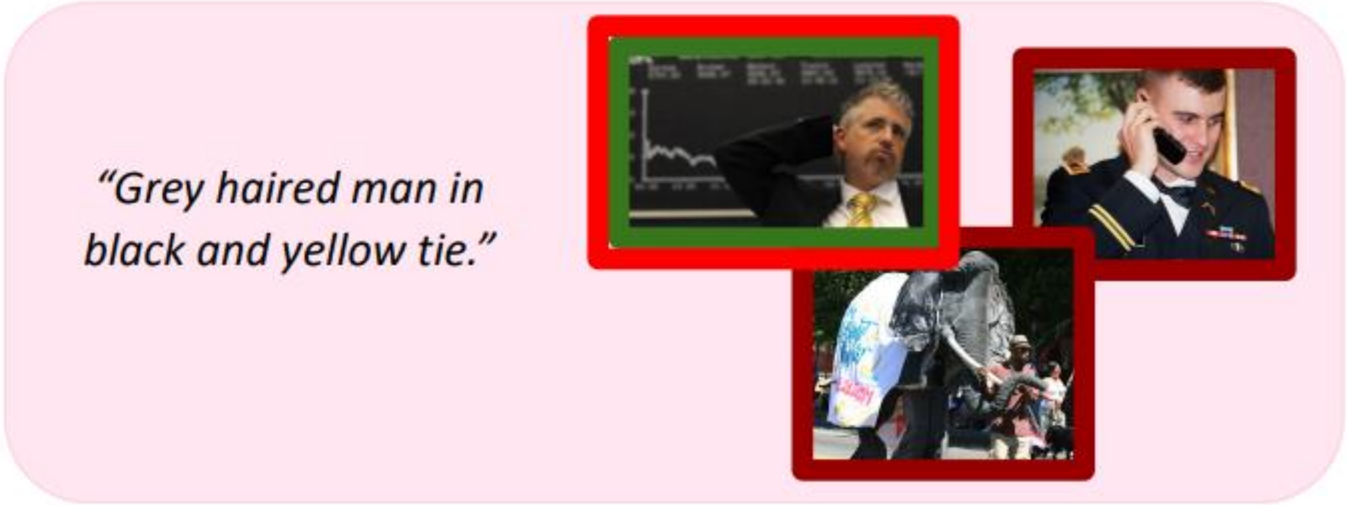
Vision-Language Pretraining: Current Trends and the Future

윤예준

목차

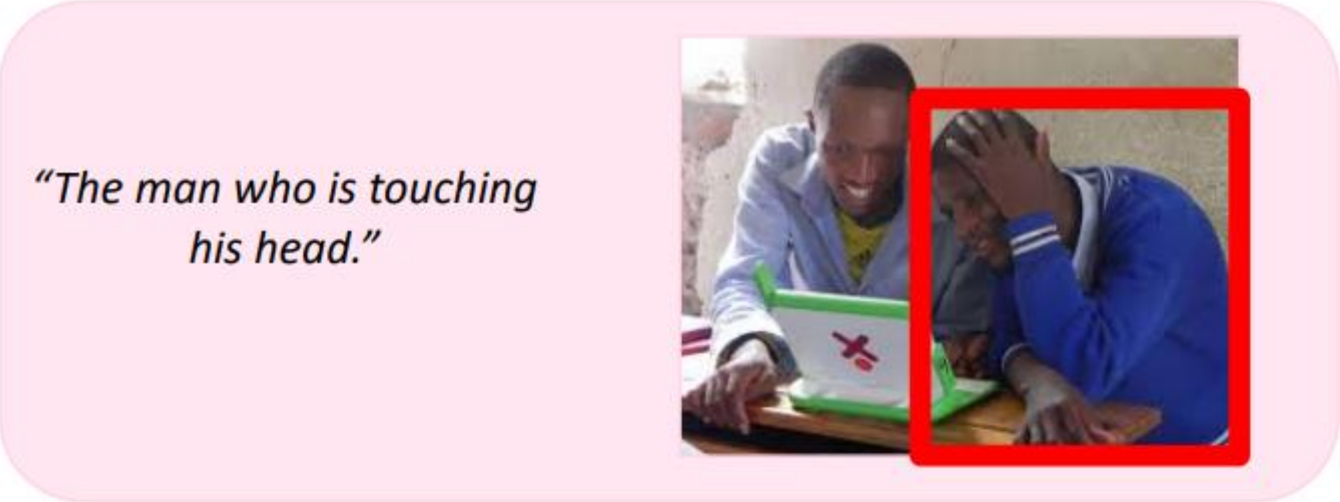
- Vision-Language Task
- Modern vision-language pretraining
- Current Trends and the Future

Image Retrieval



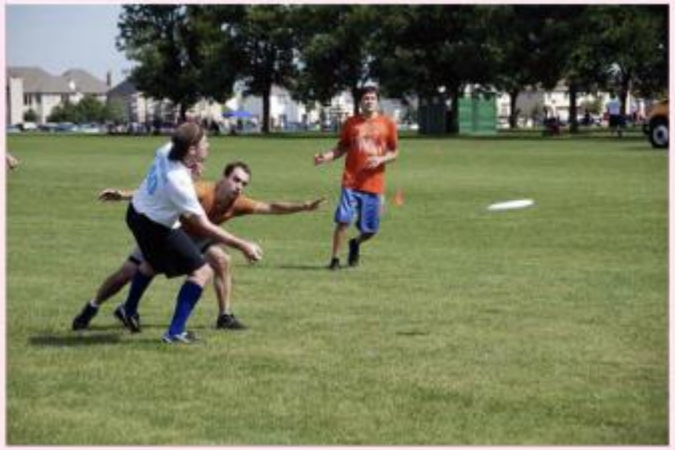
"Grey haired man in black and yellow tie."

