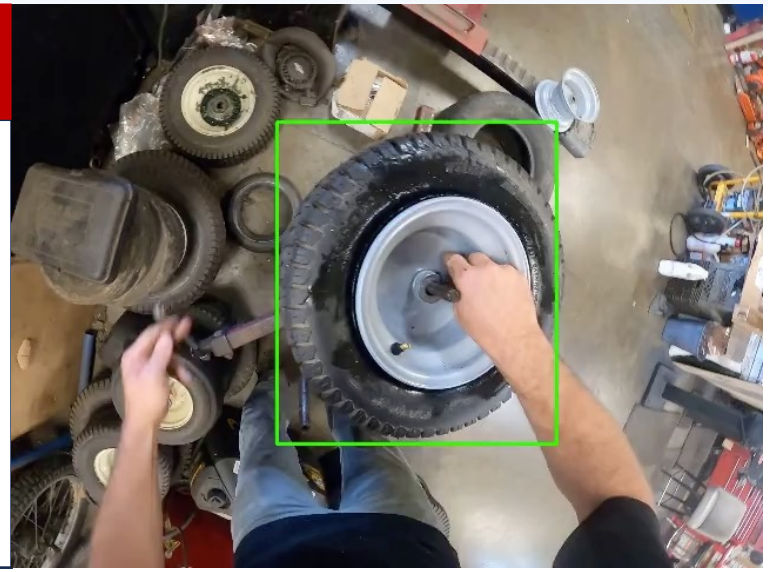
Third Person Vision

- Surveillance camera
- Fixed camera

Localizing objects in FPV of EGO 4D from *Texts*

Garage

A large tire with a gray rim in the hands of the person.



Kitchen

A small blue plate of broccoli to left of other plate.

Supermarket

A red crate on the flat shopping cart in the middle of the isle.

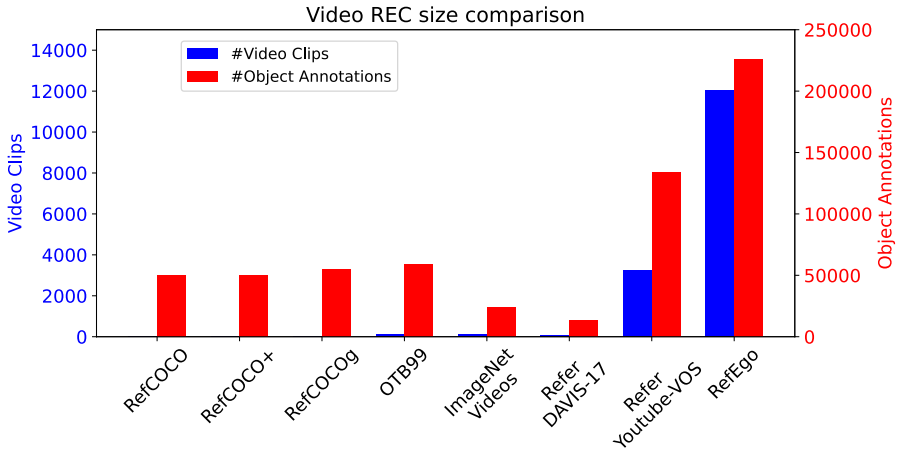
Lab

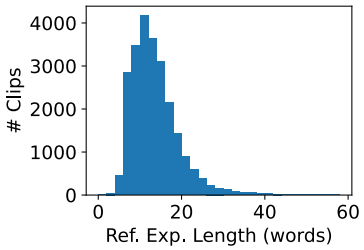
The red container near the wall, behind the two trays.

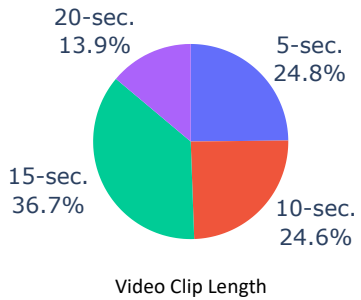
Localizing objects in FPV of EGO 4D from *Texts*

- We constructed a object localization & tracking tdataset on ***Ego4D***
  for referring expession comprehension (REC)
- 12,038 annotated clips of 41 hours total.
- 2FPS for annotation bboxes with
  two textual referring expressions for a single object.
- Objects can be out-of-frame (*no-referred-objects*).

