CVPR

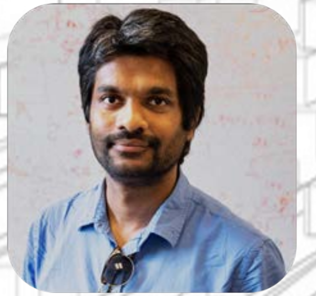
Jun 2023 9:00-12:30
CVPR Tutorials
East 8, Vancouver Convention Center

Knowledge-Driven Vision-Language Encoding

Manling Li
UIUC
(Incoming AP at Northwestern)

Xudong Lin
Columbia

Jie Lei
Meta AI



Mohit Bansal
UNC

Carl Vondrick
Columbia

Shih-Fu Chang
Columbia

Heng Ji
UIUC

How humans learn about knowledge?

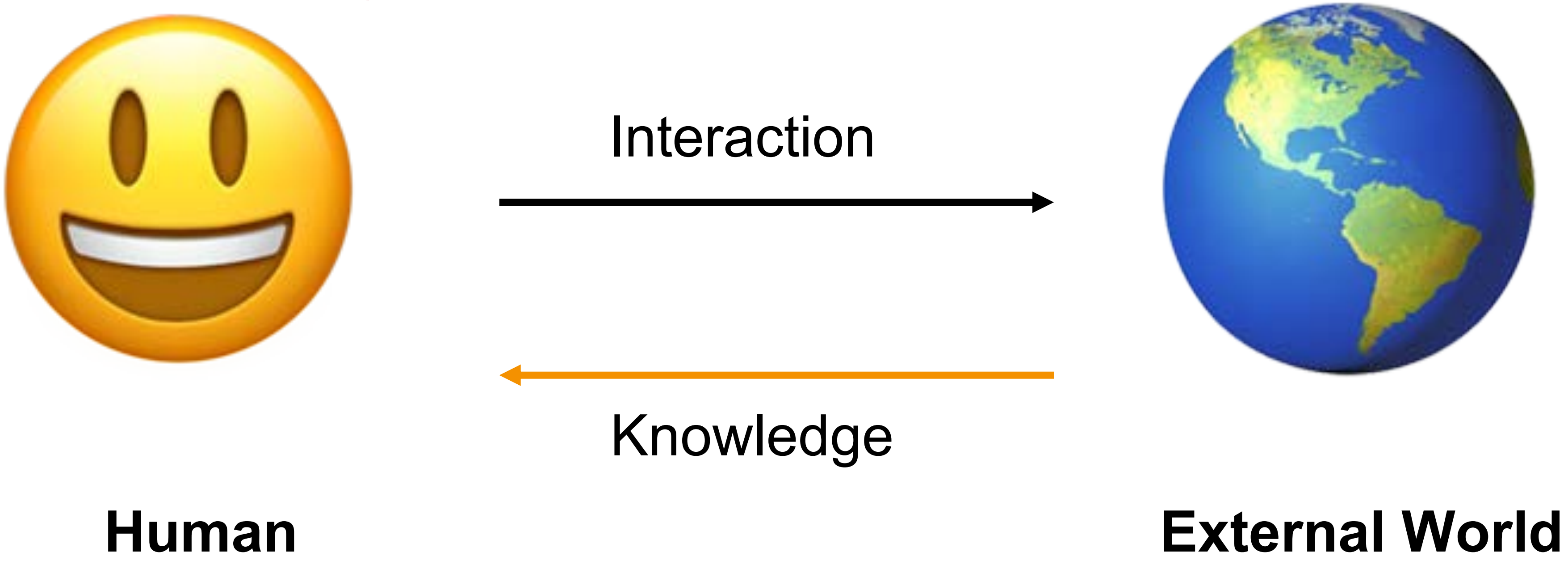


Figure Credits: Zhenhailong Wang

How humans learn about knowledge?

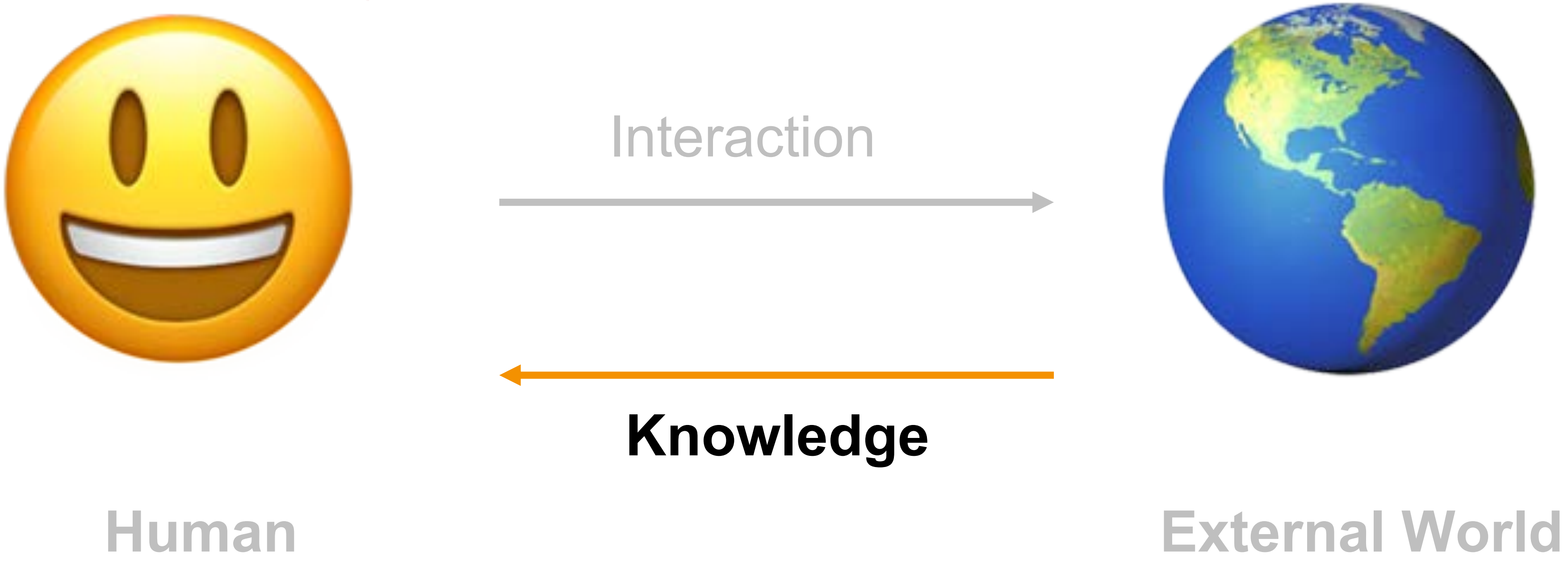
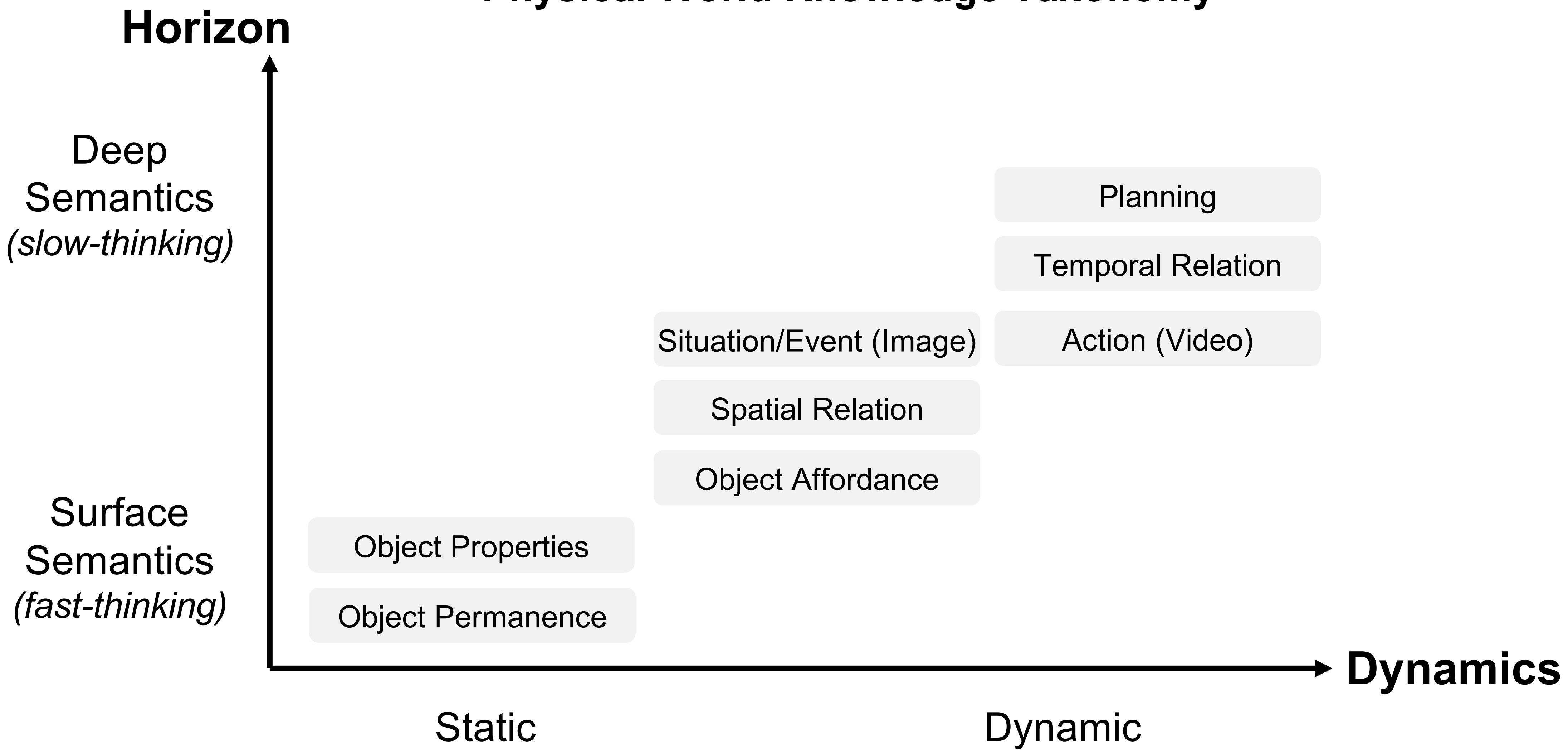


Figure Credits: Zhenhailong Wang

What is the knowledge that humans aim to learn?

Physical World Knowledge Taxonomy



Worked with Zhenhailong Wang

Machines learn knowledge through multi-sensory interactions

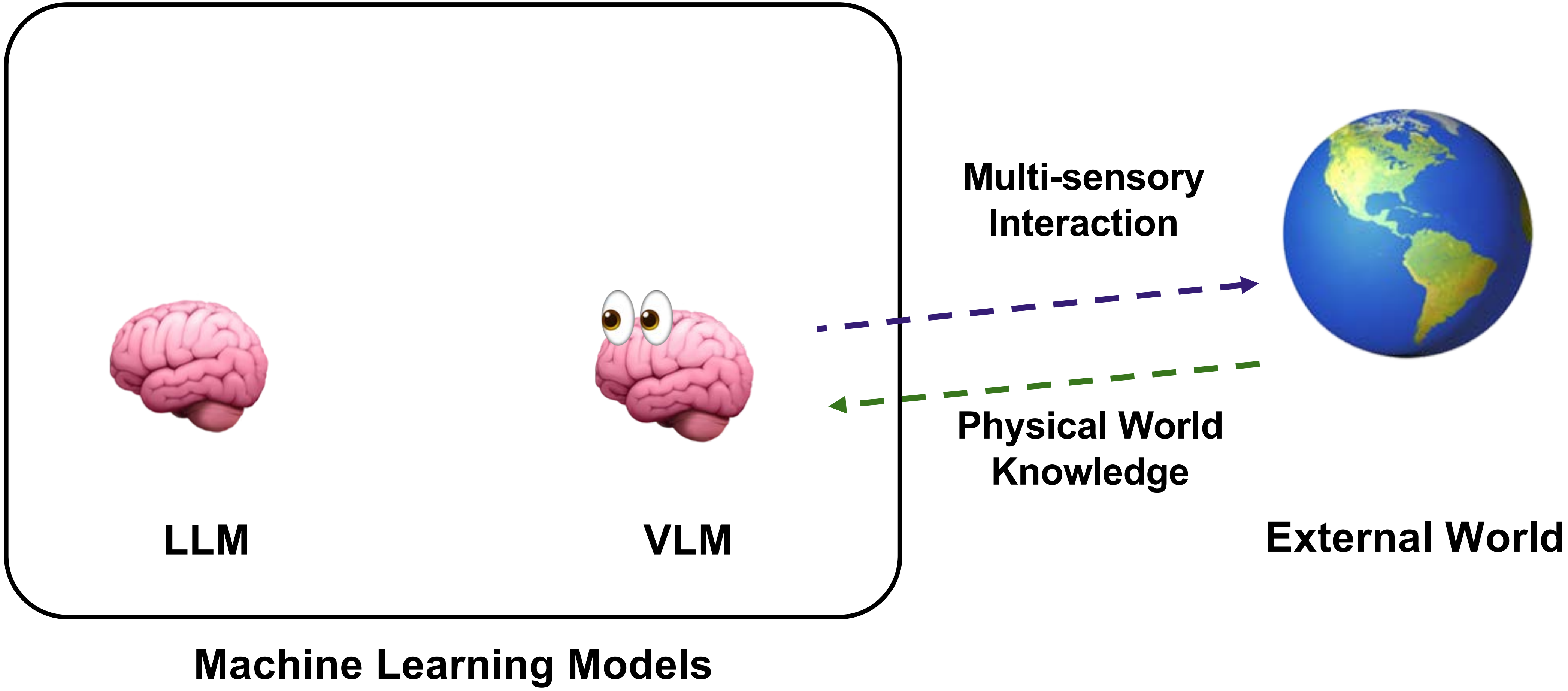


Figure Credits: Zhenhailong Wang

Machines learn knowledge through multi-sensory interactions

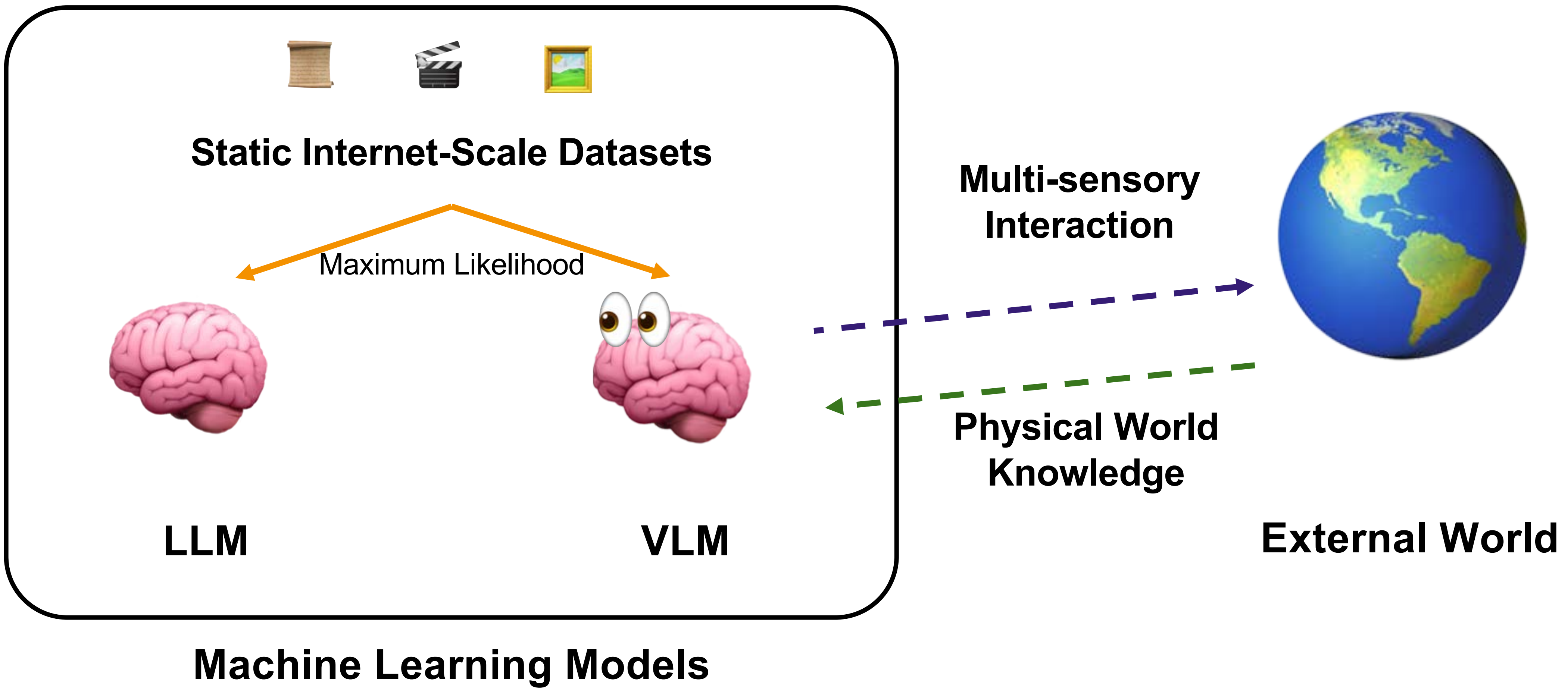
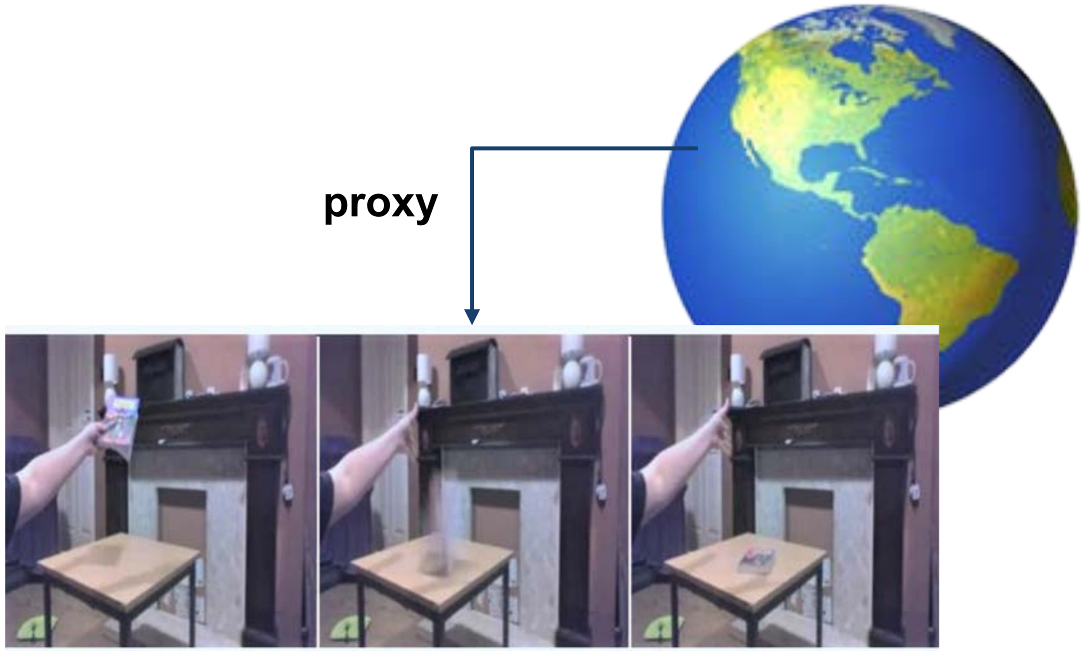


Figure Credits: Zhenhailong Wang

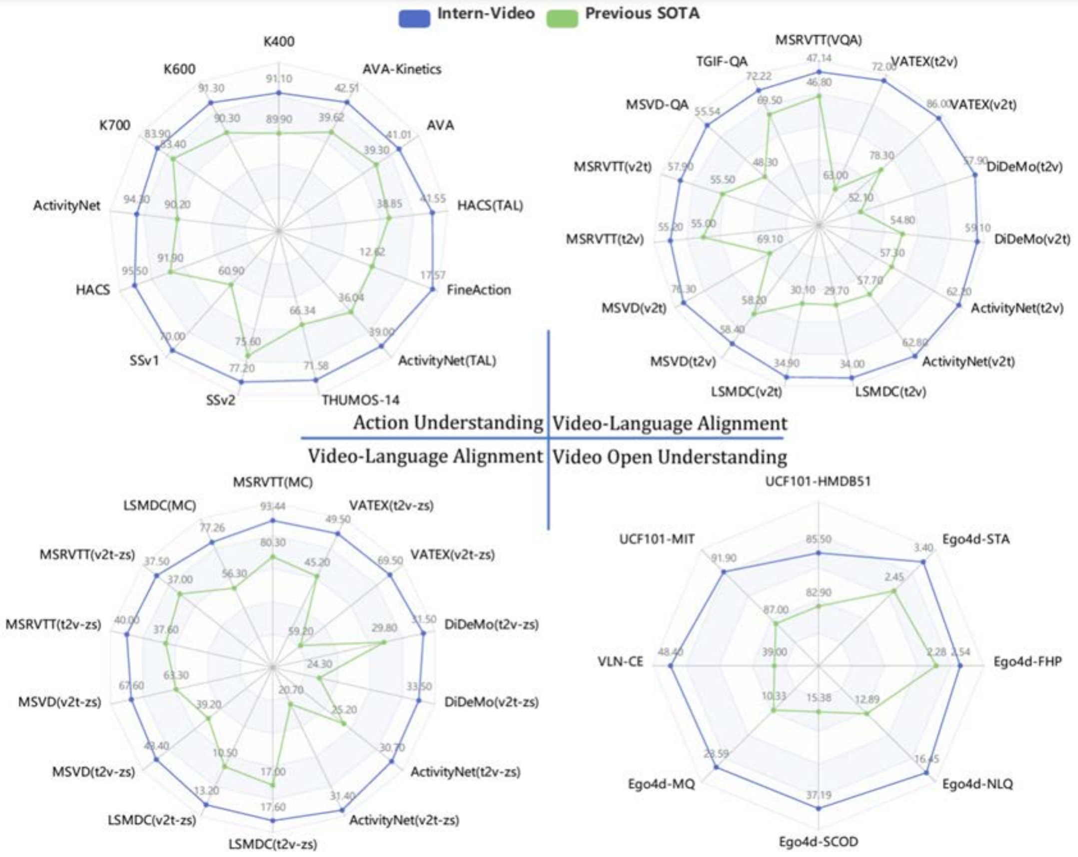
Video: A "Visual Recording" of World State Changes

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



"Book falling like a rock"

Video-Language Datasets



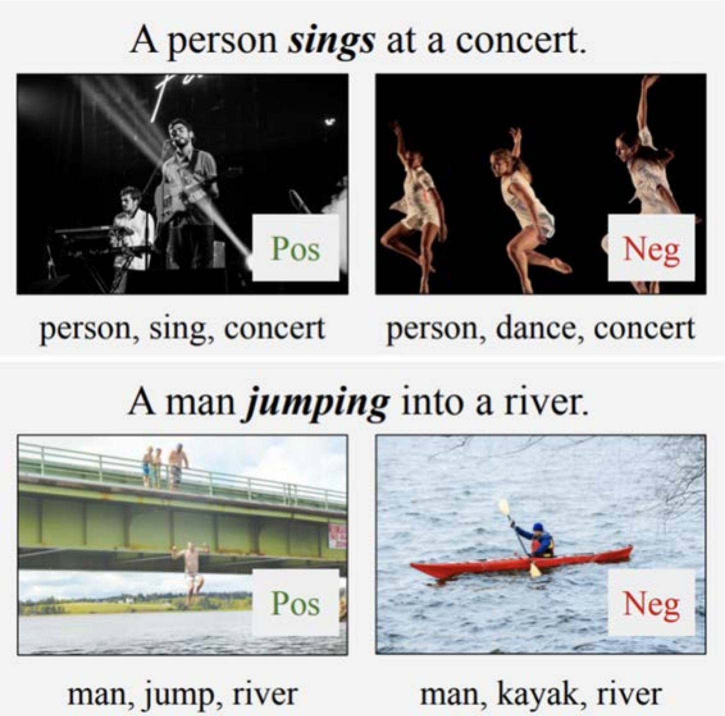
InternVideo (2023)

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.

BLIP

the grass is eating the horse	**81%**
the horse is eating the grass	78%

Bags of Words

"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

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

Method	#PT	#Train Frame	MSRVTT			DiDeMo			ActivityNet Cap		
			R1	R5	R10	R1	R5	R10	R1	R5	R10
HERO [37]	136M	310	20.5	47.6	60.9	-	-	-	-	-	-
ClipBERT [31]	0.2M	16/16/8	22.0	46.8	59.9	20.4	48.0	60.8	21.3	49.0	63.5
VideoCLIP [61]	136M	960	30.9	55.4	66.8	-	-	-	-	-	-
Frozen [4]	5M	4	31.0	59.5	70.5	31.0	59.8	72.4	-	-	-
AlignPrompt [34]	5M	8	33.9	60.7	73.2	35.9	67.5	78.8	-	-	-
All-in-one [58]	138M	9	34.4	65.4	75.8	32.7	61.4	73.5	22.4	53.7	67.7
CLIP4Clip [47]	400M	12/64/64	42.0	68.6	78.7	42.8	68.5	79.2	40.5	72.4	98.2
SINGULARITY	5M	1	36.8	65.9	75.5	47.4	75.2	84.0	43.0	70.6	81.3
SINGULARITY	17M	1	**41.5**	**68.7**	**77.0**	**53.9**	**79.4**	**86.9**	**47.1**	**75.5**	**85.5**

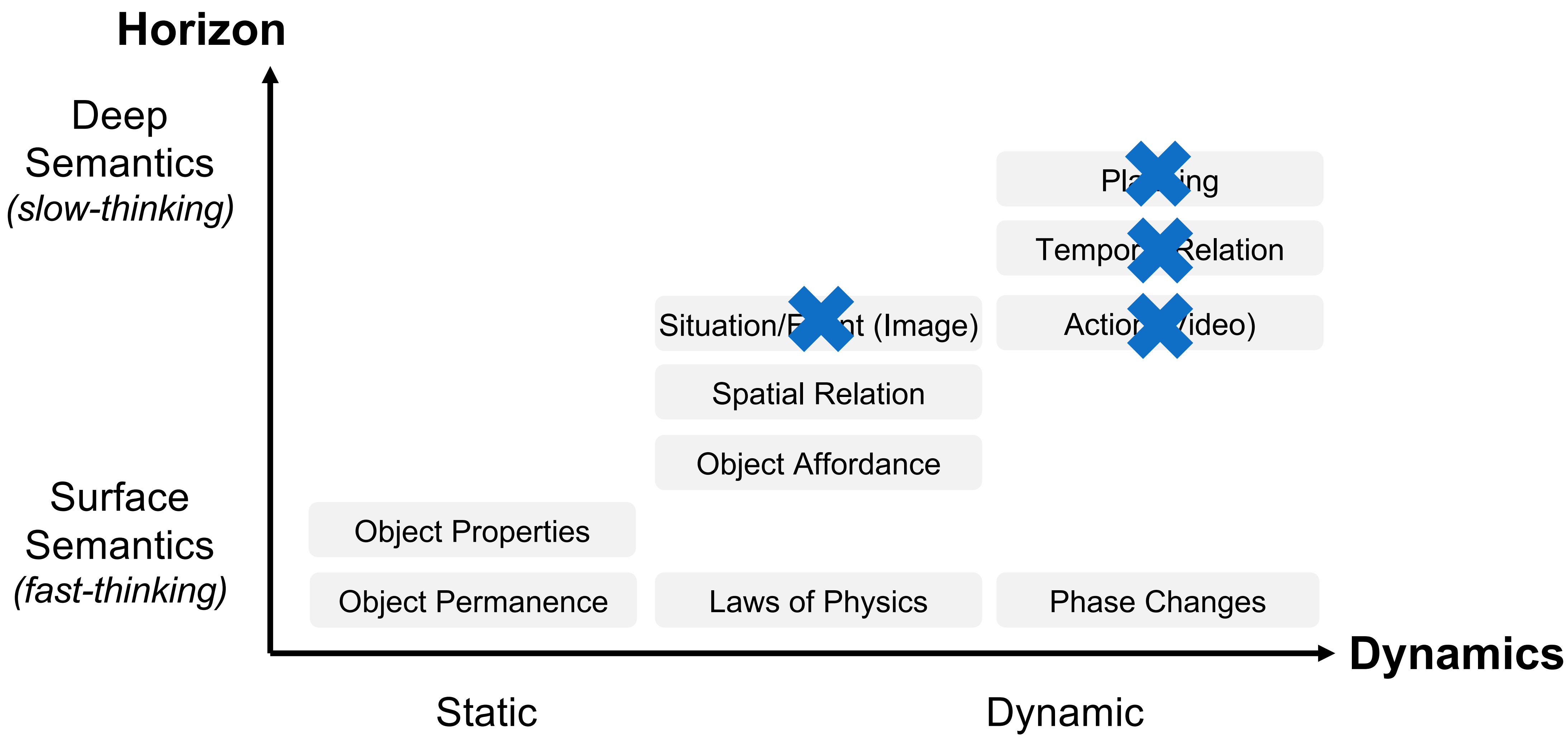
Single-Frame

"Revealing Single Frame Bias for Video-and-Language Learning" Jie Lei, et al. (ACL23)

Figure Credits: Zhenhailong Wang

Existing V+L Models

Physical Knowledge Taxonomy (🚧)



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

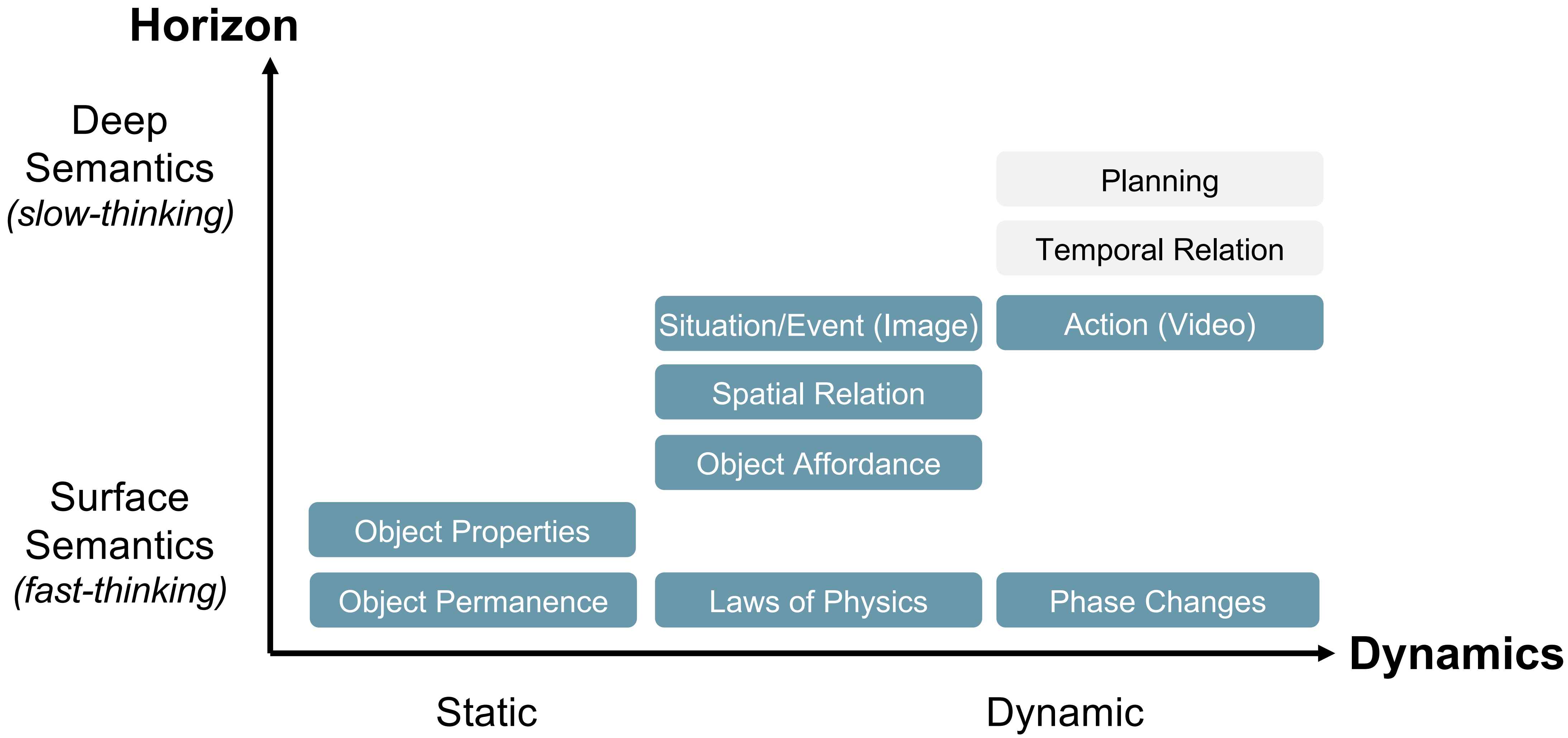
Factual Knowledge are information about **instance-level facts** extracted from raw data.

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

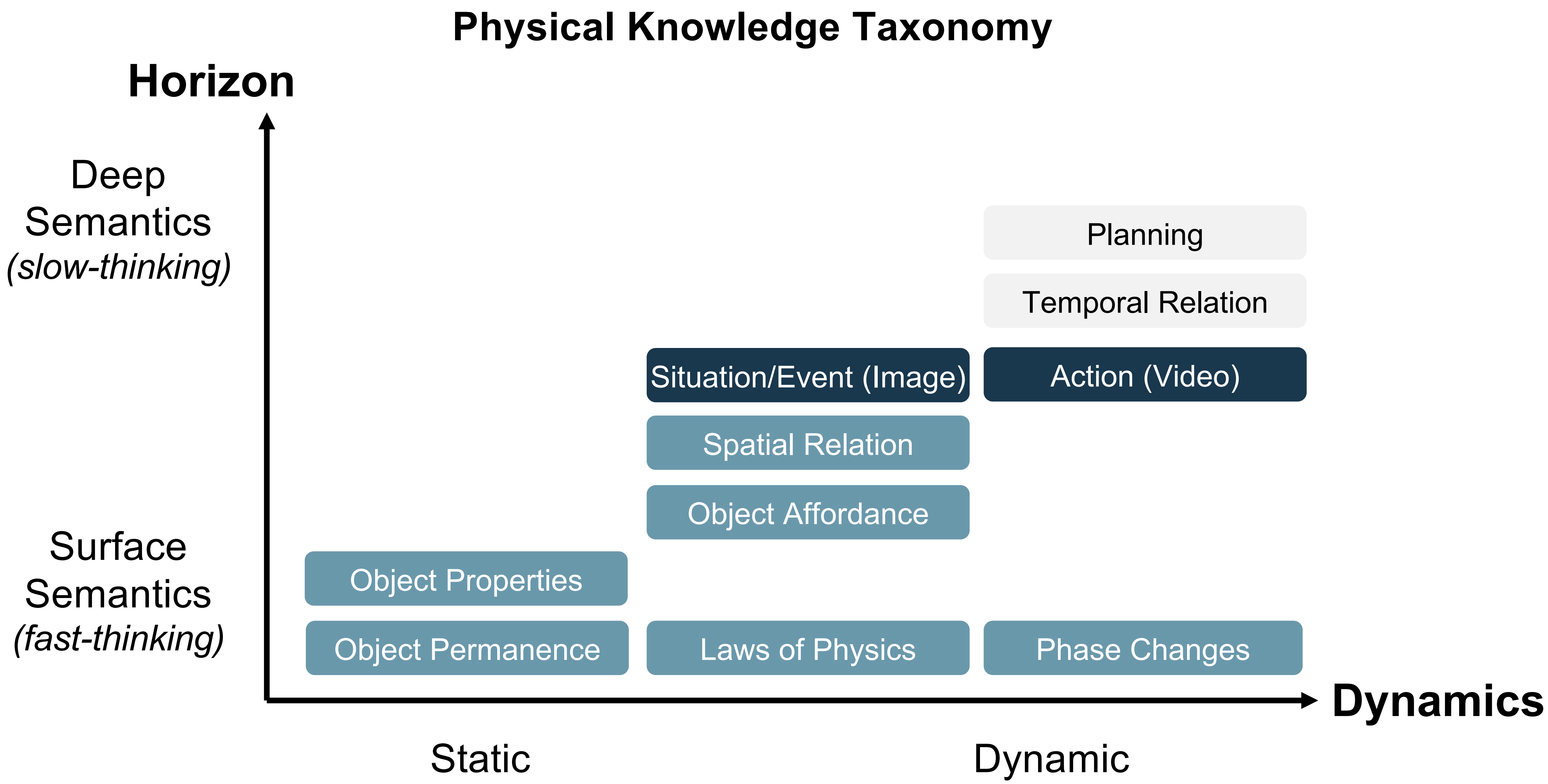
Factual Knowledge

Goal: Surface → Deep Semantic Knowledge

Physical Knowledge Taxonomy



Goal: Surface → Deep Semantic Knowledge



Challenge 1: Complex Situation

Complex Situation

Event / Action

Semantics

Surface

↓

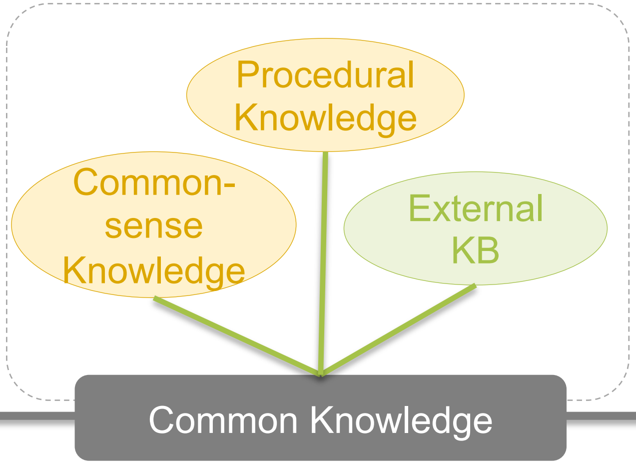
Deep

We learn three types of knowledge

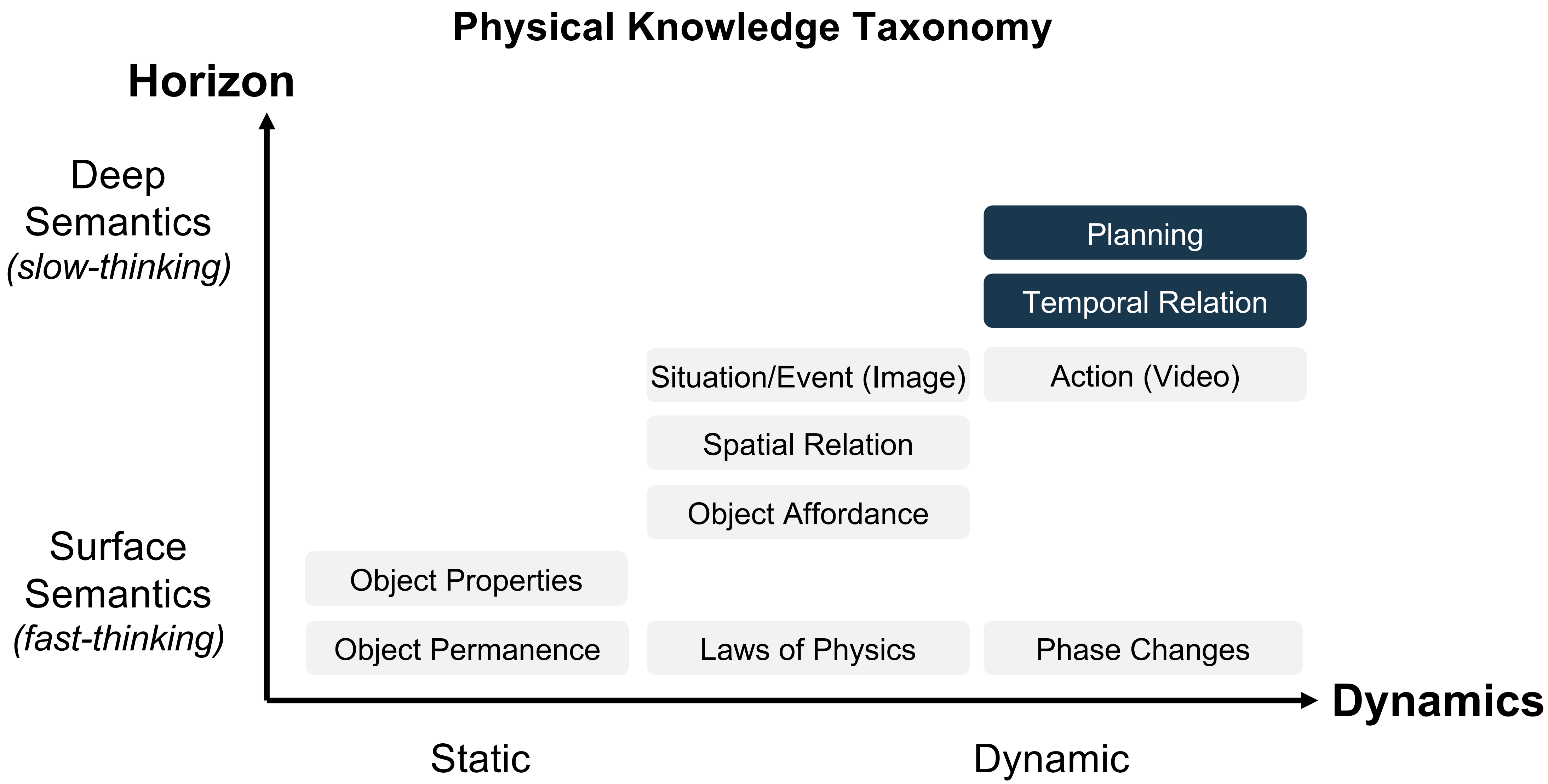
Common knowledge refers to knowledge of **common patterns** that is acquired or summarized from historical interaction with the world.

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

Factual Knowledge



Goal: Surface → Deep Semantic Knowledge



Challenge 2: Long-Horizon Planning

Complex Situation

Long-Horizon Planning

Event / Action Semantics

+

Temporal Dynamics

→

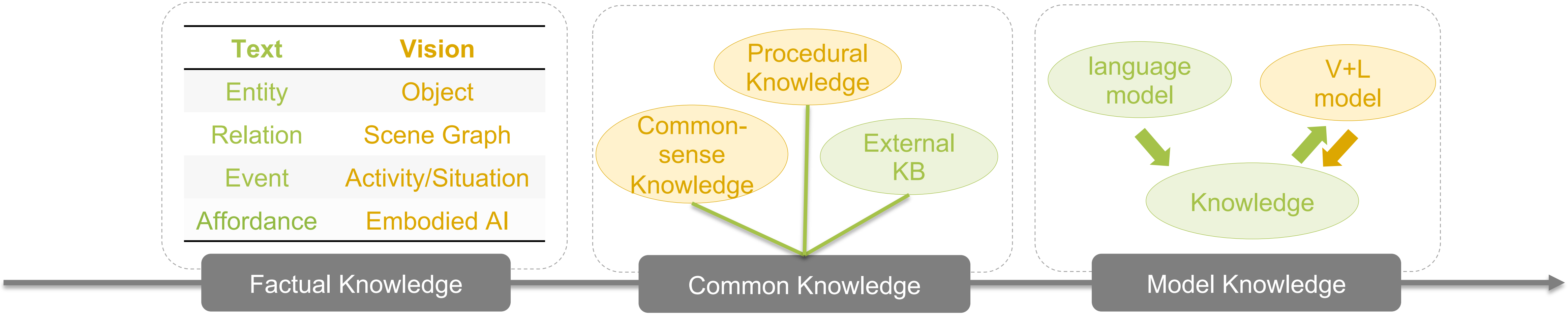
Surface

↓

Deep

We learn three types of knowledge

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



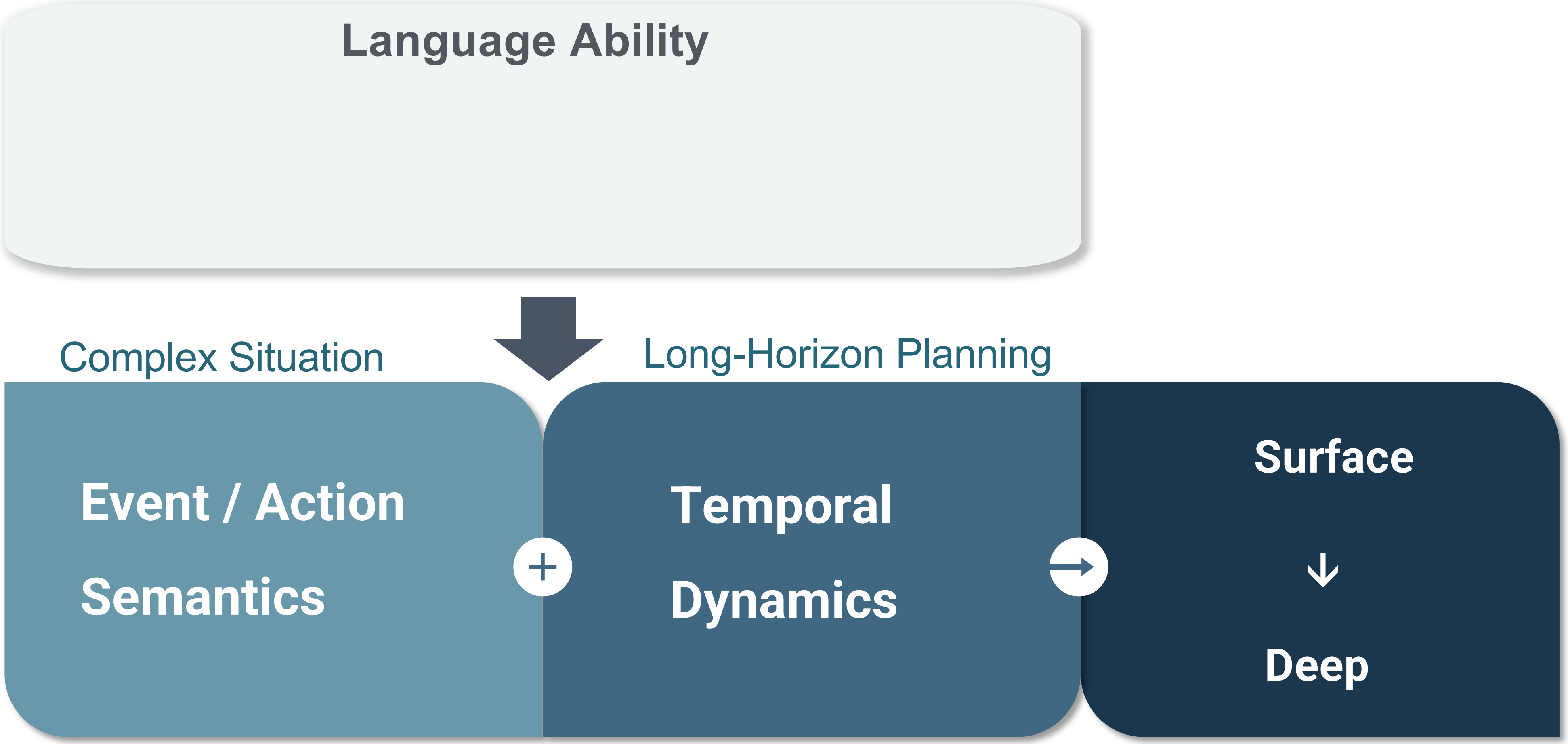
Text	Vision
Entity	Object
Relation	Scene Graph
Event	Activity/Situation
Affordance	Embodied AI

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

Language Ability

Compositional Semantic Structure

Long Context Reasoning

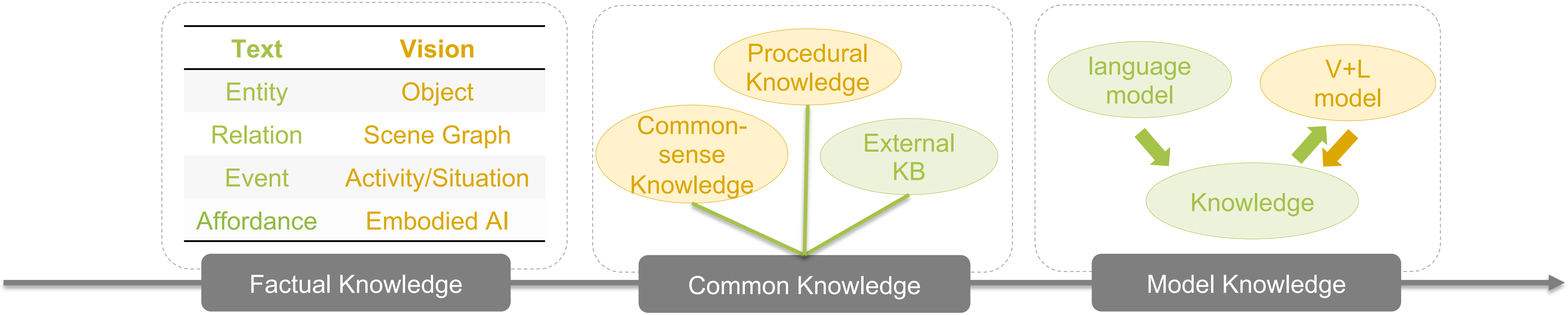
Complex Situation

Long-Horizon Planning

Event / Action Semantics + **Temporal Dynamics** → **Surface ↓ Deep**

We inject knowledge to V+L foundation models

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



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

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

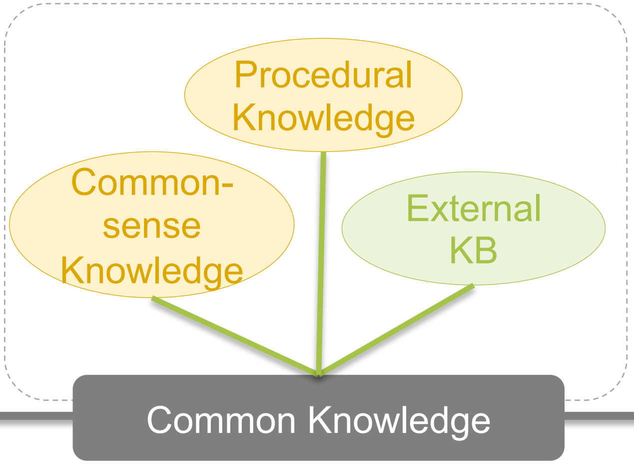
Entity	Hard Object Detection	Soft Object Prompt
SceneGraph & EventGraph	Relational Masking	Situation / Action Learning
Physical Knowledge	Affordance	Multimodal Prompt

Outline: Common Knowledge

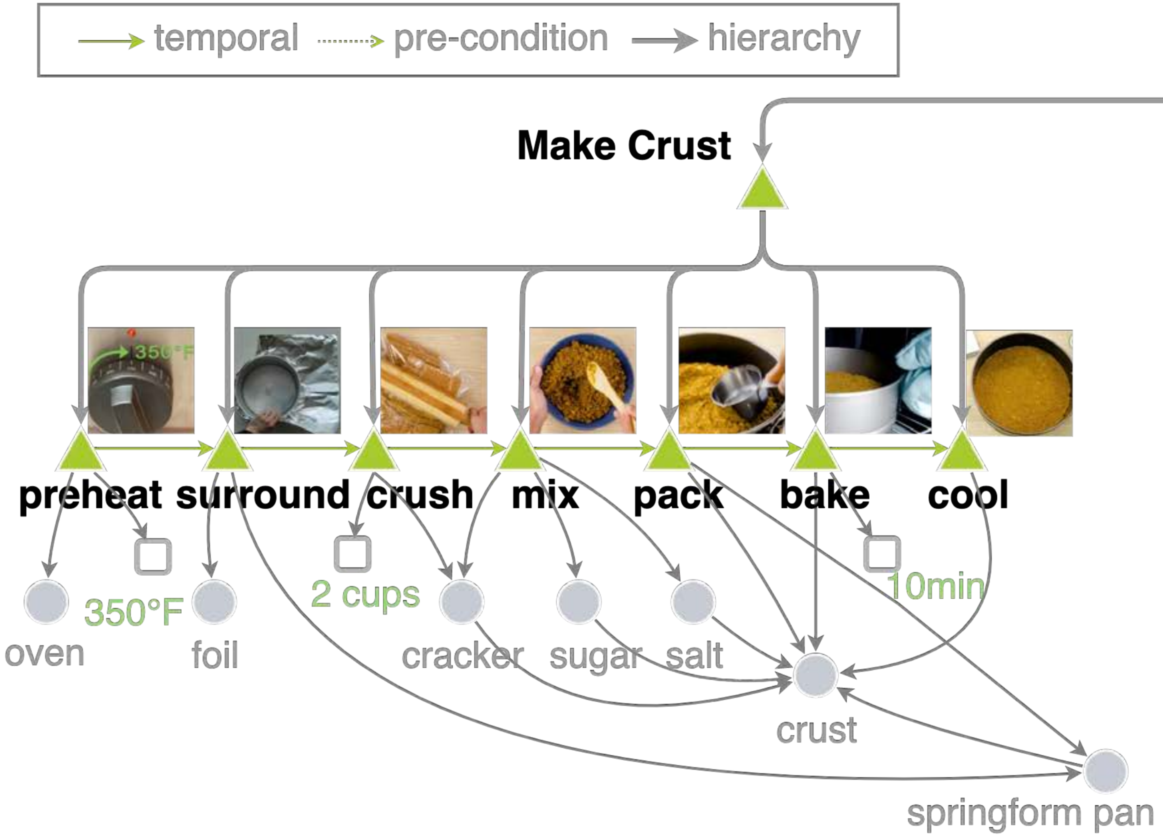
Common knowledge refers to knowledge of common patterns that is acquired or summarized from historical interaction with the world.

Text	Vision
Entity	Object
Relation	Scene Graph
Event	Activity/Situation

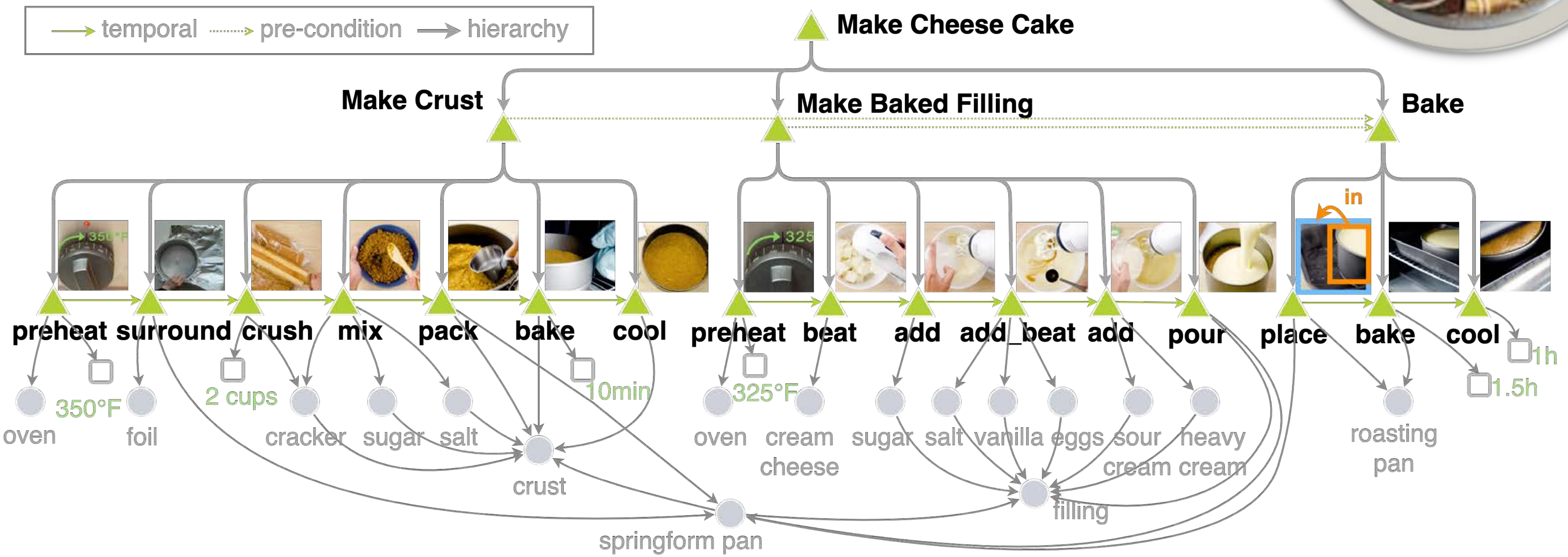
Factual Knowledge



Common Knowledge: History repeats itself

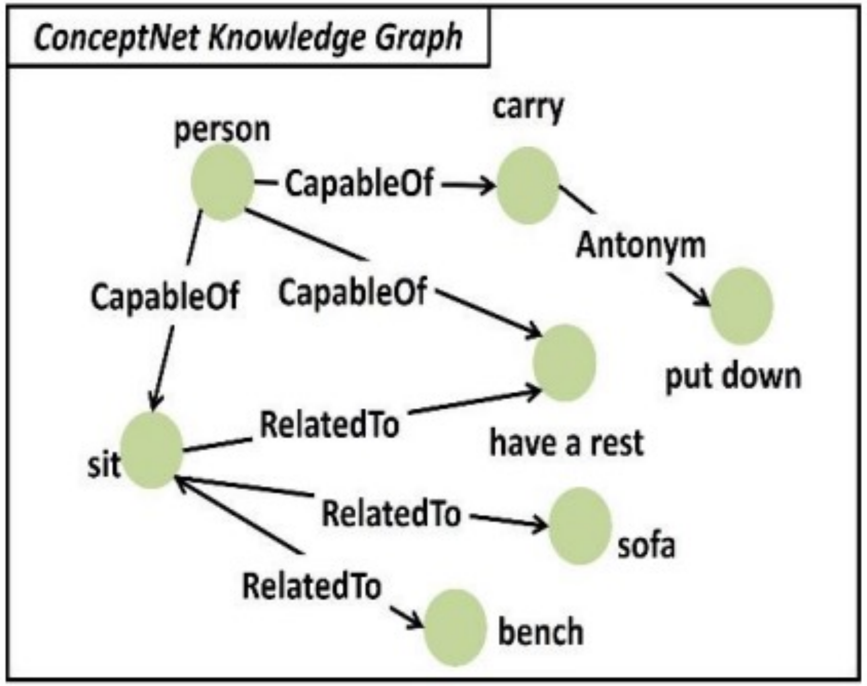


Common Knowledge: History repeats itself



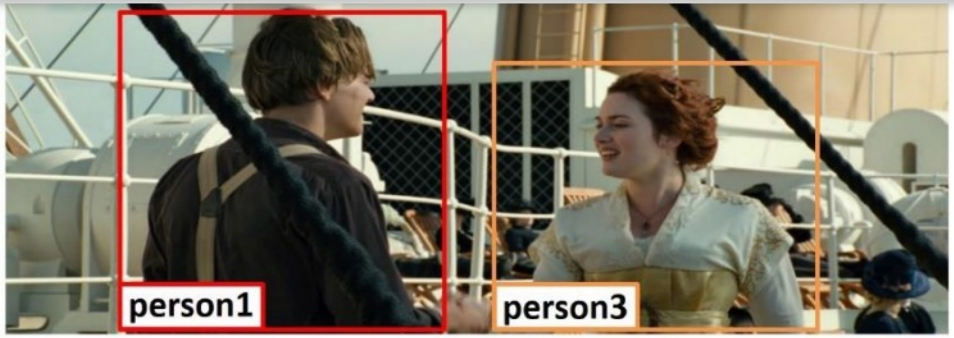
Commonsense Knowledge

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.

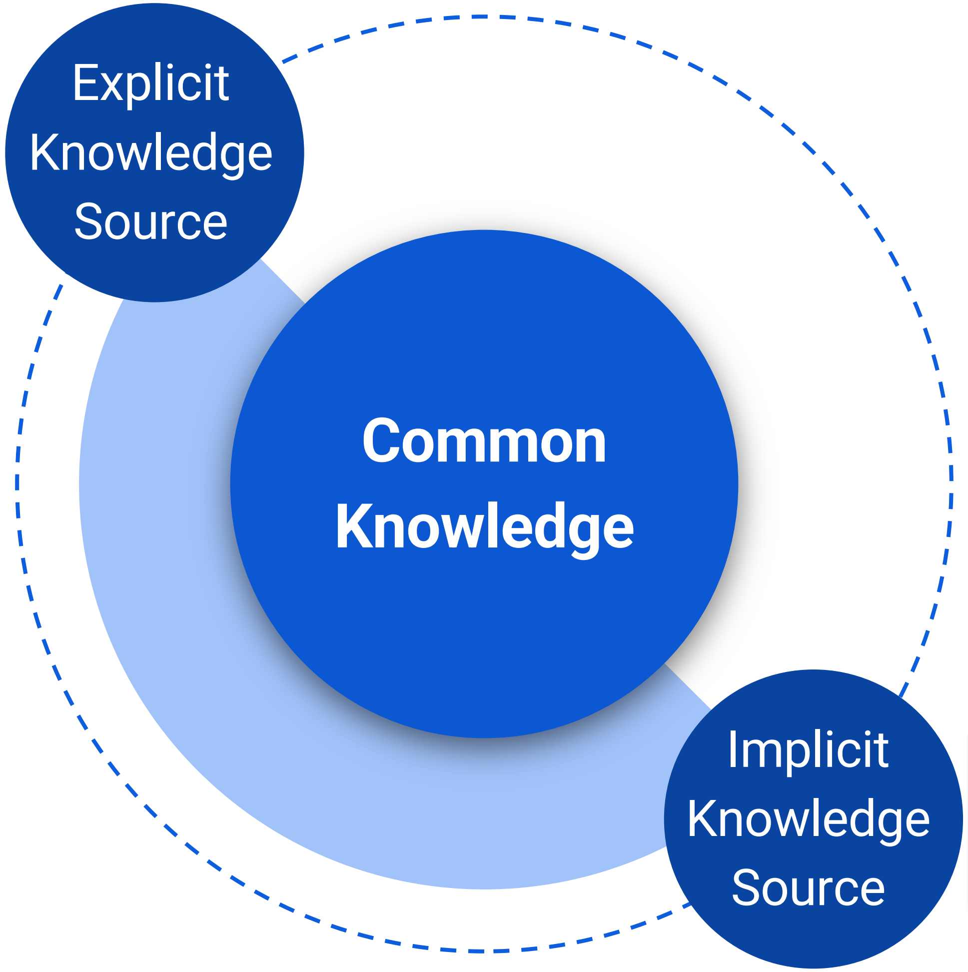
I think so because ...

(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.

How to add common knowledge?

- Two ways to learn procedural knowledge

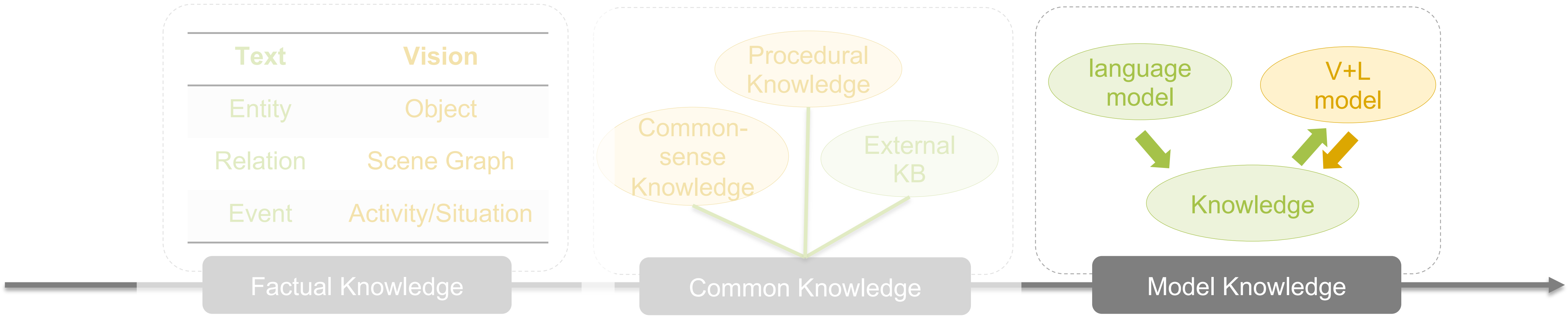
Use knowledge:
- As Data
- As Supervision
- In Model



Learning from massive data

Model Knowledge

Model Knowledge is the knowledge embedded and encoded in models.



Model Knowledge

Implicit Knowledge (L → V)	Leveraging Frozen LM	Convert multimodal info into text
Explicit Knowledge (L → V)	External KB	Large Language Models
Knowledge Distillation (V → L)	Physical Knowledge	Commonsense Knowledge

Summary: How to learn multimedia embedding?

Knowledge for Model

+

Knowledge for Data

→

Knowledge-Driven V+L Pretraining

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 in V+L Pretraining: Information about Instances

Knowledge-Driven Vision-Language Pretraining (Part II)

Manling Li
UIUC
manling2@illinois.edu

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN
Northwestern University
COLUMBIA UNIVERSITY
Meta AI

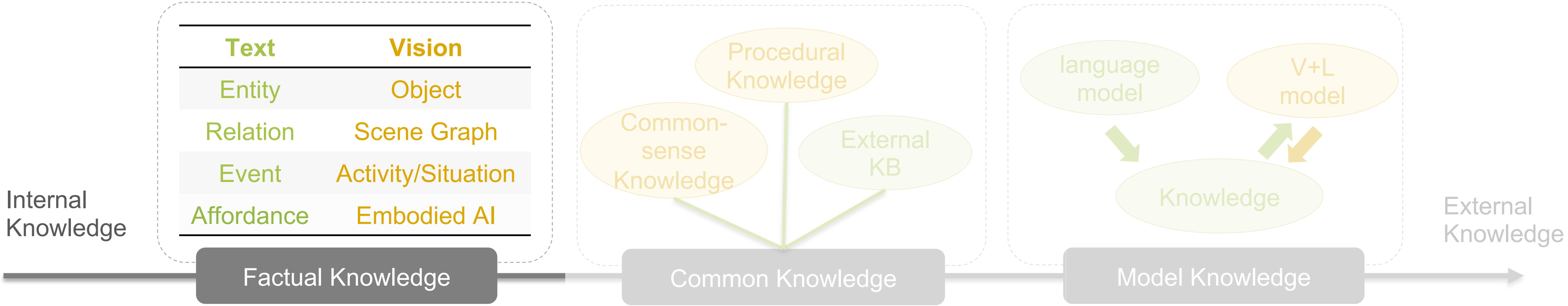
Timetable

Content	Time	Presenter
Motivation and Overview	15min	Manling Li
Factual Knowledge	30min	Manling Li
Commonsense Knowledge	15min	Manling Li
Procedural Knowledge	30min	Xudong Lin
Model Knowledge	30min	Jie Lei
Panel: Knowledge vs Large Models	15min	Mohit Bansal, Carl Vondrick, Xudong Lin
Panel: LLMs for multimodal	15min	Mohit Bansal, Carl Vondrick, Jie Lei
Panel: Image vs Video vs Audio vs Others	15min	Mohit Bansal, Carl Vondrick, Xudong Lin
Panel: Open Challenges	15min	Mohit Bansal, Carl Vondrick, Jie Lei
QA	30min	All

Factual Knowledge

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

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



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?