Grounding Referring Expressions



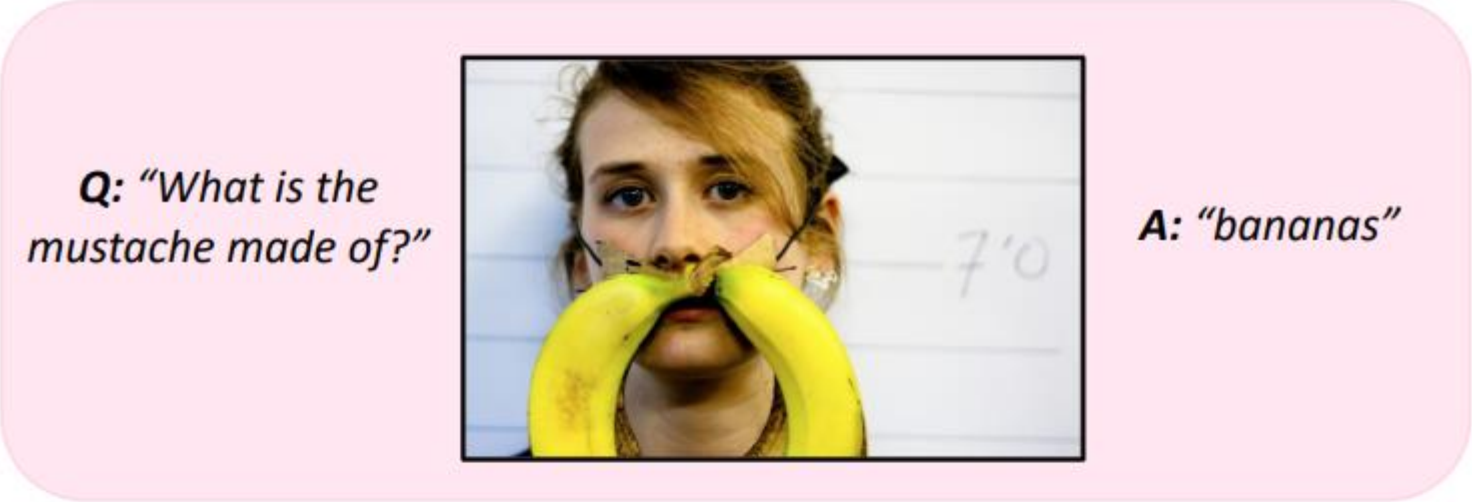
"The man who is touching his head."

Image Captioning



"A group of young people playing a game of Frisbee."

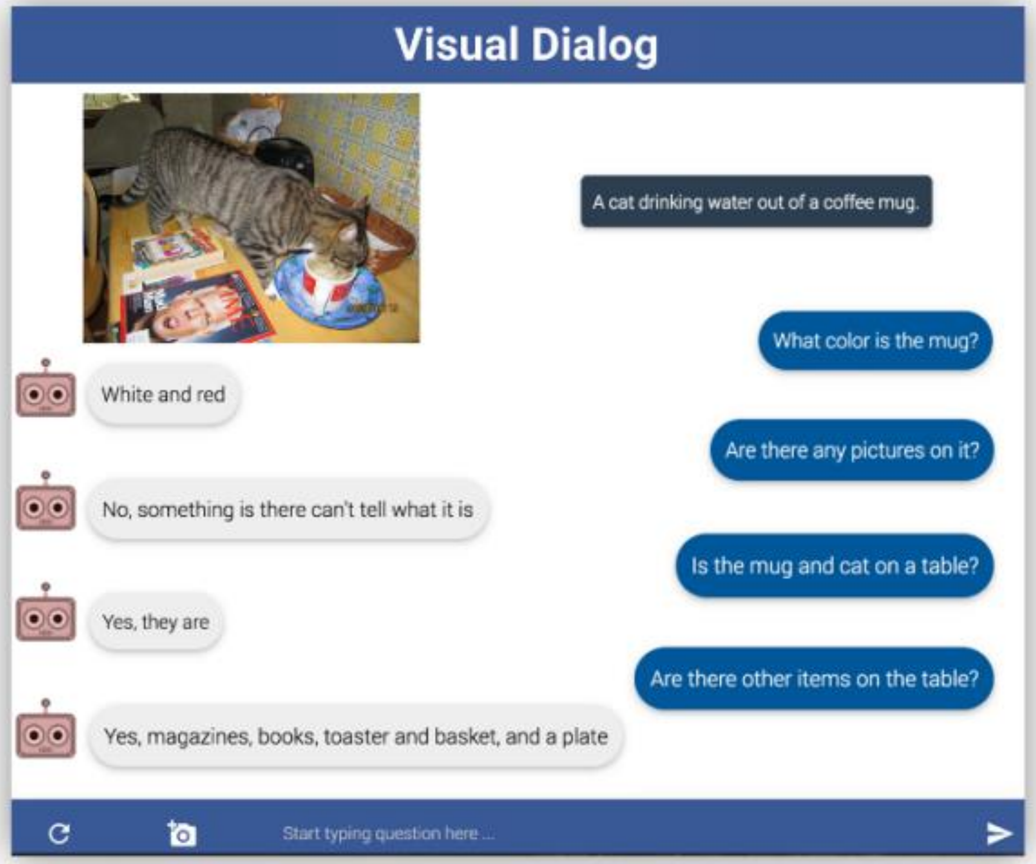
Visual Question Answering(VQA)



Q: *“What is the mustache made of?”*

A: *“bananas”*

Visual Dialog



Visual Dialog

A cat drinking water out of a coffee mug.

What color is the mug?

White and red

Are there any pictures on it?

No, something is there can't tell what it is

Is the mug and cat on a table?

Yes, they are

Are there other items on the table?

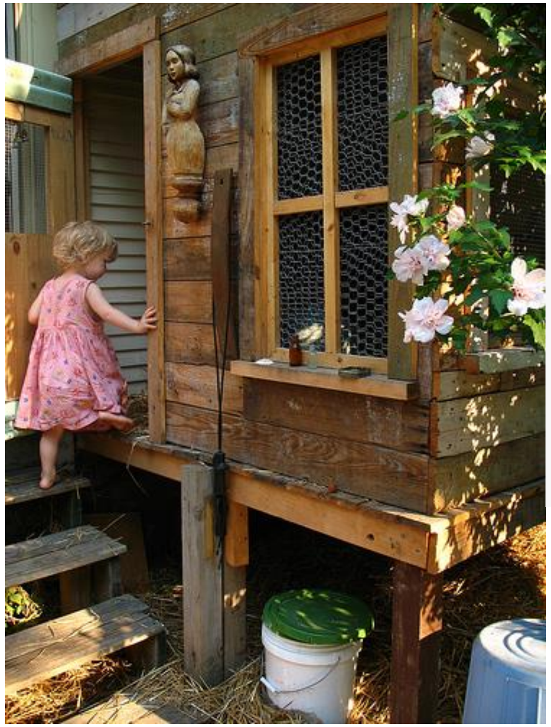
Yes, magazines, books, toaster and basket, and a plate

