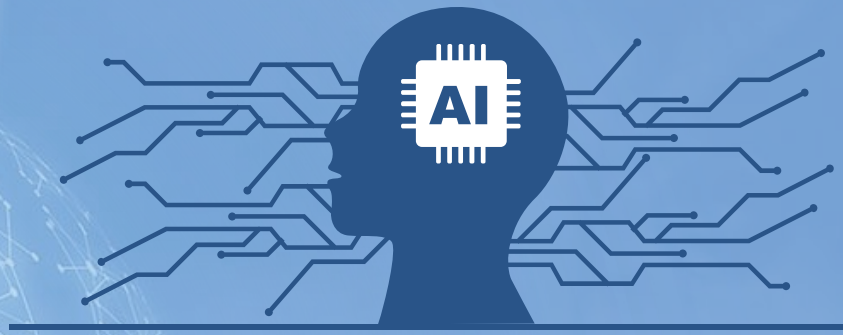
Artificial Intelligence



Knowledge

Wen-Huang Cheng (鄭文皇)

National Taiwan University

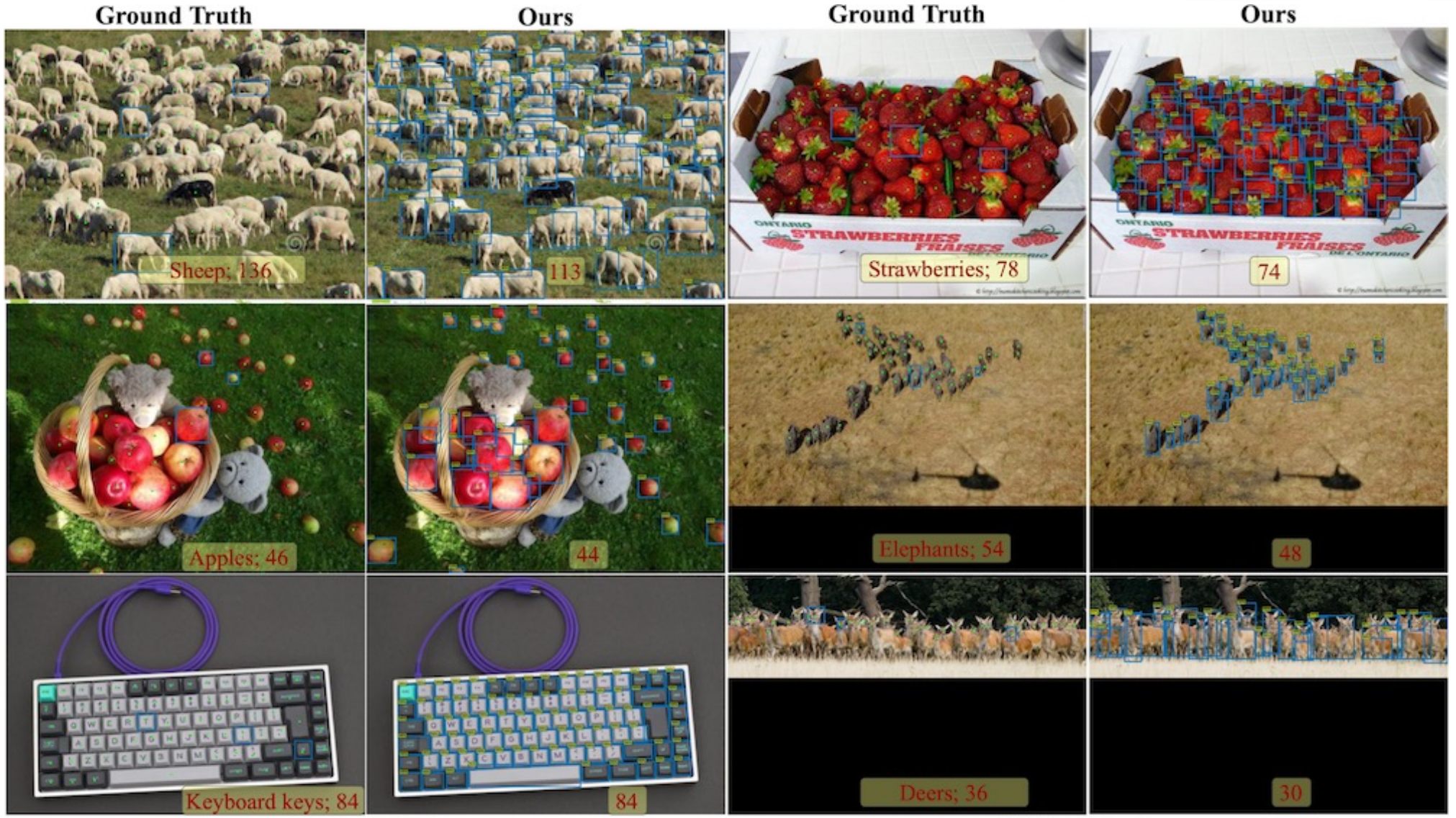
wenhuang@csie.ntu.edu.tw

Point, Segment and Count: A Generalized Framework for Object Counting

Zhizhong Huang[1] Mingliang Dai[1] Yi Zhang[2] Junping Zhang[1] Hongming Shan[3,4]
[1] Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai 200433, China
[2] College of Computer Science, Sichuan University, Chengdu 610065, China
[3] Institute of Science and Technology for Brain-inspired Intelligence and MOE Frontiers Center for Brain Science, Fudan University, Shanghai 200433, China
[4] Shanghai Center for Brain Science and Brain-inspired Technology, Shanghai 200031, China
{zzhuang19, mldai21, jpzhang, hmshan}@fudan.edu.cn, yzhang@scu.edu.cn

Code: https://github.com/Hzzone/PseCo

AI Weekly

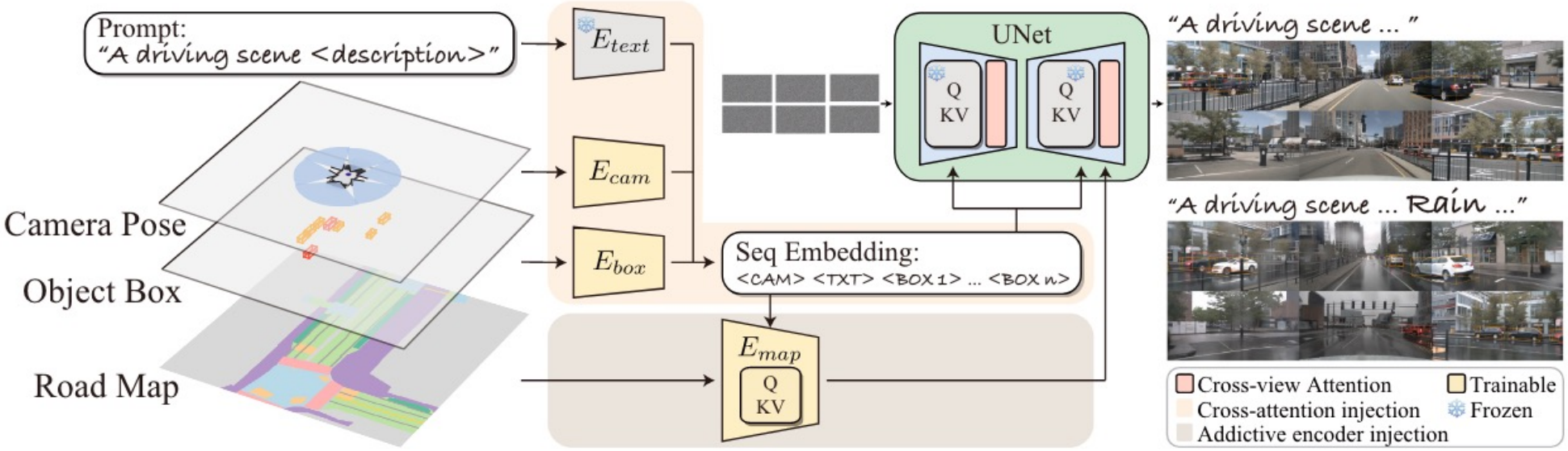

Ground Truth
Ours
Ground Truth
Ours
Sheep; 136
113
Strawberries; 78
74
Apples; 46
44
Elephants; 54
48
Keyboard keys; 84
84
Deers; 36
30
(a) Few-shot
(b) Zero-shot

ICLR

The International Conference on Learning Representations (ICLR)
2024

MAGICDRIVE: STREET VIEW GENERATION WITH DIVERSE 3D GEOMETRY CONTROL

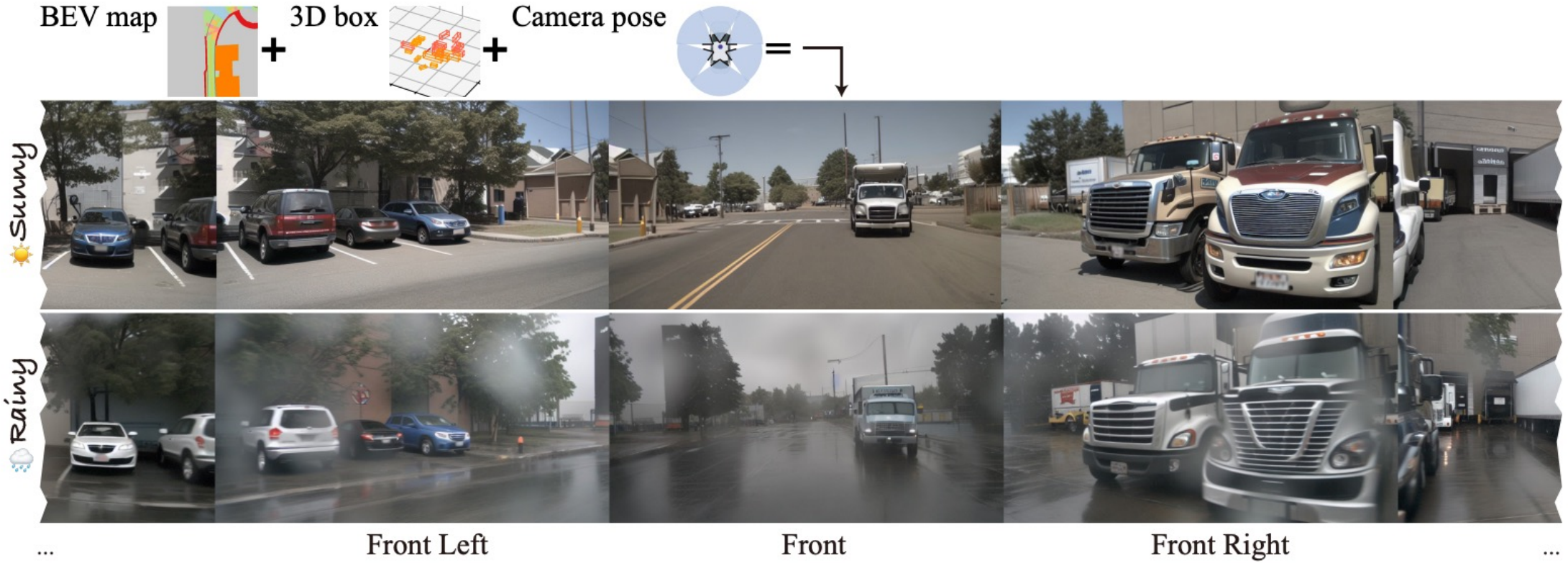
Ruiyuan Gao[1*], **Kai Chen**[2*], **Enze Xie**[3†], **Lanqing Hong**[3],
Zhenguo Li[3], **Dit-Yan Yeung**[2], **Qiang Xu**[1†]
[1]The Chinese University of Hong Kong [2]Hong Kong University of Science and Technology
[3]Huawei Noah's Ark Lab



AI Weekly

ICLR

The International Conference on Learning Representations (ICLR)
2024



This Lecture – Agenda

- **Knowledge Base**
- **Knowledge-based Agents**
  - Canonical Logic Models
    - Propositional Logic
    - Model Checking
  - Modern Vision-Language Models
    - Factual Knowledge
    - Common Knowledge
    - Model Knowledge

Knowledge

Knowledge

- Humans reason based on existing knowledge and draw conclusions. The concept of representing knowledge and drawing conclusions from it is also used in AI, and in this lecture we will explore how we can achieve this behavior.

How humans learn about knowledge?

Interaction

Knowledge

Human

External World

How humans learn about knowledge?

Interaction

Knowledge

Human

External World

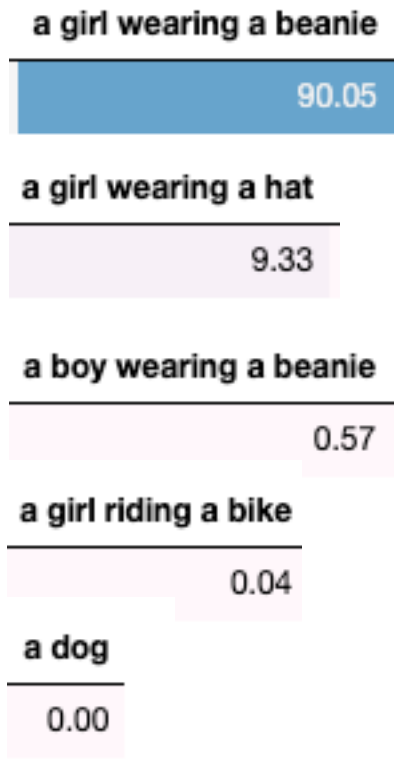
Vision-Language Models

- **Vision-language models** have demonstrated great potential in learning generic visual representations and allowing **zero-shot** transfer to a variety of downstream classification tasks via **prompting**.
  - Zero-shot: Zero-shot learning is a way to generalize on unseen labels, without having specifically trained to classify them.

Prompting (manually)

- ‘a girl wearing a beanie’.
- ‘a girl wearing a hat’.
- ‘a boy wearing a beanie’.
- ‘a girl riding a bike’.
- ‘a dog’.

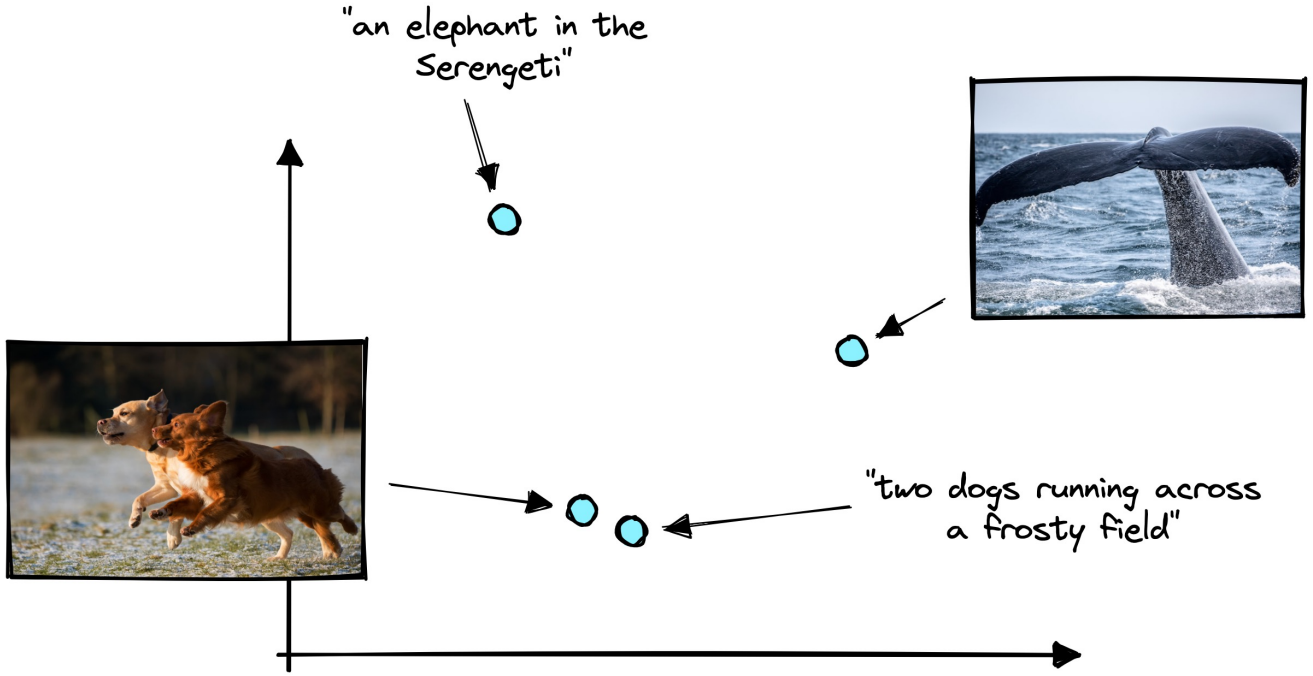
Making predictions with probability (by vision-language models)



Source: https://towardsdatascience.com/clip-the-most-influential-ai-model-from-openai-and-how-to-use-it-f8ee408958b1

Multi-modality

- The multi-modal nature of CLIP is powered by two encoder models trained to “speak the same language”. Text inputs are passed to a text encoder, and image inputs to an image encoder. These models then create a vector representation of the respective input by by encoding similar concepts in text and images into similar vectors.



Source: https://www.pinecone.io/learn/clip/

Large-scale Multi-modal Dataset: Example

Backend url:
https://knn5.laior
Index:
laion_5B

Clip retrieval works by converting the text query to a CLIP embedding , then using that embedding to query a knn index of clip image embedddings

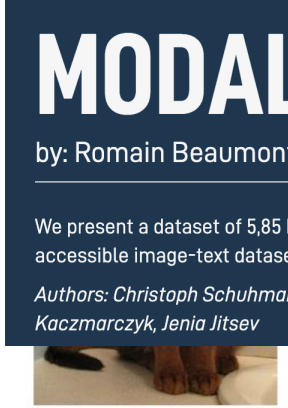
Display captions
Display full captions
Display similarities
Safe mode
Hide duplicate urls
Hide (near) duplicate images
Search over
image
Search with multilingual clip

french cat


french cat


french cat


How to tell if your feline is french. He wears a b...


イケメン猫モデル「トキ・ナンタケット」がかっこいい - NAVER まとめ


Hipster cat


網友挑戰「加幾筆畫出最創意貓咪圖片」，笑到岔氣之後我也手


cat in a suit Georgian sells tomatoes


French Bread Cat Loaf Metal Print

LAION-5B: A NEW ERA OF OPEN LARGE-SCALE MULTI-MODAL DATASETS

by: Romain Beaumont, 10 Oct, 2022

We present a dataset of 5,85 billion CLIP-filtered image-text pairs, 14x bigger than LAION-400M, previously the biggest openly accessible image-text dataset in the world.

Authors: Christoph Schuhmann, Richard Vencu, Romain Beaumont, Theo Coombes, Cade Gordon, Aarush Katta, Robert Kaczmarczyk, Jenia Jitsev

https://laion.ai/blog/laion-5b/

Observations

[Google Research 2023]

- **Open-web datasets used to train open text-to-image models suffer from significant issues.**
  - To be specific, the captions in the LAION dataset, used to train vision-language models, come from alt HTML tags (Alttext) and often describes only a narrow aspect of the image, neglecting significant visual details.
  - For example, an image of a person can have as Alttext the name of the person and the name of the photographer, but not a description of their appearance, their clothes, their position, or the background.

Source: “A Picture is Worth a Thousand Words: Principled Recaptioning Improves Image Generation,” arXiv, 2023.

Alttext Examples (1/2)

Alttext

2013 ducati monster 1100 evo diesel motorcycle photos and specifications. Black Bedroom Furniture Sets. Home Design Ideas

重新標圖

Cover of Oregon Wine Press February 2019

Alttext Examples (2/2)

Alttext

You Want To Learn About What Human Anatomy The Skeletal System

GraphicRiver Toucan Pattern 7750027

The RECAP Framework

- This work proposes to relabel the corpus with improved captions that are auto-generated by a custom I2T (image-to-text) model, so as to increase sample efficiency and allow a vision-language model to better understand the relations between captions and images.

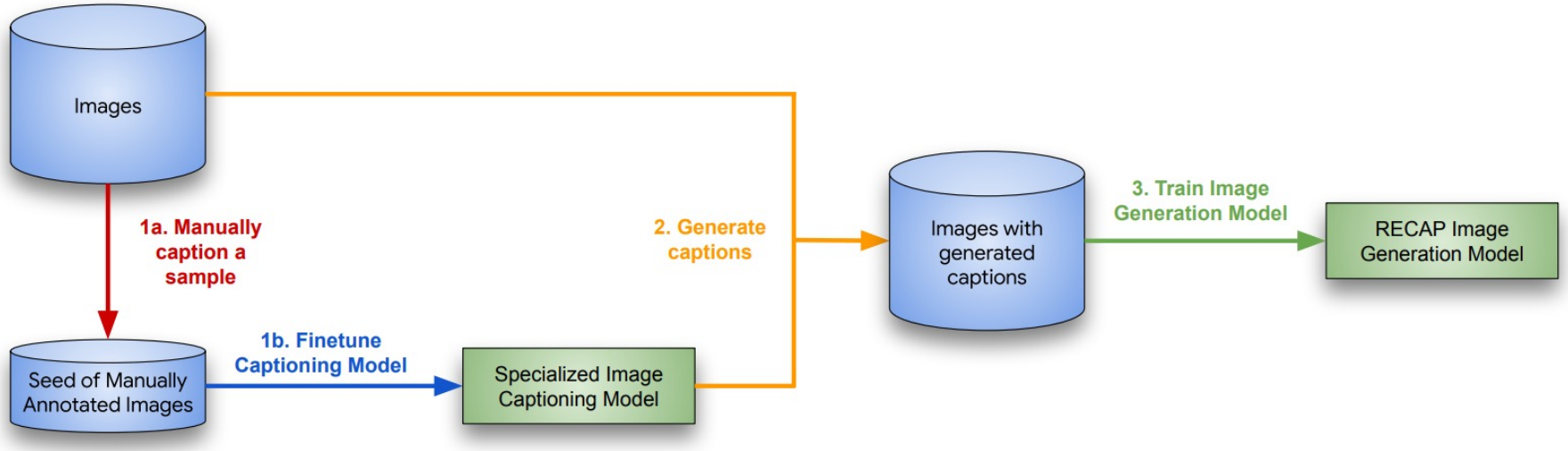


Figure 2. Schematic diagram of our method RECAP. In steps (1a) and (1b) we fine-tune an image-to-text captioning model on a small set of detailed human captions. In step (2) we use this fine-tuned model to recaption the images in the training dataset of a text-to-image model, and with this dataset, in step (3) we train an image generation model with the recaptioned dataset.

Examples

找專家標記的

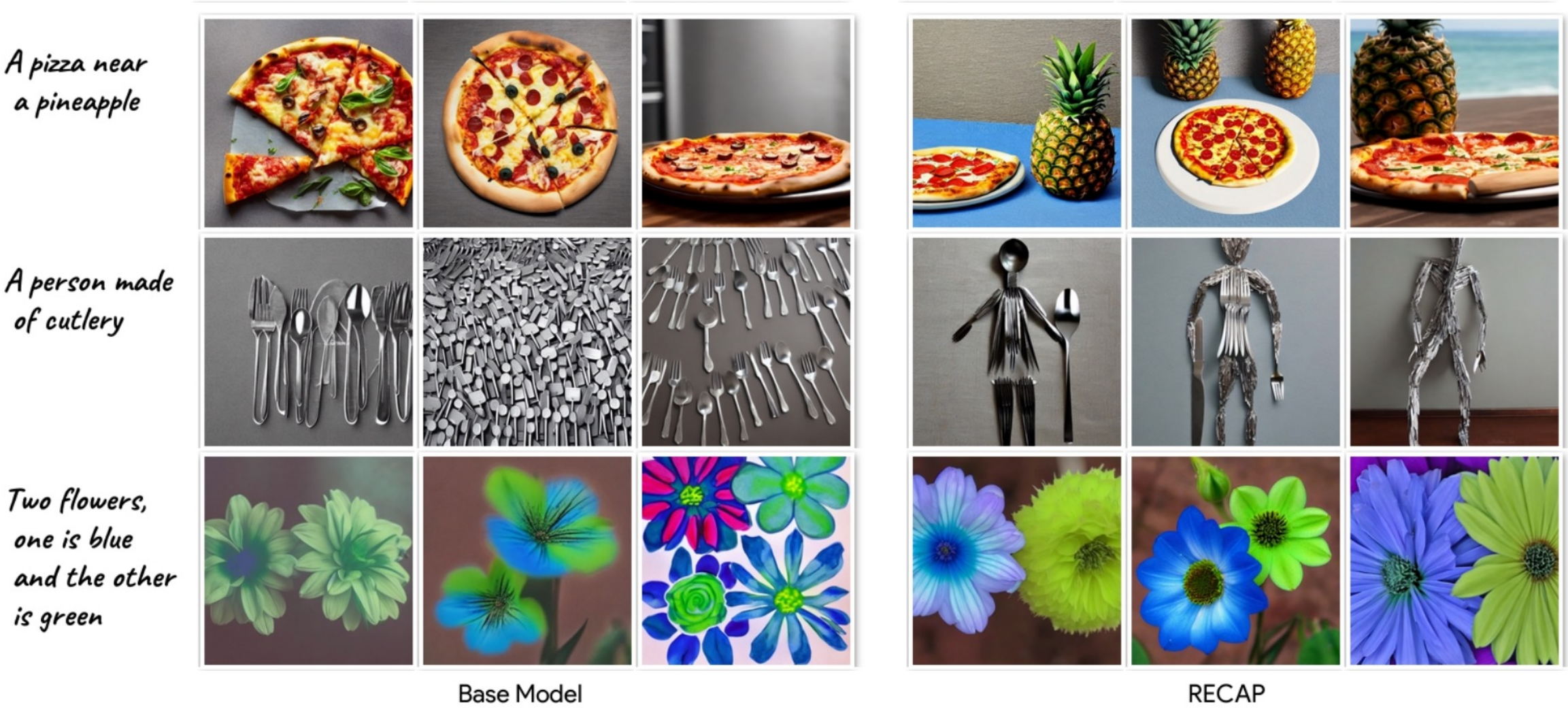
(given by human)

Short Caption	A blue willys gasser car	A glass of iced tea
Long Caption	A willys gasser car in blue color. it is placed near the car shed on the floor. in the back, there is a man seated in the chair	A glass of iced tea placed on a saucer decorated with mint leaves. it is located on the wood table

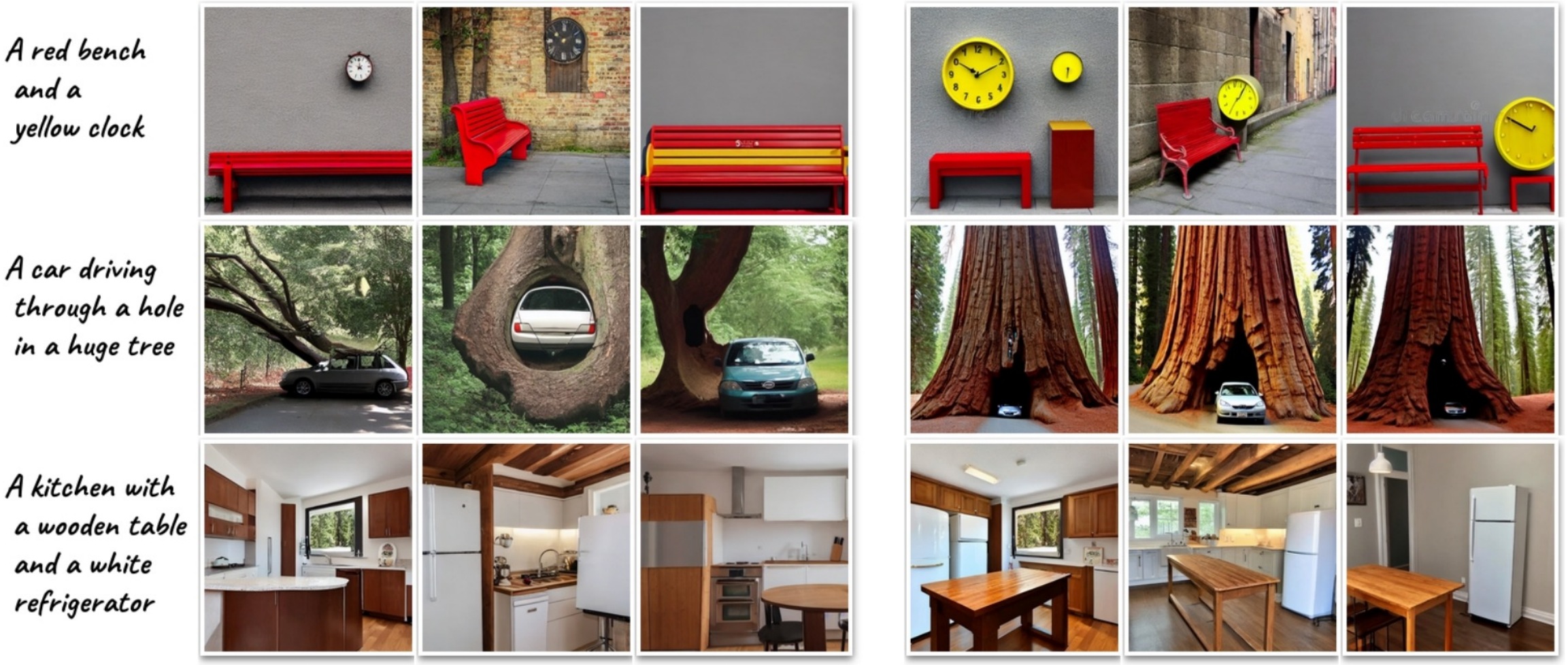
Alttext	2013 ducati monster 1100 evo diesel motorcycle photos and specifications. Black Bedroom Furniture Sets. Home Design Ideas	Cover of Oregon Wine Press February 2019
RECAP Short	A green benelli motorcycle	A poster of food served on plates
RECAP Long	A modern motorcycle with a combination of black, grey, and green colors. it is placed on a black background	A black plate filled with a variety of indian food items along with a glass of white wine. the caption reads "spice it up"

Alttext	You Want To Learn About What Human Anatomy The Skeletal System	GraphicRiver Toucan Pattern 7750027
RECAP Short	A girl making a human skeleton model	Floral and toucans seamless pattern
RECAP Long	A girl who is kneeling down and drawing a human skeleton model. the skeleton model is made of cardboard	A seamless vector pattern with toucan birds and tropical flowers. the background color is beige

Results


A pizza near a pineapple
A person made of cutlery
Two flowers, one is blue and the other is green
Base Model
RECAP

Results


A red bench and a yellow clock
A car driving through a hole in a huge tree
A kitchen with a wooden table and a white refrigerator
Base Model
RECAP

knowledge-based agents

agents that reason by operating on internal representations of knowledge

An example

Consider the following sentences

(1) If it didn't rain, Amy visited Bobby today.

(2) Amy visited Bobby or Carson today, but not both.

(3) Amy visited Carson today.

Can we answer the question “Did it rain today?”

- Amy did not visit Bobby today.
- It rained today.

Logic

sentence

an assertion about the world
in a knowledge representation language

Propositional Logic
(命題邏輯)

Proposition Symbols

used to represent a proposition

P Q R

Logical Connectives

$\neg$	$\wedge$	$\vee$
not	and	or
$\rightarrow$	$\leftrightarrow$	
implication	biconditional	

Not (¬)

Truth Table

P	$\neg P$
false	true
true	false

And ($\wedge$)

P	Q	$P \wedge Q$
false	false	false
false	true	false
true	false	false
true	true	true

Or ($\vee$)

P	Q	$P \vee Q$
false	false	false
false	true	true
true	false	true
true	true	true

Implication ($\rightarrow$)

P	Q	$P \rightarrow Q$
false	false	true
false	true	true
true	false	false
true	true	true

Implication represent a structure of “if P then Q”

For example:

if P: “It is raining” and Q: “I’m indoors”
then P→Q means “If it is raining, then I’m indoors.”

In the case of P implies Q, P is called the antecedent and Q is called the consequent.

When the antecedent is true, the whole implication is true in the case that the consequent is true

- i.e., if it is raining and I'm indoors, then the sentence "if it is raining, then I'm indoors" is true.

When the antecedent is true, the implication is false if the consequent is false

- i.e., if I'm outside while it is raining, then the sentence "If it is raining, then I'm indoors" is false.

However, when the antecedent is false, the implication is always true, regardless of the consequent.