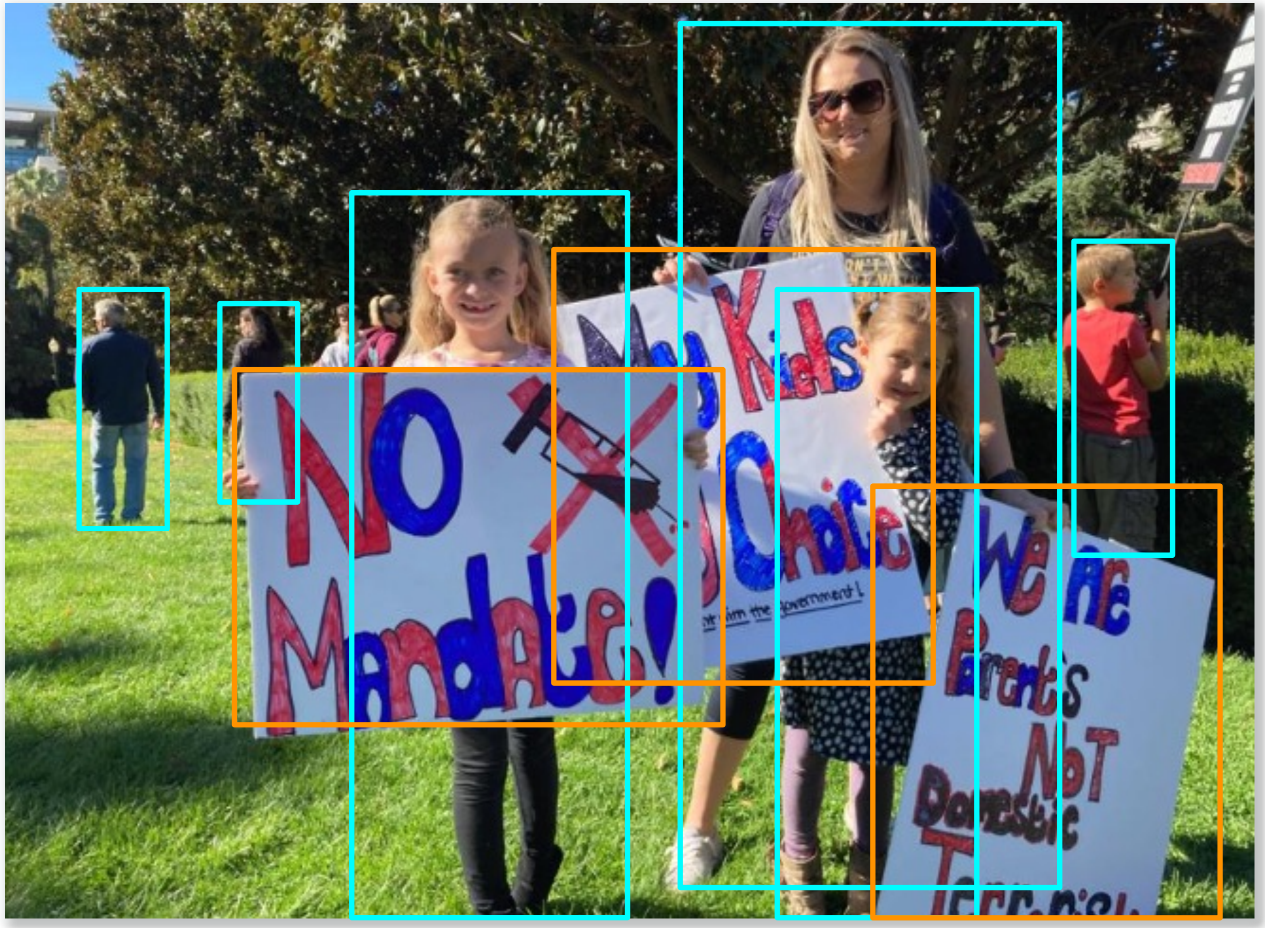
vaccine

negation

Existing object-centric info miss situational understanding

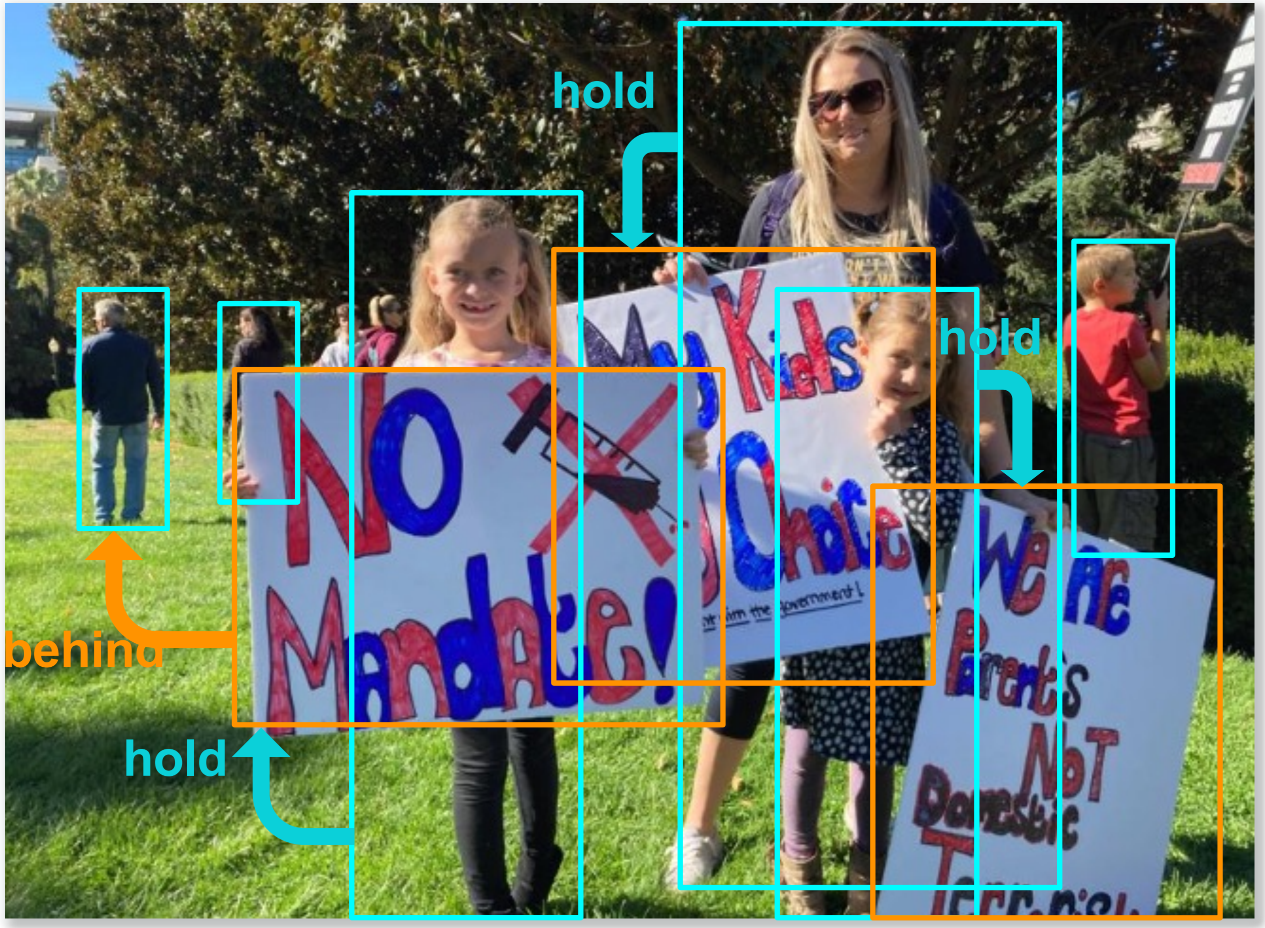
Vision

Existing object-centric info miss situational understanding



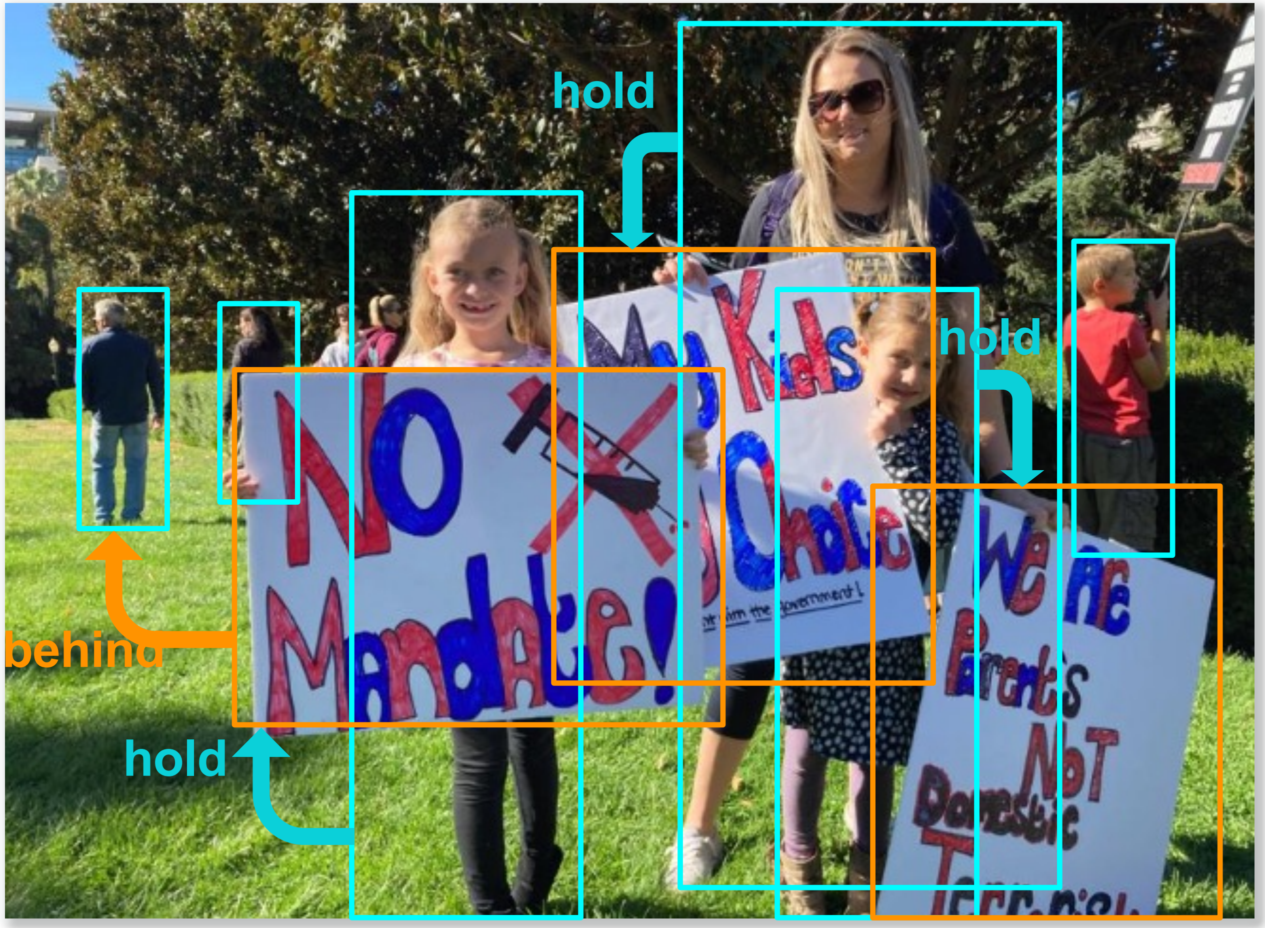
Vision
Object

Existing object-centric info miss situational understanding



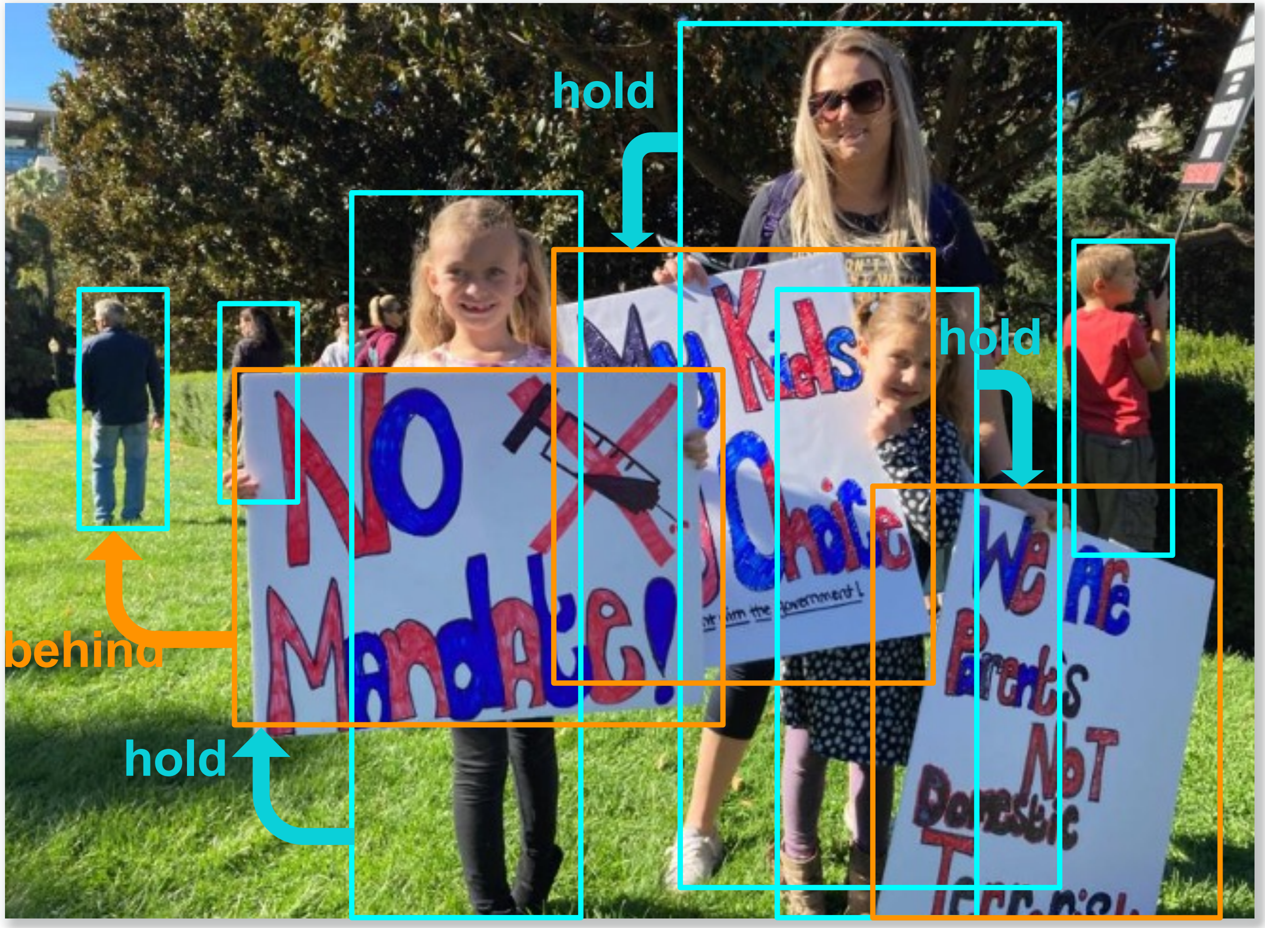
Vision
Object
Relation
Scene Graph

Existing object-centric info miss situational understanding



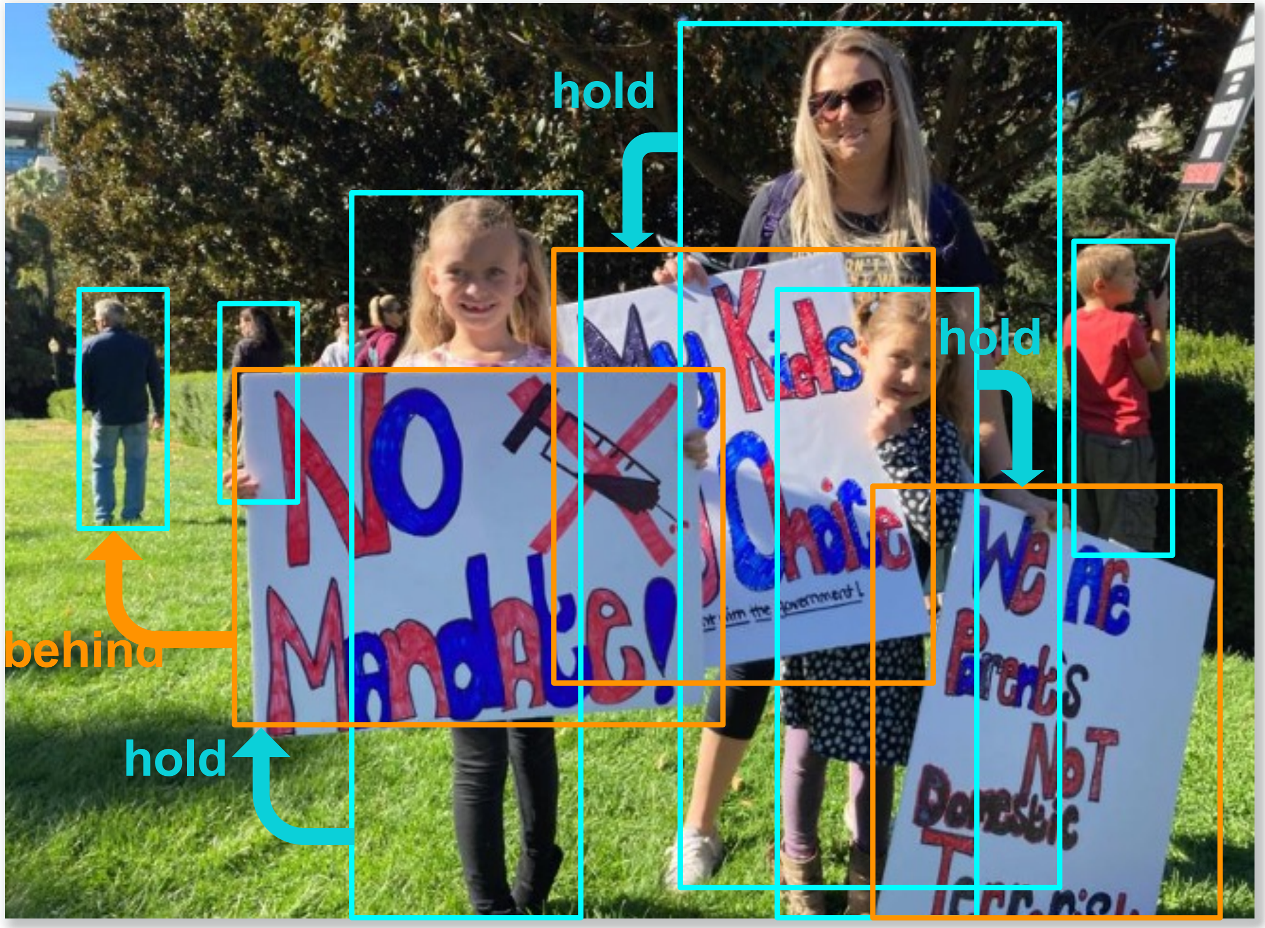
Vision	Text
Object	
Relation	
Scene Graph	

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
Object	Entity
Relation	Relation
Scene Graph	Entity-Relation Graph

Entity-centric

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”

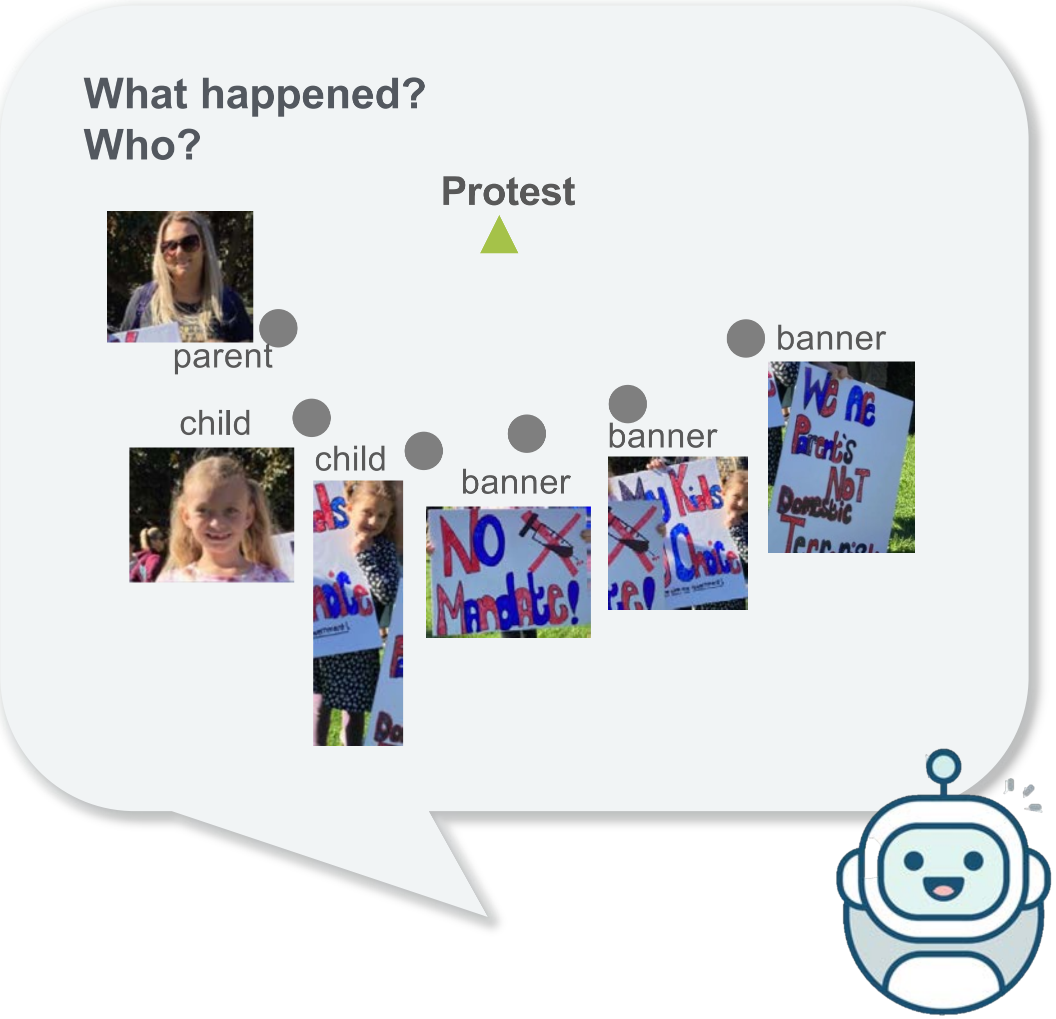
Event	Protest

What happened?

Protest

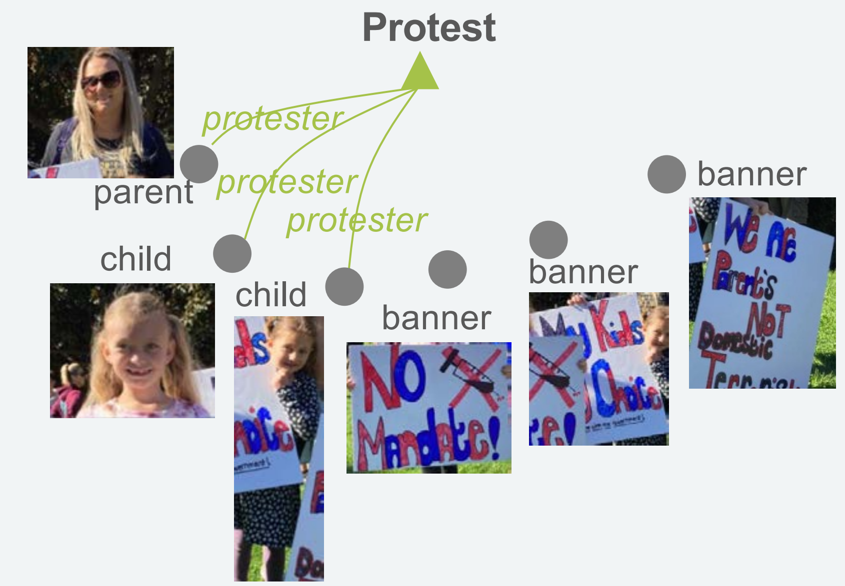
Definition of “Event”

Event	Protest
	parent
	child
	banner
	No vaccine mandate for kids



Definition of “Event”

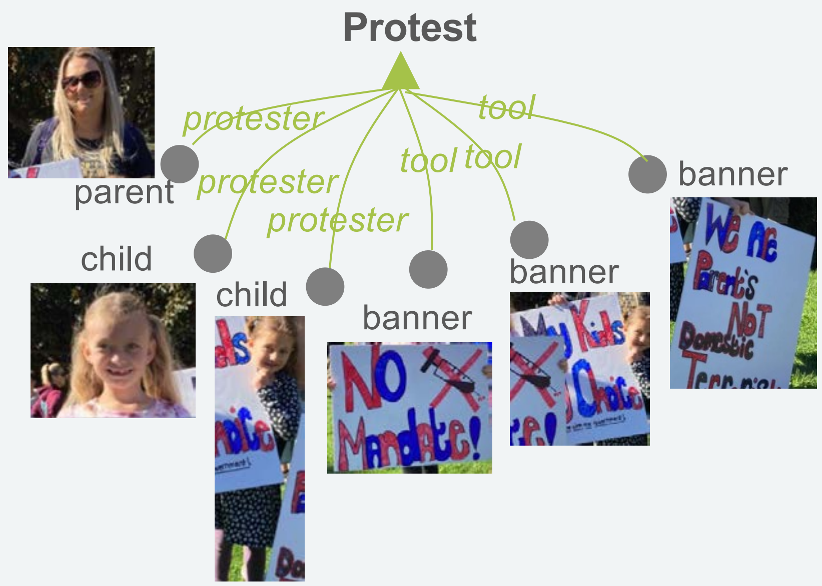
Event	Protest
Protester	parent
Protester	child
	banner
	No vaccine mandate for kids



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]

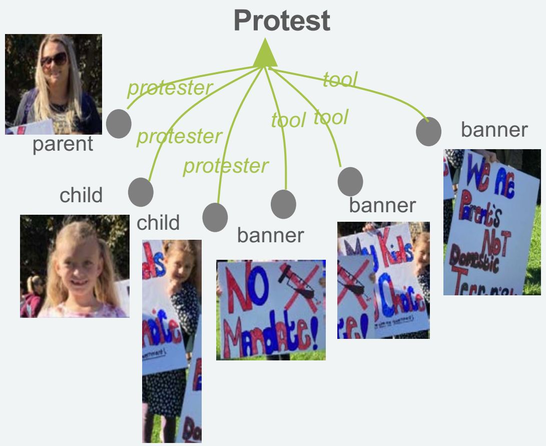
What is Multimodal Event Extraction? [Li et al, ACL'20]

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]

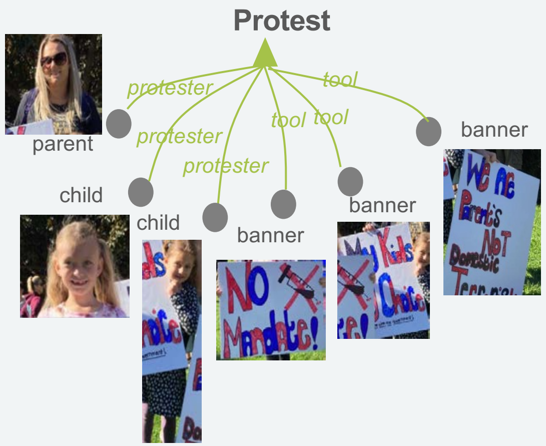
Event Extraction

Input: text, image, video, speech, …

Output: structured knowledge

1. **Event Type** ***(e.g., protest)***

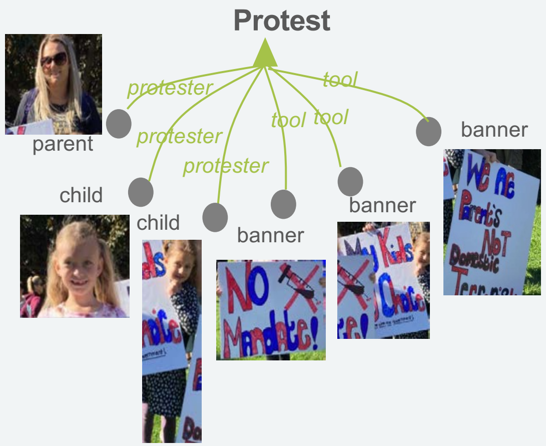
2. **Participants** ***(e.g., child)*** **& Semantic Roles** ***(e.g., protester)***



[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)***

Entity-centric	Two Objects	A Triplet	Generic	Surface
Event-centric	Multiple Objects	Semantic Structures	Situational	Deep

[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

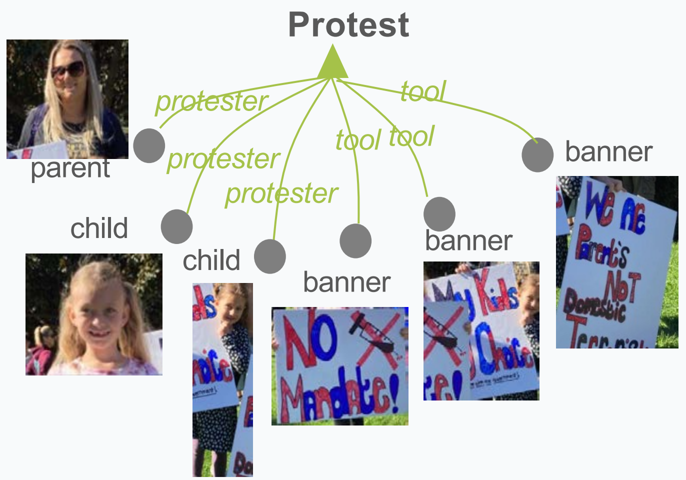
State-of-the-art Captioner (Kamath et al., 2022)

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

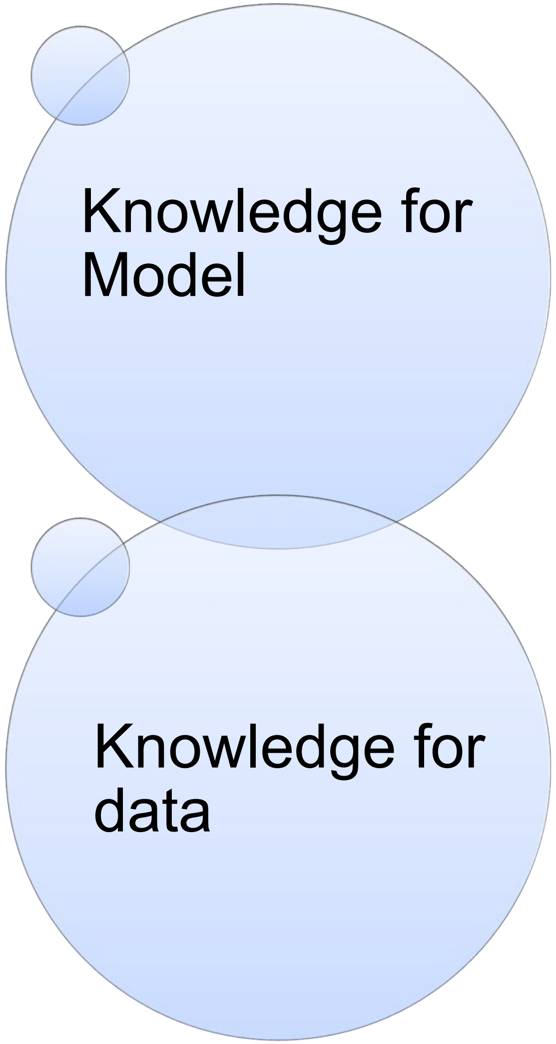
Entity-centric

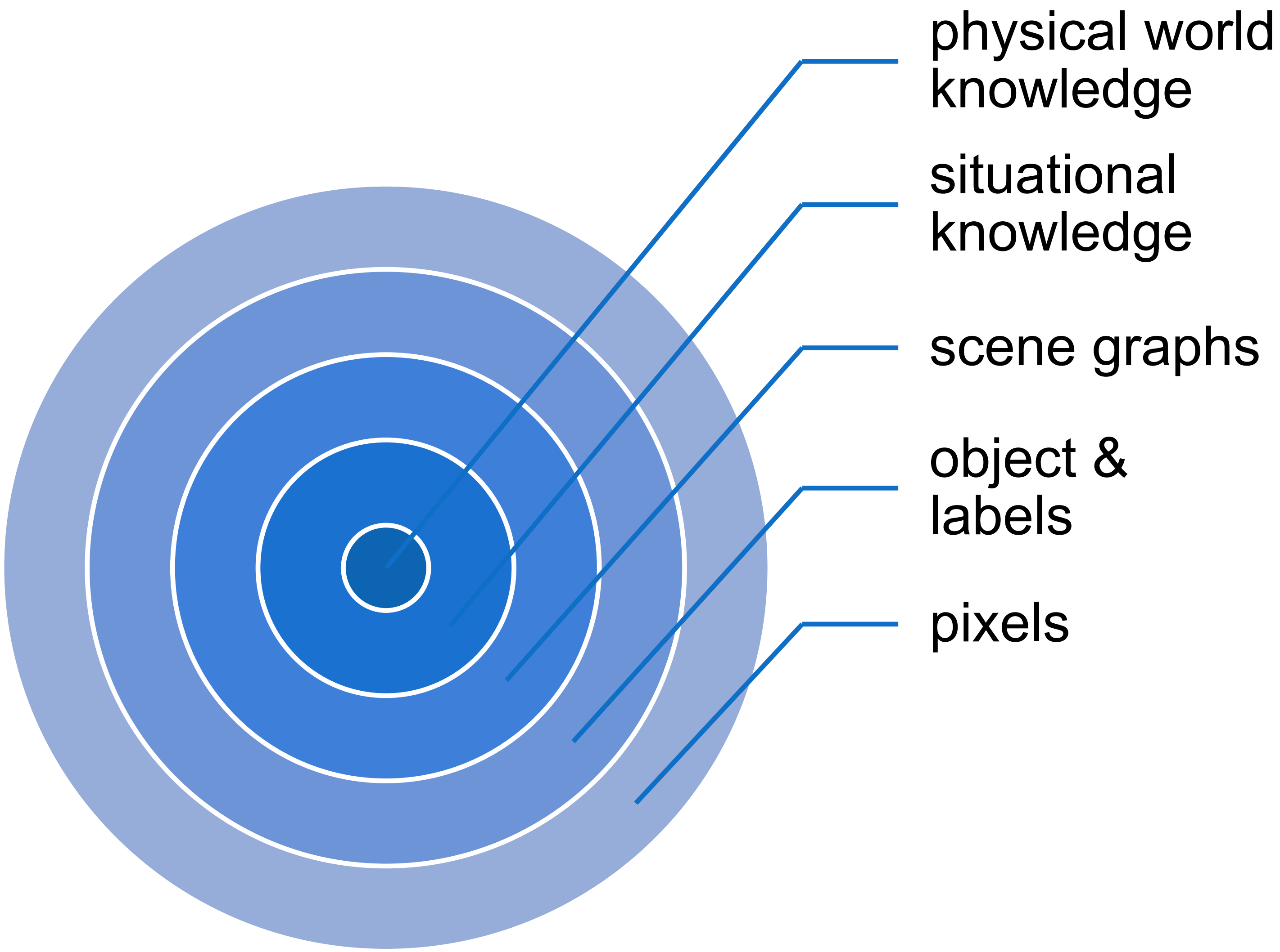
↓

Event-centric



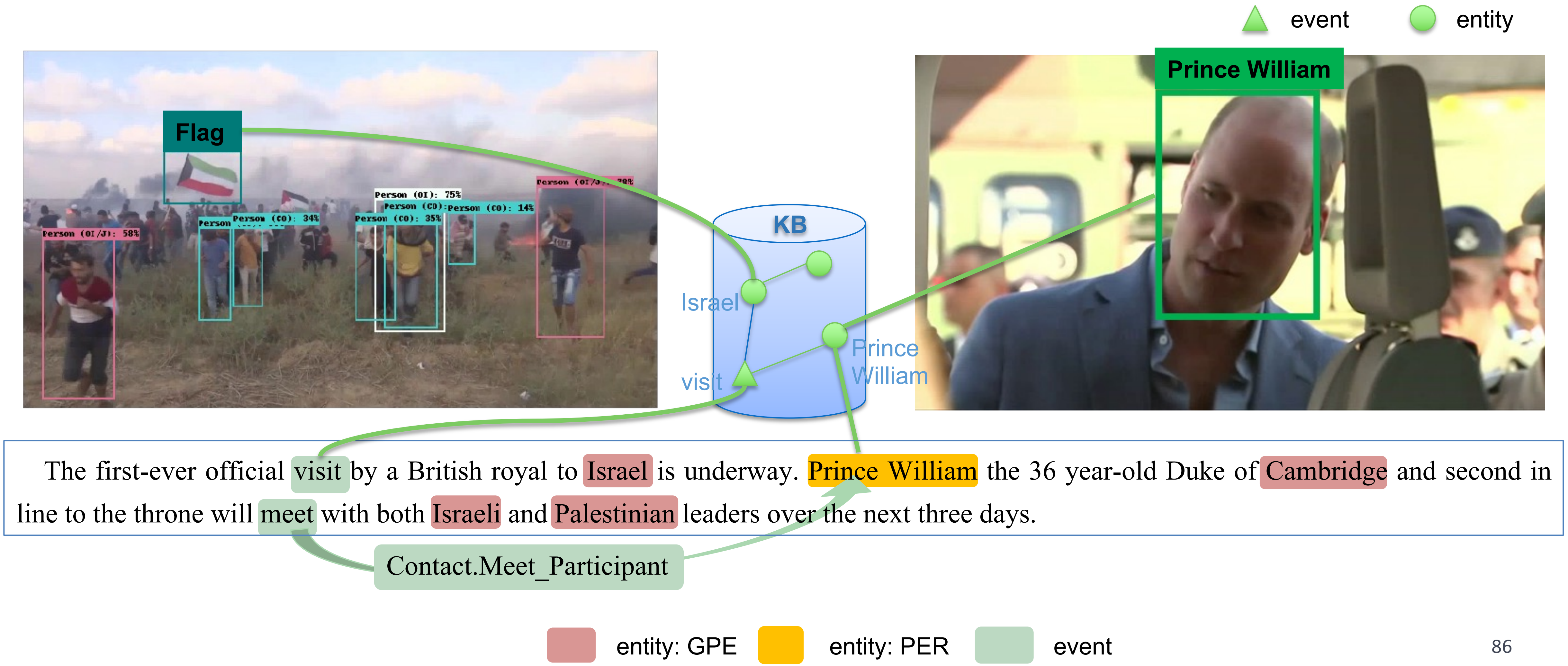
Adding knowledge to pretraining models



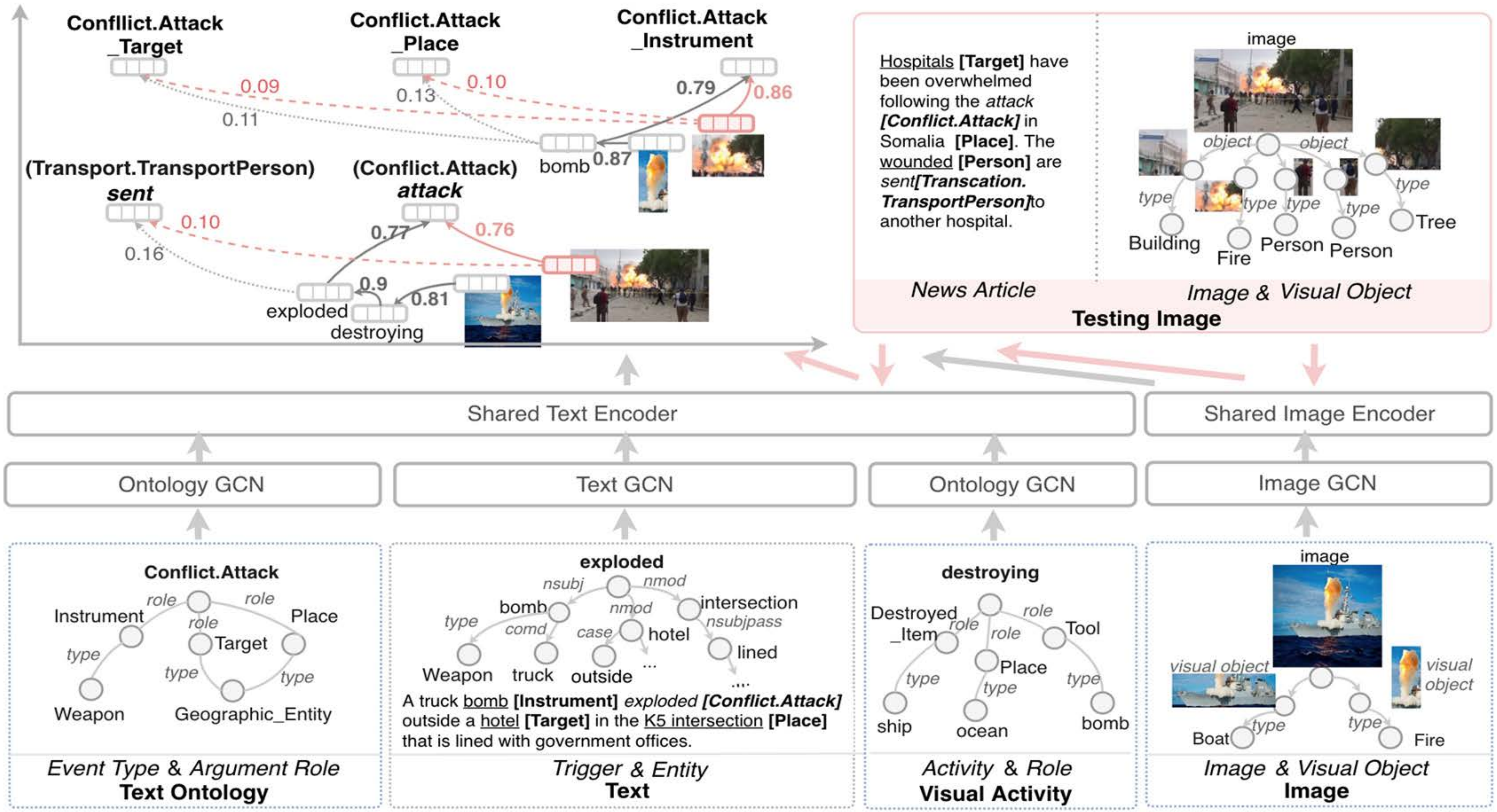


What is factual knowledge?

- Multimedia Knowledge Base with entities, relations and events.



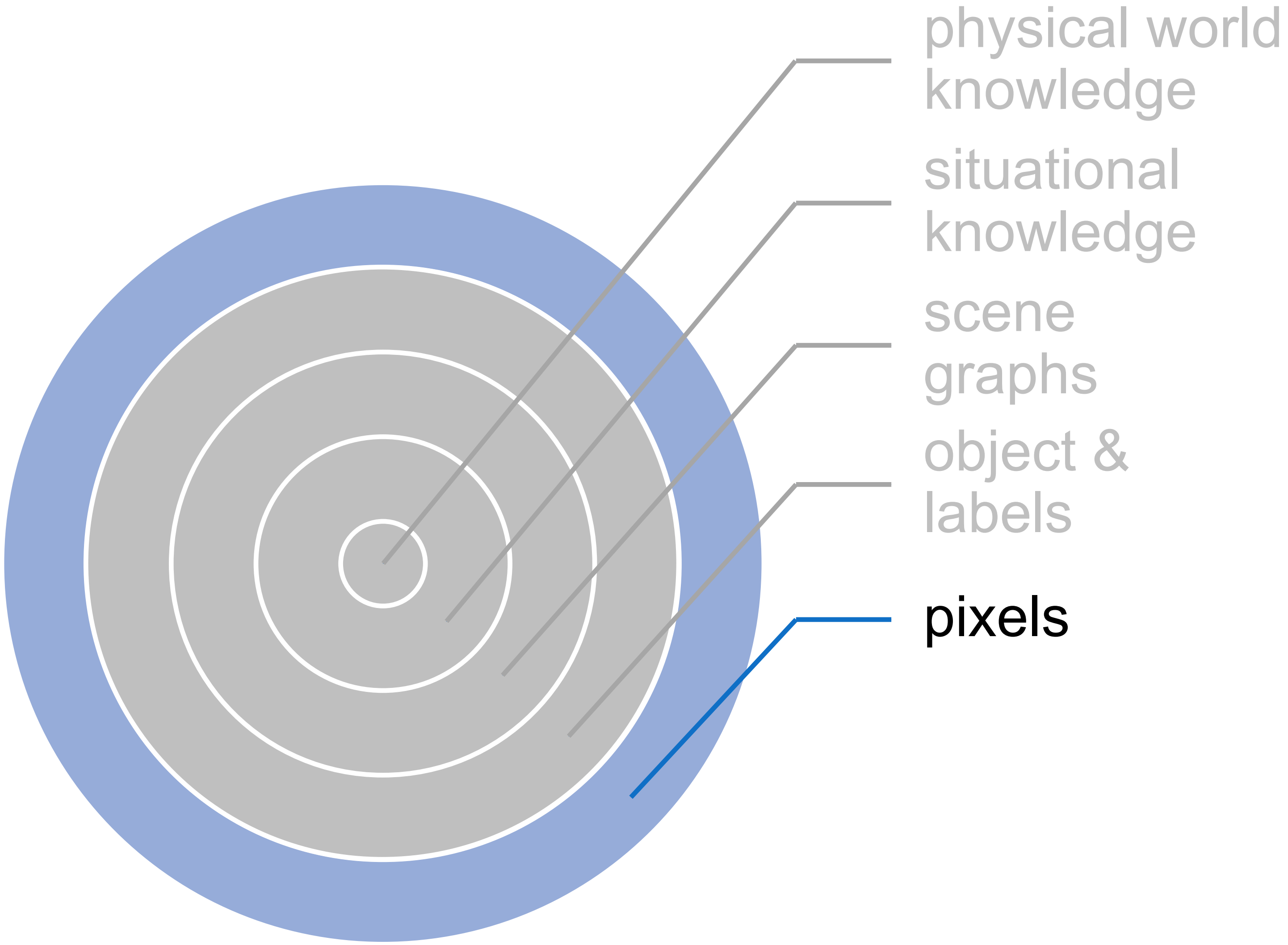
Goal: A joint representation of text and vision knowledge



Adding knowledge to pretraining models

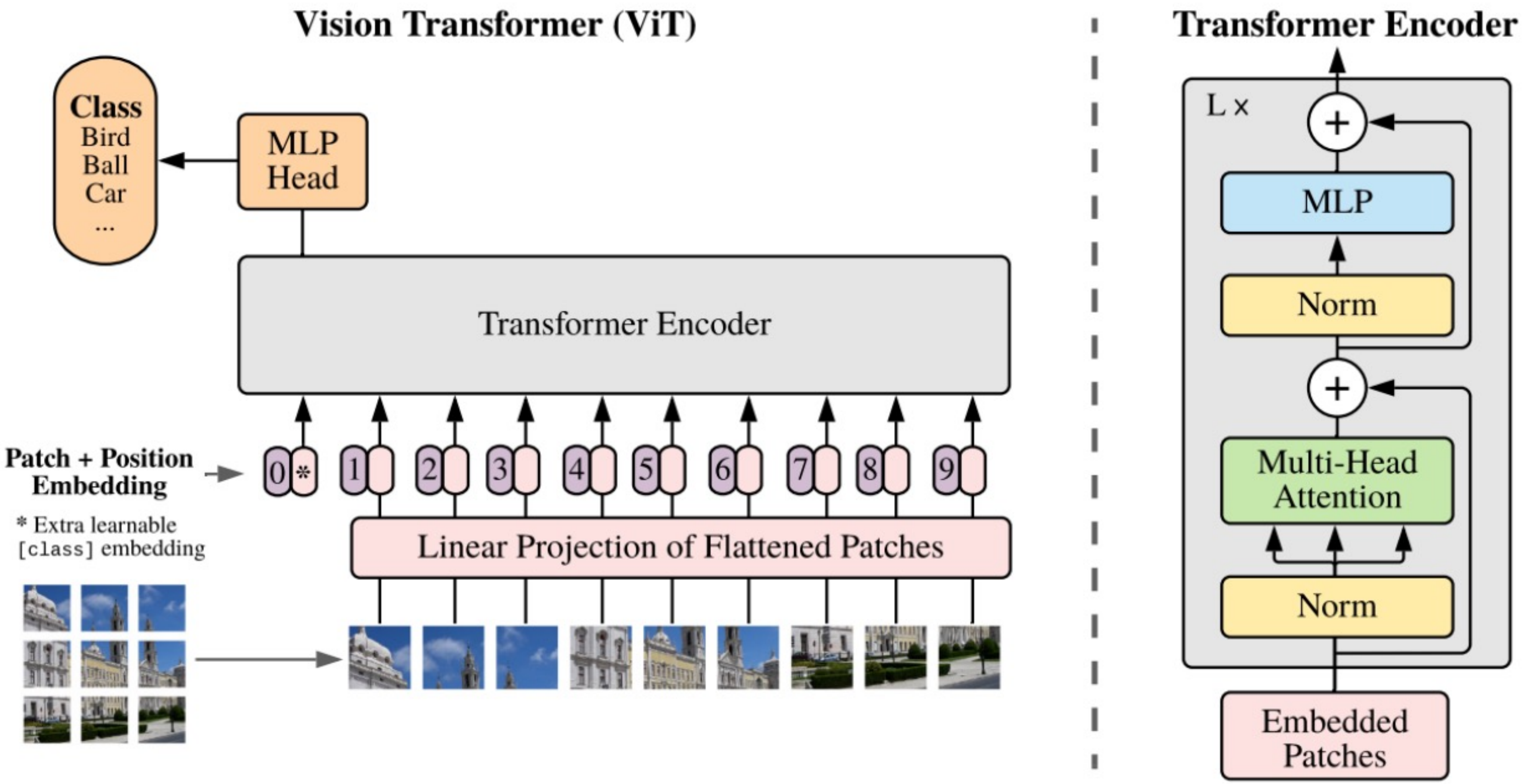
Knowledge for Model

Knowledge for data



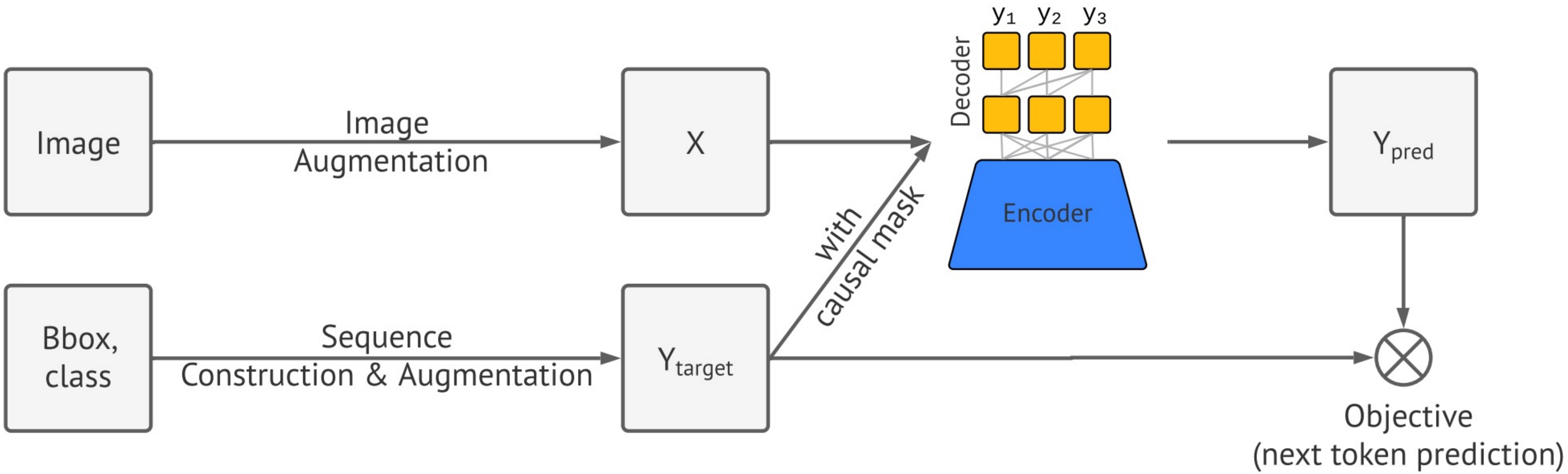
Pixels - An Image is Worth 16x16 Words

The simplest way is to split an image into patches

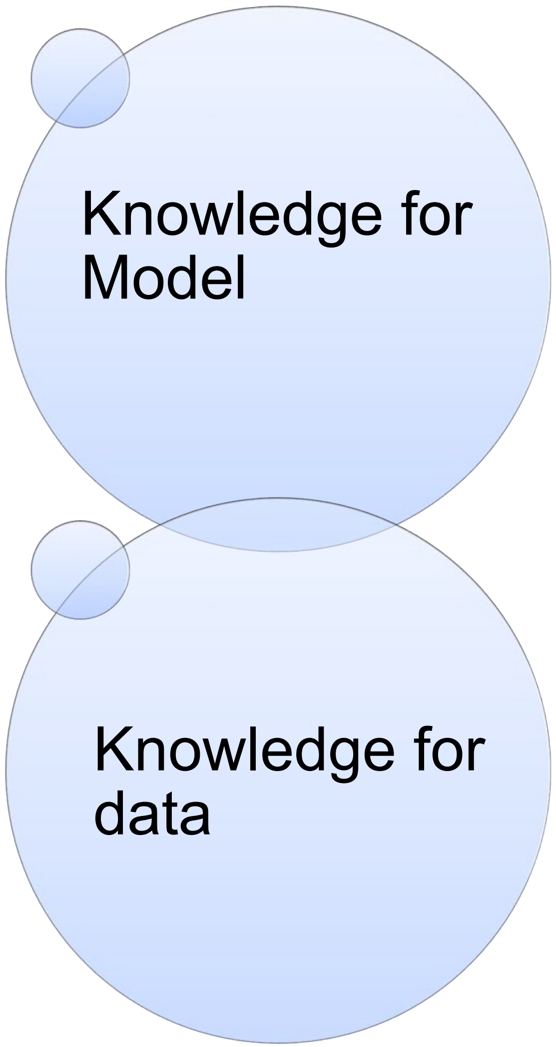


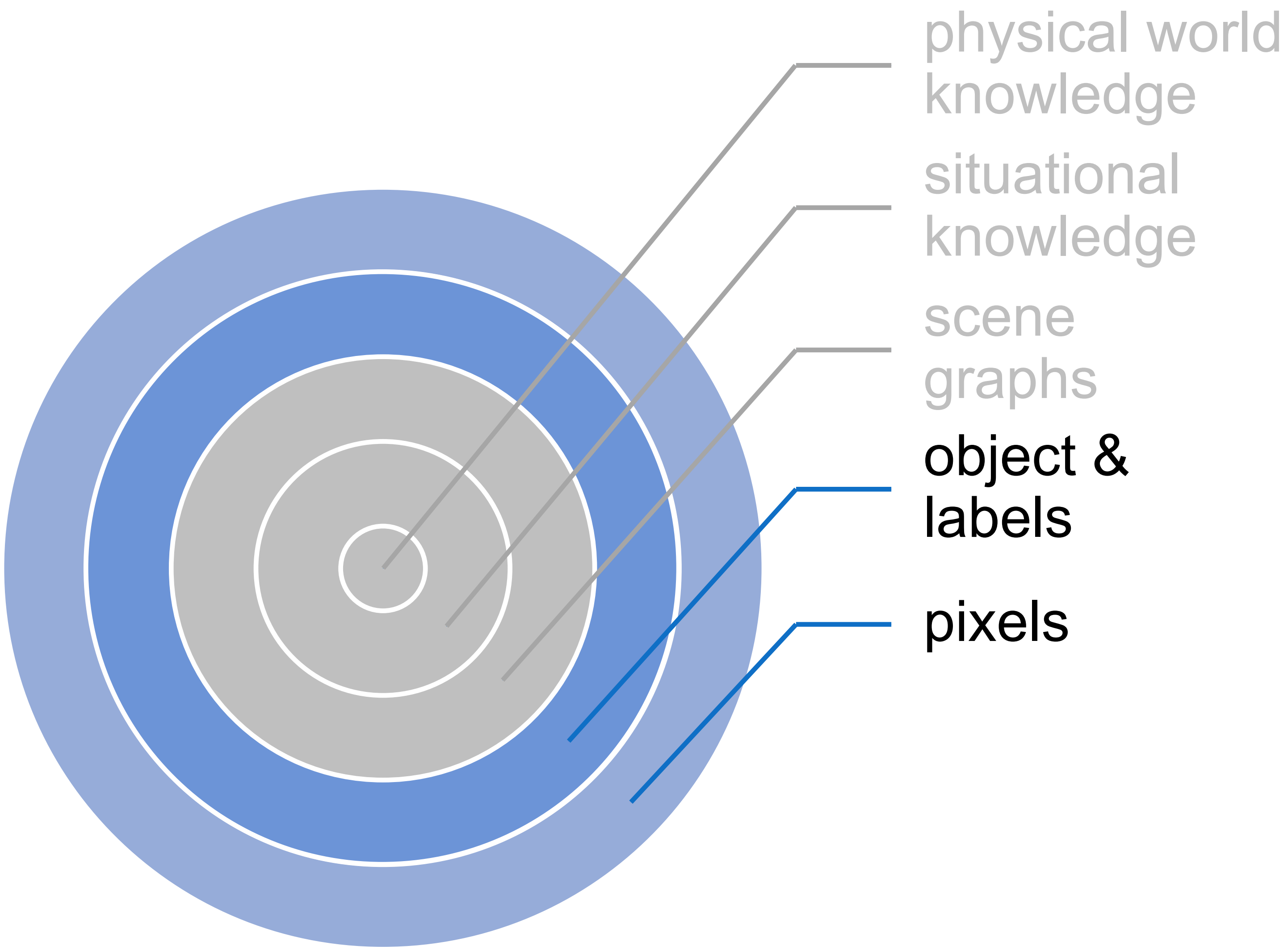
Pixels - Unified Model: Pix2Seq

Another way is to treat pixels as tokens.



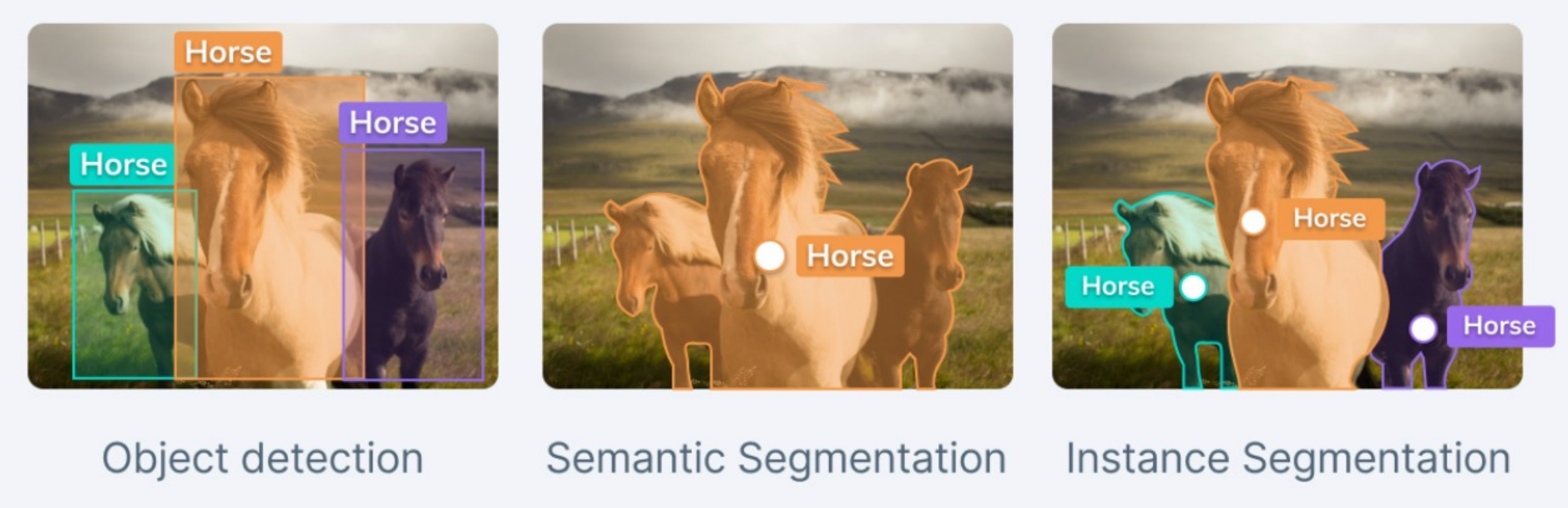
Adding knowledge to pretraining models





Entity Knowledge

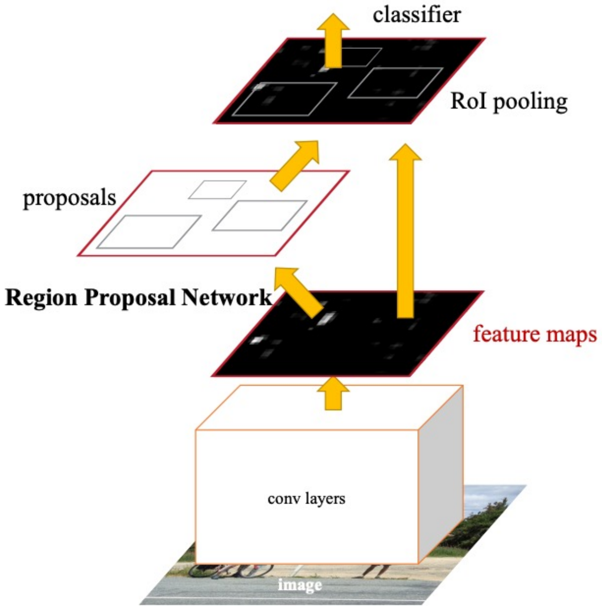
- Object Detection: Object instances at the bounding box level
- Semantic Segmentation: Object class at the pixel level
- Instance Segmentation: Object instances at the pixel level



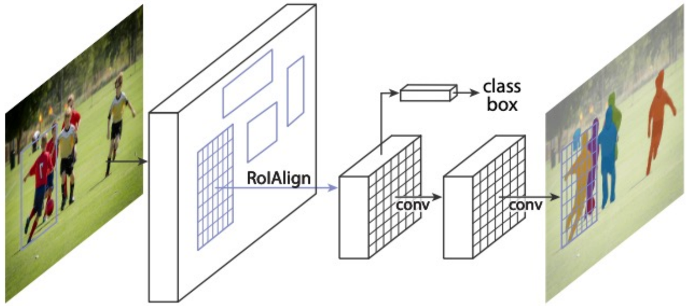
https://www.v7labs.com/blog/object-detection-guide

The way to obtain entity knowledge: Object Extraction

- Two-Stage (With Proposal)

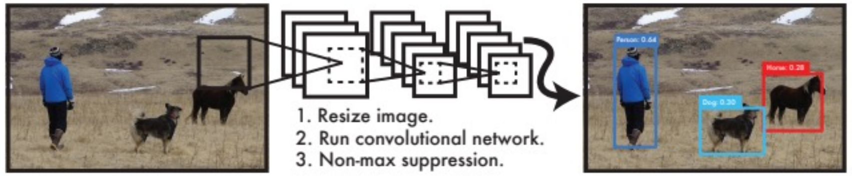


Faster RCNN

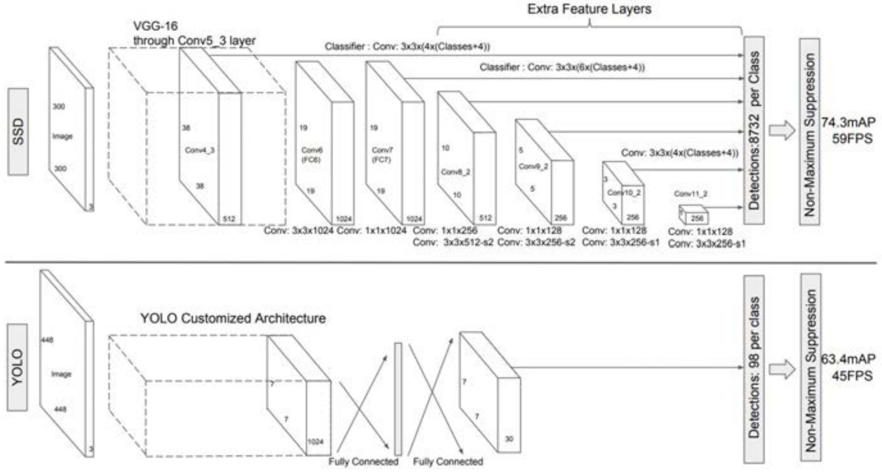


Mask RCNN

- One-Stage (Without Proposal)



YOLO

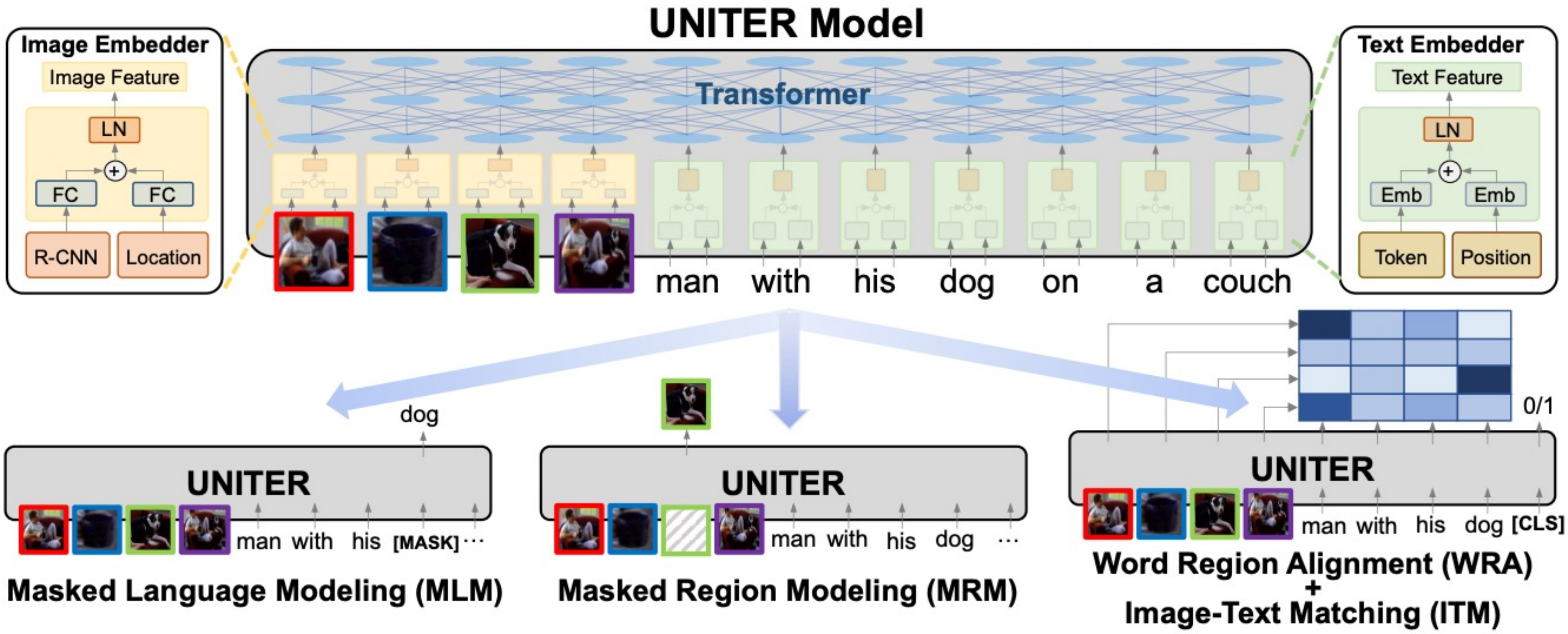


SSD

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.