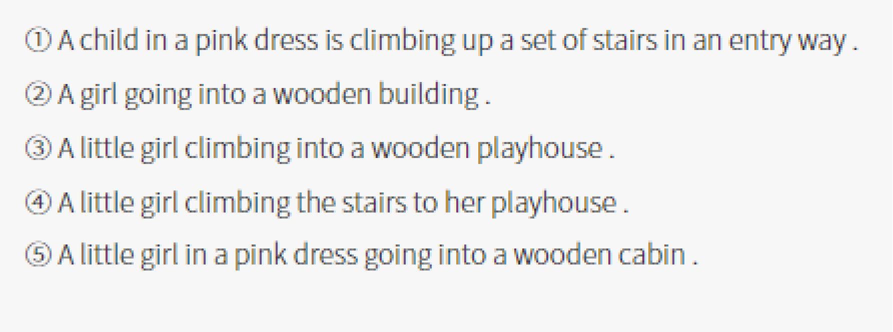
VL Datasets

- Image Retrieval: Flickr, COCO
- Grounding Referring Expression: RefCOCO, Visual7W
- Image Captioning: COCO
- Visual Question Answering: VQA v1, VQA v2, Visual Genome, GQA
- Visual Dialog: Visual Dialog, GuessWhat?!

VL Datasets

- Image Retrieval: Flickr, COCO





① A child in a pink dress is climbing up a set of stairs in an entry way .

② A girl going into a wooden building .

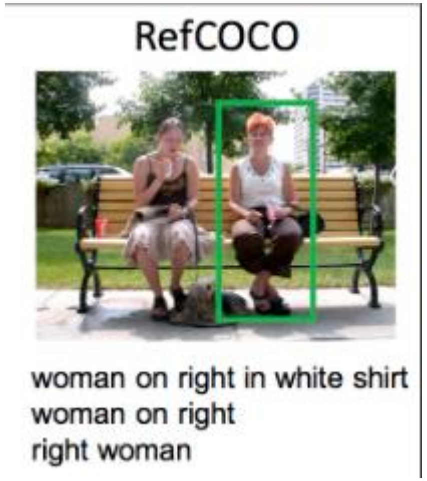
③ A little girl climbing into a wooden playhouse .

④ A little girl climbing the stairs to her playhouse .

⑤ A little girl in a pink dress going into a wooden cabin .

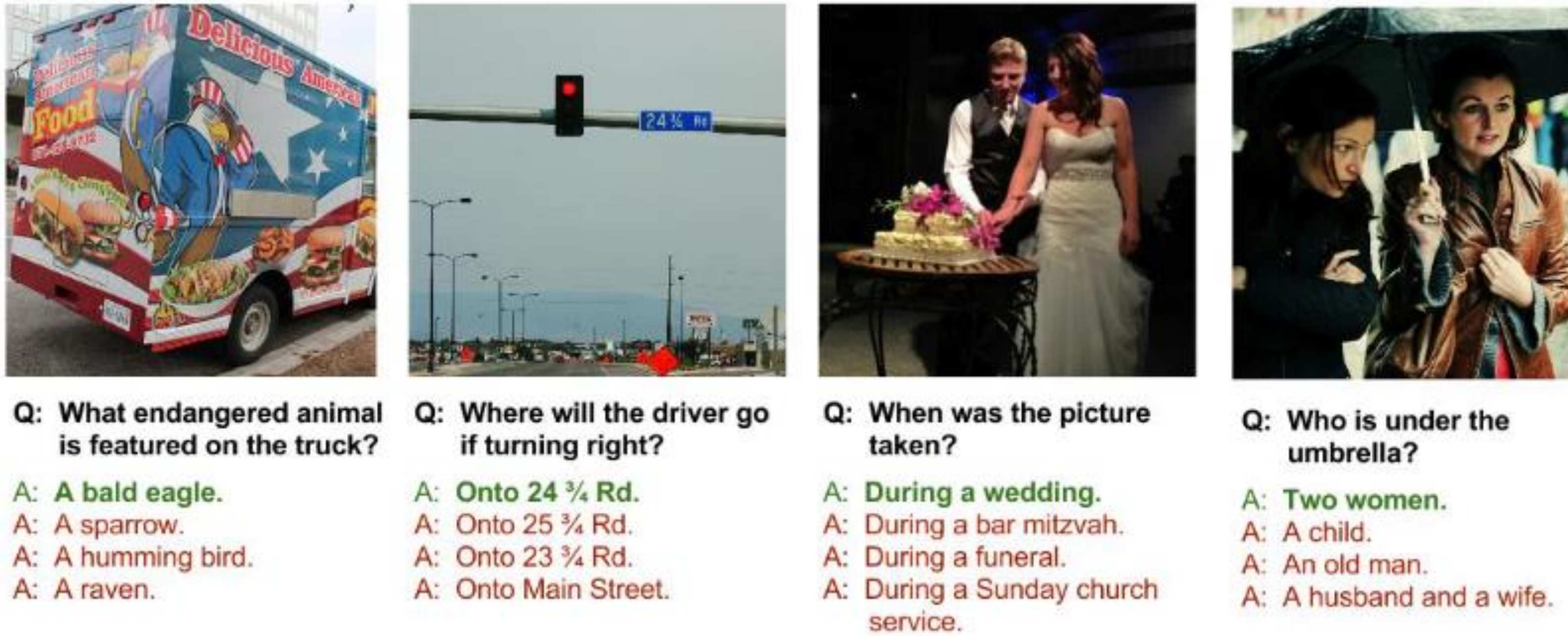
VL Datasets

- Grounding Referring Expression: RefCOCO, Visual7W



RefCOCO

woman on right in white shirt
woman on right
right woman



Q: What endangered animal is featured on the truck?
A: A bald eagle.
A: A sparrow.
A: A humming bird.
A: A raven.