- Logically, we can't learn anything from an implication (P → Q) if the antecedent (P) is false.
- Looking at our example, if it is not raining, the implication doesn't say anything about whether I'm indoors or not. I could be an indoors type and never walk outside, even when it is not raining, or I could be an outdoors type and be outside all the time when it is not raining.
- When the antecedent is false, we say that the implication is trivially true.

Biconditional ($\leftrightarrow$)

P	Q	$P \leftrightarrow Q$
false	false	true
false	true	false
true	false	false
true	true	true

model

assignment of a truth value to every propositional symbol (a "possible world")

The model is the truth-value assignment that provides information about the world

P: It is raining.

Q: It is a Tuesday.

A model $\{P = \text{true}, Q = \text{false}\}$

In fact, there are more possible models, say $2^2=4$ because we have two propositions.

knowledge base

a set of sentences known by a knowledge-based agent

Entailment (蘊涵)

if $\alpha \models \beta$ (α entails β)

then, in every model where sentence α is true, sentence β is also true.

For example, if α: “It is a Tuesday in January” and β: “It is January,” then we know that α ⊨ β.
If it is true that it is a Tuesday in January, we also know that it is January.

Entailment is different from implication.

- Implication is a logical connective between two propositions.
- Entailment, on the other hand, is a relation that means that if all the information in α is true, then all the information in β is true.

inference

the process of deriving new sentences from old ones

① If it didn't rain, Amy visited Bobby today.
② Amy visited Bobby or Carson today, but not both.
③ Amy visited Carson today.

④ Amy did not visit Bobby today.
⑤ It rained today.

(4)(5) are inferred from (1)(2)(3).

Does

$\mathrm{KB} \models \alpha$

?

Model Checking

A way to infer new knowledge
based on existing knowledge

Model Checking

- To determine if $\text{KB} \models \alpha$ (answering the question: "can we conclude that α is true based on our knowledge base"):
  - Enumerate all possible models.
  - If in every model where KB is true, α is true, then KB entails α.
  - Otherwise, KB does not entail α.

P: It is a Tuesday. Q: It is raining. R: Harry will go for a run.

KB: $(P \wedge \neg Q) \rightarrow R$ P (P is true) $\neg Q$ (Q is false)

Query: R First, we enumerate all possible models.

P	Q	R	KB
false	false	false	
false	false	true	
false	true	false	
false	true	true	
true	false	false	
true	false	true	
true	true	false	
true	true	true	

KB: $(P \wedge \neg Q) \rightarrow R$

Next, we go through every model and check whether it is true given our knowledge base.

P	Q	R	KB
false	false	false	false
false	false	true	false
false	true	false	false
false	true	true	false
true	false	false	
true	false	true	
true	true	false	
true	true	true	

KB: $(P \wedge \neg Q) \rightarrow R$

Next, we go through every model and check whether it is true given our knowledge base.

P	Q	R	KB
false	false	false	false
false	false	true	false
false	true	false	false
false	true	true	false
true	false	false	
true	false	true	
true	true	false	false
true	true	true	false

KB: $(P \wedge \neg Q) \rightarrow R$

Next, we go through every model and check whether it is true given our knowledge base.

P	Q	R	KB
false	false	false	false
false	false	true	false
false	true	false	false
false	true	true	false
true	false	false	false
true	false	true	true

Due to $(P \wedge \neg Q) \rightarrow R$ being in the KB, we know that in the case where P is true and Q is false, R must be true. Thus, we say that our KB is false for the model where R is false and true for the model where R is true.

- Looking at this table, there is only one model where our knowledge base is true.
- In this model, we see that R is also true.
- By the definition of entailment, if R is true in all models where the KB is true, then $\mathrm{KB} \models R$.

P	Q	R	KB
false	false	false	false
false	false	true	false
false	true	false	false
false	true	true	false
true	false	false	false
true	false	true	true
true	true	false	false
true	true	true	false

A Python Code Example

Let's look at how knowledge and logic can be represented as code.

```
from logic import *
# Create new classes, each having a name, or a symbol, representing each proposition.
rain = Symbol("rain") # It is raining.
bobby = Symbol("bobby") # Amy visited Bobby
carson = Symbol("carson") # Amy visited Carson

# Save sentences into the KB
knowledge = And( # Starting from the "And" logical connective, because each proposition represents knowledge that we know to be true.
    Implication(Not(rain), bobby), # ¬(It is raining) → (Amy visited Bobby)
    Or(bobby, carson), # (Amy visited Bobby) ∨ (Amy visited Carson).
    Not(And(bobby, carson)), # ¬(Amy visited Bobby ∧ Amy visited Carson) i.e. Amy did not visit both Bobby and Carson.
    carson # Amy visited Carson. Note that while previous propositions contained multiple symbols with connectors, this is a proposition consisting of one symbol. This means that we take as a fact that, in this KB, Amy visited Carson.)
```

```
Python 3.5.2 (default, Nov 17 2016, 17:05:23)
[GCC 5.4.0 20160609] on linux
Type "copyright", "credits" or "license()" for more information.
>>>
========== RESTART: /home/andrew/workspace/dcaa/logic.py ==========
>>>
```

恆真式 (Tautology)

偶然性 (Contingency)

https://www.youtube.com/watch?v=D9pkdnIrjdM

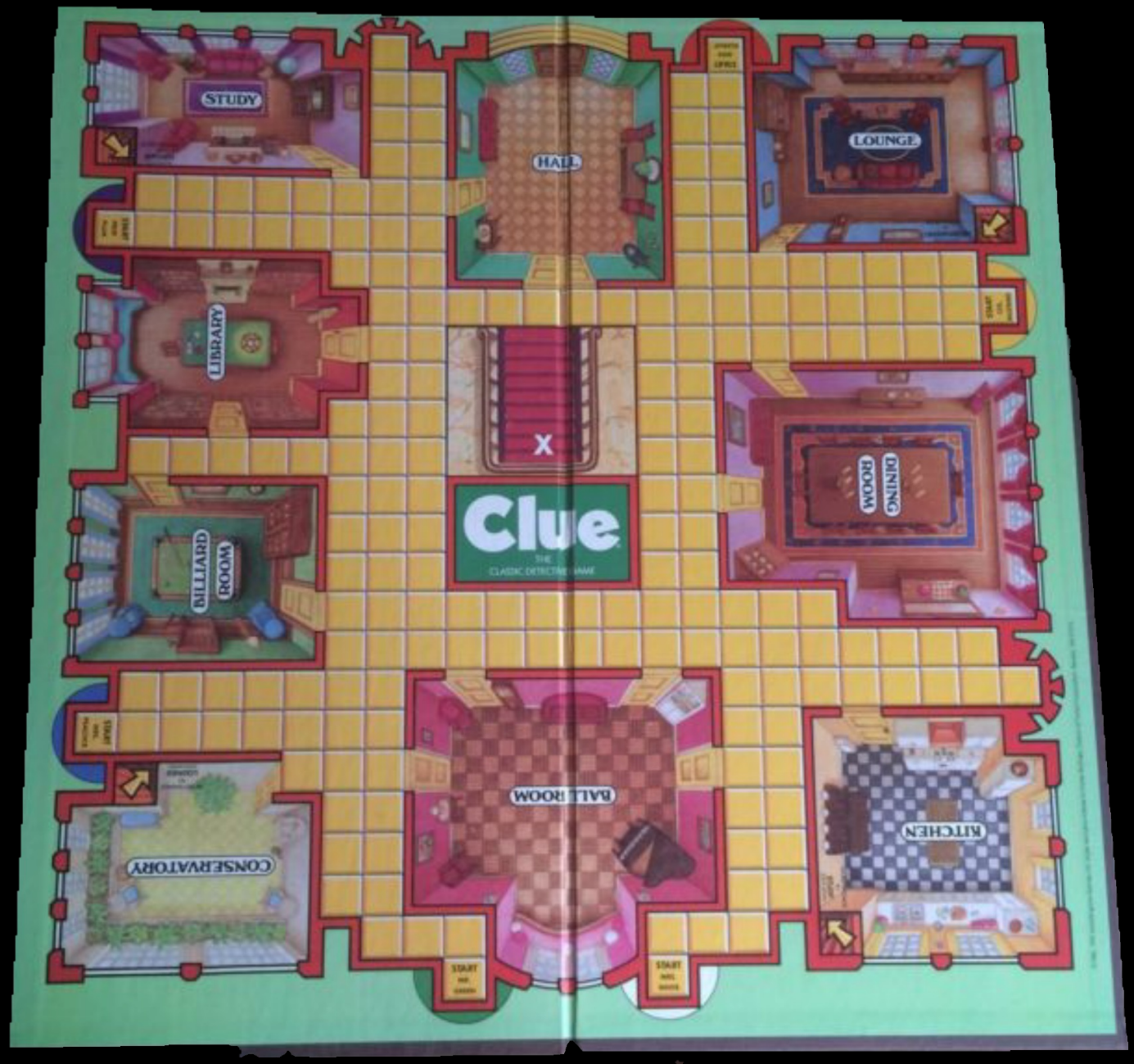
Knowledge Engineering

the process of figuring out how to represent propositions and logic in AI

Clue

In the game, a murder was committed
by a person
using a tool
in a location.

People, tools, locations are represented by cards.



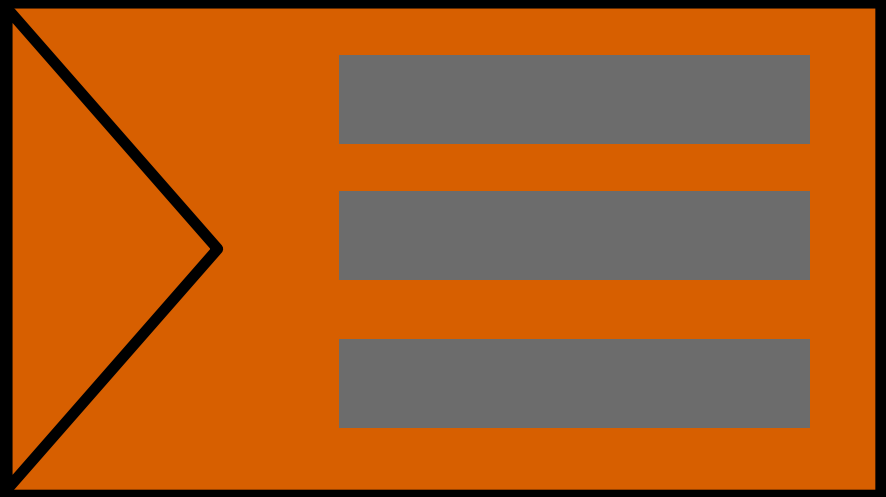
Clue

People	Rooms	Weapons
Col. Mustard	Ballroom	Knife
Prof. Plum	Kitchen	Revolver
Ms. Scarlet	Library	Wrench

People

Rooms

Weapons



One card of each category is picked at random and put it in an envelope, and it is up to the participants to uncover whodunnit.

https://www.youtube.com/watch?v=70Os18oRzMg

Clue

Propositional Symbols

mustard	*ballroom*	*knife*
plum	*kitchen*	*revolver*
scarlet	*library*	*wrench*

Clue

We can start creating our knowledge base by adding the rules of the same.

$$(mustard \vee plum \vee scarlet)$$

$$(ballroom \vee kitchen \vee library)$$

$$(knife \vee revolver \vee wrench)$$

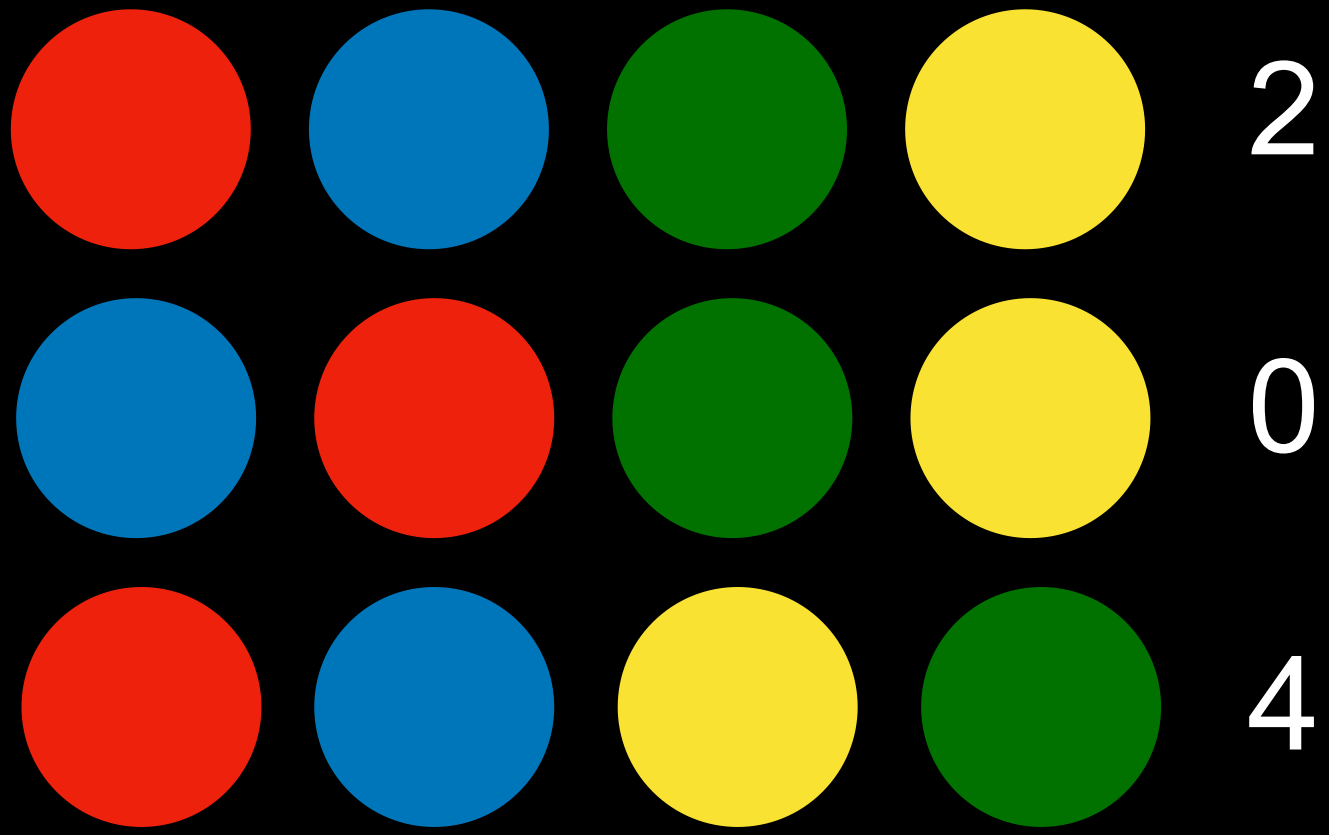
(hints obtained during the course of the game)

$$\neg plum$$

$$\neg mustard \vee \neg library \vee \neg revolver$$

Mastermind

Another type of puzzle that can be solved using propositional logic is a Mastermind game. In this game, player one arranges colors in a certain order, and then player two has to guess this order. Each turn, player two makes a guess, and player one gives back a number, indicating how many colors player two got right. Let's simulate a game with four colors.



Knowledge-Driven Vision-Language Encoding

JUNE 18-22, 2023

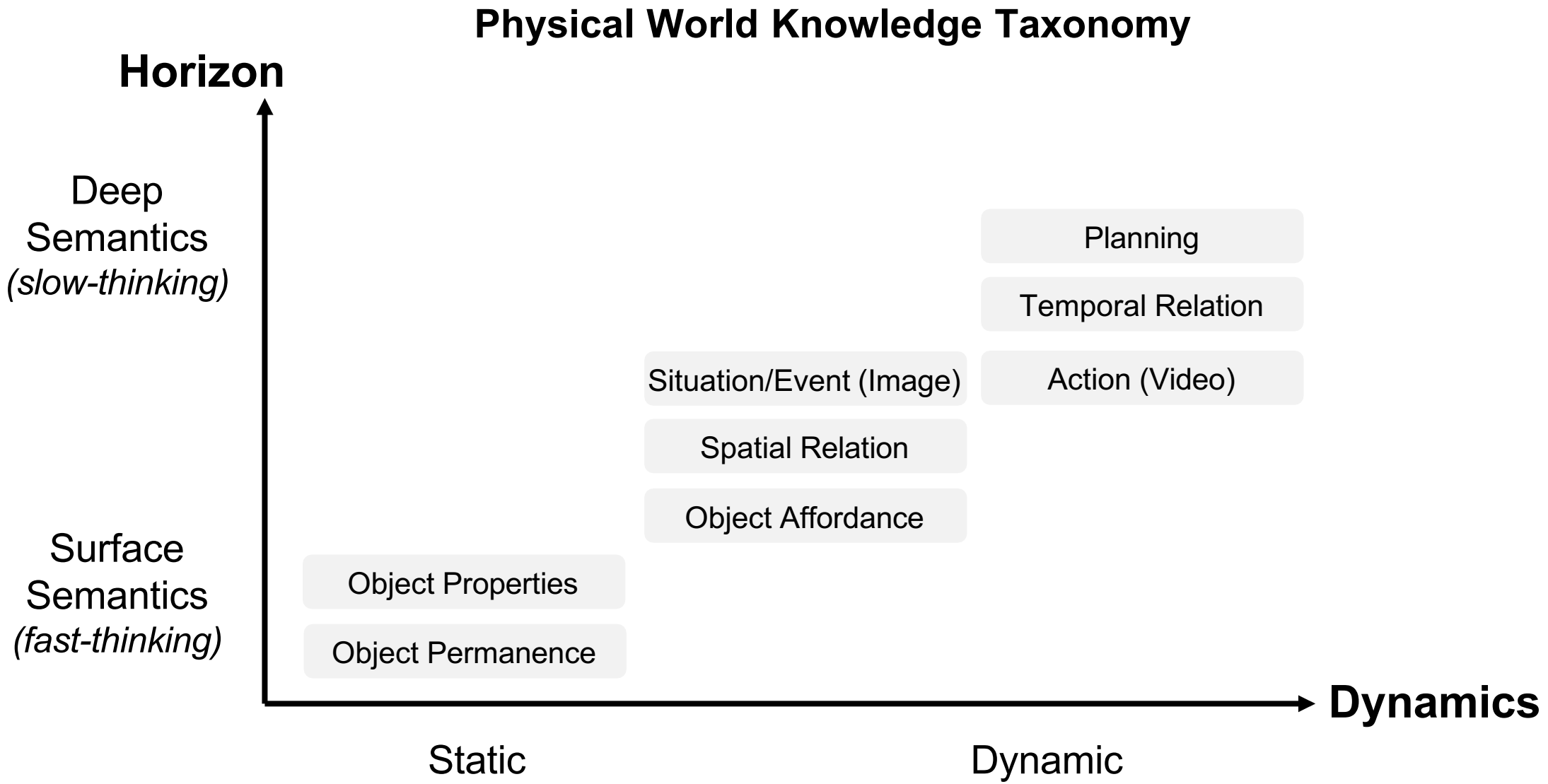
CVPR

VANCOUVER, CANADA

Tutorial

Slide credits:
Manling Li
Xudong Lin
Jie Lei

What is the knowledge that humans aim to learn?



Worked with Zhenhailong Wang

Machines learn knowledge through multi-sensory interactions

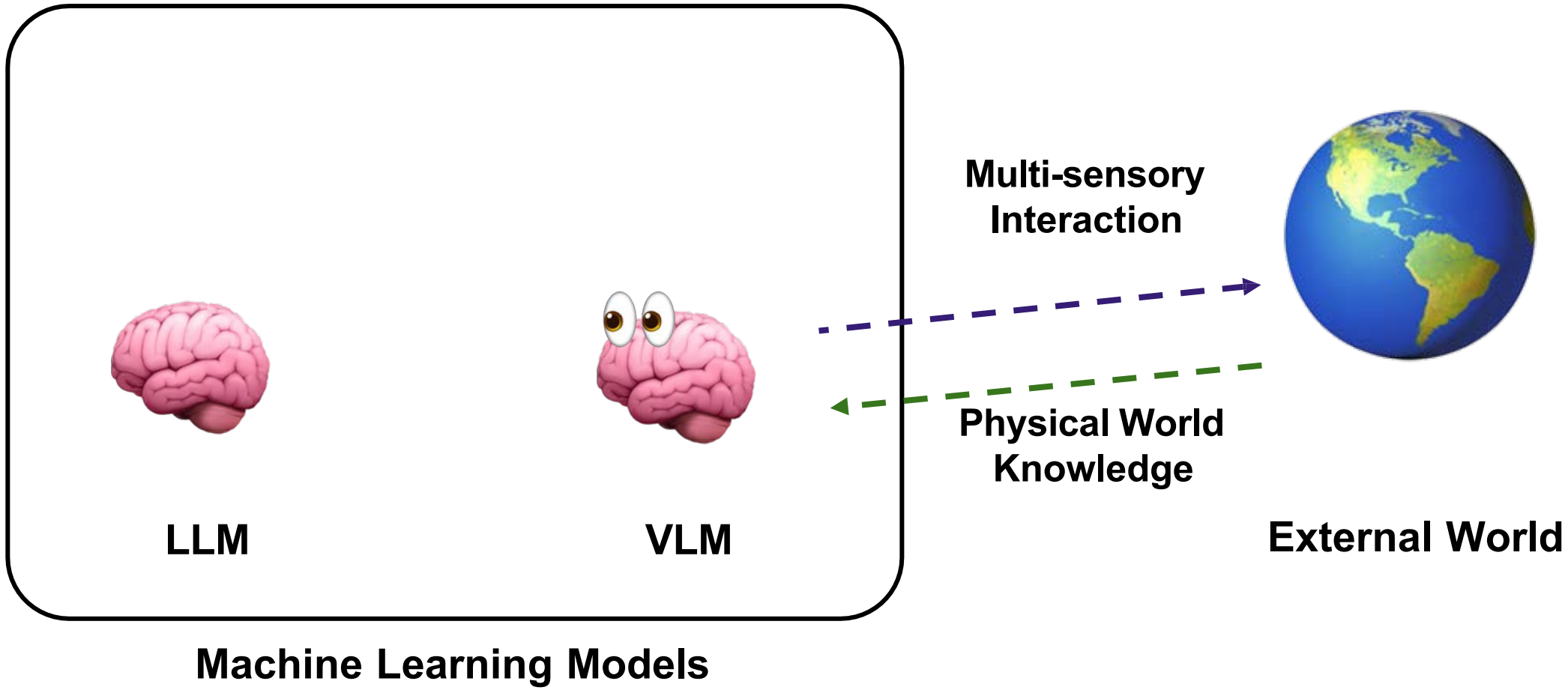


Figure Credits: Zhenhailong Wang

Machines learn knowledge through multi-sensory interactions

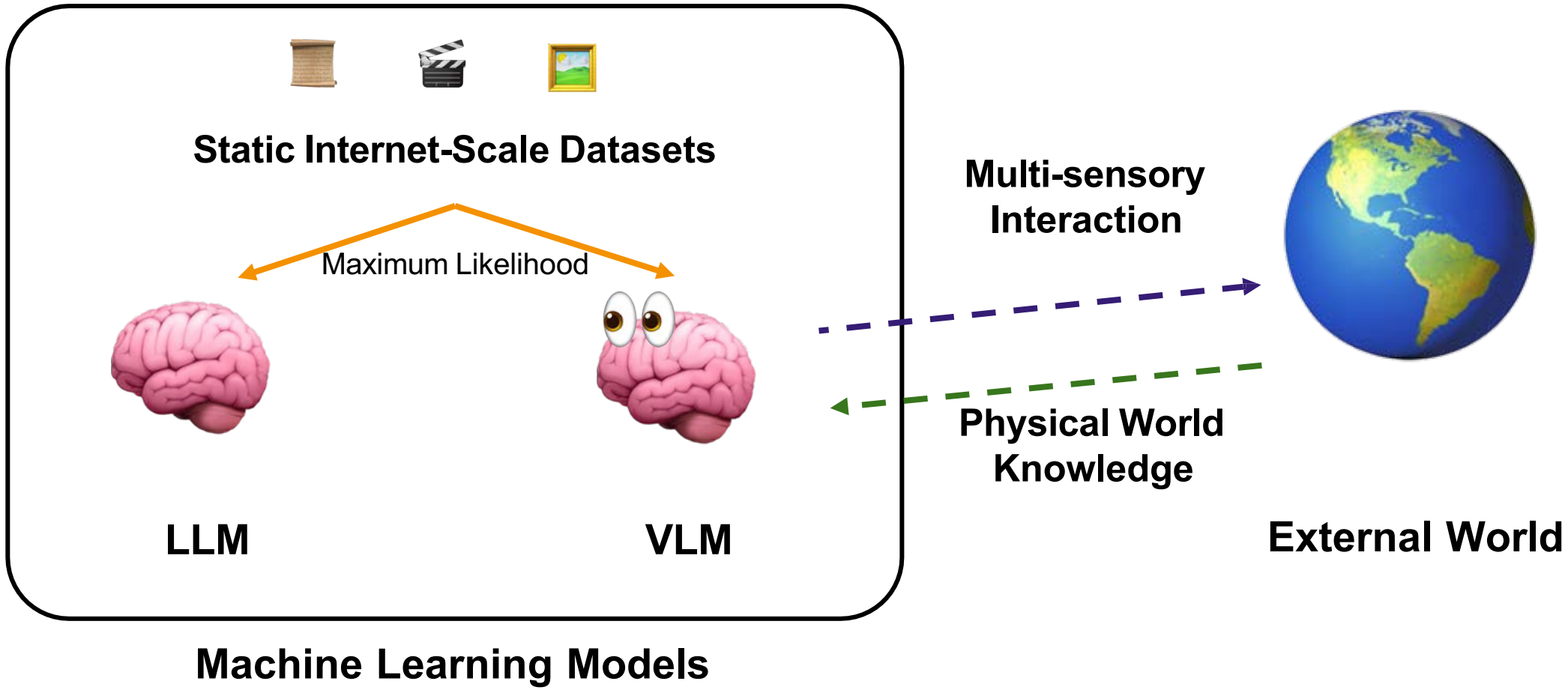
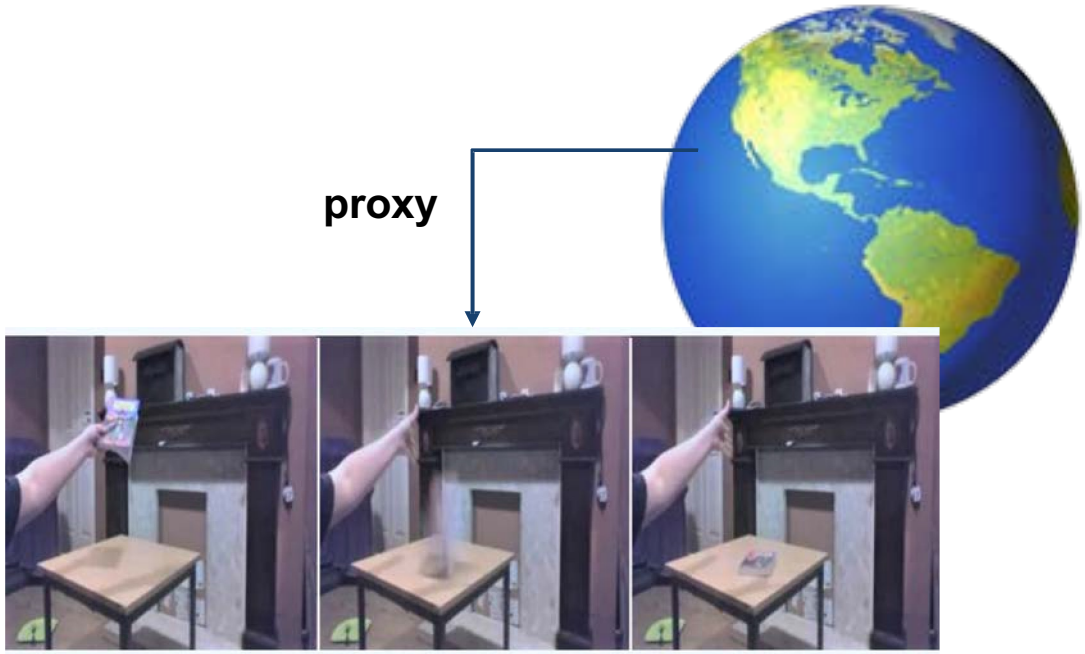


Figure Credits: Zhenhailong Wang

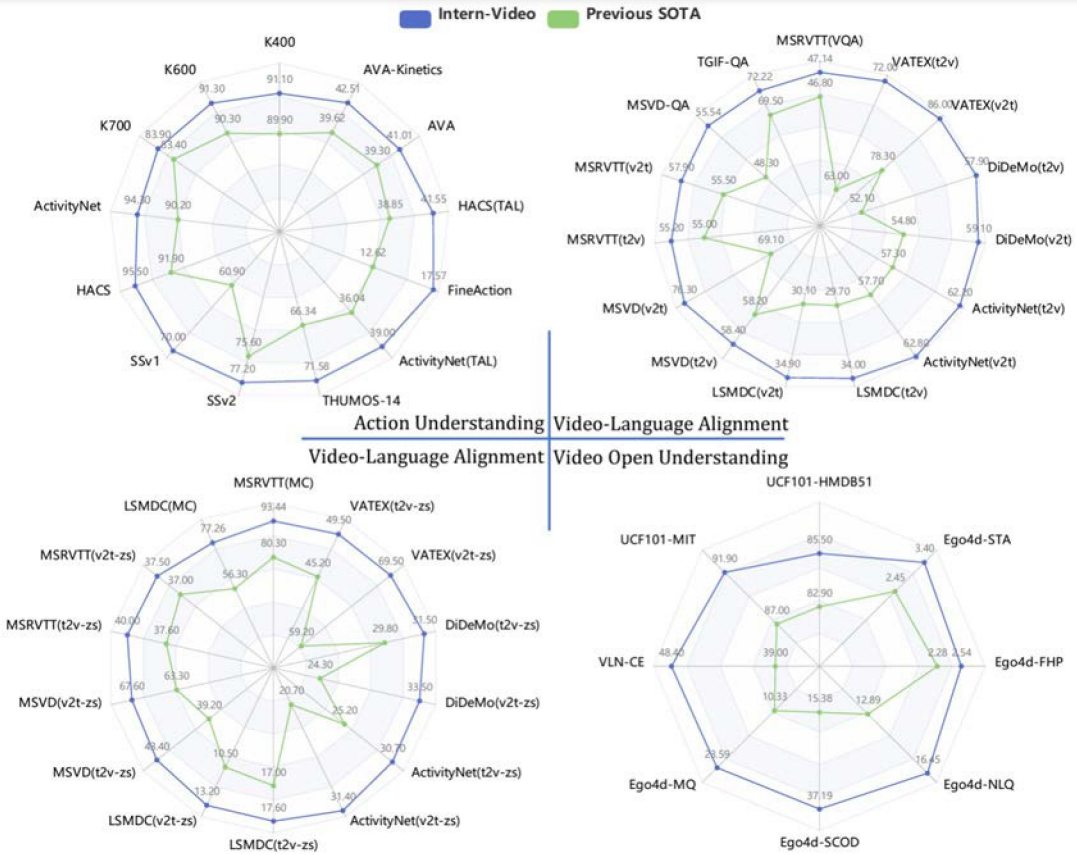
Video: A "Visual Recording" of World State Changes

Where can we learn such physical world knowledge without interactive data $ $ $?



"Book falling like a rock"

Video-Language Datasets



InternVideo (2023)

Video-Language Foundation Models

Figure Credits: Zhenhailong Wang

Existing V+L Models

Current models rely on object-centric abilities as a **shortcut** for V+L understanding.