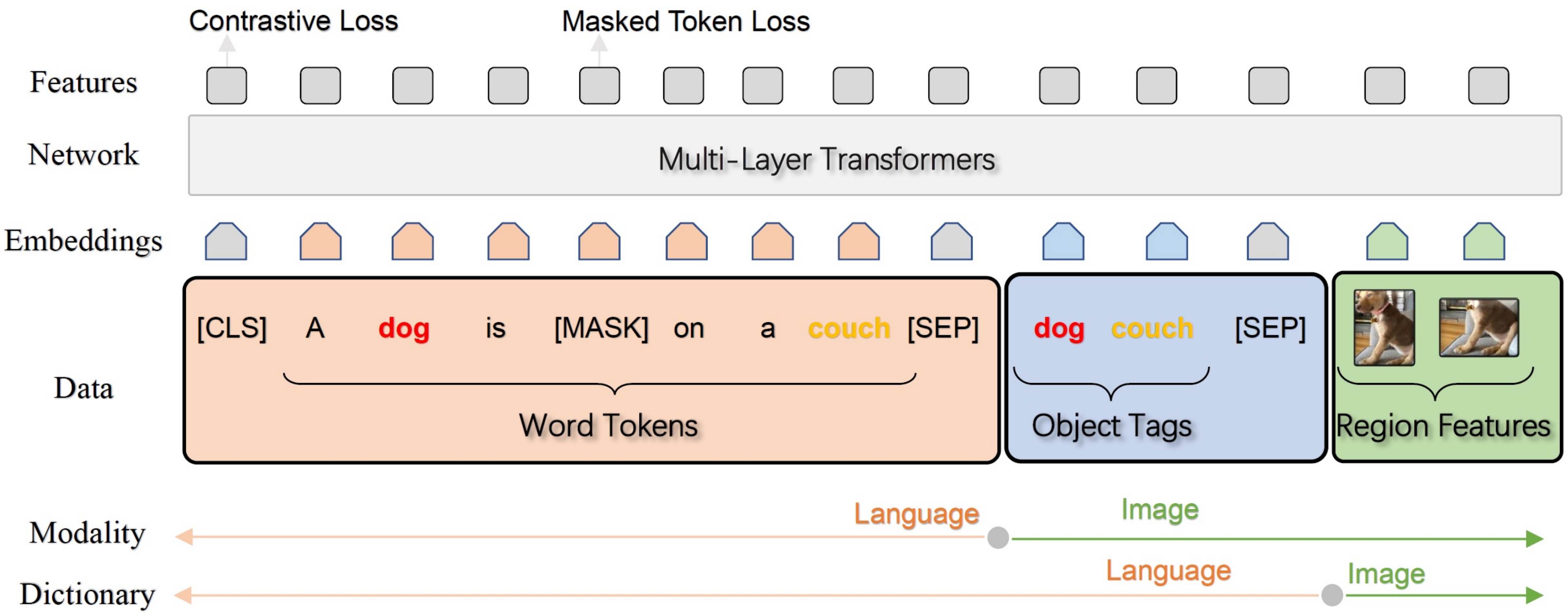
Adding objects to V+L Pretraining

Objects are used to better mask the regions.



Oscar [ECCV 2020] and VinVL [CVPR 2021]

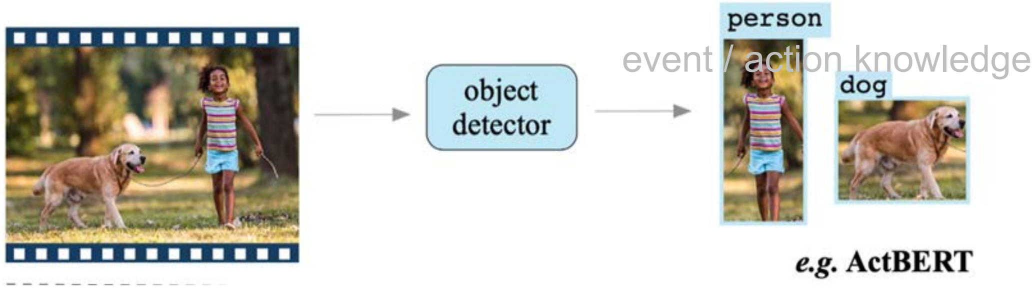
- Object knowledge is richer.
  - Add object label knowledge as anchor points



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



previous work rely on object detectors with expensive computation and limited object categories

Individually pre-trained with Video-Text Contrastive loss

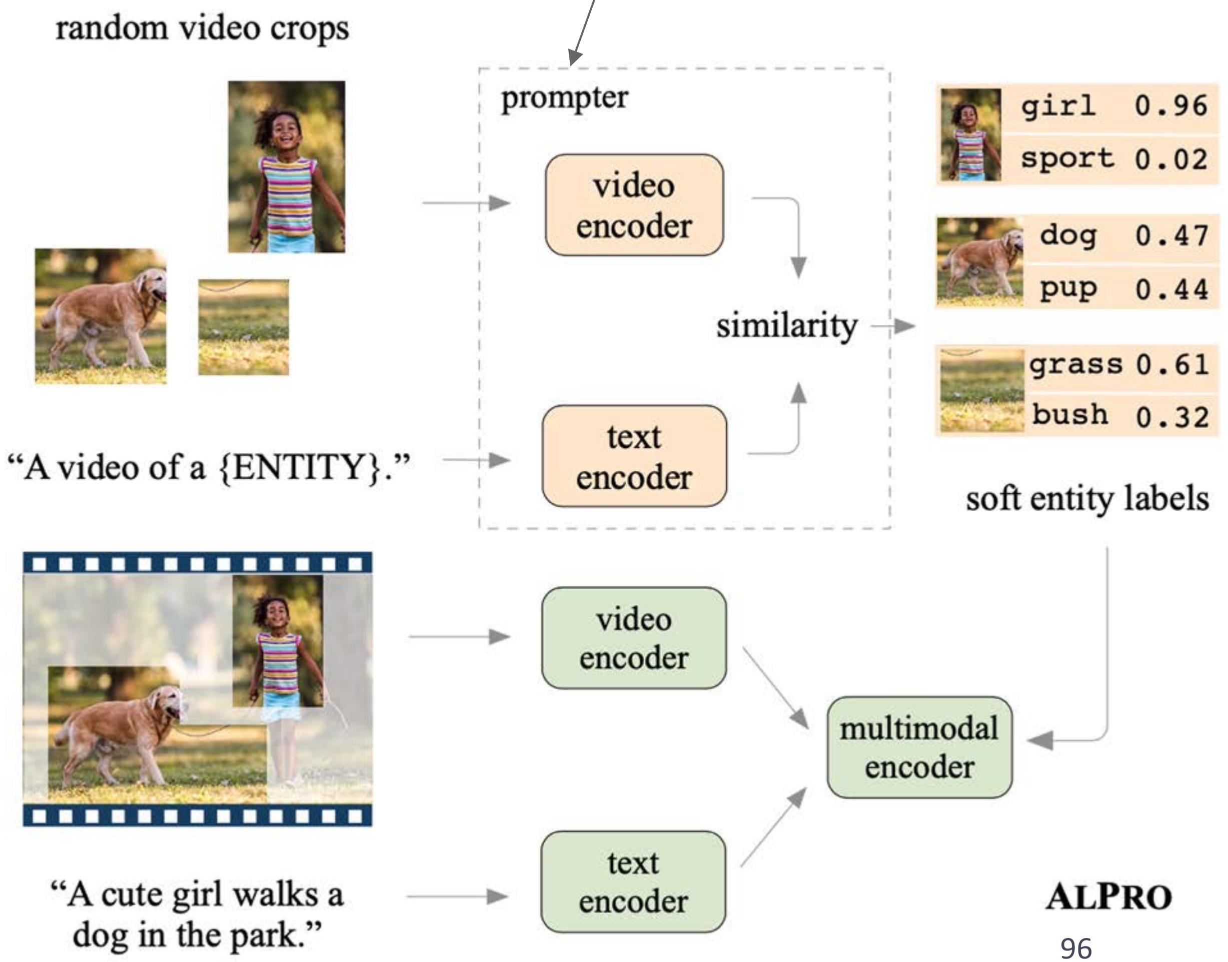
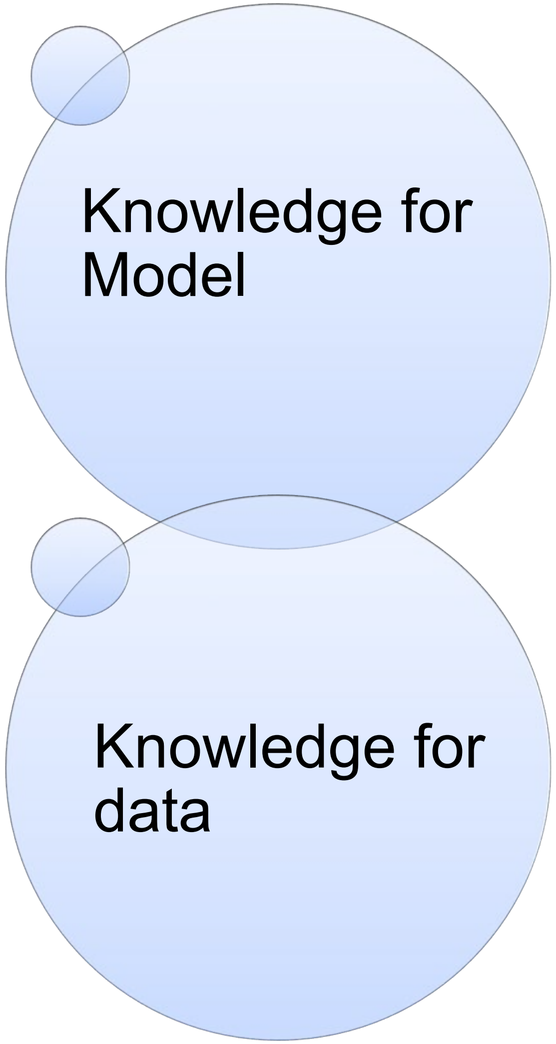
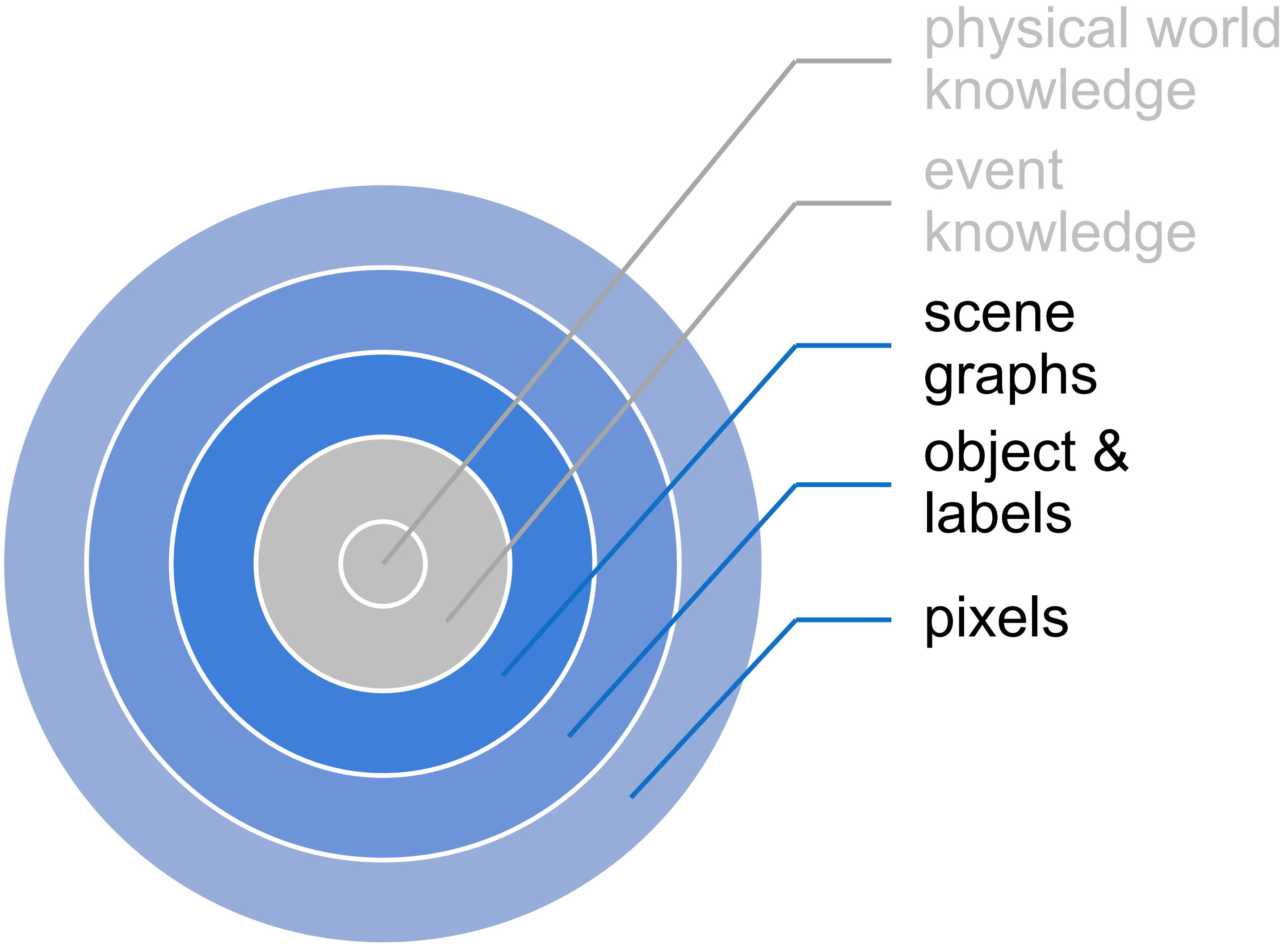


image source: Align and Prompt: Video-and-Language Pre-training with Entity Prompts

Adding knowledge to pretraining models





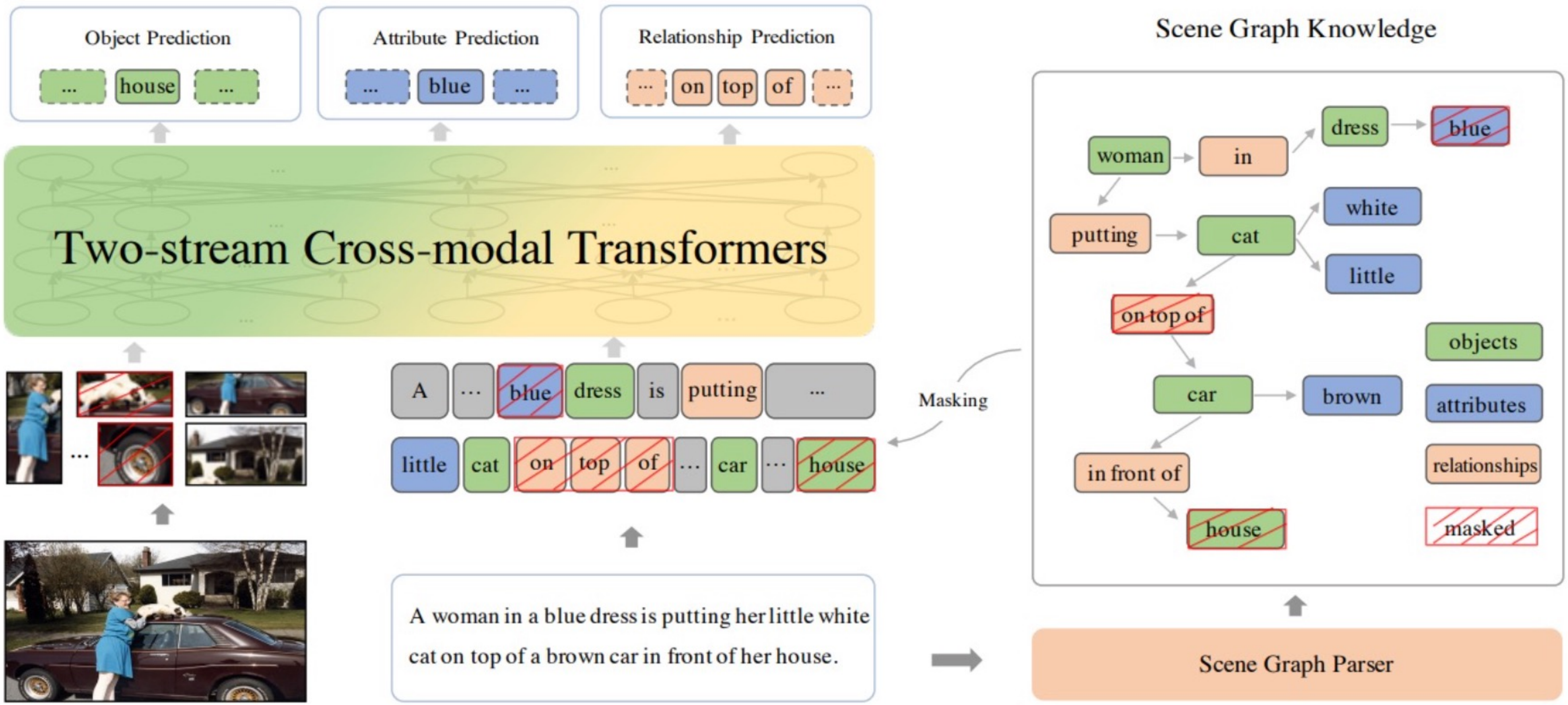
ERINE-ViL [AAAI2021]

- 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

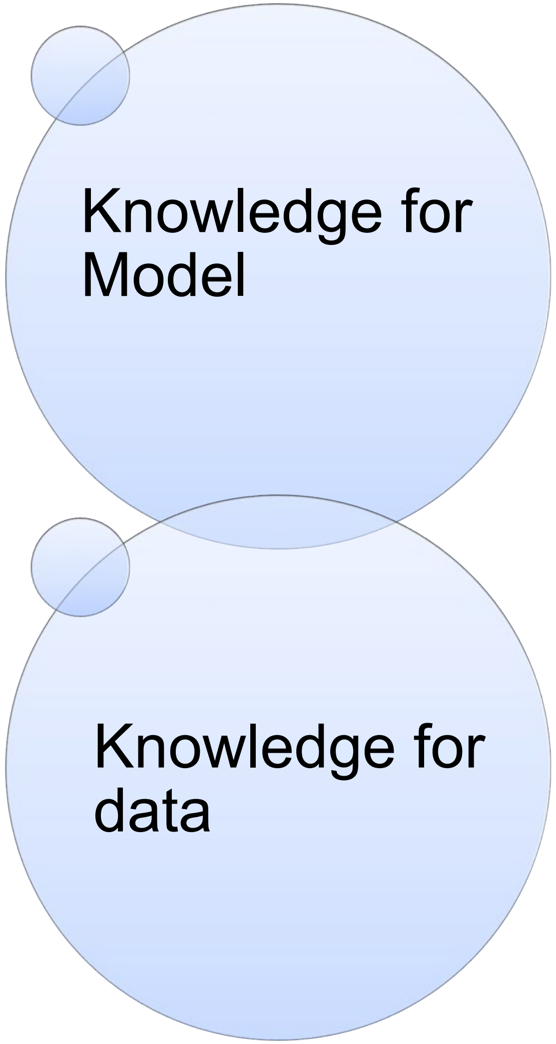
ERINE-ViL [AAAI2021]

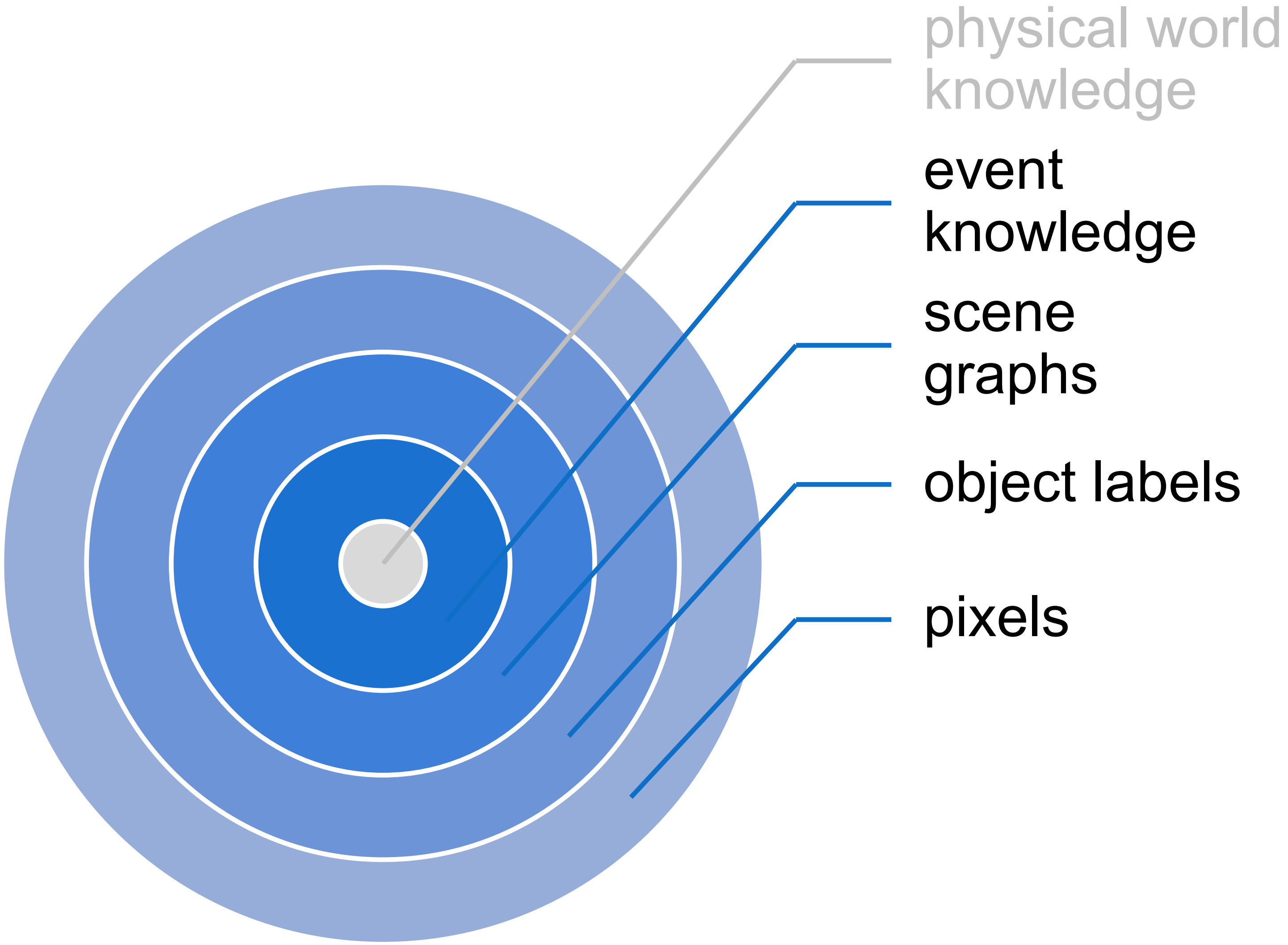
- Add scene graph knowledge as downstream tasks



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	ROLE	VALUE
AGENT	MAN	AGENT	VET
SOURCE	SHEEP	SOURCE	DOG
TOOL	SHEARS	TOOL	CLIPPER
ITEM	WOOL	ITEM	CLAW
PLACE	FIELD	PLACE	ROOM

JUMPING

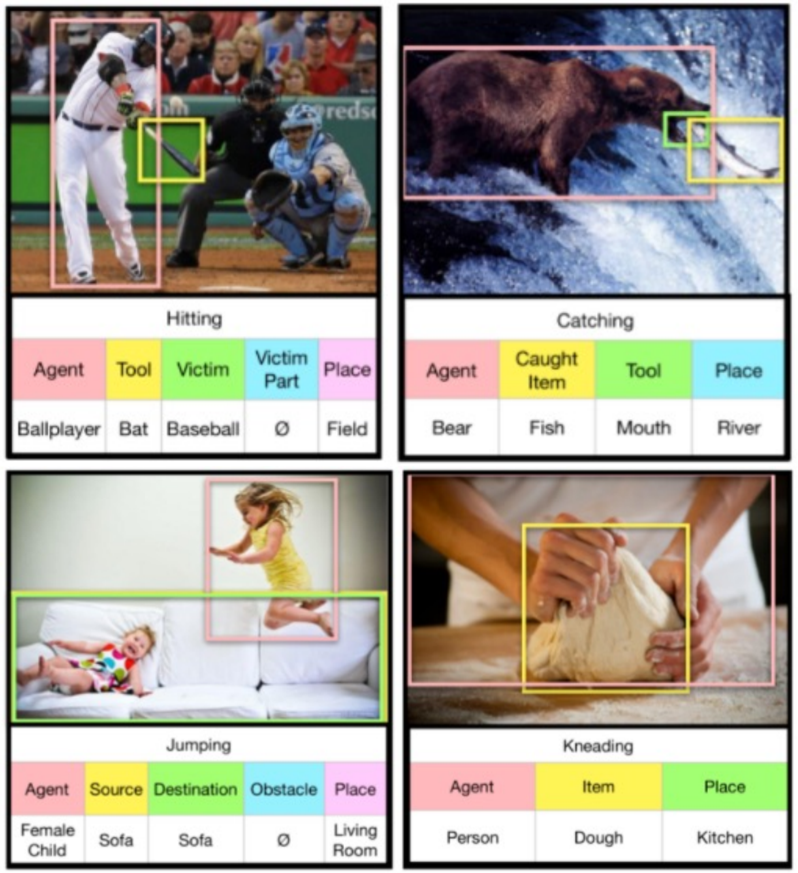
ROLE	VALUE	ROLE	VALUE
AGENT	BOY	AGENT	BEAR
SOURCE	CLIFF	SOURCE	ICEBERG
OBSTACLE	-	OBSTACLE	WATER
DESTINATION	WATER	DESTINATION	ICEBERG
PLACE	LAKE	PLACE	OUTDOOR

SPRAYING

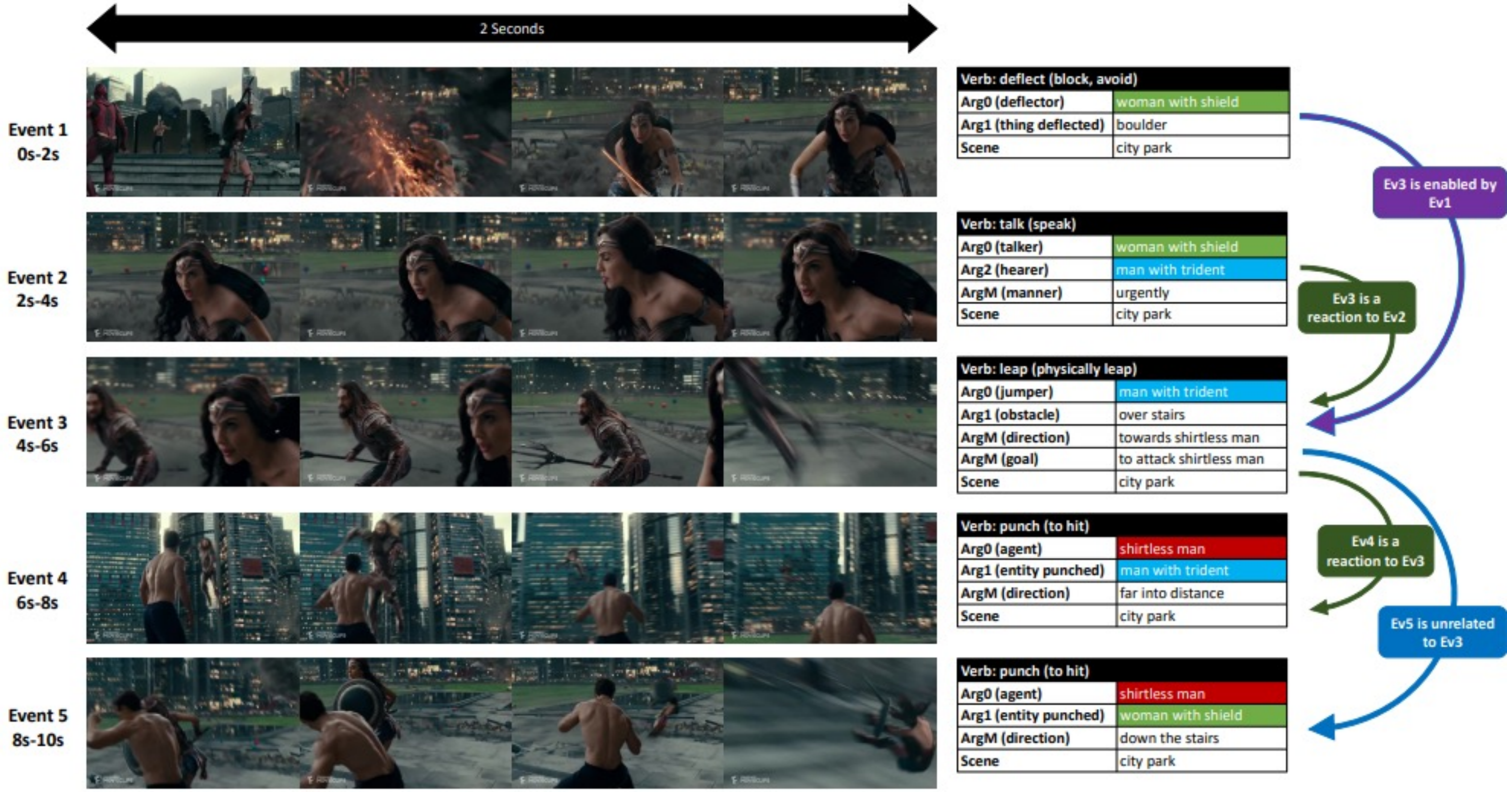
ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	FIREMAN
SOURCE	SPRAY CAN	SOURCE	HOSE
SUBSTANCE	PAINT	SUBSTANCE	WATER
DESTINATION	WALL	DESTINATION	FIRE
PLACE	ALLEYWAY	PLACE	OUTSIDE

Vision-only Event and Argument Extraction

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

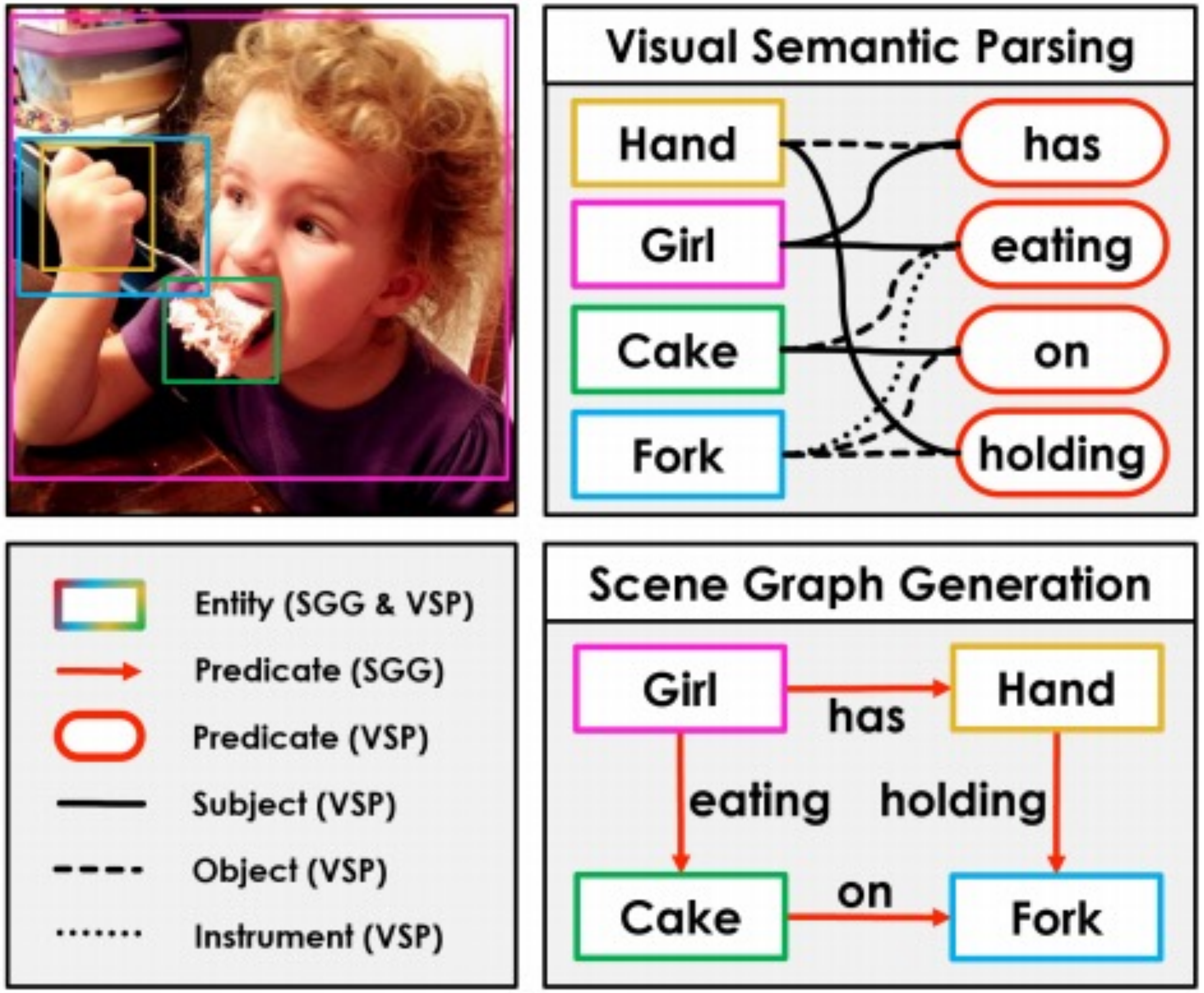


- Video Situation Recognition extends the work to videos [Sadhu et al, 2021]



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

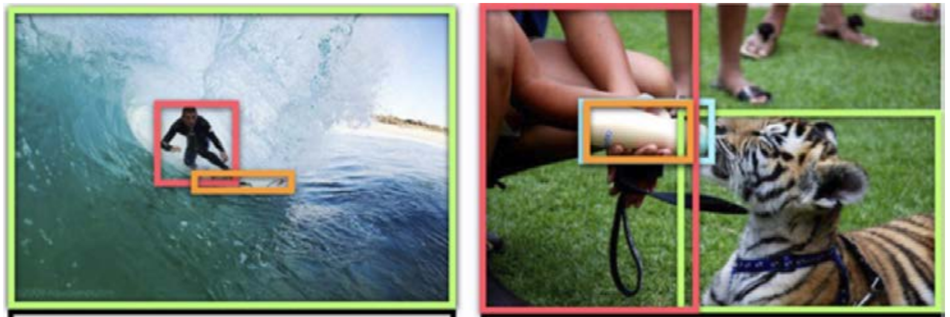
SPRAYING

ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	FIREMAN
SOURCE	SPRAY CAN	SOURCE	HOSE
SUBSTANCE	PAINT	SUBSTANCE	WATER
DESTINATION	WALL	DESTINATION	FIRE
PLACE	ALLEYWAY	PLACE	OUTSIDE

CLIPPING

ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	VET
SOURCE	SHEEP	SOURCE	DOG
TOOL	SHEARS	TOOL	CLIPPER
ITEM	WOOL	ITEM	CLAW
PLACE	FIELD	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

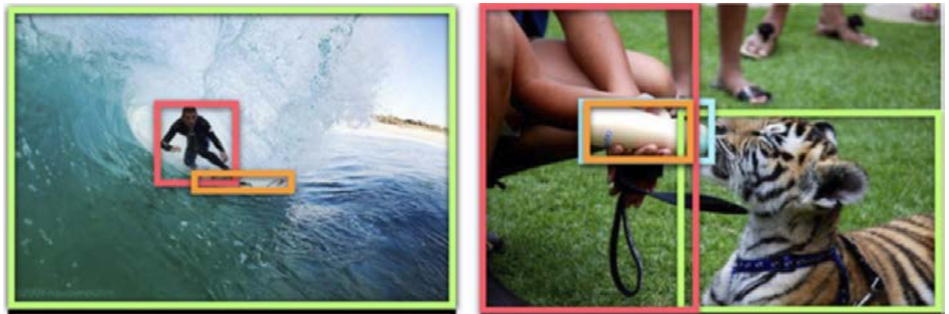
SPRAYING

ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	FIREMAN
SOURCE	SPRAY CAN	SOURCE	HOSE
SUBSTANCE	PAINT	SUBSTANCE	WATER
DESTINATION	WALL	DESTINATION	FIRE
PLACE	ALLEYWAY	PLACE	OUTSIDE

CLIPPING

ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	VET
SOURCE	SHEEP	SOURCE	DOG
TOOL	SHEARS	TOOL	CLIPPER
ITEM	WOOL	ITEM	CLAW
PLACE	FIELD	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

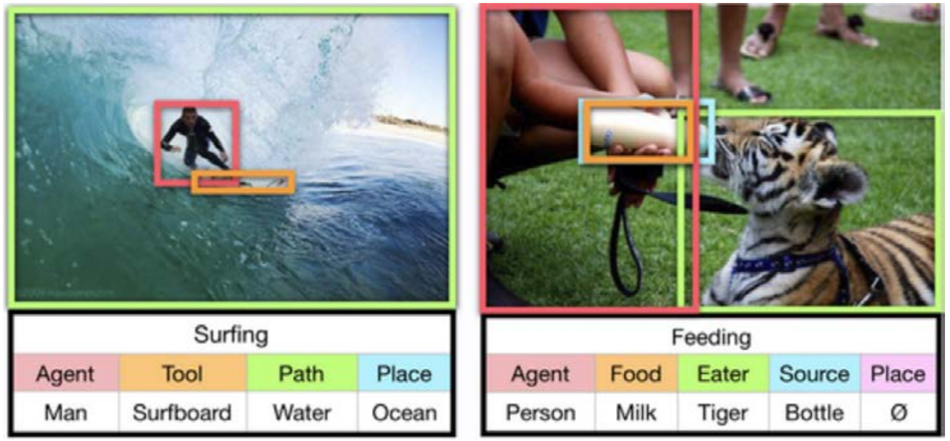
Bottleneck: Cross-modal Fusion

Transfer

Language → Vision

Existing Work: Situation Recognition

(Yatskar et al., 2016, …)



(Pratt et al., 2020, …)

Supervised Learning

Bottleneck: Lack of Annotation

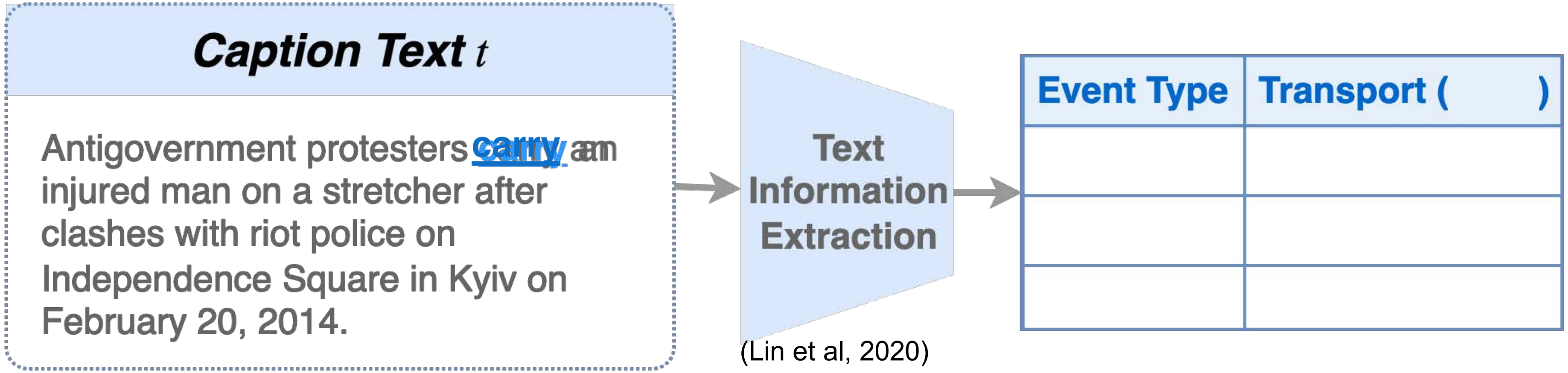
Vision-Only

Bottleneck: Cross-modal Fusion

Structural Transfer

Language → Vision

CLIP-Event: Event-Driven Vision-Language Pretraining



[Manling Li, et al., CVPR 2022]

CLIP-Event: Event-Driven Vision-Language Pretraining

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	
Target	
Instrument	

Transfer text event knowledge to images

Caption Text t

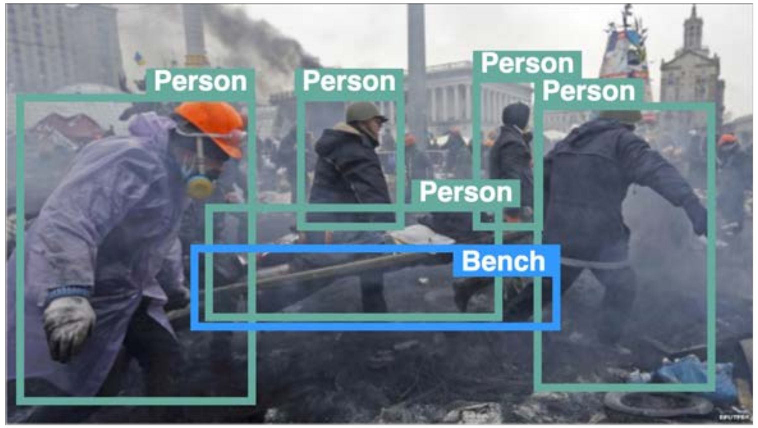
Antigovernment protesters carry an injured man on a stretcher after clashes with riot police on Independence Square in Kyiv on February 20, 2014.

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.

Positive Labels

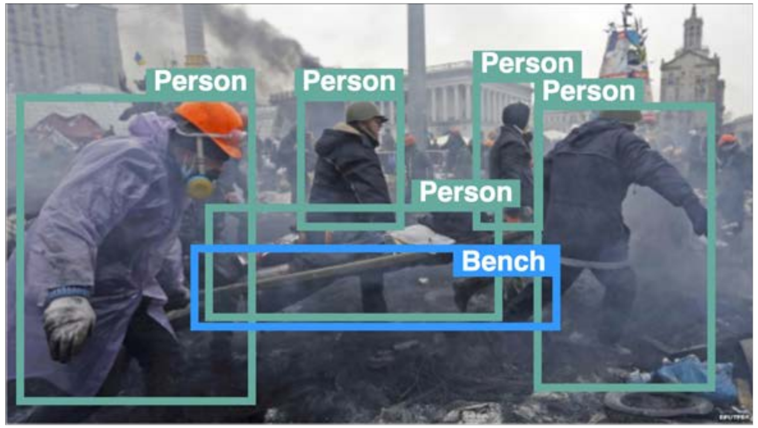
Event Type	Transport (carry)
Agent	protesters
Target	injured man
Instrument	stretcher

Confusion Matrix of existing V+L models

Negative Labels (events)

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

Image i



Hard negatives via manipulating event structures

Caption Text t

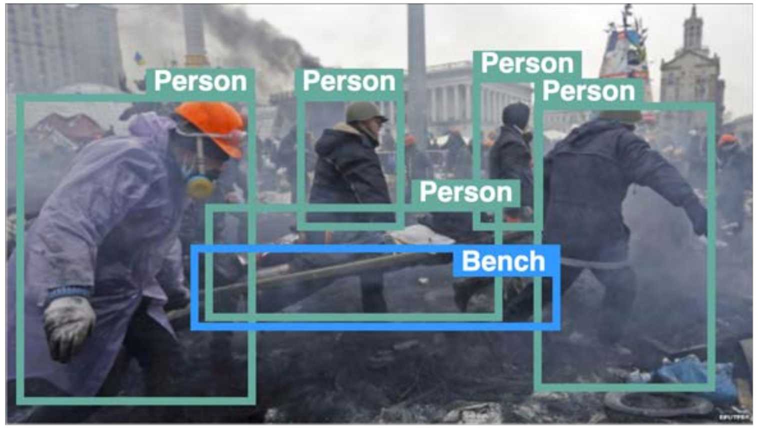
Antigovernment protesters carry an injured man on a stretcher after clashes with riot police on Independence Square in Kyiv on February 20, 2014.

Text Information Extraction

Positive Labels

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

Image i



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

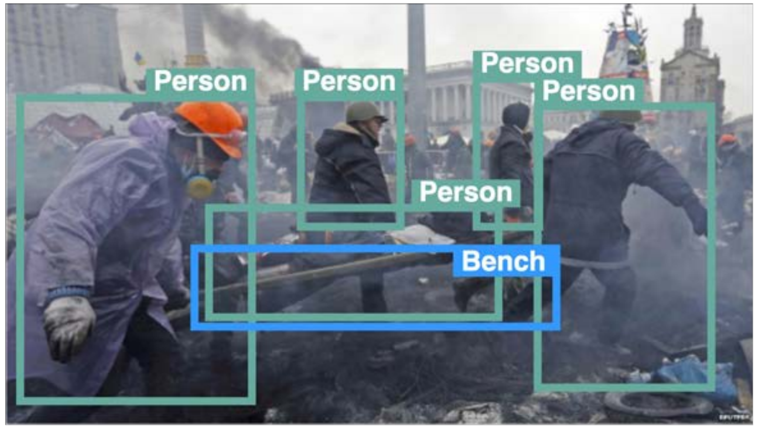
Hard negatives via manipulating event structures

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.

Text Information Extraction

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.

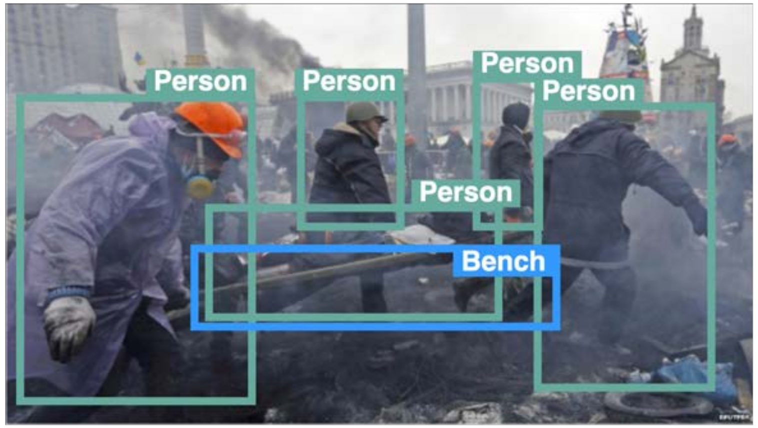
Hard negatives via manipulating event structures

***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.

Text Information Extraction

***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.

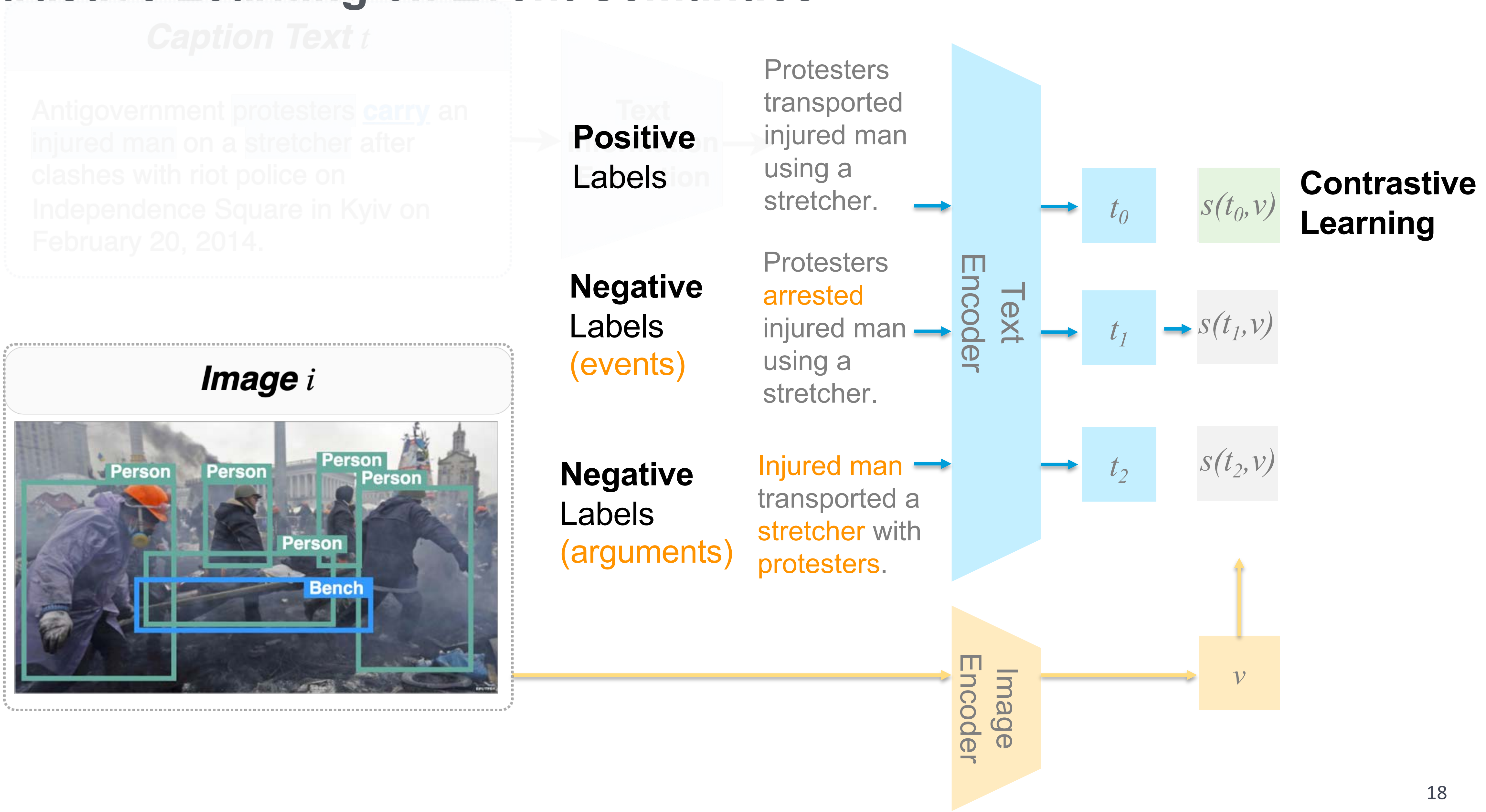
Negative Labels (arguments)

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

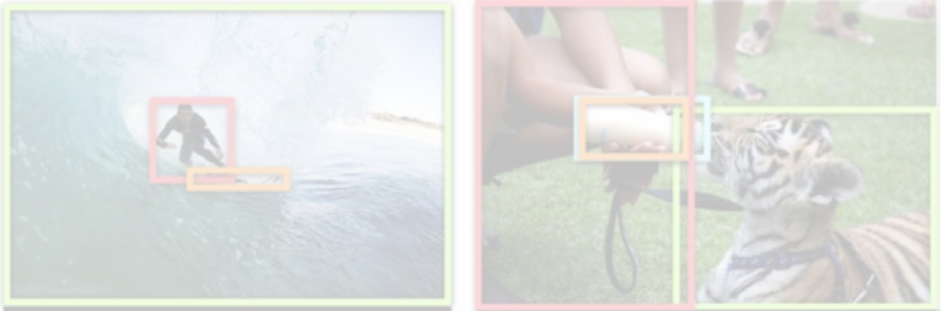
ROLE	VALUE
AGENT	FIREMAN
SOURCE	HOSE
SUBSTANCE	WATER
DESTINATION	FIRE
PLACE	OUTSIDE

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

(Yatskar et al., 2016, …)

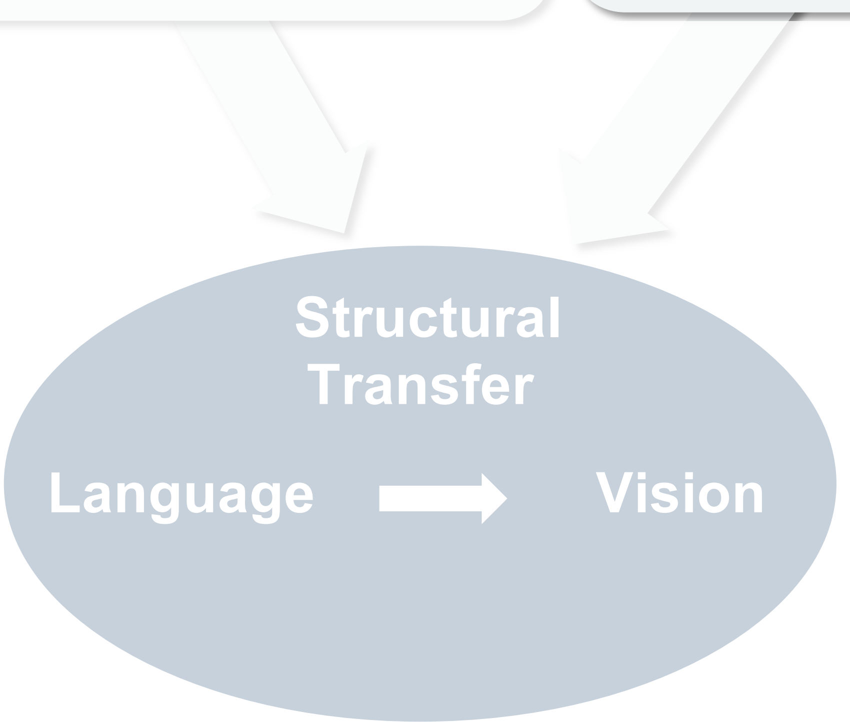
(Pratt et al., 2020, …)

Supervised Learning

Bottleneck: Lack of Annotation

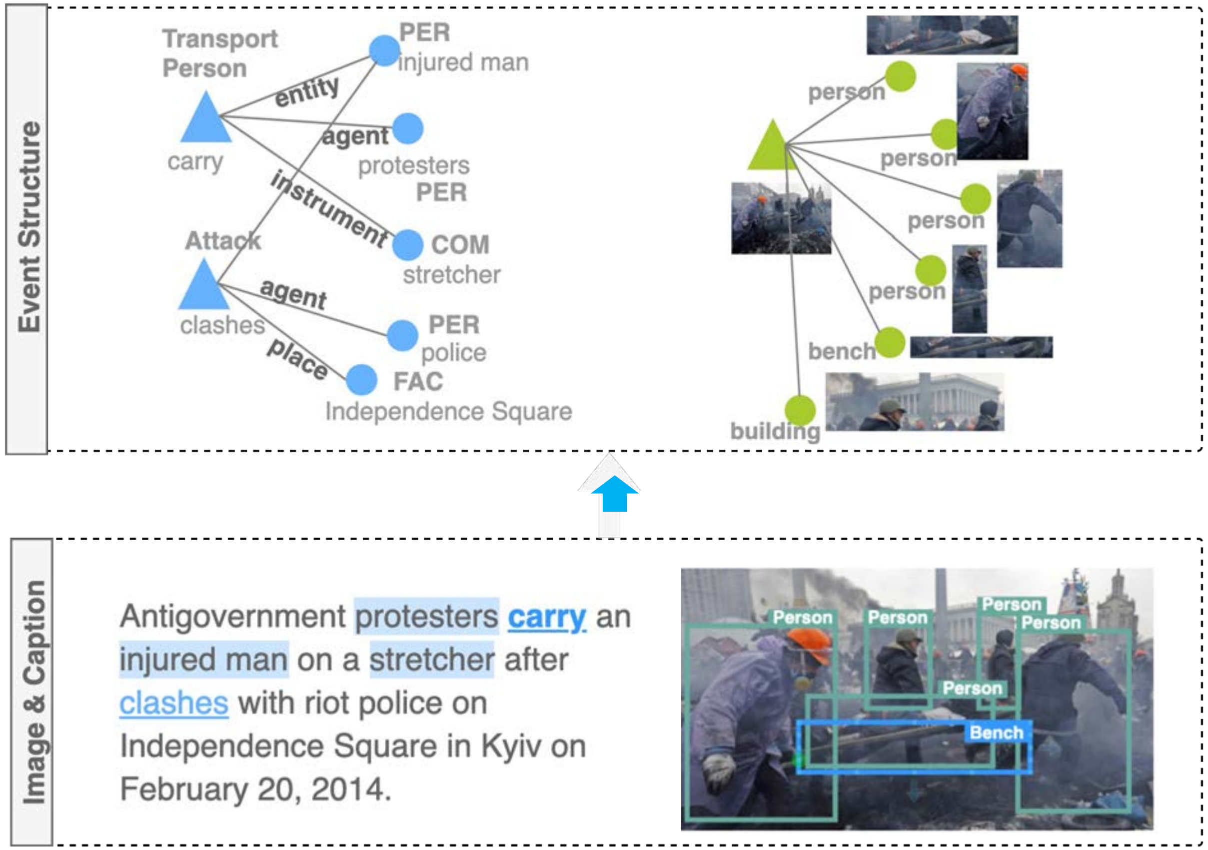
Vision-Only

Bottleneck: Cross-modal Fusion

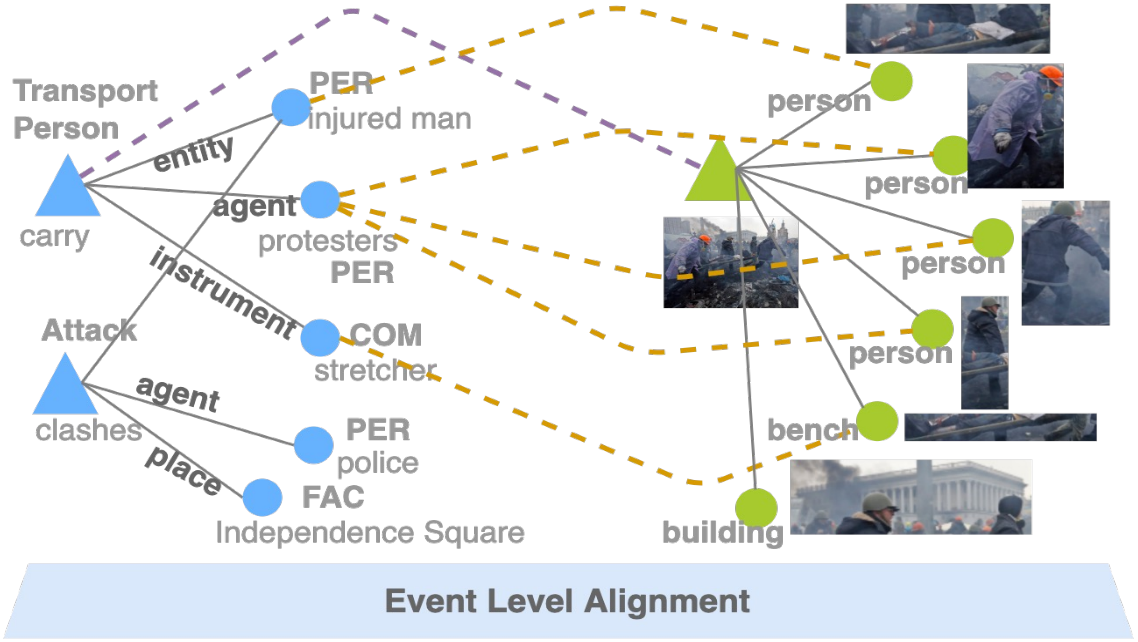


The first V+L Pretraining with Event Semantic Structures

Challenge: Structured Encoding



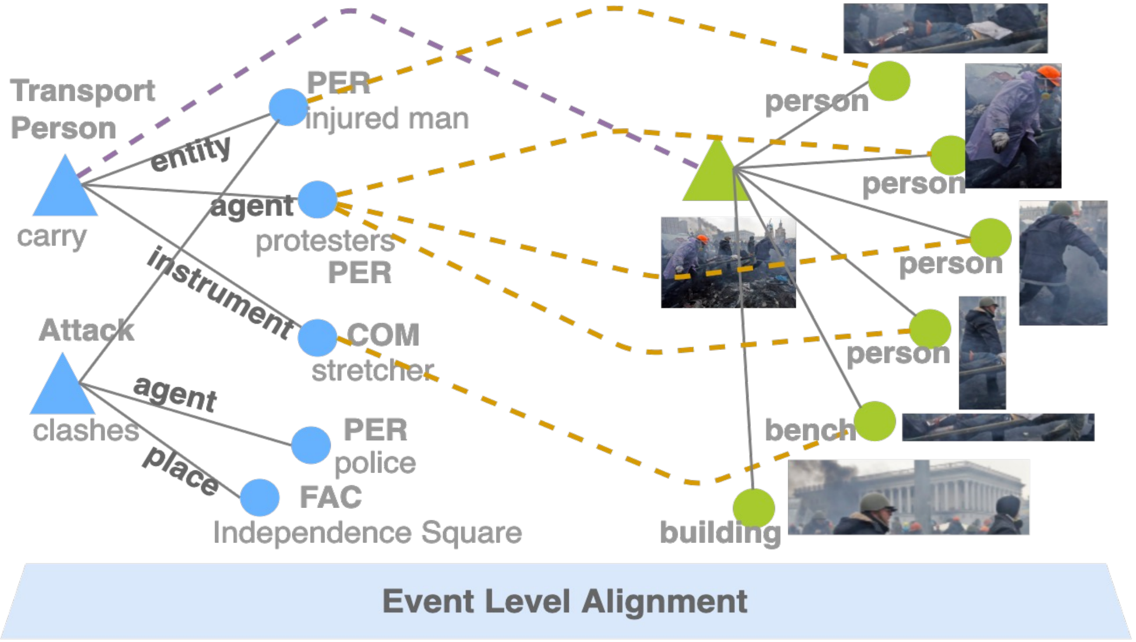
The first V+L Pretraining with Event Semantic Structures



The first V+L Pretraining with Event Semantic Structures

Structured Alignment via Optimal Transport

Text Event Graph ⟷ Image Event Graph

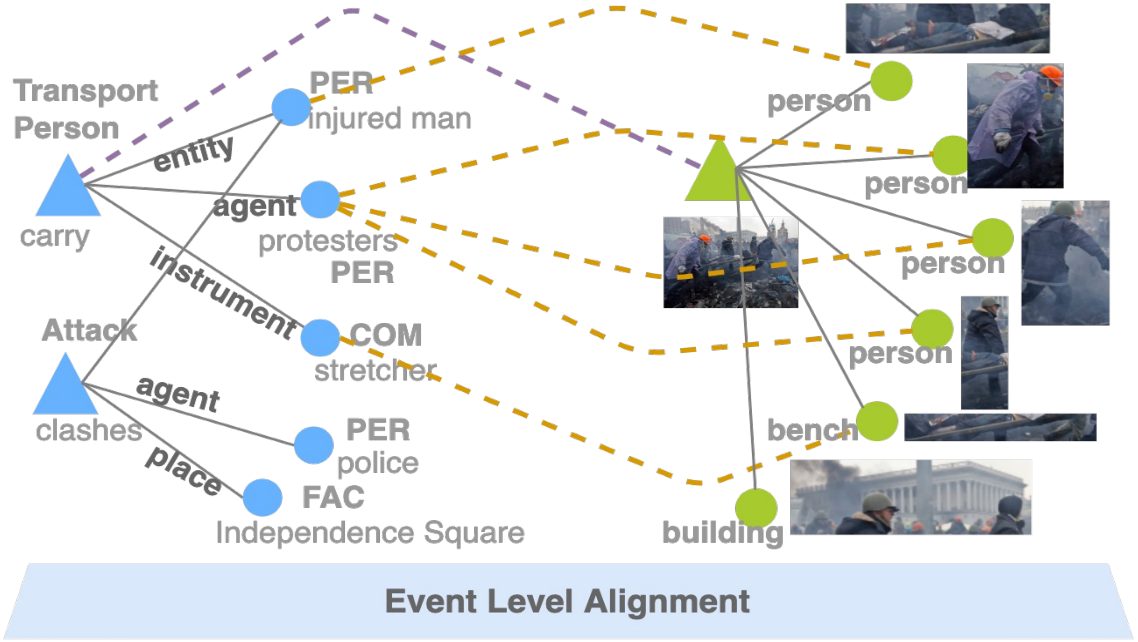


The first V+L Pretraining with Event Semantic Structures

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

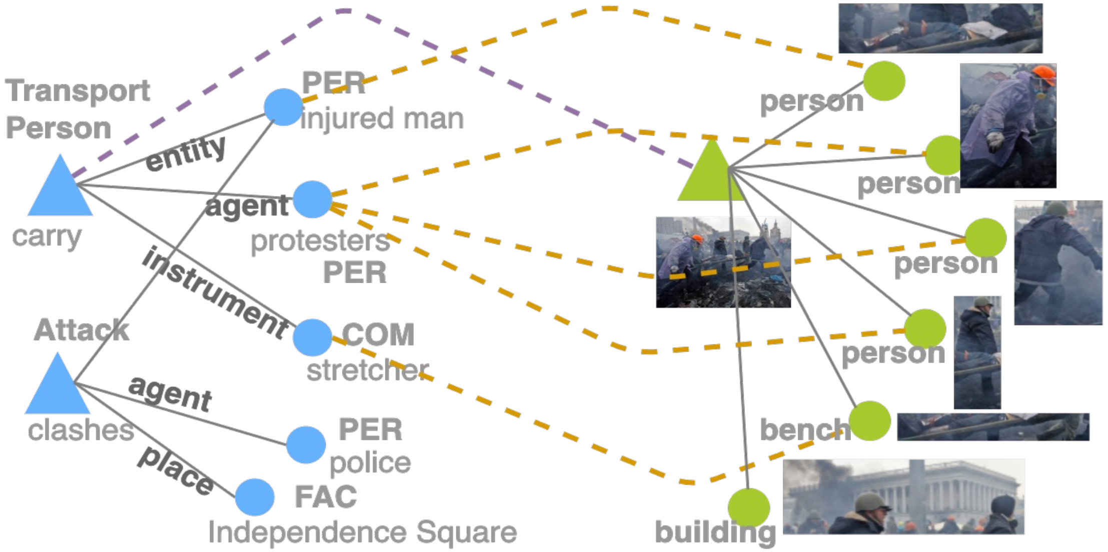
Structured Alignment via Optimal Transport

Text Event Graph ⟷ Image Event Graph

1. Define cost matrix C (embedding similarity)

2. Optimization Goal: minimize transport distance

$$D(S,T) = \min_{T} \boldsymbol{T} \cdot \boldsymbol{C}$$



Event Level Alignment

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

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

$$\begin{aligned} &\text{for } i = 0, 1, 2, \ldots \text{ until convergence,} \\ &\boldsymbol{p}^{i+1} = \mathbf{1} \oslash (\boldsymbol{K}\boldsymbol{q}^{i}), \\ &\boldsymbol{q}^{i+1} = \mathbf{1} \oslash (\boldsymbol{K}^{\top}\boldsymbol{p}^{i+1}), \end{aligned}$$

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

The first V+L Pretraining with Event Semantic Structures

Structured Alignment via Optimal Transport

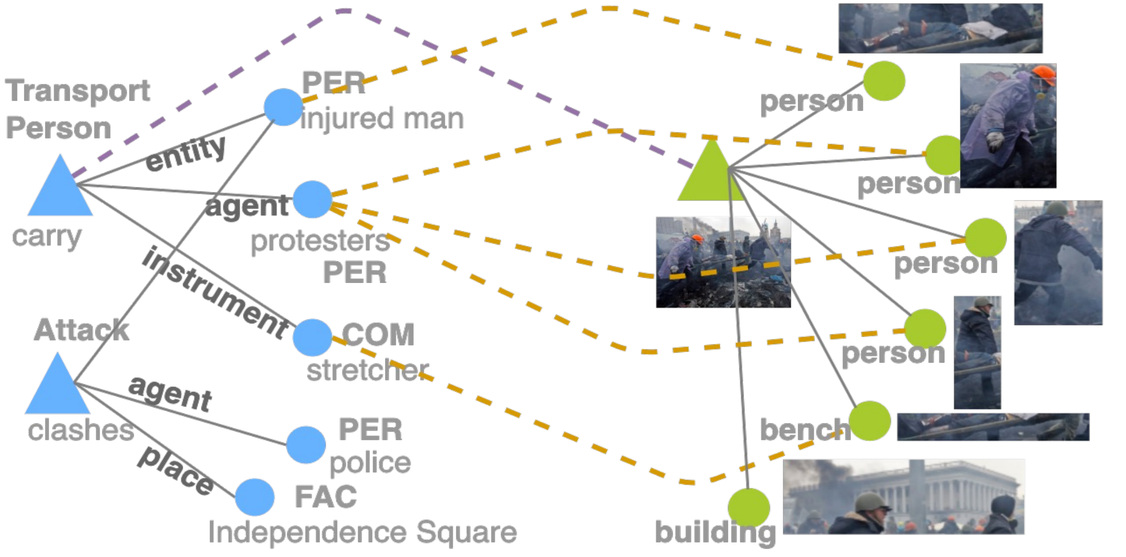
Text Event Graph ⟷ Image Event Graph

1. Define cost matrix C (embedding similarity)

2. Optimization Goal: minimize transport distance

$$D(S,T) = \min_{T} \boldsymbol{T} \cdot \boldsymbol{C}$$

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)

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

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

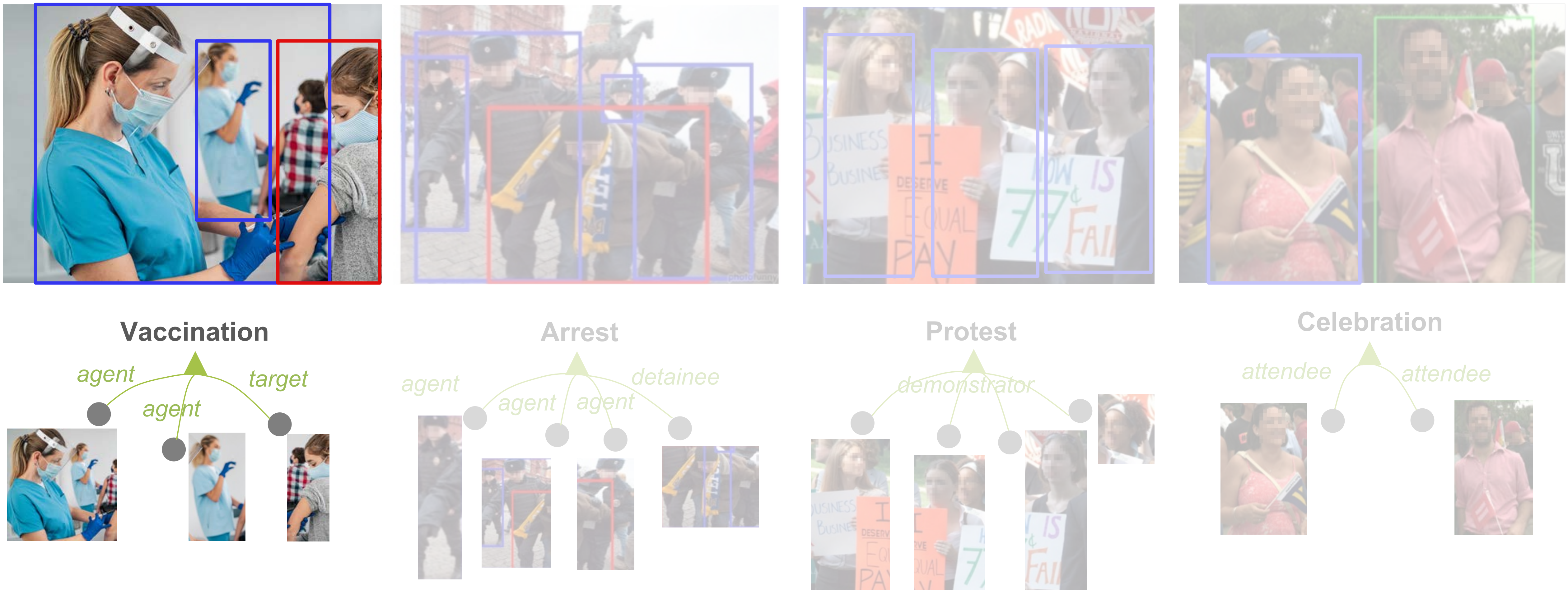
$$\boldsymbol{p}^{i+1} = \boldsymbol{1} \oslash (\boldsymbol{K}\boldsymbol{q}^{i}),$$

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

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

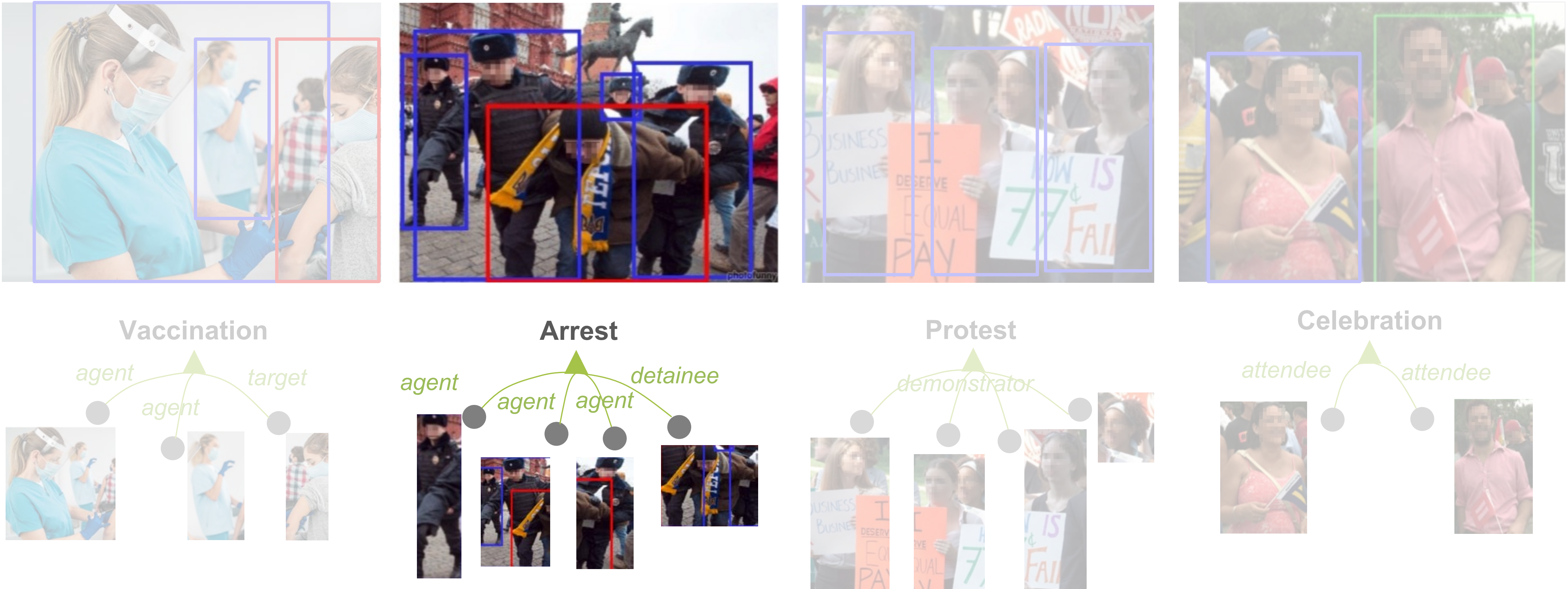
CLIP-Event on Visual Event Extraction

Supporting Zero-shot Vision Event Extraction the first time.