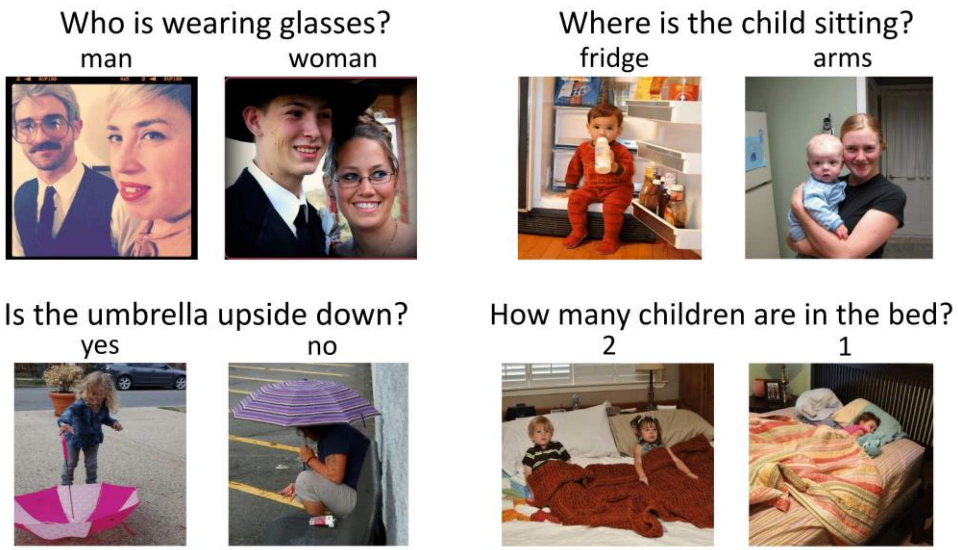
Q: Where will the driver go if turning right?
A: Onto 24 ¾ Rd.
A: Onto 25 ¾ Rd.
A: Onto 23 ¾ Rd.
A: Onto Main Street.

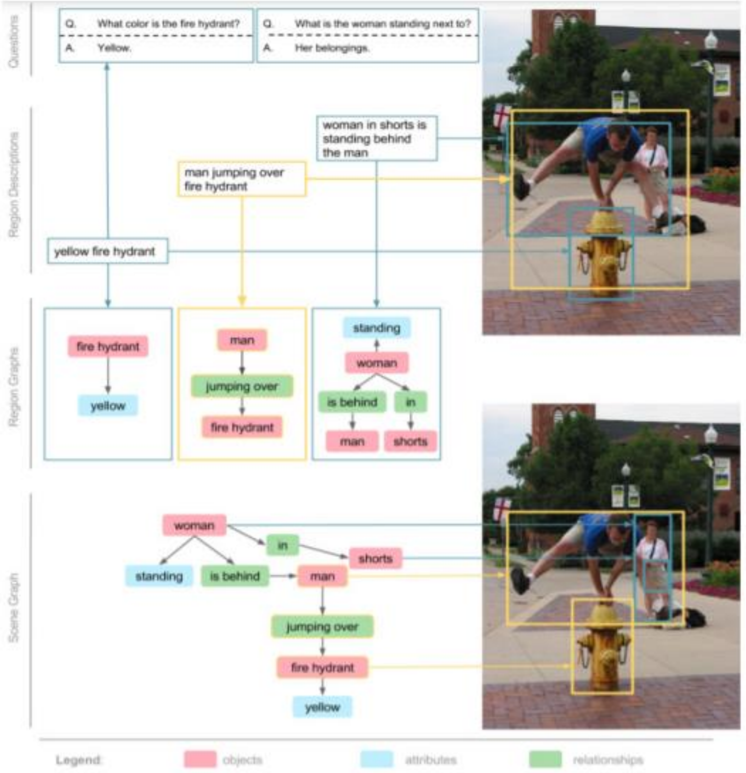
Q: When was the picture taken?
A: During a wedding.
A: During a bar mitzvah.
A: During a funeral.
A: During a Sunday church service.

Q: Who is under the umbrella?
A: Two women.
A: A child.
A: An old man.
A: A husband and a wife.

VL Datasets

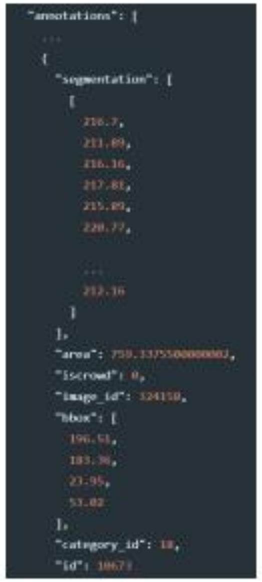
- Visual Question Answering: VQA v1, VQA v2, Visual Genome, GQA


Who is wearing glasses?
man
woman
Where is the child sitting?
fridge
arms
Is the umbrella upside down?
yes
no
How many children are in the bed?
2
1

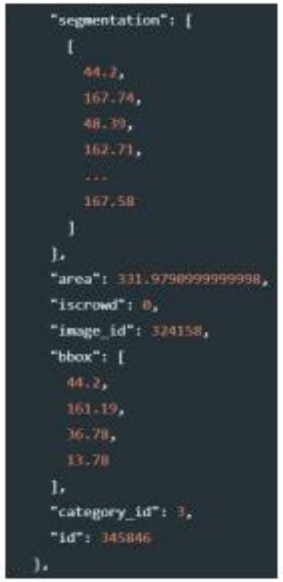


VL Datasets

- Image Captioning: COCO



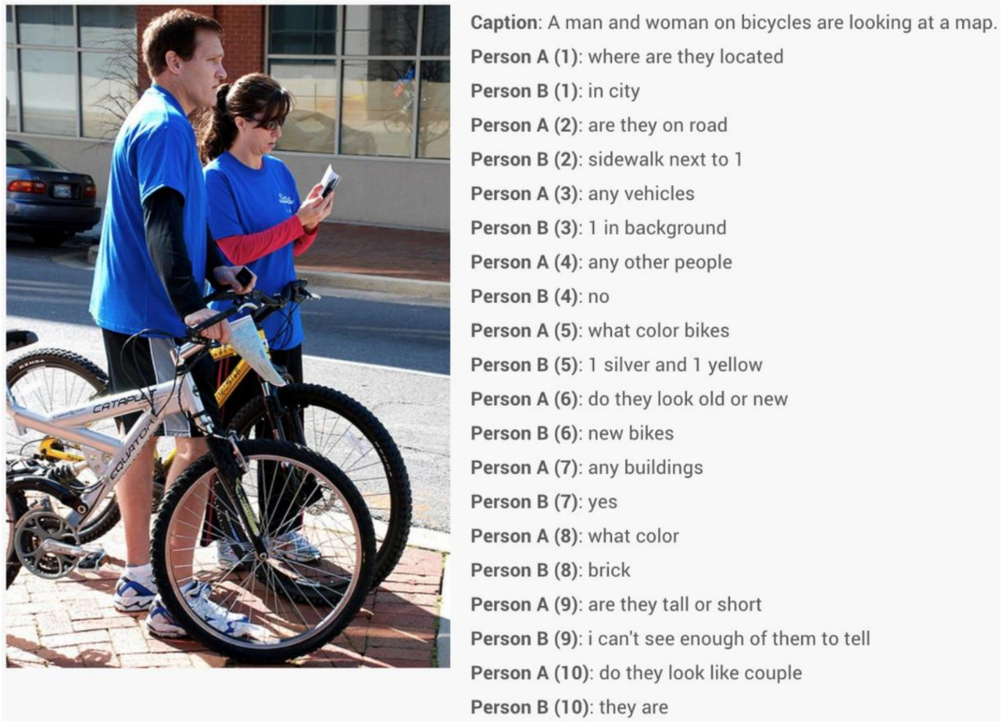
```
"annotations": [
  ...
  {
    "segmentation": [
      [
        216.7,
        211.09,
        216.16,
        217.81,
        215.09,
        220.77,


        ...
        212.16
      ]
    ],
    "area": 759.3375500000003,
    "iscrowd": 0,
    "image_id": 324158,
    "bbox": [
      196.51,
      183.36,
      23.95,
      53.02
    ],
    "category_id": 18,
    "id": 10673
```