Model	Verb Accuracy
MMT	60.8
Merged–MMT	60.7
Lang–MMT	64.5
Image–MMT	59.7

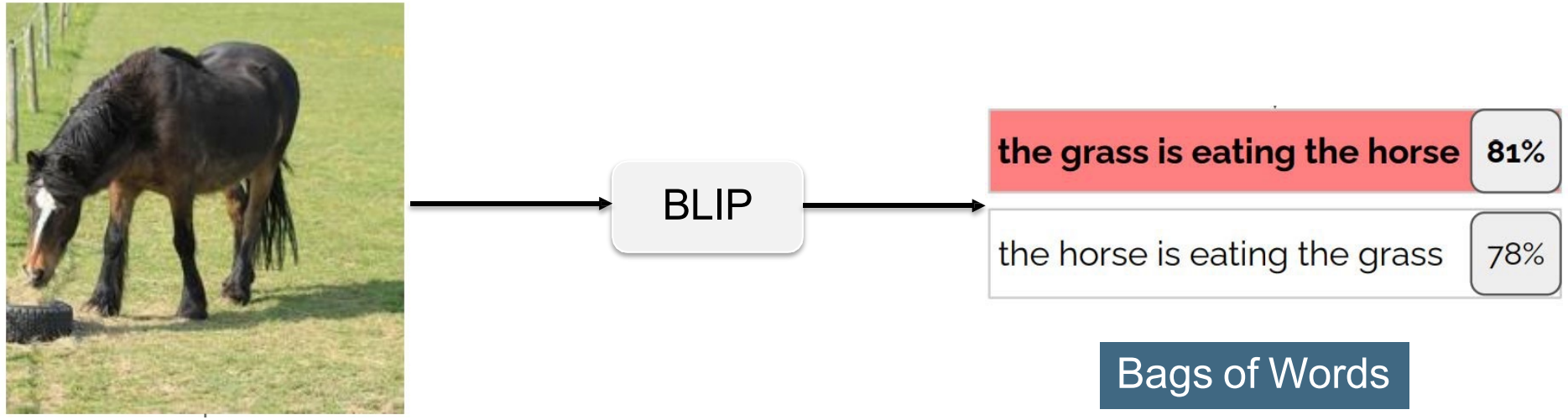
Low Verb Performance

"Probing Image–Language Transformers for Verb Understanding" Lisa Hendricks, et al. (arXiv 2021)

Figure Credits: Zhenhailong Wang

Existing V+L Models

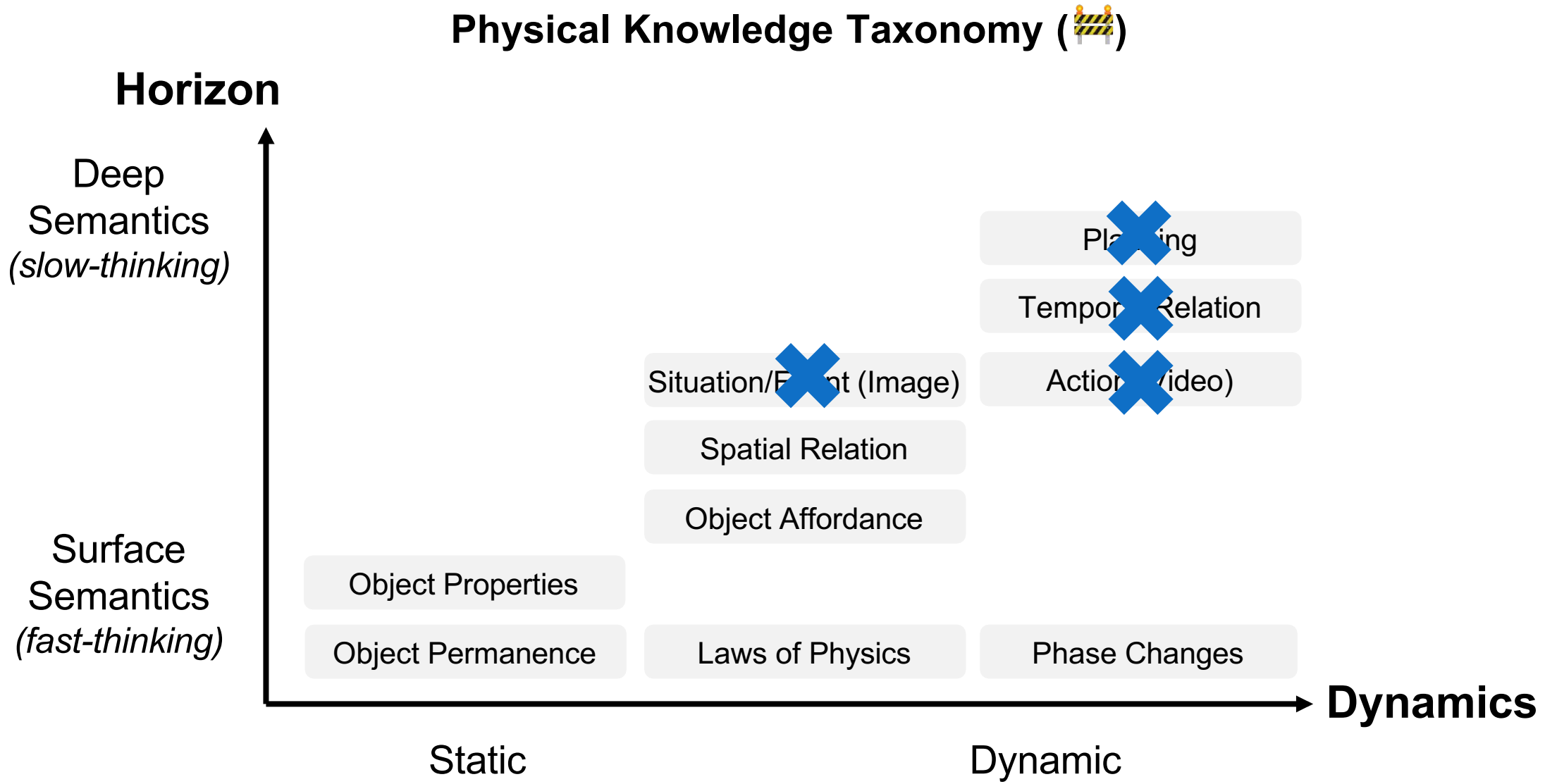
Current models rely on object-centric abilities as a **shortcut** for V+L understanding.



"When and why vision-language models behave like bags-of-words, and what to do about it" Mert Yuksekgonul, et al. (ICLR 2023)

Figure Credits: Zhenhailong Wang

Existing V+L Models




Surface
Deep


Surface
Deep
Object-Centric
Local
Static
Event-Centric
Situational
Dynamic

Knowledge can also help with V+L Pretraining

Compared to raw data, knowledge is **important and useful information.**

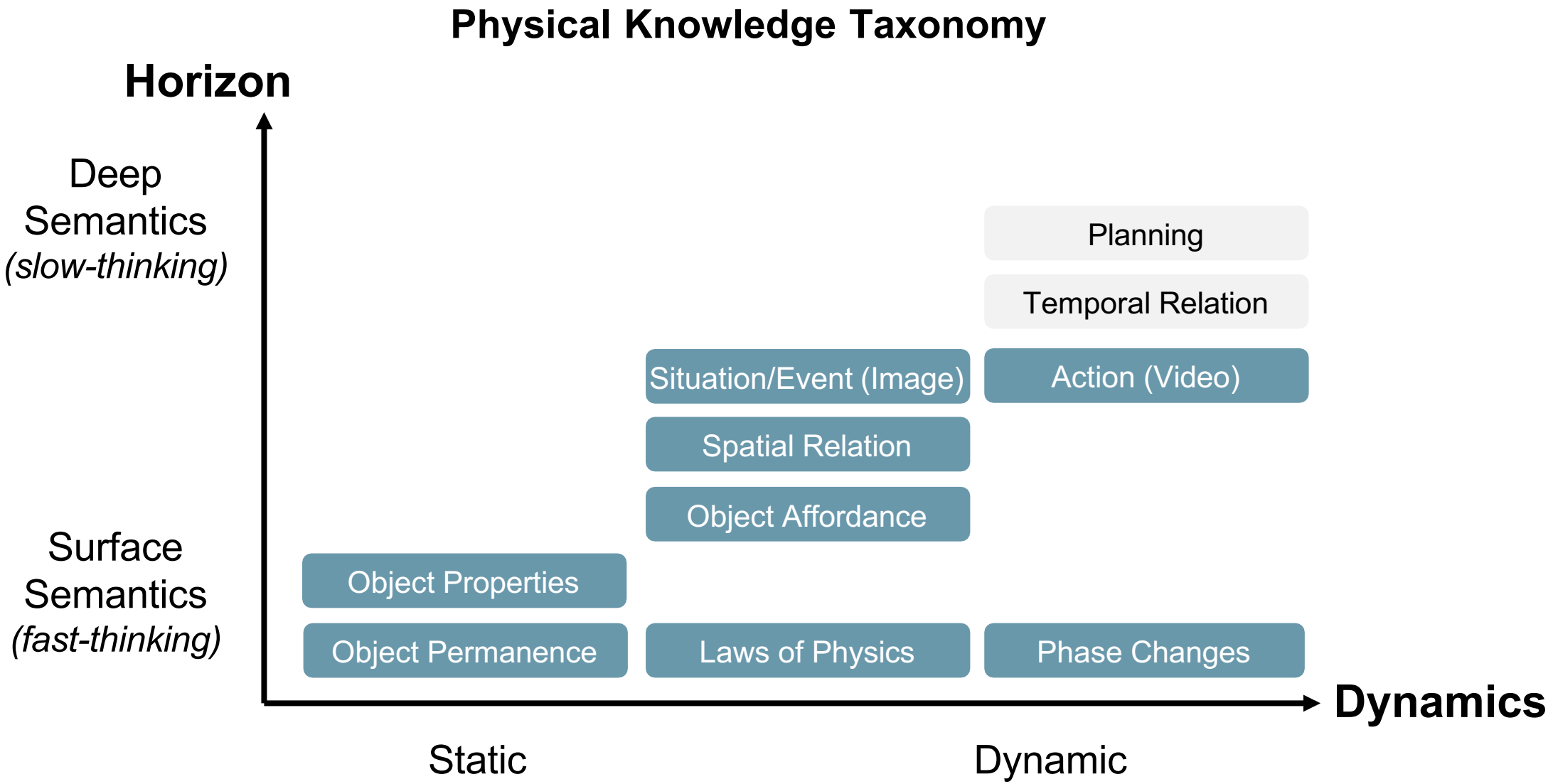
We learn three types of knowledge

Factual Knowledge are information about **instance-level facts** extracted from raw data.

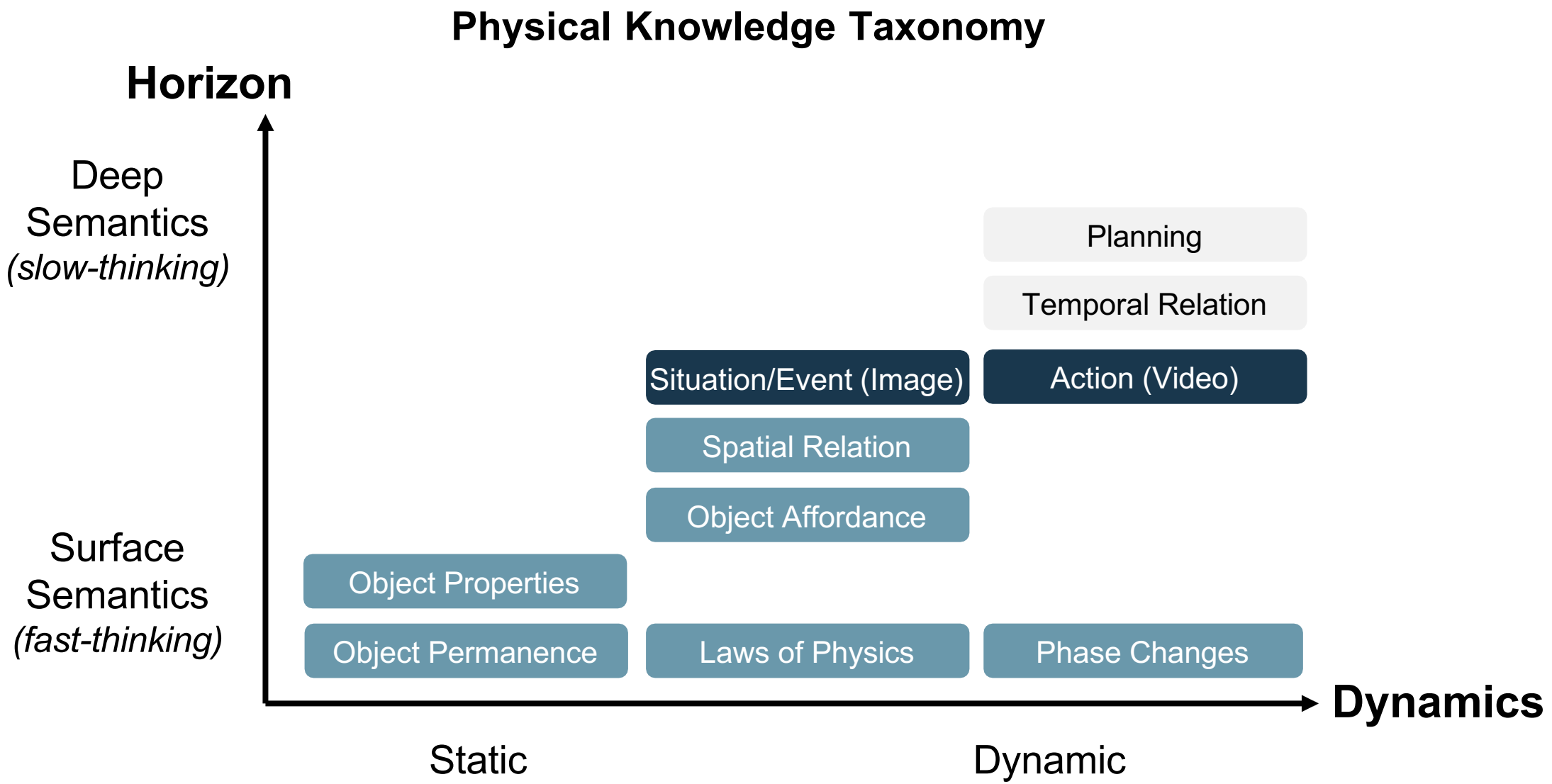
Text	Vision
Entity	Object
Relation	Scene Graph
Event	Activity/Situation
Affordance	Embodied AI

Factual Knowledge

Goal: Surface ➔ Deep Semantic Knowledge



Goal: Surface ➔ Deep Semantic Knowledge

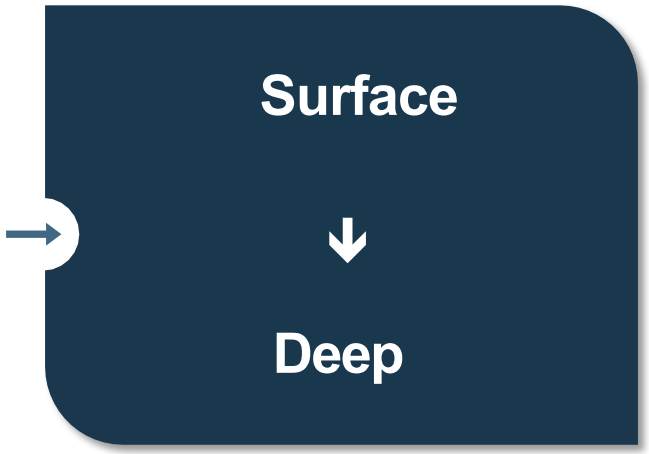


Challenge 1: Complex Situation

Complex Situation

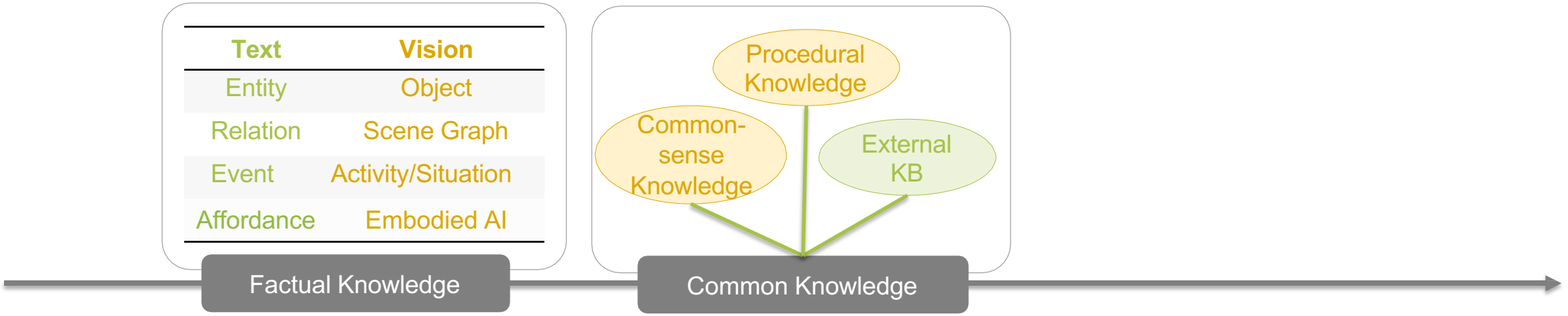
Event / Action

Semantics

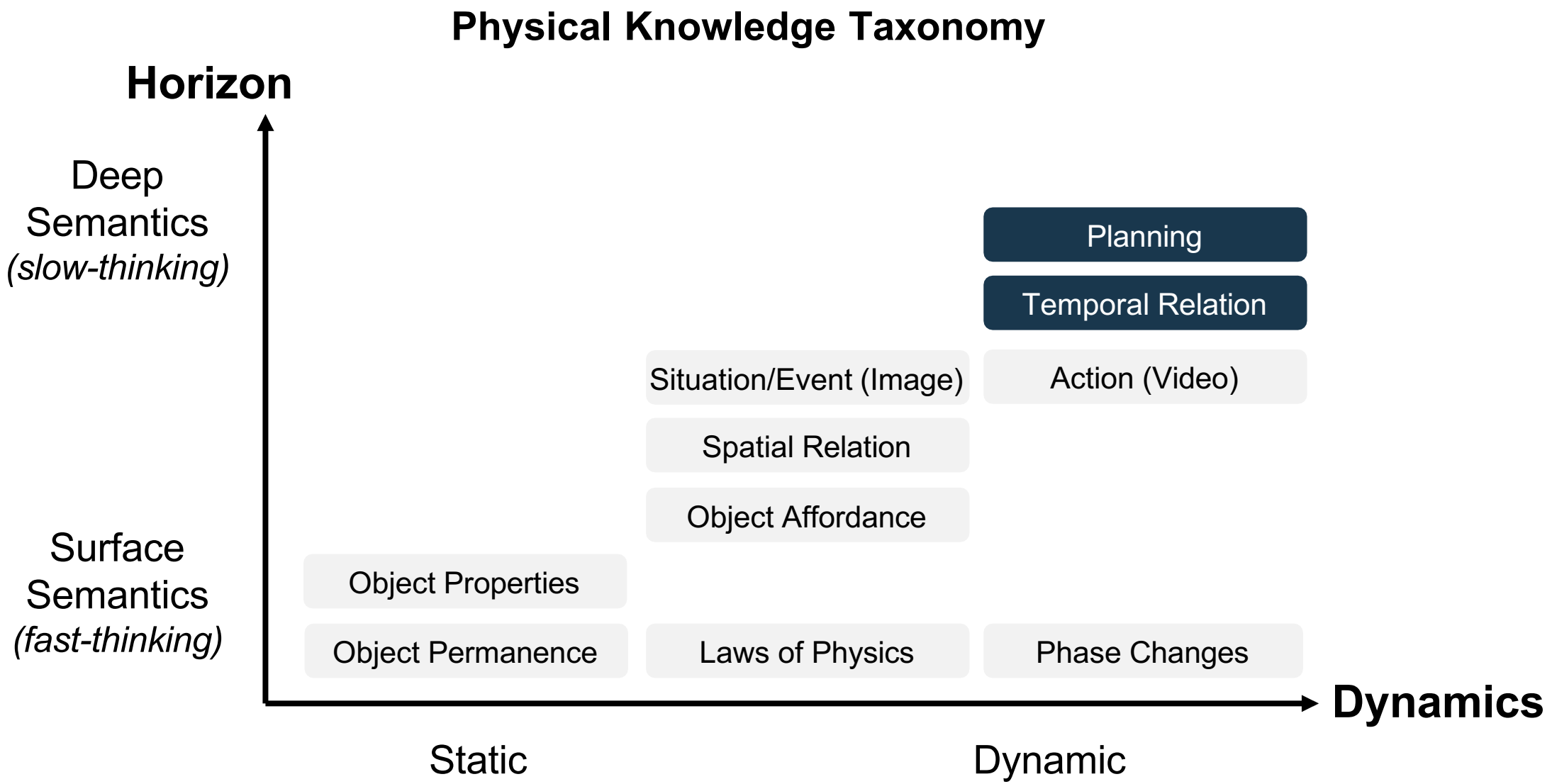


We learn three types of knowledge

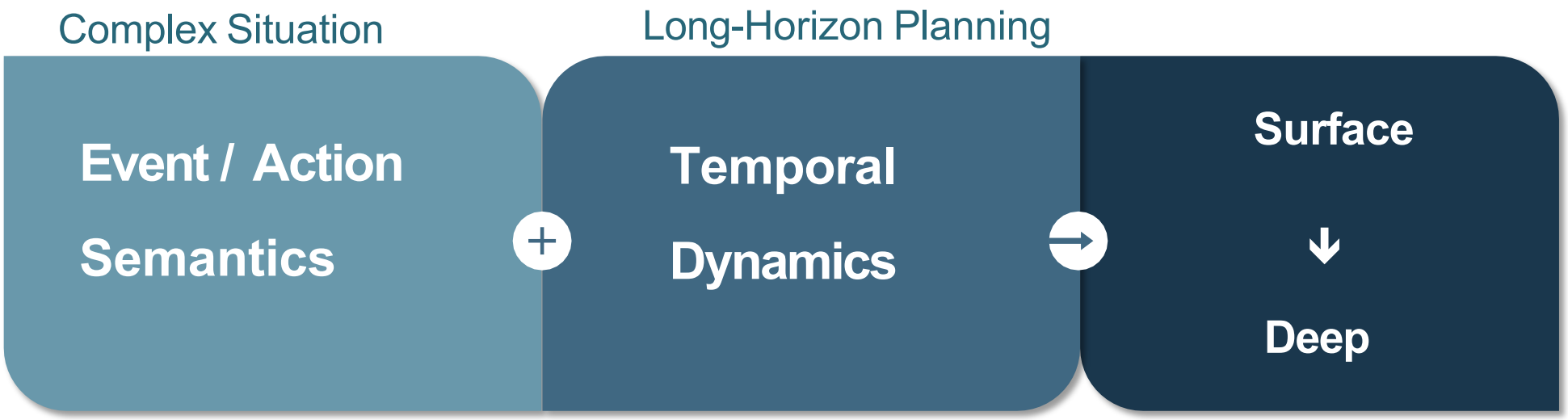
Common knowledge refers to knowledge of **common patterns** that is acquired or summarized from historical interaction with the world.



Goal: Surface ➔ Deep Semantic Knowledge

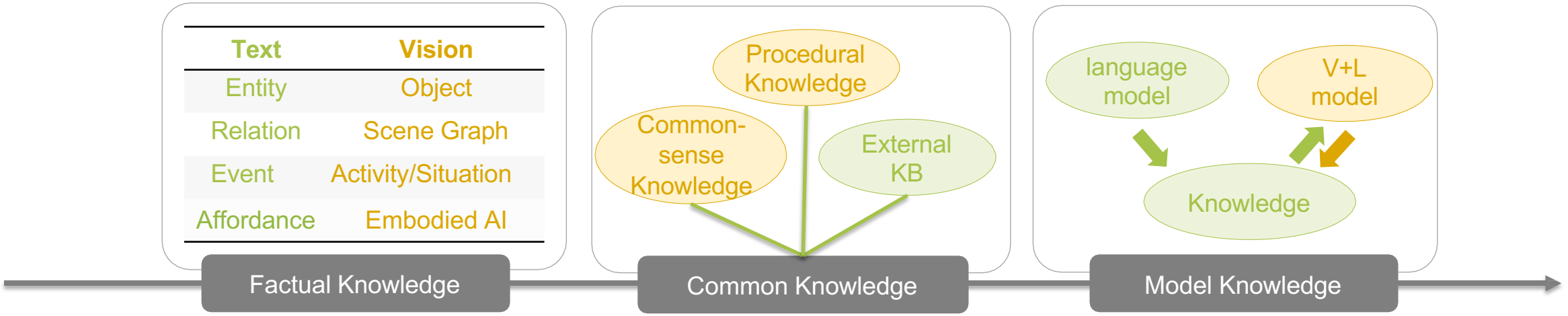


Challenge 2: Long-Horizon Planning



We learn three types of knowledge

Model Knowledge (parametric knowledge) is the knowledge embedded and encoded in models.

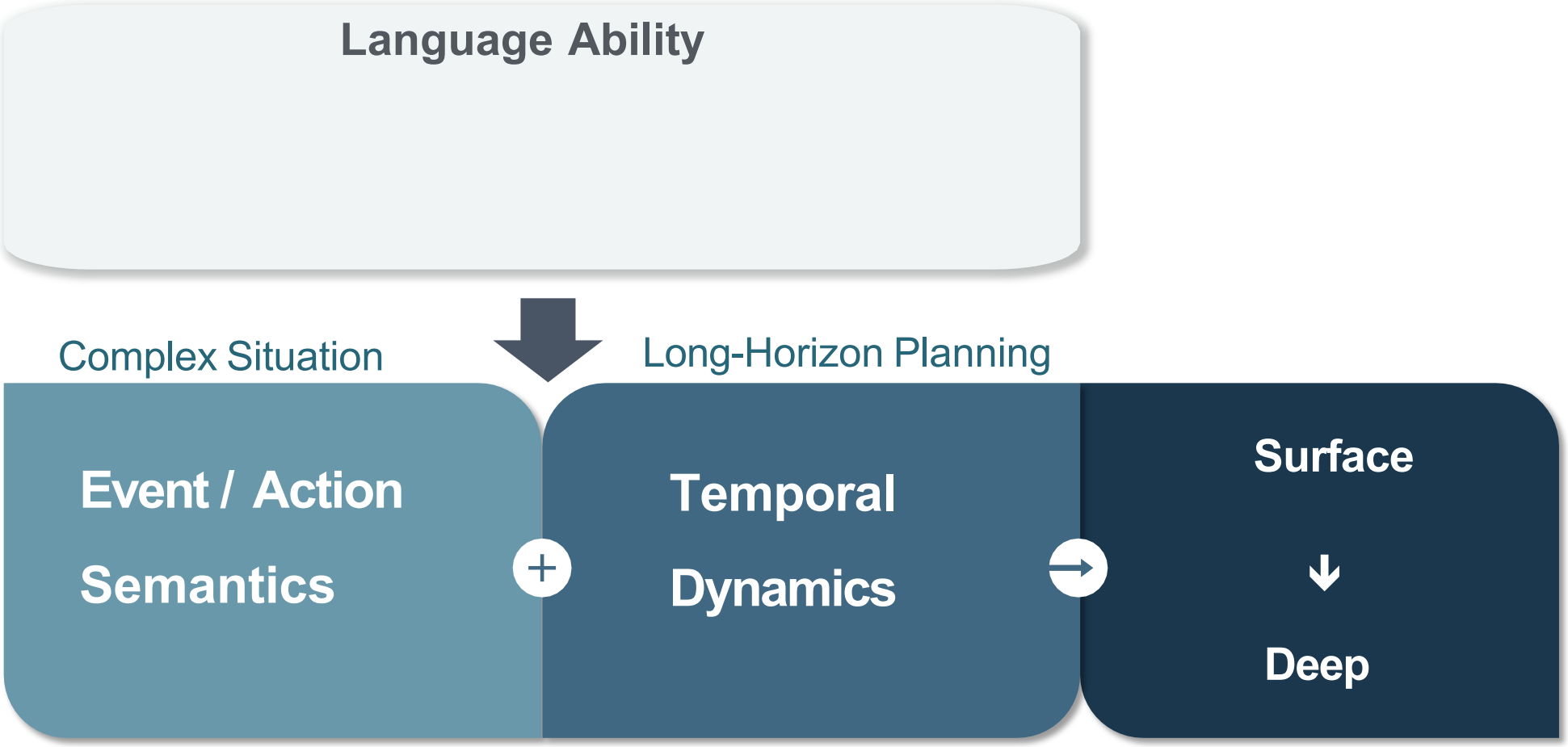


Can we borrow the ability from LLM?

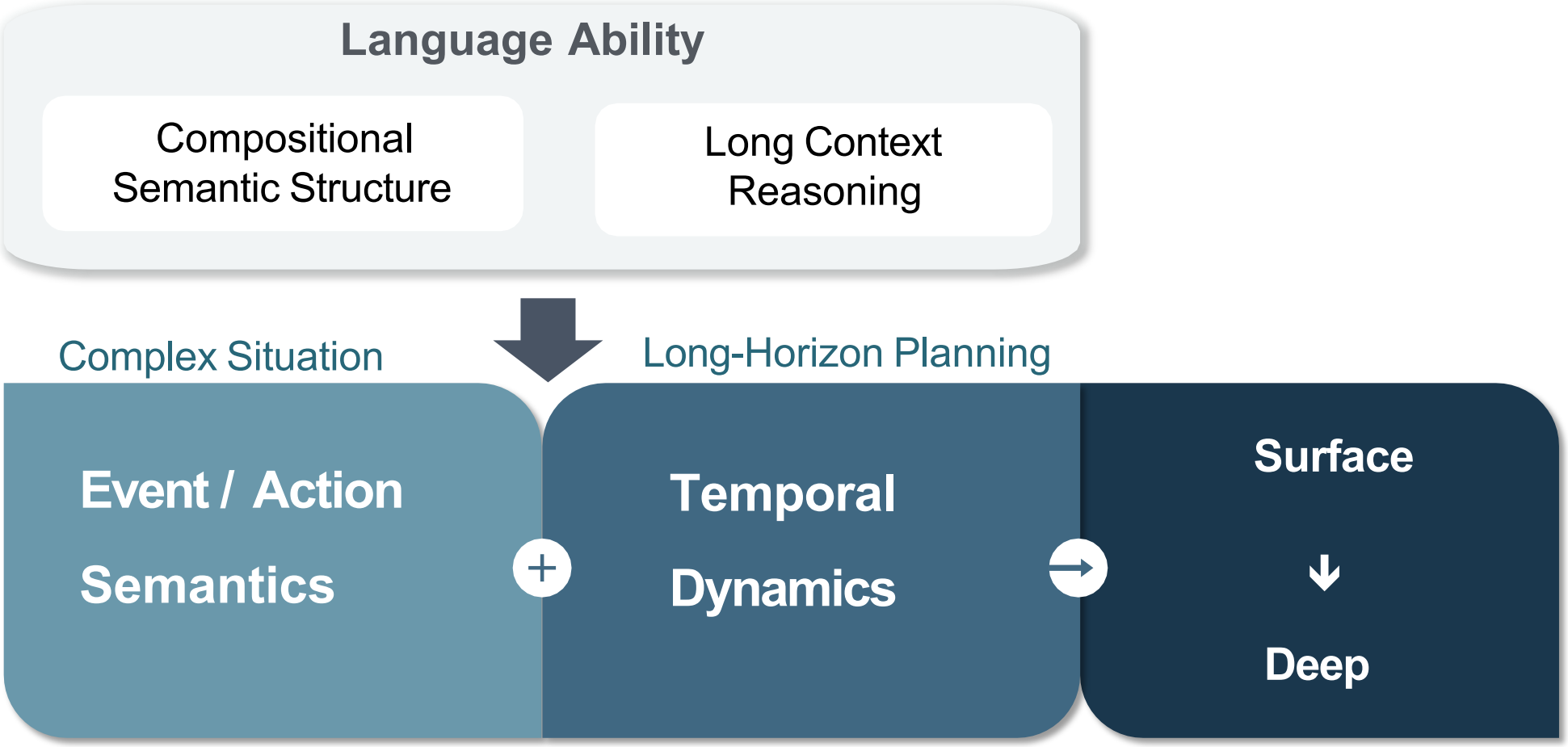
Large Language Models (LLMs) are very powerful.

https://twitter.com/xiye_nlp

One Solution: Language As Supervision

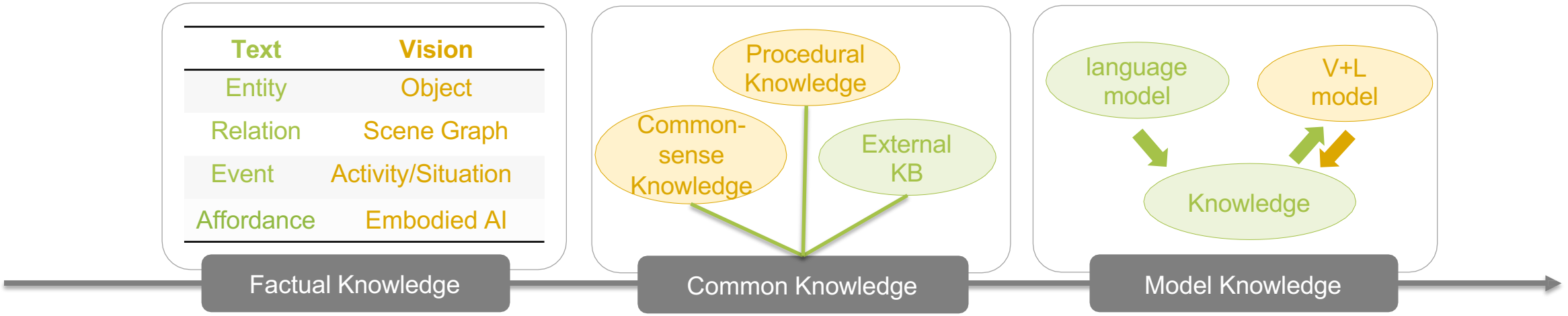


One Solution: Language As Supervision



We inject knowledge to V+L foundation models

We patch three types of knowledge into V+L foundation models.