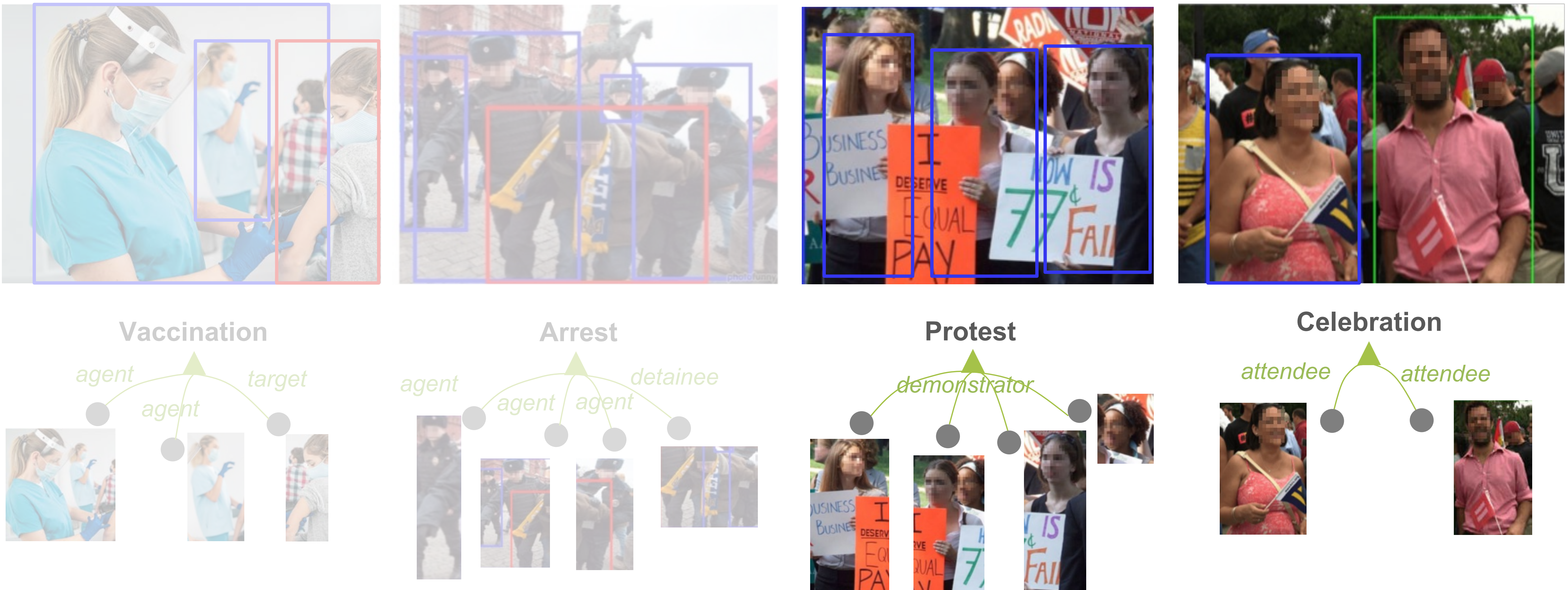
CLIP-Event on Visual Event Extraction

Supporting **Zero-shot** Vision Event Extraction the first time.

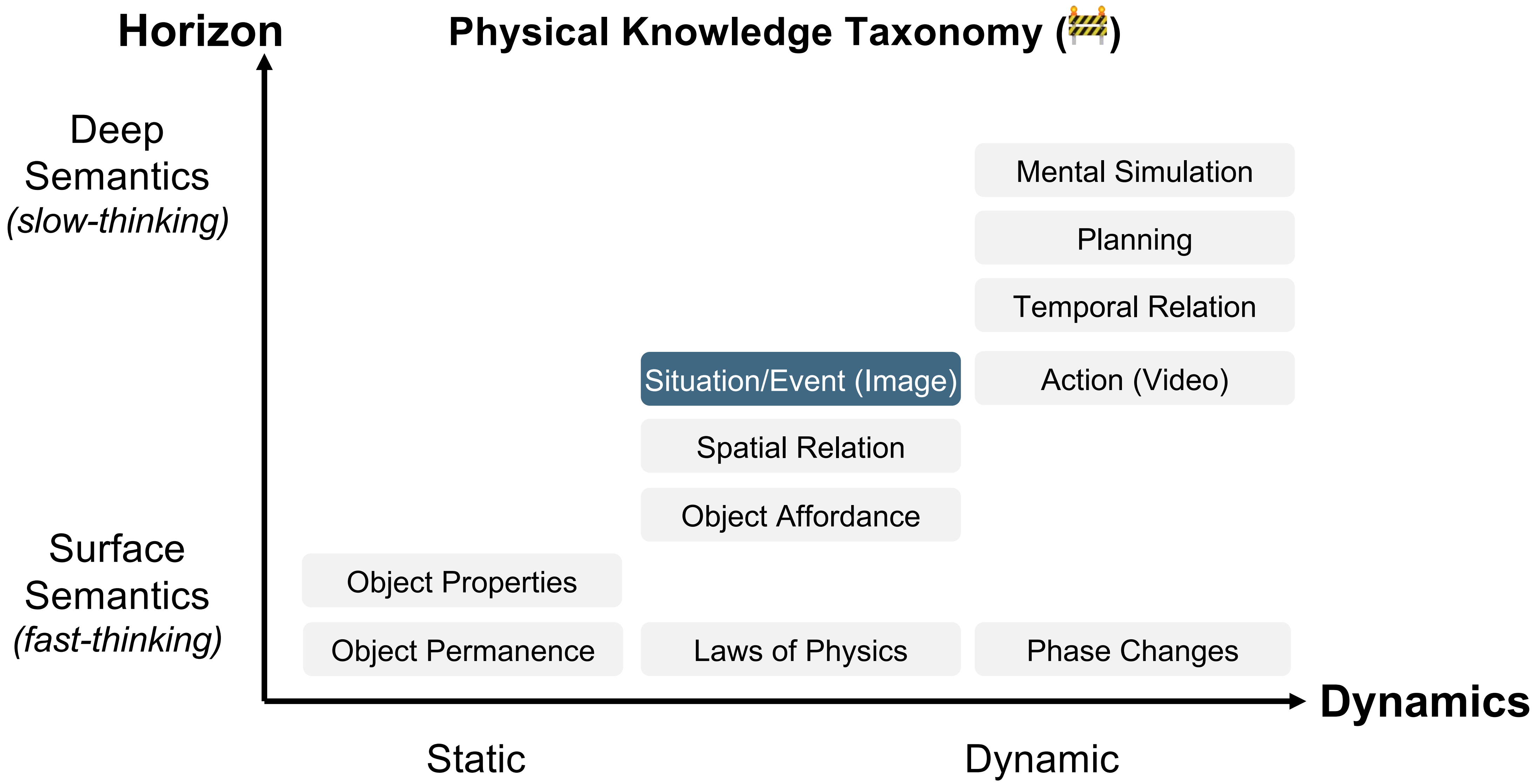


CLIP-Event on Visual Event Extraction

Supporting **Zero-shot** Vision Event Extraction the first time.



Our Research Goal: Surface → Deep Semantics



Our Research Goal: Surface → Deep Semantics

Horizon

Physical Knowledge Taxonomy (🚧)

Deep Semantics
(slow-thinking)

Mental Simulation

Planning

Temporal Relation

Situation/Event (Image)

Action (Video)

Spatial Relation

Object Affordance

Surface Semantics
(fast-thinking)

Object Properties

Object Permanence

Laws of Physics

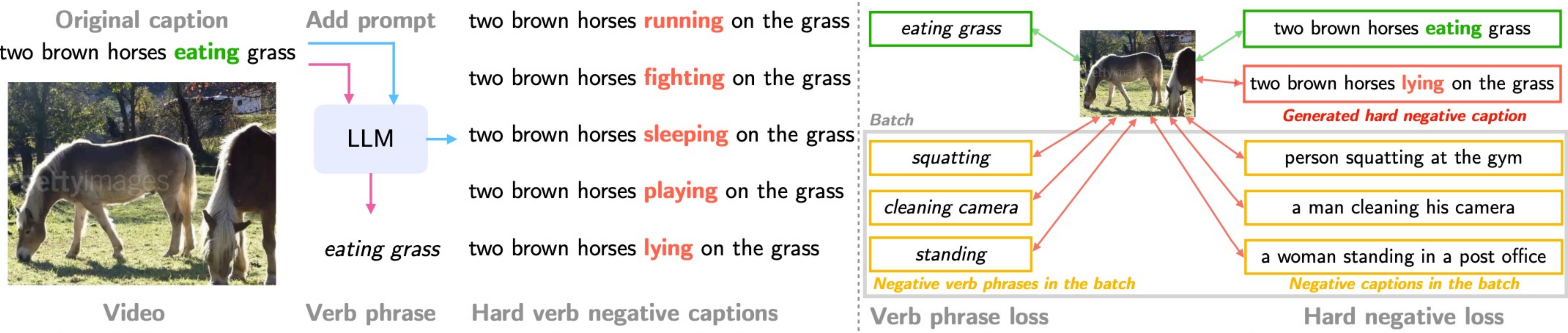
Phase Changes

Dynamics

Static

Dynamic

Verbs in Action: Improving verb understanding in video-language models



it's a video of a bald monk sitting at a temple looking at his laptop

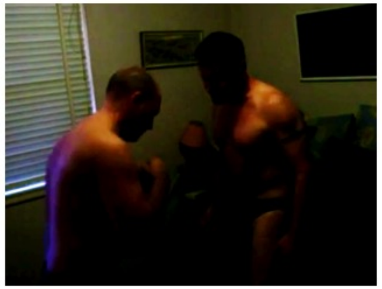
it's a video of a bald monk lying at a temple looking at his laptop
it's a video of a bald monk standing at a temple looking at his laptop
it's a video of a bald monk dancing around a temple holding his laptop
it's a video of a bald monk jumping up at a temple closing his laptop
it's a video of a bald monk running in a temple searching for his laptop

a person draws a dragon

a person carves a dragon
a person paints a dragon
a person doodles a dragon
a person sculpts a dragon
a person destroys a dragon

a girl skateboarding in a public place

a girl dancing in a public place
a girl running in a public place
a girl singing in a public place
a girl sitting on her skateboard in a public place
a girl falling off her skateboard in a public place



man is punching another man in the dark

man is arguing with another man in the dark
man is kissing another man in the dark
man is talking to another man in the daylight
man is kicking another man in the light
man is hugging another man in the dark

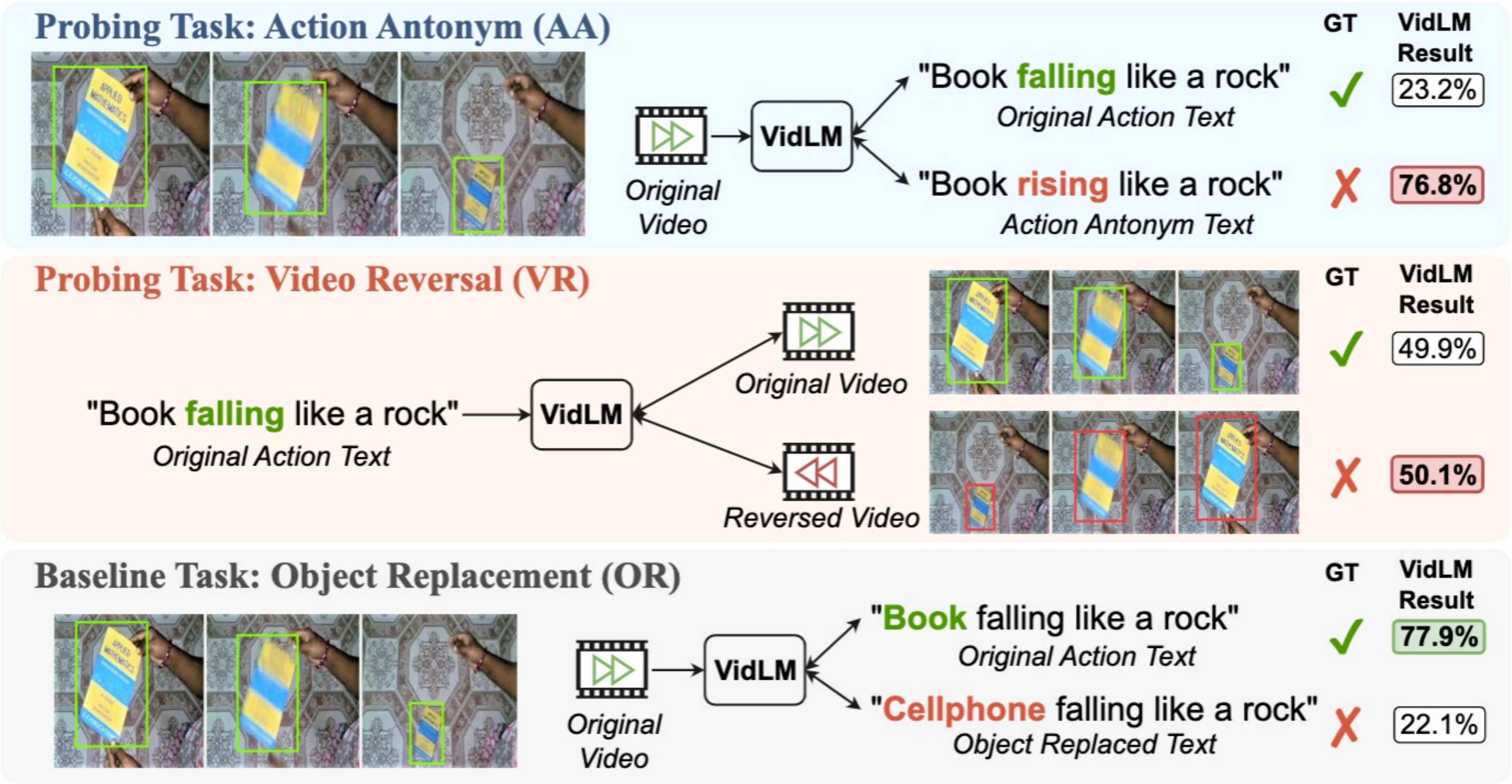
Video: A "Visual Recording" of World State Changes

Do SOTA Video-Language Models (VLM) possess fundamental **Action** Knowledge?

[Zhenhailong Wang, et al., Under Submission (NeurIPS 2023)]

Video: A "Visual Recording" of World State Changes

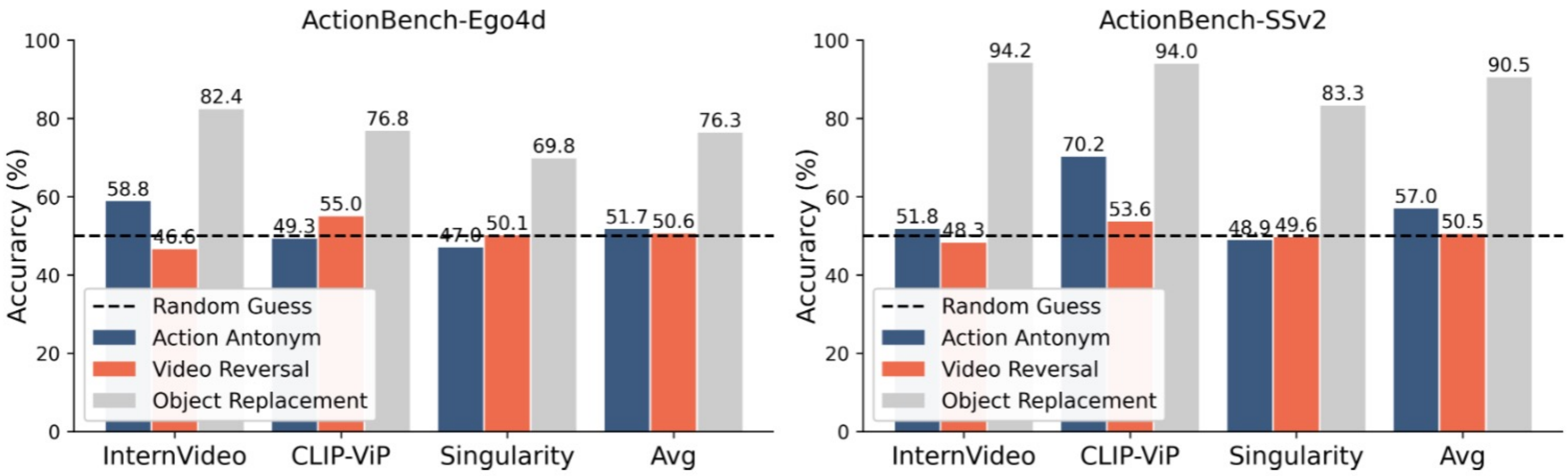
Do SOTA Video-Language Models (VLM) possess fundamental **Action** Knowledge?



Video: A “Visual Recording” of World State Changes

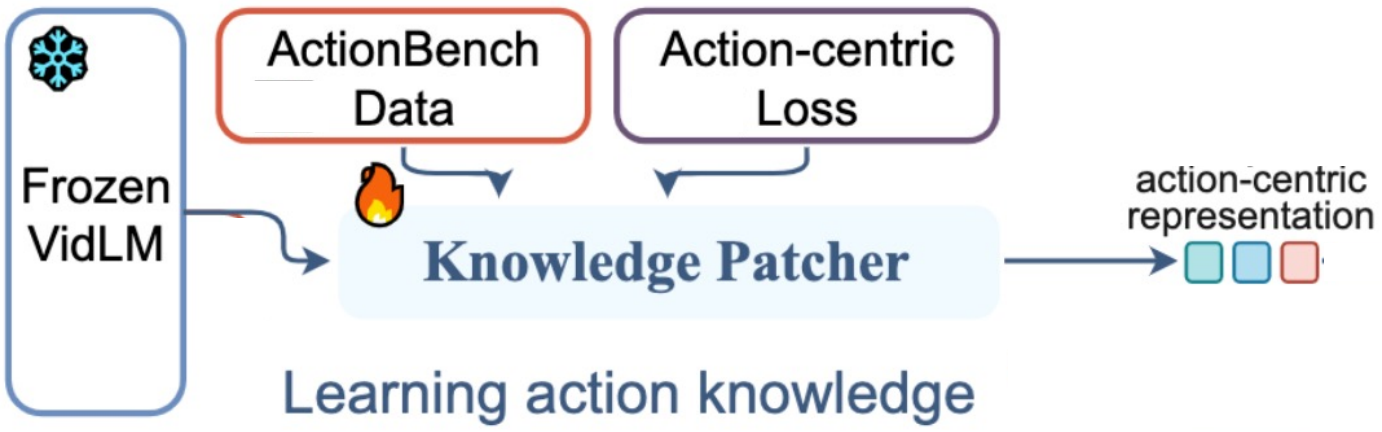
Do SOTA Video-Language Models (VLM) possess fundamental **Action** Knowledge?

- Near **random** performance on Action Antonym (AA) and Video Reversal (VR)
- Clear **biases towards objects** compared to actions



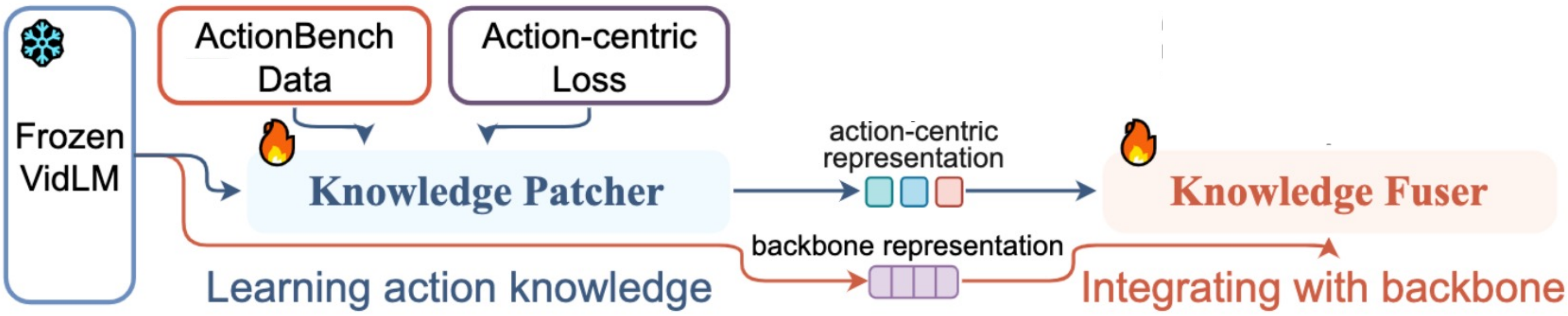
Patch & Fuse: Patching frozen VLMs with Action Knowledge

Patch frozen VidLMs with action knowledge without hurting their general VL capabilities.



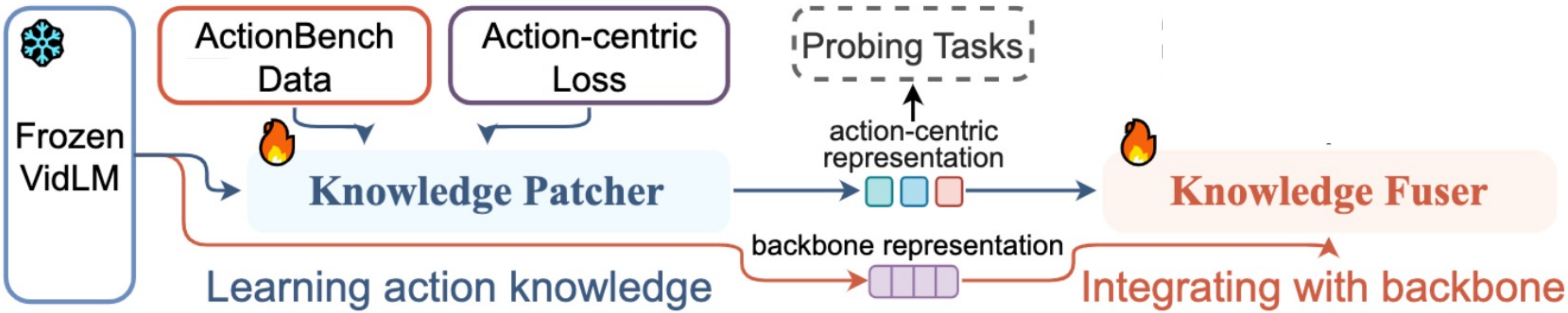
Patch & Fuse: Patching frozen VLMs with Action Knowledge

Patch frozen VidLMs with action knowledge without hurting their general VL capabilities.



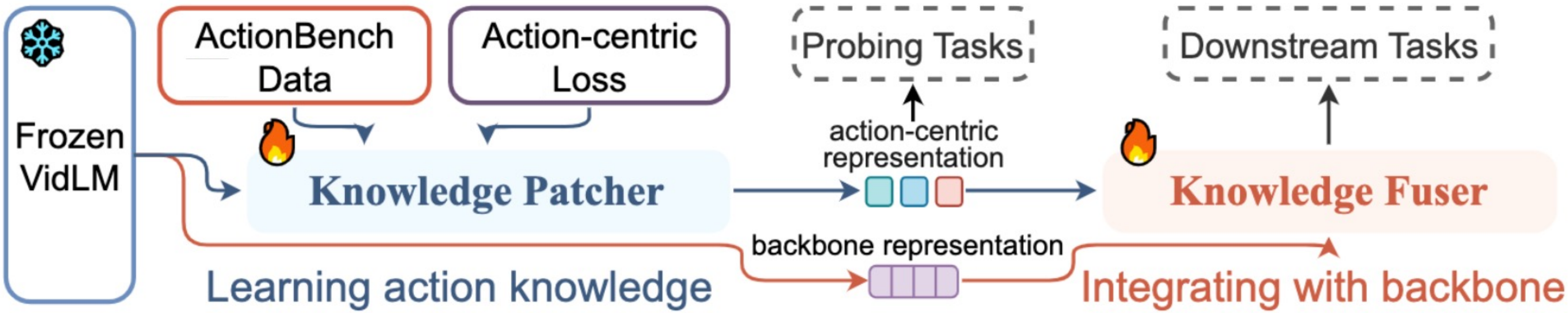
Patch & Fuse: Patching frozen VLMs with Action Knowledge

Patch frozen VidLMs with action knowledge without hurting their general VL capabilities.



Patch & Fuse: Patching frozen VLMs with Action Knowledge

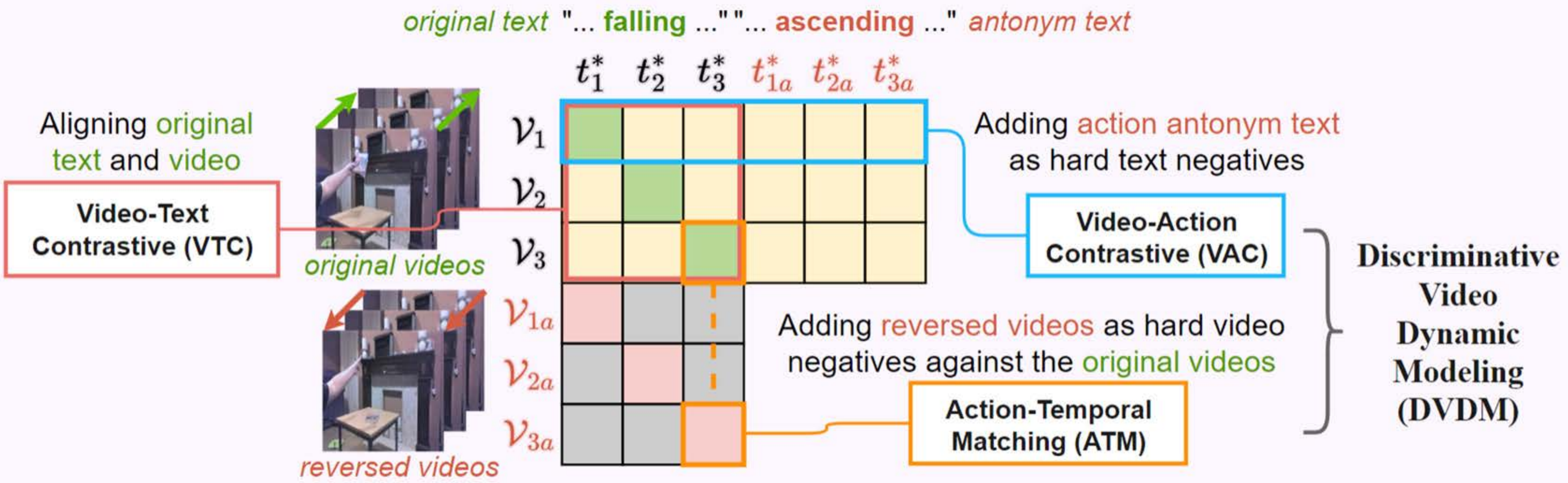
Patch frozen VidLMs with action knowledge without hurting their general VL capabilities.



Patch & Fuse: Patching frozen VLMs with Action Knowledge

Video-Action Contrastive (VAC): encourages learning the **alignment** between the **video** and the **action verbs**

Action-Temporal Matching: encourages learning the correct **temporal ordering** implied by the **action text**

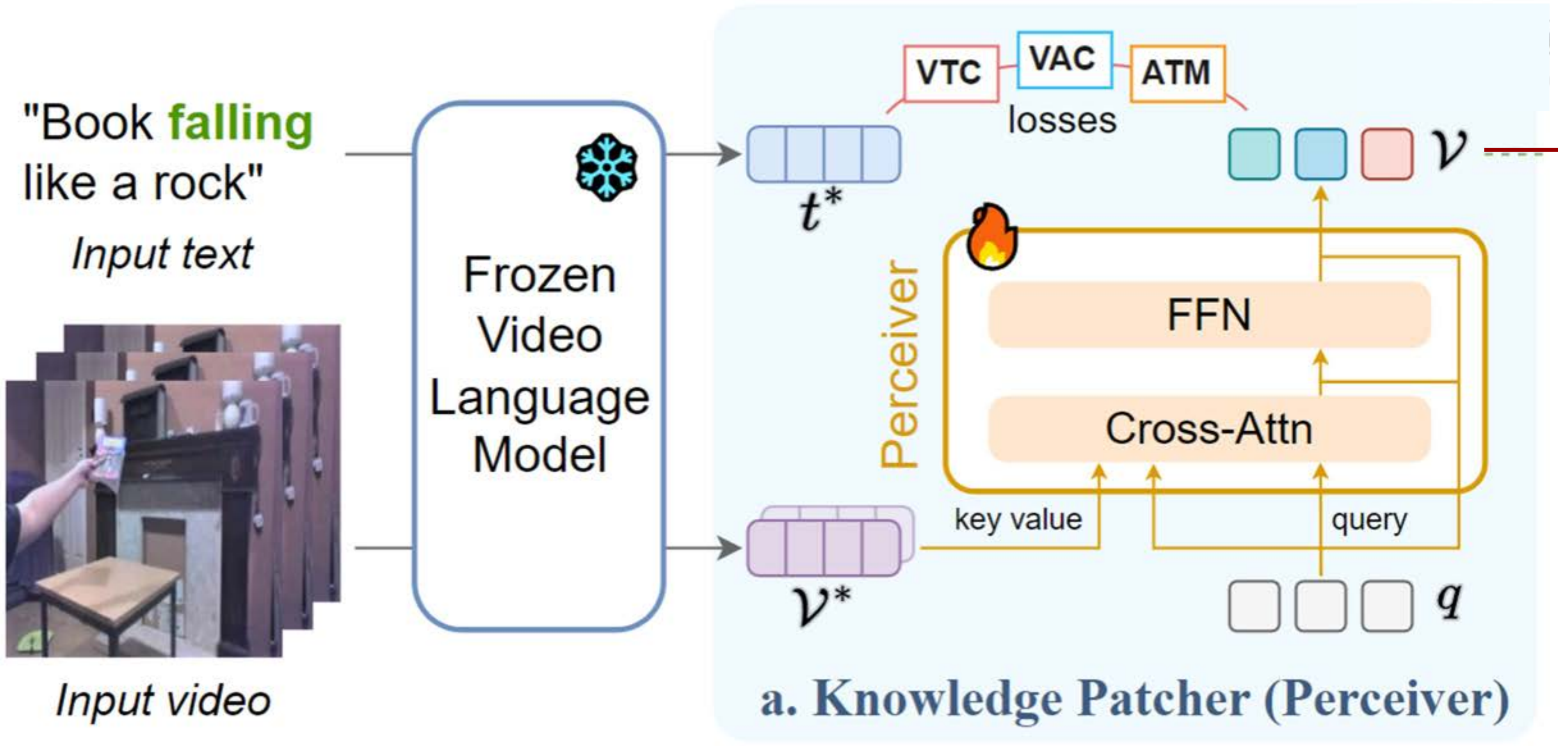


c. Training Objectives

Patch & Fuse: Patching frozen VLMs with Action Knowledge

Video-Action Contrastive (VAC): encourages learning the **alignment** between the **video** and the **action verbs**

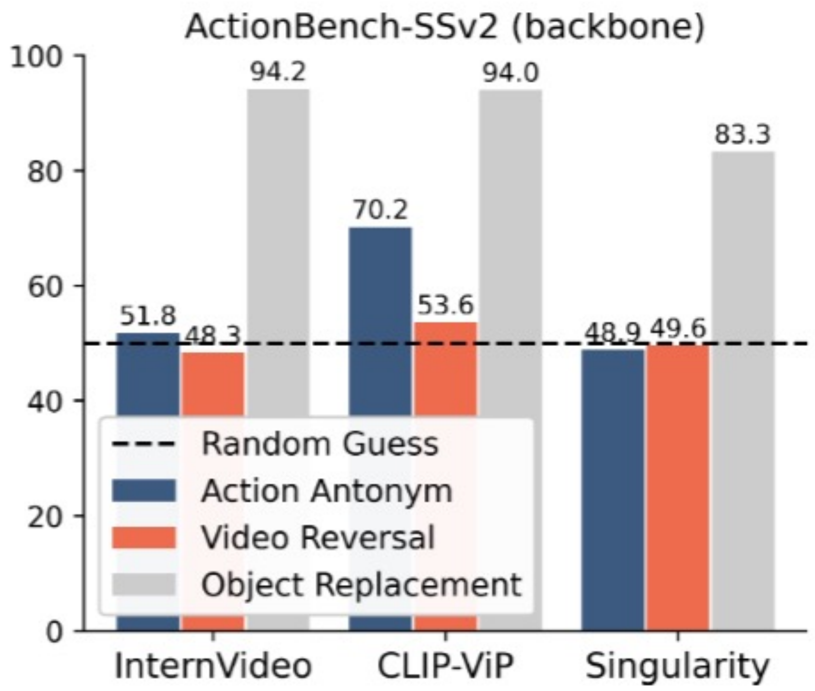
Action-Temporal Matching: encourages learning the correct **temporal ordering** implied by the **action text**



action-centric understanding

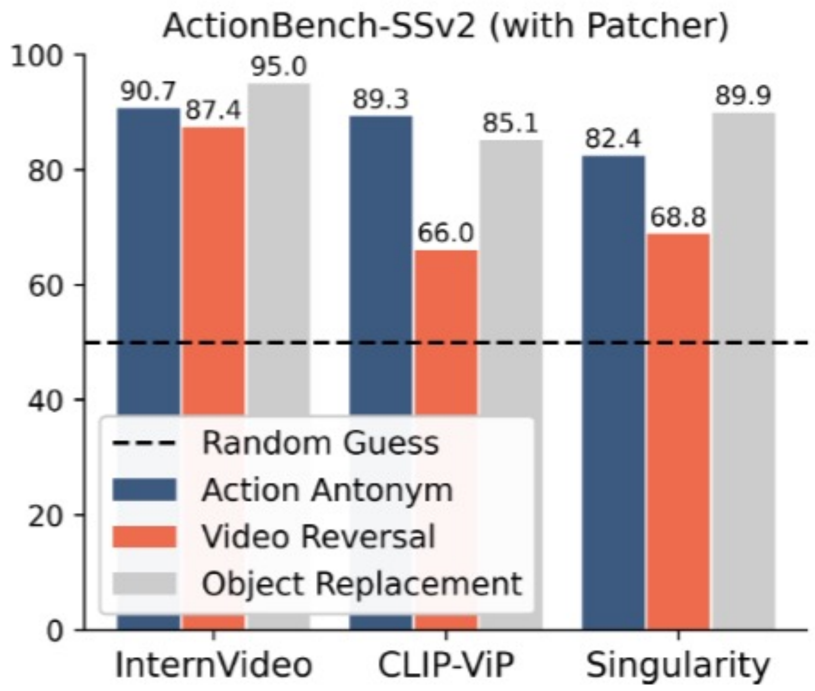
Results regarding Patch: **Before vs After** adding Knowledge Patcher

Adding the Knowledge Patcher nearly doubles the performance.



+ Knowledge Patcher

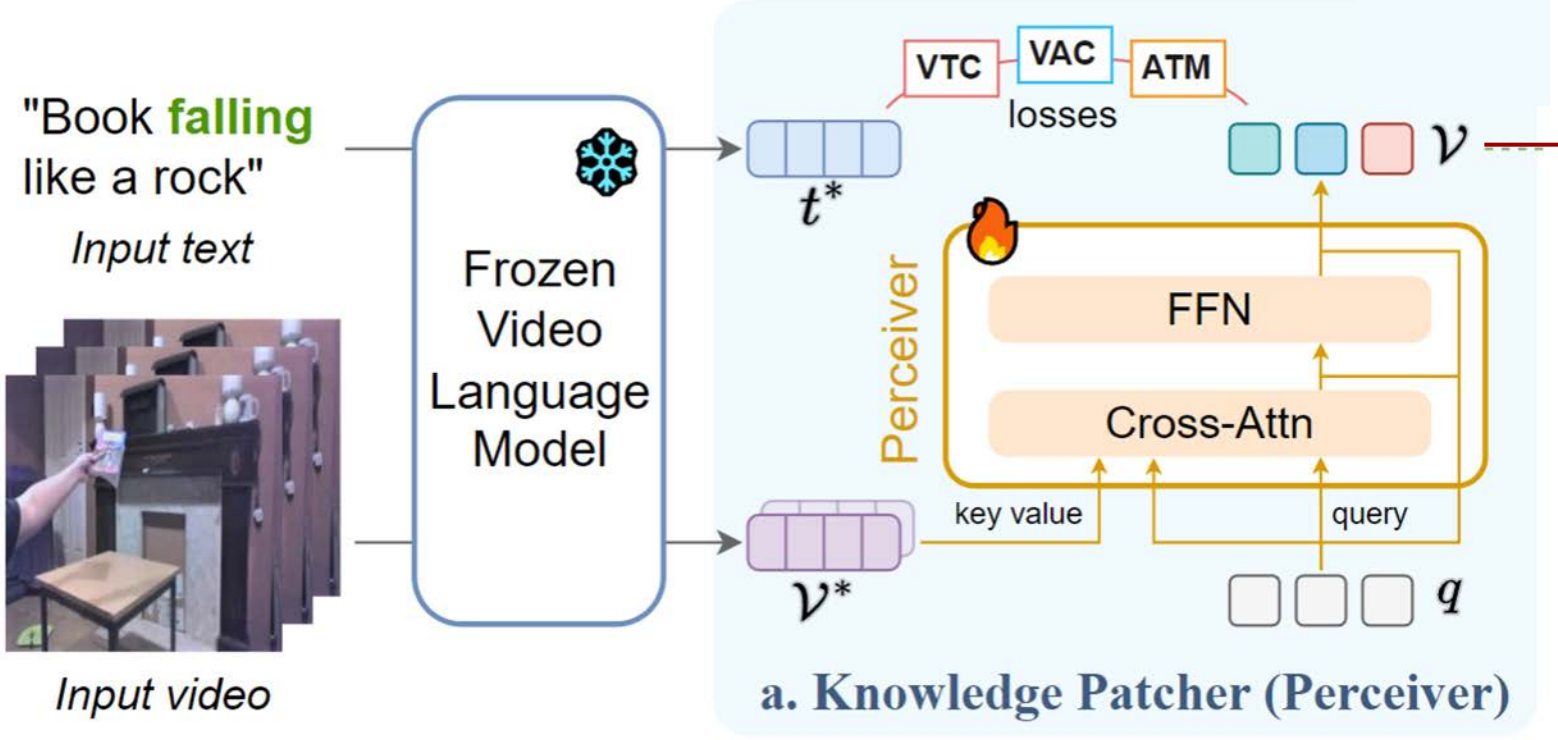
Trained with VTC+DVDM



Patch & Fuse: Retaining VL Capabilities

Video-Action Contrastive (VAC): encourages learning the **alignment** between the **video** and the **action verbs**

Action-Temporal Matching: encourages learning the correct **temporal ordering** implied by the **action text**

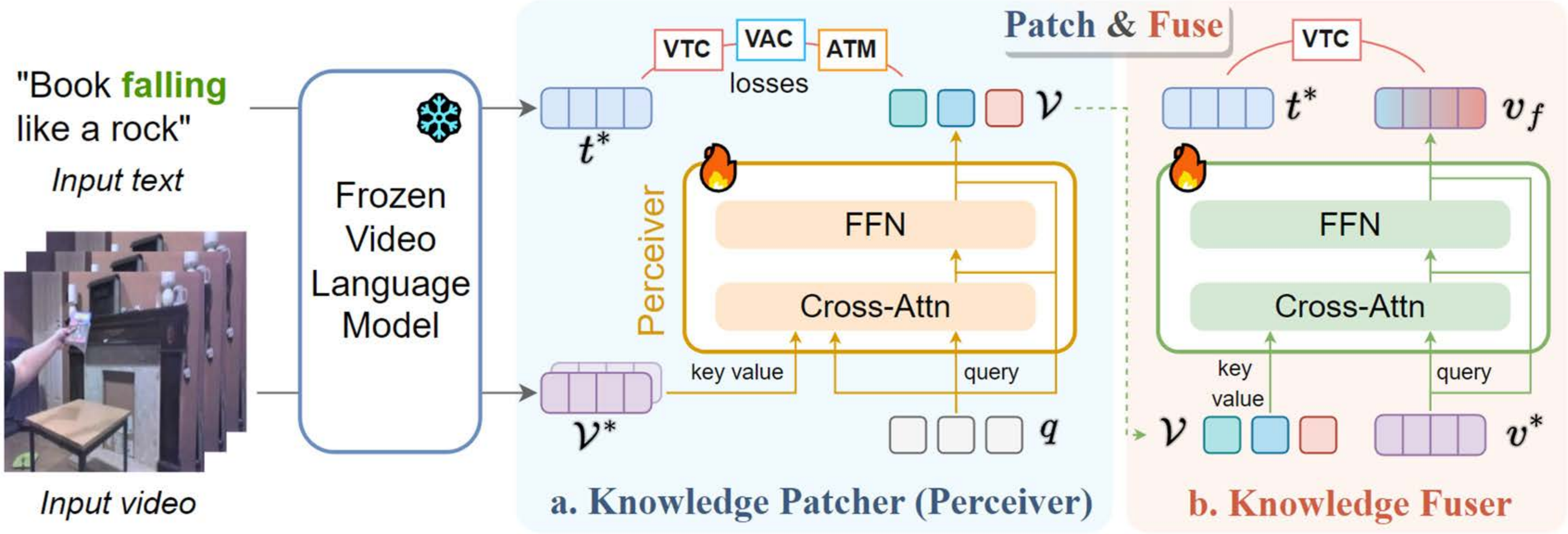


action-centric understanding

→ not favorable if a task is not action-centric

Patch & Fuse: Retaining VL Capabilities

A **unified representation** that has good understanding of both actions and objects.



Patch & Fuse: Retaining VL Capabilities

Video-Text Retrieval	Causal-Temporal VQA	Video-to-Action Retrieval	
SSv2-Label	NExT-QA	SSv2-Template	Temporal-SSv2

More object-centric

Require joint understanding of objects and actions

More action-centric

Example

Video-Text Retrieval: SSv2-label

"pushing scissors so that it falls off the table"

Example

Video-to-Action Retrieval: SSv2-template (where the main object is obfuscated)

"pushing **something** so that it falls off the table"

Results regarding Fuse: Retaining VL Capabilities

Video-Text Retrieval	Causal-Temporal VQA	Video-to-Action Retrieval	
SSv2-Label	NExT-QA	SSv2-Template	Temporal-SSv2

More object-centric

Require joint understanding of objects and actions

More action-centric

Method [Patcher Training Loss]	Video-Text Retrieval SSv2-label				Causal-Temporal VQA NExT-QA		Video-to-Action Retrieval SSv2-template		Temporal-SSv2	
	$R1_{v2t}$	$R5_{v2t}$	$R1_{t2v}$	$R5_{t2v}$	Val (Acc)	Test (Acc)	$R1$	$R5$	$R1$	$R5$
InternVideo Backbone	18.8	39.9	19.9	40.0	43.2	44.3	5.6	15.9	11.2	35.8
KP-Transformer FT [VTC]	24.1	50.0	21.7	46.0	48.1	49.6	21.1	55.9	41.1	88.9
KP-Perceiver FT [VTC]	27.0	57.4	27.1	**56.8**	48.0	49.5	24.8	59.7	42.5	91.3
Side-Tuning [83] [VTC+DVDM]	30.9	59.2	26.6	53.1	56.3	56.4	22.2	55.1	50.2	90.9
PATCH & FUSE [VTC+DVDM]	**32.3**	**61.2**	**28.0**	54.3	**56.9**	**56.6**	**26.9**	**61.5**	**51.2**	**91.9**

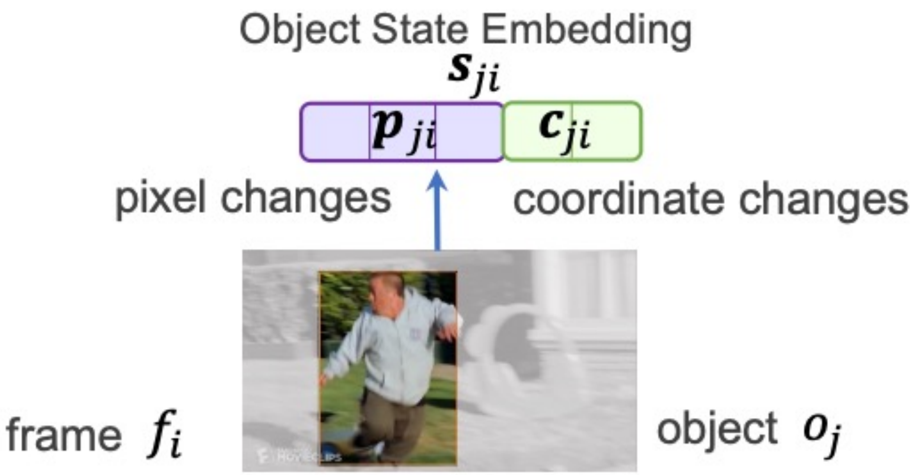
Performs competitively on **both object-centric and action-centric** tasks.

Video Events as Argument State Changes

Video Event =

Status Changes of Arguments

Status Changes of an object =
Displacement (movement of bounding box)
+
Pixel Changes (intra-boundingbox changing)



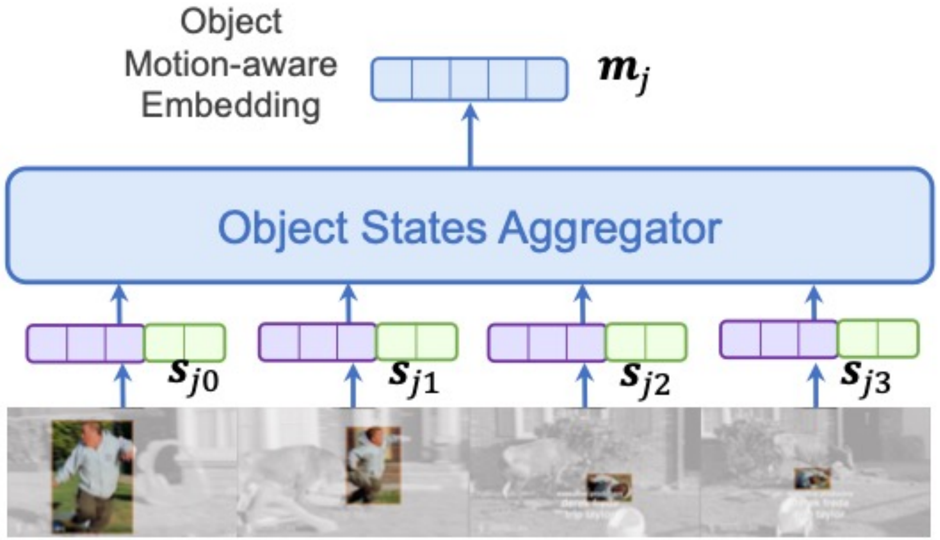
Object State Embedding of the man

Video Events as Argument State Changes

Video Event =

Status Changes of Arguments

Status Changes of an object =
Displacement (movement of bounding box)
+
Pixel Changes (intra-boundingbox changing)

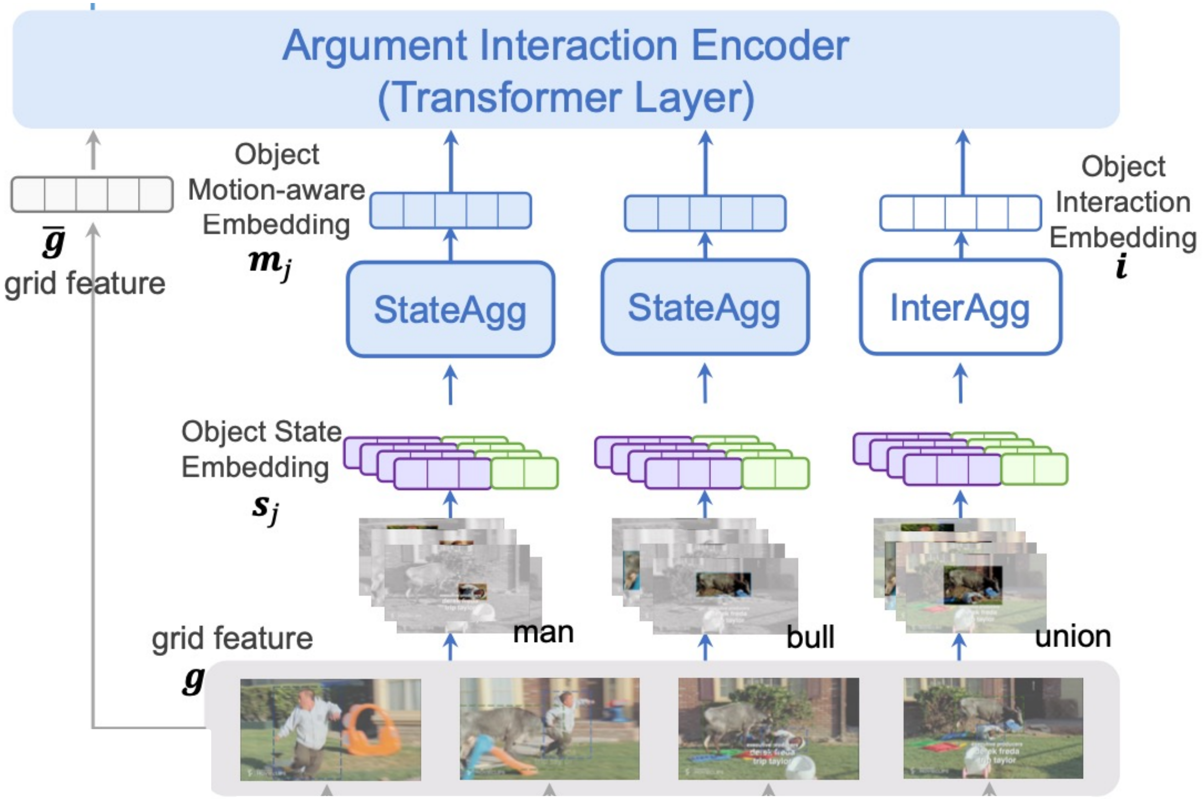


Video Events as Argument State Changes

Video Event =

Status Changes of Arguments

Status Changes of an object =
Displacement (movement of bounding box)
+
Pixel Changes (intra-boundingbox changing)

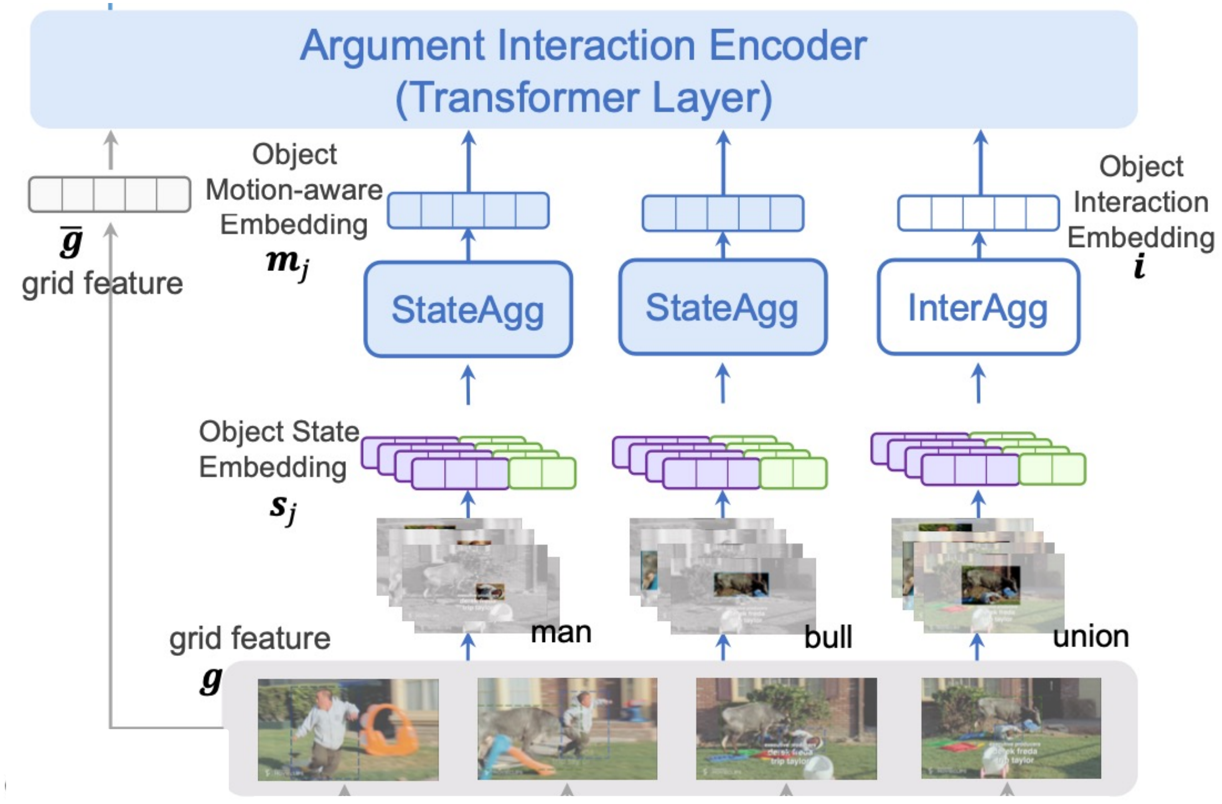


Video Events as Argument State Changes

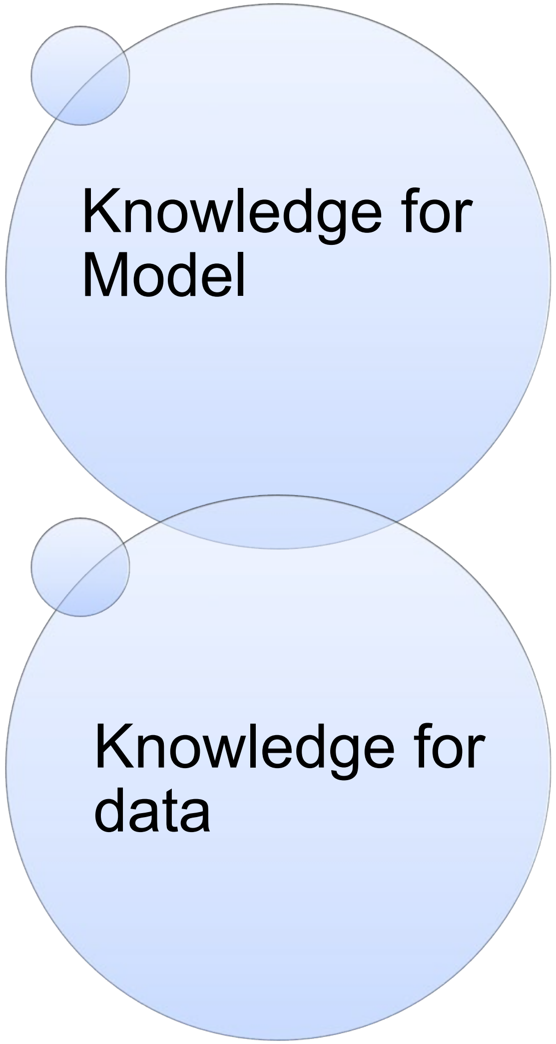
Video Event =

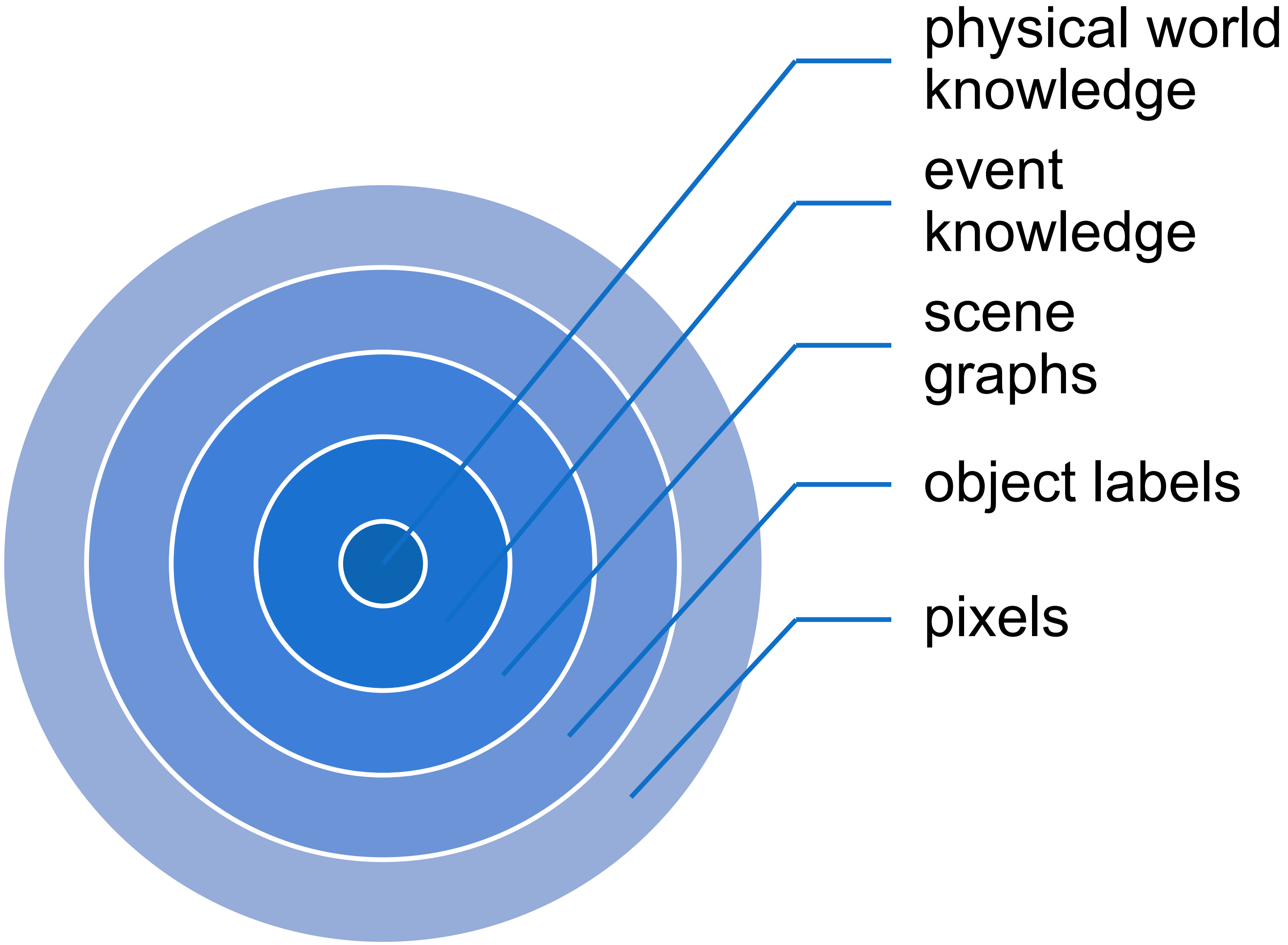
Status Changes of Arguments

Status Changes of an object =
Displacement (movement of bounding box)
+
Pixel Changes (intra-boundingbox changing)



Adding knowledge to pretraining models





What is embodied AI?

Aiming at the creation of an embodied agent (e.g., a robot) which learns, through interaction and exploration, to creatively solve challenging tasks within its environment.

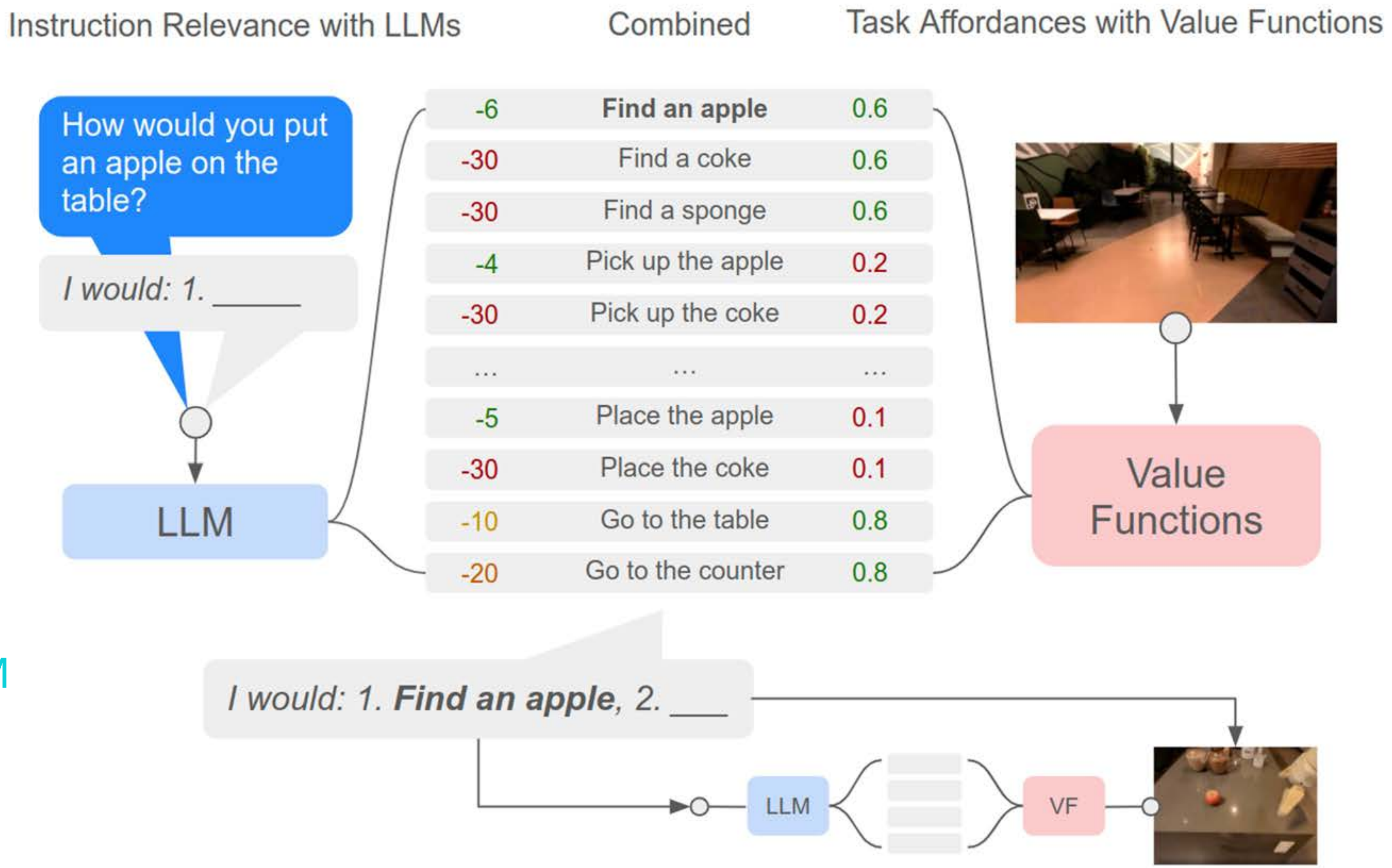
- See: perceive their environment through vision or other senses.
- Talk: hold a natural language dialog grounded in their environment.
- Listen: understand and react to audio input anywhere in a scene.
- Act: navigate and interact with their environment to accomplish goals.
- Reason: consider and plan for the long-term consequences of their actions.