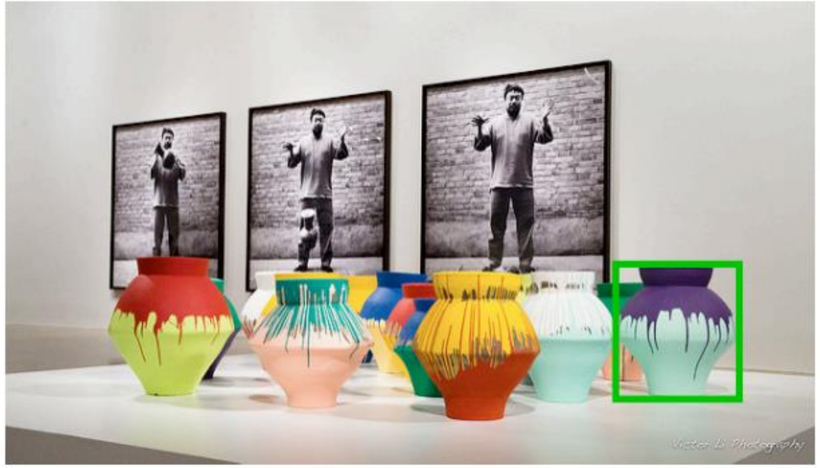
```
    "segmentation": [
      [
        44.2,
        167.74,
        48.39,
        162.71,
        ...
        167.58
      ]
    ],
    "area": 331.9790999999998,
    "iscrowd": 0,
    "image_id": 324158,
    "bbox": [
      44.2,
      161.19,
      36.78,
      13.78
    ],
    "category_id": 3,
    "id": 345846
  },
```

VL Datasets

- Visual Dialog: Visual Dialog, GuessWhat?!



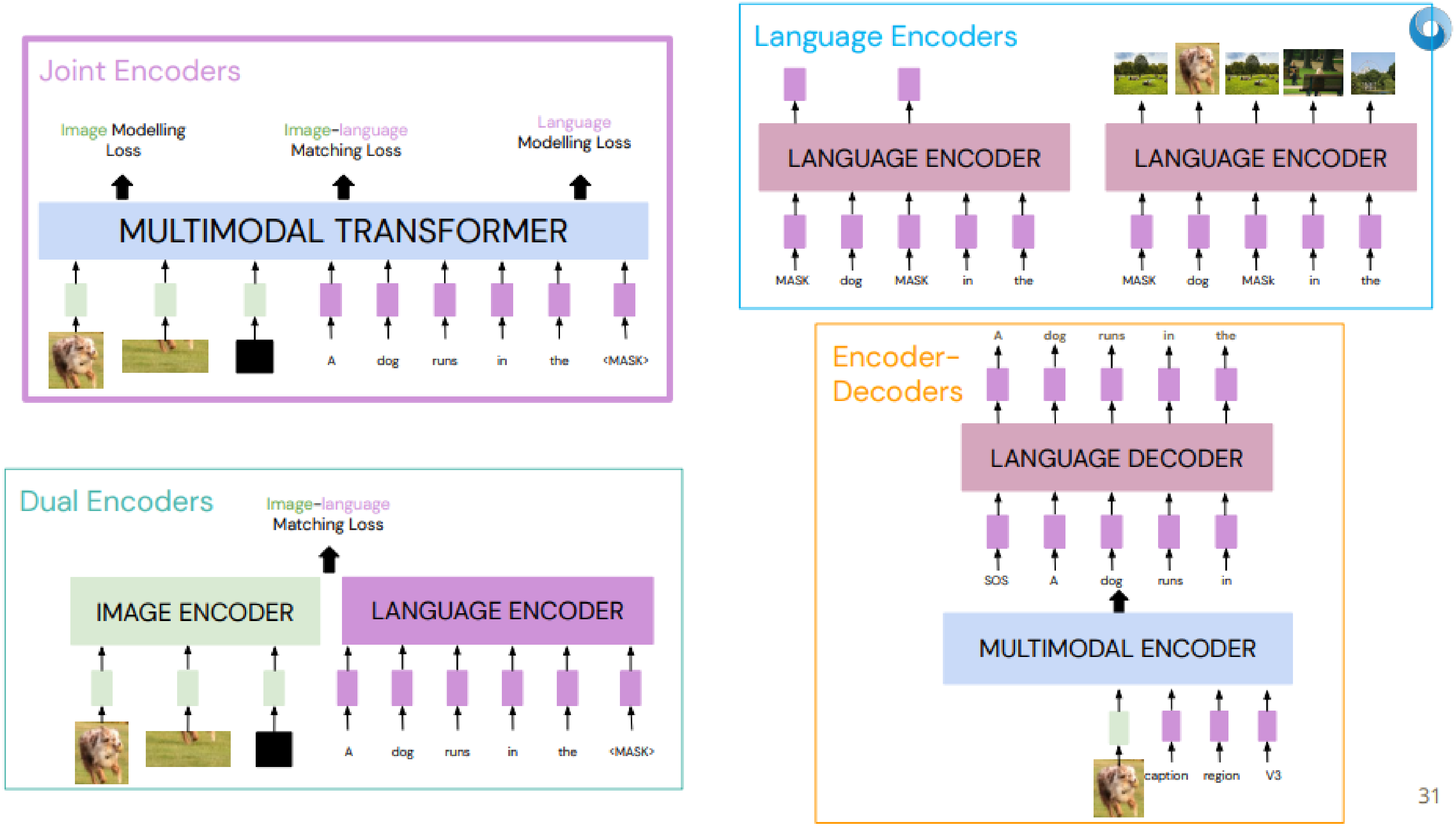
Caption: A man and woman on bicycles are looking at a map.
Person A (1): where are they located
Person B (1): in city
Person A (2): are they on road
Person B (2): sidewalk next to 1
Person A (3): any vehicles
Person B (3): 1 in background
Person A (4): any other people
Person B (4): no
Person A (5): what color bikes
Person B (5): 1 silver and 1 yellow
Person A (6): do they look old or new
Person B (6): new bikes
Person A (7): any buildings
Person B (7): yes
Person A (8): what color
Person B (8): brick
Person A (9): are they tall or short
Person B (9): i can't see enough of them to tell
Person A (10): do they look like couple
Person B (10): they are