Split	# Clips	# Images	# Images with BBox
Train	9,172	225,500	173,183
Val.	1,549	38,470	29,322
Test	1,317	31,560	23,814

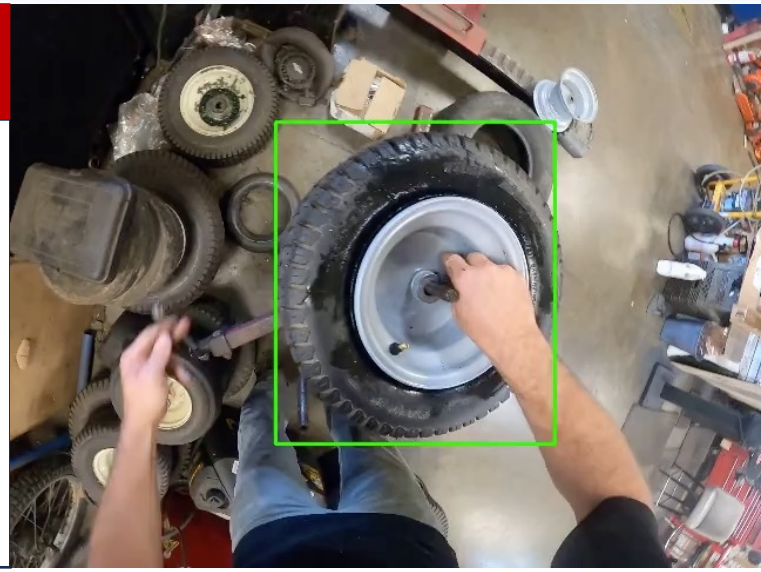






Garage

A large tire with a gray rim in the hands of the person.



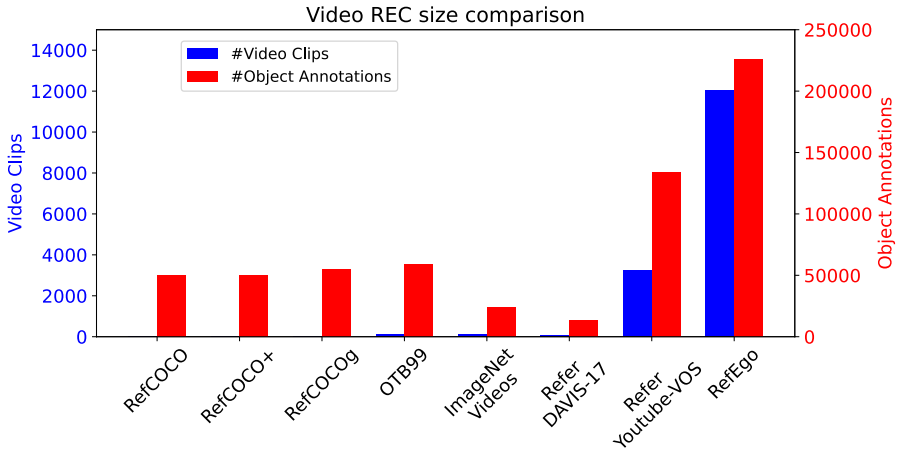
Supermarket

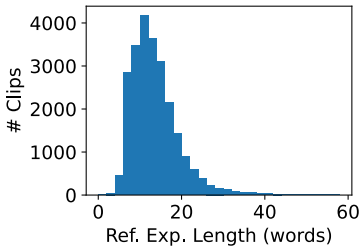
A red crate on the flat shopping cart in the middle of the isle.

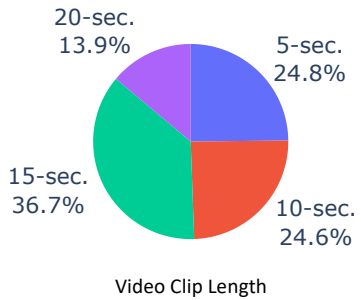
Localizing objects in FPV of EGO4D from *Texts*

- We constructed a object localization & tracking tdataset on ***Ego4D***
  for referring expession comprehension (REC)
- 12,038 annotated clips of 41 hours total.
- 2FPS for annotation bboxes with
  two textual referring expressions for a single object.
- Objects can be out-of-frame (*no-referred-objects*).