Outline: Factual Knowledge

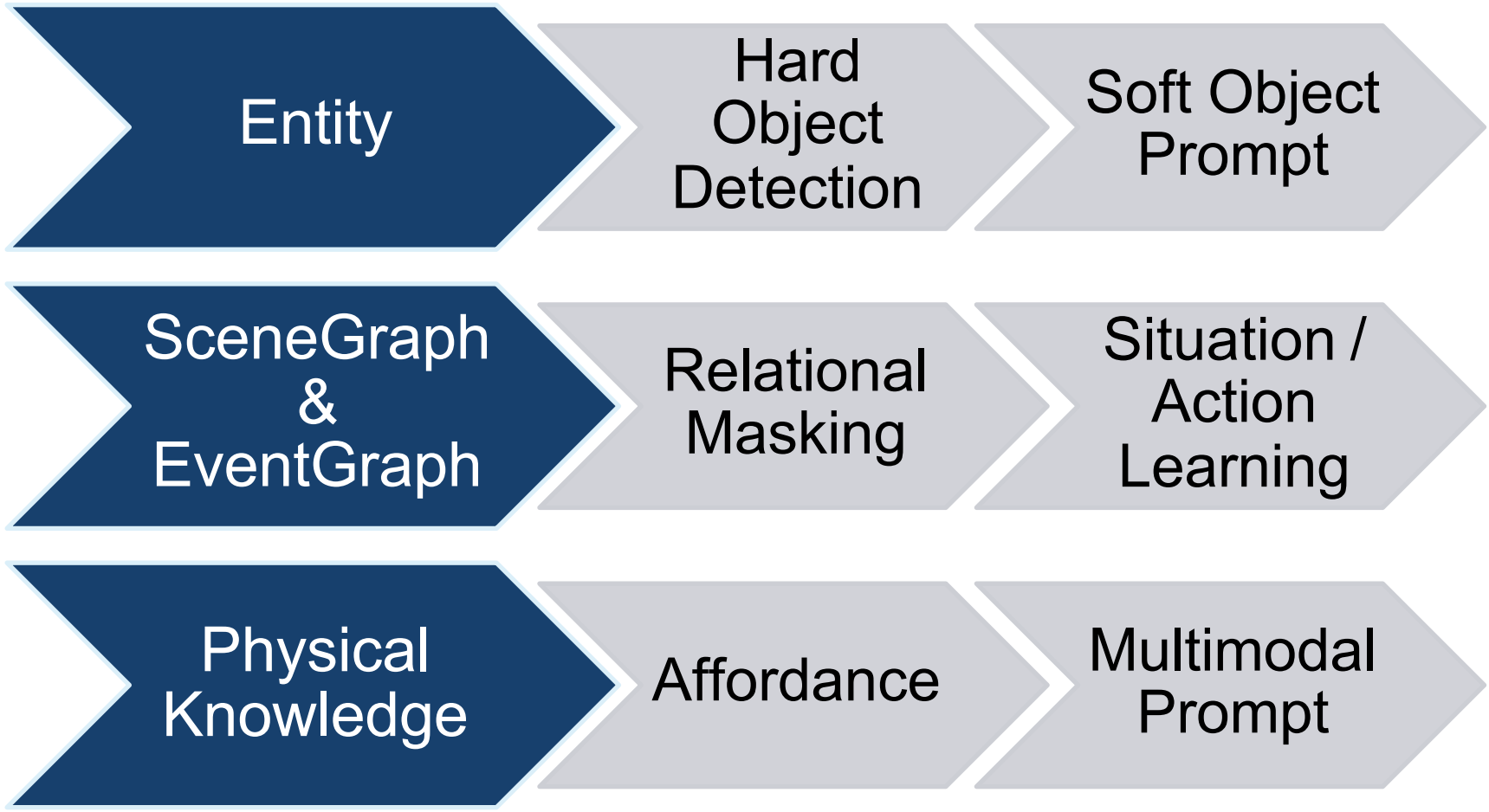
Factual Knowledge are information about instances extracted from raw data.

Text	Vision
Entity	Object
Relation	Scene Graph
Event	Activity/Situation
Affordance	Embodied AI

Factual Knowledge

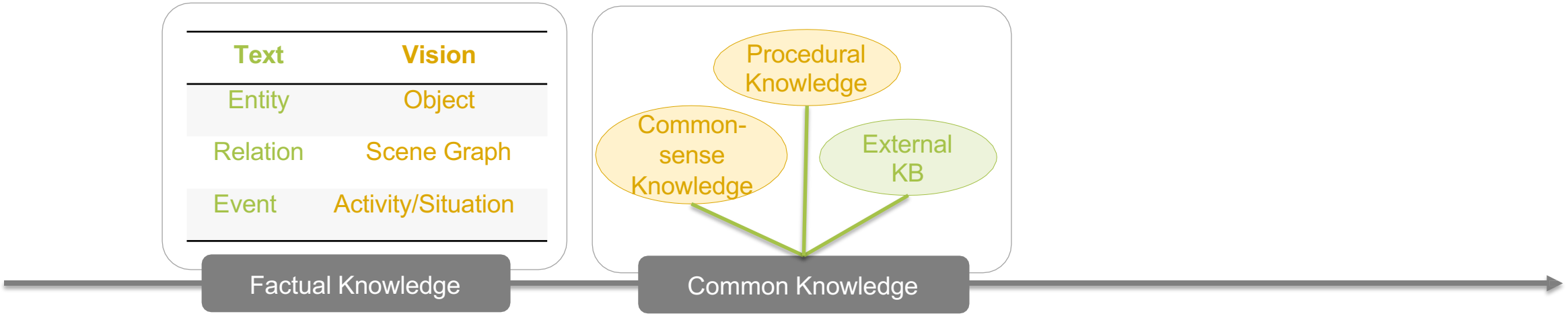
Implicit Knowledge

Outline: Factual Knowledge

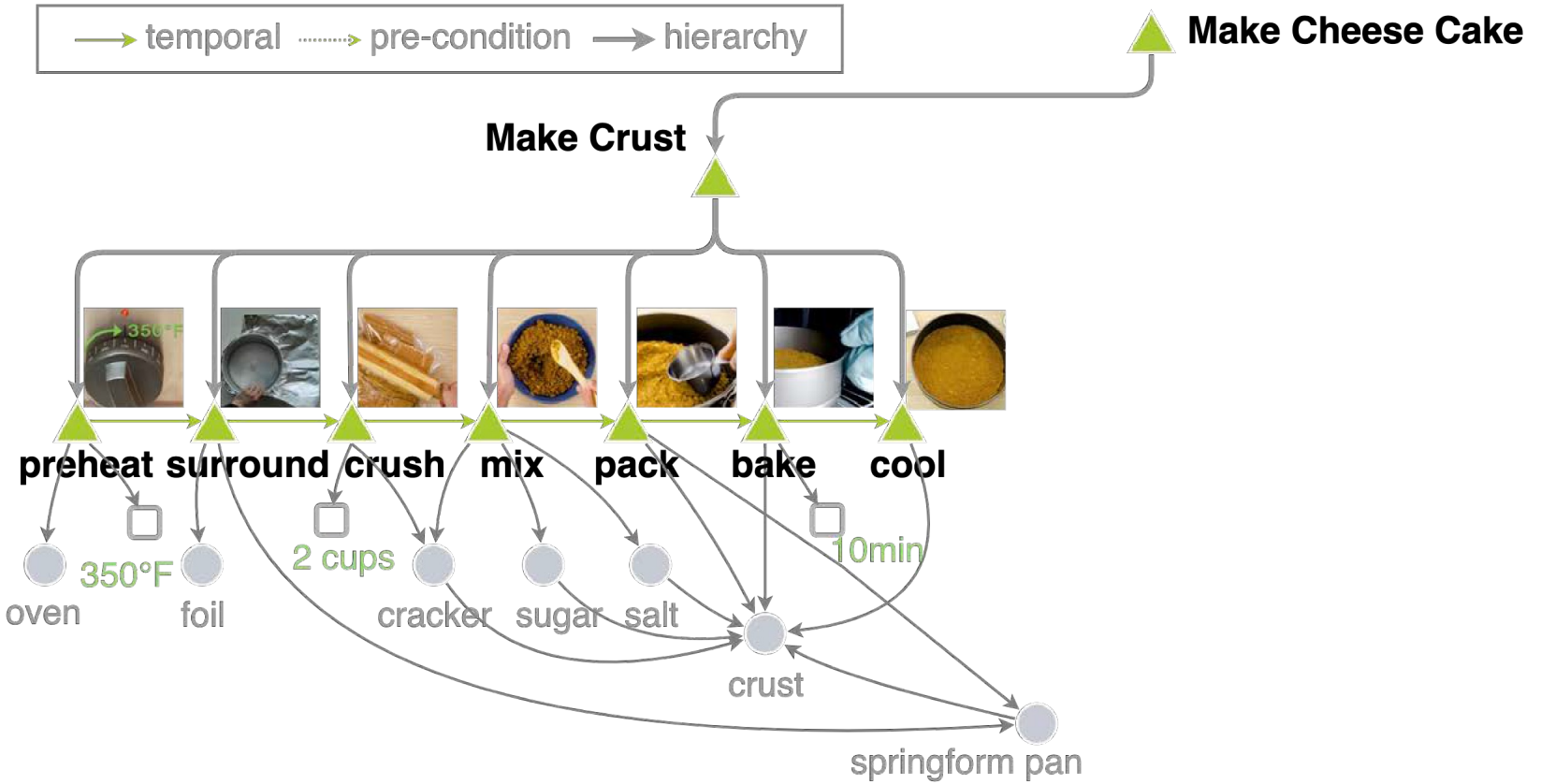


Outline: Common Knowledge

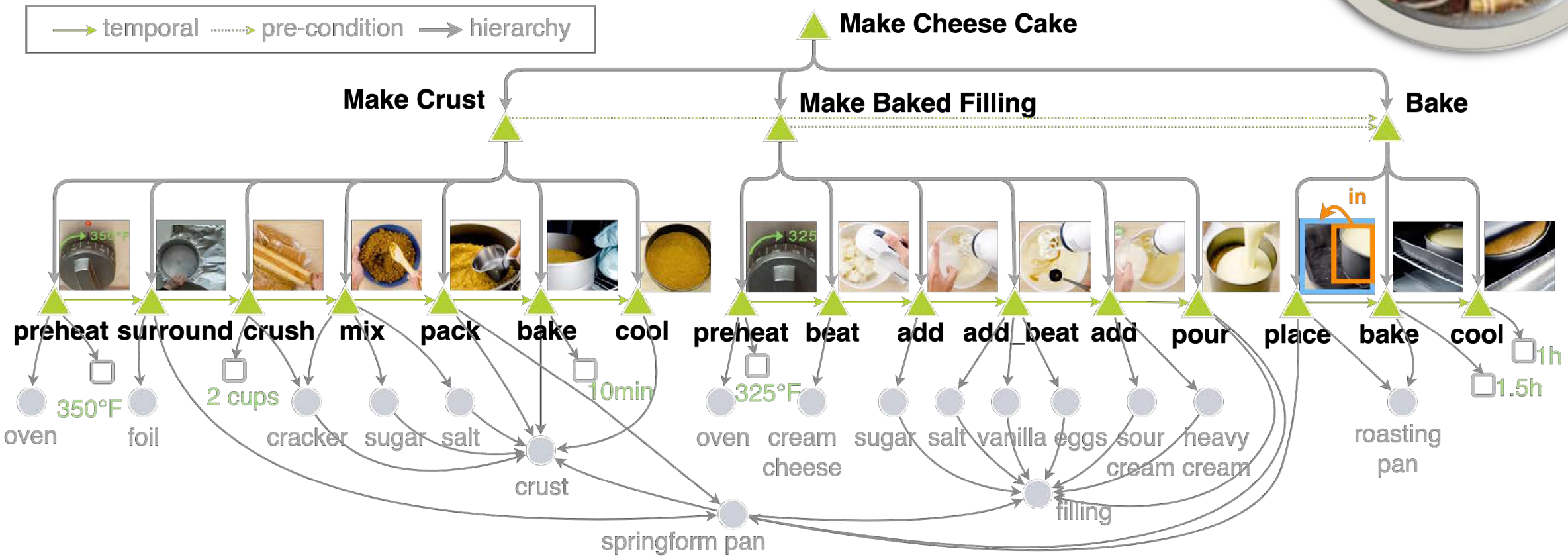
Common knowledge refers to knowledge of common patterns that is acquired or summarized from historical interaction with the world.



Common Knowledge: History repeats itself



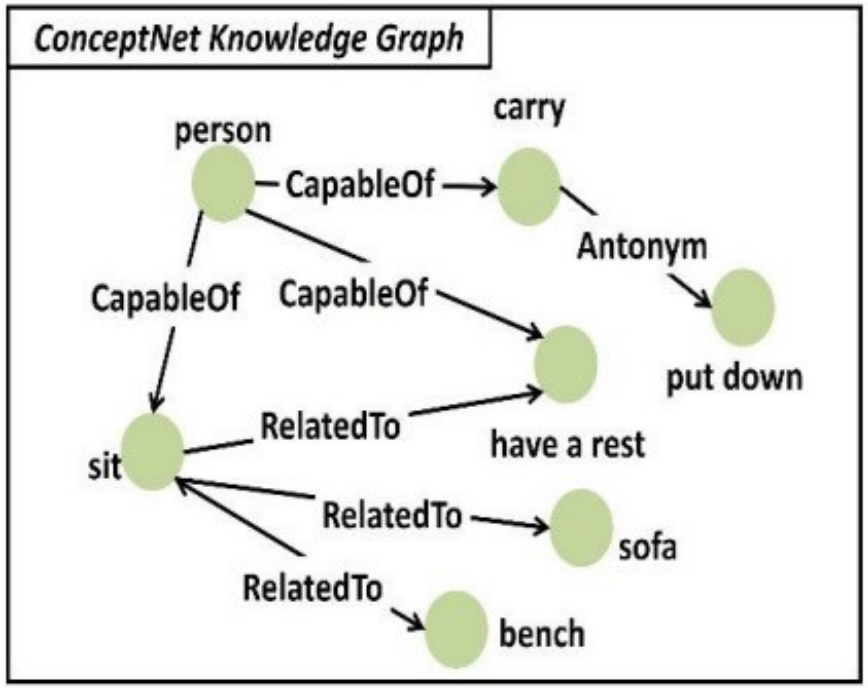
Common Knowledge: History repeats itself



Regardless of the cake flavor we choose, the preparation process is largely the same!

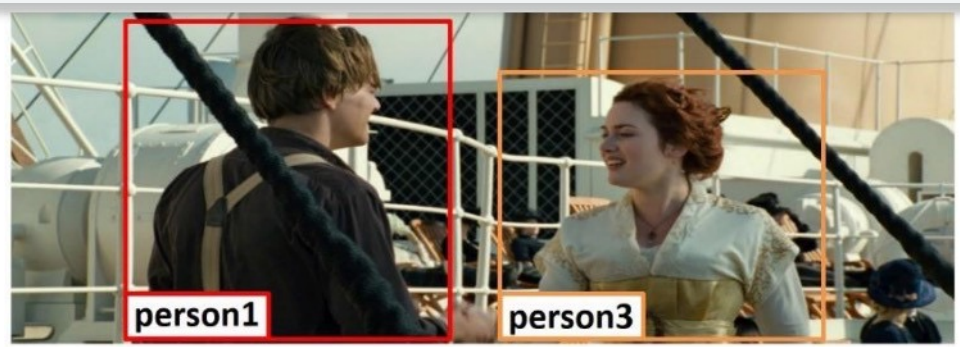
Commonsense Knowledge

Commonsense knowledge **includes facts about events occurring in time, about the effects of actions.**



ConceptNet
An open, multilingual knowledge graph

Freebase



Why are [person1] and [person3] shaking hands?

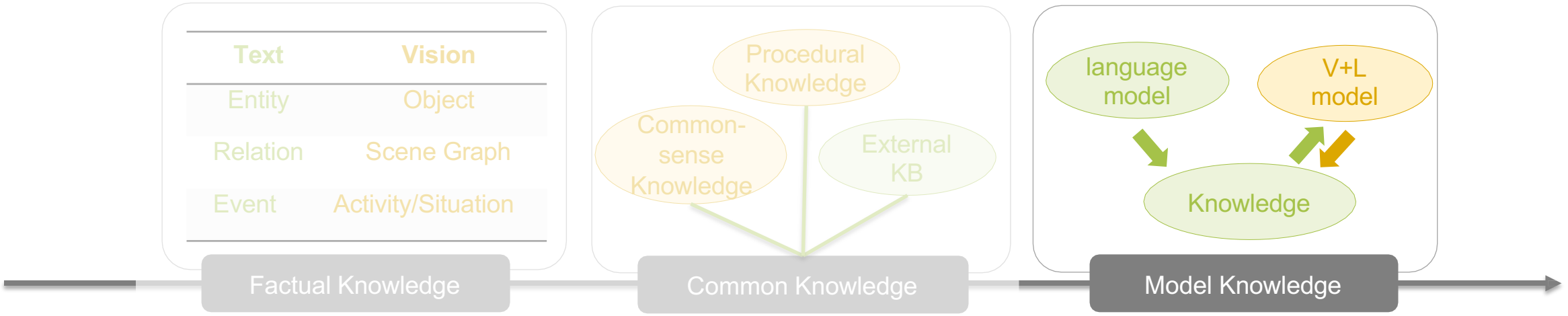
(a) [person1] and [person3] are presenting a trophy to someone.
(b) [person1] and [person3] just made a deal.
(c) [person1] and [person3] are old friends seeing each other for the first time in a long time.
(d) They have just met and are greeting each other.

I think so because ...

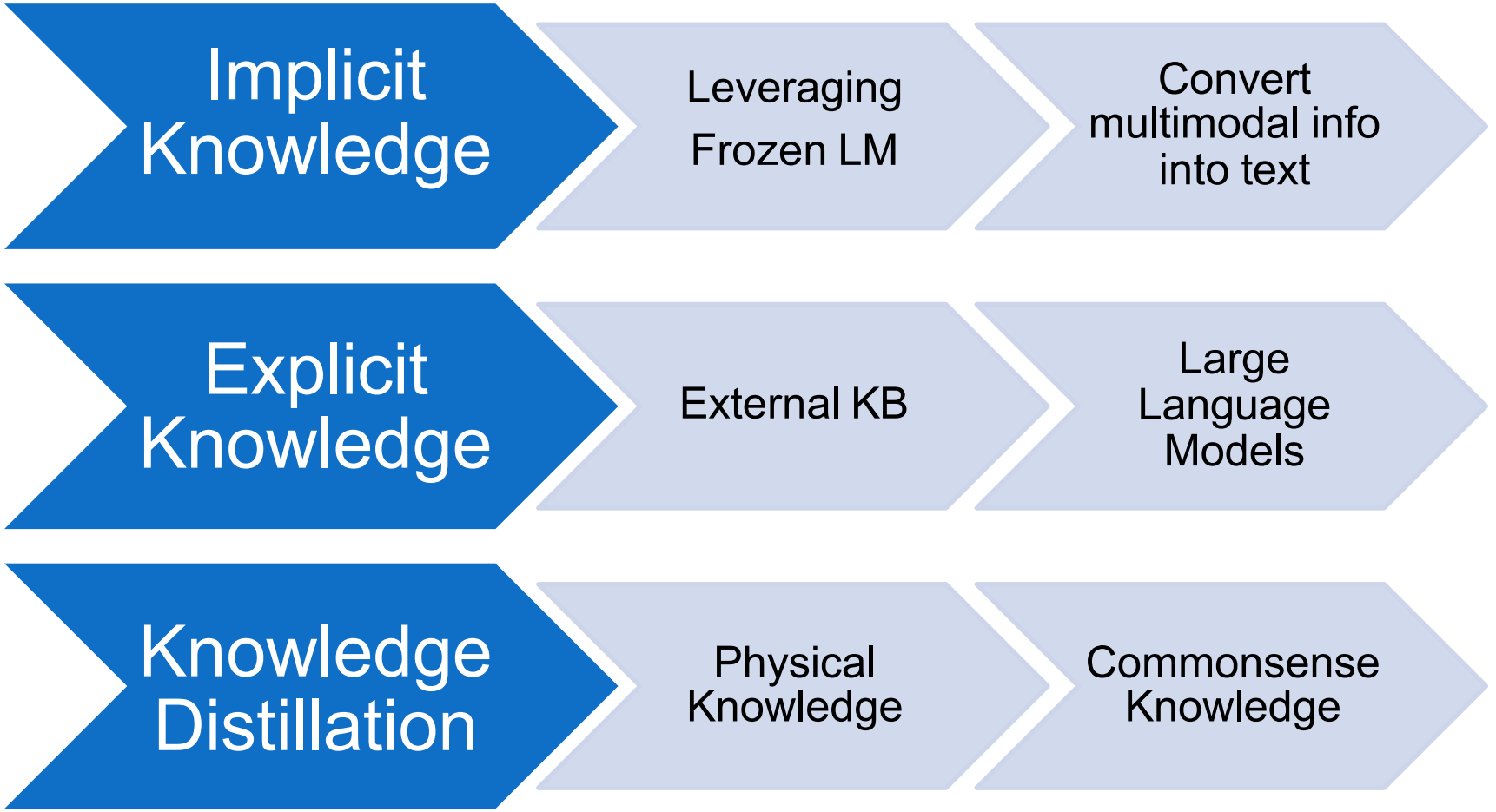
(a) People like to greet each other when they meet by shaking hands.
(b) They look like they are shaking hands to introduce themselves.
(c) They are meeting each other for the first time.
(d) Some people shake hands to greet one another by grasping each others' arms.

Model Knowledge

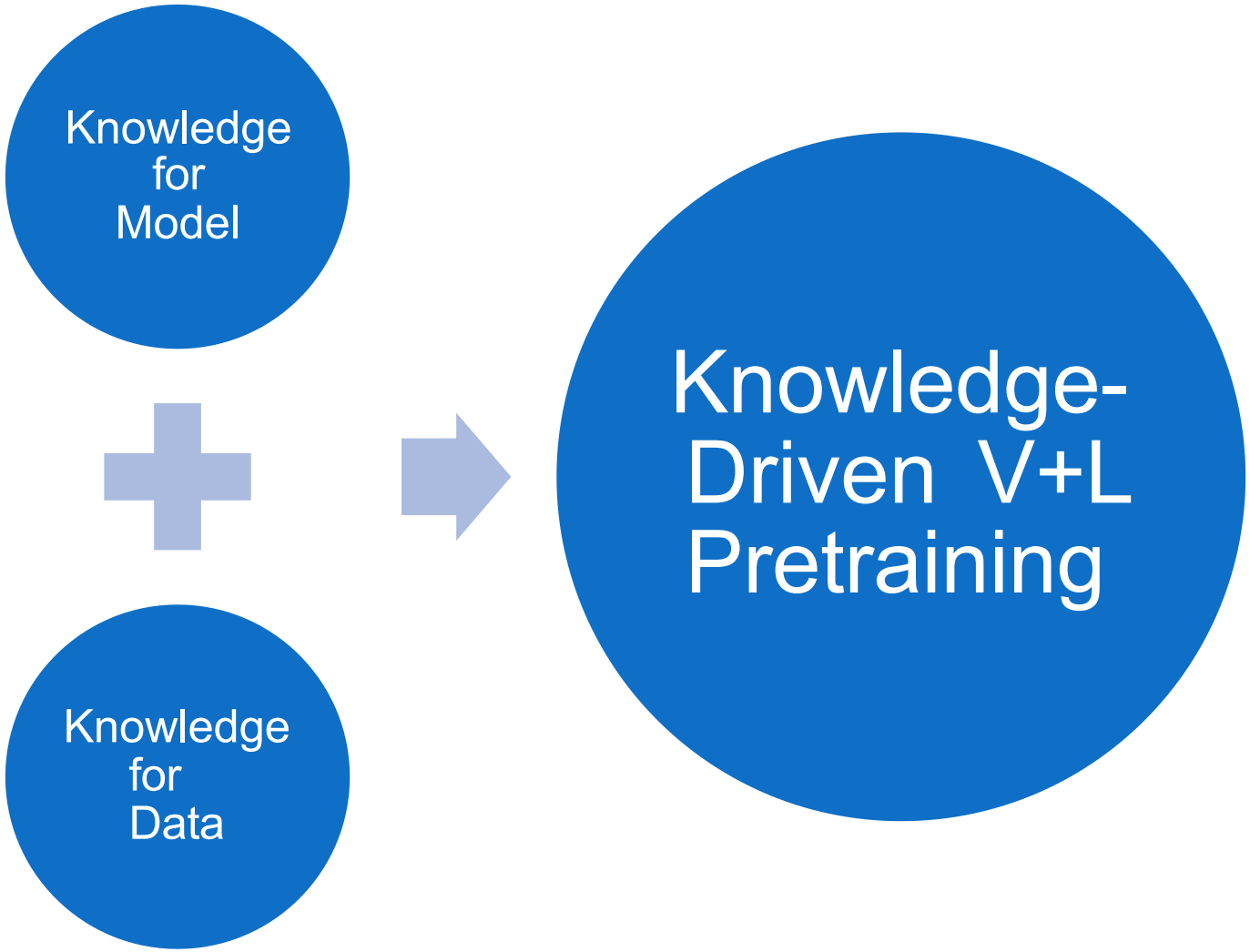
Model Knowledge is the knowledge embedded and encoded in models.



Model Knowledge



Summary: How to learn multimedia embedding?



Moving towards…

- On the model side, adding knowledge can guide the model where to focus.
  - Compositional Multi-Granularity Semantic Knowledge (such as verb, adjectives, etc)
  - Long Horizon Reasoning (such as temporal dynamics, etc)
  - Parametric Knowledge Controlling (such as parameter editing, etc)
- On the data side, knowledge is useful in the following ways:
  - In-context prompt
  - Data augmentation
  - Data selection
  - Effective Feedback

Factual Knowledge: Information about Instances

What is “Event” Knowledge?

What happened?

What is “Event” Knowledge?

What happened?

Yes! A protest.

What are they protesting for?

What is “Event” Knowledge?

What happened?

Yes! A protest.

What are they protesting for?

vaccine

What is “Event” Knowledge?

What happened?

Yes! A protest.

What are they protesting for?