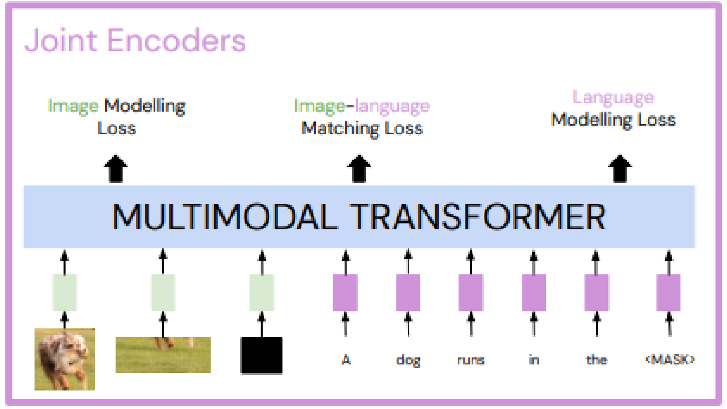
Questioner	Oracle
Is it a vase?	Yes
Is it partially visible?	No
Is it in the left corner?	No
Is it the turquoise and purple one?	Yes

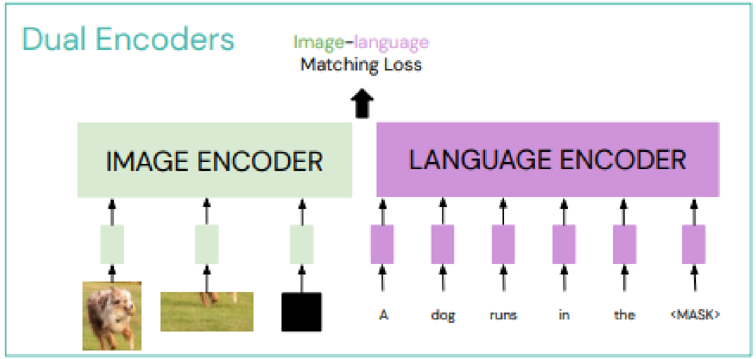
Figure 1: An example game. After a sequence of four questions, it becomes possible to locate the object (highlighted by a green bounding box).

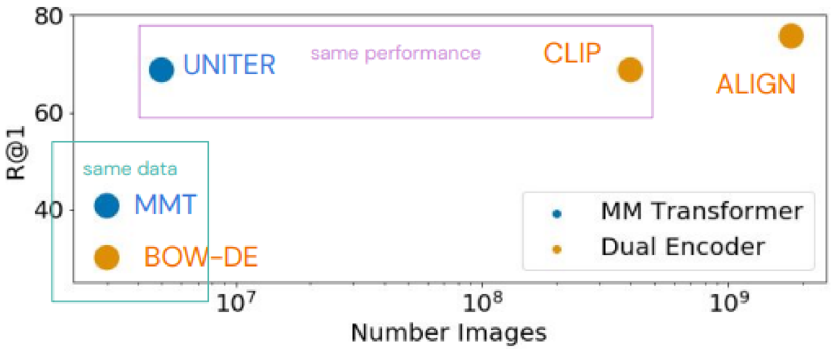
현대 모델 구조


Joint Encoders
Image Modelling Loss
Image-language Matching Loss
Language Modelling Loss
MULTIMODAL TRANSFORMER
A
dog
runs
in
the
<MASK>
Dual Encoders
Image-language Matching Loss
IMAGE ENCODER
LANGUAGE ENCODER
A
dog
runs
in
the
<MASK>
Language Encoders
LANGUAGE ENCODER
MASK
dog
MASK
in
the
LANGUAGE ENCODER
MASK
dog
MASk
in
the
Encoder-Decoders
A
dog
runs
in
the
LANGUAGE DECODER
SOS
A
dog
runs
in
MULTIMODAL ENCODER
caption
region
V3
31

현대 모델 구조

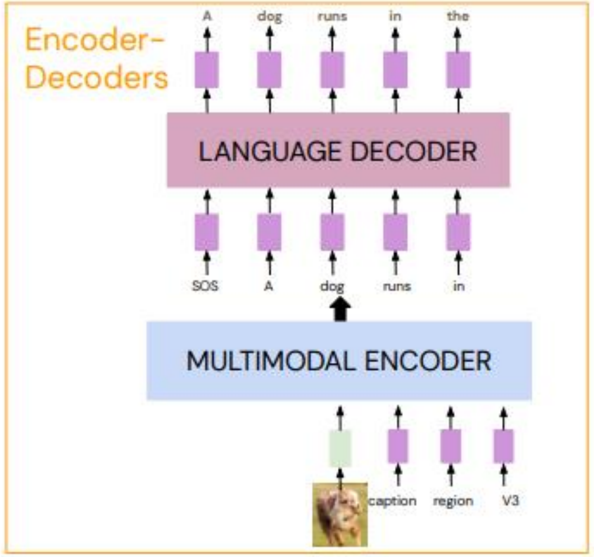






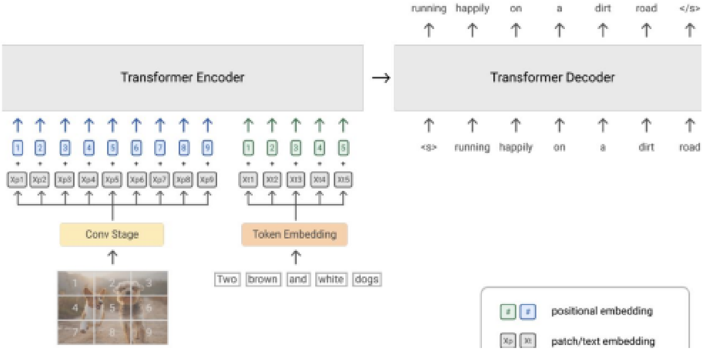
BOW-DE: [Miech & Alayrac et al. CVPR 2021]
MMT: [Hendricks et al. TACL 2021]
UNITER: [Chen et al. ECCV 2020]
CLIP: [Radford et al. Arxiv 2021]
ALIGN: [Jia et al. Arxiv 2021]