Split	# Clips	# Images	# Images with BBox
Train	9,172	225,500	173,183
Val.	1,549	38,470	29,322
Test	1,317	31,560	23,814







Kitchen

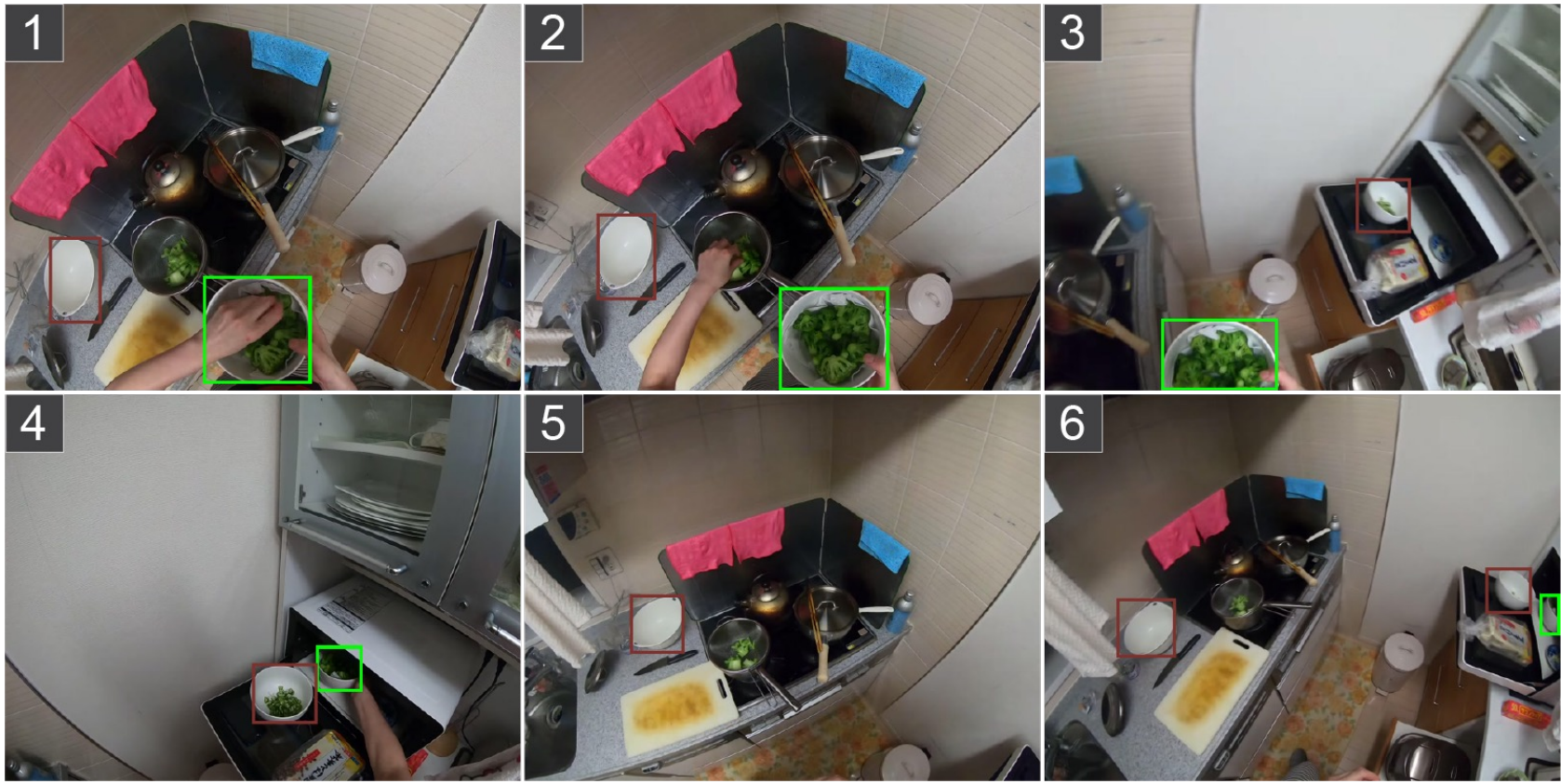
A small blue plate of broccoli to left of other plate.