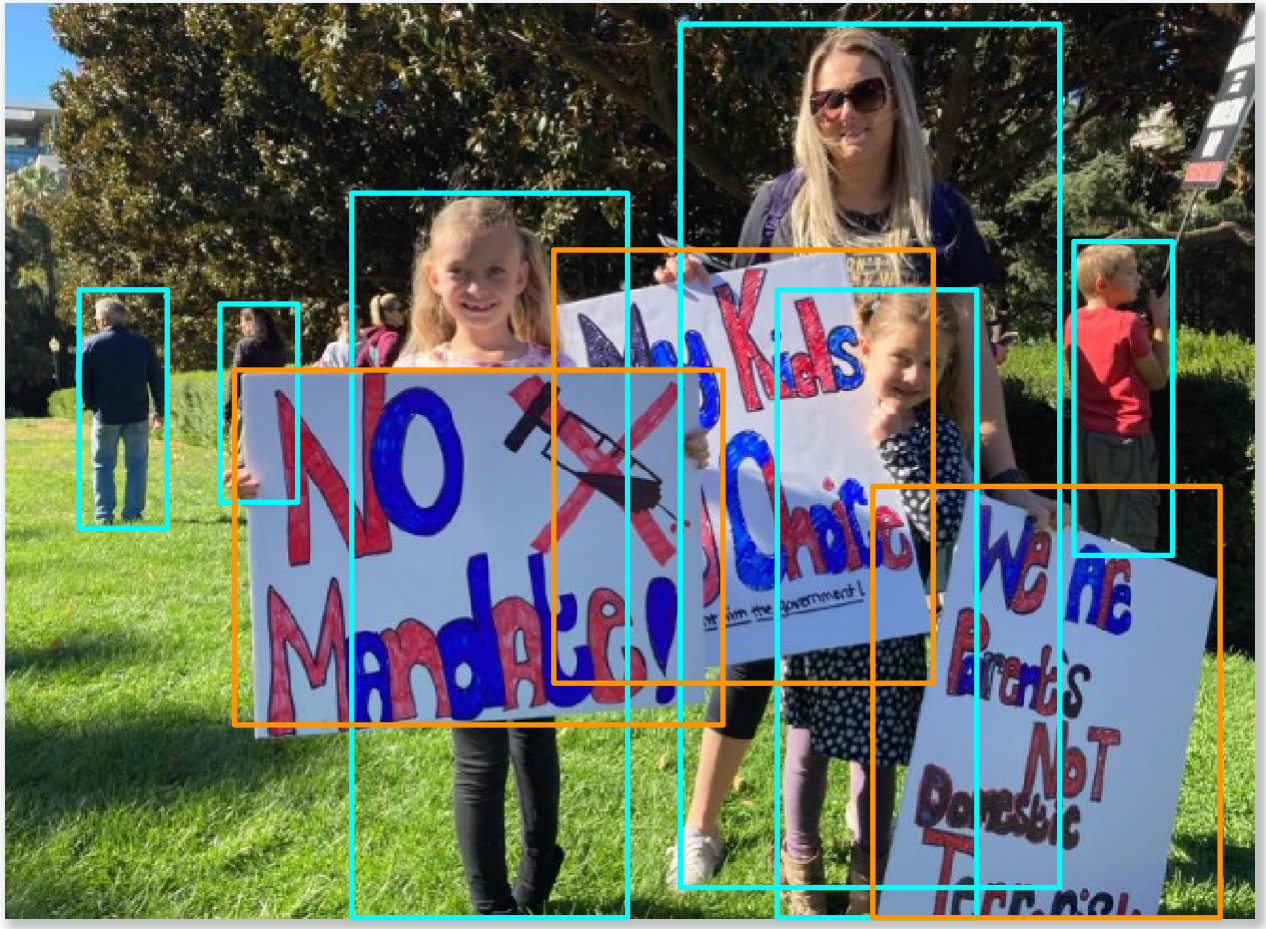
vaccine

negation

Existing object-centric info miss situational understanding

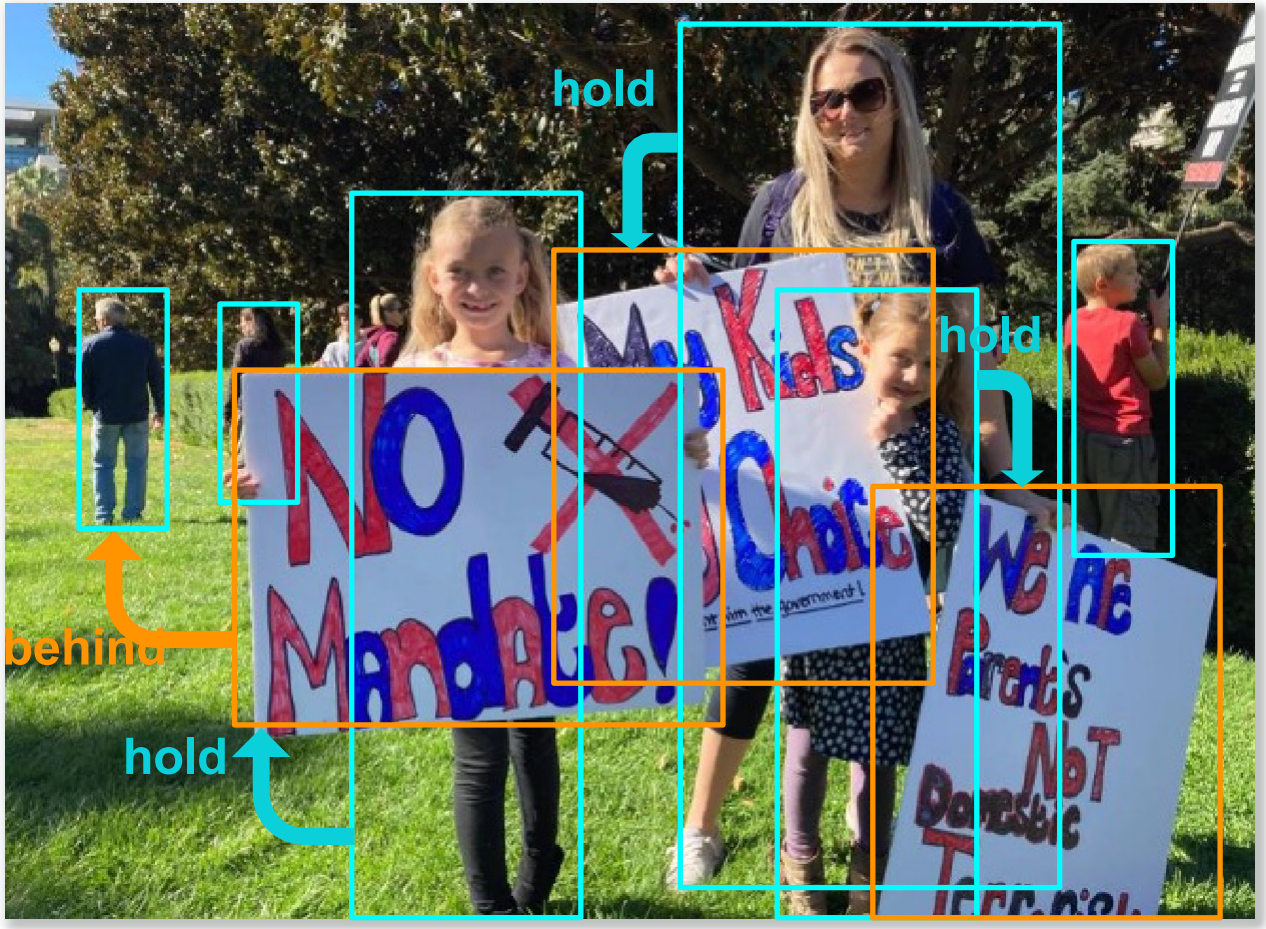
Vision

Existing object-centric info miss situational understanding



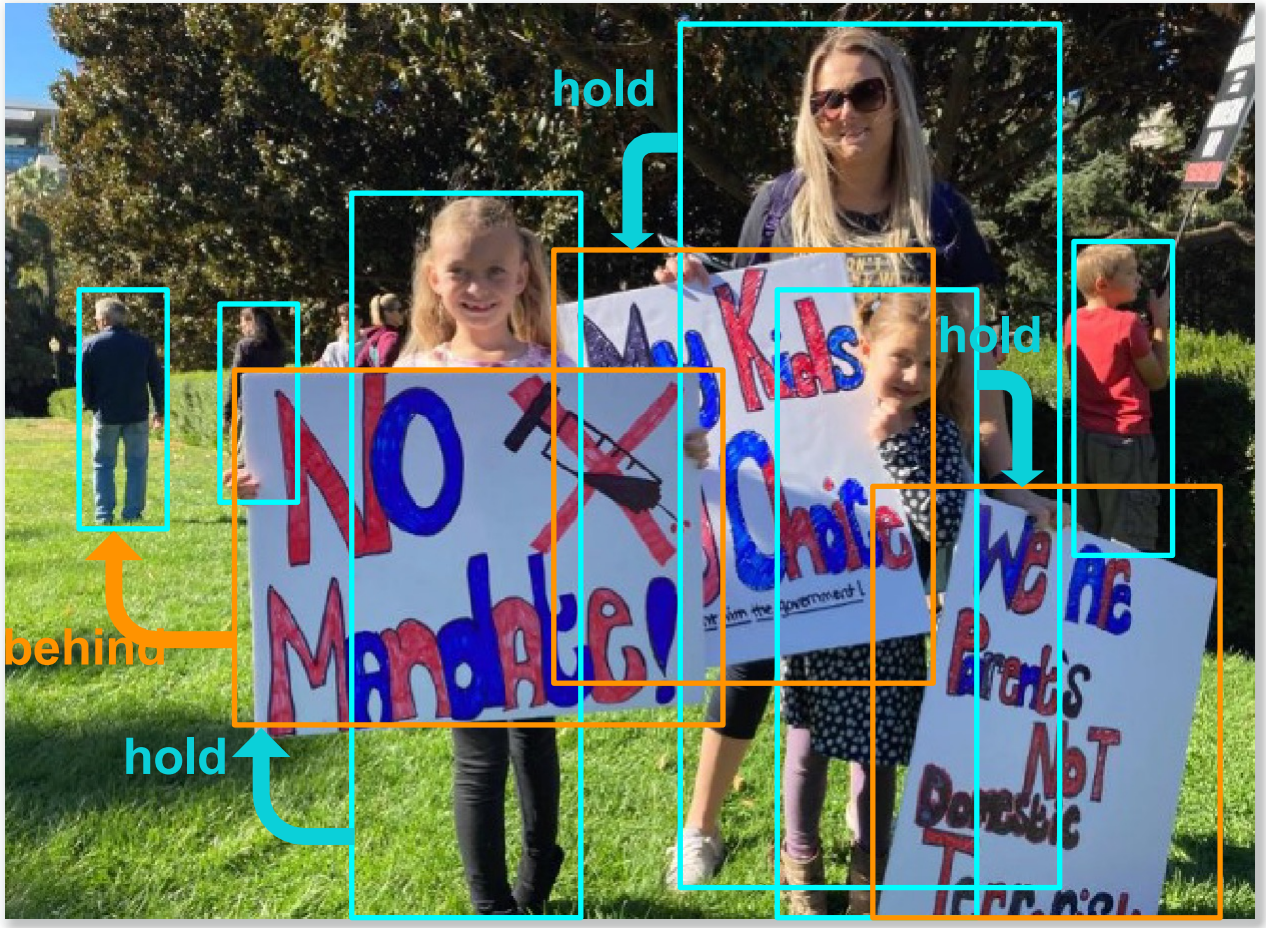
Vision
Object

Existing object-centric info miss situational understanding



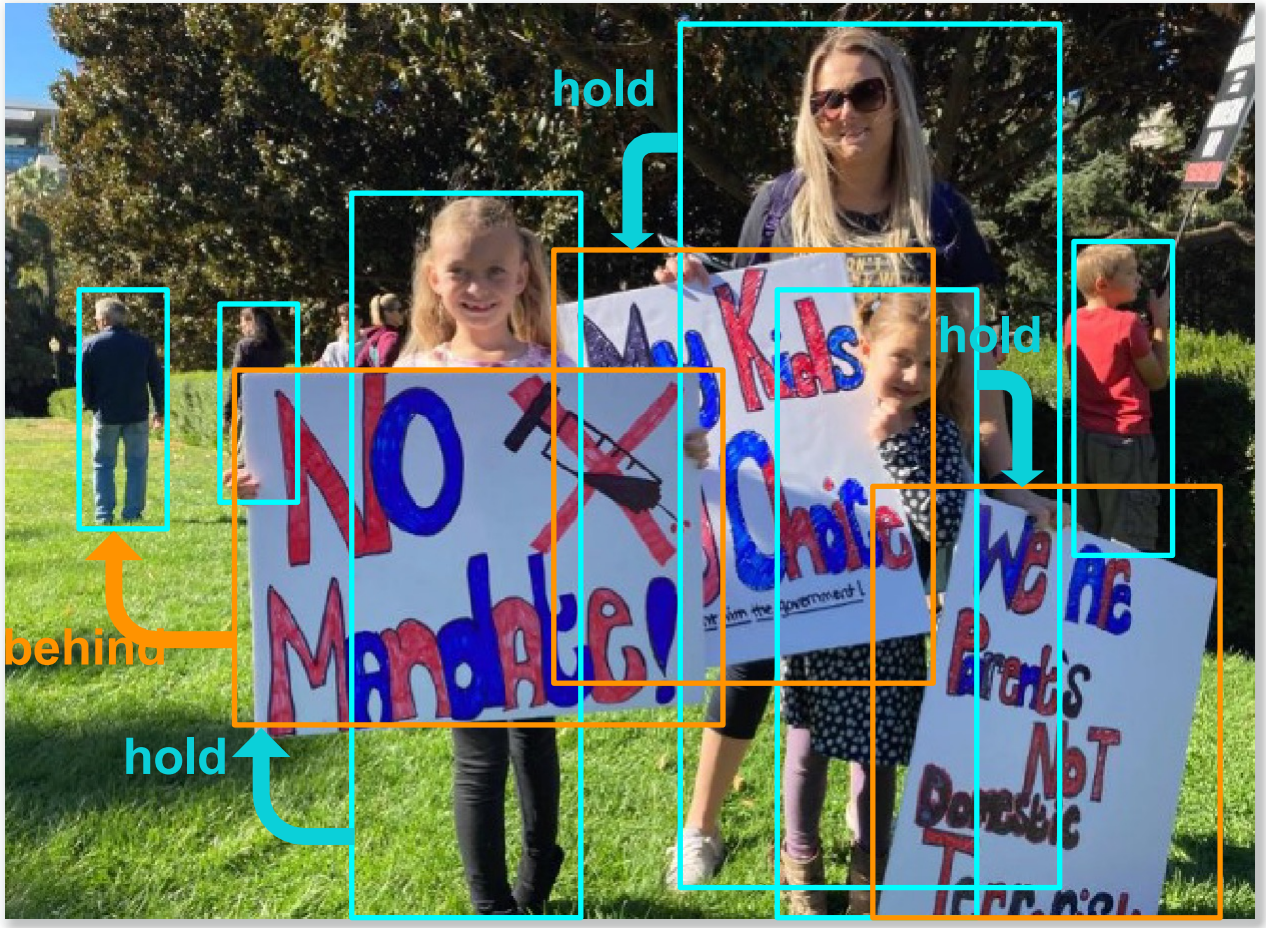
Vision
Object
Relation
Scene Graph

Existing object-centric info miss situational understanding



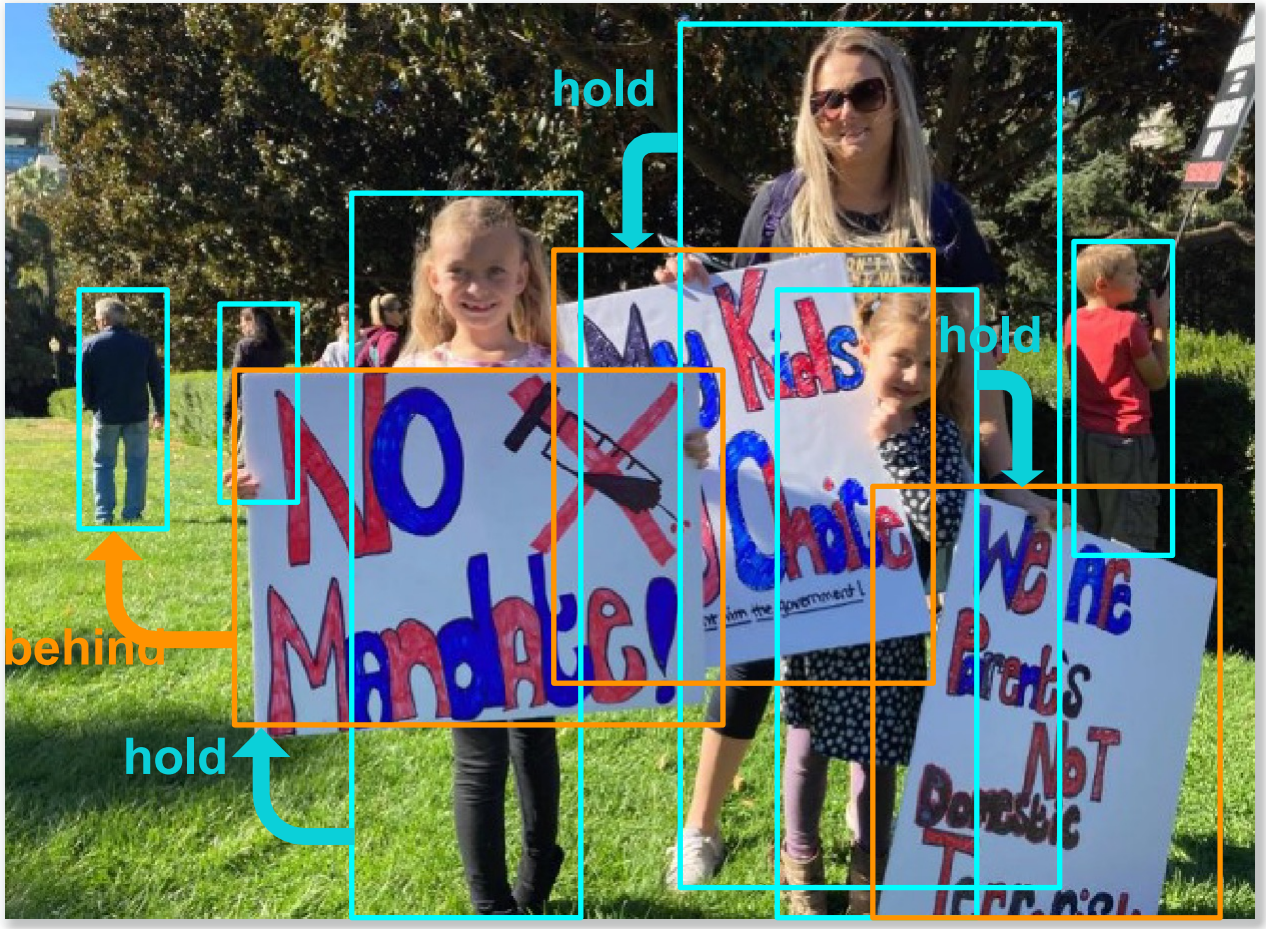
Vision	Text
Object	
Relation	
Scene Graph	

Existing object-centric info miss situational understanding



Vision	Text
Object	Entity
Relation	
Scene Graph	

Existing object-centric info miss situational understanding



Vision	Text
Object	Entity
Relation	Relation
Scene Graph	Entity-Relation Graph

Existing object-centric info miss situational understanding

	Vision	Text
Entity-centric	Object	Entity
	Relation	Relation
	Scene Graph	Entity-Relation Graph

Existing object-centric info miss situational understanding

Vision	Text
Object	Entity
Relation	Relation
Scene Graph	Entity-Relation Graph

Entity-centric

State-of-the-art Captioner (Kamath et al., 2022)

Answer

a woman holding a sign in front of a group of people.

a woman holding a sign while standing in a park.

a woman holding a sign in front of a crowd.

Definition of “Event”

Event	Protest

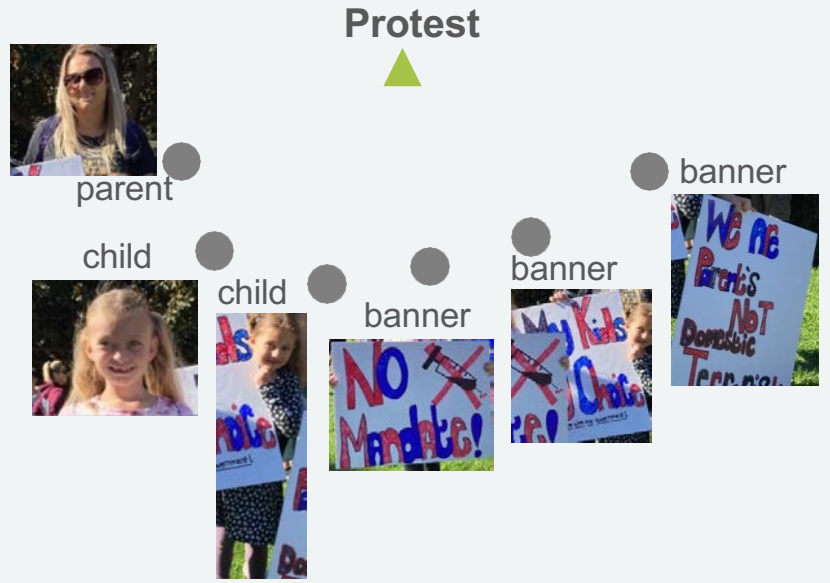
What happened?

Protest

Definition of “Event”

Event	Protest
	parent
	child
	banner
	No vaccine mandate for kids

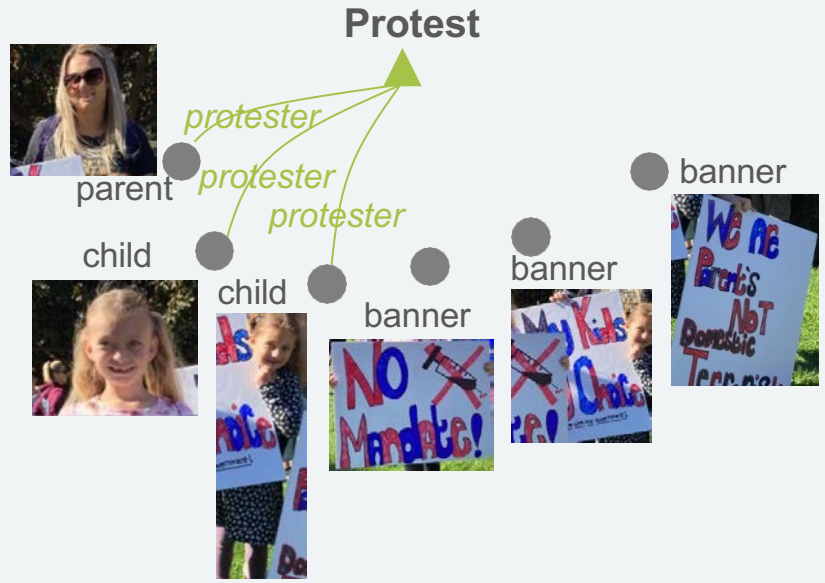
What happened?
Who?



Definition of “Event”

Event	Protest
Protester	parent
Protester	child
	banner
	No vaccine mandate for kids

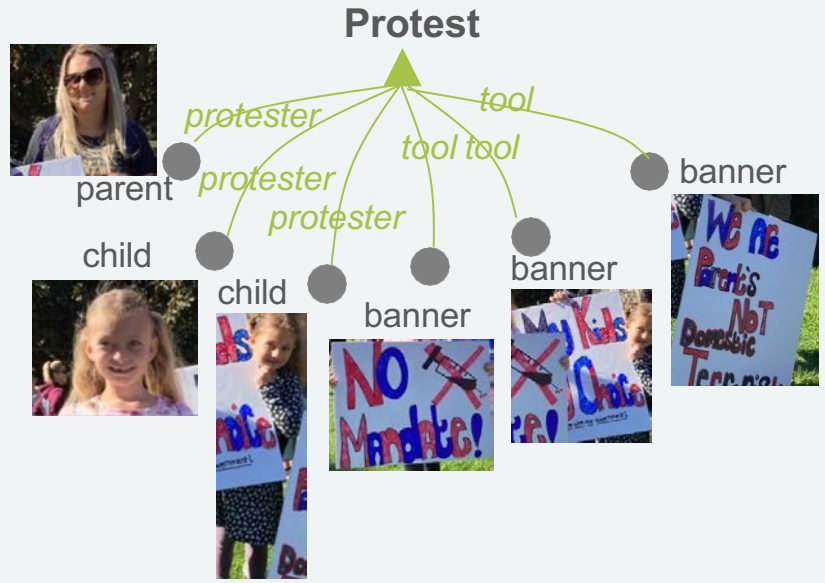
What happened?
Who?



Definition of “Event”

Event	Protest
Protester	parent
Protester	child
Tool	banner
Topic	No vaccine mandate for kids

What happened?
Who?



A New Task of Multimodal Event Extraction [ACL'20]

Event Extraction

Input: text, image, video, speech, …

Output: structured knowledge

1. **Event Type** ***(e.g., protest)***
2. **Participants** ***(e.g., child)*** **& Semantic Roles** ***(e.g., protester)***

[Manling Li, et al., Cross-media Structured Common Space for Multimedia Event Extraction. ACL 2020]

A New Task of Multimodal Event Extraction [ACL'20]

Event Extraction

Input: text, image, video, speech, …

Output: structured knowledge

1. **Event Type** *(e.g., protest)*
2. **Participants** *(e.g., child)* **& Semantic Roles** *(e.g., protester)*

[Manling Li, et al., Cross-media Structured Common Space for Multimedia Event Extraction. ACL 2020]

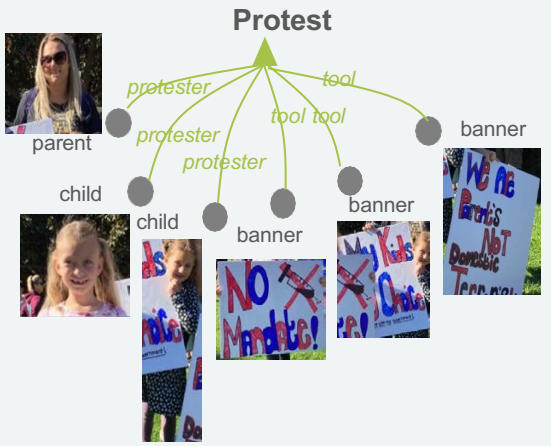
What is Multimodal Event Extraction? [Li et al, ACL'20]

Event Extraction

Input: text, image, video, speech, …

Output: structured knowledge

1. Event Type *(e.g., protest)*
2. Participants *(e.g., child)* & Semantic Roles *(e.g., protester)*



[Manling Li, et al., Cross-media Structured Common Space for Multimedia Event Extraction. ACL 2020]

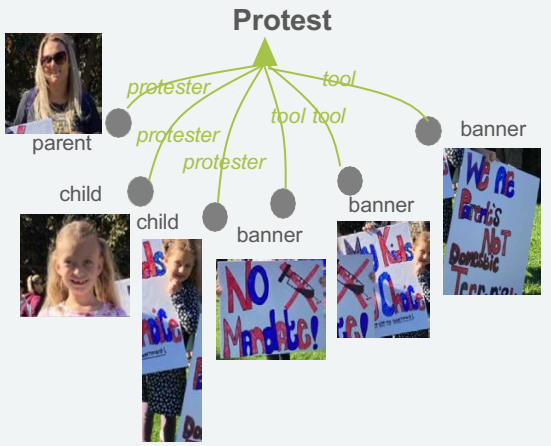
What is Multimodal Event Extraction? [Li et al, ACL'20]

Event Extraction

Input: text, image, video, speech, …

Output: structured knowledge

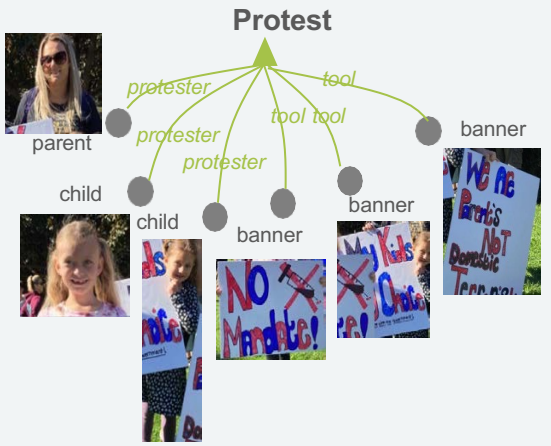
1. **Event Type** *(e.g., protest)*
2. **Participants** *(e.g., child)* **& Semantic Roles** *(e.g., protester)*



[Manling Li, et al., Cross-media Structured Common Space for Multimedia Event Extraction. ACL 2020]

What is Multimodal Event Extraction? [Li et al, ACL'20]

Event Extraction



Input: text, image, video, speech, …

Output: structured knowledge

1. **Event Type** *(e.g., **protest**)*
2. **Participants** *(e.g., **child**)* **& Semantic Roles** *(e.g., **protester**)*

[Manling Li, et al., Cross-media Structured Common Space for Multimedia Event Extraction. ACL 2020]

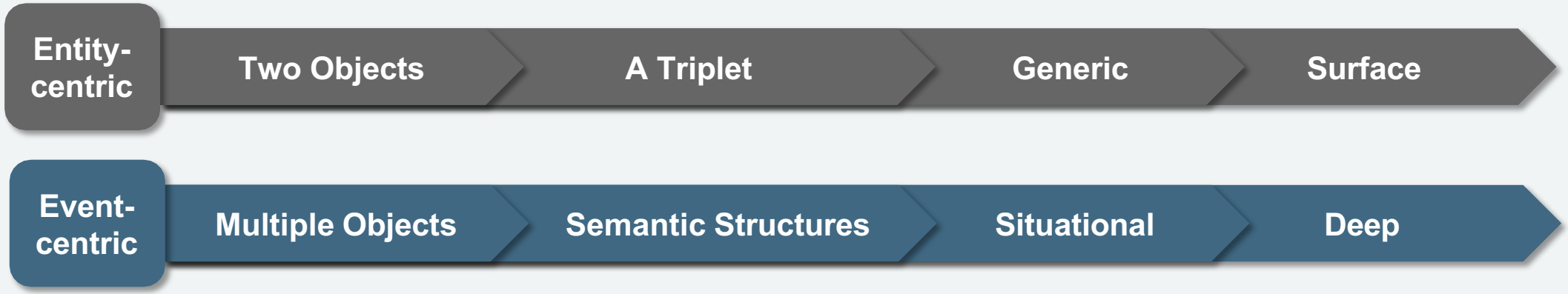
What is Multimodal Event Extraction? [Li et al, ACL'20]

Event Extraction

Input: text, image, video, speech, …

Output: structured knowledge

1. **Event Type** *(e.g., **protest**)*
2. **Participants** *(e.g., **child**)* **& Semantic Roles** *(e.g., **protester**)*



[Manling Li, et al., Cross-media Structured Common Space for Multimedia Event Extraction. ACL 2020]

Goal: Entity-Centric → Event-Centric

	Text	Vision
Entity-centric	Entity	Object
	Relation	Relation
	Entity-Relation Graph	Scene Graph

Goal: Entity-Centric → Event-Centric

Text	Vision
Entity	Object
Relation	Relation
Entity-Relation Graph	Scene Graph
Verb	Activity
Event Structure	**Image Event Graph**

Event-centric

Goal: Entity-Centric → Event-Centric

	Text	Vision
Entity-centric	Entity	Object
	Relation	Relation
	Entity-Relation Graph	Scene Graph
Event-centric	Verb	Activity
	Event Structure	**Image Event Graph**

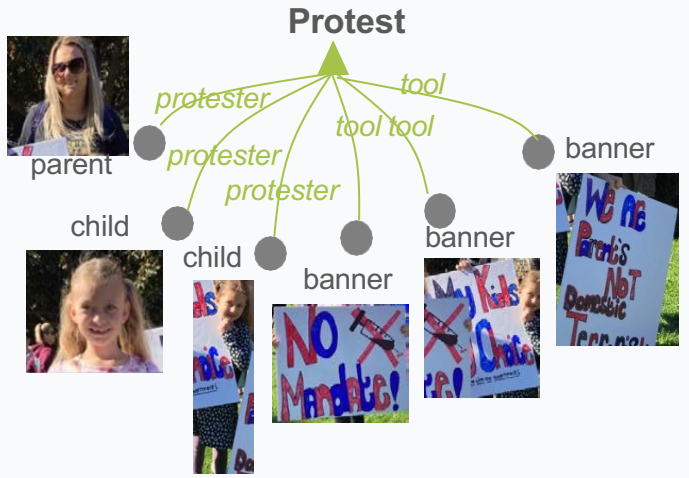
Goal: Entity-Centric → Event-Centric

State-of-the-art Captioner (Kamath et al., 2022)

Answer
a woman holding a sign in front of a group of people.
a woman holding a sign while standing in a park.
a man holding a sign in front of a crowd.

Entity-centric

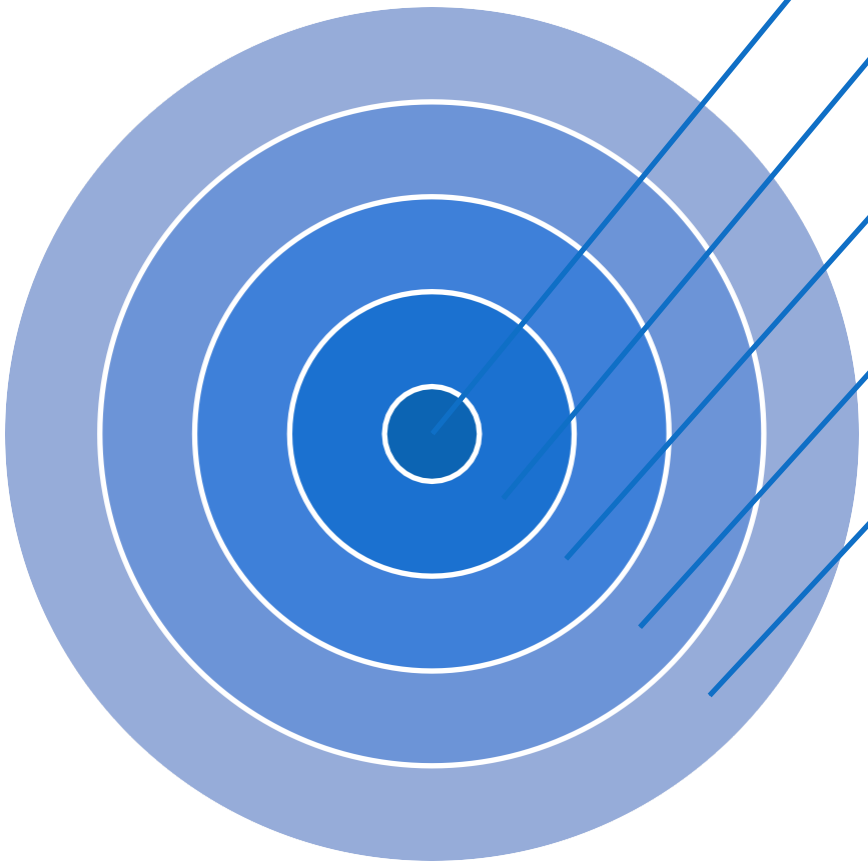
Event-centric



Adding knowledge to pretraining models

Knowledge for Model

Knowledge for data



physical world knowledge

situational knowledge

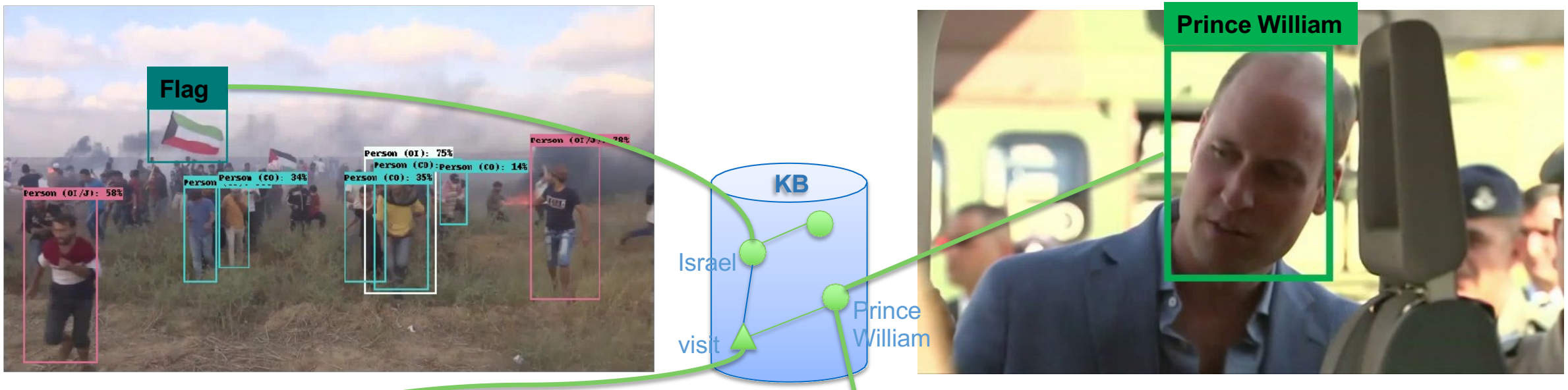
scene graphs

object & labels

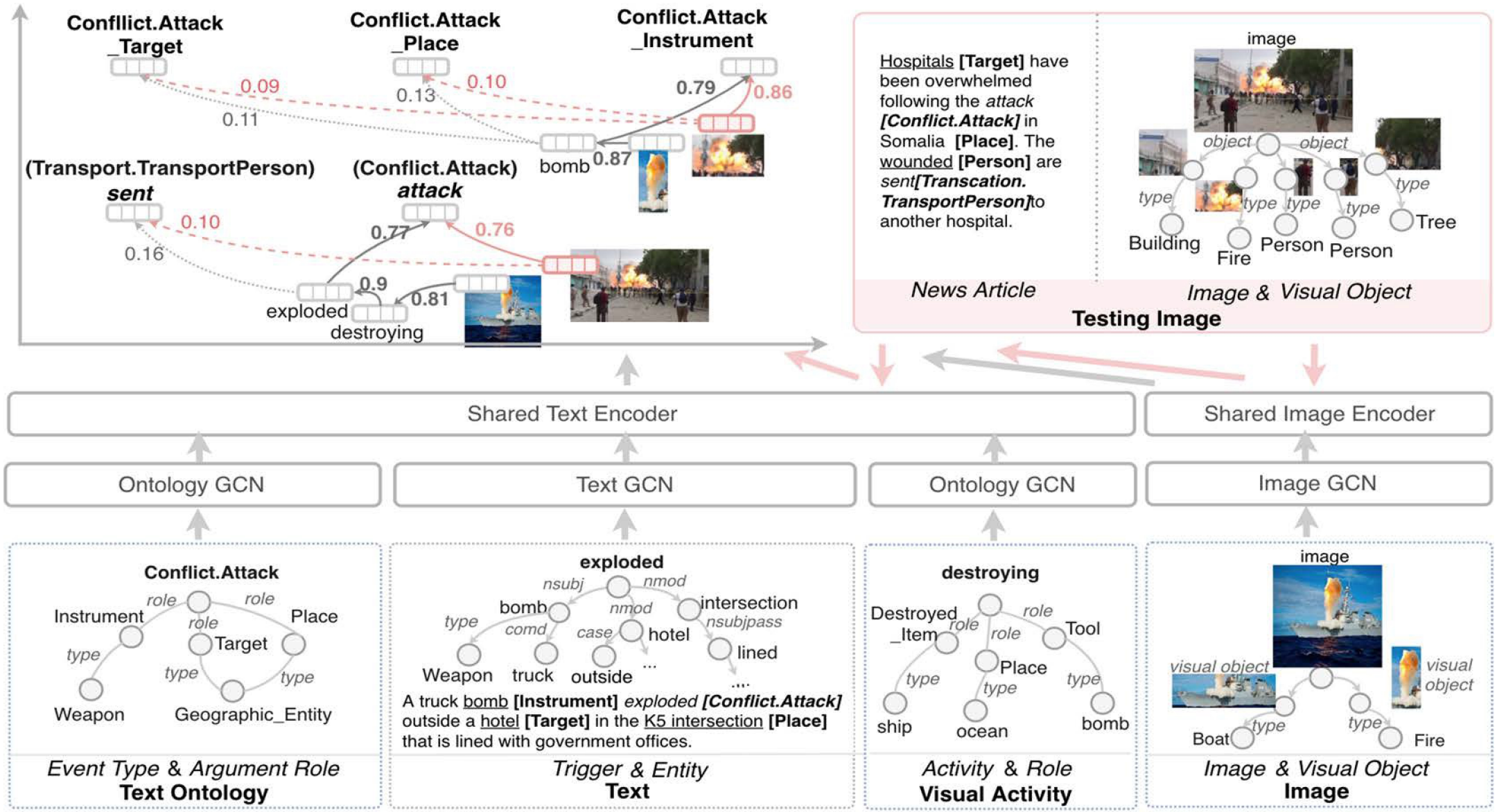
pixels

What is factual knowledge?