현대 모델 구조



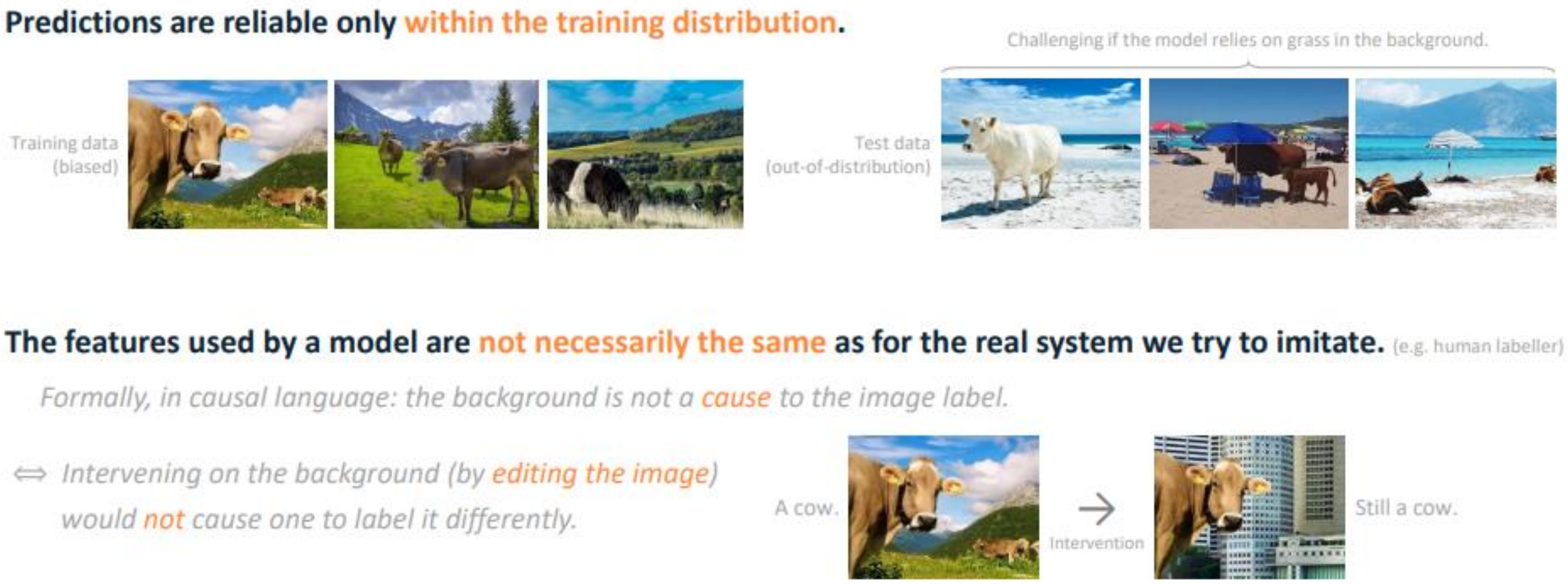
SimVLM [Wang et al, 2022]

Unifies tasks as text generation.
Removes object detection supervision.
Trains on large-scale noisy image-text data (ALIGN).



Current Trends and the Future

Statistical learning has limitations.



Predictions are reliable only within the training distribution.

Challenging if the model relies on grass in the background.

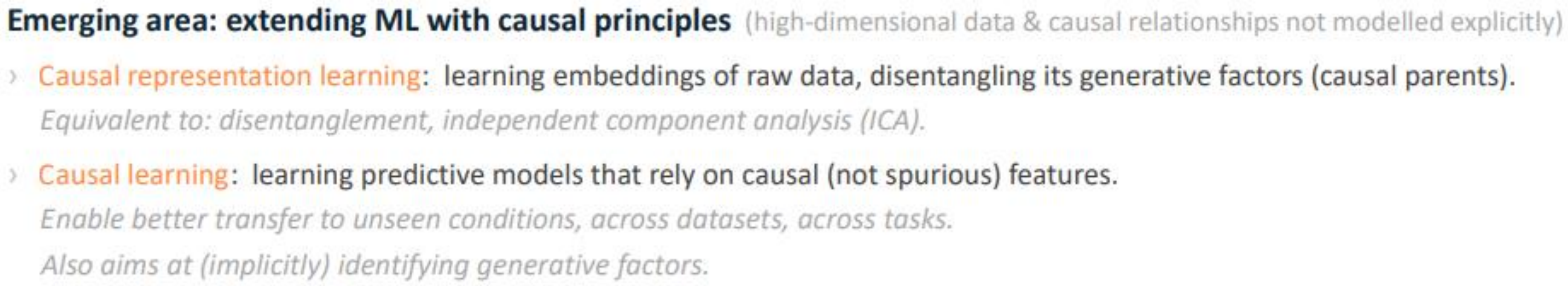
The features used by a model are not necessarily the same as for the real system we try to imitate. (e.g. human labeller)

Formally, in causal language: the background is not a cause to the image label.

⟺ *Intervening on the background (by editing the image) would not cause one to label it differently.*

Causal learning

- Learning the data-generating mechanisms of a task (and not just the correlations in a specific dataset).