Lab

The red container near the wall, behind the two trays.

REC in First-Person Vision of Ego4D

- Is the referred object always in the images in real-world FPV?
  - REC assumes that the referred object is always in the image, while RefEgo presumably includes images where the referred object is out of frame



RefEgo: Referring Expression Comprehension Dataset from First-Person Perception of Ego4D, Kurita, Katsura and Onami (ICCV2023).

REC in First-Person Vision of EGO4D

- **RefEgo model can work the detecter whether the referred objects are in the images or not.**
  - **This is a difficult problem in existing REC and other models**

MDETR (RefCOCOg) off-the-shelf

MDETR (RefEgo) w/ Binary Head

False positive detections!

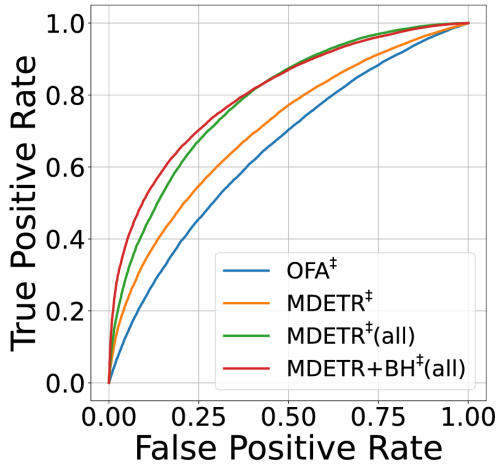
A red crate on the flat shopping cart in the middle of the isle.

MDETR (RefCOCOg) off-the-shelf

MDETR (RefEgo) w/ Binary Head

False positive & inconsistent detections!

The green and white bicycle on top.



ROC curve

Model	AUC
(Chance rate)	50.0
OFA‡	64.9
MDETR‡	70.7
MDETR‡ (all)	78.8
MDETR+BH‡ (all)	80.5

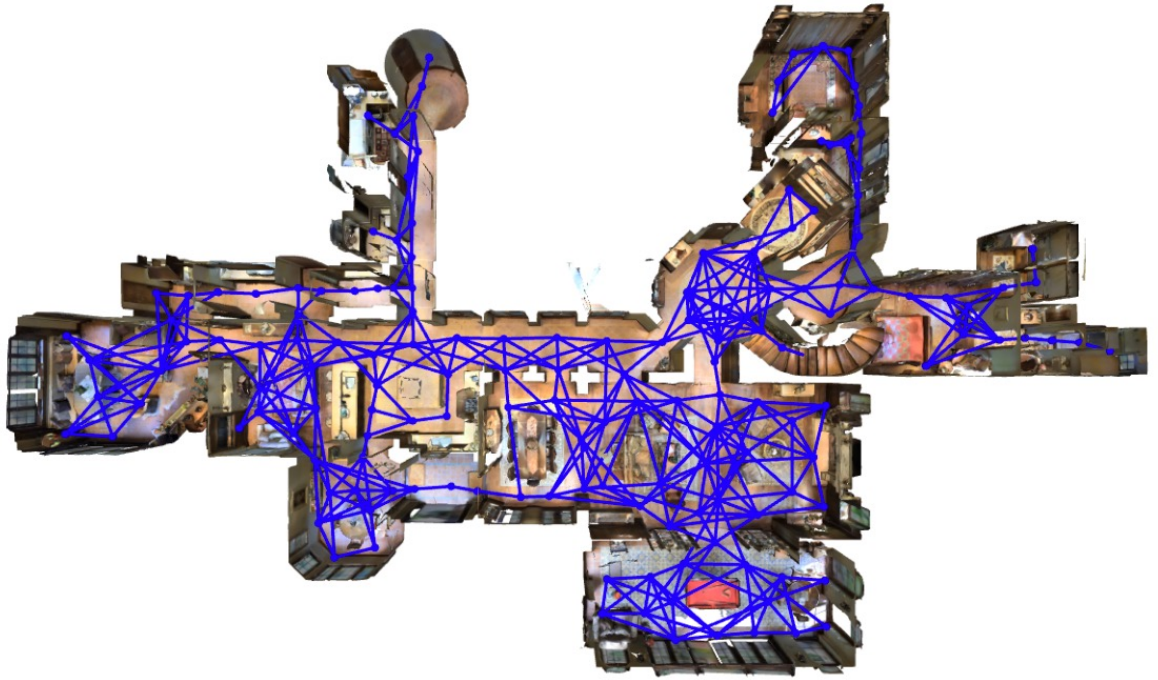
- Bettre accuracy for detecting whether the textually-referred objects are in the give image or not.