- Multimedia Knowledge Base with entities, relations and events.



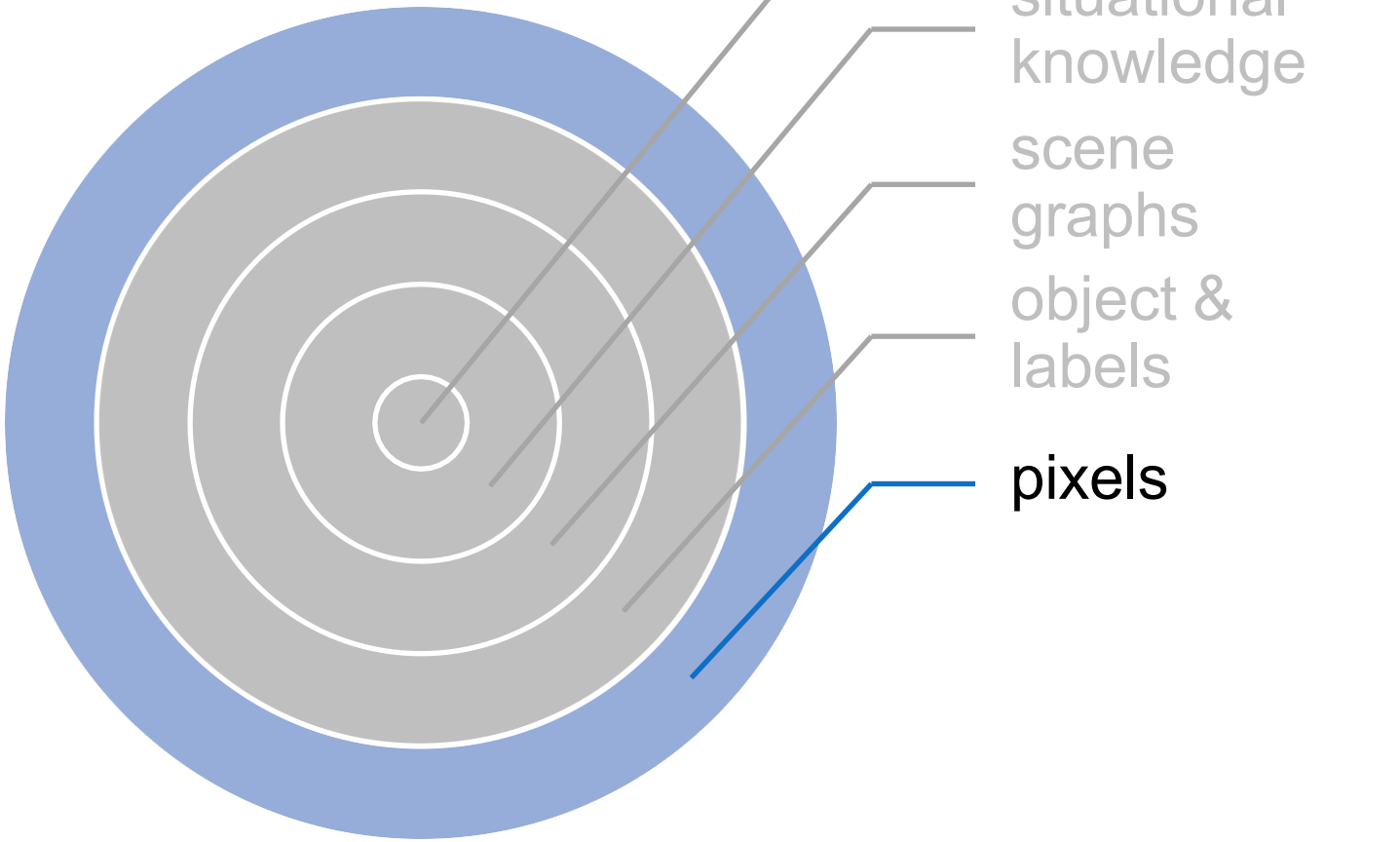
Goal: A joint representation of text and vision knowledge



Adding knowledge to pretraining models

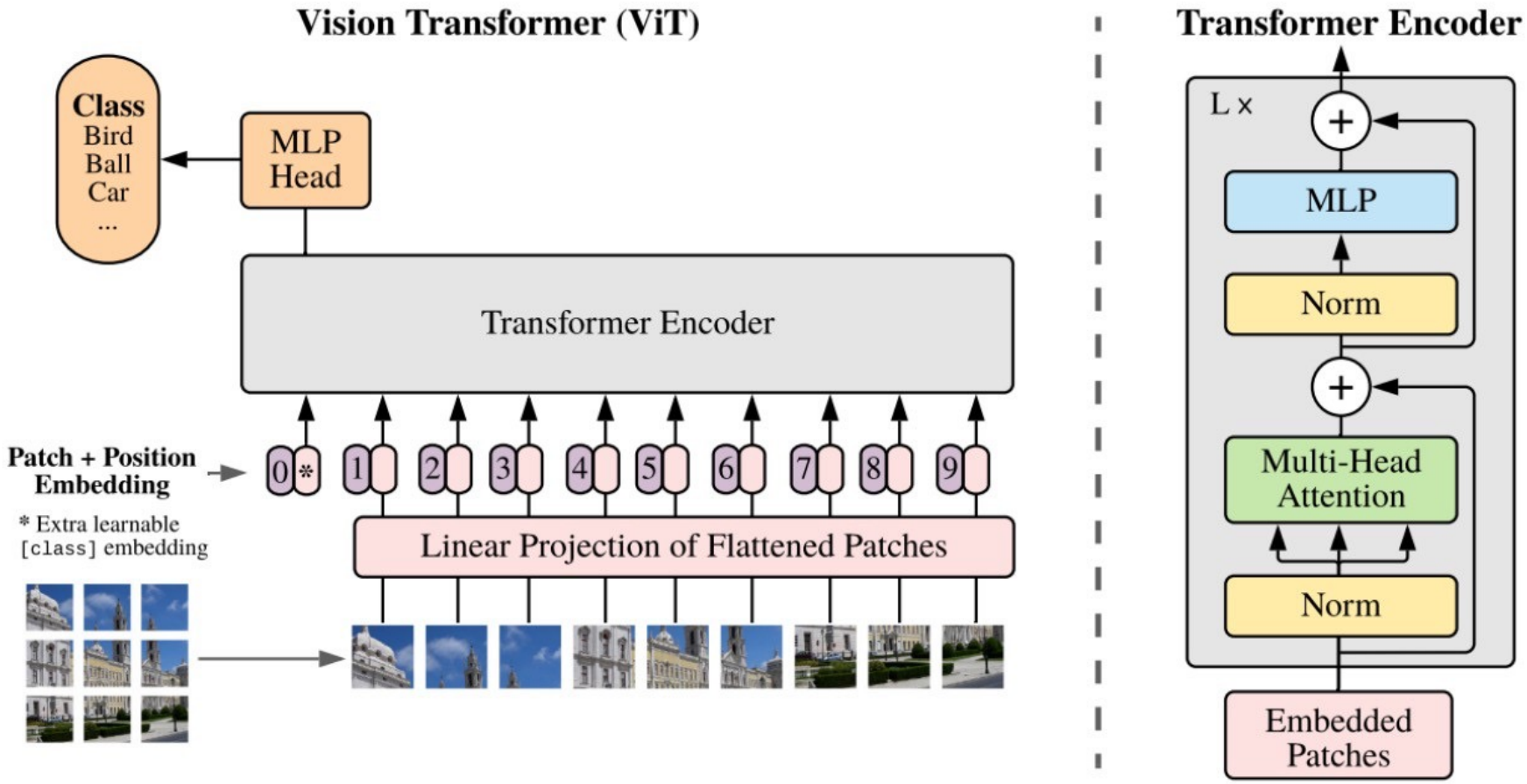
Knowledge for Model

Knowledge for data



Pixels - An Image is Worth 16x16 Words

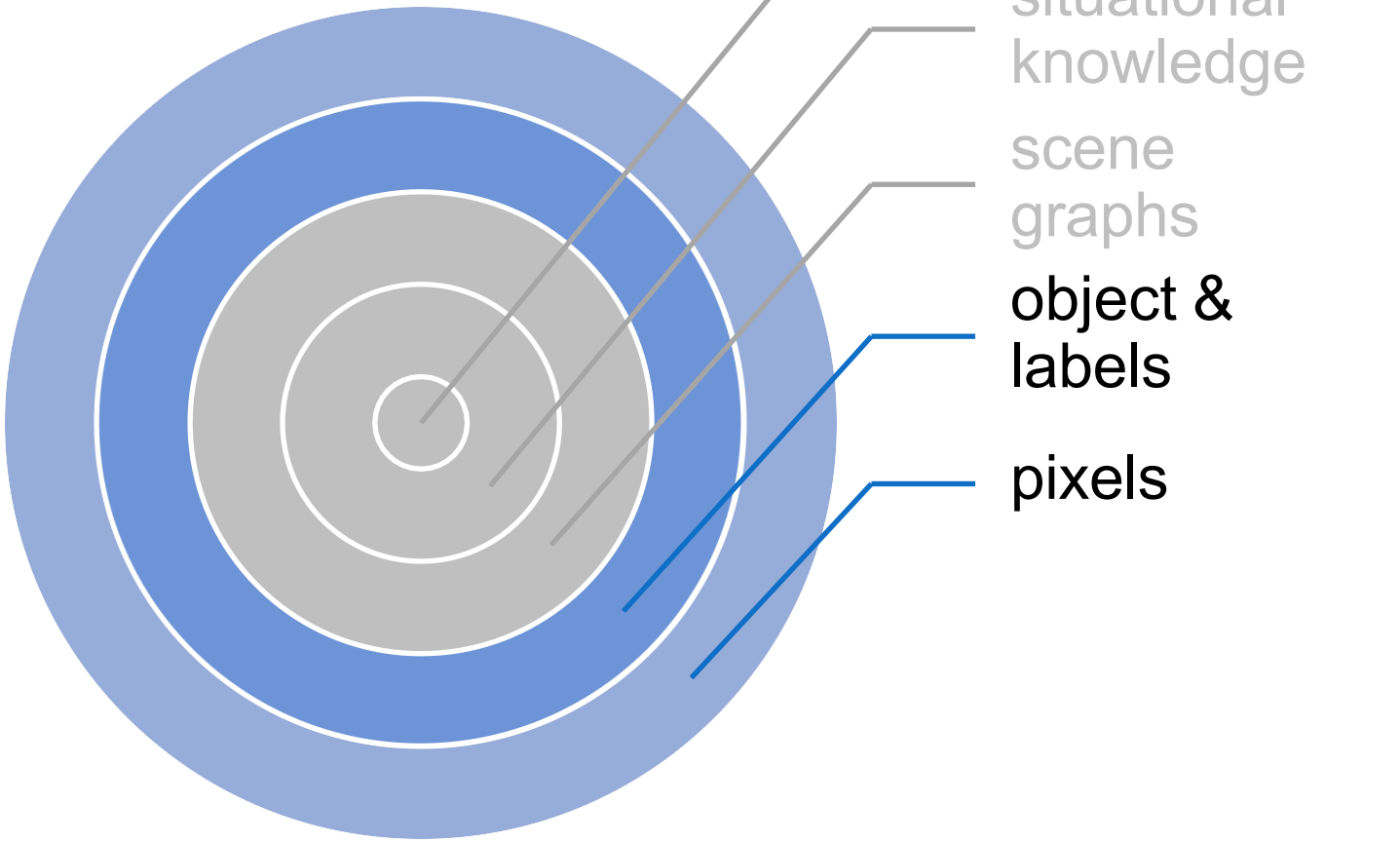
The simplest way is to split an image into patches



Adding knowledge to pretraining models

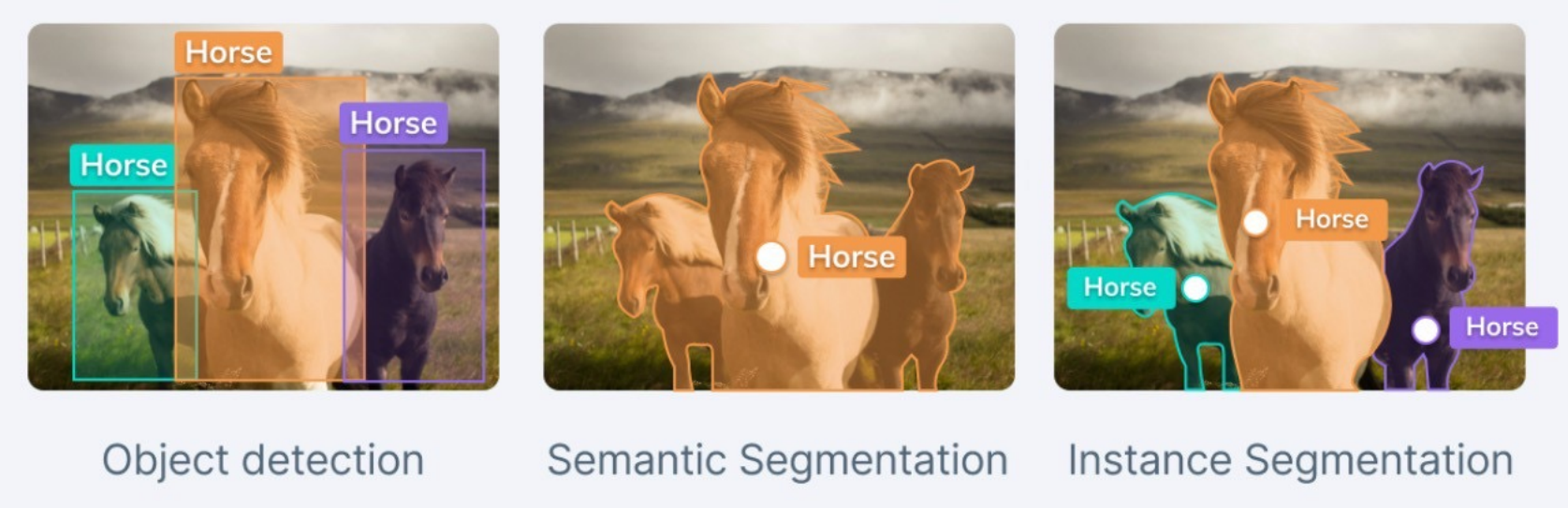
Knowledge for Model

Knowledge for data



Entity Knowledge

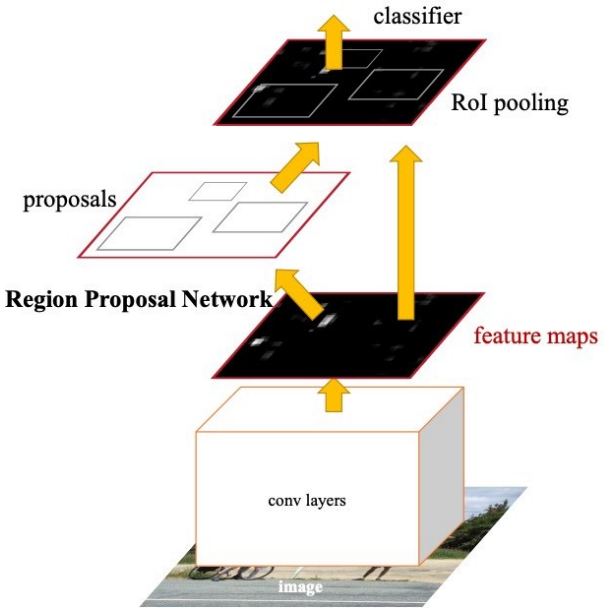
- Object Detection: Object instances at the bounding box level
- Semantic Segmentation: Object class at the pixel level
- Instance Segmentation: Object instances at the pixel level



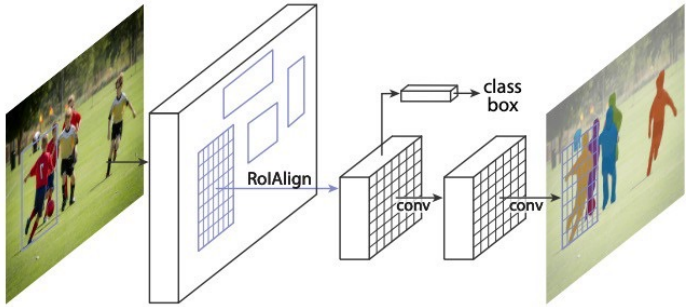
https://www.v7labs.com/blog/object-detection-guide

The way to obtain entity knowledge: Object Extraction

Two-Stage (With Proposal)

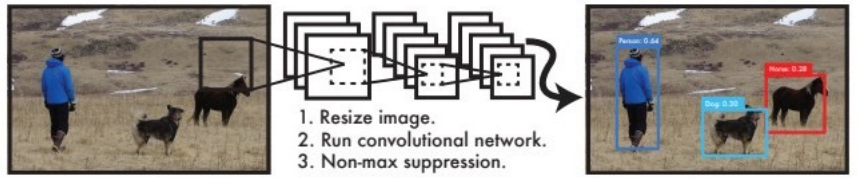


Faster RCNN

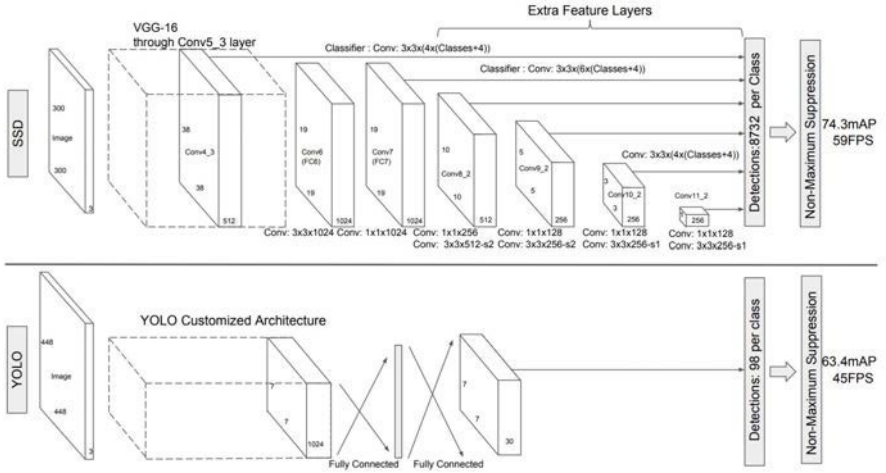


Mask RCNN

One-Stage (Without Proposal)



YOLO



SSD

Ren, S., He, K., Girshick, R., & Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS 2015*.

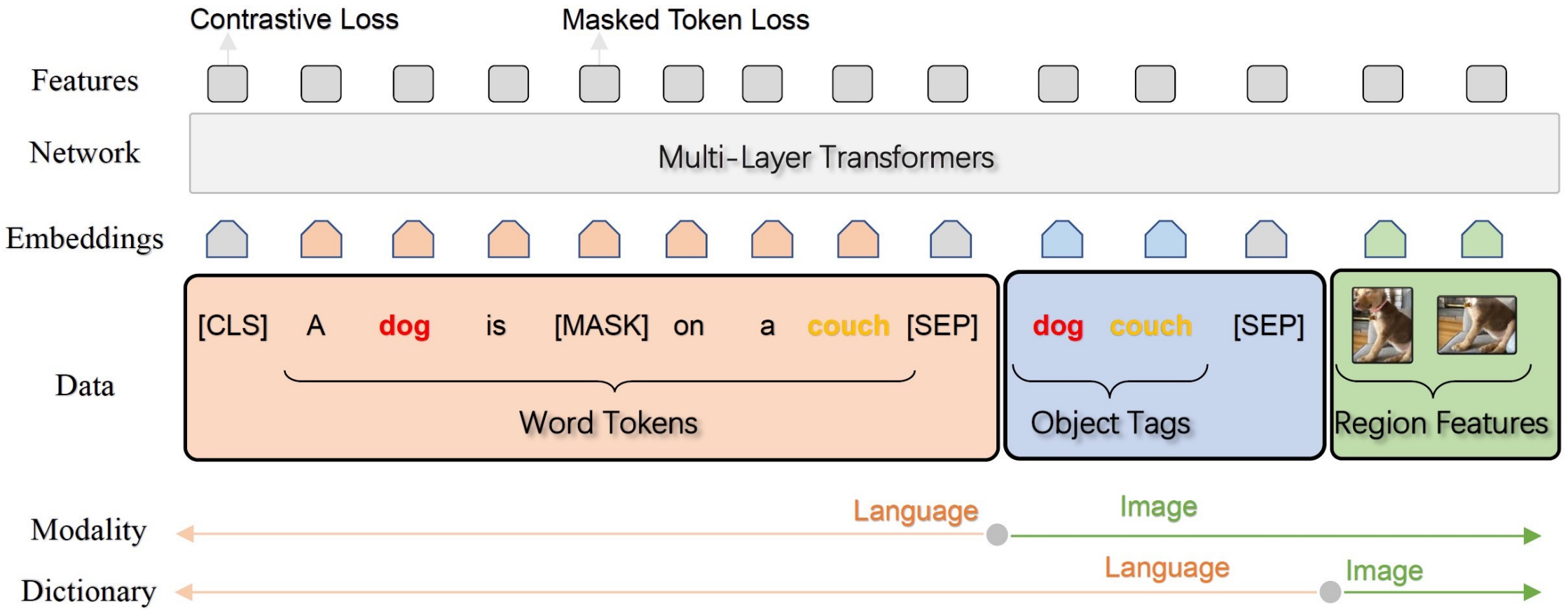
He, Kaiming, et al. "Mask r-cnn." *CVPR* 2017.

Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." *CVPR* 2016.

Liu, Wei, et al. "Ssd: Single shot multibox detector." *ECCV* 2016.

Oscar [ECCV 2020] and VinVL [CVPR 2021]

- Object knowledge is richer.
  - Add object label knowledge as anchor points



Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks, ECCV 2020

VinVL: Making Visual Representations Matter in Vision-Language Models. CVPR 2021

Soft Prompt Entity Knowledge [CVPR2022]

- **[Align and Prompt 2021]** Align and Prompt: Video-and-Language Pre-training with Entity Prompts
  - Adding regional entity prediction task

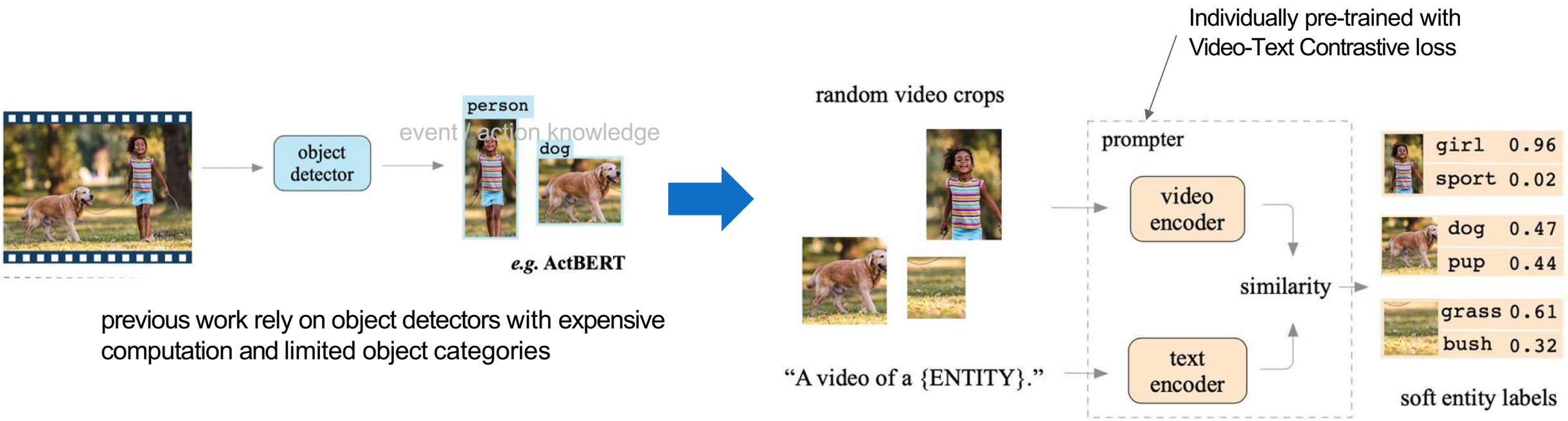
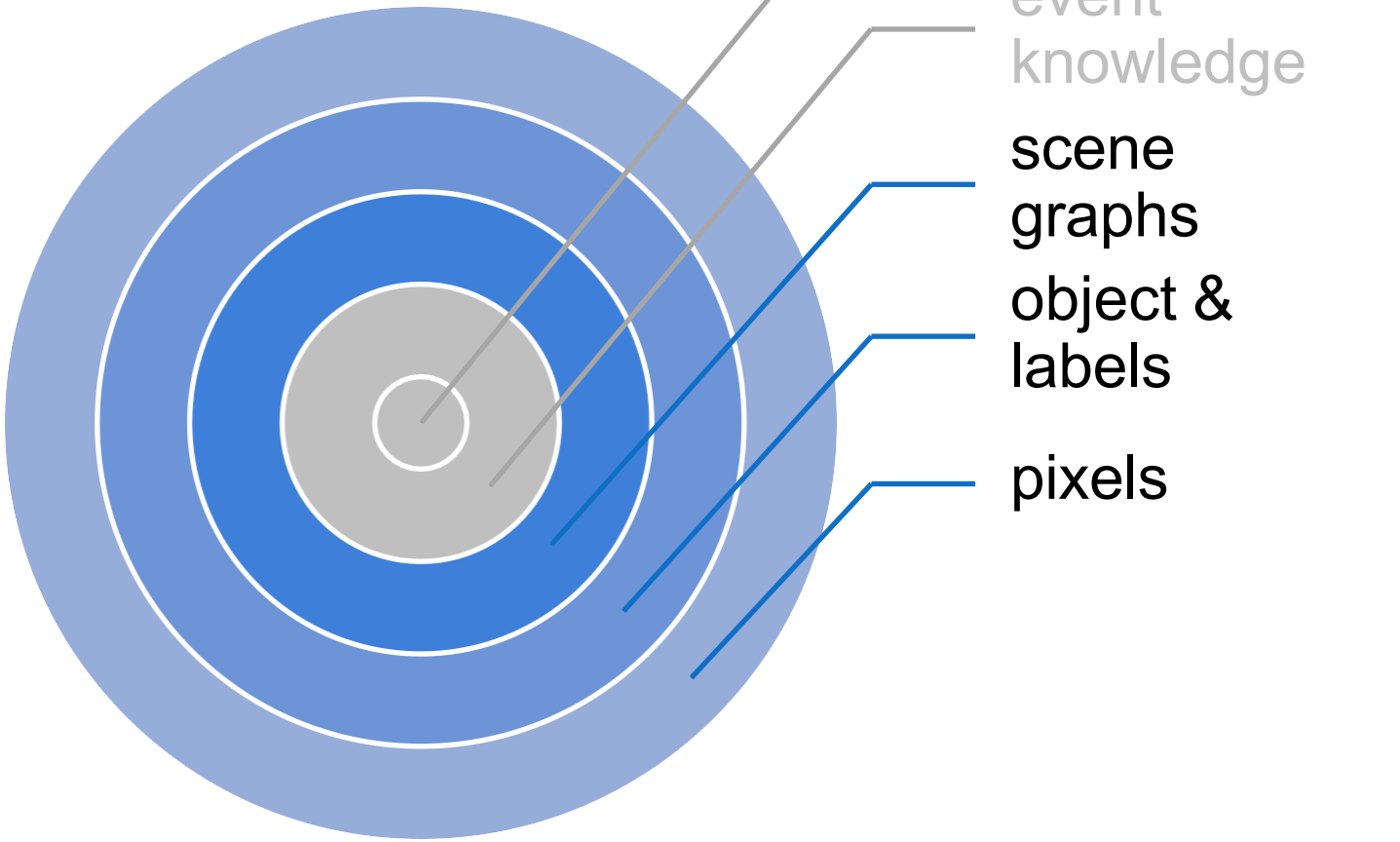


image source: Align and Prompt: Video-and-Language Pre-training with Entity Prompts