SayCan: Grounding Language in Affordances

SayCan: Grounding Language in Affordances

We need to ground the language model in tasks that are feasible within a specific real-world context.

SayCan: Grounding Language in Affordances



Note that the LM scores options instead of doing generation

Combining SayCan with Chain of Thought Prompting

Chain-of-thought prompting can improve examples with negation.

CHAIN-OF-THOUGHT PLANNING ROLLOUT WITH PALM-SAYCAN.

Human: Can you bring a fruit-flavored drink without caffeine?
Explanation: The user has asked for a drink that is fruit-flavored and does not have caffeine, I will bring the lime soda.
Robot: 1. find a lime soda, 2. pick up the lime soda, 3. bring it to you, 4. put down the lime soda, 5. done

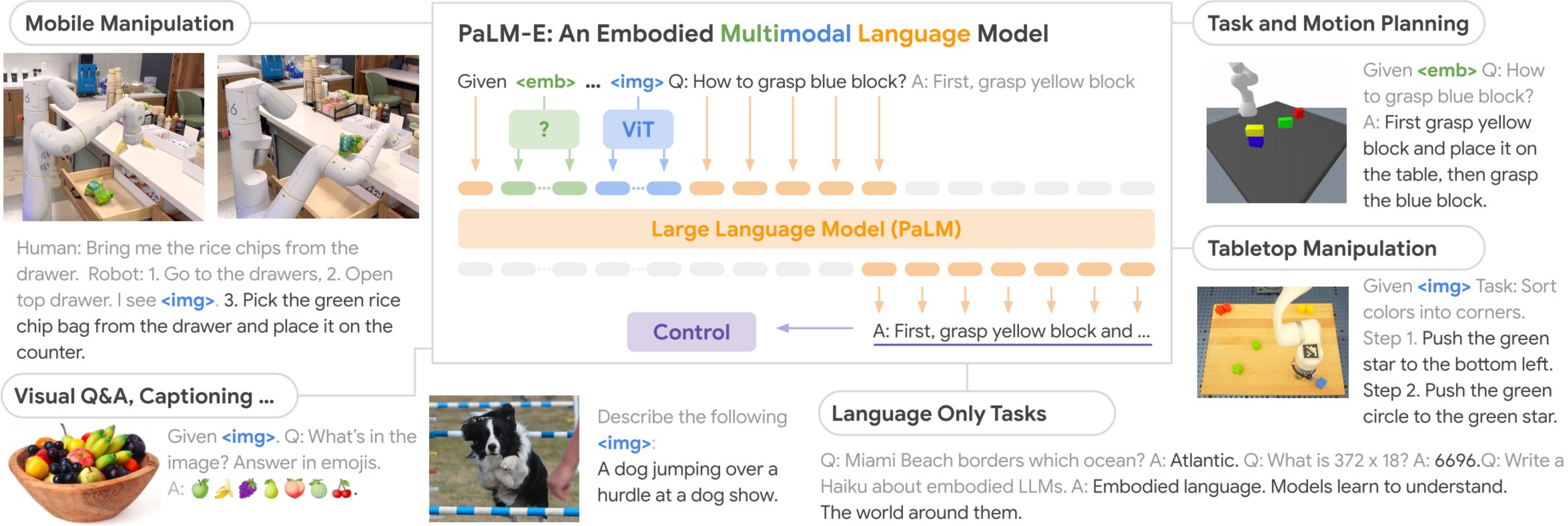
Human: Bring me the spicy chips.
Explanation: The user has asked for chips that are spicy, I will bring the Jalapeno chips.
Robot: 1. find the jalapeno chips, 2. pick up the jalapeno chips, 3. bring it to you, 4. put down the jalapeno chips, 5.done

Human: The snack is too small, bring me something more filling.
Explanation: The user has asked for a snack that is more filling, I will bring the multigrain chips.
Robot: 1. find the multigrain chips, 2. pick up the multigrain chips, 3. bring it to you, 4. put down the multigrain chips, 5. done

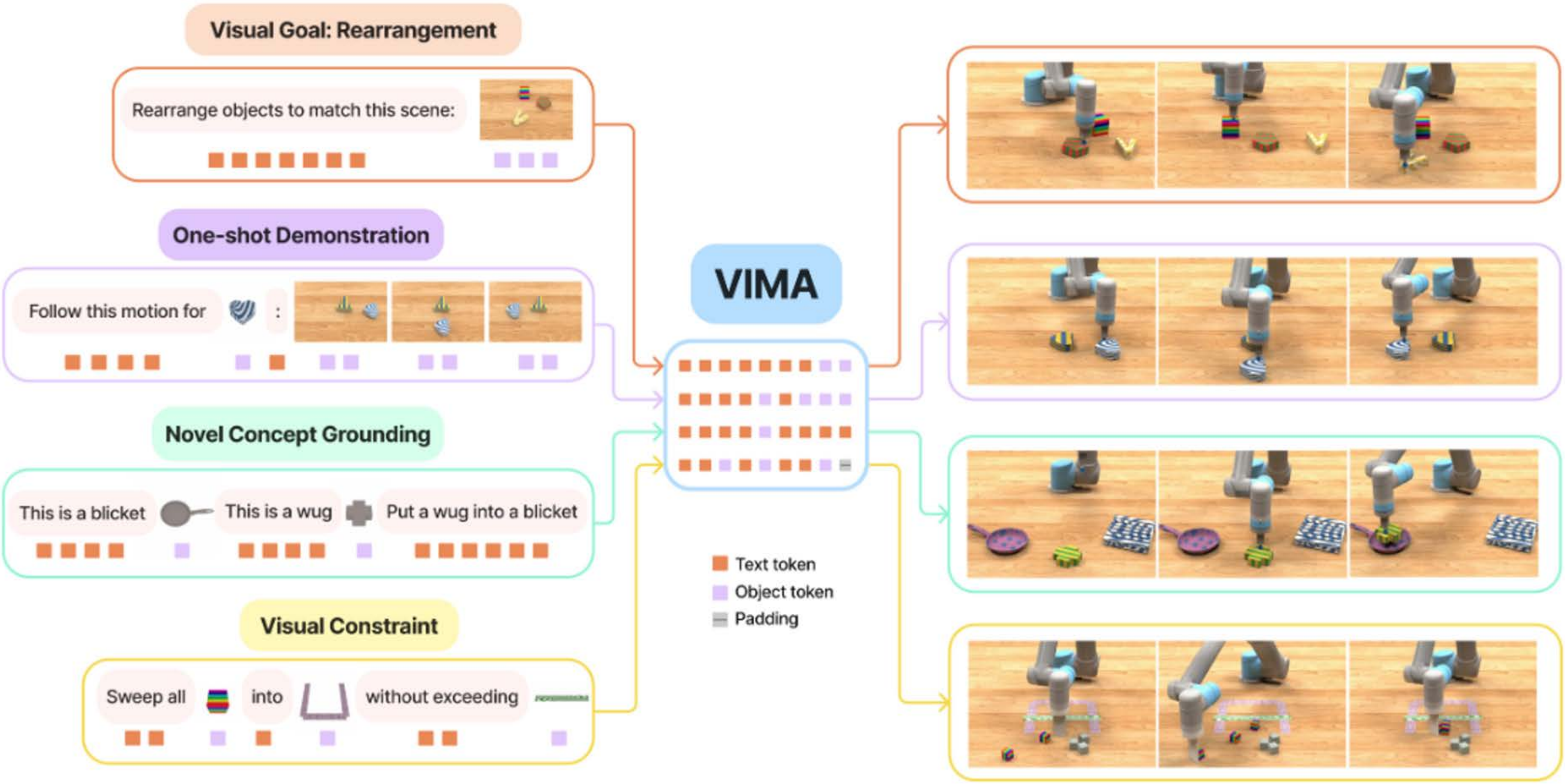
Table 4: Chain-of-thought planning rollout with PaLM-SayCan. The highlighted part is the chain of thought generated by PaLM-SayCan.

PaLM-E: An Embodied Multimodal Language Model

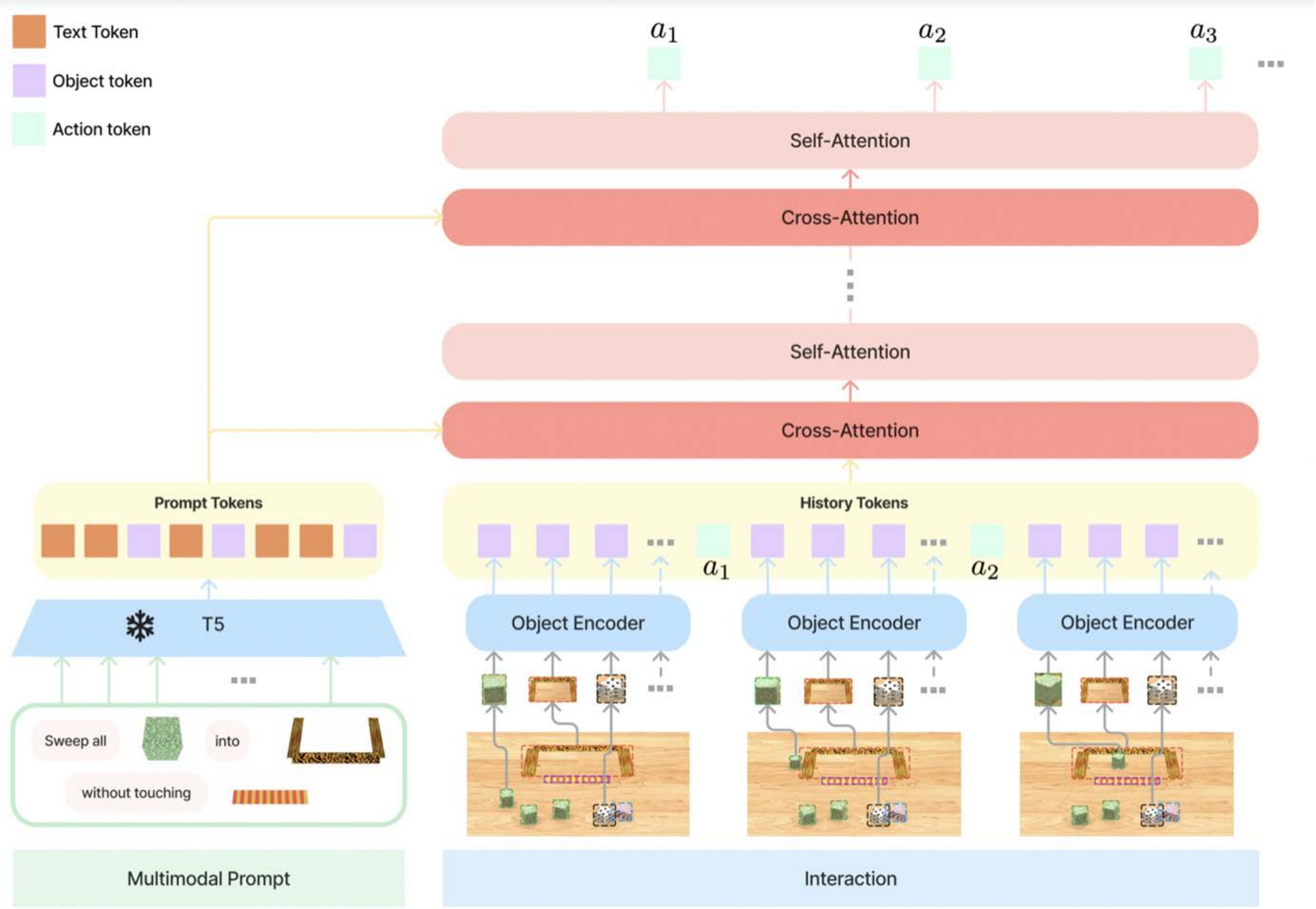
Encoding embodied observations as language tokens.



VIMA: Robot Manipulation with Multimodal Prompts



VIMA: Robot Manipulation with Multimodal Prompts

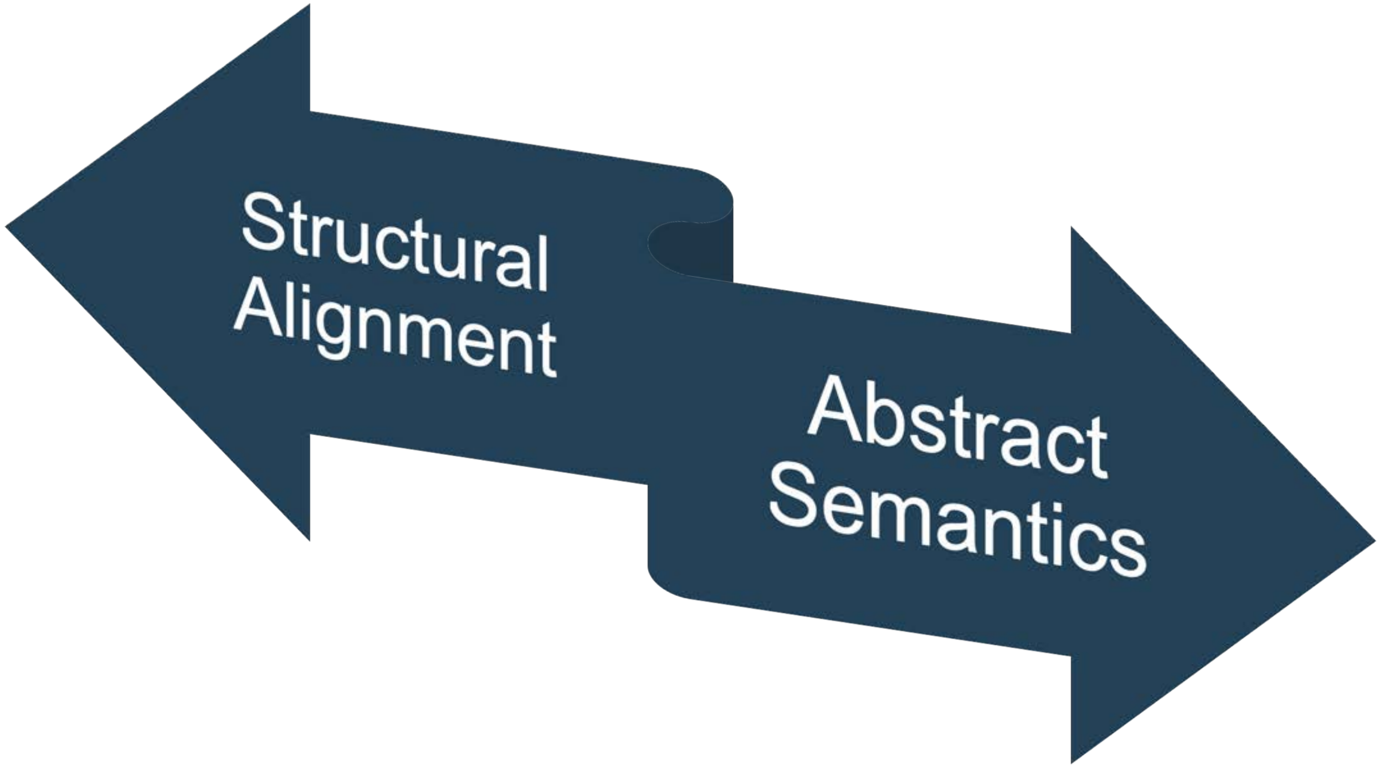


Zero-shot Generalization: The secret ingredient

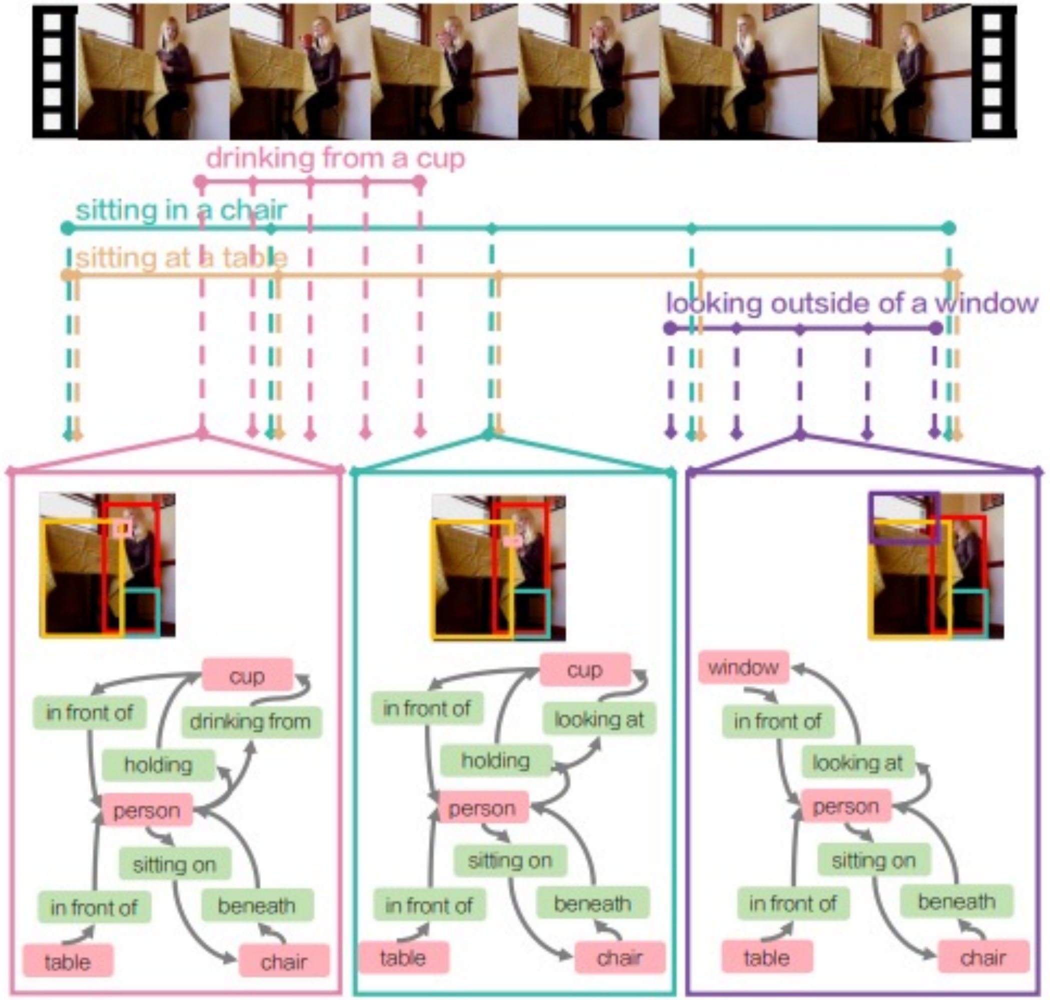
- *We note that this can only be achieved with both cross-attention and object token sequence representation — altering any component will degrade the performance significantly,especially in the low model capacity regime.*
- *The data efficiency can be attributed to VIMA's object-centric representation, which is less prone to overfitting than learning directly from pixels in the low-data regime.*

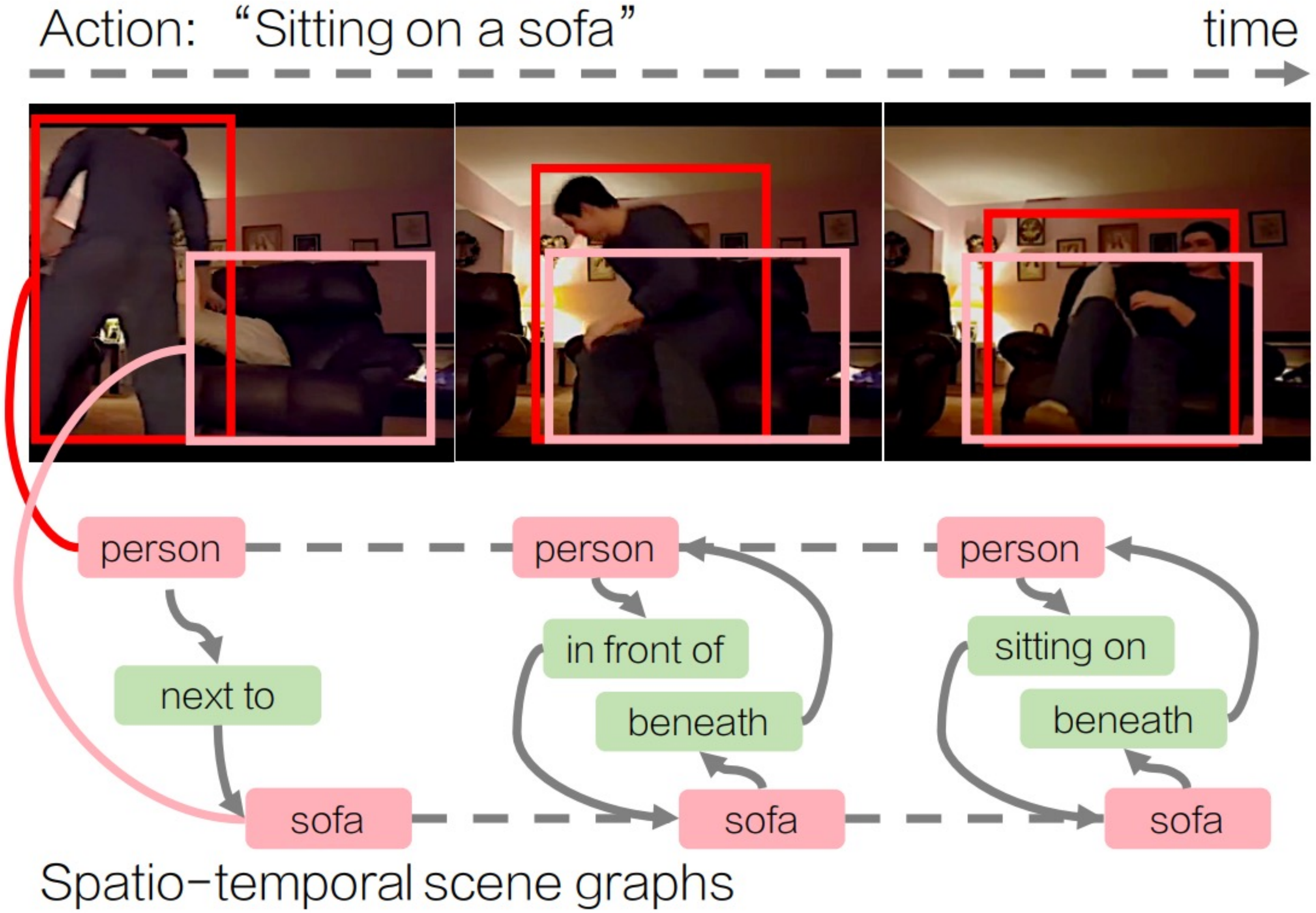
Future Challenges

- Structured: Capturing semantic structure
- Abstract: Understanding abstract and complicated concepts



Future Direction 1: Structure-Aware Encoding





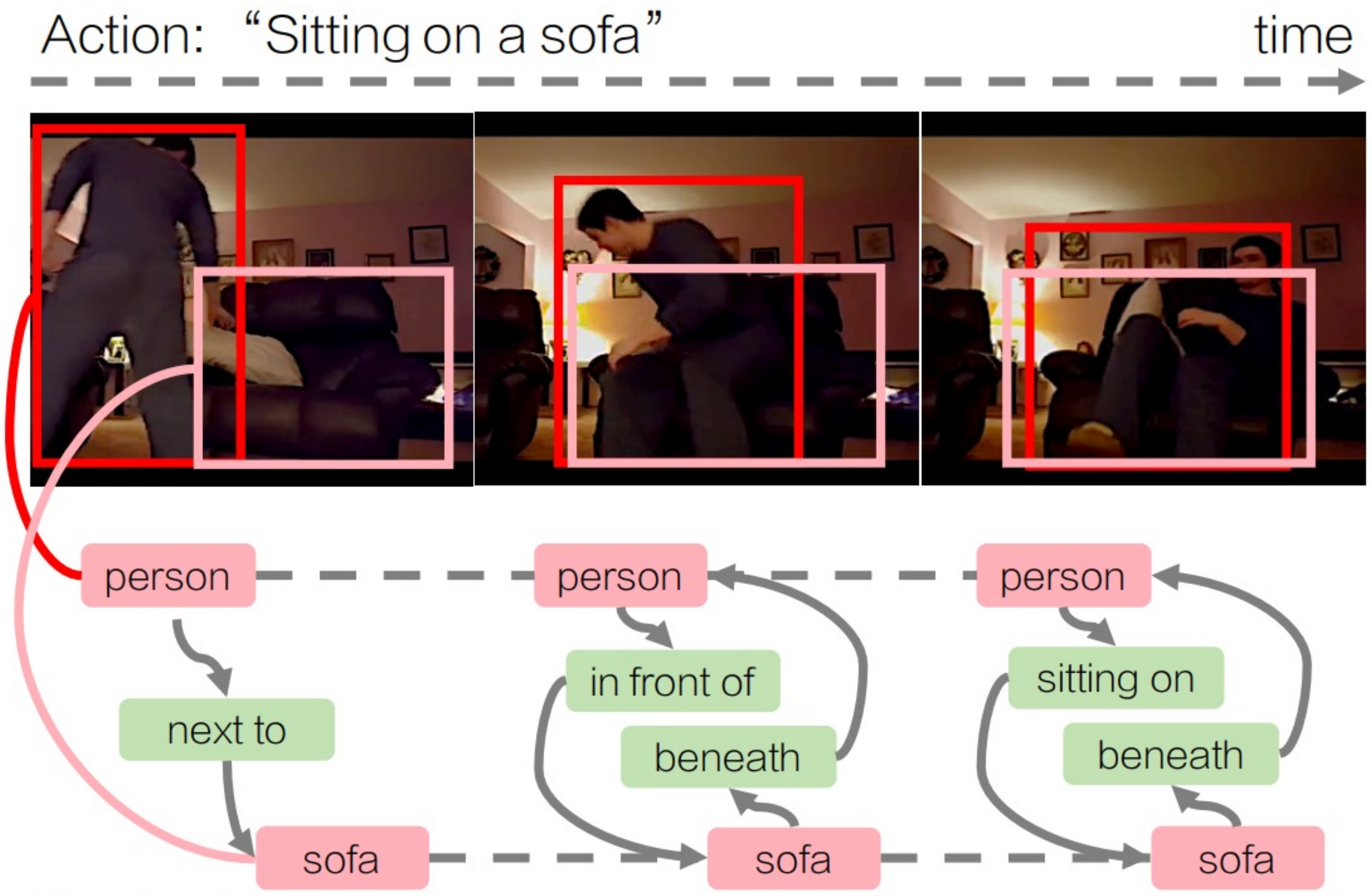
[Ji el al, 2019]

Future Direction 1: Structure-Aware Encoding

Text

Vision

- Better in capturing details with visual features



Spatio-temporal scene graphs

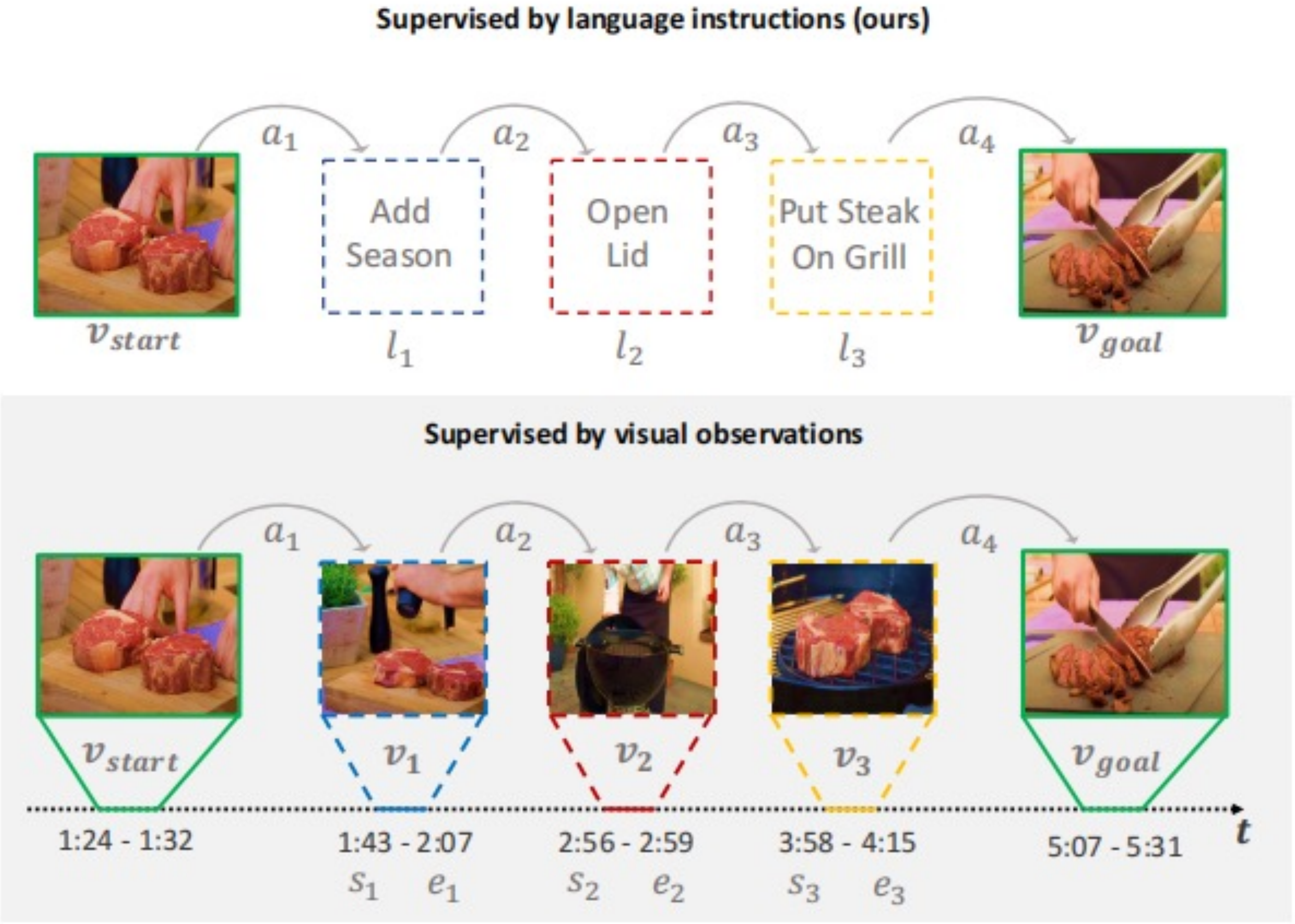
Future Direction 1: Structure-Aware Encoding

Text

- Strong ability in reasoning and semantic structure understanding

Vision

- Better in capturing details with visual features



P3IV: Probabilistic Procedure Planning from Instructional Videoswith Weak Supervision

Future Direction 2: Abstract Semantics

Deep Semantic Understanding:

Discover knowledge (important information) that humans are actively seeking or communicating.

Future Direction 2: Abstract Semantics

Text generation paradigm (e.g., GPT-3) is taking over the NLP world.

But it is flat and surface-to-surface.

Bounded Knowledge | **Short Context** | **Surface-to-Surface**

Future Direction 2: Abstract Semantics

Text generation paradigm (e.g., GPT-3) is taking over the NLP world.

But it is flat and surface-to-surface.

Bounded Knowledge

Short Context

Surface-to-Surface

Surface → Deep

Concrete → Abstract

Static → Dynamic

Perception → Cognition

Future Direction 2: Abstract Semantics

Text generation paradigm (e.g., GPT-3) is taking over the NLP world.

But it is flat and surface-to-surface.

Bounded Knowledge

Short Context

Surface-to-Surface

Compositional

Novel Concept

\+

Long

Horizon

Surface → Deep

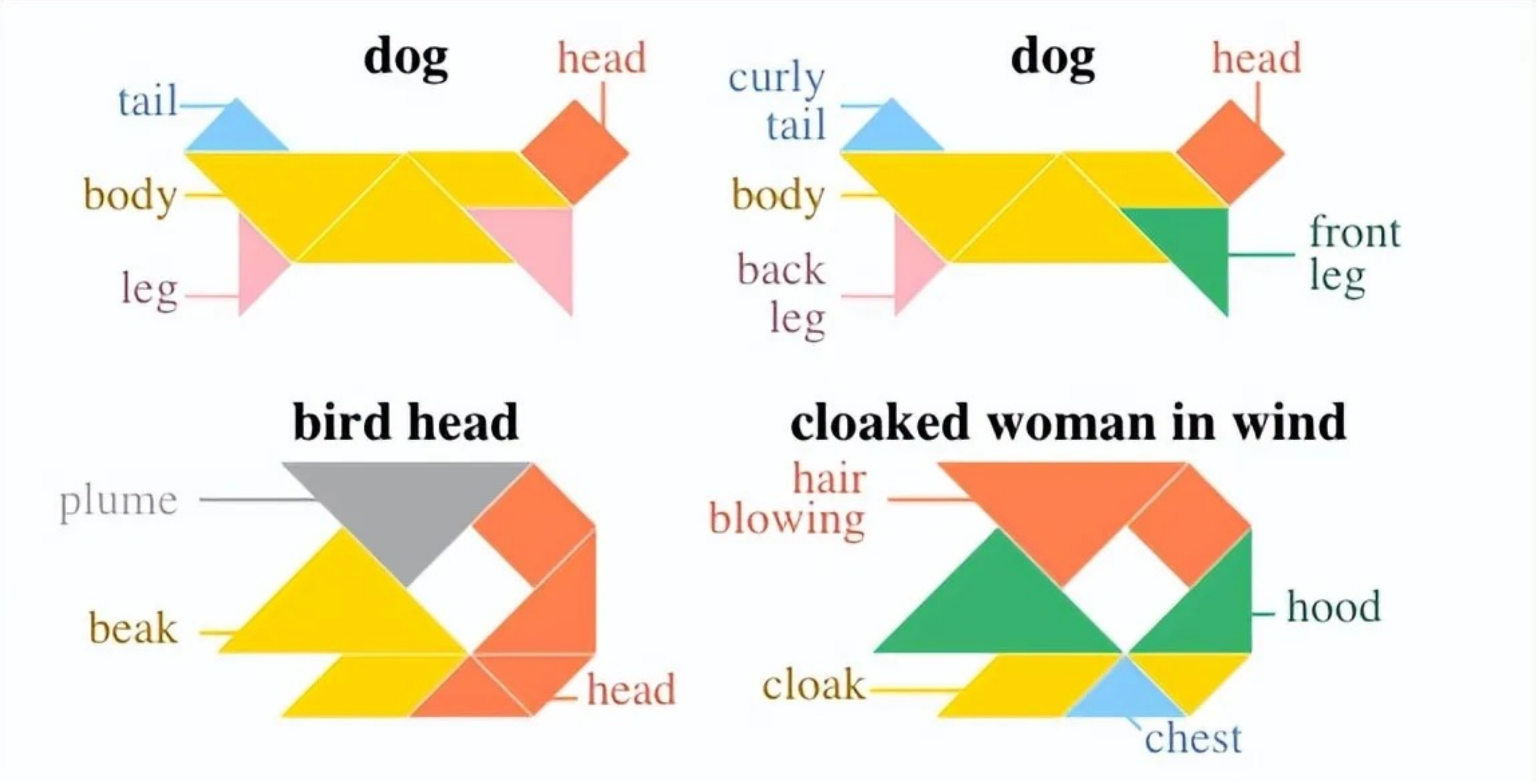
Concrete → Abstract

Static → Dynamic

Perception → Cognition

Future Direction 2: Abstract Semantics

Abstract



Abstract Visual Reasoning with Tangram Shapes (EMNLP 2022 Best Paper)

Future Direction 2: Abstract Semantics

Abstract

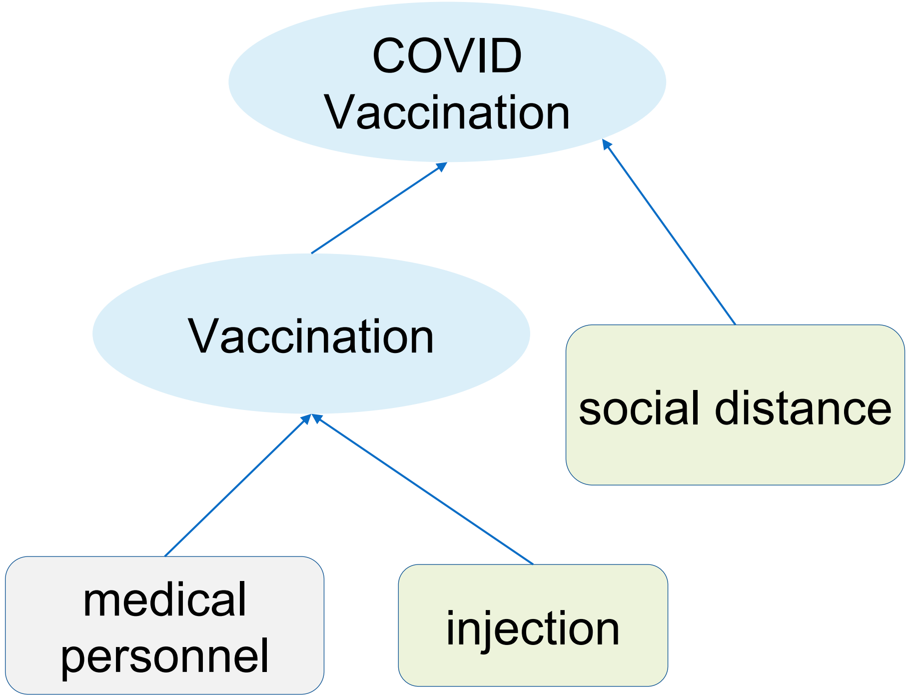
Love

Happiness

Emotion <-> Music

Future Direction 2: Compositional Semantics

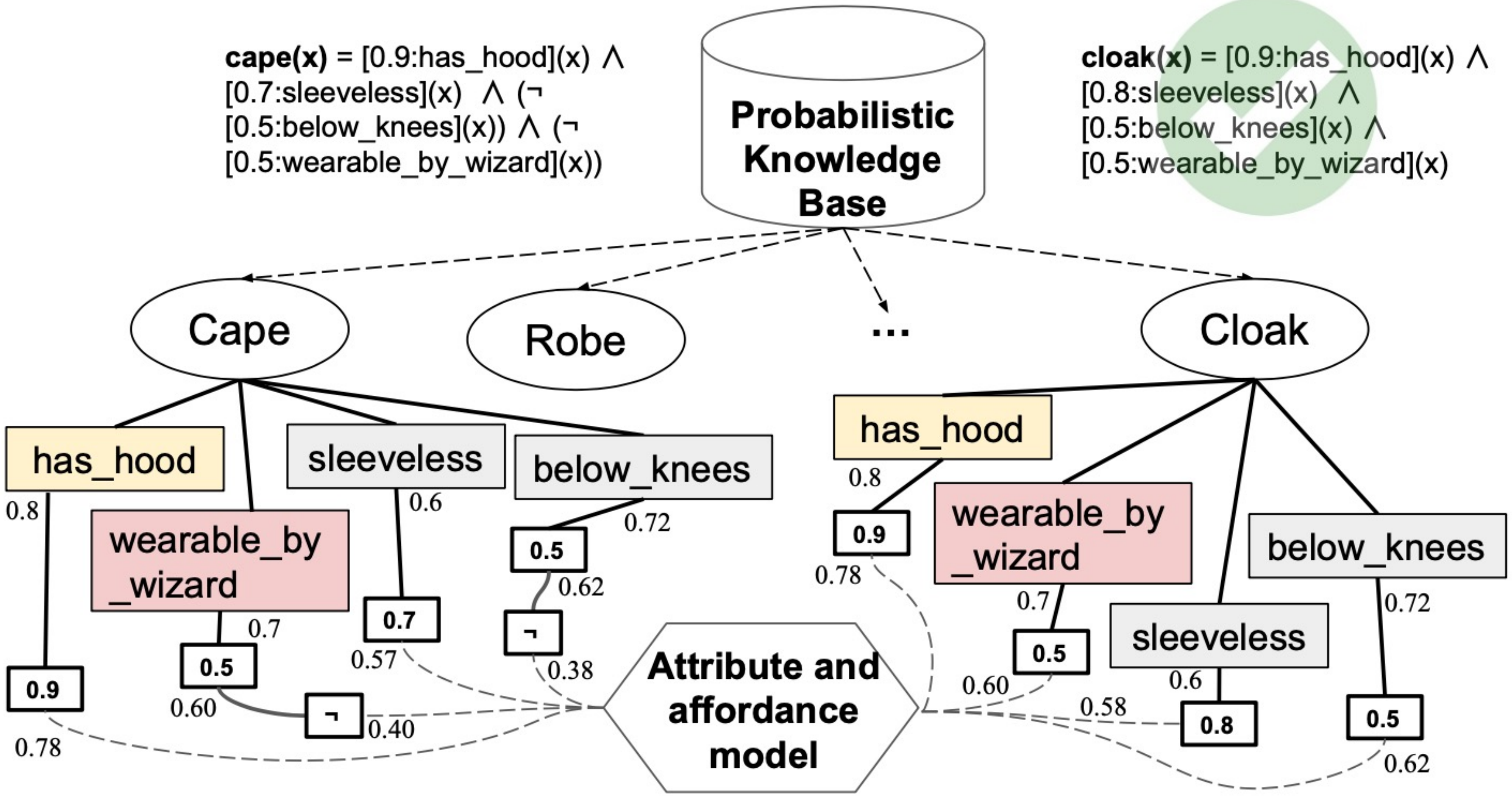
Compositional



Future Direction 2: Abstract Semantics

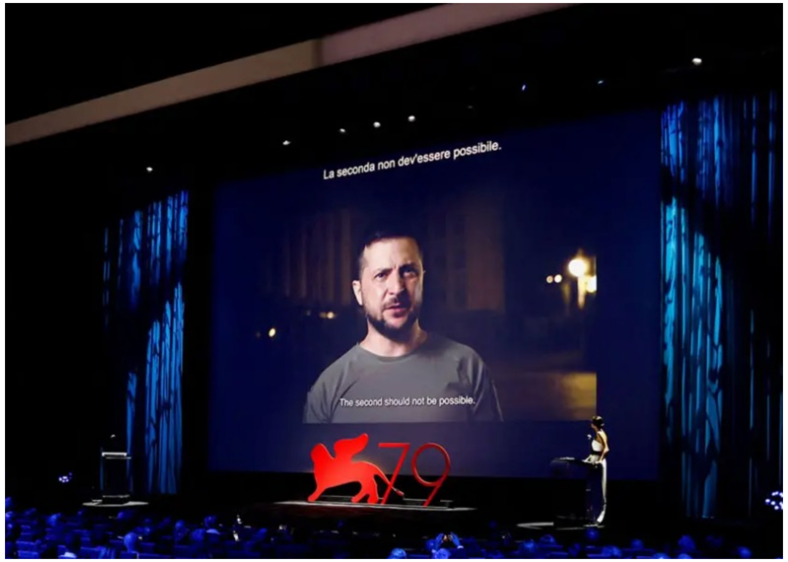
Reasoning & Explanability

Image and object bounding box



Future Direction 2: Abstract Semantics

Alternative Interpretation & Propaganda



New York Post — Lean Right

Zelensky describes Ukraine war as 'horror film' at Venice Film Festival opening

Scroll India — Lean Left

Ukrainian President addresses Venice Film Festival, calls war 'drama based on real-life'

- Information Surgery
- Video Framing
- Propagation Path
- Propaganda Detection
- Misinformation Detection
- Reframing

Using LLMs to help

LLMs have **good compositional ability** from low-level observations to high level concepts.

Z

We observe the following features in an image: "Dog standing, looking back," "Man in camouflage uniform holding baby in warm clothes," "Rubble and overturned vehicle in background," "Outdoor scene with water, mud, and man-made objects".

---

what is the activity happening in the image?

Based on the provided features, it is difficult to determine the exact activity happening in the image. However, we can make some observations and assumptions:

- There is a dog standing and looking back, suggesting that the dog may be accompanying the man and baby.
- The man is wearing a camouflage uniform and holding a baby in warm clothes, which suggests that he may be a military or security personnel.
- There is rubble and an overturned vehicle in the background, suggesting that there may have been some kind of conflict or disaster in the area.
- The outdoor scene includes water, mud, and man-made objects, which could suggest that the area is a construction site, a disaster zone, or a military training ground.

Based on these observations, it is possible that the man is a military or security personnel who is holding the baby while on duty in a conflict or disaster zone. The dog may be accompanying them on their mission. However, without more context, it is impossible to determine the exact activity happening in the image.

Other Open Questions

- Evaluation:
  - Is the performance boost because of adding knowledge?
- Explanability and Reasoning
  - How can large models leverage knowledge?
- Continuous learning for large models
  - How can large models continue learning more complicated concepts?
- Bias in learning knowledge
  - Debiasing with knowledge guidance

Jun 2023
CVPR Tutorials
Knowledge-Driven Vision-Language Encoding

CVPR

Commonsense Knowledge in V+L Pretraining

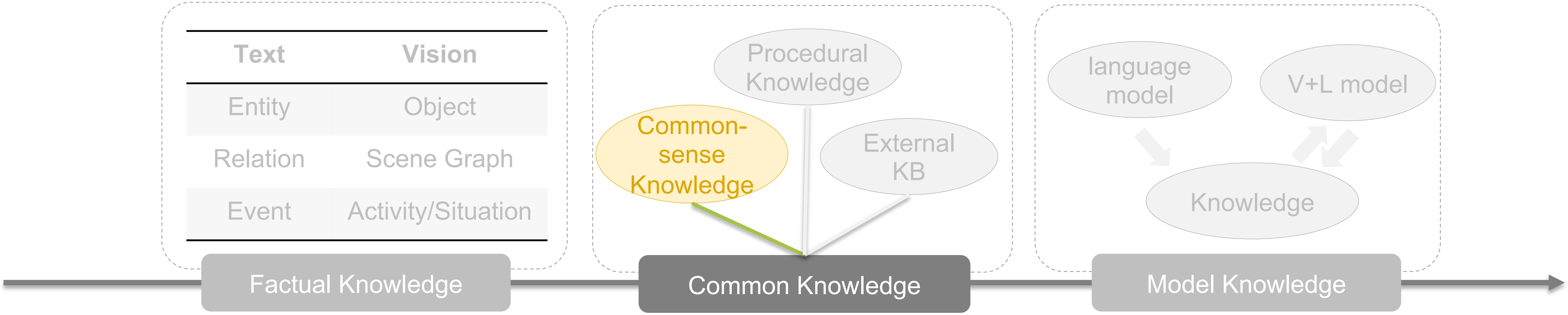
Knowledge-Driven Vision-Language Pretraining (Part III)

Manling Li
UIUC
manling2@illinois.edu

University of Illinois Urbana-Champaign
Northwestern University
Columbia University
Meta AI

Commonsense Knowledge

Commonsense Knowledge is the basic facts and behaviors of the everyday world.



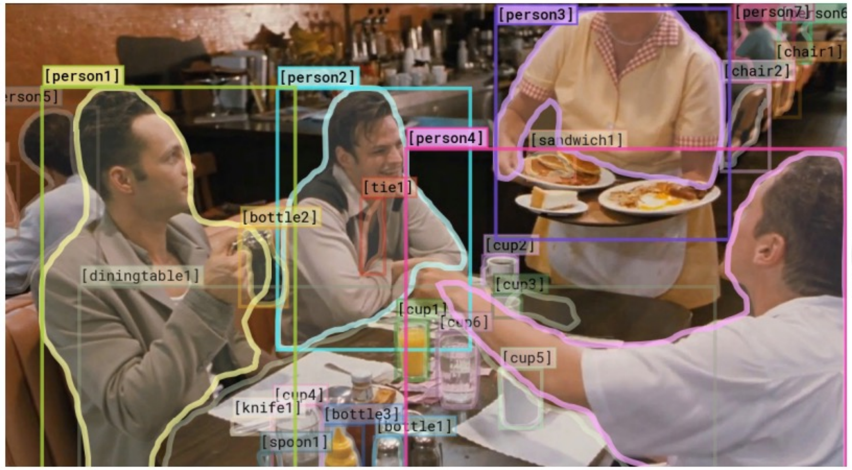
Outline

- What is Commonsense Knowledge
  - VCR
  - VisualCOMET
  - Physical Knowledge
- Knowledge → VLM
  - Knowledge Graph Riddles
  - Knowledge Graph Embedding
- VLM → Knowledge
  - Physical Knowledge
  - Unimodal vs Multimodal

Part 1: What is Visual Commonsense Knowledge?

Visual Commonsense Knowledge

Visual Commonsense Reasoning (VCR): From Recognition to Cognition

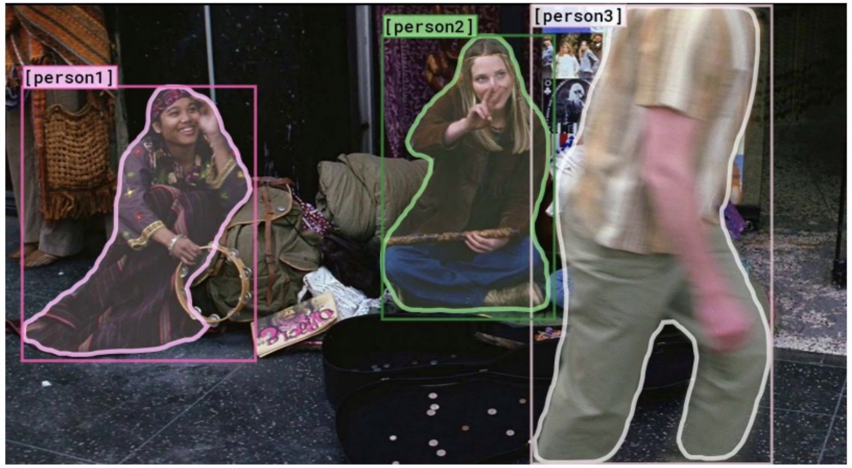


Why is [person4] pointing at [person1]?

a) **He is telling [person3] that [person1] ordered the pancakes.**
b) He just told a joke.
c) He is feeling accusatory towards [person1].
d) He is giving [person1] directions.

I chose a) because…

a) [person1] has the pancakes in front of him.
b) [person4] is taking everyone's order and asked for clarification.
c) [person3] is looking at the pancakes and both she and [person2] are smiling slightly.
d) **[person3] is delivering food to the table, and she might not know whose order is whose.**



How did [person2] get the money that's in front of her?

a) [person2] is selling things on the street.
b) **[person2] earned this money playing music.**
c) She may work jobs for the mafia.
d) She won money playing poker.

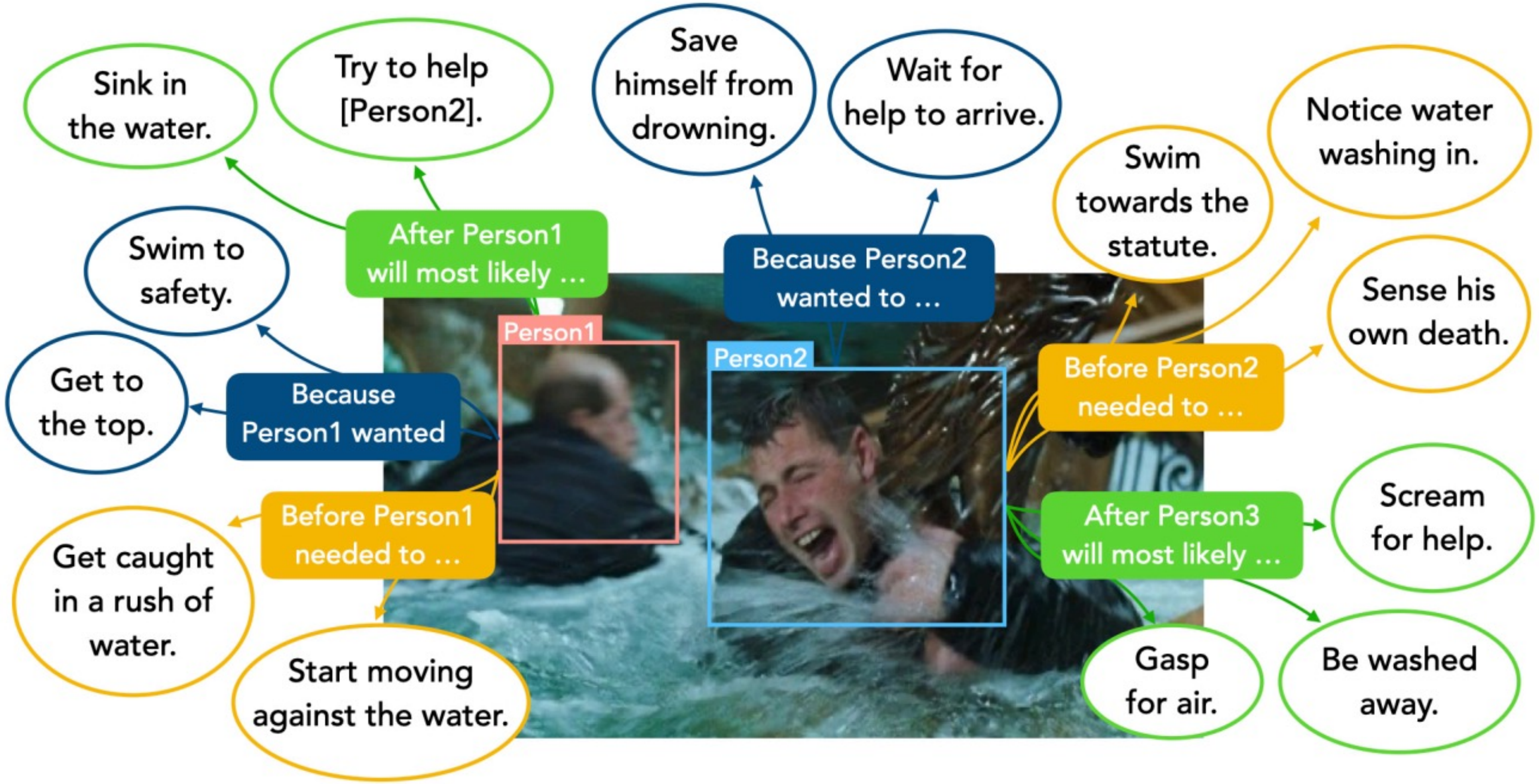
I chose b) because…

a) She is playing guitar for money.
b) [person2] is a professional musician in an orchestra.
c) **[person2] and [person1] are both holding instruments, and were probably busking for that money.**
d) [person1] is putting money in [person2]'s tip jar, while she plays music.

From Recognition to Cognition: Visual Commonsense Reasoning. 2019

Visual Commonsense Knowledge

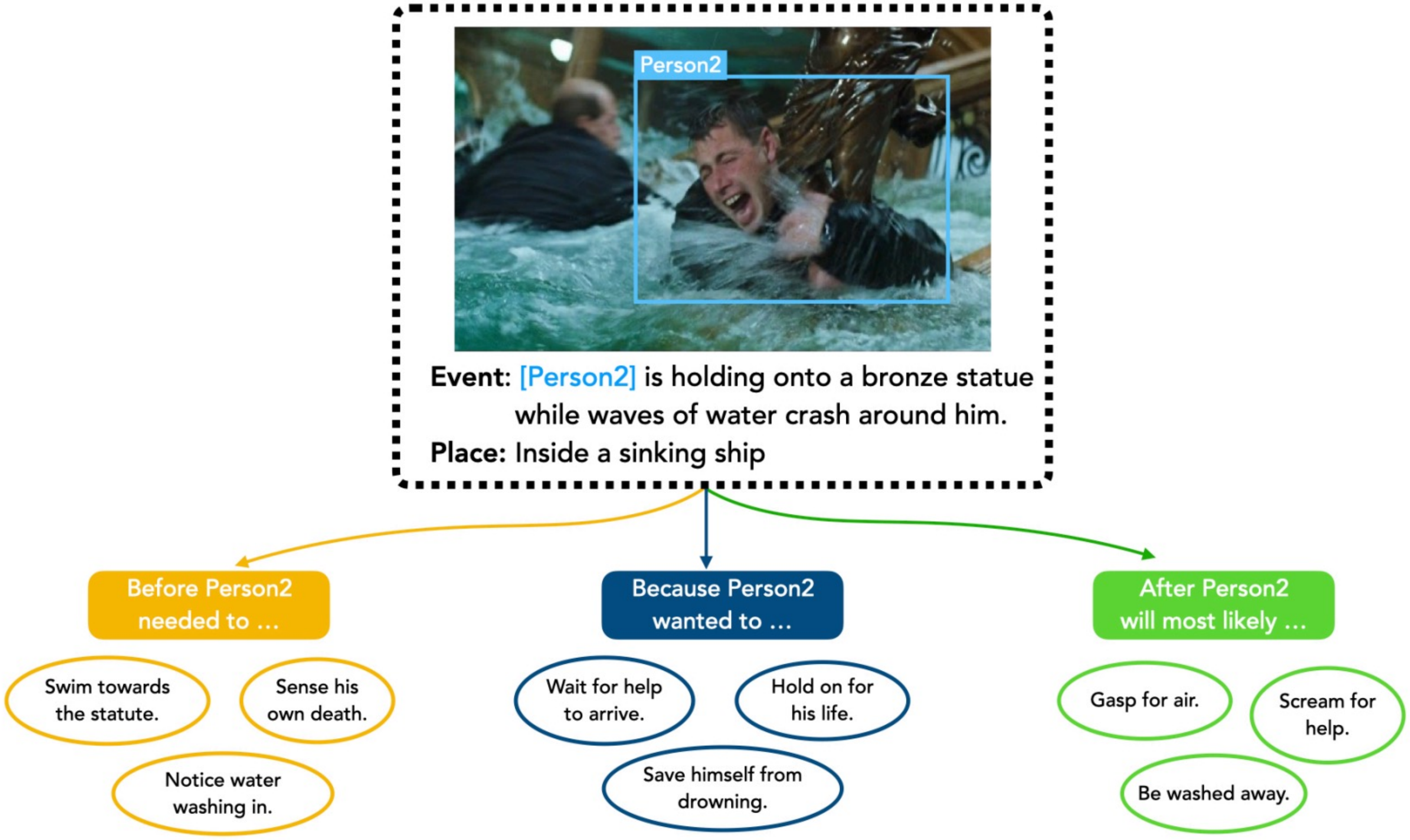
VisualCOMET: Cognitive Image Understanding via Visual Commonsense Graphs



VisualCOMET: Reasoning About the Dynamic Context of a Still Image https://mosaickg.apps.allenai.org/visual_comet

Visual Commonsense Knowledge

VisualCOMET Task Formulation: Generate the entire visual commonsense graph



Visual Commonsense Knowledge

Large Dataset Collection: There are in total 139,377 distinct Visual Commonsense Graphs over 59,356 images involving 1,465,704 commonsense inferences.

	Train	Dev	Test	Total
# Images/Places	47,595	5,973	5,968	**59,356**
# Events at Present	111,796	13,768	13,813	**139,377**
# Inferences on Events Before	467,025	58,773	58,413	584,211
# Inferences on Events After	469,430	58,665	58,323	586,418
# Inferences on Intents at Present	237,608	28,904	28,568	295,080
# Total Inferences	1,174,063	146,332	145,309	**1,465,704**

Physical Commonsense Knowledge

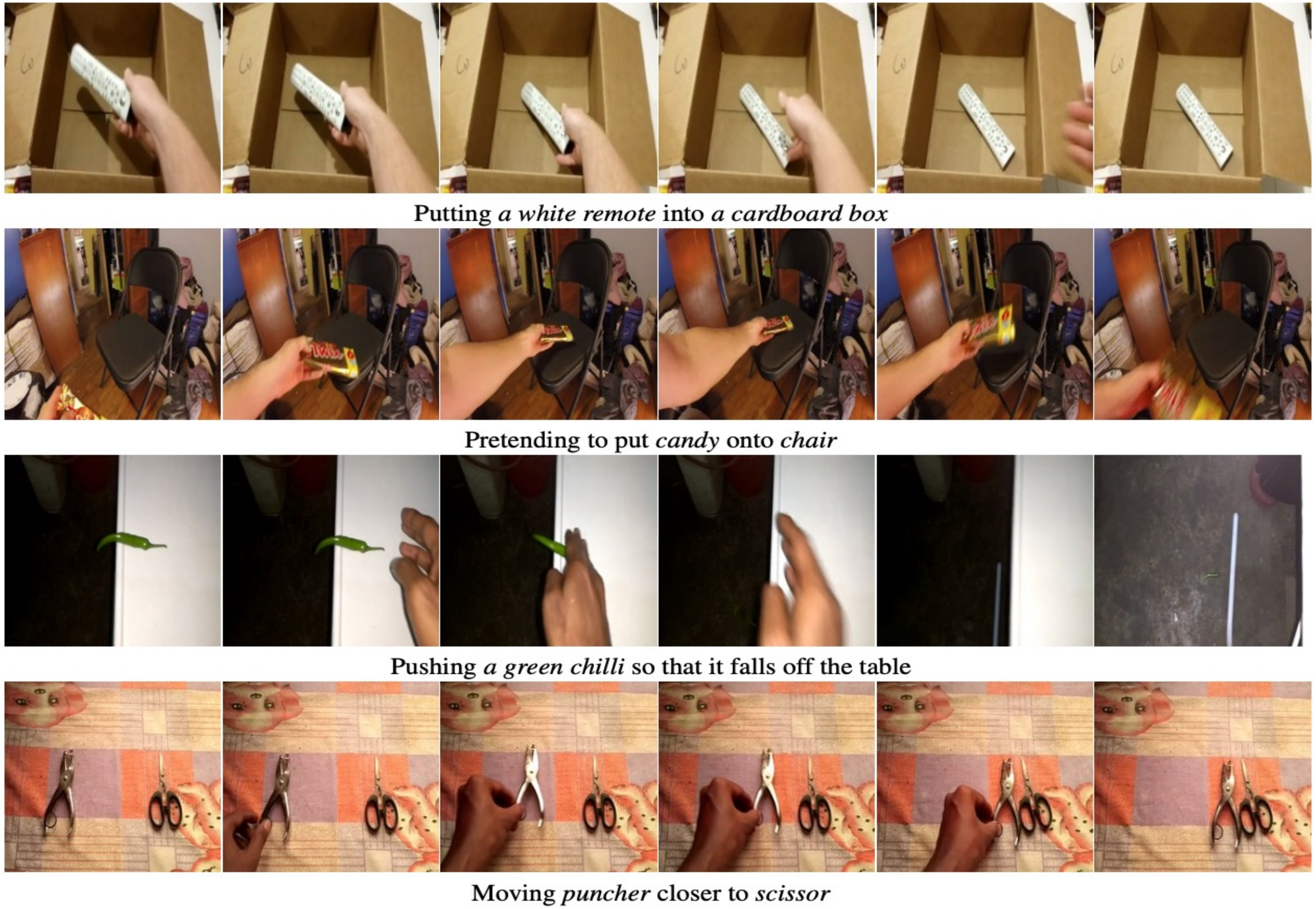
Physical Commonsense Knowledge can be learned via natural language.

a. Shape, Material, and Purpose		
[Goal]	Make an outdoor pillow	
[Sol1]	Blow into a tin can and tie with rubber band	✘
[Sol2]	Blow into a trash bag and tie with rubber band	✔
[Goal]	To make a hard shelled taco,	
[Sol1]	put seasoned beef, cheese, and lettuce onto the hard shell.	✘
[Sol2]	put seasoned beef, cheese, and lettuce into the hard shell.	✔
[Goal]	How do I find something I lost on the carpet?	
[Sol1]	Put a solid seal on the end of your vacuum and turn it on.	✘
[Sol2]	Put a hair net on the end of your vacuum and turn it on.	✔

b. Commonsense Convenience		
[Goal]	How to make sure all the clocks in the house are set accurately?	
[Sol1]	Get a solar clock for a reference and place it just outside a window that gets lots of sun. Use a system of call and response once a month, having one person stationed at the solar clock who yells out the correct time and have another person move to each of the indoor clocks to check if they are showing the right time. Adjust as necessary.	✘
[Sol2]	Replace all wind-ups with digital clocks. That way, you set them once, and that's it. Check the batteries once a year or if you notice anything looks a little off.	✔

PIQA: Reasoning about Physical Commonsense in Natural Language. AAAI 2020

The “Something Something” Dataset

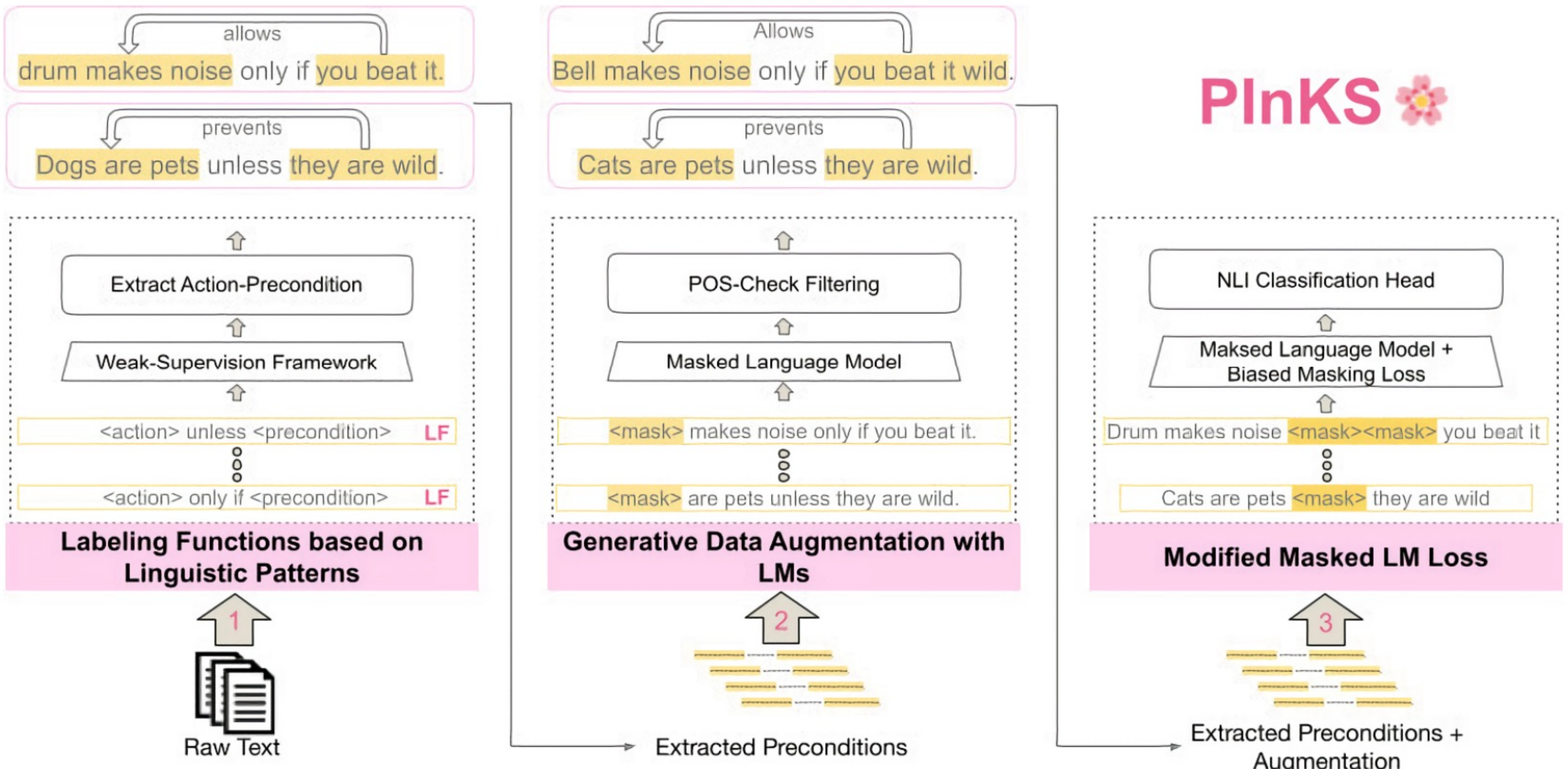


10 selected classes
Dropping [something]
Moving [something] from right to left
Moving [something] from left to right
Picking [something] up
Putting [something]
Poking [something]
Tearing [something]
Pouring [something]
Holding [something]
Showing [something] (almost no hand)

The “something something” video database for learning and evaluating visual common sense

PInKS: Preconditioned Commonsense Inference

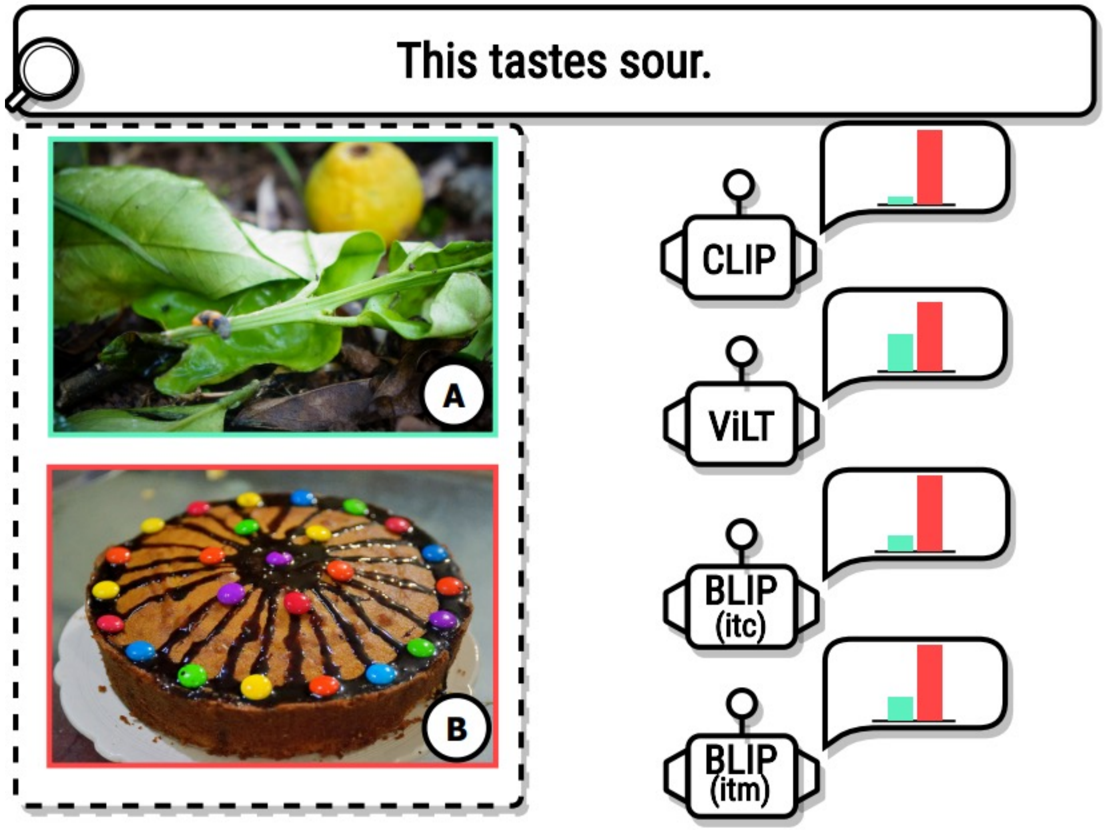
Text	Label	Action	Precondition
A drum makes noise only if you beat it.	Allow	A drum makes noise	you beat it.
Your feet might come into contact with something if it is on the floor.	Allow	Your feet might come into contact with something	it is on the floor.
Pears will rot if not refrigerated	Prevent	Pears will rot	refrigerated
Swimming pools have cold water in the winter unless they are heated.	Prevent	Swimming pools have cold water in the winter	they are heated.