RefEgo: Referring Expression Comprehension Dataset from First-Person Perception of Ego4D, Kurita, Katsura and Onami (ICCV2023).

Vision and language navigation (VLN)

Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

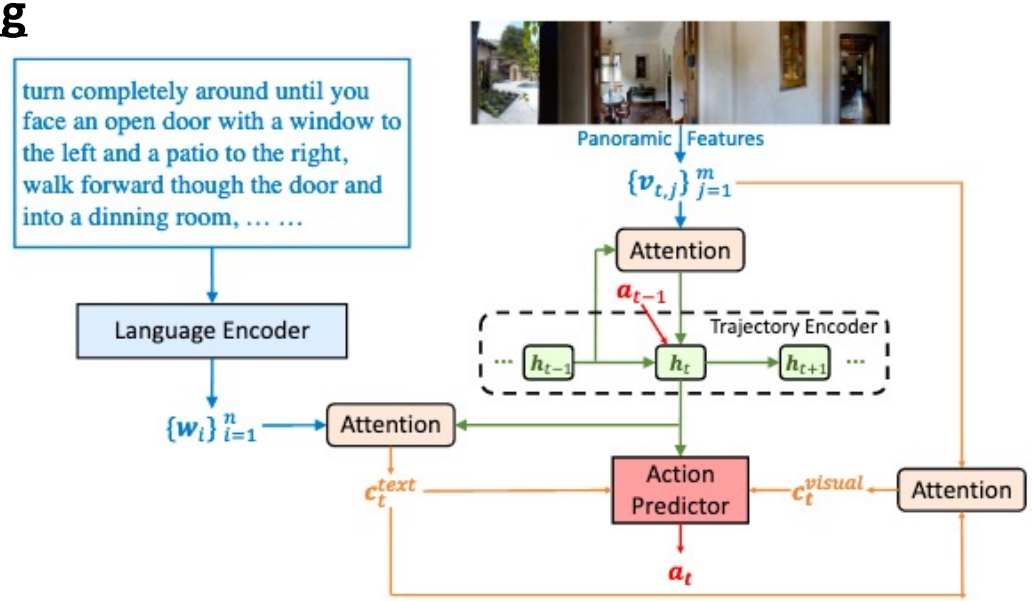
- A navigation task for a robotic agent to reach the goal place following the textual instruction.
- Photorealistic environment.
  Based on real houses, corrected with 3D scans.
- The dataset was released in 2017.
- Possible application to the robotic navigation



Language model-based approach

We proposed the image-captioning-style language model approach for VLN.

Existing



Reinforced Cross-modal Matching [Wang *et al*. 2018]

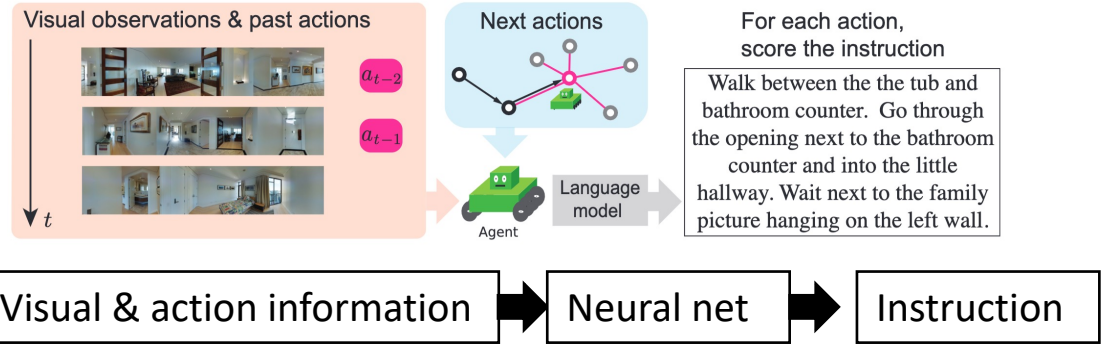
The problem of this commonplace method:

As a common problem for vision-and-language tasks, the neural network models often depend on either visual or textual information, and ignore the other.