Adding knowledge to pretraining models

Knowledge for Model

Knowledge for data



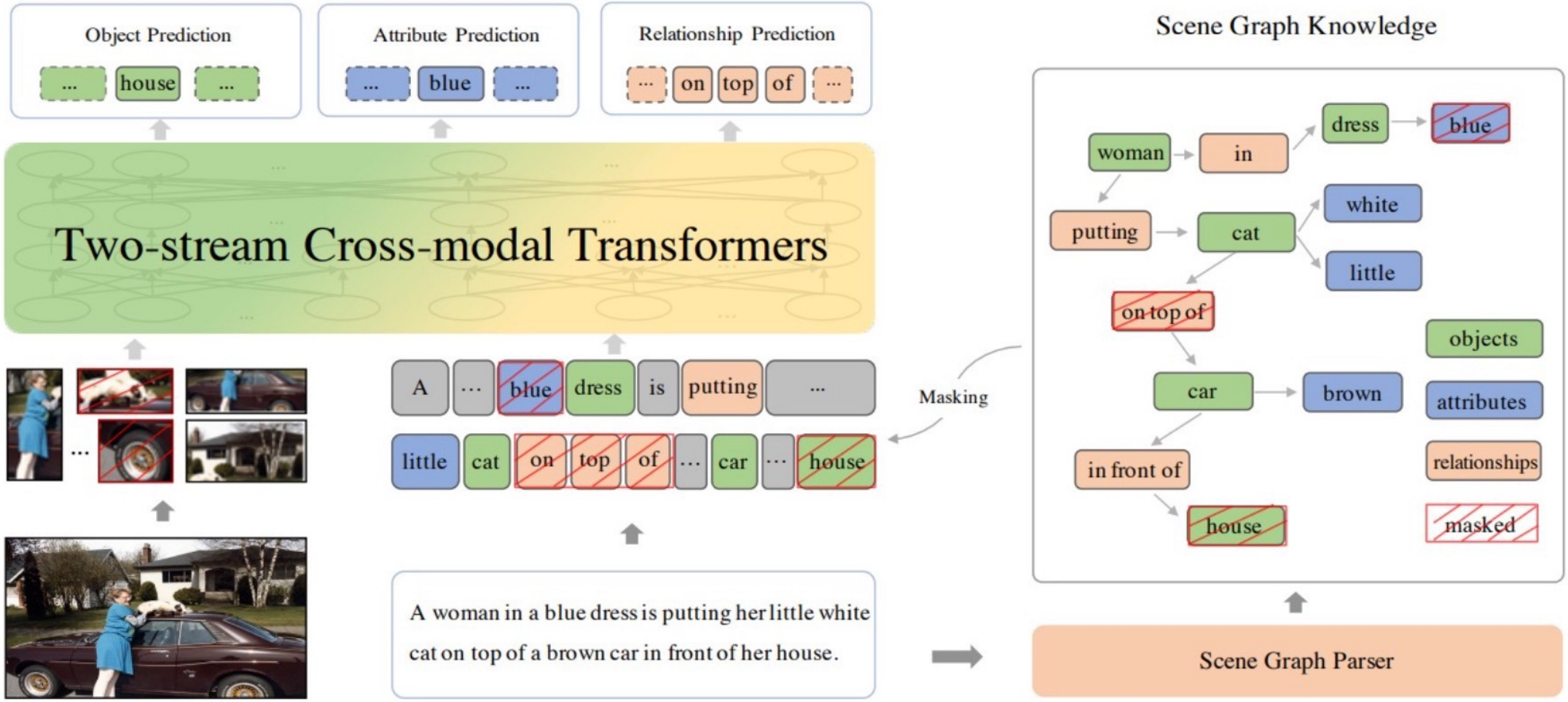
ERINE-ViL [AAAI2021]

- Add scene graph knowledge as downstream tasks
  - Object prediction
  - Attribute prediction
  - Relationship prediction

ERNIE-ViL: Knowledge Enhanced Vision-Language Representations Through Scene Graph, AAAI 2021

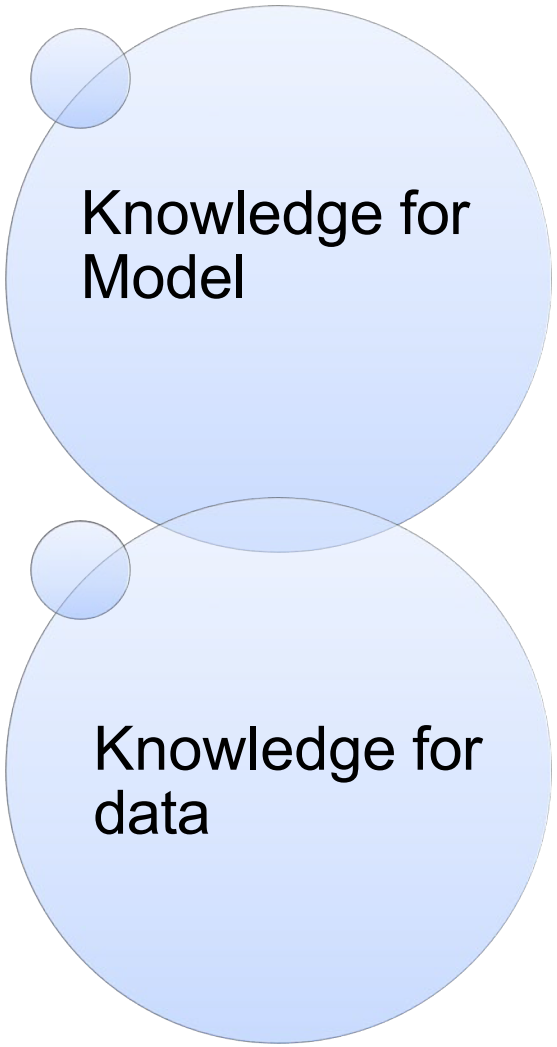
ERINE-ViL [AAAI2021]

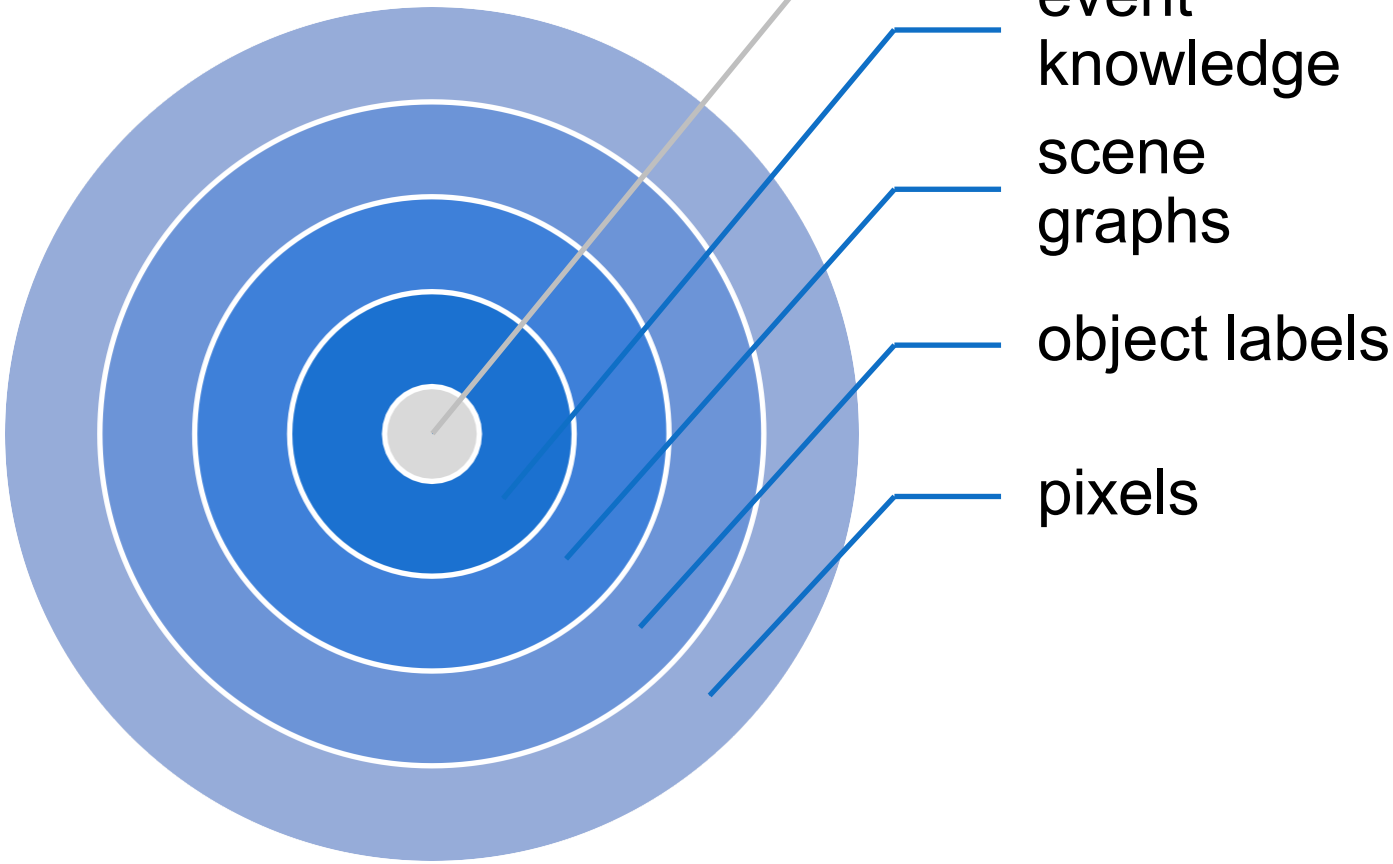
- Add scene graph knowledge as downstream tasks



ERNIE-ViL: Knowledge Enhanced Vision-Language Representations Through Scene Graph, AAAI 2021

Adding knowledge to pretraining models





Vision vs. NLP for Event Extraction

- Vision does not study newsworthy, complex events
  - Focusing on daily life and sports (Perera et al., 2012; Chang et al., 2016; Zhang et al., 2007; Ma et al., 2017)
  - Without localizing a complete set of arguments for each event (Gu et al., 2018; Li et al., 2018; Duarte et al., 2018; Sigurdsson et al., 2016; Kato et al., 2018; Wu et al., 2019a)
- Most related: Situation Recognition (Yatskar et al., 2016)
  - Classify an image as one of 500+ FrameNet verbs
  - Identify 192 generic semantic roles via a 1-word description


CLIPPING

ROLE	VALUE
AGENT	MAN
SOURCE	SHEEP
TOOL	SHEARS
ITEM	WOOL
PLACE	FIELD

ROLE	VALUE
AGENT	VET
SOURCE	DOG
TOOL	CLIPPER
ITEM	CLAW
PLACE	ROOM


JUMPING

ROLE	VALUE
AGENT	BOY
SOURCE	CLIFF
OBSTACLE	-
DESTINATION	WATER
PLACE	LAKE

ROLE	VALUE
AGENT	BEAR
SOURCE	ICEBERG
OBSTACLE	WATER
DESTINATION	ICEBERG
PLACE	OUTDOOR

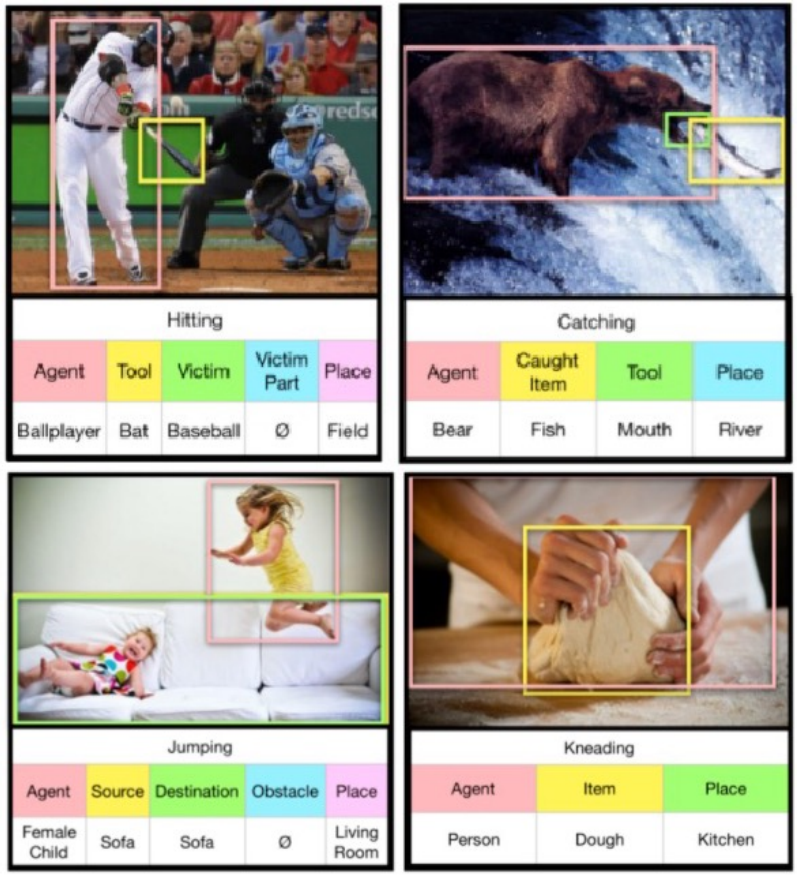

SPRAYING

ROLE	VALUE
AGENT	MAN
SOURCE	SPRAY CAN
SUBSTANCE	PAINT
DESTINATION	WALL
PLACE	ALLEYWAY

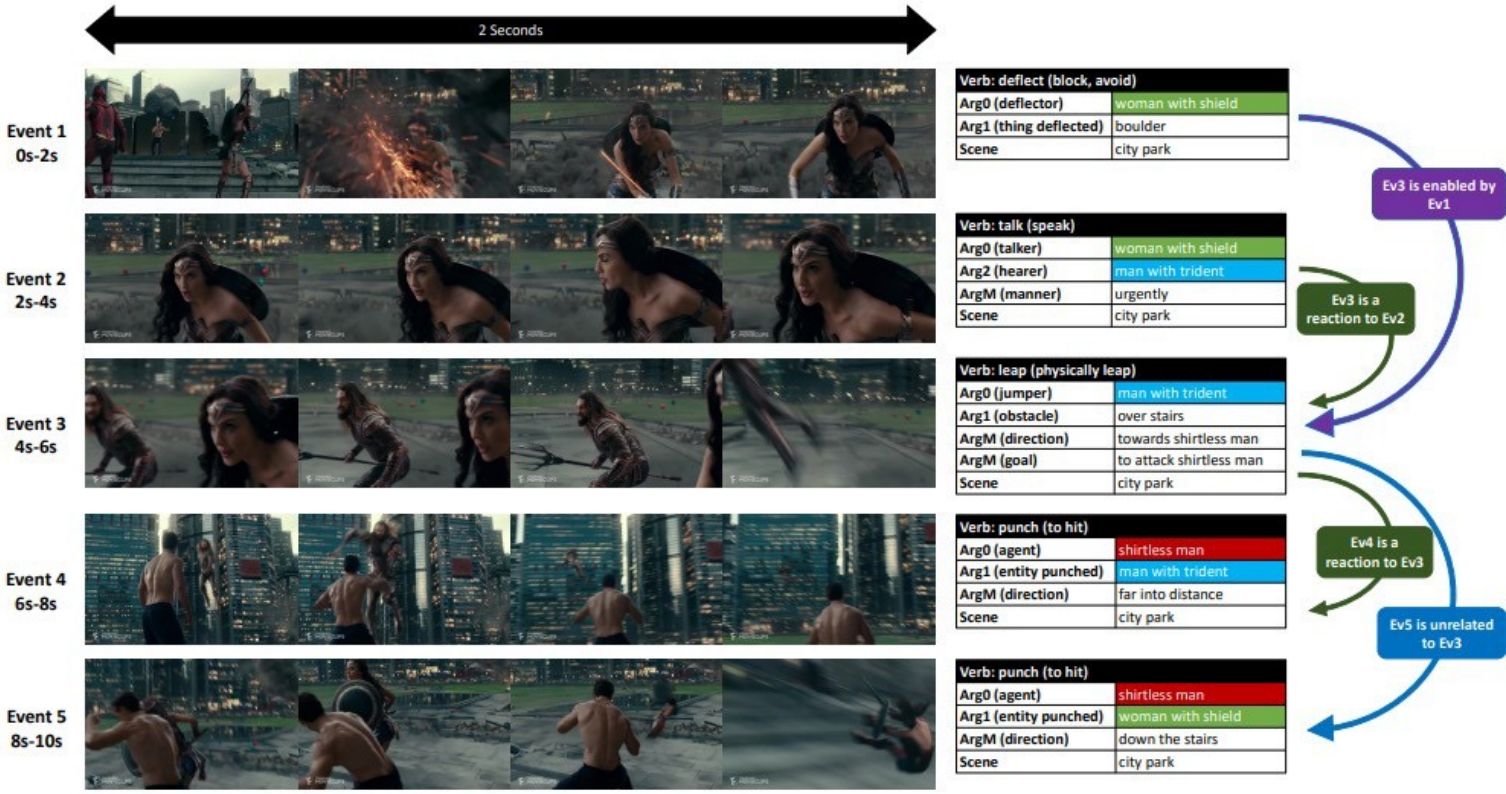
ROLE	VALUE
AGENT	FIREMAN
SOURCE	HOSE
SUBSTANCE	WATER
DESTINATION	FIRE
PLACE	OUTSIDE

Vision-only Event and Argument Extraction

- Grounded Situation Recognition adds visual argument localization [Pratt et al, 2020]

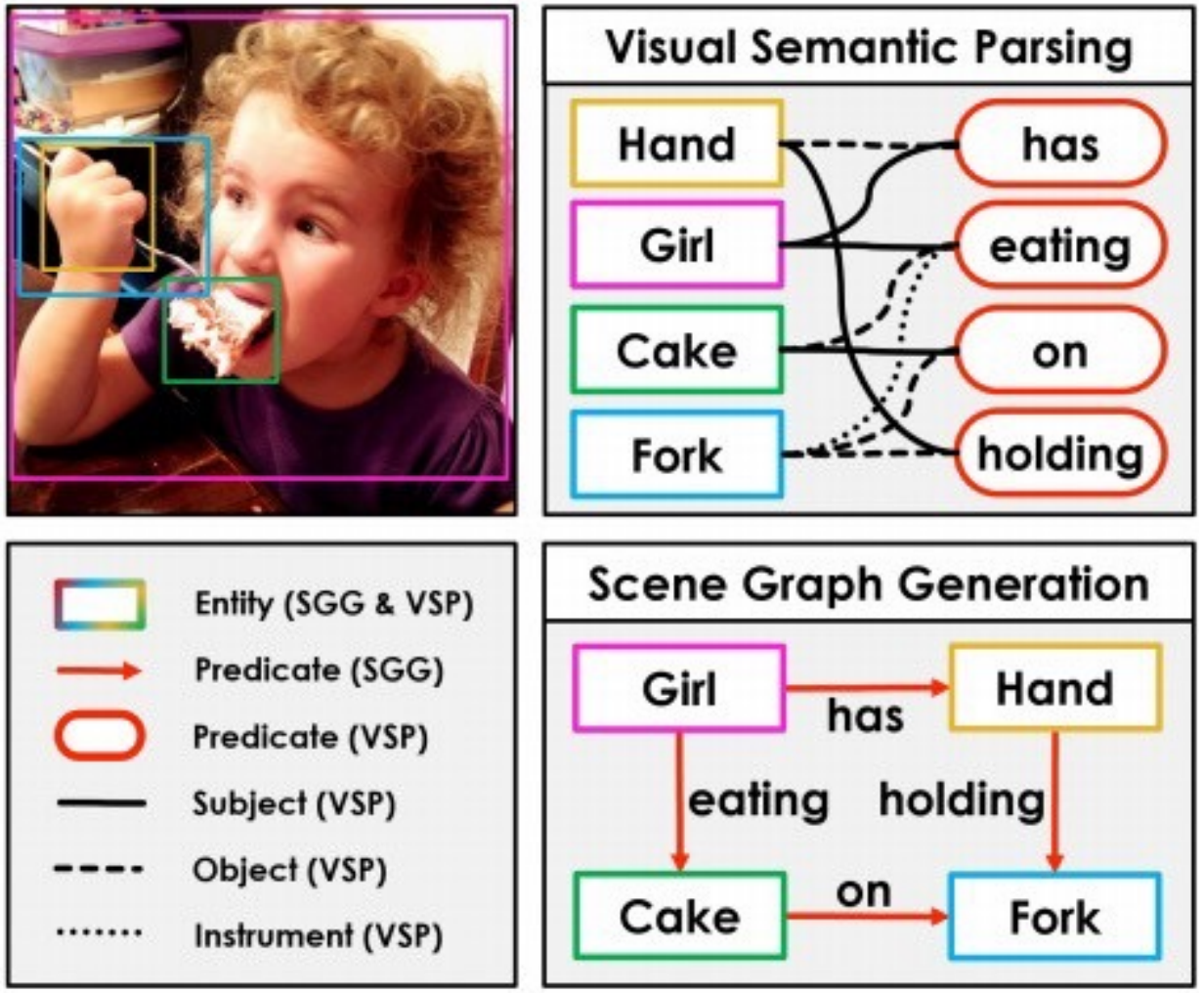


- Video Situation Recognition extends the work to videos [Sadhu et al, 2021]



Vision-only Event and Argument Extraction

- Another line of work is based on scene graphs [Xu et al, 2017; Li et al, 2017; Yang et al, 2018; Zellers et al, 2018].
  - extracting <subject, predicate, object>
  - structure is simpler than the aforementioned multi-argument event
- Visual Semantic Parsing is using predicate as event, and subject, object, instrument as argument [Zareian el al, 2020]
  - Added bounding box grounding



Existing Work: Situation Recognition




SPRAYING

ROLE	VALUE
AGENT	MAN
SOURCE	SPRAY CAN
SUBSTANCE	PAINT
DESTINATION	WALL
PLACE	ALLEYWAY

ROLE	VALUE
AGENT	FIREMAN
SOURCE	HOSE
SUBSTANCE	WATER
DESTINATION	FIRE
PLACE	OUTSIDE

CLIPPING

ROLE	VALUE
AGENT	MAN
SOURCE	SHEEP
TOOL	SHEARS
ITEM	WOOL
PLACE	FIELD

ROLE	VALUE
AGENT	VET
SOURCE	DOG
TOOL	CLIPPER
ITEM	CLAW
PLACE	ROOM

(Yatskar et al., 2016, …)

Surfing

Agent	Tool	Path	Place
Man	Surfboard	Water	Ocean

Feeding

Agent	Food	Eater	Source	Place
Person	Milk	Tiger	Bottle	Ø

(Pratt et al., 2020, …)

Supervised Learning

Bottleneck:
Lack of Annotation

Vision-Only

Bottleneck:
Cross-modal Fusion

Existing Work: Situation Recognition




SPRAYING

ROLE	VALUE
AGENT	MAN
SOURCE	SPRAY CAN
SUBSTANCE	PAINT
DESTINATION	WALL
PLACE	ALLEYWAY

ROLE	VALUE
AGENT	FIREMAN
SOURCE	HOSE
SUBSTANCE	WATER
DESTINATION	FIRE
PLACE	OUTSIDE

CLIPPING

ROLE	VALUE
AGENT	MAN
SOURCE	SHEEP
TOOL	SHEARS
ITEM	WOOL
PLACE	FIELD

ROLE	VALUE
AGENT	VET
SOURCE	DOG
TOOL	CLIPPER
ITEM	CLAW
PLACE	ROOM

(Yatskar et al., 2016, …)


Surfing

Agent	Tool	Path	Place
Man	Surfboard	Water	Ocean

Feeding

Agent	Food	Eater	Source	Place
Person	Milk	Tiger	Bottle	Ø

(Pratt et al., 2020, …)

Supervised Learning

Bottleneck:
Lack of Annotation

Vision-Only

Bottleneck:
Cross-modal Fusion

Transfer

Language → **Vision**

Existing Work: Situation Recognition




SPRAYING

ROLE	VALUE
AGENT	MAN
SOURCE	SPRAY CAN
SUBSTANCE	PAINT
DESTINATION	WALL
PLACE	ALLEYWAY

ROLE	VALUE
AGENT	FIREMAN
SOURCE	HOSE
SUBSTANCE	WATER
DESTINATION	FIRE
PLACE	OUTSIDE

CLIPPING

ROLE	VALUE
AGENT	MAN
SOURCE	SHEEP
TOOL	SHEARS
ITEM	WOOL
PLACE	FIELD

ROLE	VALUE
AGENT	VET
SOURCE	DOG
TOOL	CLIPPER
ITEM	CLAW
PLACE	ROOM

(Yatskar et al., 2016, …)


Surfing

Agent	Tool	Path	Place
Man	Surfboard	Water	Ocean

Feeding

Agent	Food	Eater	Source	Place
Person	Milk	Tiger	Bottle	Ø

(Pratt et al., 2020, …)

Supervised Learning

Bottleneck:
Lack of Annotation

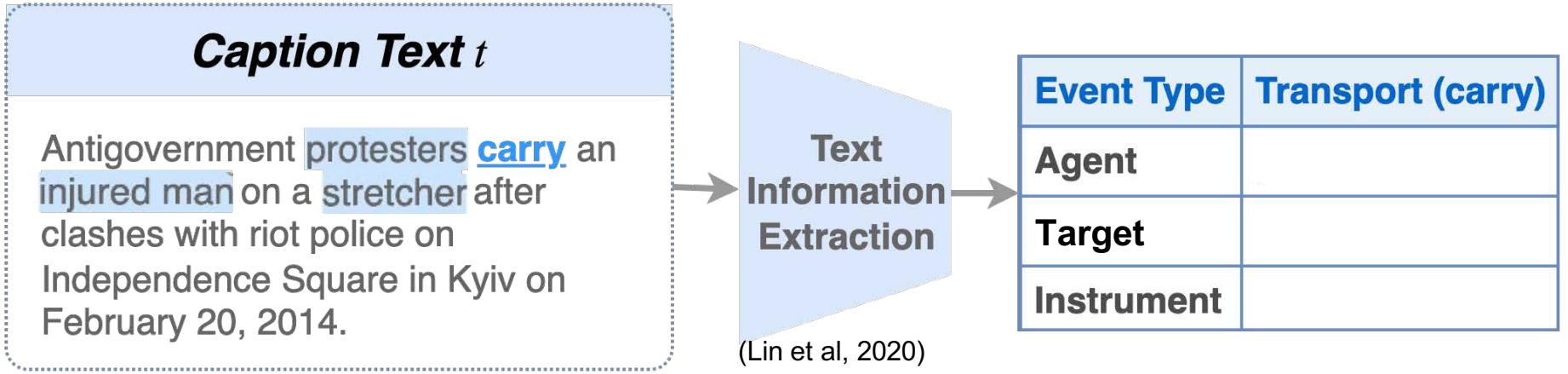
Vision-Only

Bottleneck:
Cross-modal Fusion

Structural Transfer

Language → **Vision**

CLIP-Event: Event-Driven Vision-Language Pretraining



(Lin et al, 2020)

Transfer text event knowledge to images

Caption Text t

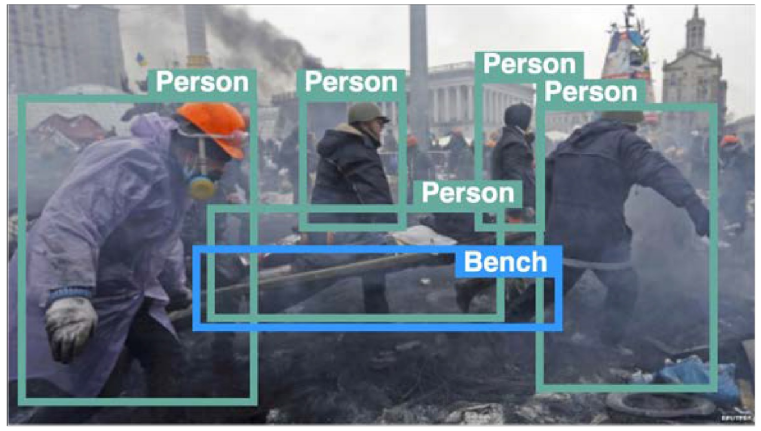
Antigovernment protesters carry an injured man on a stretcher after clashes with riot police on Independence Square in Kyiv on February 20, 2014.

Text Information Extraction

(Lin et al, 2020)

Event Type	Transport (carry)
Agent	protesters
Target	injured man
Instrument	stretcher

Image i



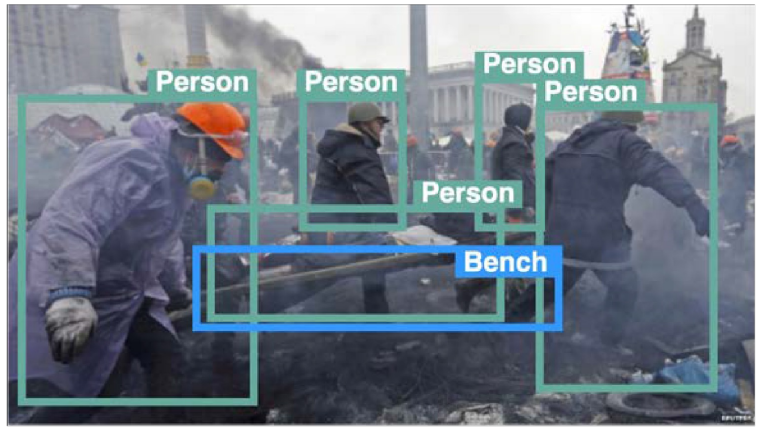
Weakly Supervision

Hard negatives via manipulating event structures

Caption Text t

Antigovernment protesters carry an injured man on a stretcher after clashes with riot police on Independence Square in Kyiv on February 20, 2014.

Image i



Positive Labels

Event Type	Transport (carry)
Agent	protesters
Target	injured man
Instrument	stretcher

Confusion Matrix of existing V+L models

Negative Labels (events)

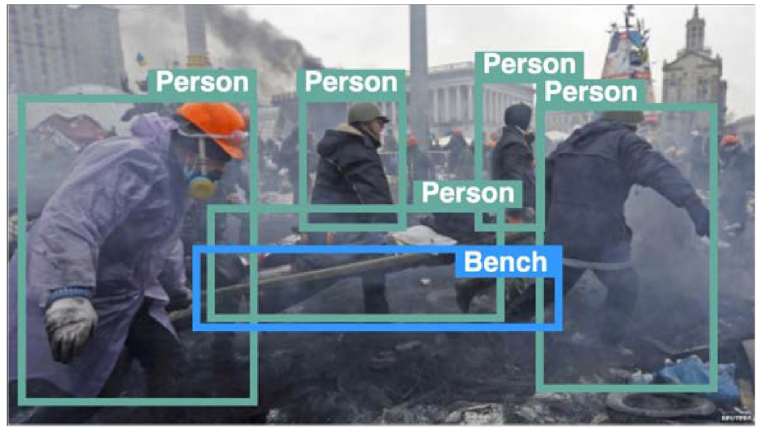
Event Type	Arrest (arrest)
Agent	protesters
Target	injured man
Instrument	stretcher

Hard negatives via manipulating event structures

Caption Text t

Antigovernment protesters carry an injured man on a stretcher after clashes with riot police on Independence Square in Kyiv on February 20, 2014.

Image i



Positive Labels

Event Type	Transport (carry)
Agent	protesters
Target	injured man
Instrument	stretcher

Negative Labels (events)

Event Type	Arrest (arrest)
Agent	protesters
Target	injured man
Instrument	stretcher

Negative Labels (arguments)

Event Type	Transport (carry)
Agent	protesters
Target	injured man
Instrument	stretcher

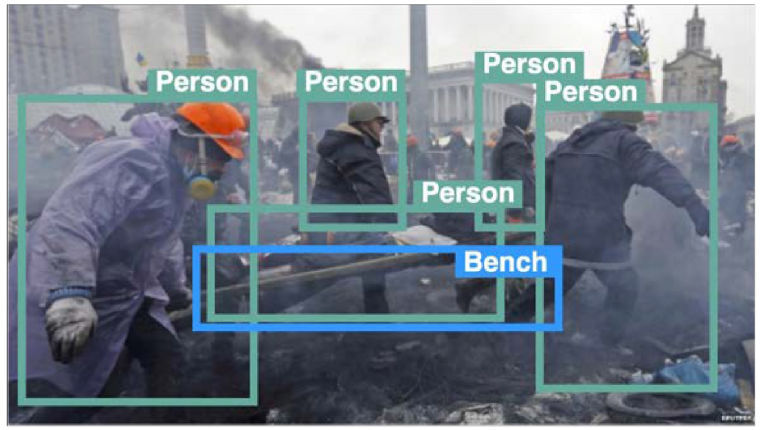
Role Switching

Hard negatives via manipulating event structures

Caption Text t

Antigovernment protesters carry an injured man on a stretcher after clashes with riot police on Independence Square in Kyiv on February 20, 2014.