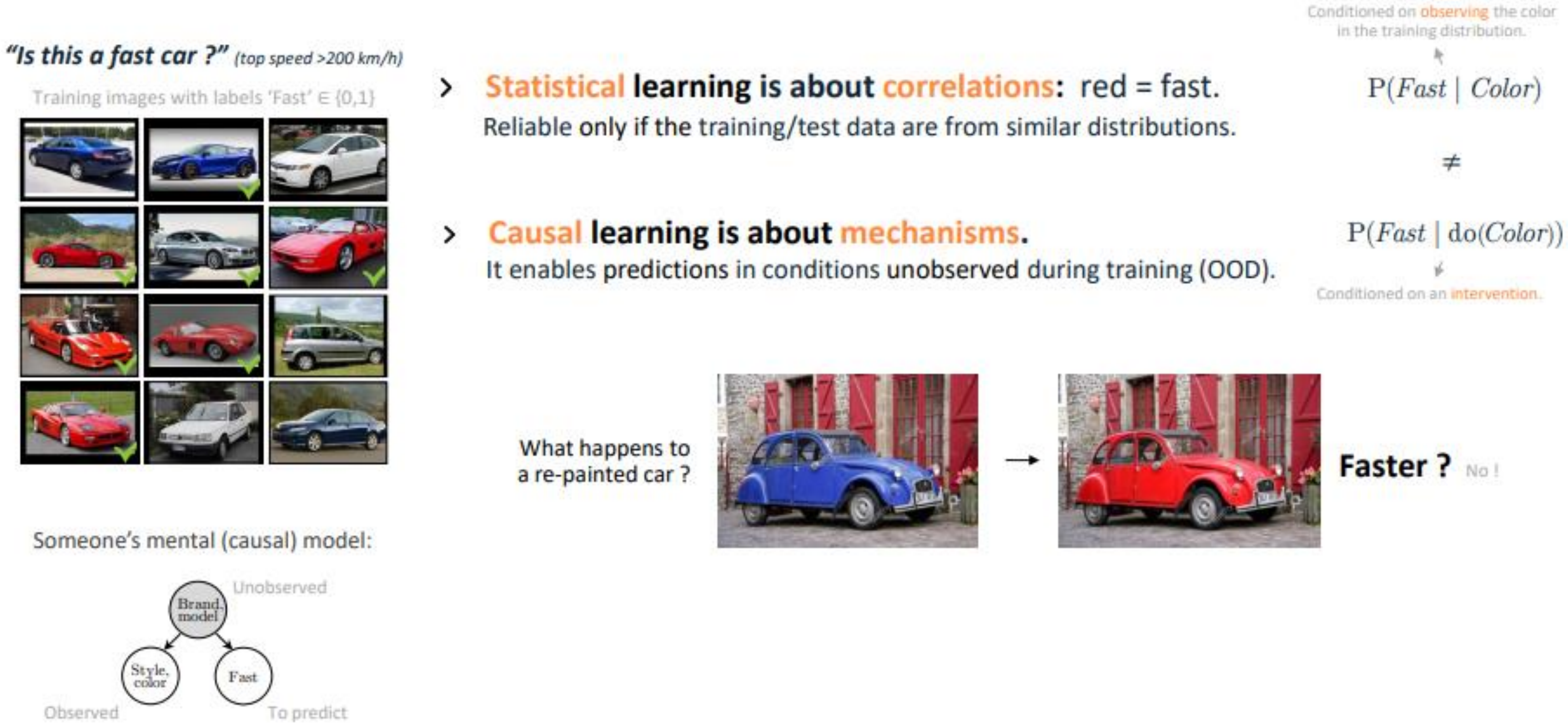


Emerging area: extending ML with causal principles (high-dimensional data & causal relationships not modelled explicitly)

› Causal representation learning: learning embeddings of raw data, disentangling its generative factors (causal parents).
Equivalent to: disentanglement, independent component analysis (ICA).

› Causal learning: learning predictive models that rely on causal (not spurious) features.
Enable better transfer to unseen conditions, across datasets, across tasks.
Also aims at (implicitly) identifying generative factors.

statistical vs. causal learning



"Is this a fast car ?" *(top speed >200 km/h)*

Training images with labels 'Fast' ∈ {0,1}

- **Statistical learning is about correlations:** red = fast.
  Reliable only if the training/test data are from similar distributions.

- **Causal learning is about mechanisms.**
  It enables predictions in conditions unobserved during training (OOD).

Conditioned on observing the color in the training distribution.

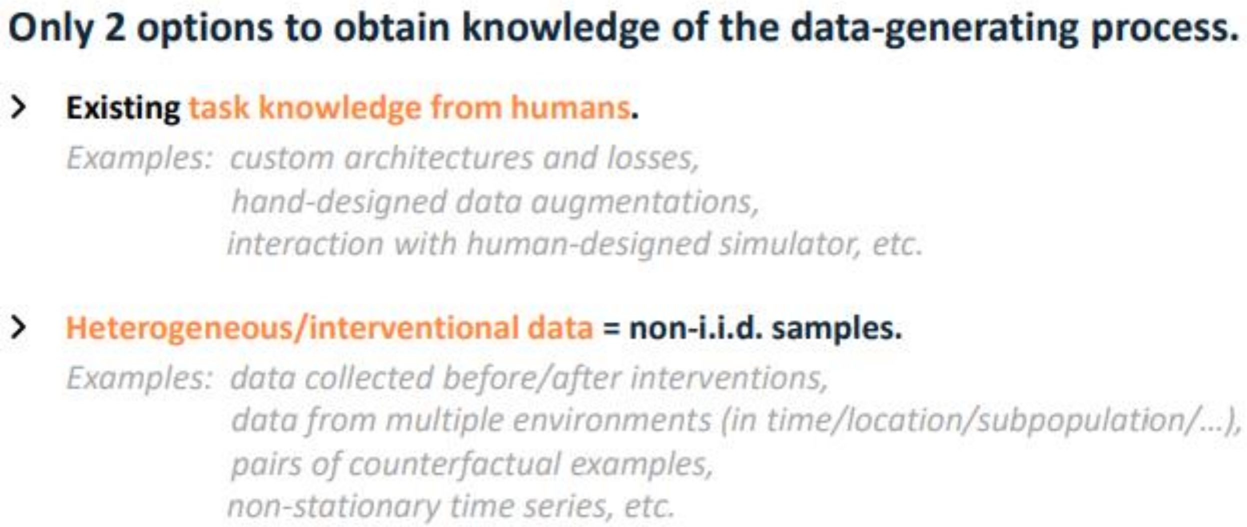
$$\mathrm{P}(\mathit{Fast} \mid \mathit{Color}) \neq \mathrm{P}(\mathit{Fast} \mid \mathrm{do}(\mathit{Color}))$$

Conditioned on an intervention.

What happens to a re-painted car ? → **Faster ?** No !

Someone's mental (causal) model:

statistical vs. causal learning



Only 2 options to obtain knowledge of the data-generating process.

- **Existing task knowledge from humans.**

  Examples: custom architectures and losses,
  hand-designed data augmentations,
  interaction with human-designed simulator, etc.

- **Heterogeneous/interventional data = non-i.i.d. samples.**

  Examples: data collected before/after interventions,
  data from multiple environments (in time/location/subpopulation/...),
  pairs of counterfactual examples,
  non-stationary time series, etc.