➡ The model performance often degrades in unseen conditions.

Proposed

The neural network is the conditional language-generation model given the visual and action information.



Visual & action information ➡ Neural net ➡ Instruction

The neural netowork score the given instruction under the visual information and given action.
Both visual and textual information is crucial for this network.

The first work that directly utilize the language-generation model for the navigation task.

➡ The kangugae model-based approach does NOT degrades in unseen conditions

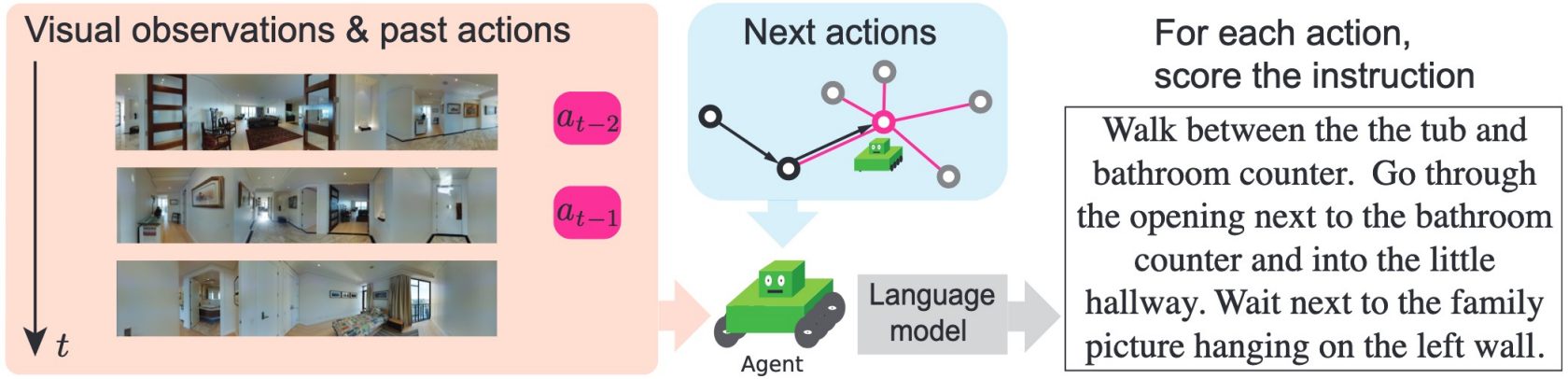
Generative language-grounded policy (GLGP)

Instruction X 、 Environment （vision & action history） h_t .
The existing study directly models $p(a_t \mid h_t, X)$,
while we assume the same action prior $p(a_t \mid h_t) = 1/|A|$ (A : action set) and obtain

$$p(a_t|h_t, X) = \frac{p(X|a_t, h_t)p'(a_t|h_t)}{\sum_{a'_t \in \mathcal{A}} p(X|a'_t, h_t)p'(a'_t|h_t)} = \frac{p(X|a_t, h_t)}{\sum_{a'_t \in \mathcal{A}} p(X|a'_t, h_t)}$$

with a language model $p(X|a_t, h_t)$.
Language model $p(X|a_t, h_t)$ selects next actions by the likelihood of the instruction X as action prediction.

$$L = -\sum_{t=1}^{T} \left\{ \log p(X|a_t, h_t) + \log \sum_{a'_t \in A} p(X|a'_t, h_t) \right\}$$