Part 2: How can commonsense knowledge be learned via V+L pretraining?

Failures

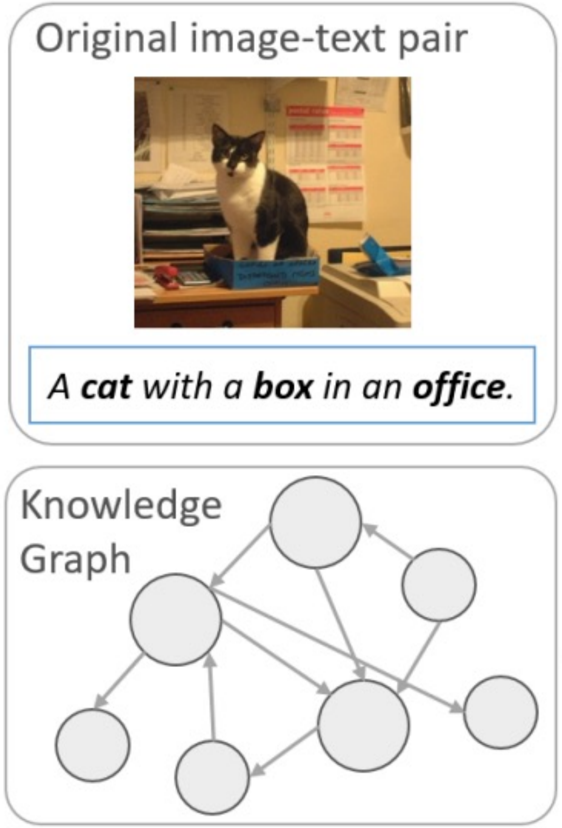
Current V+L models lack abilities to capture commonsense knowledge:



Improving Commonsense in Vision-Language Models via Knowledge Graph Riddles. 2022

DANCE: Improving Commonsense in Vision-Language Models

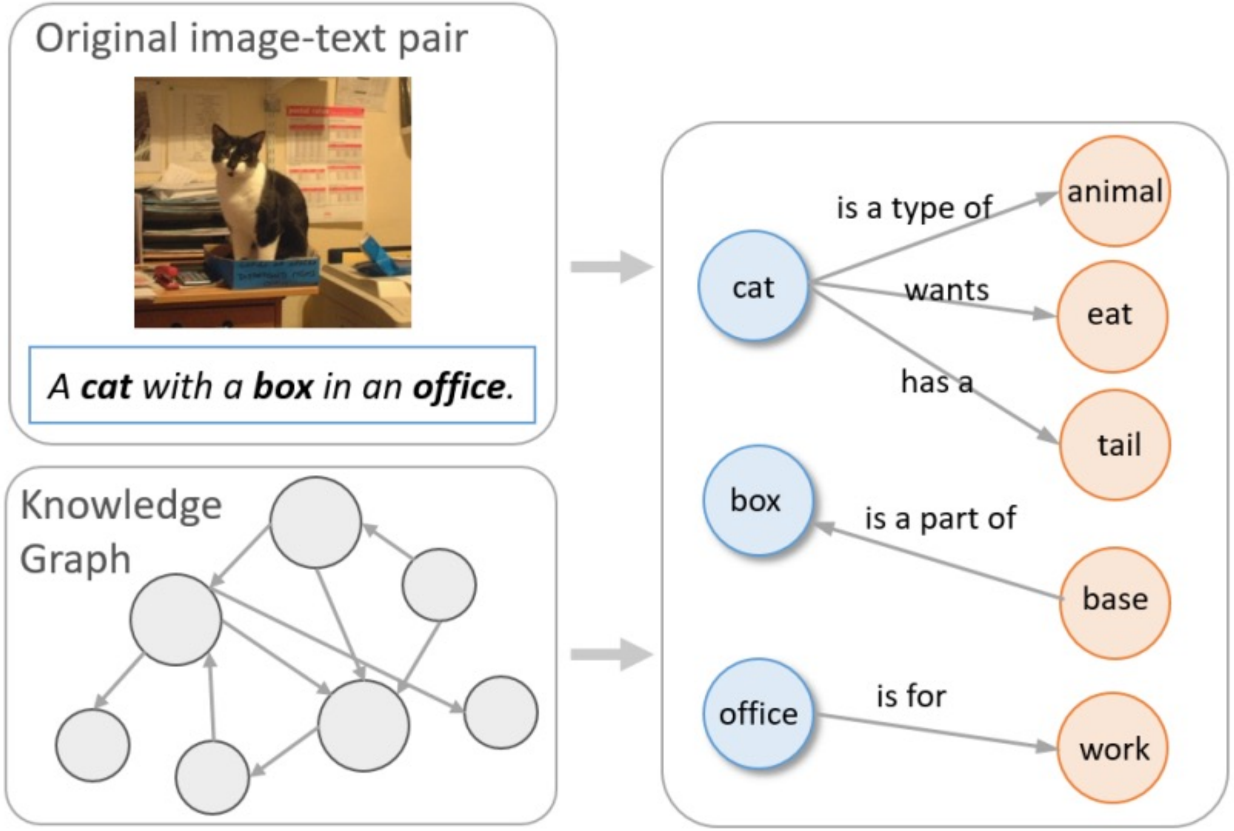
DANCE: Data Augmentation with kNowledge graph linearization for CommonsensE capability



Improving Commonsense in Vision-Language Models via Knowledge Graph Riddles. 2022

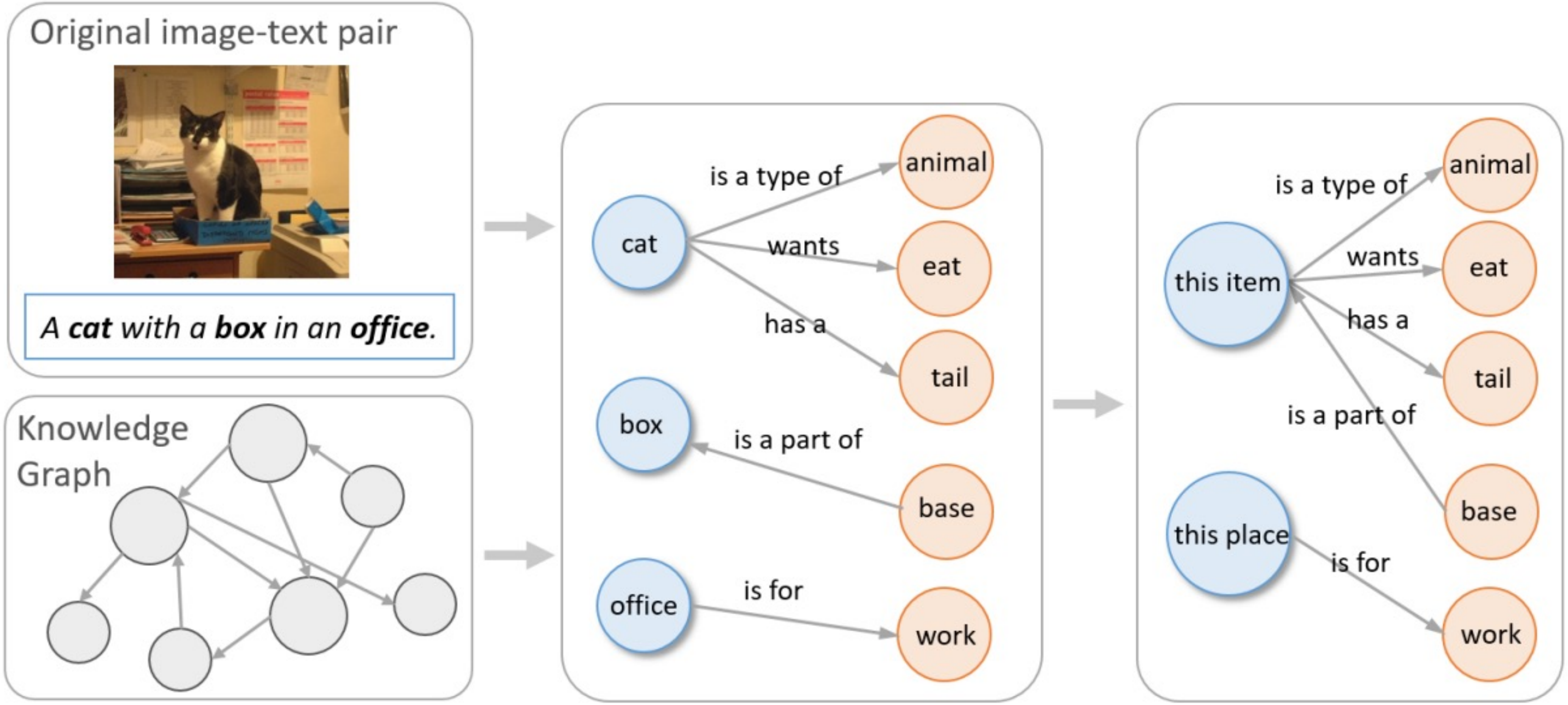
DANCE: Improving Commonsense in Vision-Language Models

DANCE: Data Augmentation with kNowledge graph linearization for CommonsensE capability



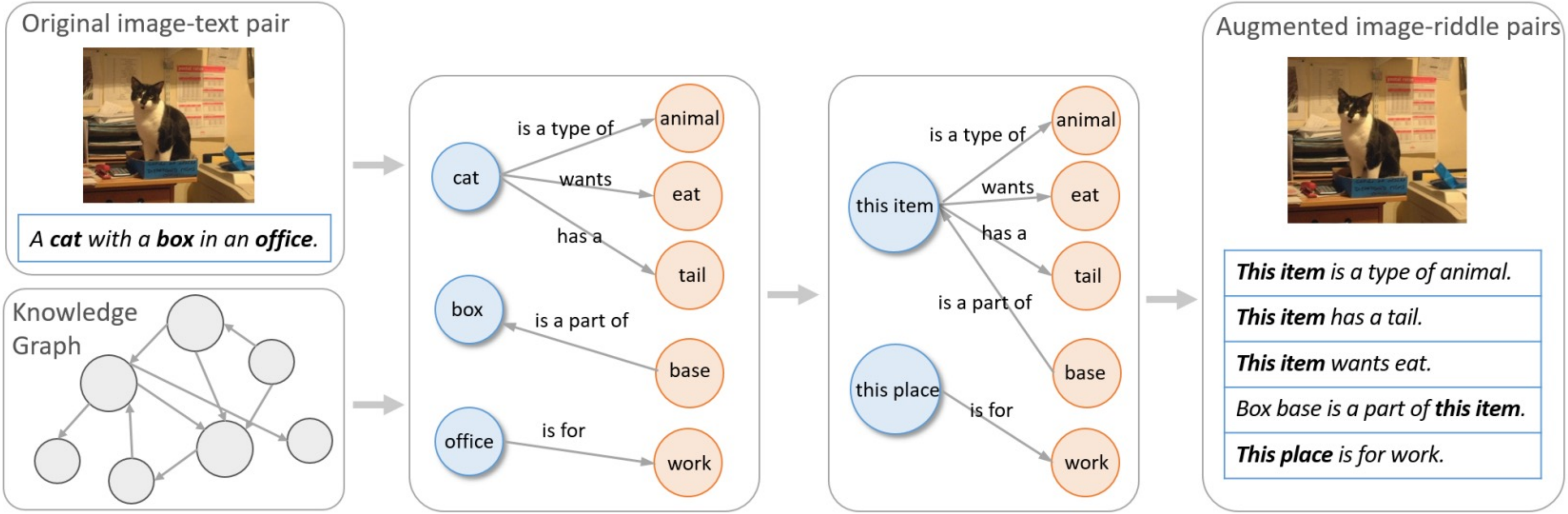
DANCE: Improving Commonsense in Vision-Language Models

DANCE: Data Augmentation with kNowledge graph linearization for CommonsensE capability

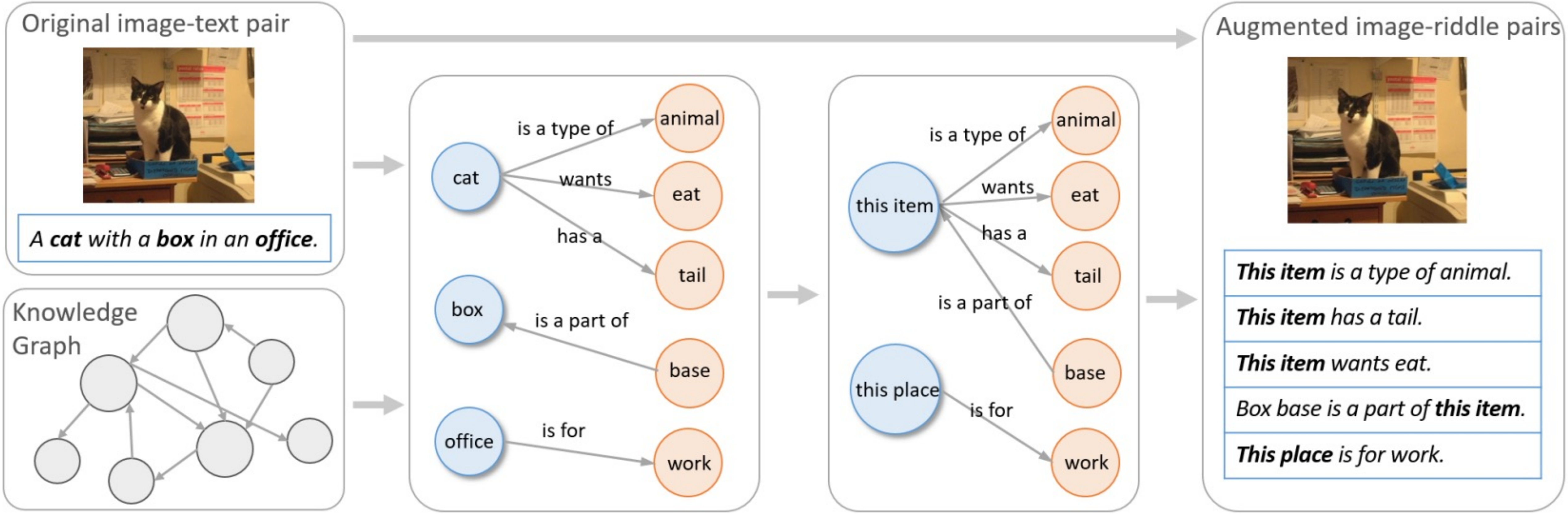


DANCE: Improving Commonsense in Vision-Language Models

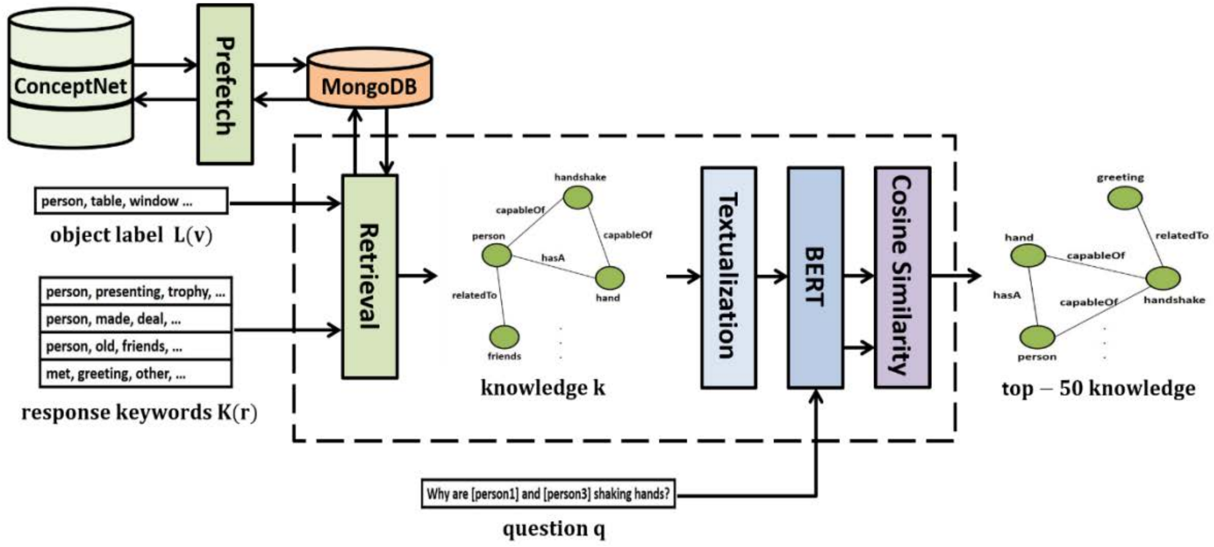
DANCE: Data Augmentation with kNowledge graph linearization for CommonsensE capability

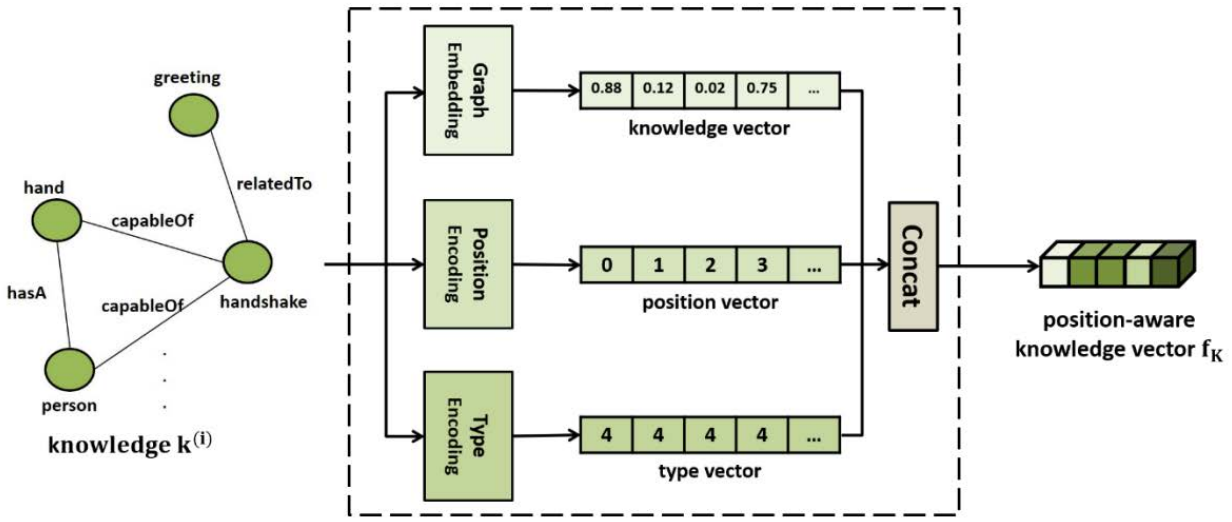


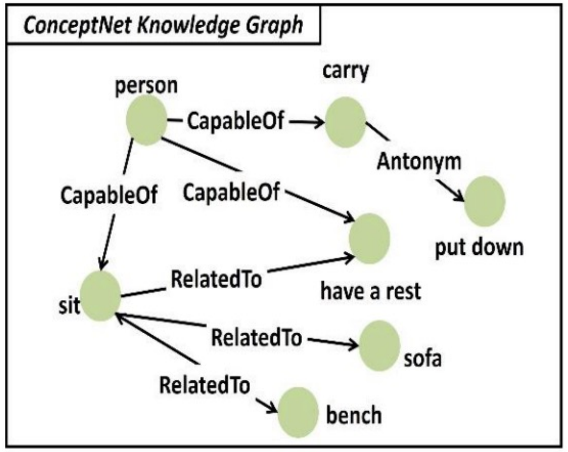
DANCE: Data Augmentation with kNowledge graph linearization for CommonsensE capability



Vision–Language Knowledge Co-Embedding







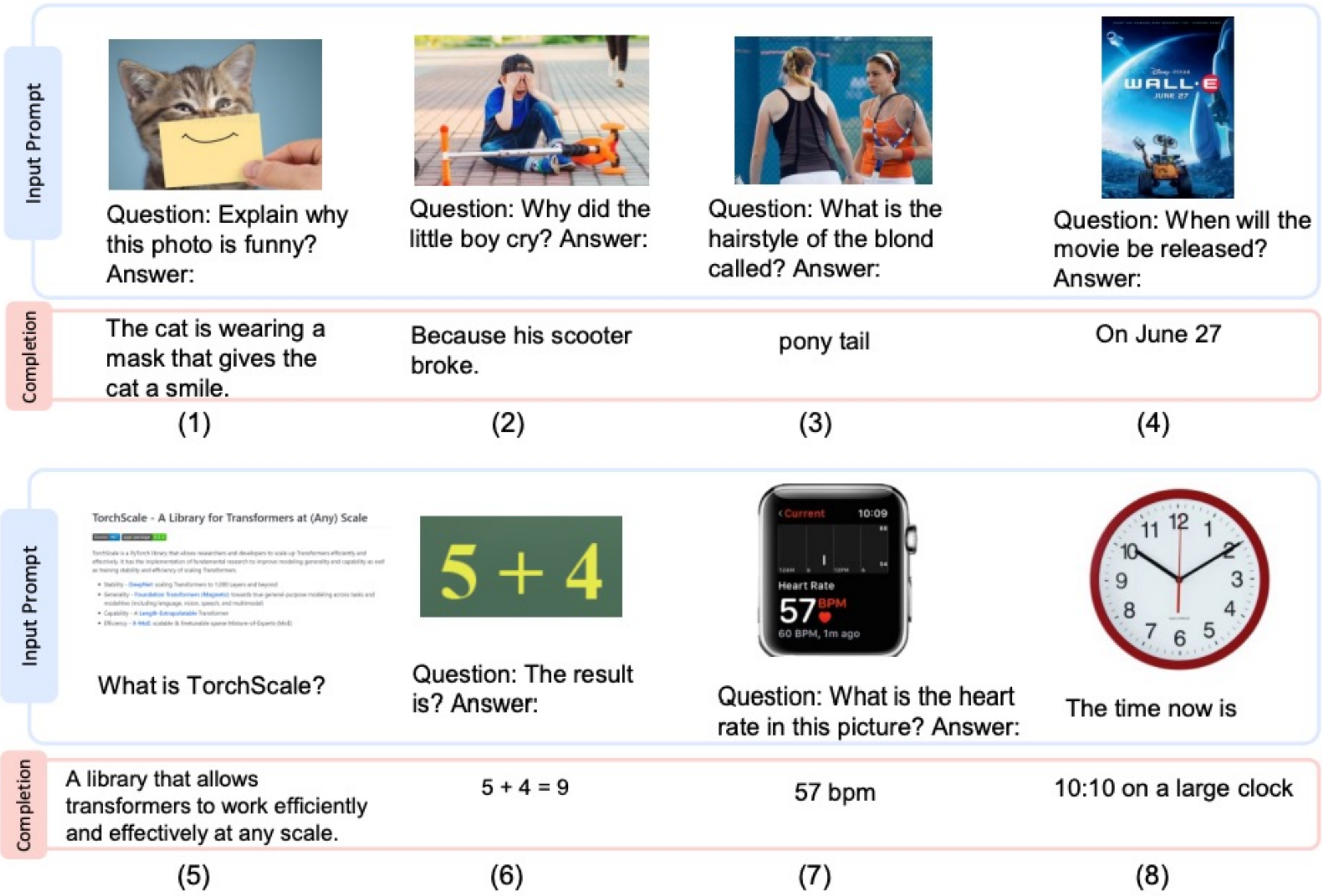
What is [person1] doing now?

ⓐ He is riding on a horse.	0.1%
ⓑ [person1] is being handed a fork to eat his food.	0.2%
ⓒ He is sitting and resting after carrying a heavy item into the room.	**98%**
ⓓ He is in the process of turning to look at [person1].	1.7%

Because...

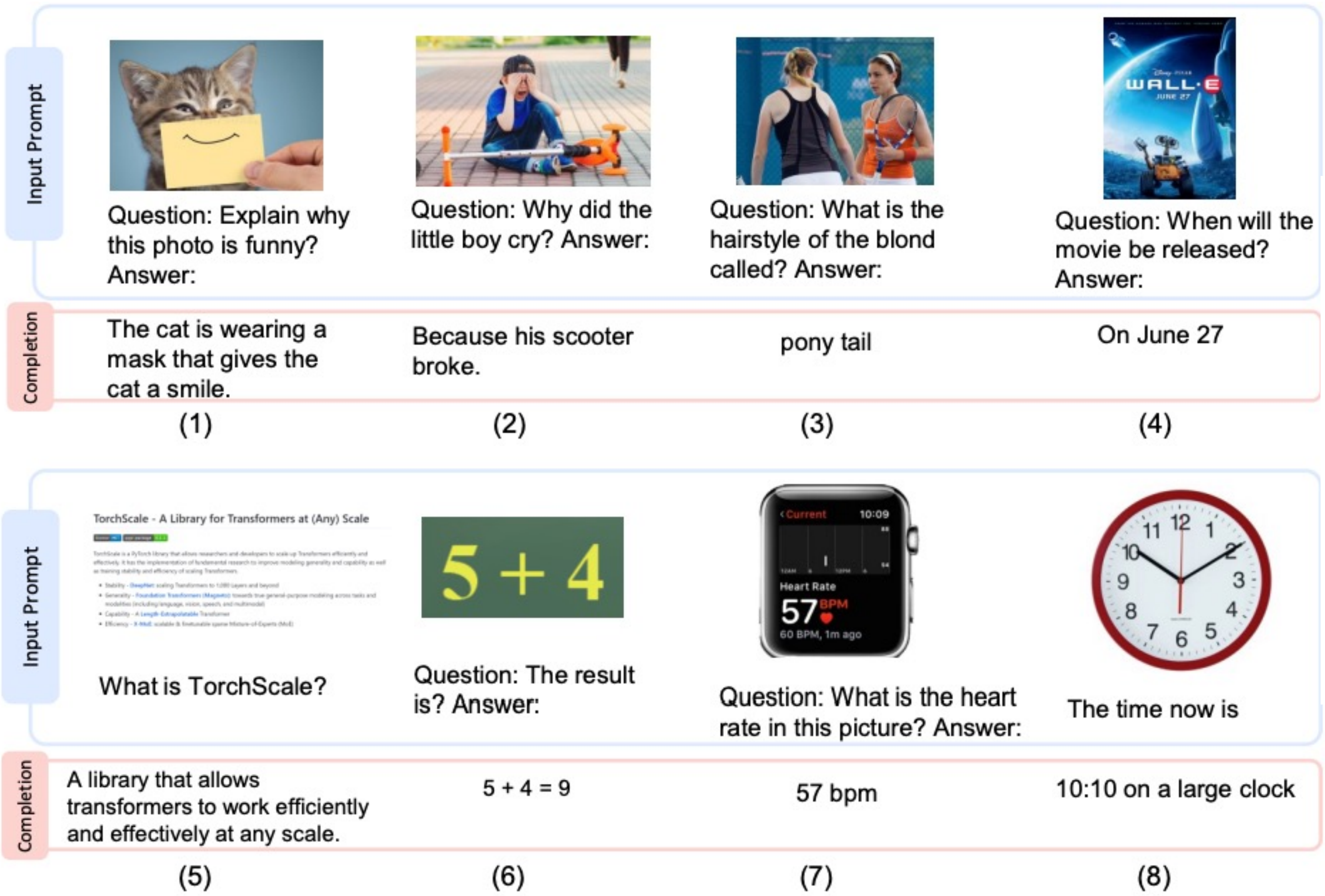
ⓐ He is sitting next to a cart with a large object on it. He is tired from pulling the kart because it is heavy.	0.2%
ⓑ He is using a brace, used to carry things.	14%
ⓒ He has brought many items into the fort which make him feel at ease.	0.8%
ⓓ He had a bench and has put it down.	**85%**

Borrowing Knowledge from Language



Language Is Not All You Need: Aligning Perception with Language Models. arXiv, 2023

Borrowing Knowledge from Language



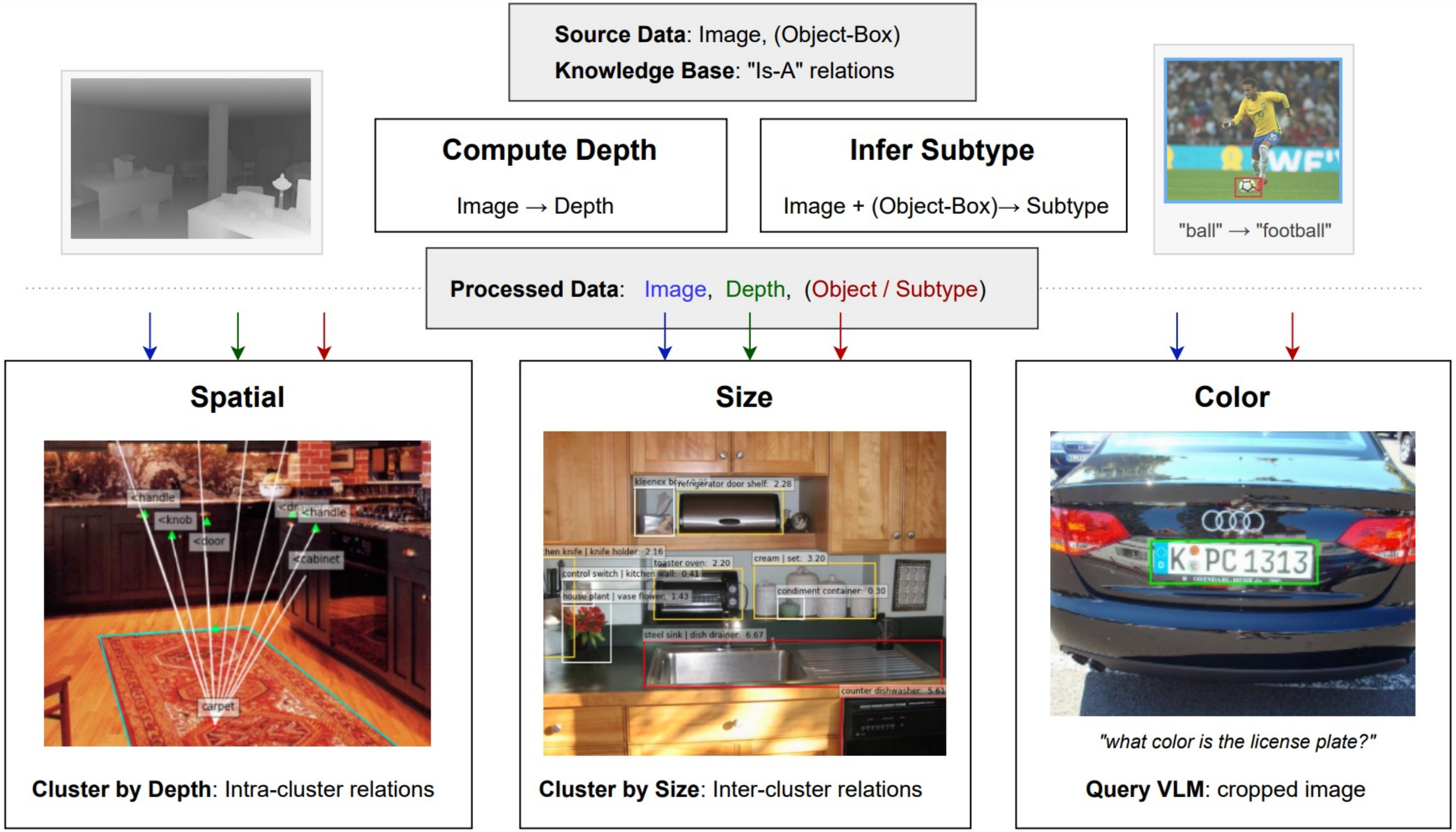
- Language tasks
  - Language understanding
  - Language generation
  - OCR-free text classification
- Cross-modal transfer
  - Commonsense reasoning
- Nonverbal reasoning
  - IQ Test (Raven's Progressive Matrices)
- Perception-language tasks
  - Image captioning
  - Visual question answering
  - Web page question answering
- Vision tasks
  - Zero-shot image classification
  - Zero-shot image classification with descriptions

Language Is Not All You Need: Aligning Perception with Language Models. arXiv, 2023

Part 3: Are VLMs commonsense KBs?

Probing “Visible” Physical Commonsense Knowledge

Visually accessible knowledge representing color, size and space



VIPHY: Probing “Visible” Physical Commonsense Knowledge

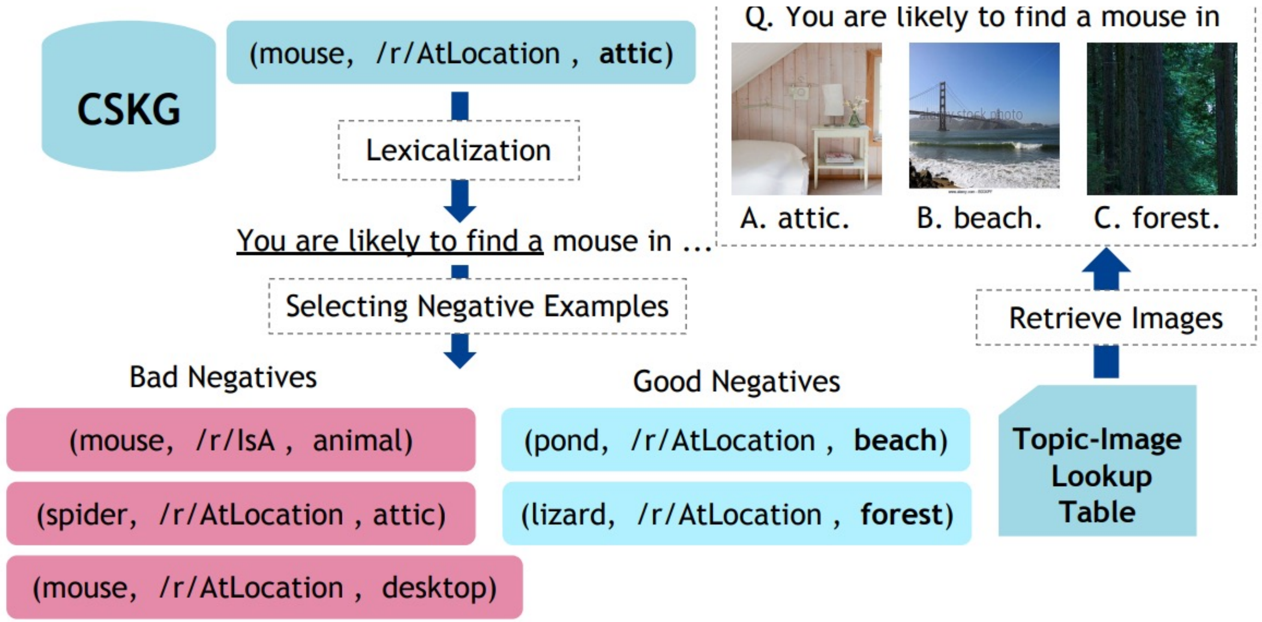
Probing “Visible” Physical Commonsense Knowledge

Visually accessible knowledge representing color, size and space

Task	Setting	Prompt
Color	ZS	O is of `[MASK]` color
	FT	`[CLS]` color of O
	QA	What is the color of O? (a) .. (b) ..
Size	ZS	O_1 is `[MASK]` than O_2 in size
	FT	`[CLS]` size of O_1 in comparison to O_2
	QA	what is the size of O_1 in comparison to O_2? (a) .. (b) ..
Spatial	ZS	in a S, the O_1 is located `[MASK]` the O_2
	FT	`[CLS]` in a S, the O_1 is located in comparison to O_2
	QA	in a S, where is O_1 is located in comparison to O_2? (a) .. (b) ..

Are Visual-Linguistic Models Commonsense KBs?

CS dimension	Starting prompt	Answer candidates	# Instances
part-whole	Furry animals have	A_1: effect of chilling innovation. **A_2: millions of hair.** A_3: hole in.	1,165
taxonomic	Recruit is a way to	A_1: rate. **A_2: enlist.** A_3: slope.	1,323
distinctness	Shade is not	A_1: flat. A_2: postal worker. **A_3: sunny.**	828
similarity	Throw up is a synonym of	A_1: rutinic acid. A_2: random. **A_3: vomit.**	644
quality	A wet floor is	**A_1: slippery.** A_2: light brown. A_3: abbreviated to unido.	1,840
utility	A fork is used for	A_1: speed of transit. A_2: confuse voters. **A_3: picking up food.**	2,090
creation	Music is created by	A_1: olive oil mill. A_2: mapping process. **A_3: instruments.**	100
temporal	Going for a haircut requires	**A_1: finding barber.** A_2: hard examinations. A_3: write persuasively.	1,889
spatial	You are likely to find a document folder in	**A_1: file drawer.** A_2: madagascar jungle. A_3: minerals.	1,599
desire	You would thank someone because you want to	A_1: accomplish mutual goal. **A_2: feel good.** A_3: cool off.	1,781



Are Visual-Linguistic Models Commonsense KBs?

CS dimension	Starting prompt	Answer candidates	# Instances
part-whole	Furry animals have	A_1: effect of chilling innovation. **A_2: millions of hair.** A_3: hole in.	1,165
taxonomic	Recruit is a way to	A_1: rate. **A_2: enlist.** A_3: slope.	1,323
distinctness	Shade is not	A_1: flat. A_2: postal worker. **A_3: sunny.**	828
similarity	Throw up is a synonym of	A_1: rutinic acid. A_2: random. **A_3: vomit.**	644
quality	A wet floor is	**A_1: slippery.** A_2: light brown. A_3: abbreviated to unido.	1,840
utility	A fork is used for	A_1: speed of transit. A_2: confuse voters. **A_3: picking up food.**	2,090
creation	Music is created by	A_1: olive oil mill. A_2: mapping process. **A_3: instruments.**	100
temporal	Going for a haircut requires	**A_1: finding barber.** A_2: hard examinations. A_3: write persuasively.	1,889
spatial	You are likely to find a document folder in	**A_1: file drawer.** A_2: madagascar jungle. A_3: minerals.	1,599
desire	You would thank someone because you want to	A_1: accomplish mutual goal. **A_2: feel good.** A_3: cool off.	1,781

dim.: spatial

You are likely to find vegetables in:
A. workplace.
B. stationary shop.
C. *garden.*

dim.: part-whole

A boat has:

A. reached legal age.
B. *sails*
C. different rules.

dim.: quality

A hill can be:

A. *steep.*
B. about to change.
C. important for normal living.

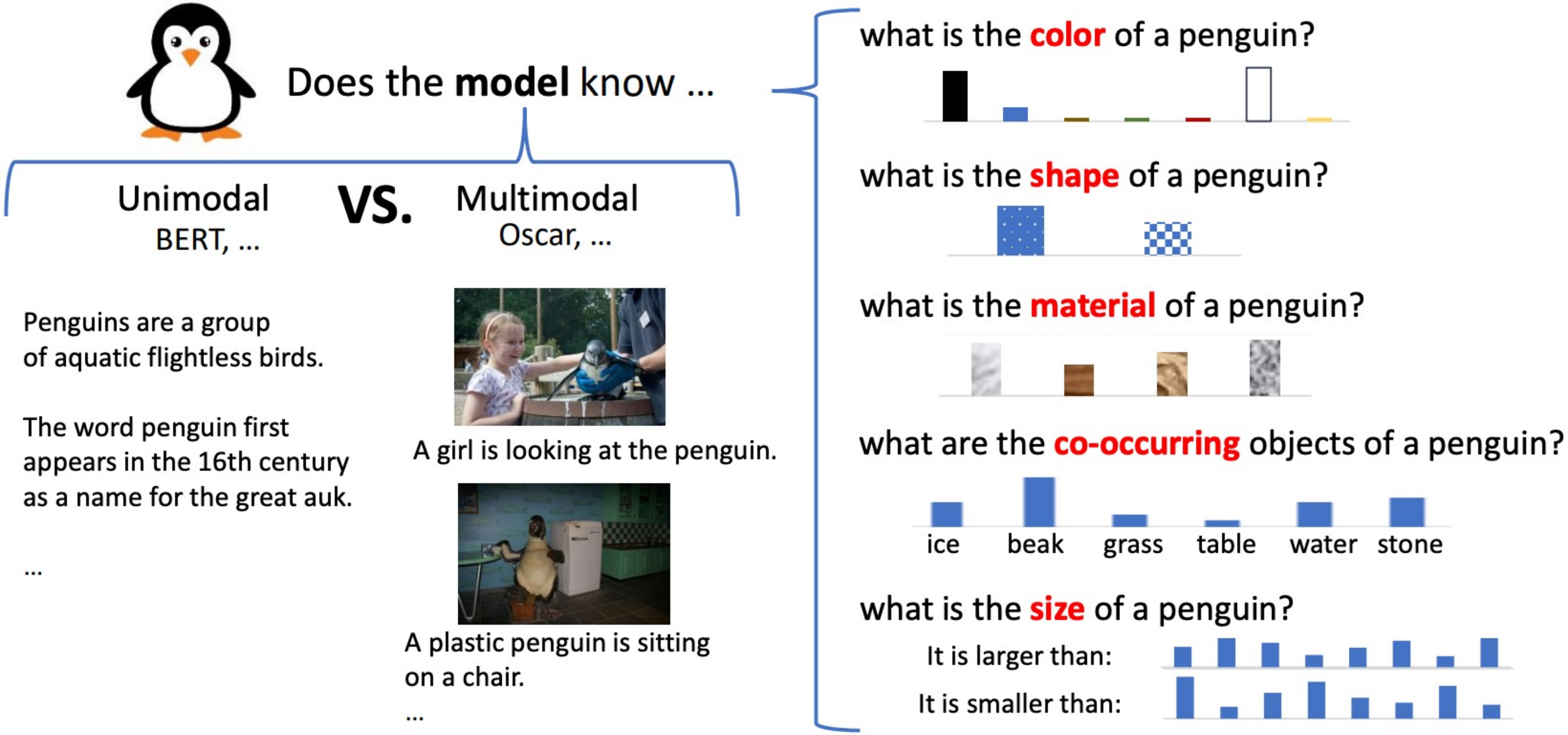
Are Visual-Linguistic Models Commonsense KBs?

Visual Commonsense Knowledge is more difficult than textual knowledge.

row	Images	part-whole 1, 165	taxonomic 1, 323	distinctness 828	similarity 644	quality 1, 840	utility 2, 090	creation 100	temporal 1, 189	spatial 1, 599	desire 1, 781	All 13, 259
1 RoBERTa	–	68.5	61.8	80.2	**67.4**	69.7	**74.2**	72.0	**60.9**	54.8	**65.9**	**67.5**
2 BERT	–	62.8	71.2	80.1	54.8	68.1	72.4	74.0	53.7	52.4	60.4	65.0
3 BERT_{CC}	–	68.4	62.0	66.6	51.1	66.0	65.4	62.0	53.6	**63.7**	58.3	61.9
4 UNITER_BERT_{T}	–	70.1	**74.5**	**81.4**	62.4	**72.0**	73.8	**79.0**	54.5	53.9	61.5	66.5
5 UNITER_{T}	–	**70.9**	59.8	71.3	51.2	69.9	71.5	71.0	52.7	61.5	62.5	64.0
6 VILBERT_{T}	–	63.9	60.3	64.9	46.7	66.1	71.2	58.0	52.2	61.0	62.8	60.7
7 UNITER_{TV}	retrieved	63.0	54.0	65.9	46.4	62.4	65.4	62.0	49.2	57.4	58.5	58.4
8 VILBERT_{TV}	retrieved	55.0	49.9	55.9	42.2	57.4	60.5	52.0	47.2	52.9	56.6	53.0
9 $\text{UNITER}_{T\tilde{V}}$	dummy	61.5	51.6	63.4	42.2	63.6	66.4	55.0	49.4	58.2	59.7	57.1
10 $\text{VILBERT}_{T\tilde{V}}$	dummy	60.4	58.9	64.9	43.9	63.4	65.5	55.0	48.4	56.8	62.0	57.9
11 UNITER_{V}	retrieved	36.4	36.6	40.1	38.5	34.2	36.6	32.0	34.8	36.2	34.3	36.0
12 VILBERT_{V}	retrieved	37.8	35.1	37.7	39.8	36.8	35.7	41.0	33.0	37.6	34.0	36.8
13 $\text{UNITER}_{\tilde{V}}$	dummy	30.8	26.3	45.7	28.6	29.2	28.7	19.0	28.7	29.6	30.7	29.7
14 $\text{VILBERT}_{\tilde{V}}$	dummy	34.8	35.8	50.5	40.4	30.4	31.1	30.0	29.4	33.5	30.1	34.6

Unimodal vs Multimodal models?

Unimodal and multimodal models' abilities to capture visual commonsense knowledge



Unimodal vs Multimodal models?

ViComTe dataset on five relation types: color, shape, material, size, and visual co-occurrence

Relation	**# Classes**	**# (subj, obj) Pairs**	**Ex Template**	**Ex (subj, obj) Pair**
color	12	2877	[subj] *can be of color* [obj]	(*sky*, *blue*)
shape	12	706	[subj] *has shape* [obj] .	(*egg*, *oval*)
material	18	1423	[subj] *is made of* [obj] .	(*sofa*, *cloth*)
size (smaller)	107	2000	[subj] *is smaller than* [obj] .	(*book*, *elephant*)
size (larger)	107	2000	[subj] *is larger than* [obj] .	(*face*, *spoon*)
co-occurrence	5939	2108	[subj] *co-occurs with* [obj] .	(*fence*, *horse*)

Unimodal vs Multimodal models?

Unimodal and multimodal models' abilities to capture visual commonsense knowledge

Source	Group	Spearman ρ	# Subjs	Avg # Occ	Top5 # Occ	Btm5 # Occ	Acc@1
VG	All	64.3 ± 23.9	355	1252.6	64.6	308.6	
	SINGLE	62.2 ± 24.0	131	494.9	64.6	1181.6	80.2
	MULTI	69.3 ± 20.7	136	1156.1	2062.2	347.0	
	ANY	58.4 ± 27.1	88	2529.6	8452.4	1213.4	
Wikipedia	All	33.4 ± 30.6	302	543.6	1758.0	49.8	
	SINGLE	29.6 ± 29.9	110	352.2	345.8	35.0	35.5
	MULTI	33.9 ± 30.9	119	500.8	1242.0	27.6	
	ANY	38.2 ± 30.4	73	902.0	3000.2	161.2	

Visual Commonsense in Pretrained Unimodal and Multimodal Models. NAACL 2022

Future Direction:
Adding commonsense knowledge to pretraining

Knowledge for Model

?

- In-context prompt
- data augmentation
- data selection

Knowledge for data

Knowledge → VLM

Knowledge Graph Riddles

Knowledge Graph Embedding

VLM → Knowledge

Physical knowledge

Unimodal vs Multimodal

Future Direction: Physical Knowledge Enhanced LM/VLM

Humans learn a huge amount of knowledge about the external world via **multisensory experience and interactions**, however, current **LLM/VLM** are trained with **static datasets**, thus **lacks understanding of the physical world**.

Spatial Relation

Z: Put object A to the left of object B. Then, put object B in front of object A. Then, put object C to the left of object A. Which object is directly behind object B?

Object C is directly behind object B.

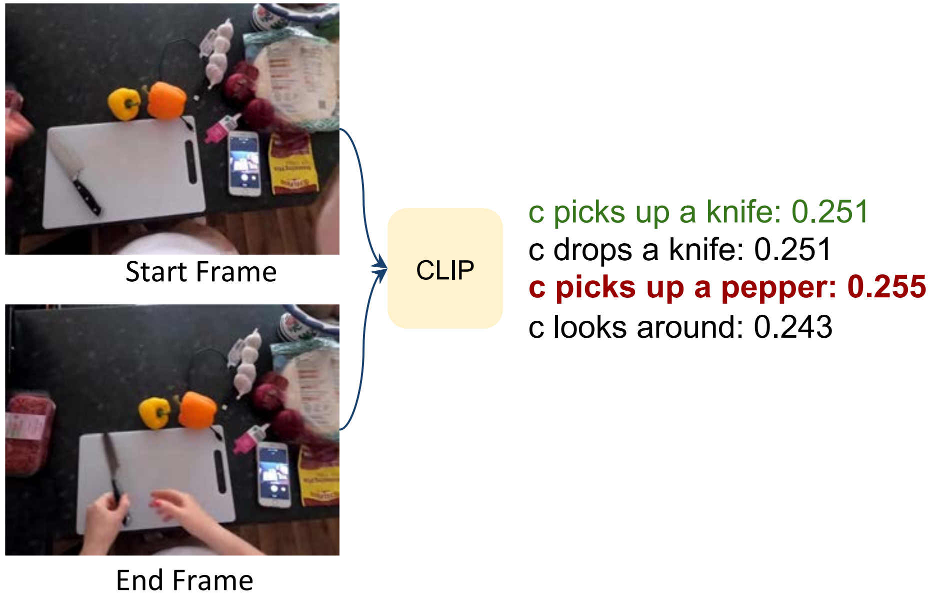
Knowledge requiring embodiment

Z: Imagine you are a human being. Put your left hand on the back of your head. Can you still see your left hand?

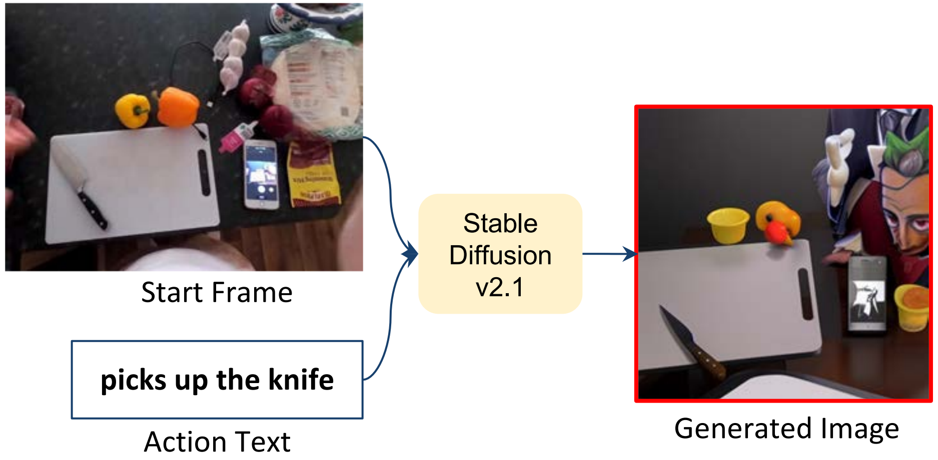
Yes, I can still see my left hand as it is positioned on the back of my head.

Future Direction:
Physical Knowledge Enhanced LM/VLM

Humans learn a huge amount of knowledge about the external world via **multisensory experience and interactions**, however, current **LLM/VLM** are trained with **static datasets**, thus **lacks understanding of the physical world**.



Action classification

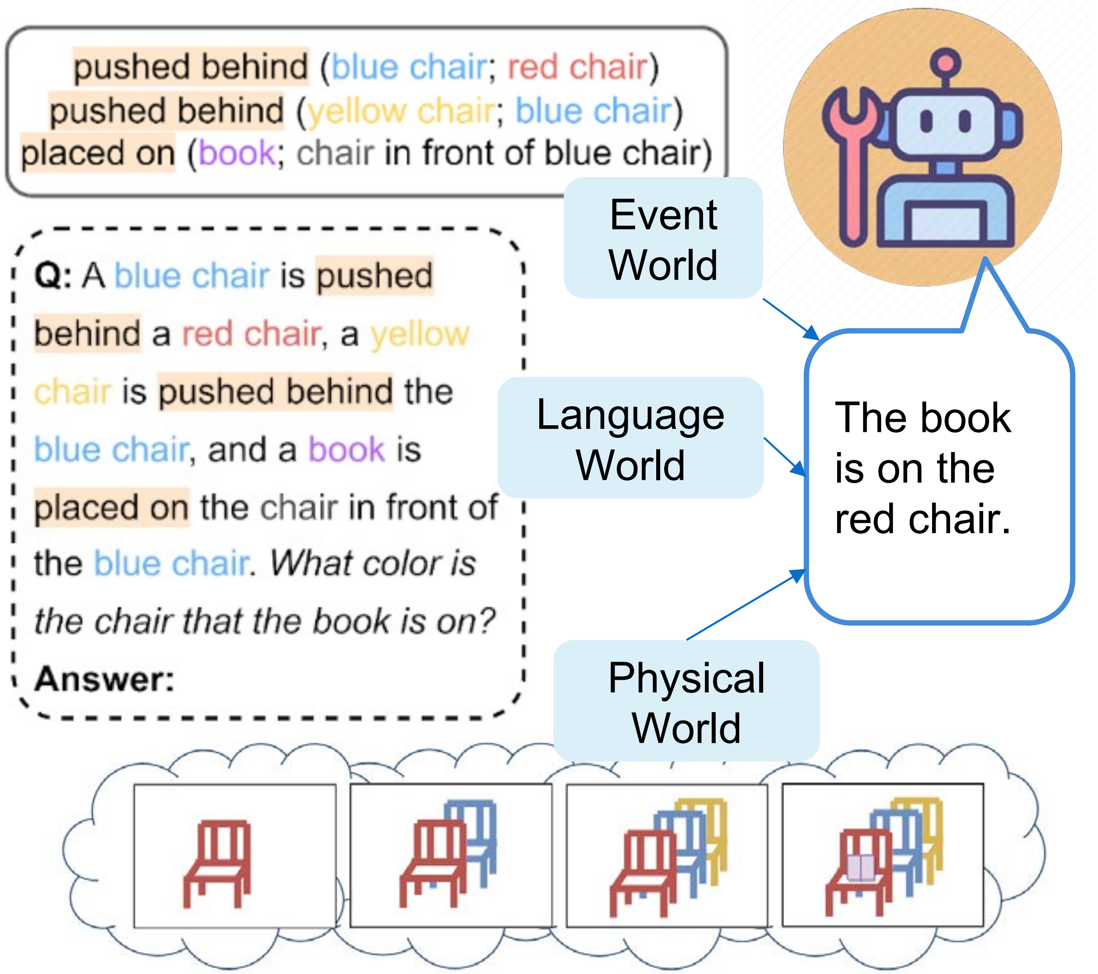


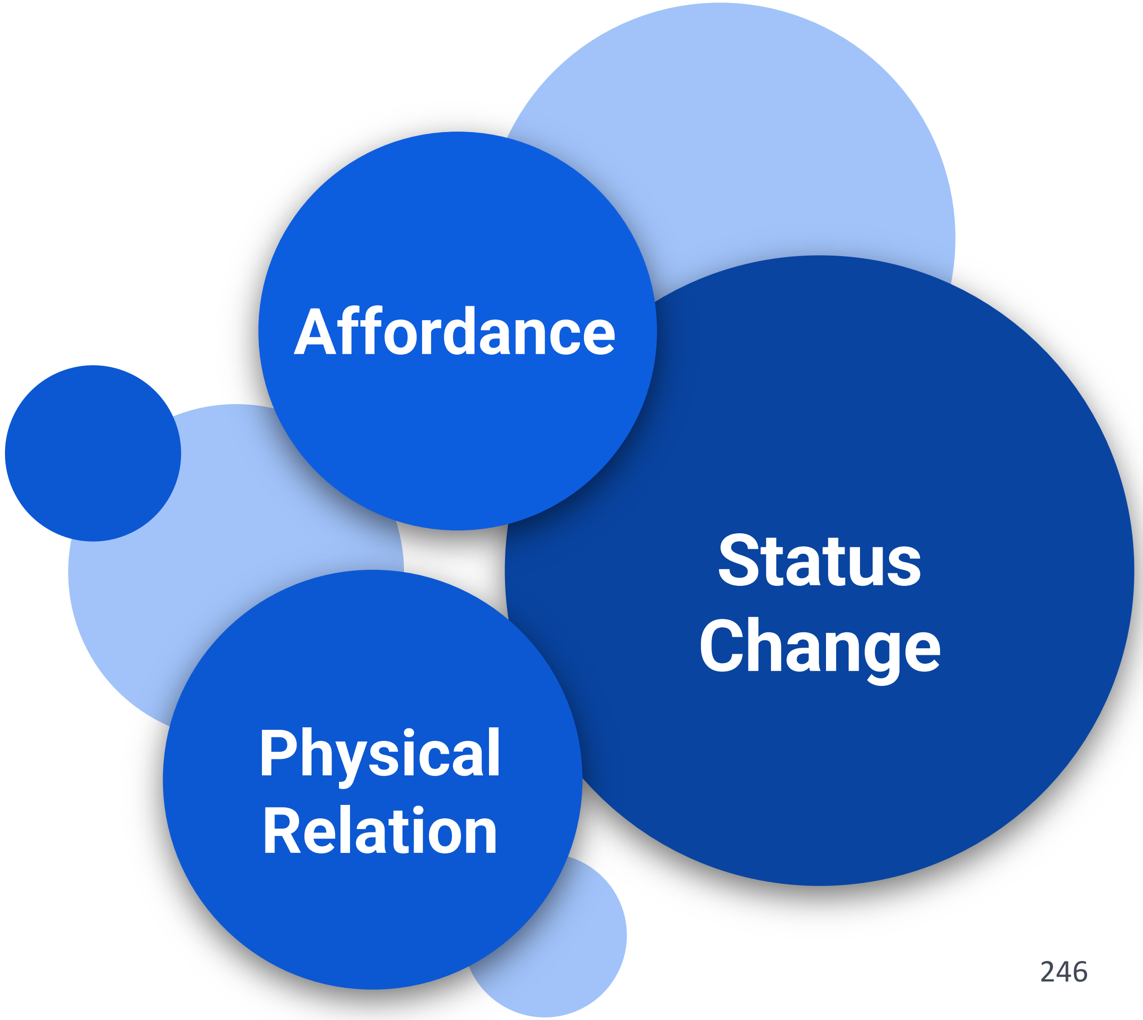
Causal Effect of Actions

Physical Interactions involving actions and objects

Future Direction:
Physical Knowledge Enhanced LM/VLM

From Reading/Seeing to Doing: From passive perception to interaction with the world.





Future Direction 2:

Disentangling Perception & Reasoning

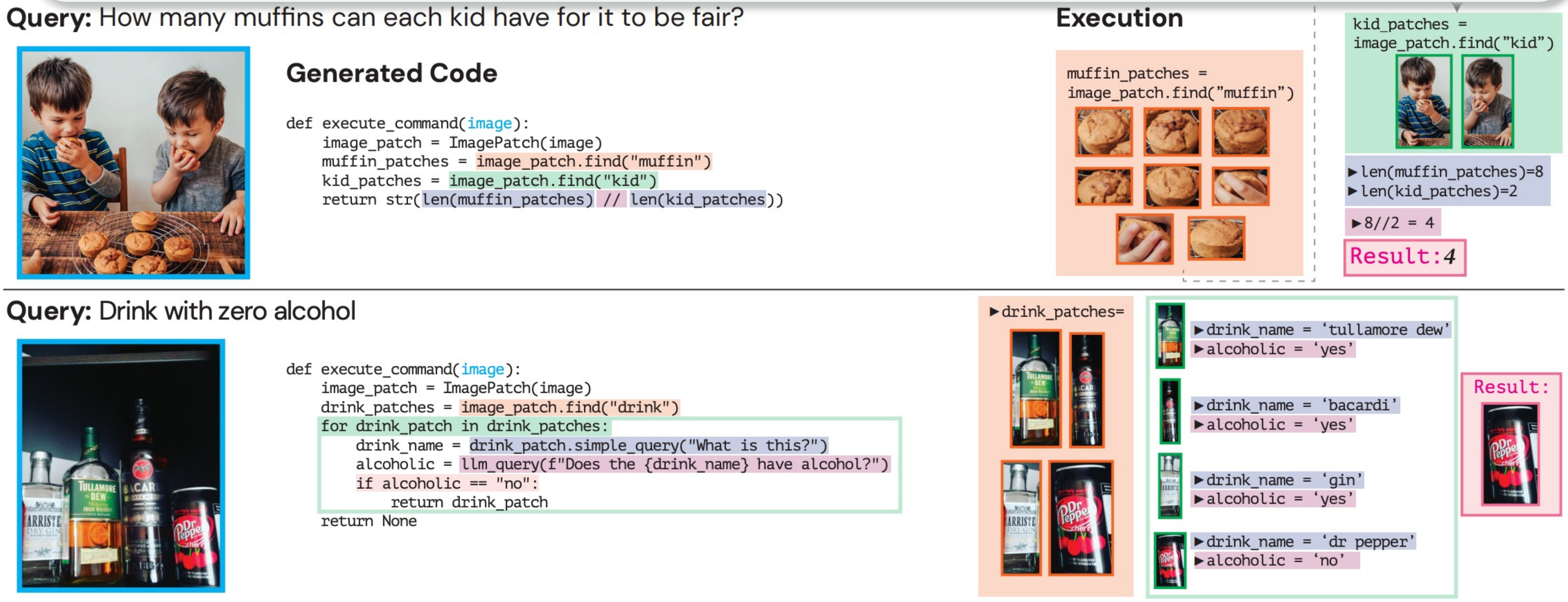
Initial Exploration: ViperGPT

VIPER-style reasoning shows the potential of **treating perception models as *tools*** and **LLMs as *reasoner*** to solve difficult problems.

[Dídac Surís*, Sachit Menon*, Carl Vondrick. arXiv 2023]

Initial Exploration: ViperGPT

Bottleneck:
- Reasoning is **limited** to the *commonsense* and *code generation* capabilities of LLMs;
- Can only handle a single image context.



[Dídac Surís*, Sachit Menon*, Carl Vondrick. arXiv 2023]

Example: Answering Questions using Tools

Q: At which festival can you see a castle in the background: Oktoberfest in Domplatz Austria or Tanabata festival in Hiratsuka, Japan?

A: You can see a castle in the background at Oktoberfest in Domplatz, Austria

WebQA

```
question = "At which festival can you see a castle in the background: Oktoberfest in
Domplatz Austria or Tanabata festival in Hiratsuka, Japan?"
input_data = [(<img1>, "J24 029 Dom, Oktoberfest"), ...]

# In[1]:
# Filter out irrelevant information for the question
input_data = [
            data_instance
            for data_instance in input_data
            if solve(f"is {data_instance[1]} relevant to the question: {question}")
            # recursive call
            # we can even offload these to a better model (e.g., GPT-3.5)
            # if ask_gpt_yes_or_no(f"is {data_instance[1]} relevant to the question:
{question}")
]
# Out[1]:
input_data == [(<img1>, "J24 029 Dom, Oktoberfest"), (<img6>, "Tanabata festival in
Hiratsuka")]

# In[2]:
# solve("Which data instance with a image has a castle on the background?")
from multimodal_models import CLIP
img_features = [CLIP.image_encoder(img) for img, text in input_data]
text = "There is a castle in the background"
text_feature = CLIP.text_encoder(text)
has_castle = [cosine_similarity(text_feature, feat) for feat in img_features]
idx_more_likely_to_have_castle = argmax(has_castle)
# Out[2]:
idx_more_likely_to_have_castle == 0
input_data[idx_more_likely_to_have_castle] == (<img1>, "J24 029 Dom, Oktoberfest")

# In[3]:
# Synthesize solution from caption
answer = "You can see a castle in the background at Oktoberfest in Domplatz, Austria"
```

Figure Credits: Xingyao Wang

Future Direction 2: Disentangling Perception & Reasoning

- **Perception:**
  - Vision-only model (e.g., object detection)
  - Vision-language model (e.g., captioning, QA)
- **Reasoner:** Language-only model – We need **divide-and-conquer**!
  - (1) decompose a problem (e.g., a sub-function call)
  - (2) use tool to solve a problem (e.g., access external database, fetch relevant information)
  - (3) update the conclusion (e.g., store something back into the database)

Figure Credits: Xingyao Wang

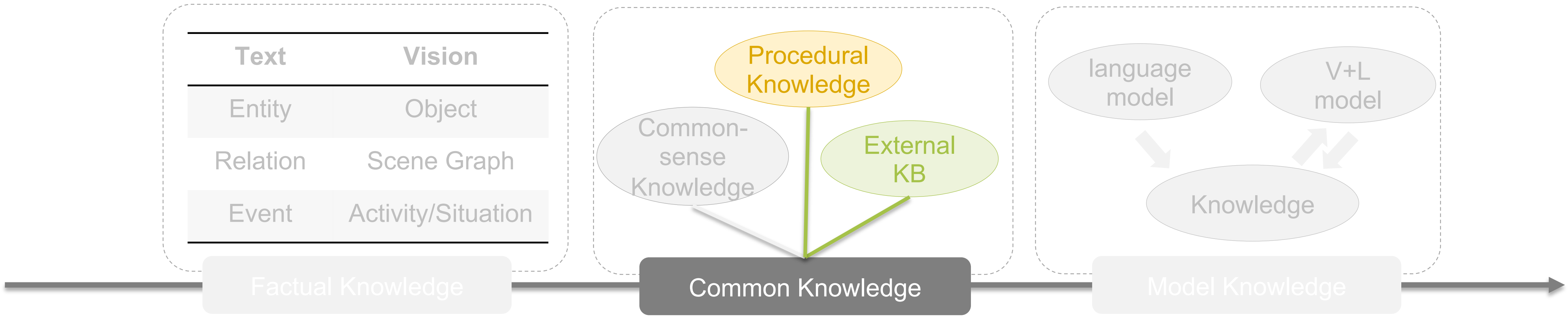
Procedural Knowledge

Knowledge-Driven Vision-Language Encoding (Part IV)

Xudong Lin
Columbia University
xudong.lin@columbia.edu

Content

Learning patterns of procedure with human-curated patterns and data.