Image i



Positive Labels

Event Type	Transport (carry)
Agent	protesters
Target	injured man
Instrument	stretcher

prompt → Protesters transported injured man using a stretcher.

Negative Labels (events)

Event Type	Arrest (arrest)
Agent	protesters
Target	injured man
Instrument	stretcher

prompt → Protesters arrested injured man using a stretcher.

Negative Labels (arguments)

Event Type	Transport (carry)
Agent	injured man
Target	stretcher
Instrument	protesters

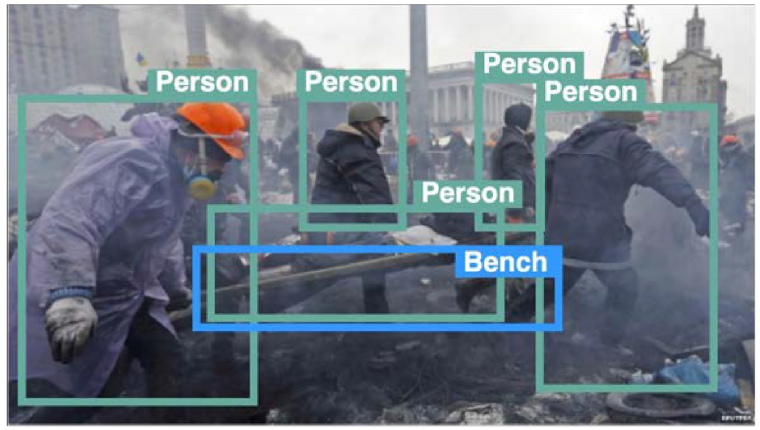
prompt → Injured man transported a stretcher with protesters.

Hard negatives via manipulating event structures

Caption Text t

Antigovernment protesters carry an injured man on a stretcher after clashes with riot police on Independence Square in Kyiv on February 20, 2014.

Image i



Positive Labels

Event Type	Transport (carry)
Agent	protesters
Entity	injured man
Instrument	stretcher

prompt

Protesters transported injured man using a stretcher.

Negative Labels (events)

Event Type	Arrest (arrest)
Agent	protesters
Entity	injured man
Instrument	stretcher

prompt

Protesters arrested injured man using a stretcher.

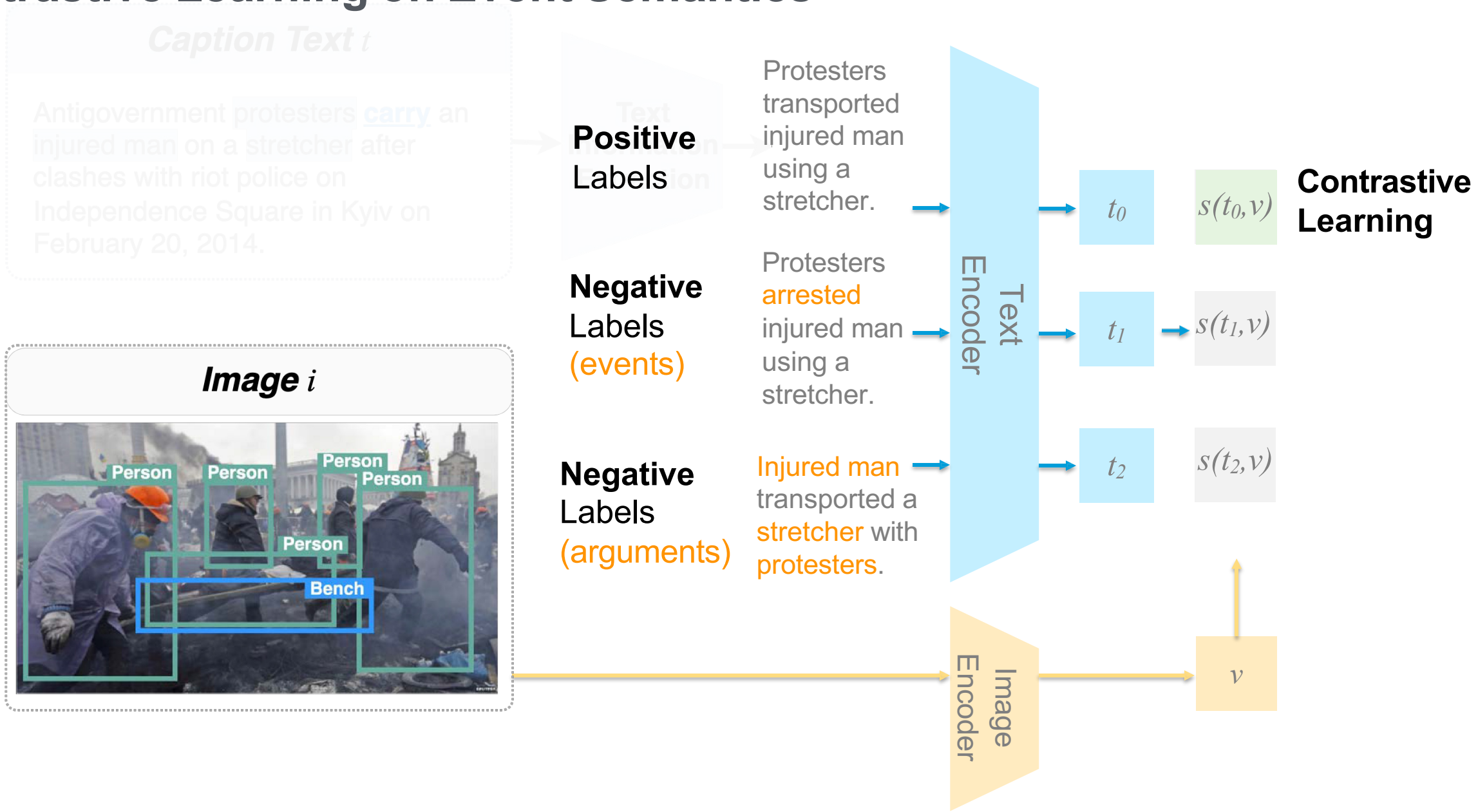
Negative Labels (arguments)

Event Type	Transport (carry)
Agent	injured man
Entity	stretcher
Instrument	protesters

prompt

Injured man transported a stretcher with protesters.

Contrastive Learning on Event Semantics



Bottlenecks of Vision Semantic Structure Learning




SPRAYING

ROLE	VALUE
AGENT	MAN
SOURCE	SPRAY CAN
SUBSTANCE	PAINT
DESTINATION	WALL
PLACE	ALLEYWAY

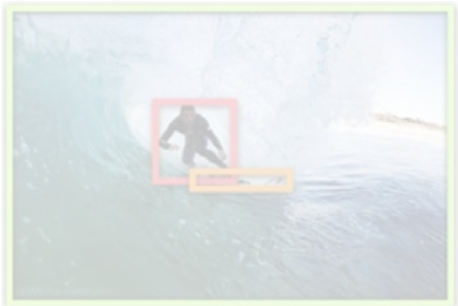
ROLE	VALUE
AGENT	FIREMAN
SOURCE	HOSE
SUBSTANCE	WATER
DESTINATION	FIRE
PLACE	OUTSIDE

CLIPPING

ROLE	VALUE
AGENT	MAN
SOURCE	SHEEP
TOOL	SHEARS
ITEM	WOOL
PLACE	FIELD

ROLE	VALUE
AGENT	VET
SOURCE	DOG
TOOL	CLIPPER
ITEM	CLAW
PLACE	ROOM

(Yatskar et al., 2016, …)




Surfing

Agent	Tool	Path	Place
Man	Surfboard	Water	Ocean

Feeding

Agent	Food	Eater	Source	Place
Person	Milk	Tiger	Bottle	Ø

(Pratt et al., 2020, …)

Supervised Learning

Bottleneck:
Lack of Annotation

Vision-Only

Bottleneck:
Cross-modal Fusion

Structural Transfer

Language → **Vision**

Bottlenecks of Vision Semantic Structure Learning




SPRAYING

ROLE	VALUE
AGENT	MAN
SOURCE	SPRAY CAN
SUBSTANCE	PAINT
DESTINATION	WALL
PLACE	ALLEYWAY

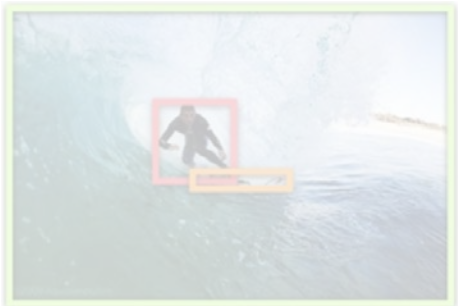
ROLE	VALUE
AGENT	FIREMAN
SOURCE	HOSE
SUBSTANCE	WATER
DESTINATION	FIRE
PLACE	OUTSIDE

CLIPPING

ROLE	VALUE
AGENT	MAN
SOURCE	SHEEP
TOOL	SHEARS
ITEM	WOOL
PLACE	FIELD

ROLE	VALUE
AGENT	VET
SOURCE	DOG
TOOL	CLIPPER
ITEM	CLAW
PLACE	ROOM

(Yatskar et al., 2016, …)




Surfing

Agent	Tool	Path	Place
Man	Surfboard	Water	Ocean

Feeding

Agent	Food	Eater	Source	Place
Person	Milk	Tiger	Bottle	Ø

(Pratt et al., 2020, …)

Supervised Learning

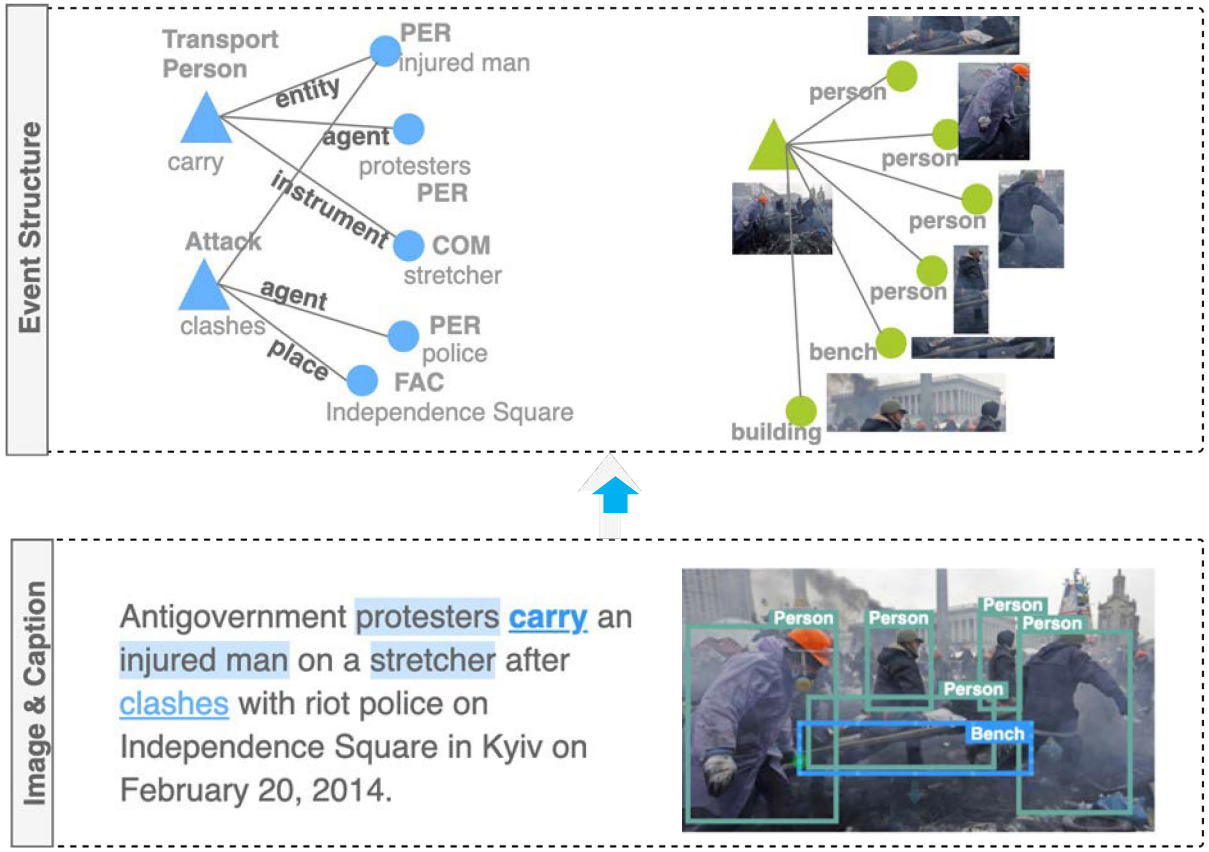
Bottleneck:
Lack of Annotation

Vision-Only

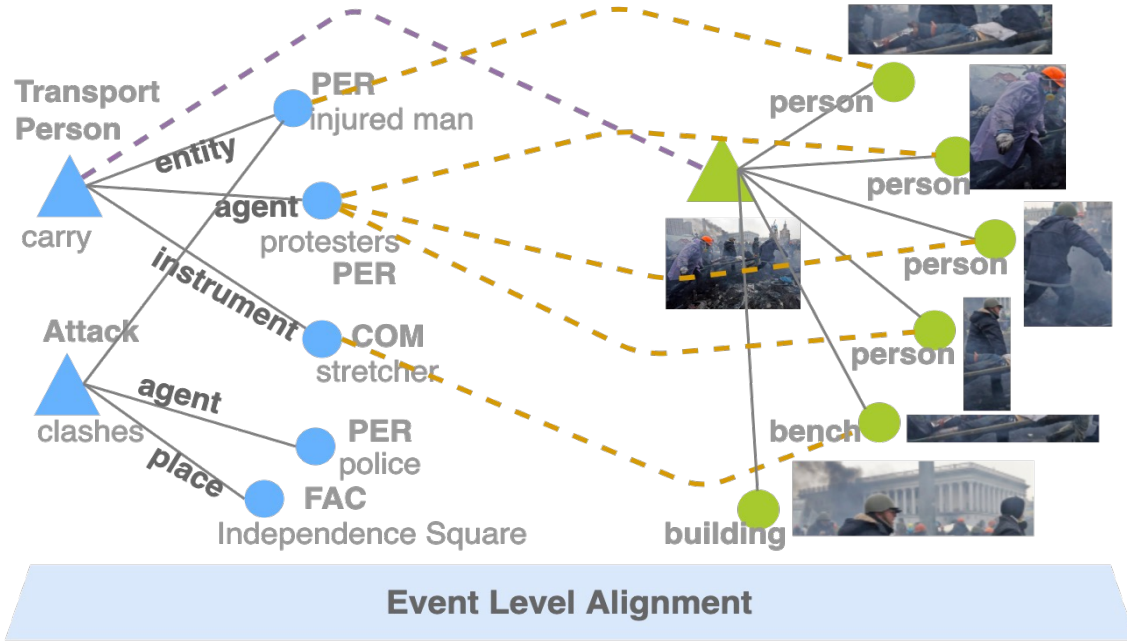
Bottleneck:
Cross-modal Fusion

The first V+L Pretraining with Event Semantic Structures

Challenge: Structured Encoding



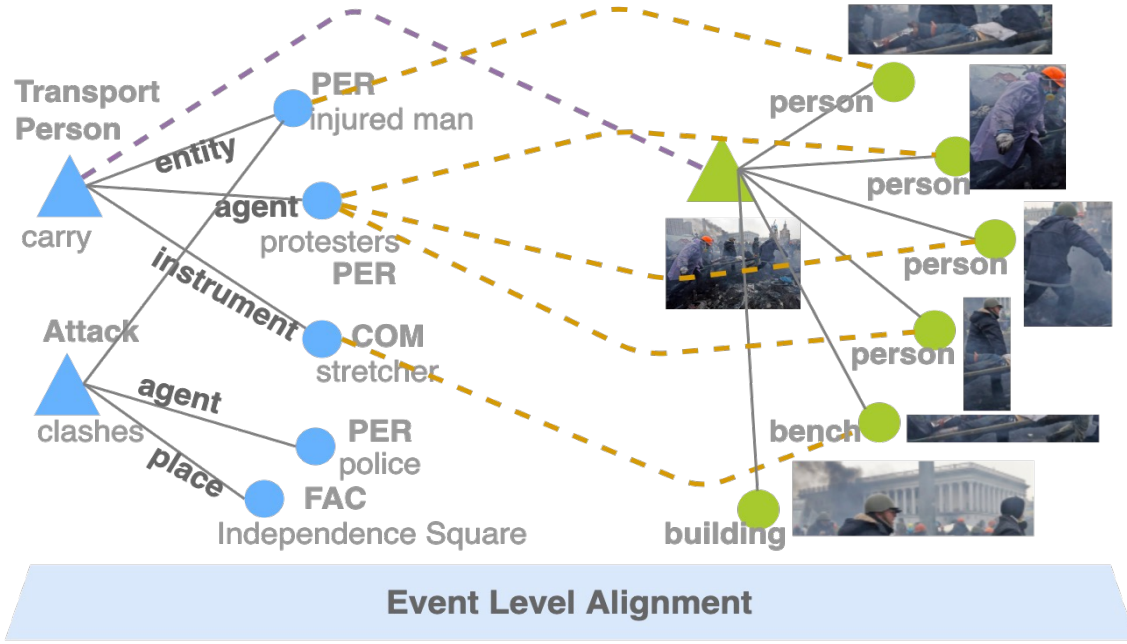
The first V+L Pretraining with Event Semantic Structures



The first V+L Pretraining with Event Semantic Structures

Structured Alignment via Optimal Transport

Text Event Graph ⟺ Image Event Graph

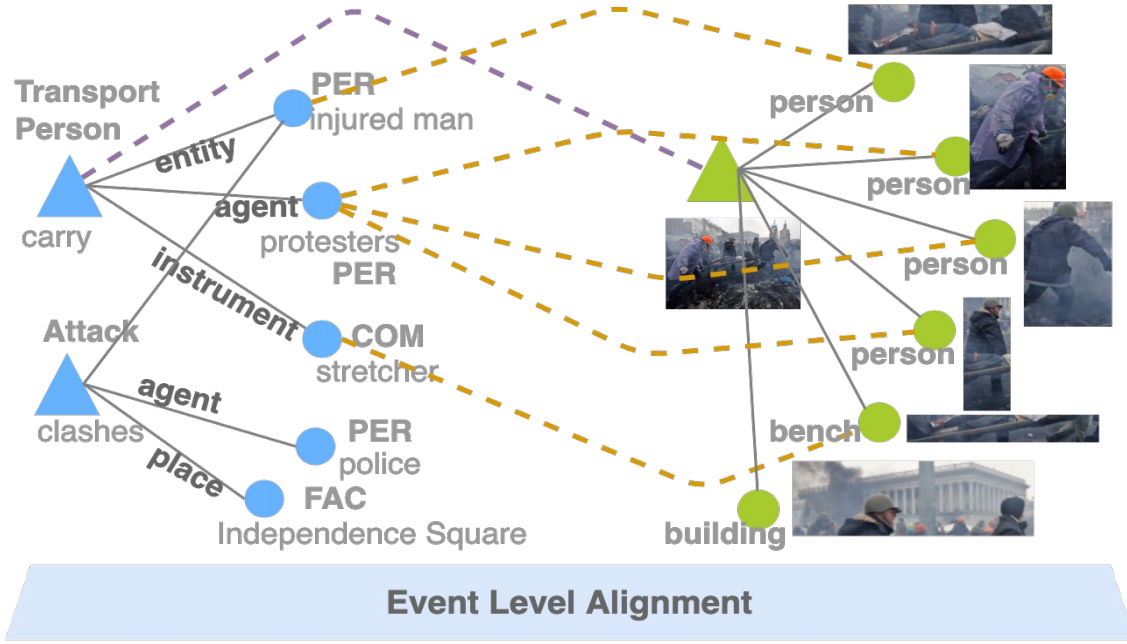


The first V+L Pretraining with Event Semantic Structures

Structured Alignment via Optimal Transport

Text Event Graph ⟺ Image Event Graph

1 Define cost matrix C (embedding similarity)



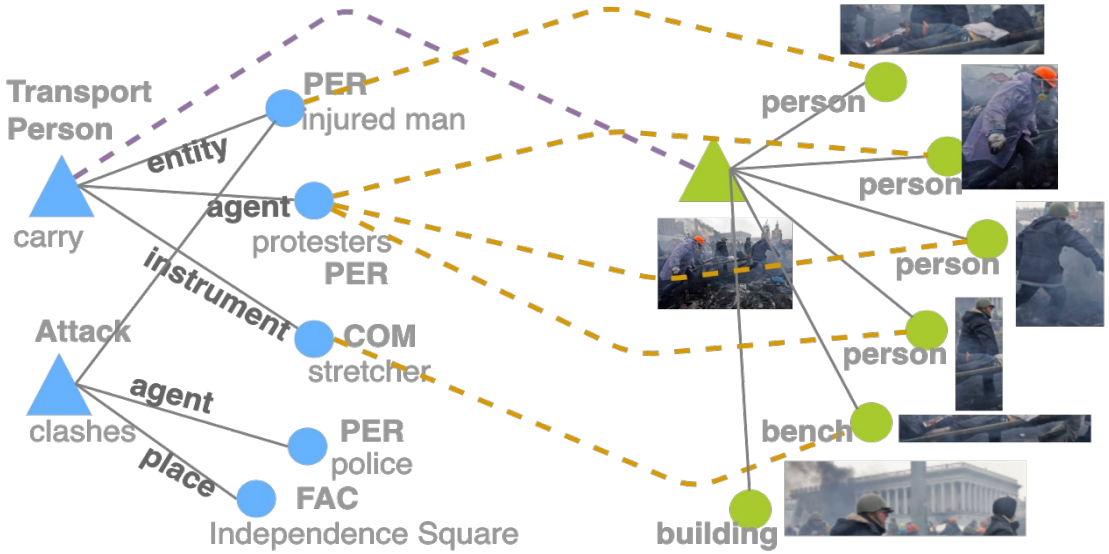
The first V+L Pretraining with Event Semantic Structures

Structured Alignment via Optimal Transport

Text Event Graph ⟷ Image Event Graph

1. Define cost matrix C (embedding similarity)

2. Optimization Goal: minimize transport distance
$D(S,T) = \min \boldsymbol{T} * \boldsymbol{C}$



Event Level Alignment

The optimal $\boldsymbol{T}$ is approximated by a differentiable Sinkhorn Knopp algorithm (Sinkhorn, 1964; Cuturi, 2013)

$$\boldsymbol{T} = \mathrm{diag}(\boldsymbol{p}) \exp(-\boldsymbol{C}/\gamma) \mathrm{diag}(\boldsymbol{q})$$

$$\text{for } i = 0, 1, 2, \ldots \text{ until convergence,}$$

$$\boldsymbol{p}^{i+1} = \mathbf{1} \oslash (\boldsymbol{K}\boldsymbol{q}^{i}),$$

$$\boldsymbol{q}^{i+1} = \mathbf{1} \oslash (\boldsymbol{K}^{\top}\boldsymbol{p}^{i+1}),$$

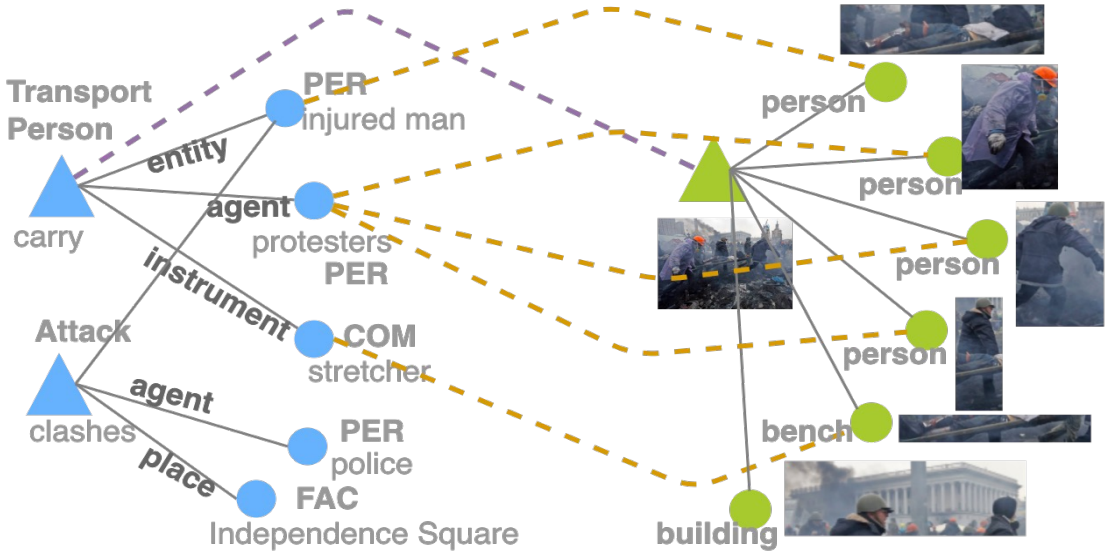
$$\boldsymbol{T}^{k} := \mathrm{diag}(\boldsymbol{p}^{k})\boldsymbol{K}\mathrm{diag}(\boldsymbol{q}^{k})$$

The first V+L Pretraining with Event Semantic Structures

Structured Alignment via Optimal Transport

Text Event Graph ⟷ Image Event Graph

1. Define cost matrix C (embedding similarity)
2. Optimization Goal: minimize transport distance $D(S,T) = \min \boldsymbol{T} * \boldsymbol{C}$
3. Optimize the transport plan T within k iterations



Event Level Alignment

The optimal T is approximated by a differentiable Sinkhorn Knopp algorithm (Sinkhorn, 1964; Cuturi, 2013)

$$\boldsymbol{T} = \mathrm{diag}(\boldsymbol{p}) \exp(-\boldsymbol{C}/\gamma) \mathrm{diag}(\boldsymbol{q})$$

$$\text{for } i = 0, 1, 2, \ldots \text{ until convergence,}$$

$$\boldsymbol{p}^{i+1} = \mathbf{1} \oslash (\boldsymbol{K}\boldsymbol{q}^{i}),$$

$$\boldsymbol{q}^{i+1} = \mathbf{1} \oslash (\boldsymbol{K}^{\top}\boldsymbol{p}^{i+1}),$$

$$\boldsymbol{T}^{k} := \mathrm{diag}(\boldsymbol{p}^{k})\boldsymbol{K}\mathrm{diag}(\boldsymbol{q}^{k})$$

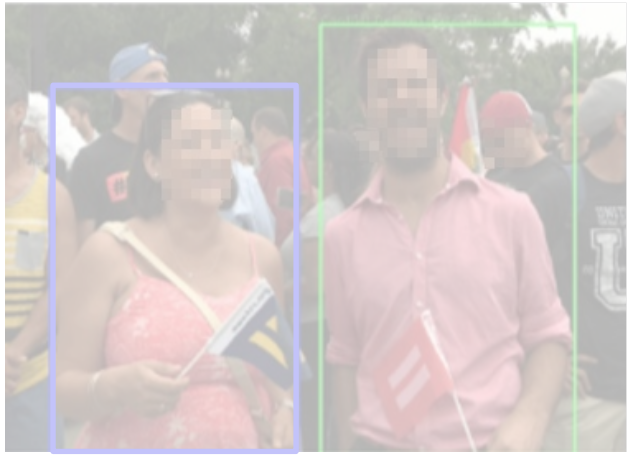
CLIP-Event on Visual Event Extraction

Supporting Zero-shot Vision Event Extraction the first time.

Vaccination

Arrest

Protest

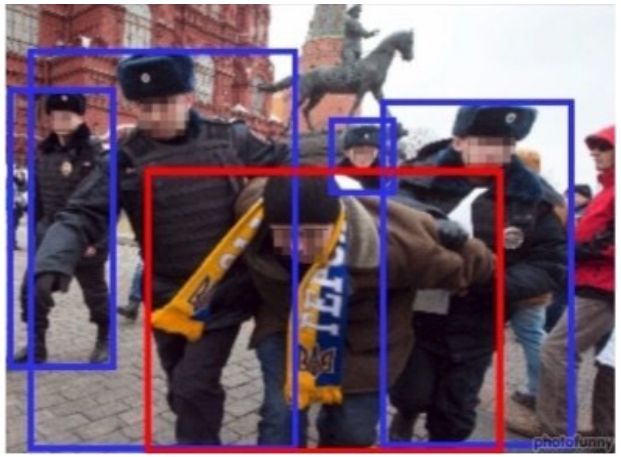


Celebration

CLIP-Event on Visual Event Extraction

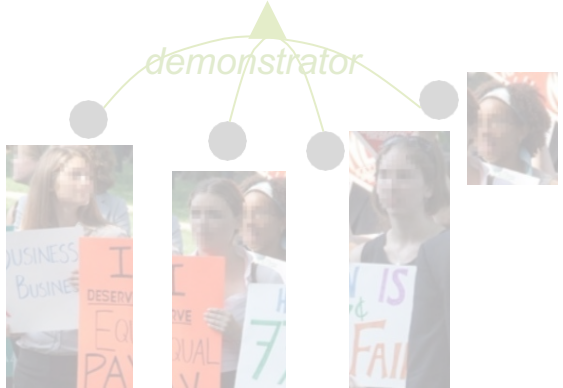
Supporting Zero-shot Vision Event Extraction the first time.

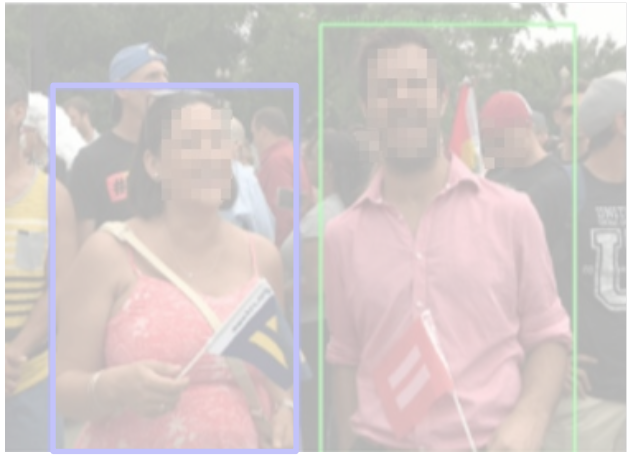
Vaccination



Arrest

Protest

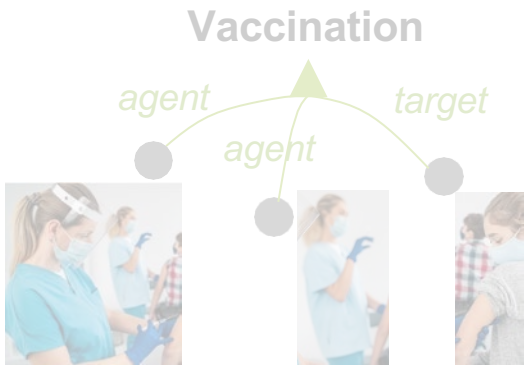


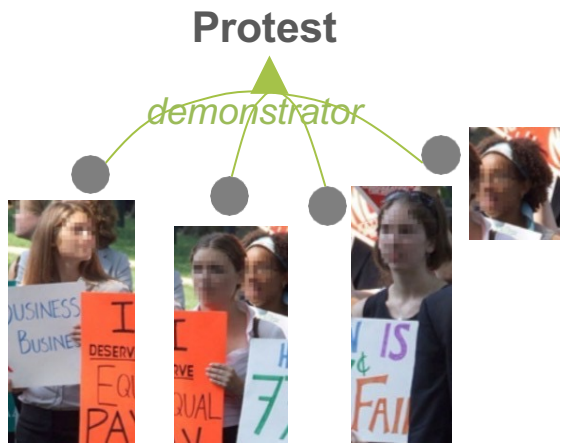


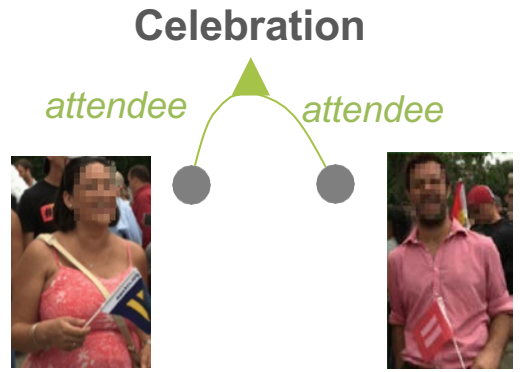
Celebration

CLIP-Event on Visual Event Extraction

Supporting Zero-shot Vision Event Extraction the first time.







To be continued…