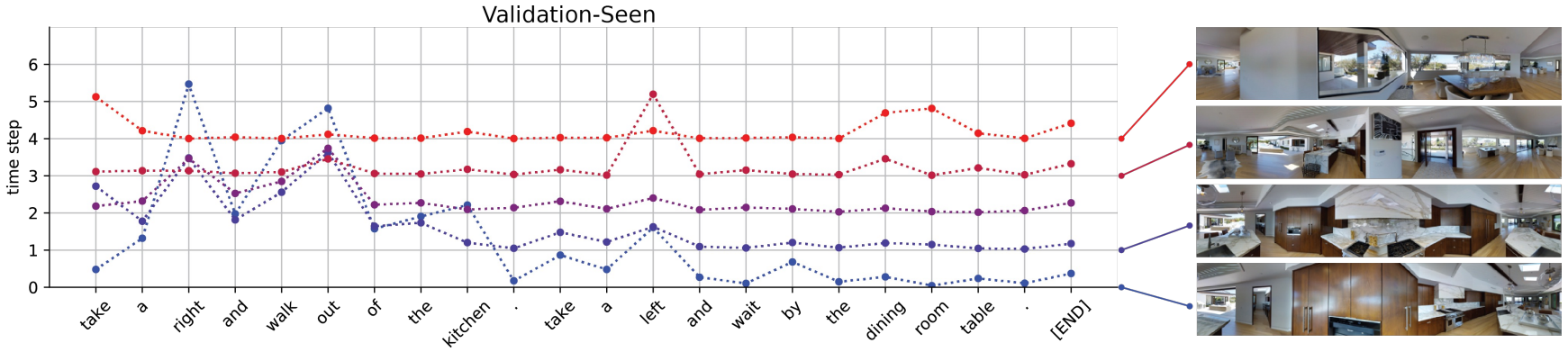
Generative Language-Grounded Policy in Vision-and-Language Navigation with Bayes' Rule
Shuhei Kurita and Kyunghyun Cho ICLR2021

1-TENT Visualizations

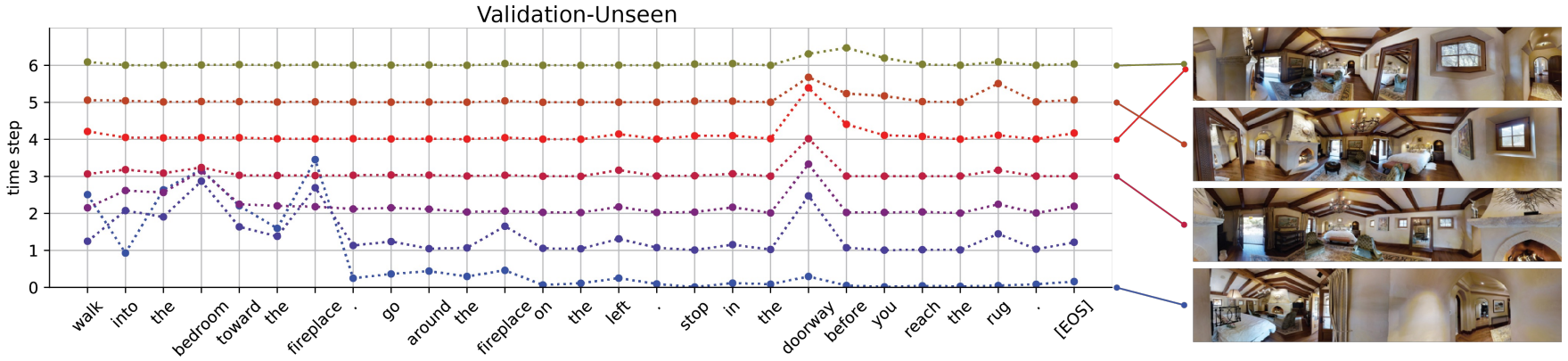
Action-decision visualization grounded on token-wise likelihood

$$S(w_k) = -\sum_{a_t \in \mathcal{A}} q(a_t, w_k) \log_{|\mathcal{A}|} q(a_t, w_k),$$

$$q(a_t, w_k) = \frac{p(w_k | a_t, h_t, w_{:k-1})}{\sum_{a_t \in \mathcal{A}} p(w_k | a_t, h_t, w_{:k-1})}$$



Take a right and walk out of the kitchen. Take a left and wait by the dining room table.