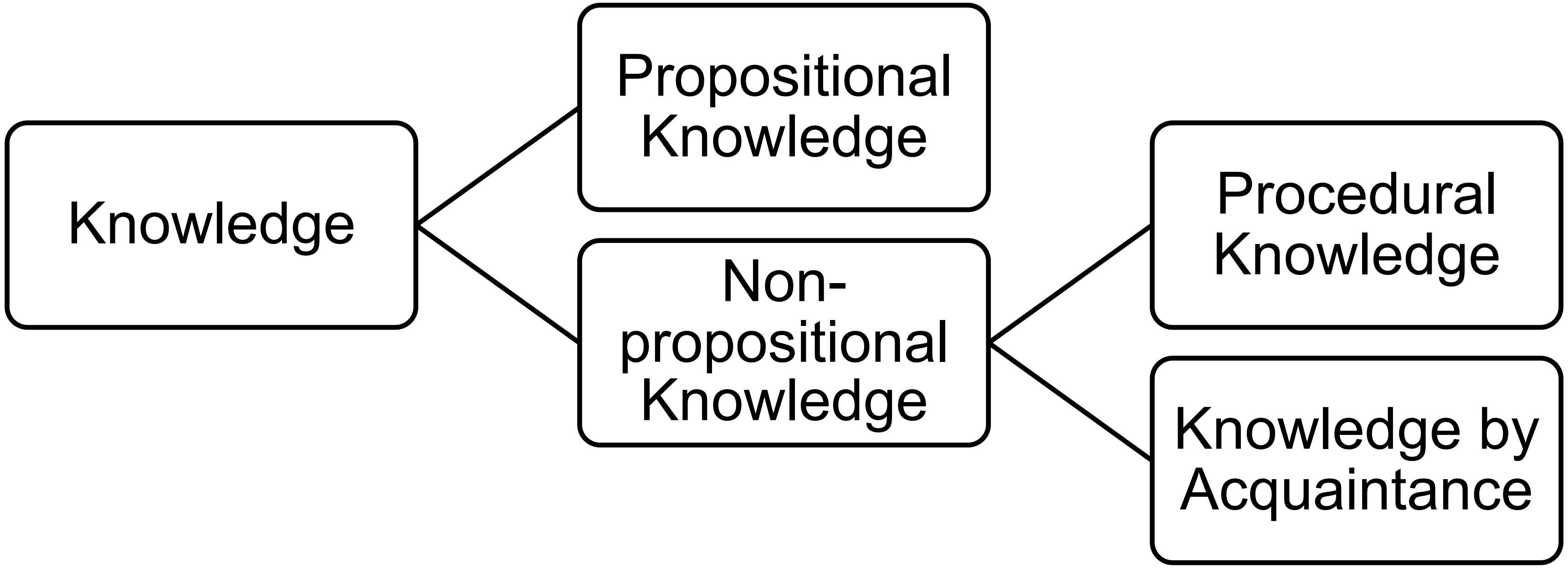


Agenda

- What is Procedural Knowledge?
- Tasks requiring Procedural knowledge.

What is Procedural Knowledge?

- Psychology View

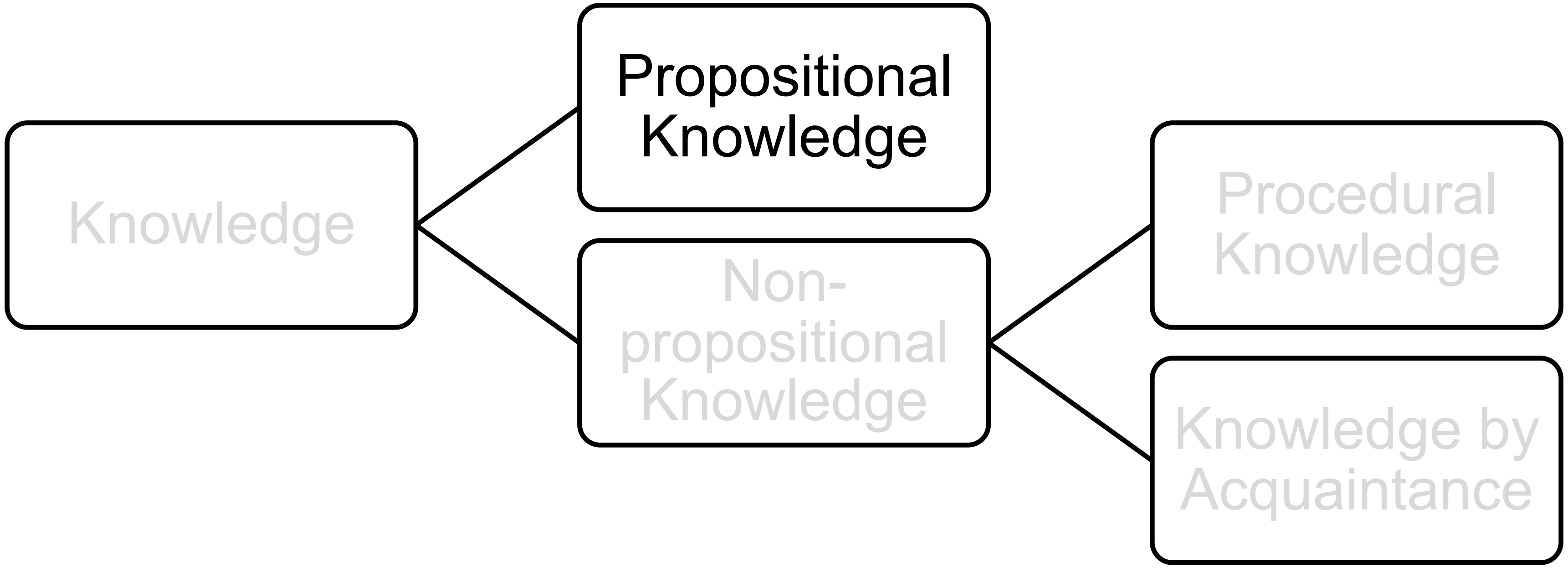


https://www.rep.routledge.com/articles/thematic/knowledge-concept-of/v-1/sections/the-varieties-of-knowledge

What is Procedural Knowledge?

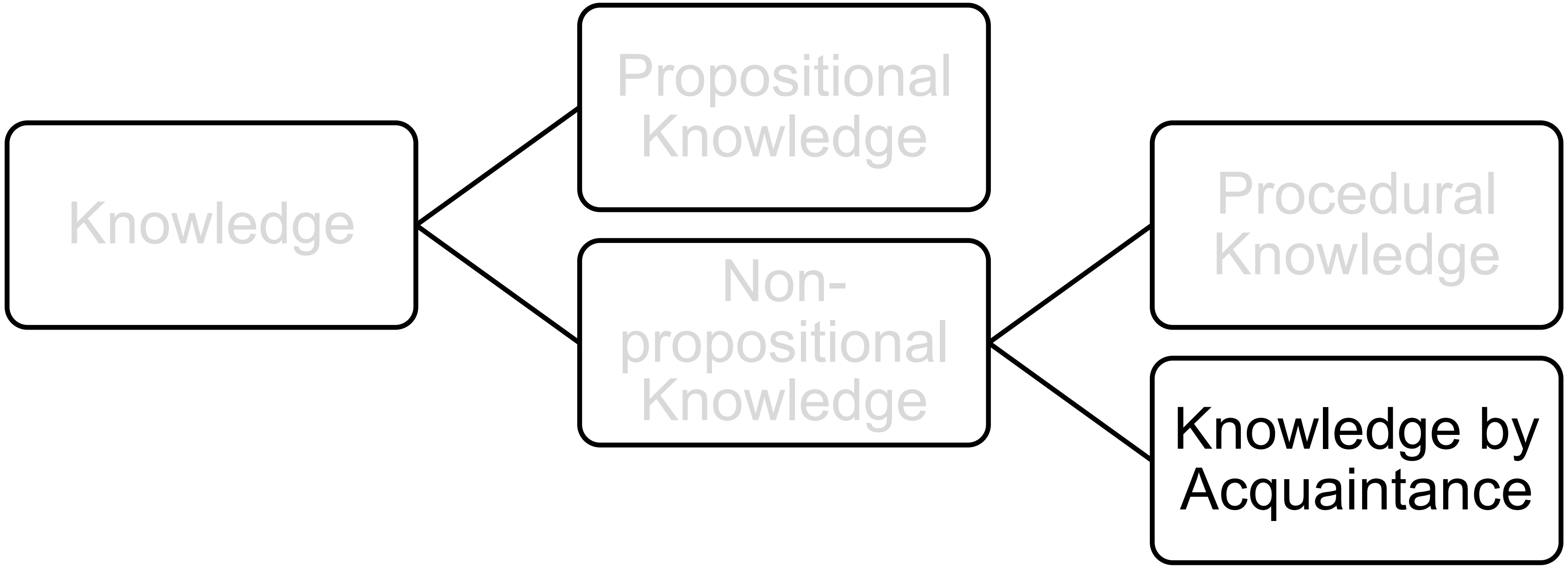
- Psychology View

I know Ottawa is the
capitol of Canada.



What is Procedural Knowledge?

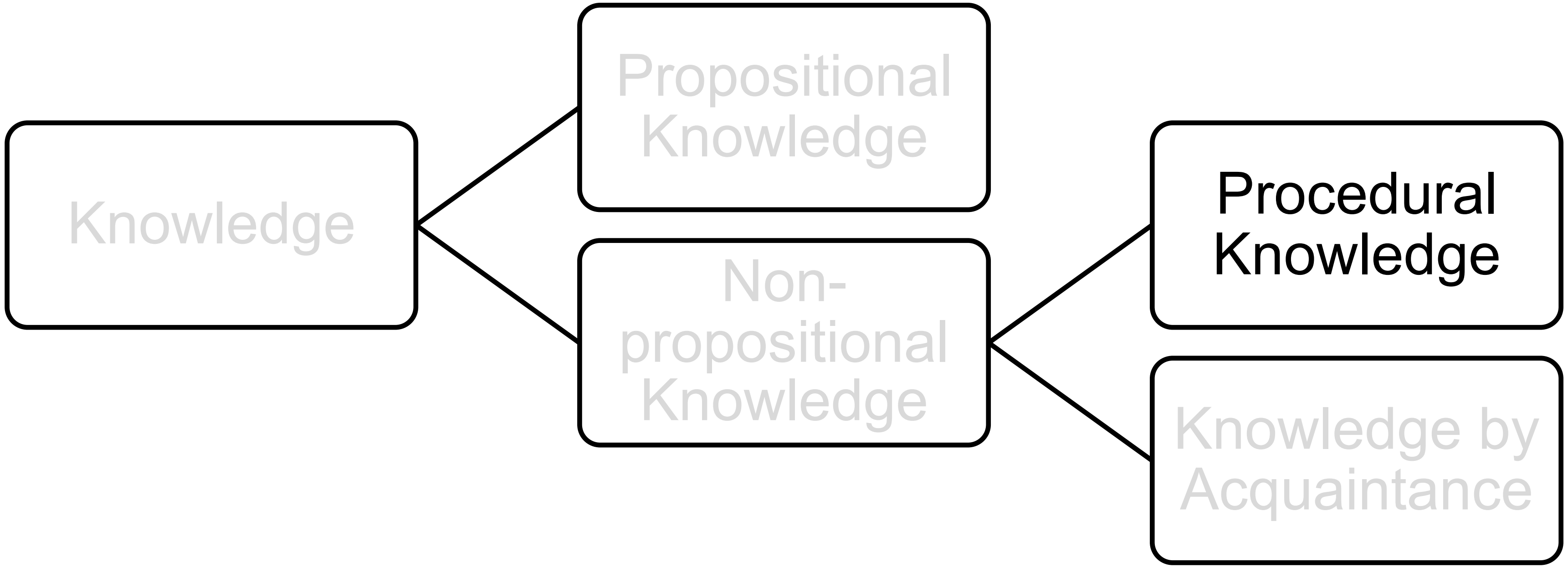
- Psychology View



I know someone.

What is Procedural Knowledge?

- Psychology View



I know **how** to do something.

Tasks Requiring Procedural Knowledge

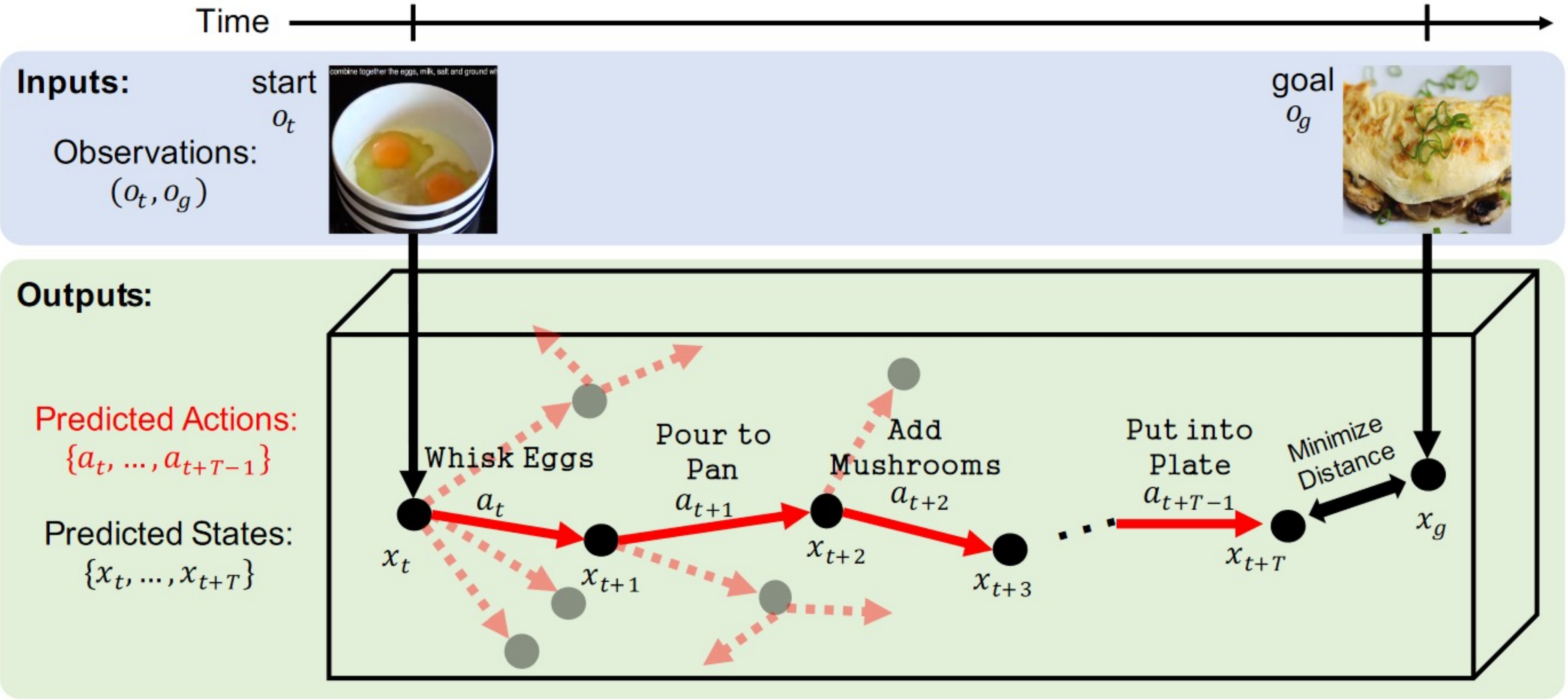
- Procedural planning

Given a start image and an end image, generate a sequence of actions.

Chang, Chien-Yi, et al. "Procedure planning in instructional videos." *European Conference on Computer Vision*. Springer, Cham, 2020.

Tasks Requiring Procedural Knowledge

- Procedural planning



Given a start image and an end image, generate a sequence of actions.

Chang, Chien-Yi, et al. "Procedure planning in instructional videos." *European Conference on Computer Vision*. Springer, Cham, 2020.

Tasks Requiring Procedural Knowledge

- Step forecasting

Time

What is the next step?

Given the historical video, predict the next step.

Frames are from Gordon Ramsay's **Fillet of Beef Wellington**

Sener, Fadime, and Angela Yao. "Zero-shot anticipation for instructional activities." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.
Lin, Xudong, et al. "Learning to recognize procedural activities with distant supervision." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

Tasks Requiring Procedural Knowledge

- Step forecasting

Time

What is the next step?

Assembling: Shingle the prosciutto on the plastic wrap; Spread mushroom over prosciutto; …

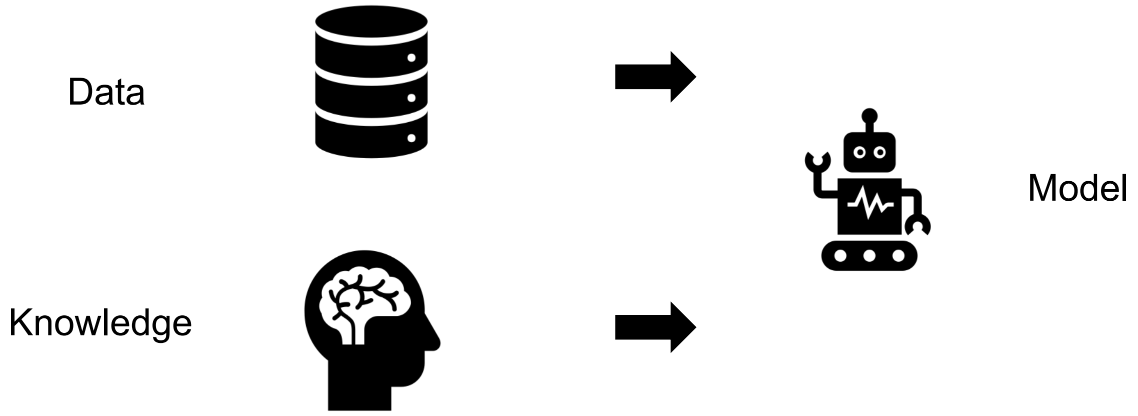
Given the historical video, predict the next step.

Frames are from Gordon Ramsay's **Fillet of Beef Wellington**

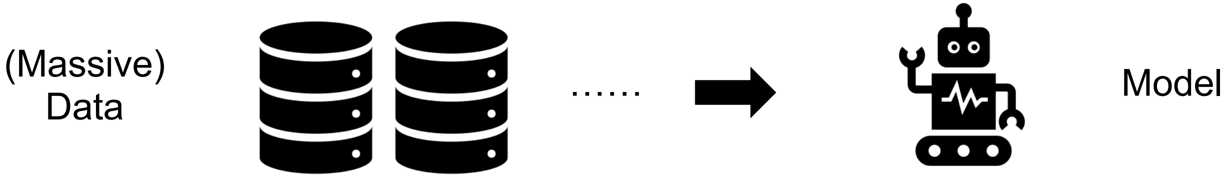
Sener, Fadime, and Angela Yao. "Zero-shot anticipation for instructional activities." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.
Lin, Xudong, et al. "Learning to recognize procedural activities with distant supervision." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

Agenda

- Explicit Knowledge Source: Learning with the help of external knowledge

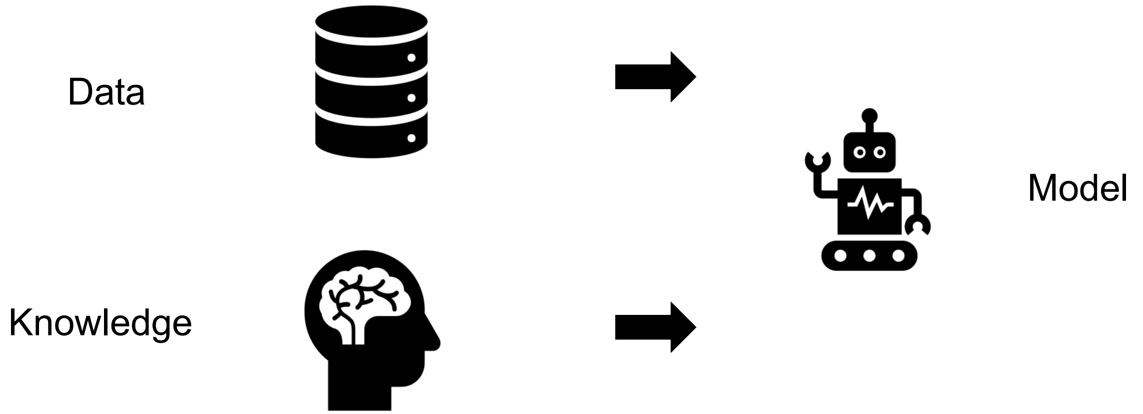


- Implicit Knowledge Source: Learning procedural knowledge from data



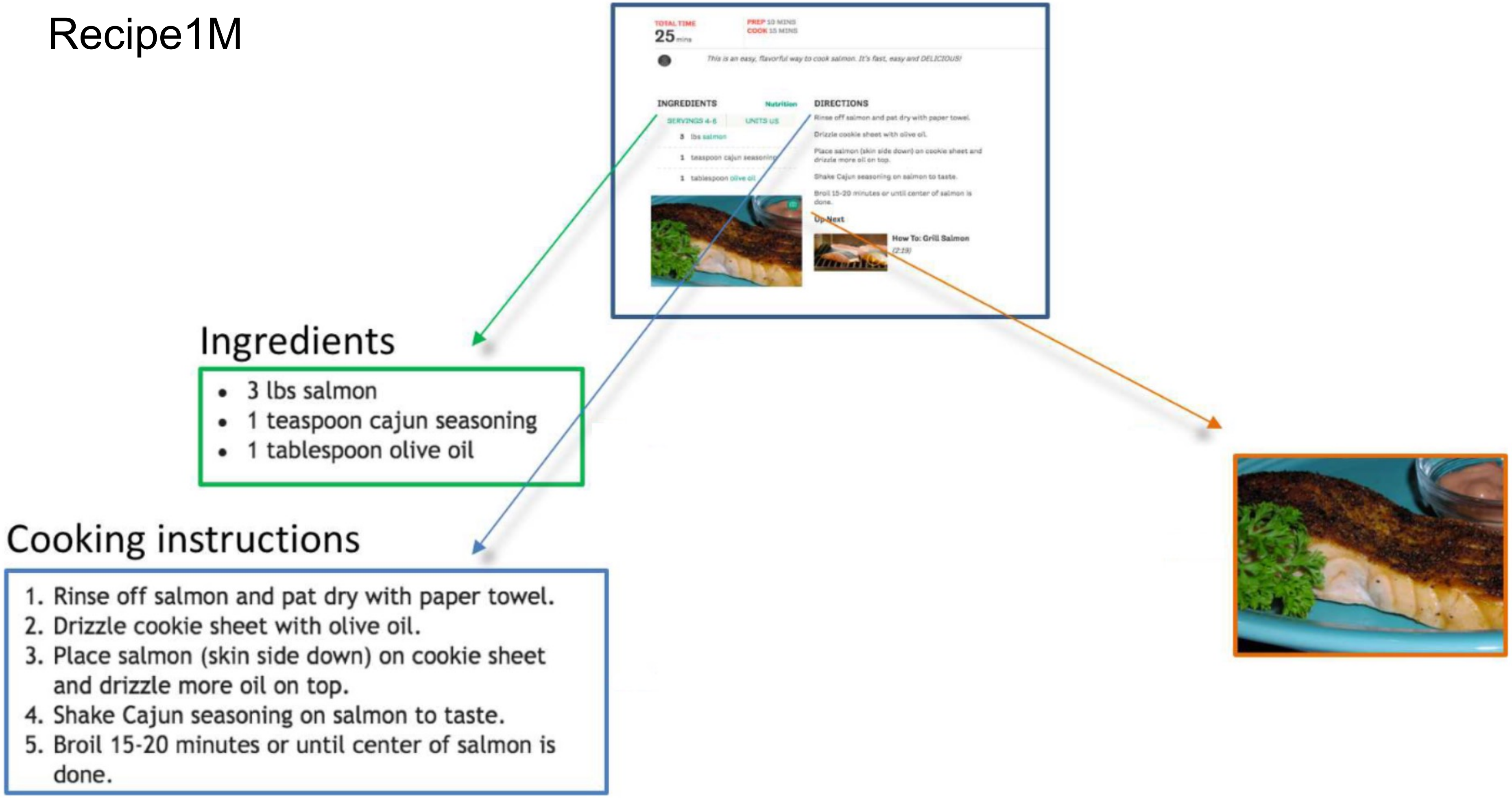
Agenda

- Explicit Knowledge Source: Learning with the help of external knowledge



Explicit Knowledge Source

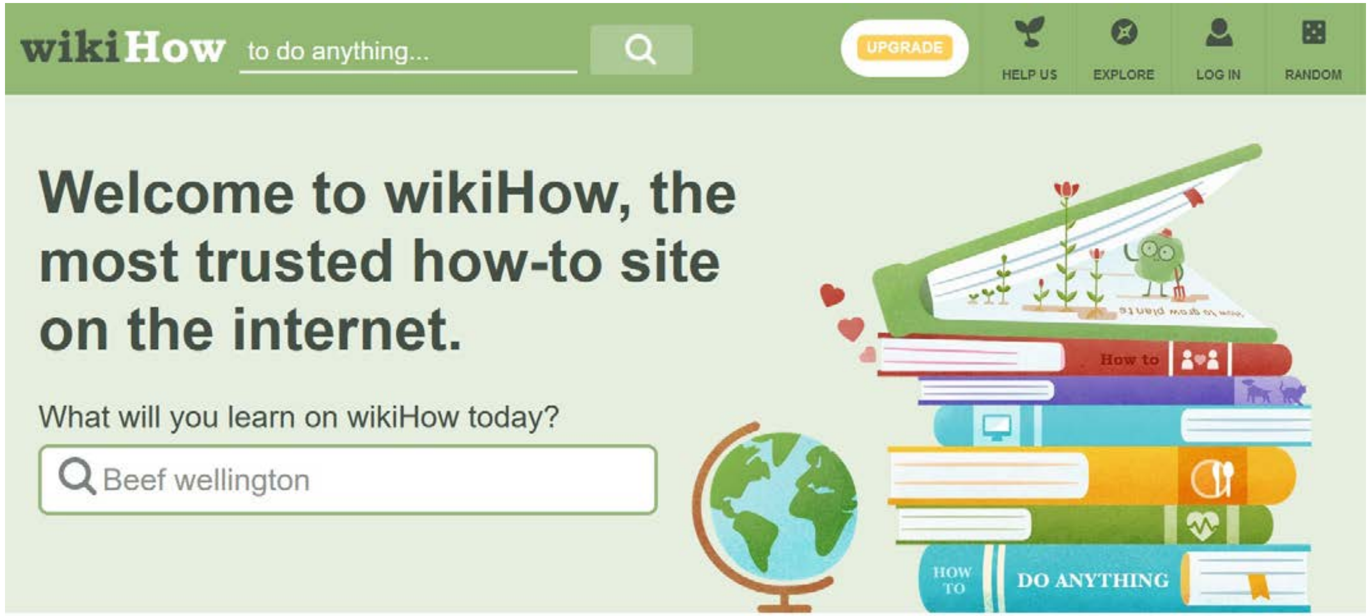
- Procedural knowledge can be easily curated from the Internet
  - Recipe1M



Salvador, Amaia, et al. "Learning cross-modal embeddings for cooking recipes and food images." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

Explicit Knowledge Source

- Procedural knowledge can be easily curated from the Internet
  - Recipe1M
  - wikiHow



Step 1. Sear the filiet mignon to brown.
Over high heat, coat bottom of a heavy skillet with olive oil. Once pan is nearly smoking, sear tenderloin until well-browned on all sides.

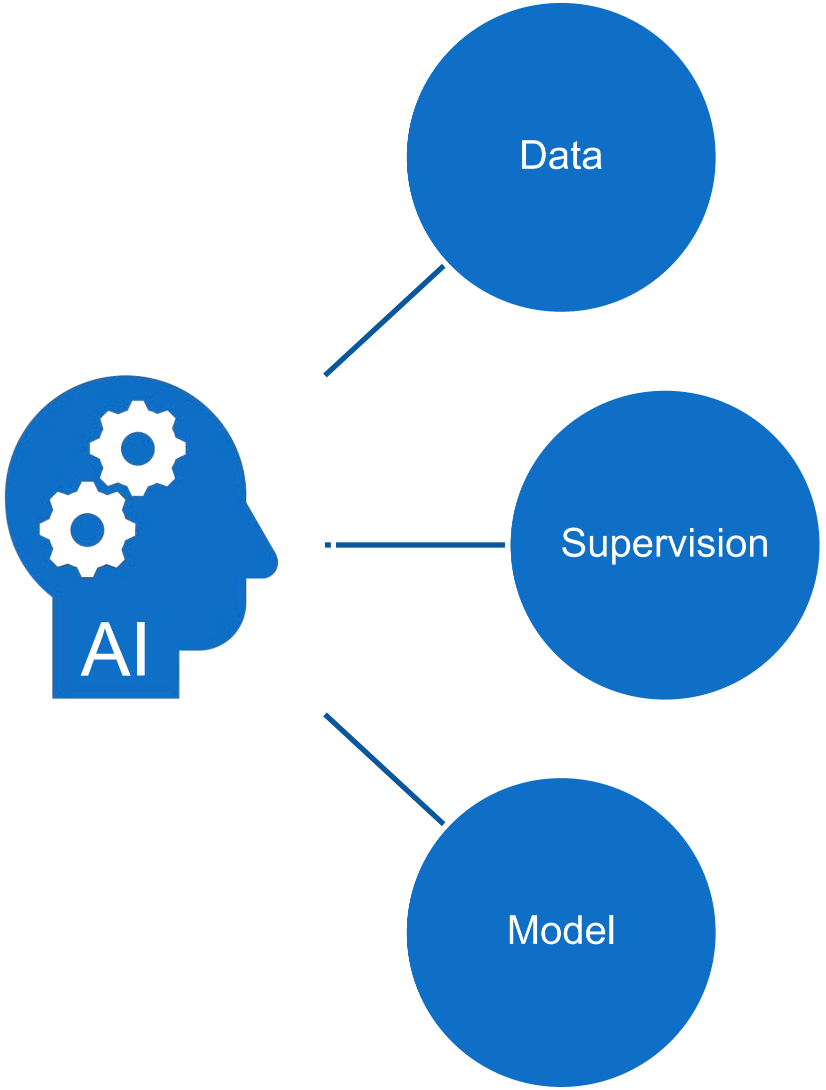
Step 2. Fry the mushroom until they are dried.
To skillet, add butter and melt over medium heat. Add mushroom mixture and cook until liquid has evaporated.

Step 3. Assembling.
Shingle the prosciutto on the plastic wrap into a rectangle that's big enough to cover the whole tenderloin. Spread the duxelles evenly and thinly over the prosciutto.

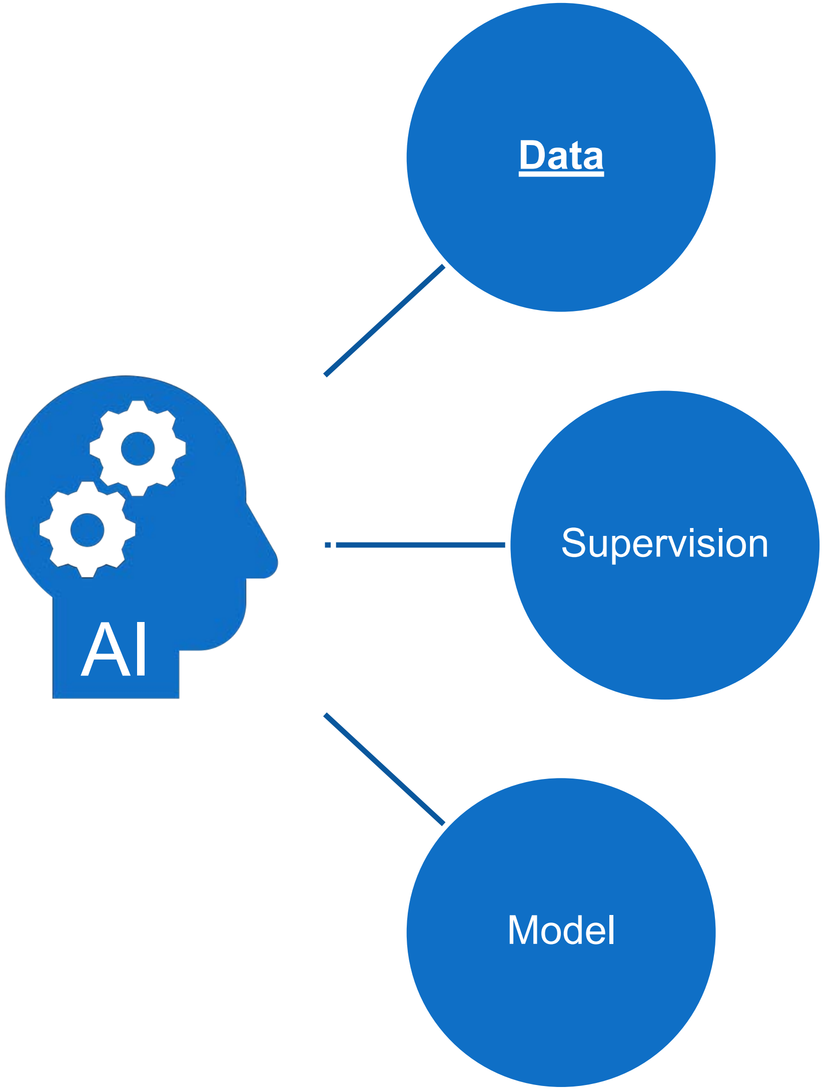
……

https://www.wikihow.com/Main-Page
Koupaee, Mahnaz, and William Yang Wang. "Wikihow: A large scale text summarization dataset." *arXiv preprint arXiv:1810.09305* (2018).

How to Utilize the Knowledge Source?



How to Utilize the Knowledge Source?



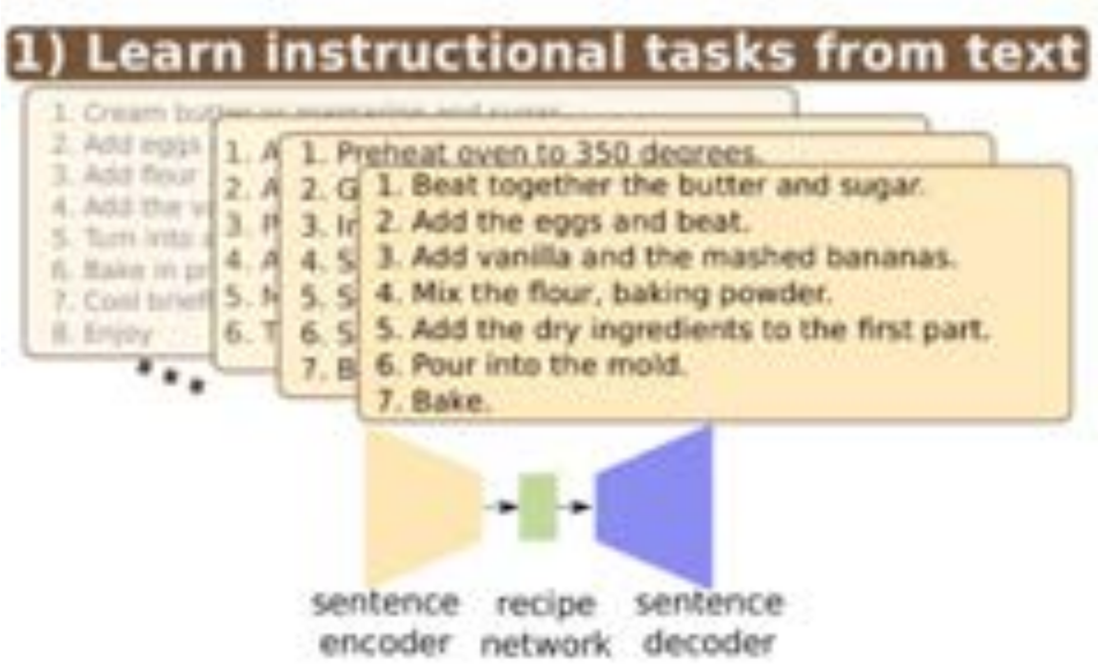
Zero-Shot Anticipation for Instructional Activities

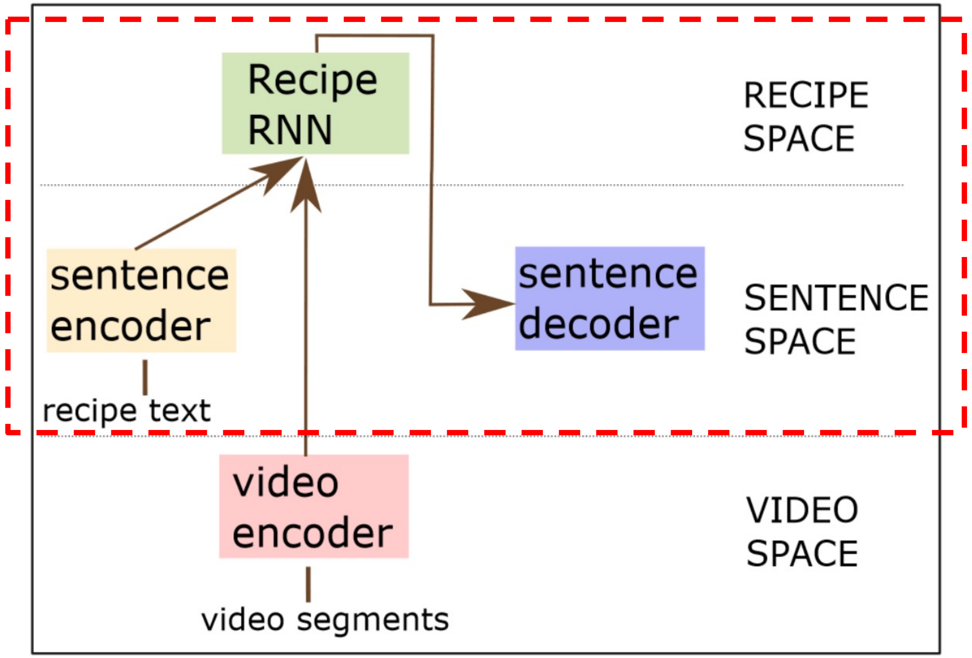
- Key Idea: Obtain training data from knowledge base.

Sener, Fadime, and Angela Yao. "Zero-shot anticipation for instructional activities." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.
Sener, Fadime, Rishabh Saraf, and Angela Yao. "Transferring Knowledge from Text to Video: Zero-Shot Anticipation for Procedural Actions." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

Zero-Shot Anticipation for Instructional Activities

- Sentence encoder encodes a step sentence into a step vector.
- Recipe network is a RNN modeling procedures.
- Sentence decoder decodes step sentences.



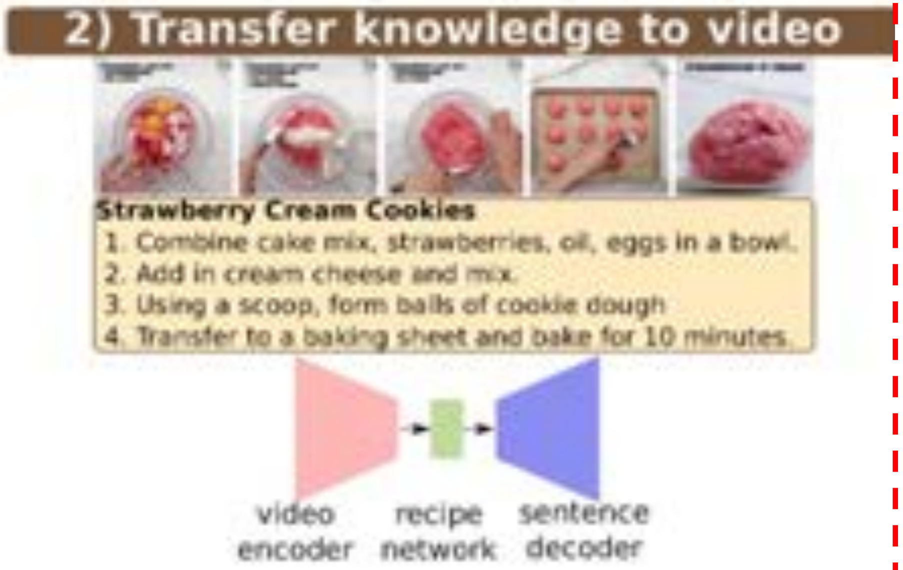


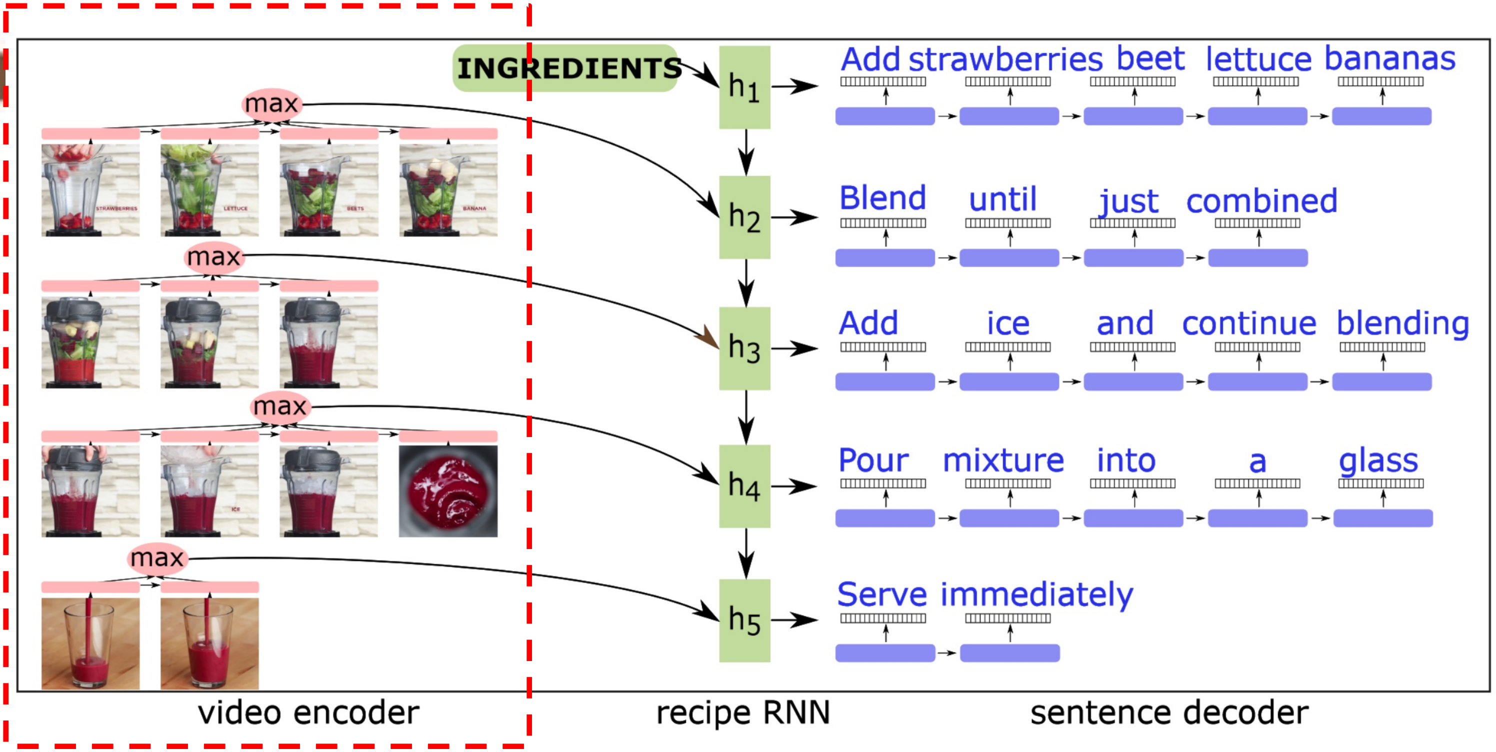
Model Overview

Sener, Fadime, and Angela Yao. "Zero-shot anticipation for instructional activities." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.
Sener, Fadime, Rishabh Saraf, and Angela Yao. "Transferring Knowledge from Text to Video: Zero-Shot Anticipation for Procedural Actions." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

Zero-Shot Anticipation for Instructional Activities

- Only train the video encoder to project video into step vectors with annotated data.



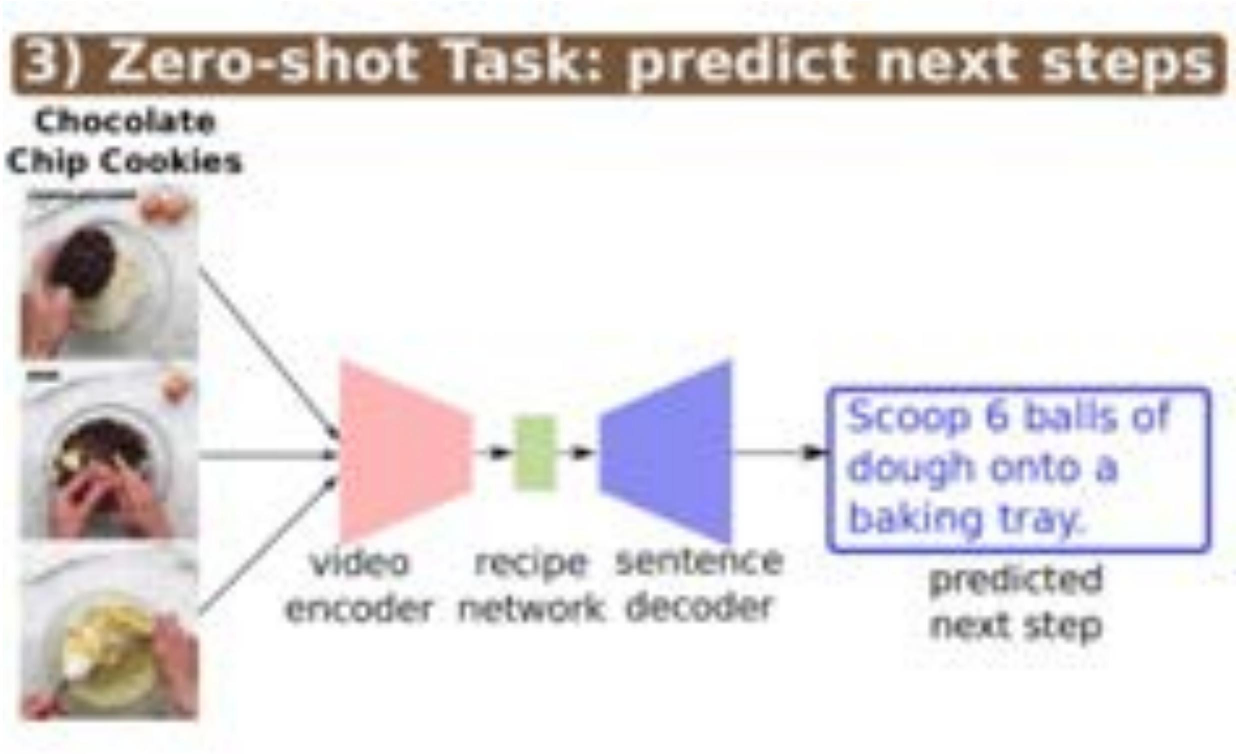


Sener, Fadime, and Angela Yao. "Zero-shot anticipation for instructional activities." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

Sener, Fadime, Rishabh Saraf, and Angela Yao. "Transferring Knowledge from Text to Video: Zero-Shot Anticipation for Procedural Actions." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

Zero-Shot Anticipation for Instructional Activities

- Generalize on new tasks.



Sener, Fadime, and Angela Yao. "Zero-shot anticipation for instructional activities." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.
Sener, Fadime, Rishabh Saraf, and Angela Yao. "Transferring Knowledge from Text to Video: Zero-Shot Anticipation for Procedural Actions." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

Zero-Shot Anticipation for Instructional Activities

- Strong zero-shot performance on the proposed Tasty video dataset

The larger knowledge base used, the better!

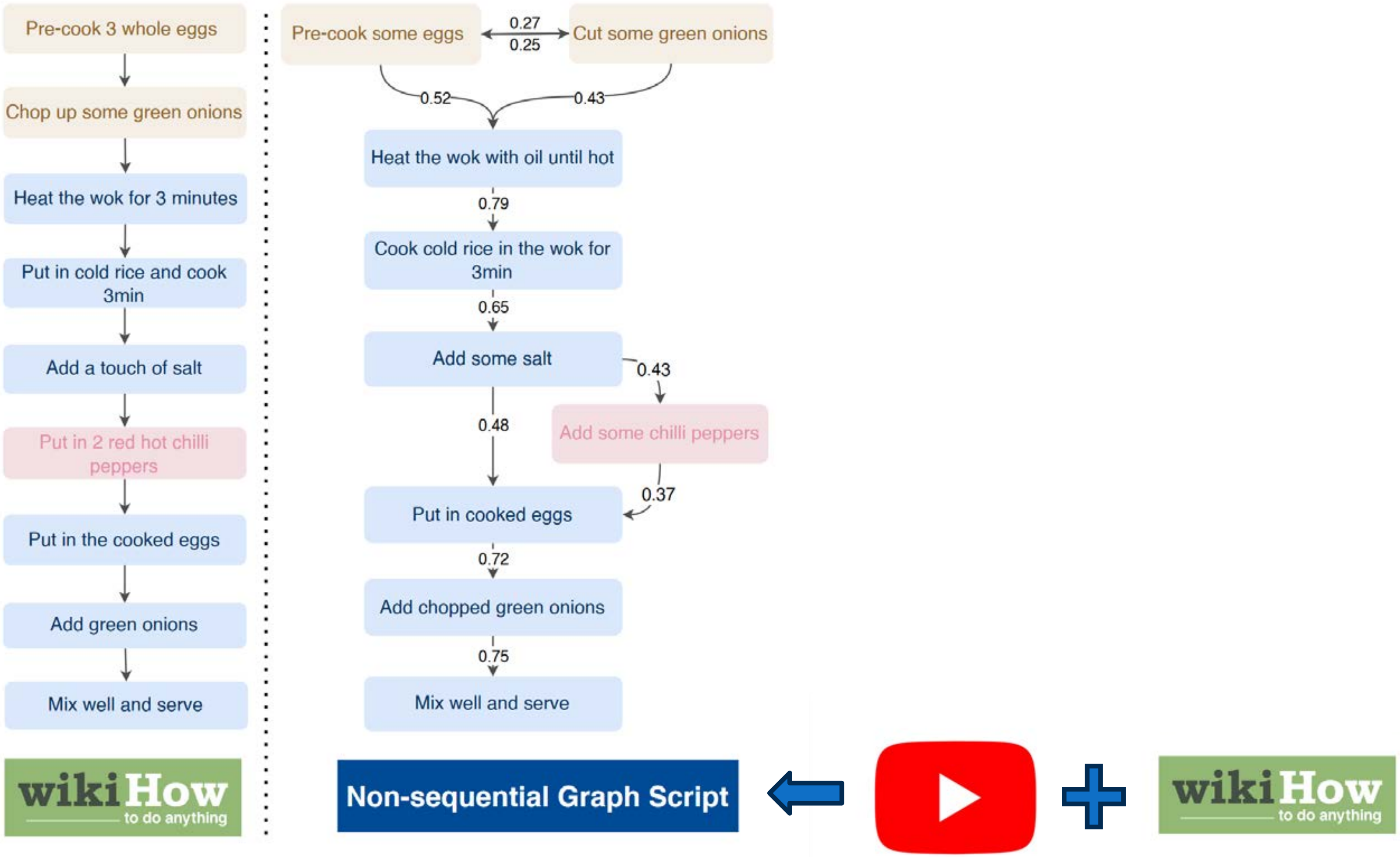
Method	ING	VERBS	BLEU1	BLEU4	METEOR
S2VT [53] (GT)	7.59	19.18	18.03	1.10	9.12
S2VT [53], next (GT)	1.54	10.66	9.14	0.26	5.59
End-to-end [60]	-	-	-	0.54	5.48
Ours Visual (GT)	20.40	19.18	19.05	1.48	11.78
Ours Visual	16.66	17.08	17.59	1.23	11.00
Ours Text (100%)	26.09	27.19	26.78	3.30	17.97
Ours Text (50%)	23.01	24.90	25.05	2.42	16.98
Ours Text (25%)	19.43	23.83	23.54	2.03	16.05
Ours Text (0%)	5.80	9.42	10.58	0.24	6.80
Ours Text noING	9.04	22.00	20.11	0.92	13.07
Ours joint video-text	22.27	23.35	21.75	2.33	14.09

- Limitation
  - Domain is limited to cooking.
  - Rely on annotated data samples for training video encoder.

Sener, Fadime, and Angela Yao. "Zero-shot anticipation for instructional activities." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

Non-Sequential Graph Script Induction via Multimedia Grounding

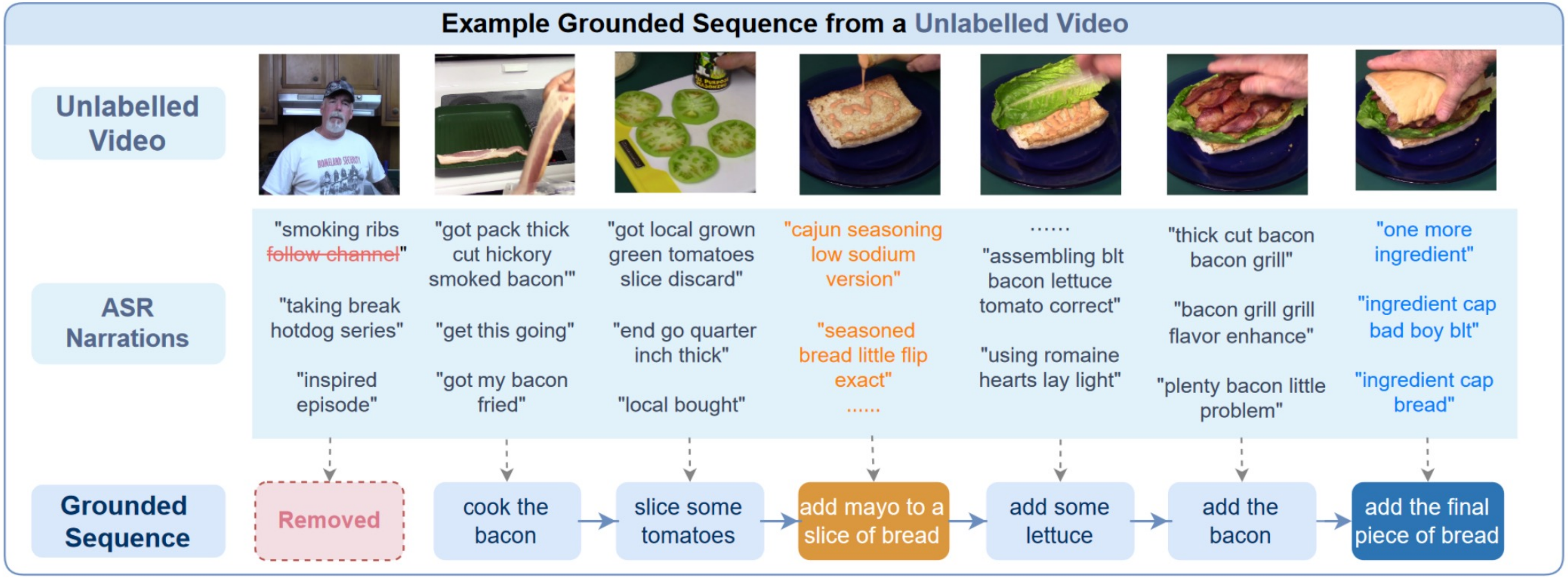
- Key Idea: Obtain non-sequential script by grounding wikiHow steps to video observations.



Zhou, Yu, et al. "Non-Sequential Graph Script Induction via Multimedia Grounding." Proceedings of the Conference of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), 2023

Non-Sequential Graph Script Induction via Multimedia Grounding

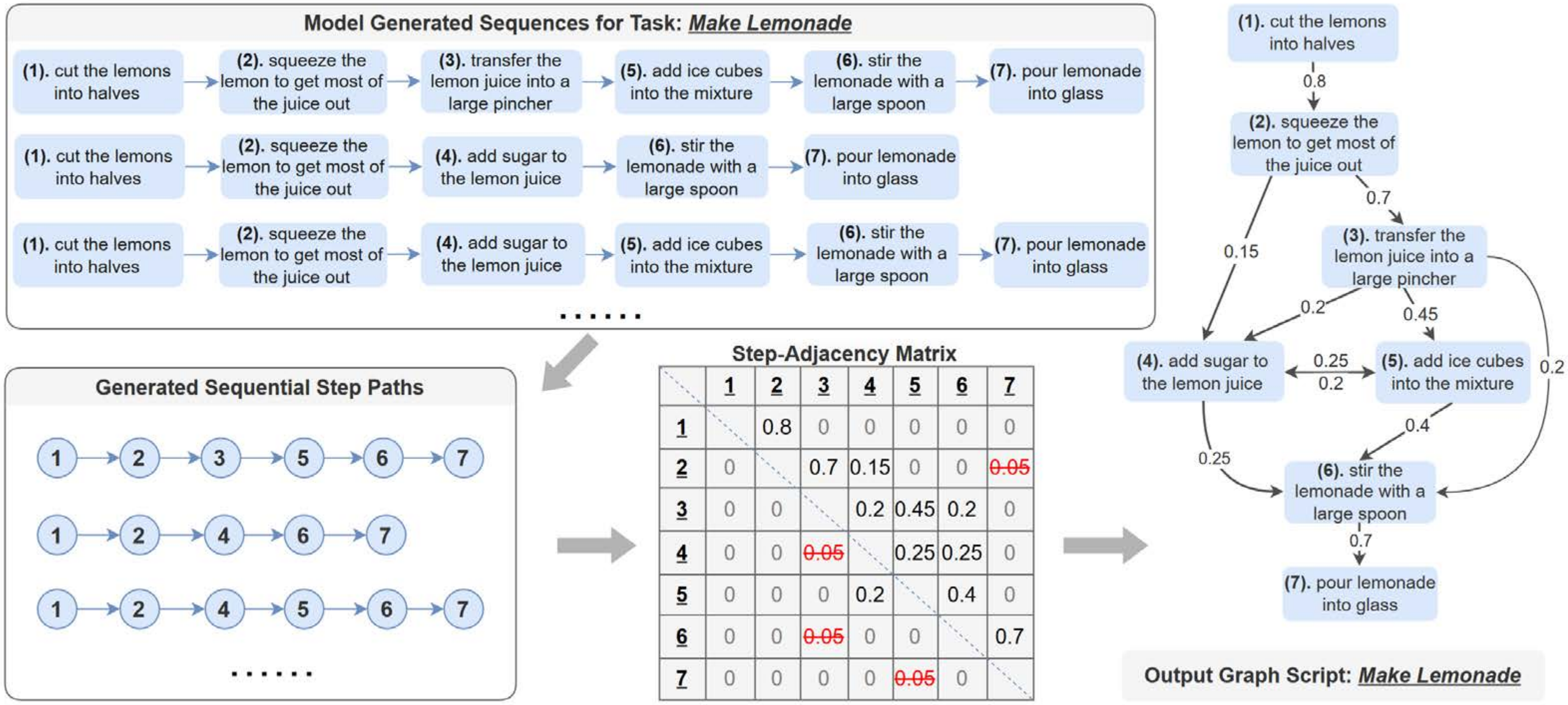
- Video observations contain real-world variance in procedure, which makes wikiHow scripts <u>non-sequential</u>.



Zhou, Yu, et al. "Non-Sequential Graph Script Induction via Multimedia Grounding." Proceedings of the Conference of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), 2023

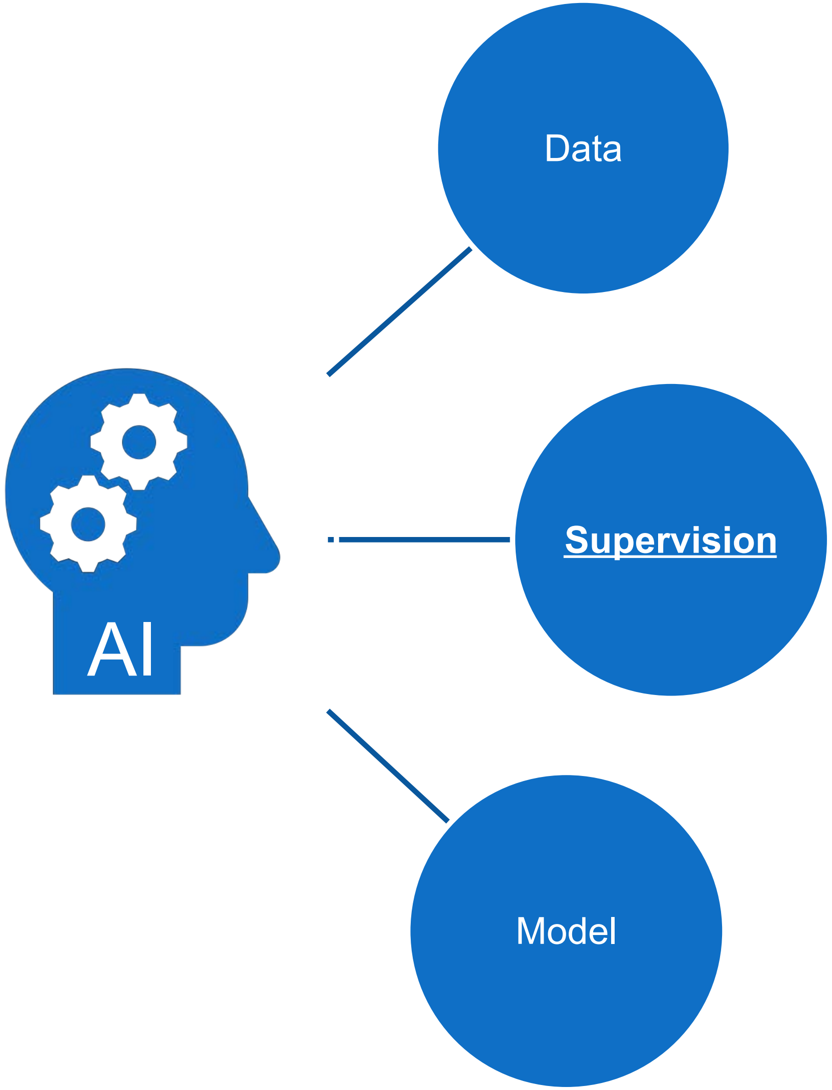
Non-Sequential Graph Script Induction via Multimedia Grounding

- Graphs can be constructed by merging multiple decoded sequence.
- Limitation: closed-vocabulary; text-only graph.



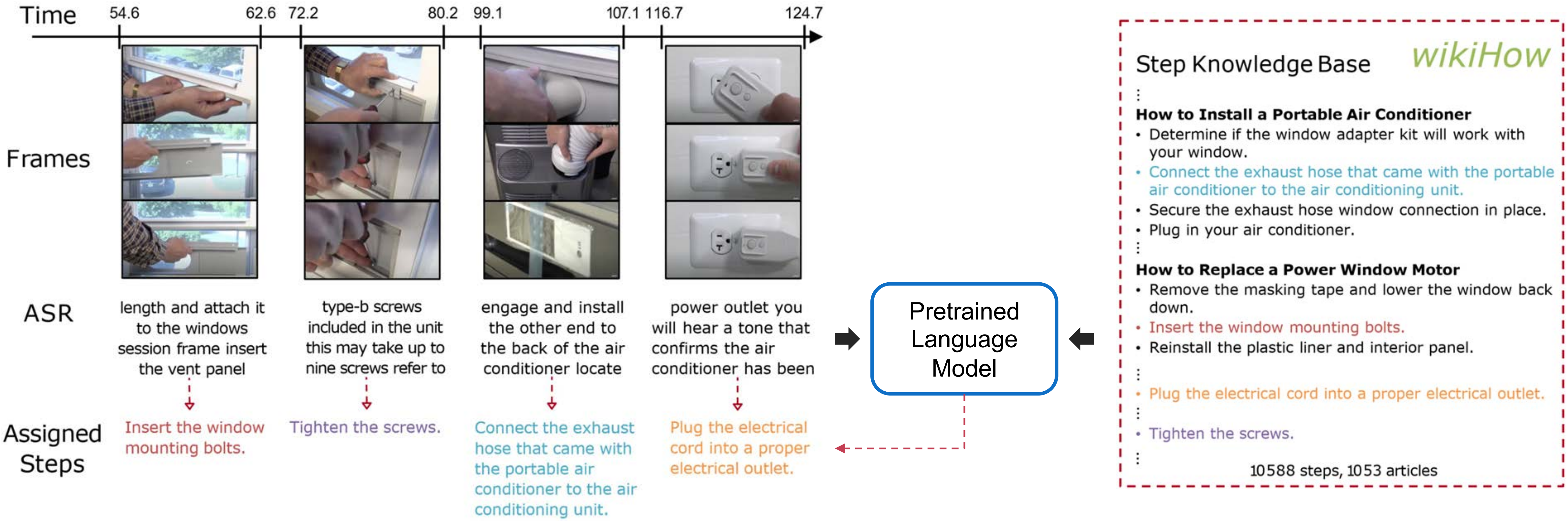
Zhou, Yu, et al. "Non-Sequential Graph Script Induction via Multimedia Grounding." Proceedings of the Conference of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), 2023

How to Utilize the Knowledge Source?



Learning To Recognize Procedural Activities with Distant Supervision

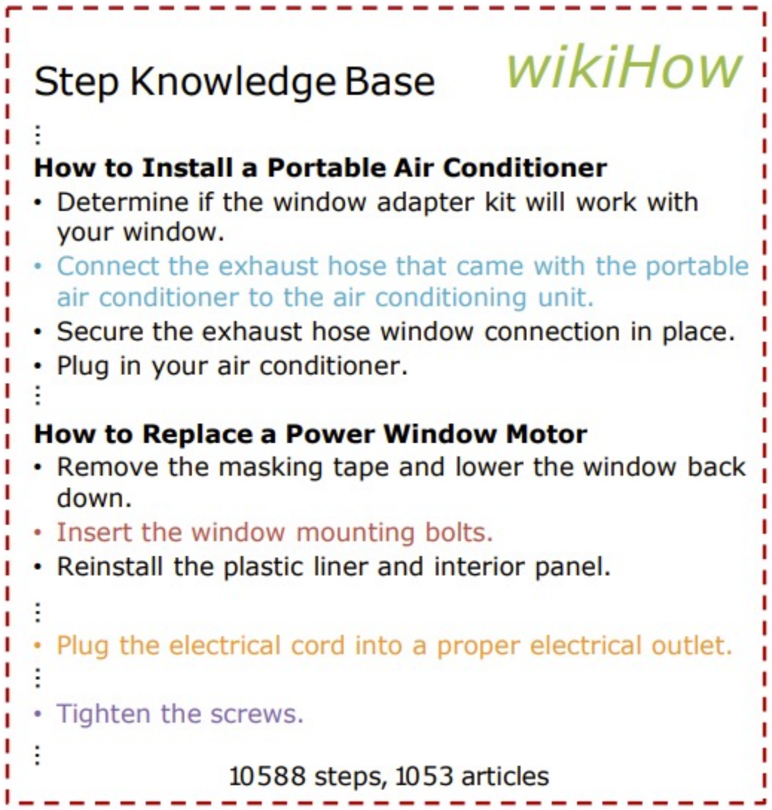
- Key Idea: Leverage pretrained language model to align knowledge base and videos with speech to obtain supervision.