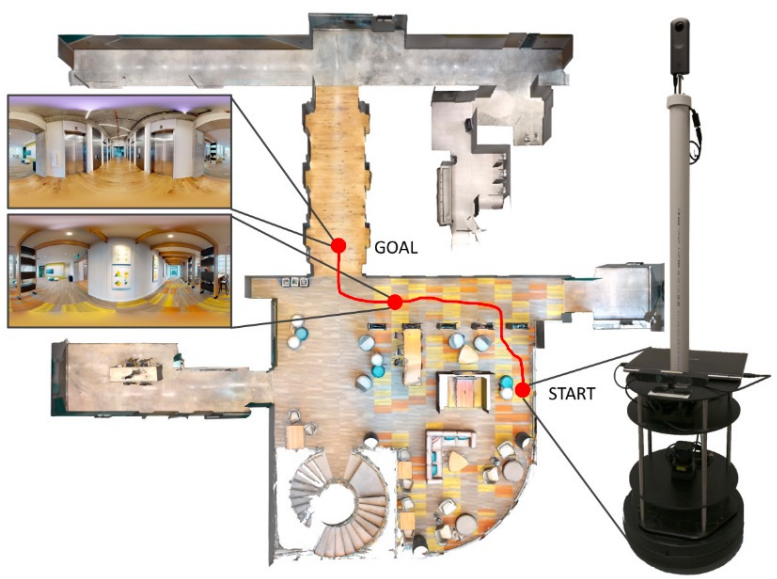


Walk into the bedroom toward the fireplace. Go around the fireplace on the left.
Stop in the doorway before you reach the rug.

VLN application on robotics

Is it possible to deploy VLN models on real robots? ***– Yes!***

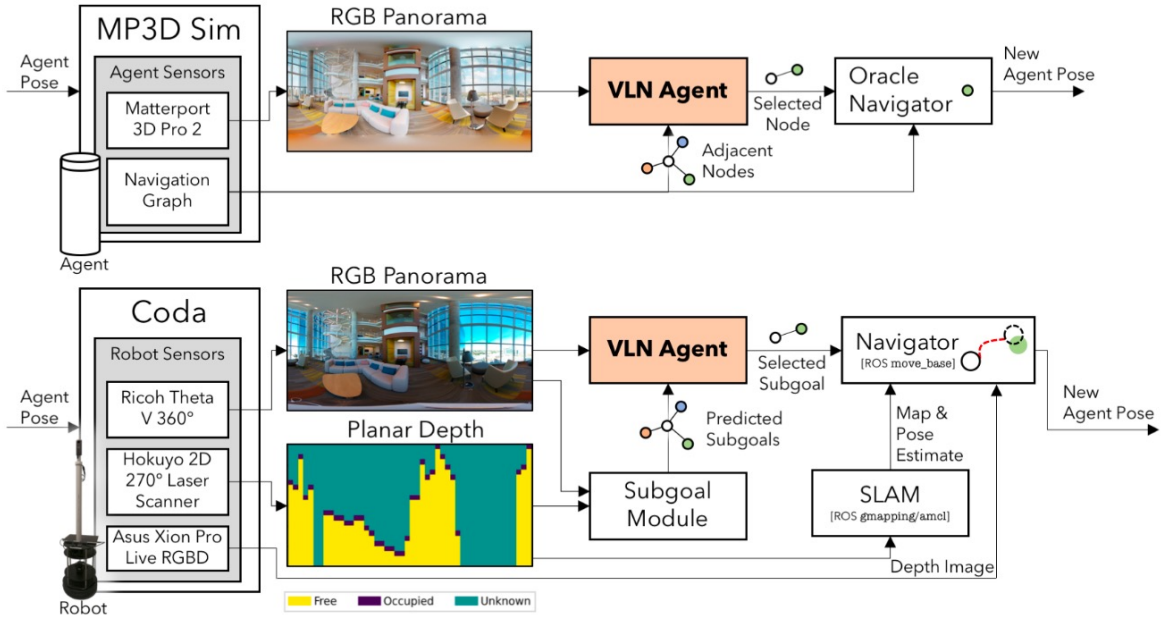
Sim-to-Real Transfer for Vision-and-Language Navigation
P. Anderson *et al*. (2020) *– The same first author with the original VLN paper.*



Go between the first and second bookshelves, turn to your left and walk straight down the hallway, then you should turn to your right at the hallway with the elevators and stop when the fire extinguisher box is on your left.

Walk between the two bookshelves, turn left, walk past the last set of pictures on the wall, turn right and wait by the elevators.

Turn left and head toward and past the blue bookcase. Turn left again and walk down the long hallway until you get to the opening on the right. Turn right and head toward the elevators and you're there.



Possible future direction:
Integration of high-level machine-learning agents and low-level robotic manipulation.