Lin, Xudong, et al. "Learning to recognize procedural activities with distant supervision." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

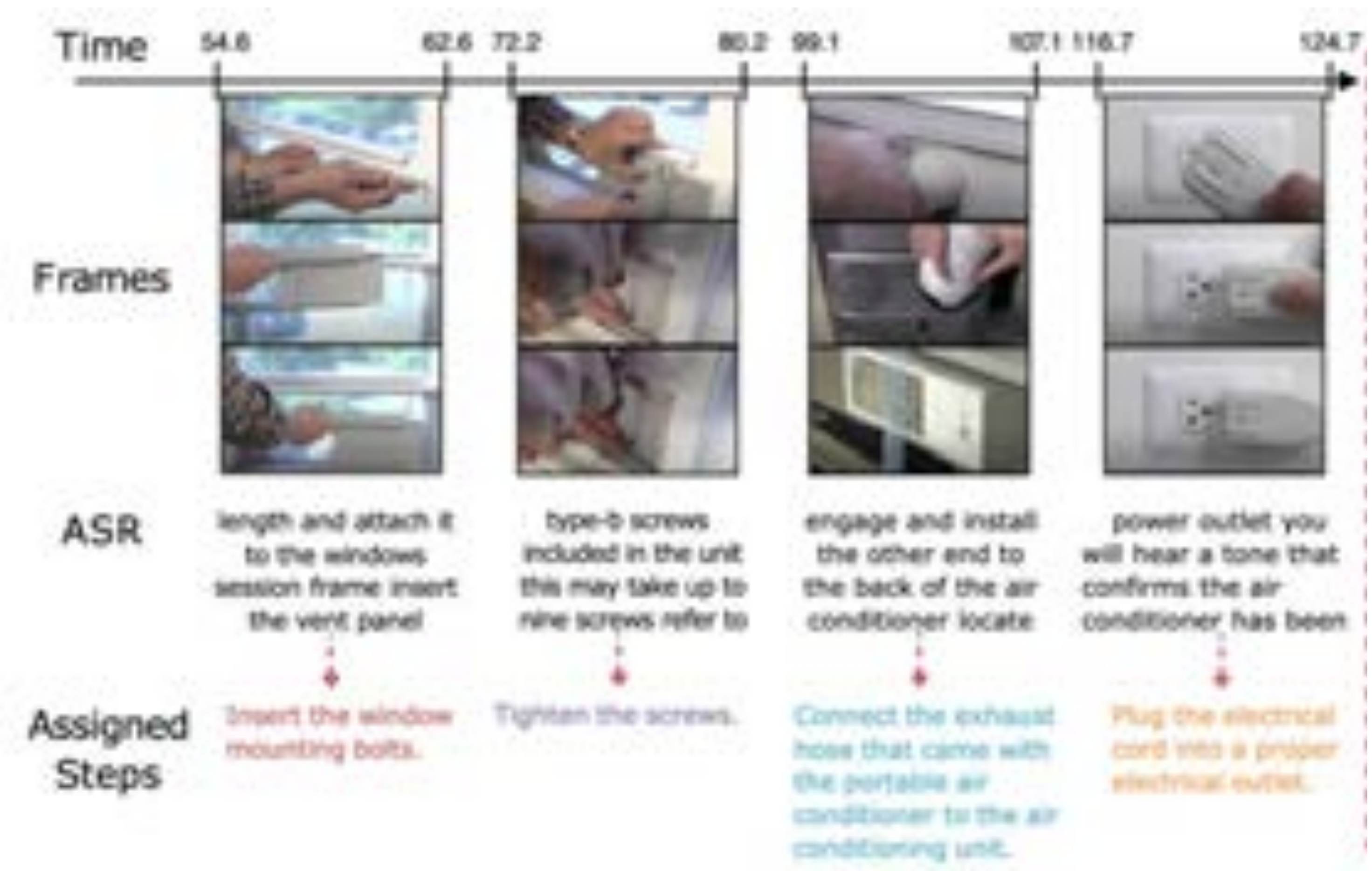
Learning To Recognize Procedural Activities with Distant Supervision

- Step Knowledge Base Construction
  - Use 1053 tasks, each of which has at least 100 examples in the HowTo100M dataset
  - Find the corresponding articles on WikiHow
  - Collect sentences for each step in each of the tasks



Lin, Xudong, et al. "Learning to recognize procedural activities with distant supervision." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

Learning To Recognize Procedural Activities with Distant Supervision

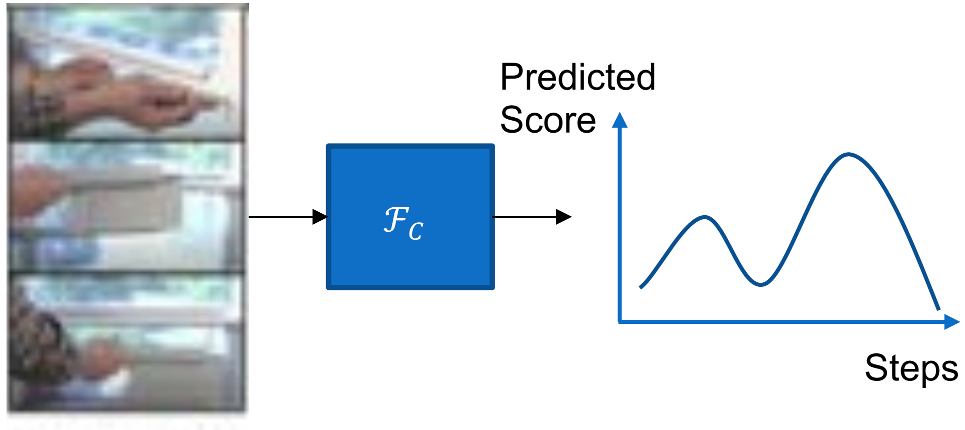


- Distant supervision creation
  - Leverage a pretrained language model to produce embeddings for both **steps** and **ASR sentences** from the video.
  - Then calculate similarity between each ASR sentence and all the steps.

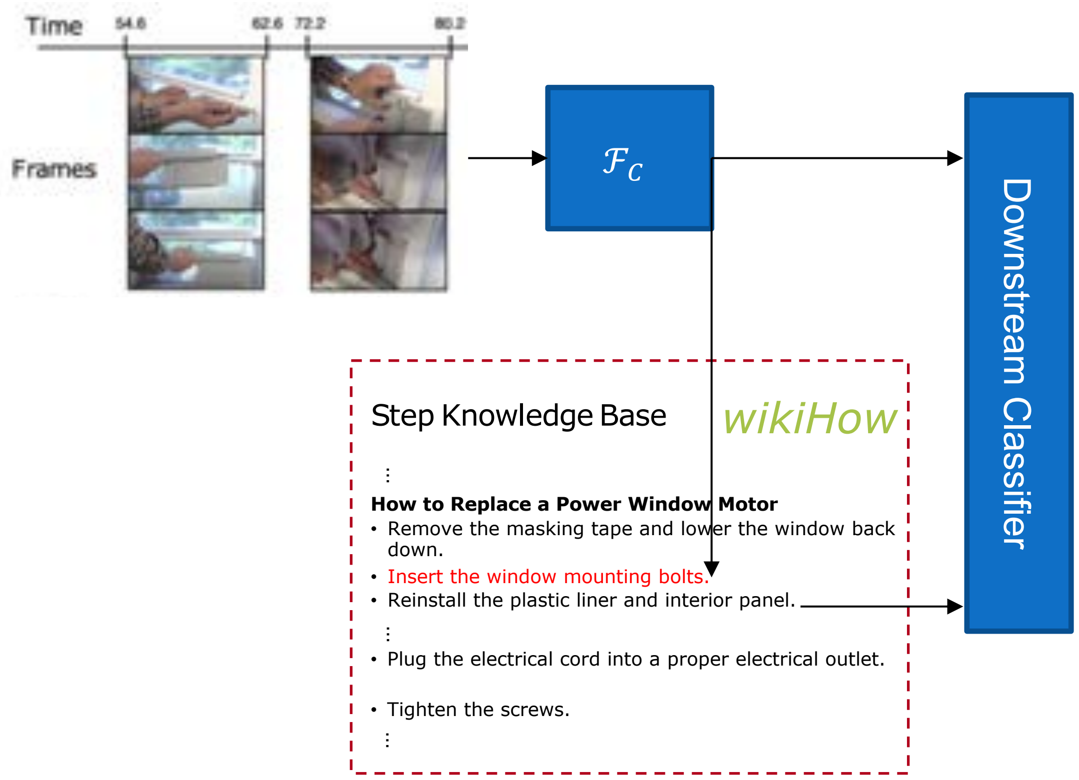
Lin, Xudong, et al. "Learning to recognize procedural activities with distant supervision." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

Learning To Recognize Procedural Activities with Distant Supervision

Pretraining: Learning to align videos and the step knowledge base



Finetuning: Training a classifier with both step-level video representation and ordering information from the knowledge base



Lin, Xudong, et al. "Learning to recognize procedural activities with distant supervision." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

Learning To Recognize Procedural Activities with Distant Supervision

- Step Forecasting on COIN
  - Wikihow Knowledge provides high-quality distant supervision!
  - Ordering information in the knowledge base further helps!

Long-term Model	Segment Model	Pretraining Supervision	Pretraining Dataset	Acc (%)
Basic Transformer	S3D [39]	Unsupervised: MIL-NCE on ASR	HT100M	28.1
Basic Transformer	SlowFast [17]	Supervised: action labels	Kinetics	25.6
Basic Transformer	TimeSformer [8]	Supervised: action labels	Kinetics	34.7
Basic Transformer	TimeSformer [8]	Unsupervised: k-means on ASR	HT100M	34.0
Basic Transformer	**TimeSformer**	**Unsupervised: distant supervision (ours)**	HT100M	**38.2**
Transformer w/ KB Transfer	**TimeSformer**	**Unsupervised: distant supervision (ours)**	HT100M	**39.4**

- The supervision from the wikihow knowledge base also helps

Recognition of procedural activities on COIN

Long-term Model	Segment Model	Pretraining Supervision	Pretraining Dataset	Acc (%)
TSN (RGB+Flow) [57]	Inception [54]	Supervised: action labels	Kinetics	73.4*
Basic Transformer	S3D [39]	Unsupervised: MIL-NCE on ASR	HT100M	70.2*
Basic Transformer	**TimeSformer**	**Unsupervised: distant supervision (ours)**	HT100M	**88.9**
Transformer w/ KB Transfer	**TimeSformer**	**Unsupervised: distant supervision (ours)**	HT100M	**90.0**

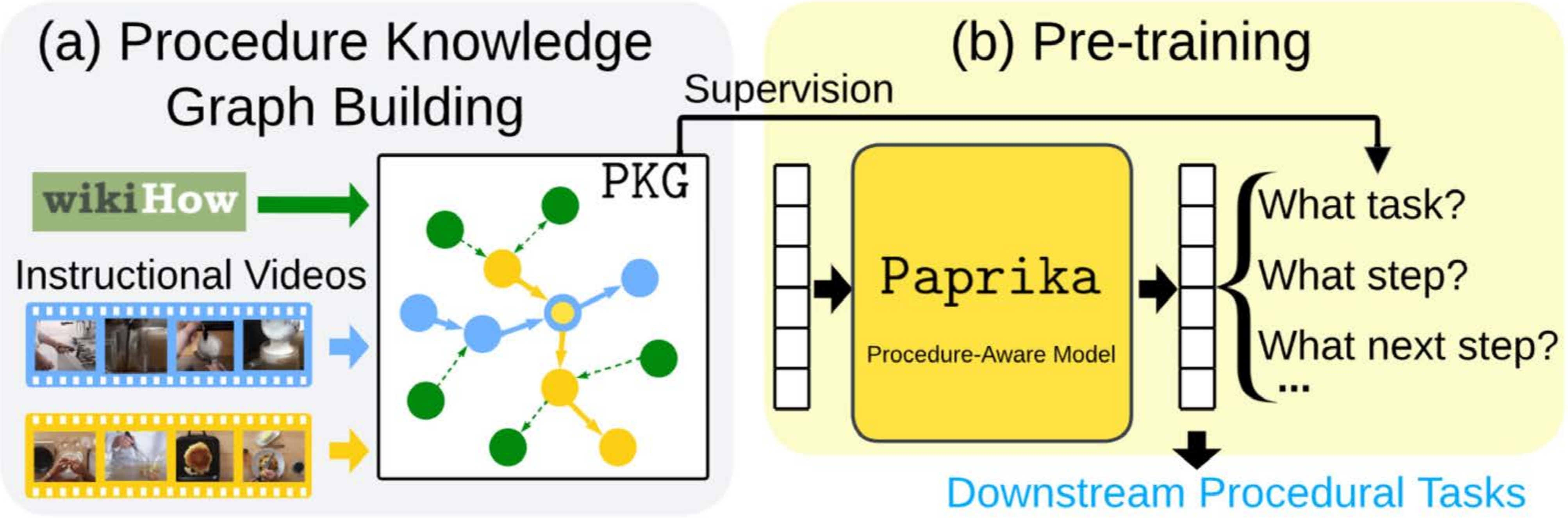
Egocentric video classification

Segment Model	Pretraining Supervision	Pretraining Dataset	Action (%)	Verb (%)	Noun (%)
ViViT-L [6]	Supervised: action labels	Kinetics	44.0	66.4	56.8
TimeSformer [8]	Supervised: action labels	Kinetics	42.3	66.6	54.4
TimeSformer	**Unsupervised: distant supervision (ours)**	HT100M	44.4	67.1	**58.1**

- Limitation: Didn’t employ ordering information in the pretraining model.

Procedure-Aware Pretraining for Instructional Video Understanding

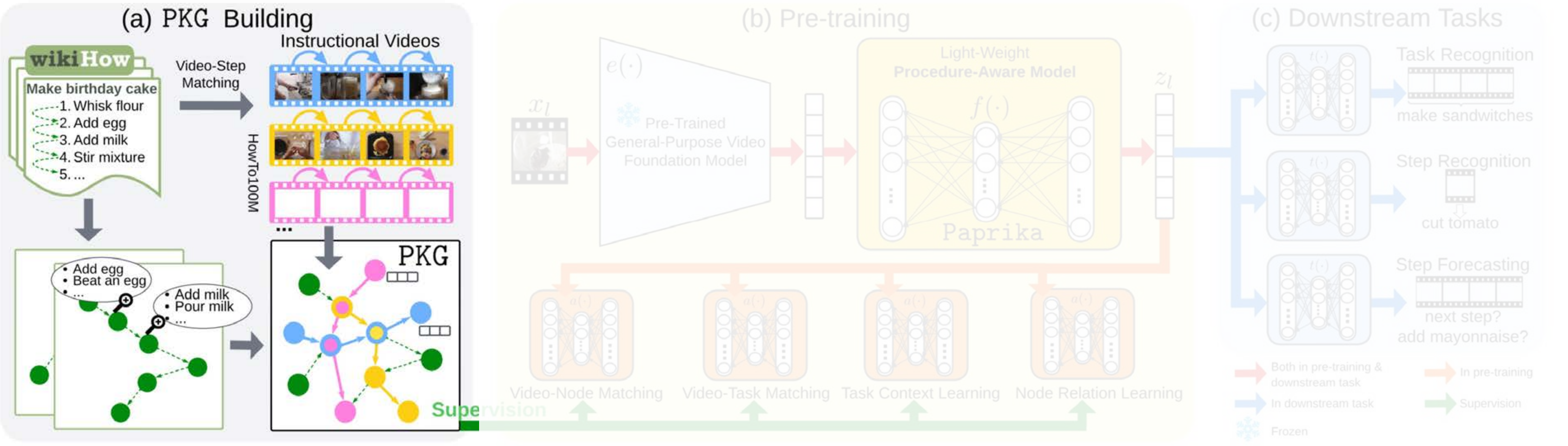
- Key Idea: Construct procedural knowledge graph and then use it to obtain supervision.



Zhou, Honglu, et al. "Procedure-aware pretraining for instructional video understanding." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

Procedure-Aware Pretraining for Instructional Video Understanding

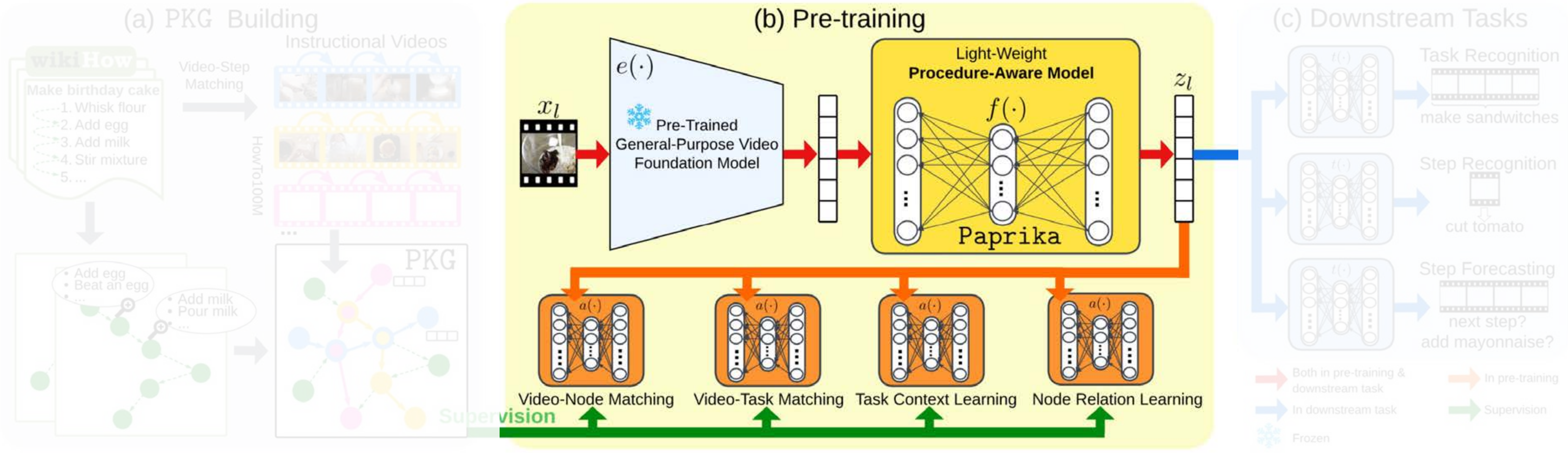
- Construct procedural knowledge graph by grounding wikiHow steps to instructional videos;



Zhou, Honglu, et al. "Procedure-aware pretraining for instructional video understanding." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

Procedure-Aware Pretraining for Instructional Video Understanding

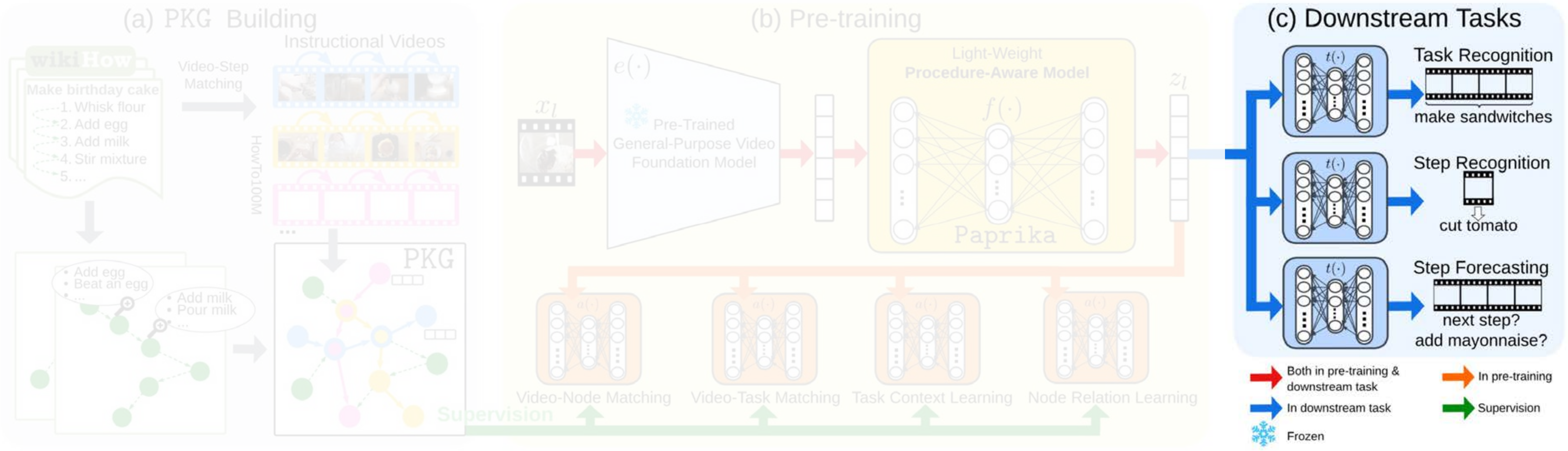
- Use procedural knowledge graph to supervise a procedure-aware model;



Zhou, Honglu, et al. "Procedure-aware pretraining for instructional video understanding." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

Procedure-Aware Pretraining for Instructional Video Understanding

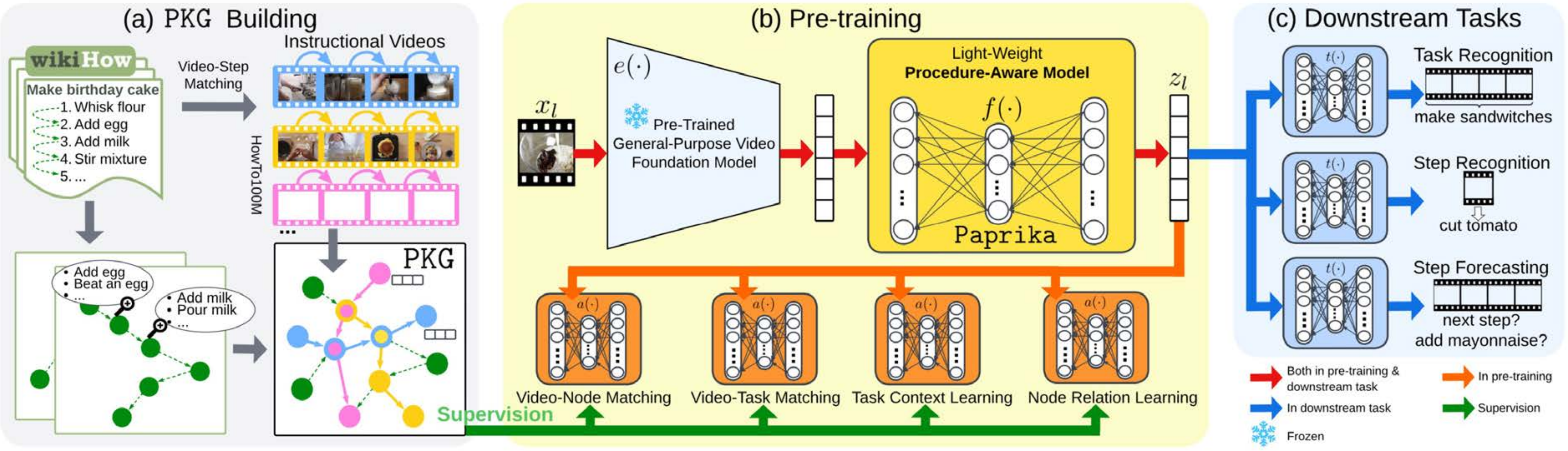
- The representation produced by the procedure-aware model can be directly used for downstream tasks.



Zhou, Honglu, et al. "Procedure-aware pretraining for instructional video understanding." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

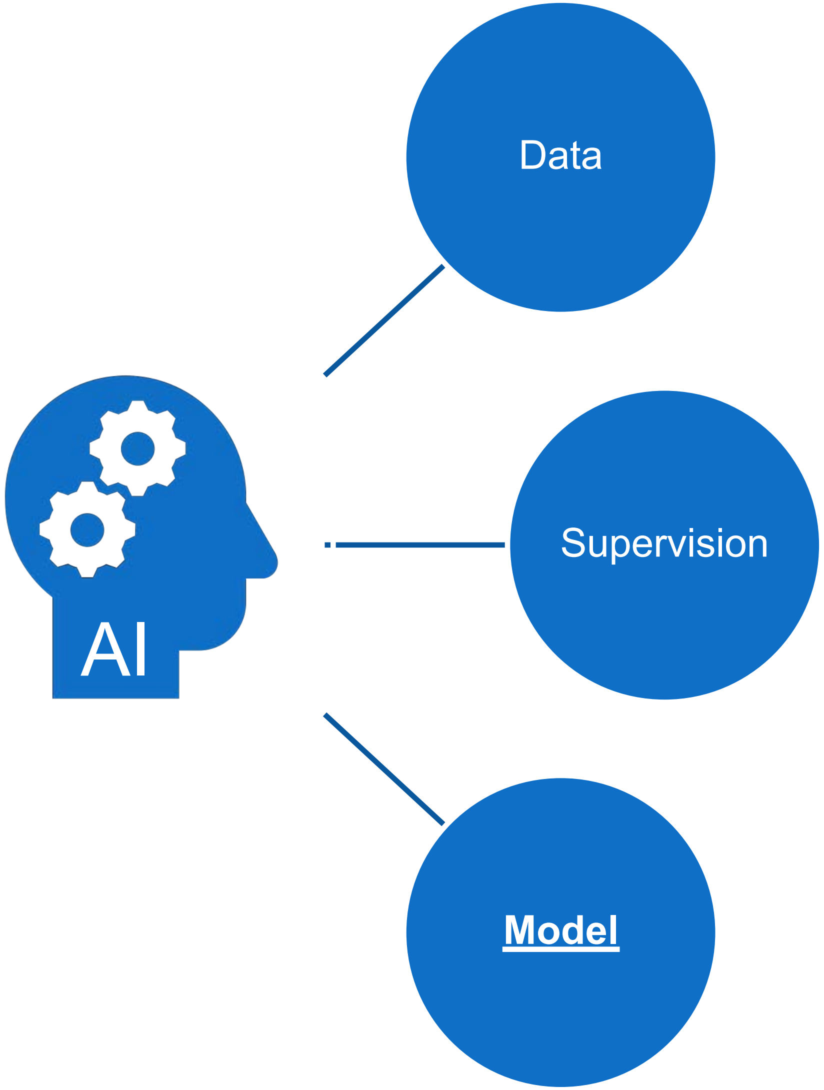
Procedure-Aware Pretraining for Instructional Video Understanding

- Limitation: closed-vocabulary; text-only graph.



Zhou, Honglu, et al. "Procedure-aware pretraining for instructional video understanding." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

How to Utilize the Knowledge Source?



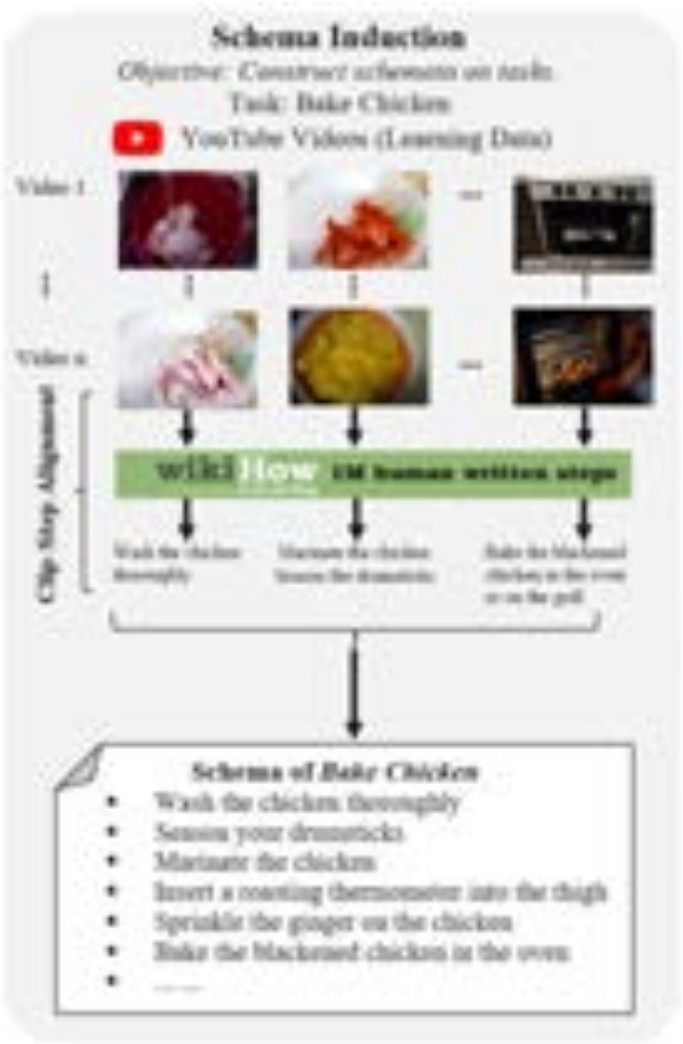
Induce, Edit, Retrieve: Language Grounded Multimodal Schema for Instructional Video Retrieval

- Key Idea: Learning multimodal schema to represent procedural knowledge.

Yang, Yue, et al. "Induce, edit, retrieve: Language grounded multimodal schema for instructional video retrieval." *arXiv preprint arXiv:2111.09276* (2021).

Induce, Edit, Retrieve: Language Grounded Multimodal Schema for Instructional Video Retrieval

- Schema Induction
  - For each task, find corresponding steps from wikiHow and videos from YouTube.
  - For each segment in each video, retrieve most relevant steps with existing video-text matching models.



Yang, Yue, et al. "Induce, edit, retrieve: Language grounded multimodal schema for instructional video retrieval." *arXiv preprint arXiv:2111.09276* (2021).

Induce, Edit, Retrieve: Language Grounded Multimodal Schema for Instructional Video Retrieval

- Schema Editing
  - For an unseen task, find the most similar seen task based on both textual and visual similarity.

Yang, Yue, et al. "Induce, edit, retrieve: Language grounded multimodal schema for instructional video retrieval." *arXiv preprint arXiv:2111.09276* (2021).

Induce, Edit, Retrieve: Language Grounded Multimodal Schema for Instructional Video Retrieval

- Schema Editing
  - For an unseen task, find the most similar seen task based on both textual and visual similarity
  - Replace object towards the unseen task.

Yang, Yue, et al. "Induce, edit, retrieve: Language grounded multimodal schema for instructional video retrieval." *arXiv preprint arXiv:2111.09276* (2021).

Induce, Edit, Retrieve: Language Grounded Multimodal Schema for Instructional Video Retrieval

- Schema Editing
  - For an unseen task, find the most similar seen task based on both textual and visual similarity
  - Replace object towards the unseen task.
  - Delete steps that are not relevant in the new task with a pretrained language model.

Yang, Yue, et al. "Induce, edit, retrieve: Language grounded multimodal schema for instructional video retrieval." *arXiv preprint arXiv:2111.09276* (2021).

Induce, Edit, Retrieve: Language Grounded Multimodal Schema for Instructional Video Retrieval

- Schema Editing
  - For an unseen task, find the most similar seen task based on both textual and visual similarity
  - Replace object towards the unseen task.
  - Delete steps that are not relevant in the new task with a pretrained language model.
  - Replace tokens least likely associated with the task in each step by prompting a pretrained language model.

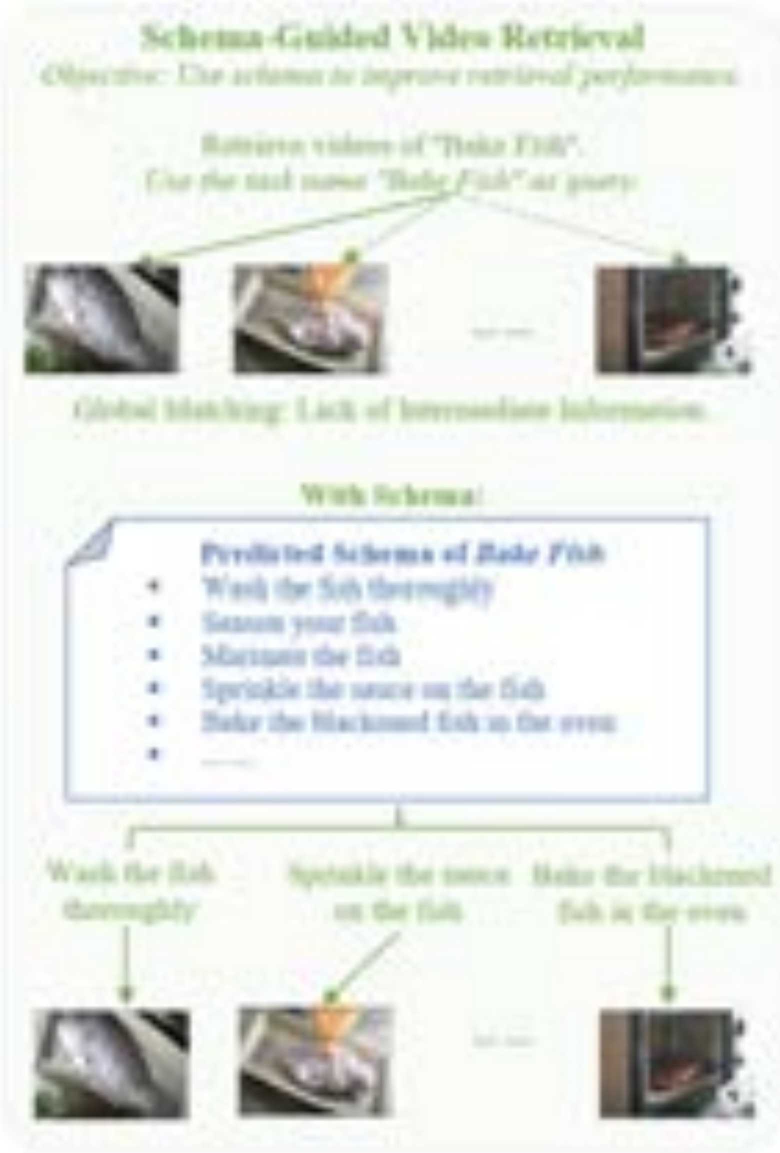
Yang, Yue, et al. "Induce, edit, retrieve: Language grounded multimodal schema for instructional video retrieval." *arXiv preprint arXiv:2111.09276* (2021).

Induce, Edit, Retrieve: Language Grounded Multimodal Schema for Instructional Video Retrieval

- The learned schema provides step-level information to better retrieve videos.

	Method	Howto-GEN P@1↑	Howto-GEN R@5↑	Howto-GEN R@10↑	Howto-GEN Med r↓	Howto-GEN MRR↑	COIN P@1↑	COIN R@5↑	COIN R@10↑	COIN Med r↓	COIN MRR↑	Youcook2 P@1↑	Youcook2 R@5↑	Youcook2 R@10↑	Youcook2 Med r↓	Youcook2 MRR↑
	MIL-NCE [31]	45.2	31.0	43.1	15.0	.198	48.3	37.1	52.8	9.5	.227	27.0	18.2	26.5	32.0	.126
Step Aggregation	T5 [30]	44.0	29.9	41.0	19.0	.190	46.1	35.3	50.7	10.0	.219	21.3	16.0	24.7	61.5	.108
	GPT-2 [39]	46.0	31.5	43.3	16.0	.200	48.9	39.2	53.4	8.0	.233	31.5	19.0	27.3	44.5	.130
	GPT-3 [2]	49.3	33.3	45.7	13.0	.211	53.3	42.1	59.0	8.0	.252	37.1	22.4	34.6	27.0	.160
	GOSC [30]	54.7	37.0	49.8	11.0	.231	53.9	41.6	55.1	8.0	.248	30.3	20.7	34.8	28.0	.146
	wikiHow	51.9	35.4	47.8	11.0	.222	53.9	40.8	56.1	7.0	.246	31.5	21.0	34.2	24.5	.149
	IER (Ours)	54.4	37.3	50.1	10.0	.231	**57.2**	42.2	57.8	**7.0**	.256	**41.6**	**25.8**	**38.8**	**20.0**	**.175**
	IER³ (Ours)	**55.0**	**37.4**	**50.6**	**10.0**	**.234**	56.1	**42.3**	**59.1**	8.0	**.258**	40.4	25.1	**38.8**	**20.0**	.172
	Oracle	56.5	38.0	50.8	10.0	.237	60.0	43.4	59.3	7.0	.262	52.8	33.5	47.1	14.0	.215

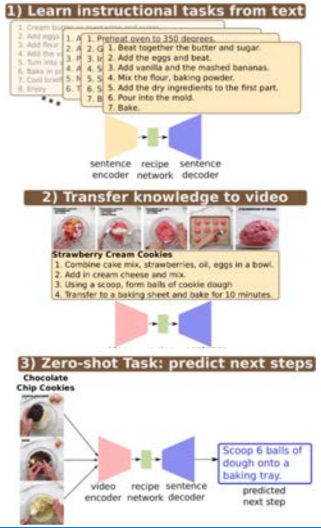
Even comparable with oracle (using manual step annotation for each query)

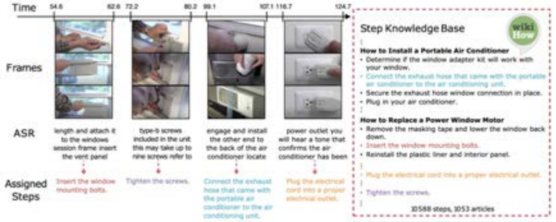


- Limitation
  - Schema is restricted to step sequence without considering graph structures, e.g., optional/exchangeable steps.
  - Only evaluated on text-video retrieval.

Yang, Yue, et al. "Induce, edit, retrieve: Language grounded multimodal schema for instructional video retrieval." *arXiv preprint arXiv:2111.09276* (2021).

Summary of Methods Using Explicit Knowledge





	Sener & Yao ICCV 2019	Lin et al. CVPR 2022	Yang et al. 2021
Knowledge as data	✓		
Knowledge as supervision		✓	
Knowledge for model		✓	✓
Sequential knowledge	✓	✓	✓
Multimodal knowledge			✓

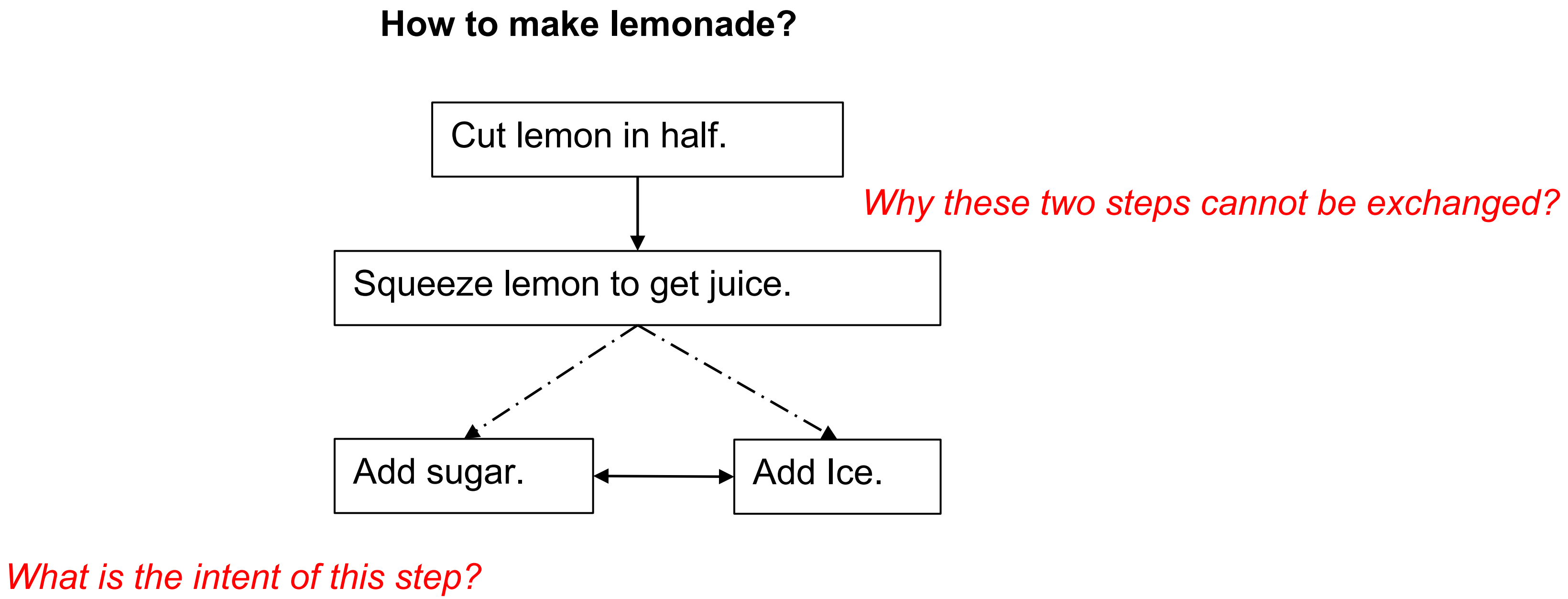
Summary of Methods Using Explicit Knowledge

	Sener & Yao ICCV 2019	Lin et al. CVPR 2022	Yang et al. 2021	Zhou et al. ACL 2023	Zhou et al. CVPR 2023
Knowledge as data	✓			✓	
Knowledge as supervision		✓			✓
Knowledge for model		✓	✓		
Sequential knowledge	✓	✓	✓		
Multimodal knowledge			✓		

- What is next?

Future Challenge: Interpret but Not Memorize

- Do models understand **why** the steps are ordered as in the knowledge base?



Open Problem: Open-vocabulary

- Can the knowledge be automatically extended to open-vocabulary?
  - **Generalize to new tasks;**
  - **Discover new steps and add them in the knowledge base…**

How to make lemonade?

Cut lemon in half.

↓

Squeeze lemon to get juice.

↙ ↘

Add sugar. ↔ Add Ice.

↓

Generalize to new tasks:
Make Berrynade

New step
discovered from
video: Carbonate

https://www.youtube.com/watch?v=GaC14YpDJHw

Agenda

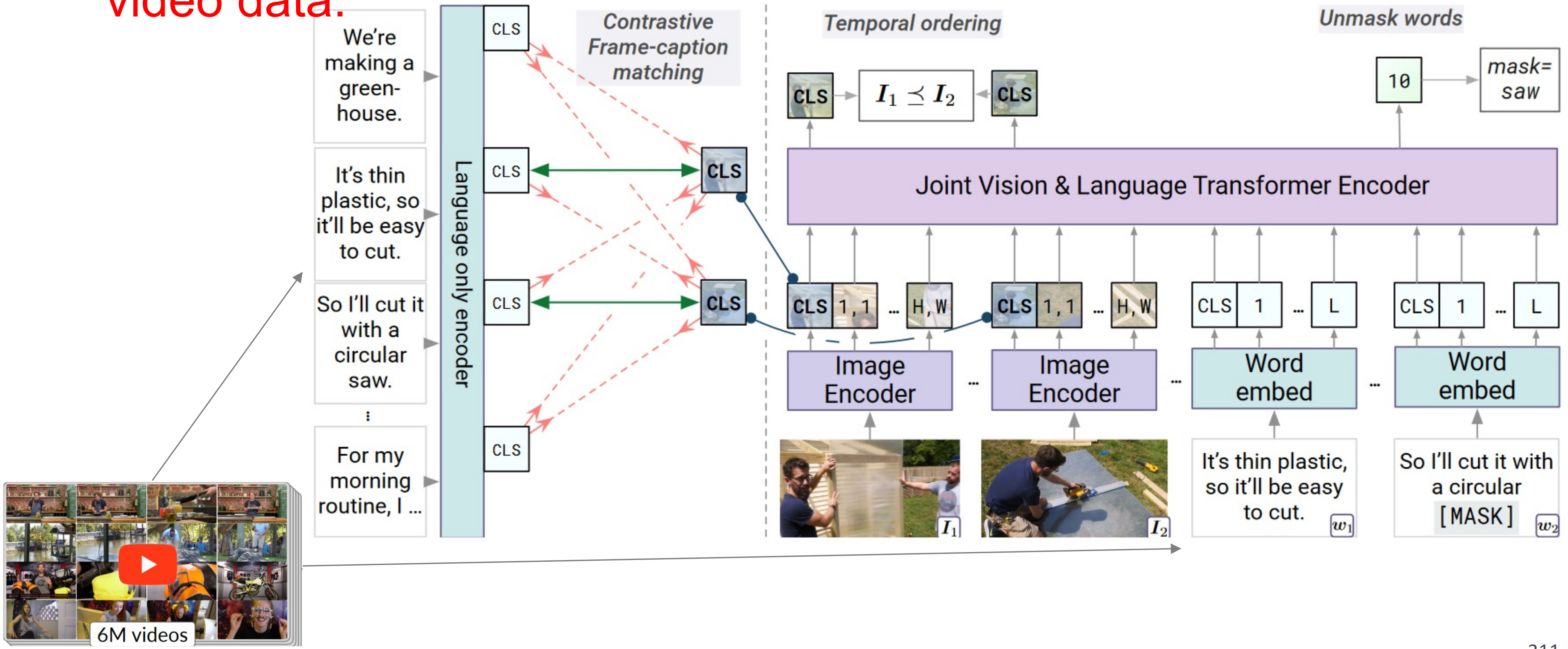
- Explicit Knowledge Source: Learning with the help of external knowledge
- Implicit Knowledge Source: Learning procedural knowledge from data

(Massive)
Data

Model

MERLOT: Multimodal Neural Script Knowledge Models

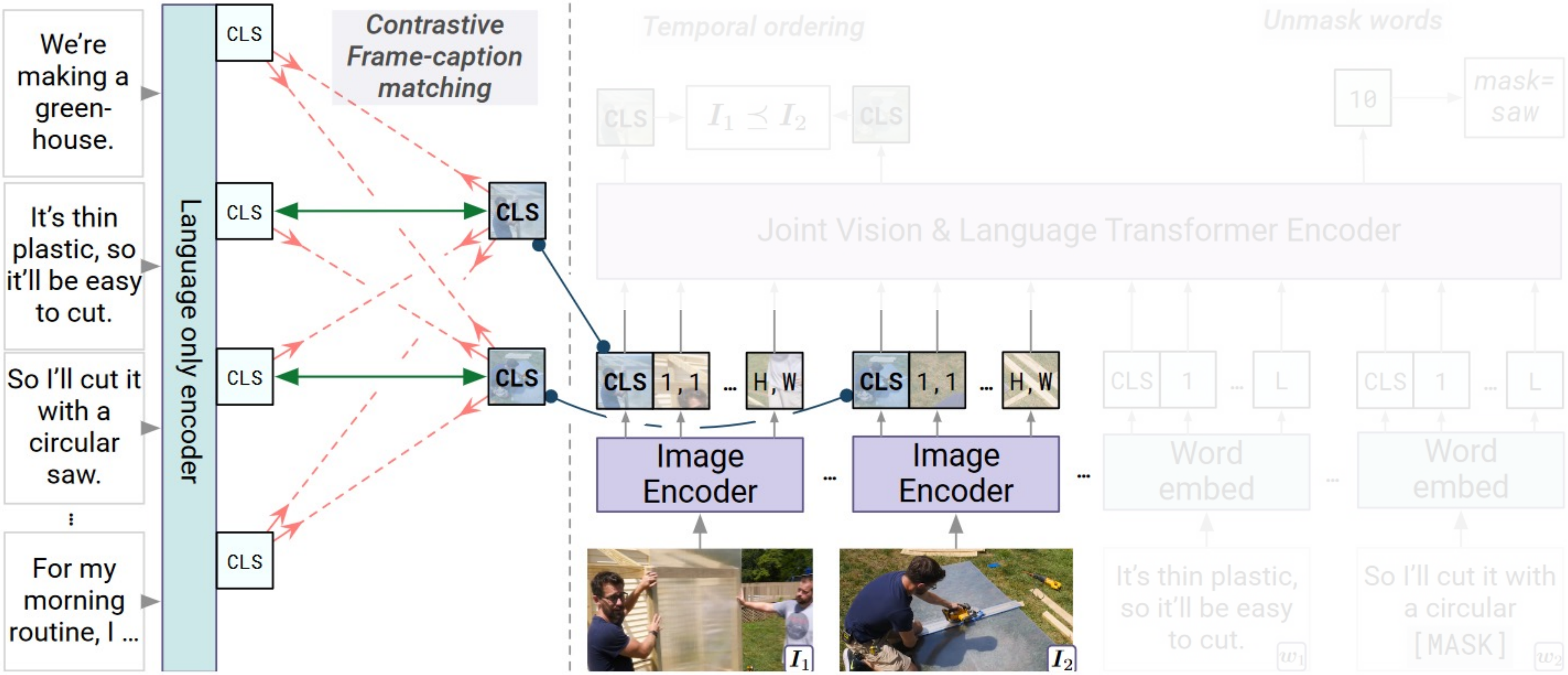
- Key Idea: Learning temporal reasoning ability through massive video data.



Zellers, Rowan, et al. "Merlot: Multimodal neural script knowledge models." *Advances in Neural Information Processing Systems* 34 (2021): 23634-23651.

MERLOT: Multimodal Neural Script Knowledge Models

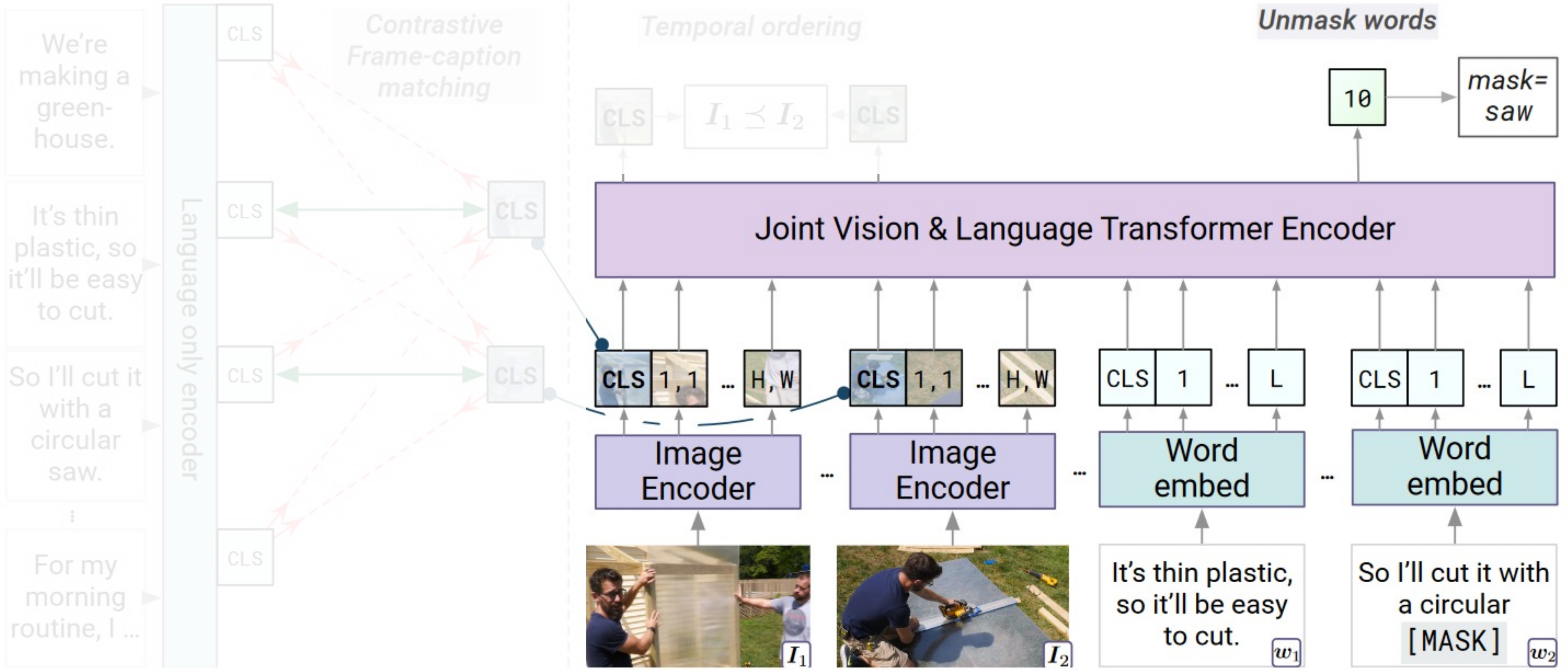
- Objective 1: Alignment between frame representations and text representations



Zellers, Rowan, et al. "Merlot: Multimodal neural script knowledge models." *Advances in Neural Information Processing Systems* 34 (2021): 23634-23651.

MERLOT: Multimodal Neural Script Knowledge Models

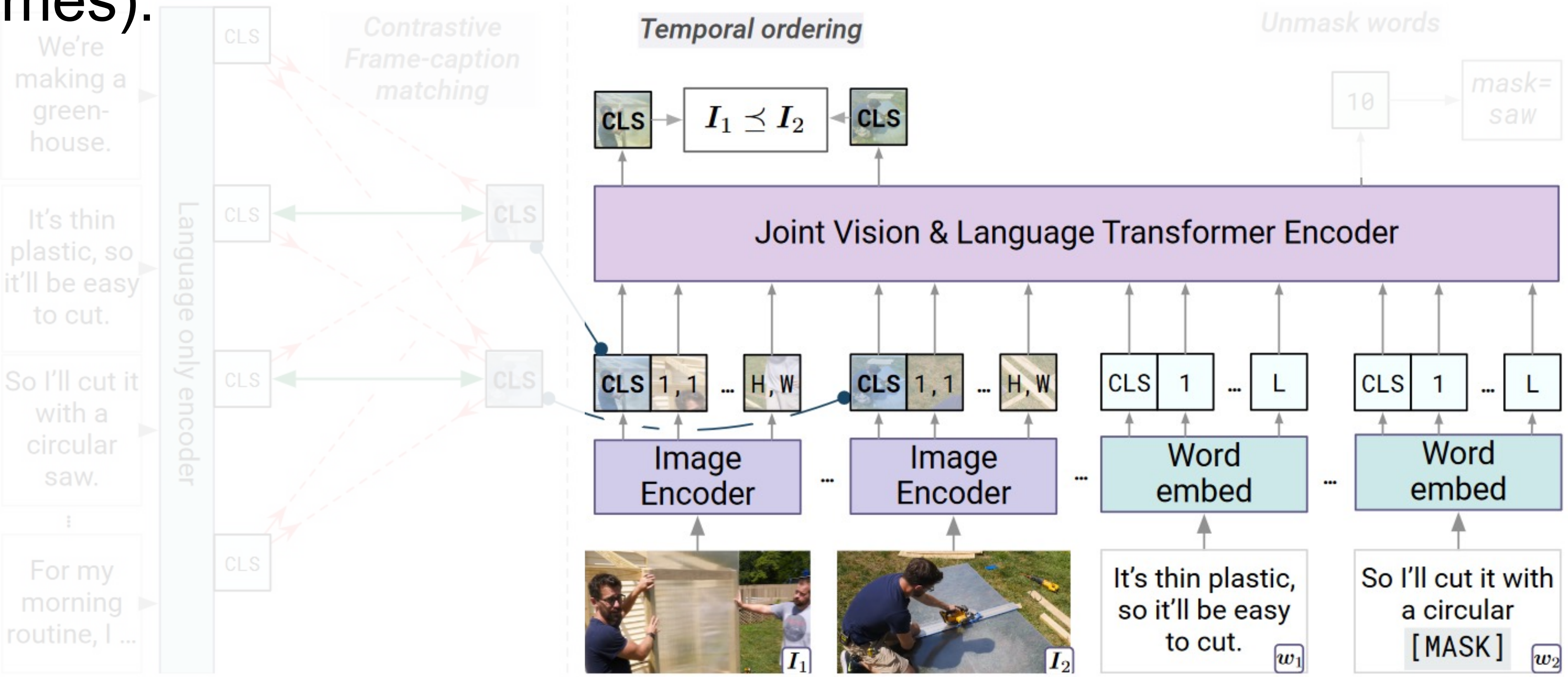
- Objective 2: Masked Token Modeling.



Zellers, Rowan, et al. "Merlot: Multimodal neural script knowledge models." *Advances in Neural Information Processing Systems* 34 (2021): 23634-23651.

MERLOT: Multimodal Neural Script Knowledge Models

- Objective 3: Temporal Ordering (Binary classification between each pair of frames).



Zellers, Rowan, et al. "Merlot: Multimodal neural script knowledge models." *Advances in Neural Information Processing Systems* 34 (2021): 23634-23651.

MERLOT: Multimodal Neural Script Knowledge Models

- The model learns strong temporal reasoning ability and joint video-language reasoning ability.

Ordering Images from Visual Stories

	Spearman (↑)	Pairwise acc (↑)	Distance (↓)
CLIP [89]	.609	78.7	.638
UNITER [22]	.545	75.2	.745
MERLOT	**.733**	**84.5**	**.498**

State-of-the-art over various video-language tasks

Tasks	Split	Vid. Length	ActBERT [127]	ClipBERT$_{8x2}$ [67]	SOTA	MERLOT
MSRVTT-QA	Test	Short	-	37.4	41.5 [118]	**43.1**
MSR-VTT-MC	Test	Short	88.2	-	88.2 [127]	**90.9**
TGIF-Action	Test	Short	-	82.8	82.8 [67]	**94.0**
TGIF-Transition	Test	Short	-	87.8	87.8 [67]	**96.2**
TGIF-Frame QA	Test	Short	-	60.3	60.3 [67]	**69.5**
LSMDC-FiB QA	Test	Short	48.6	-	48.6 [127]	**52.9**
LSMDC-MC	Test	Short	-	-	73.5 [121]	**81.7**
ActivityNetQA	Test	Long	-	-	38.9 [118]	**41.4**
Drama-QA	Val	Long	-	-	81.0 [56]	**81.4**
TVQA	Test	Long	-	-	76.2 [56]	**78.7**
TVQA+	Test	Long	-	-	76.2 [56]	**80.9**
VLEP	Test	Long	-	-	67.5 [66]	**68.4**

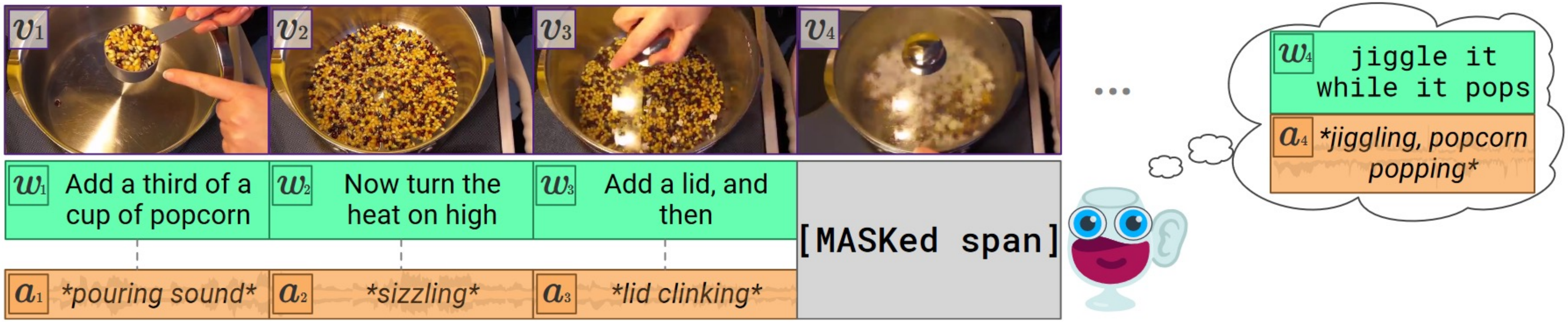
Predict future event given historical videos

- Limitation: short temporal span; importance of the temporal ordering loss is unclear.

Zellers, Rowan, et al. "Merlot: Multimodal neural script knowledge models." *Advances in Neural Information Processing Systems* 34 (2021): 23634-23651.

MERLOT Reserve: Neural Script Knowledge through Vision and Language and Sound

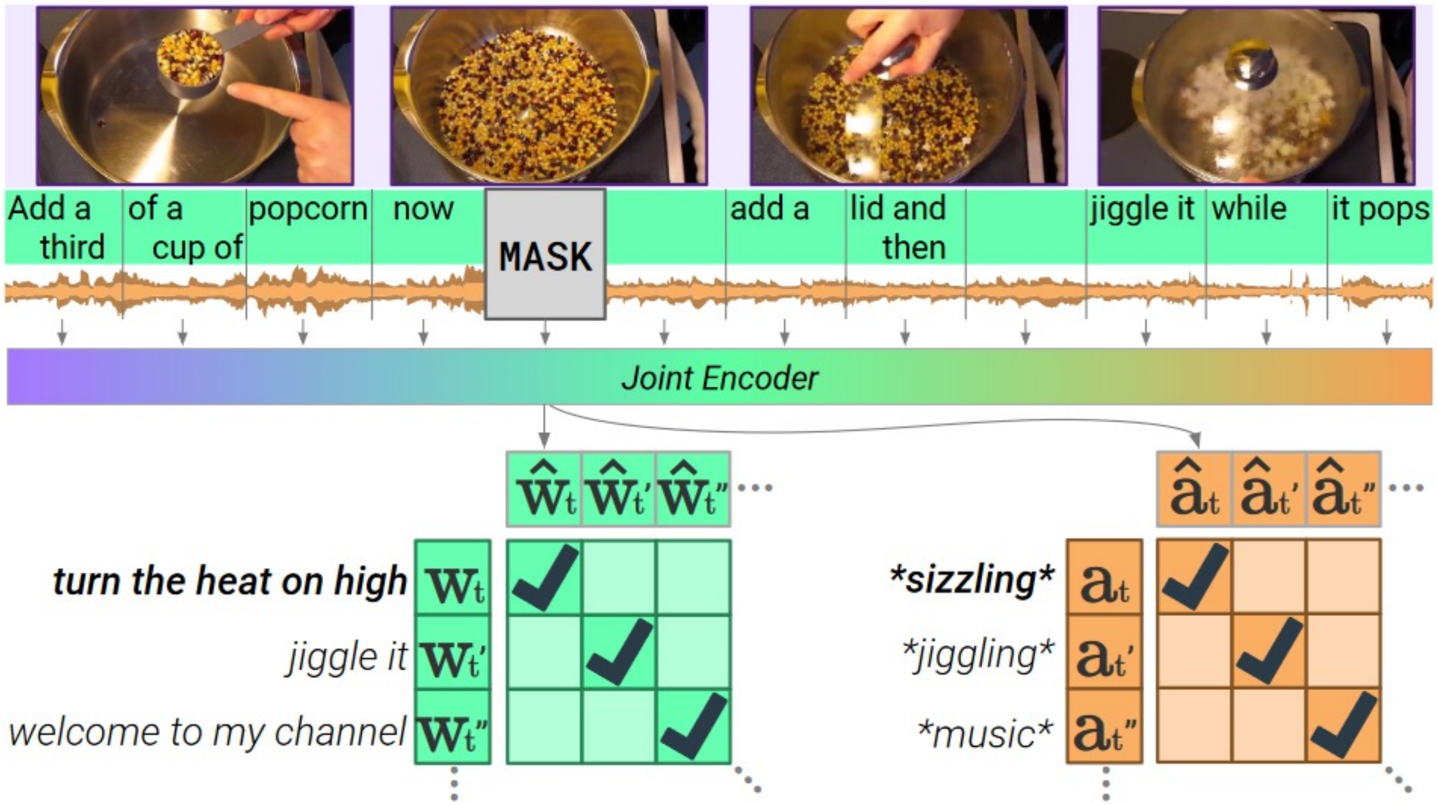
- Key Idea: Jointly learn script knowledge with video, language and audio.



Zellers, Rowan, et al. "Merlot reserve: Neural script knowledge through vision and language and sound." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

MERLOT Reserve: Neural Script Knowledge through Vision and Language and Sound

- Key objective design: <u>contrastive loss</u> between predicted and actual representation of the masked audio/text



Zellers, Rowan, et al. "Merlot reserve: Neural script knowledge through vision and language and sound." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

MERLOT Reserve: Neural Script Knowledge through Vision and Language and Sound

- Audio brings extra supervision and information towards stronger video understanding and video-language performance.
- Limitation: improvement on learned procedural knowledge may be less significant.

Action Recognition

	Model	Kinetics-600 (%) Top-1	Top-5
Vision Only	VATT-Base[2]	80.5	95.5
	VATT-Large [2]	83.6	96.6
	TimeSFormer-L [9]	82.2	95.6
	Florence [125]	87.8	97.8
	MTV-Base [122]	83.6	96.1
	MTV-Large [122]	85.4	96.7
	MTV-Huge [122]	89.6	98.3
	RESERVE-B	88.1	95.8
	RESERVE-L	89.4	96.3
+Audio	RESERVE-B	89.7	96.6
	RESERVE-L	**91.1**	**97.1**

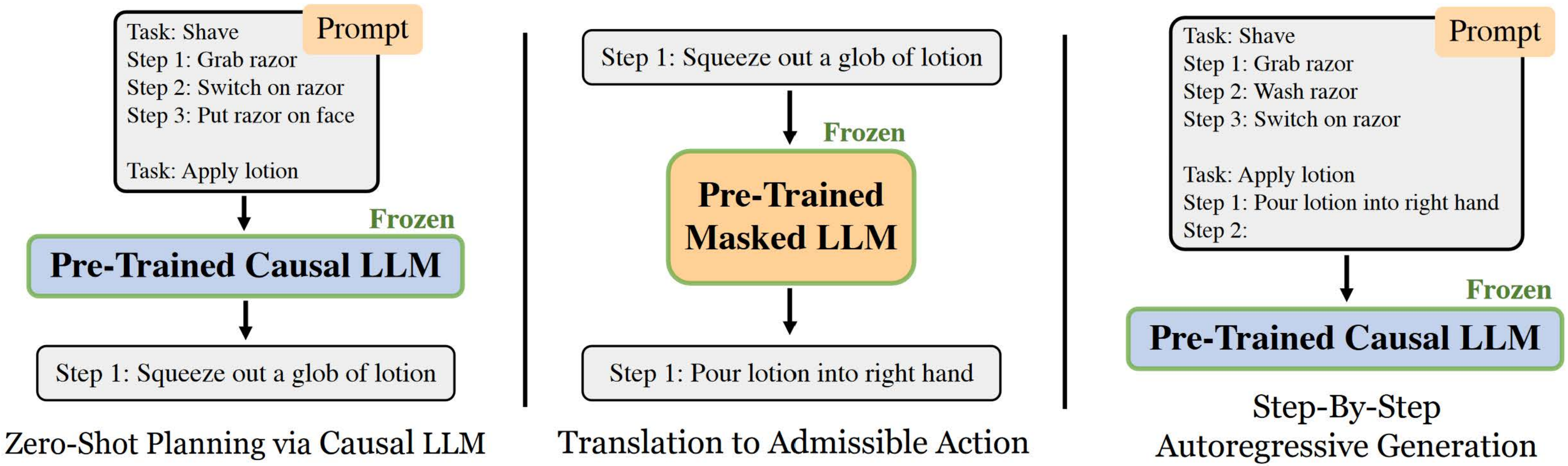
Requiring understand procedures of actions/objects

	Model	Situated Reasoning (STAR) (test acc; %) Interaction	Sequence	Prediction	Feasibility	Overall
	Supervised SoTA (ClipBERT [74])	39.8	43.6	32.3	31.4	36.7
zero-shot	Random	25.0	25.0	25.0	25.0	25.0
	CLIP (VIT-B/16) [92]	39.8	40.5	35.5	36.0	38.0
	CLIP (RN50x16) [92]	39.9	41.7	36.5	**37.0**	38.7
	Just Ask (ZS)[123]					
	RESERVE-B	44.4	40.1	38.1	35.0	39.4
	RESERVE-L	42.6	41.1	37.4	32.2	38.3
	RESERVE-B (+audio)	**44.8**	**42.4**	**38.8**	36.2	**40.5**
	RESERVE-L (+audio)	43.9	42.6	37.6	33.6	39.4

Zellers, Rowan, et al. "Merlot reserve: Neural script knowledge through vision and language and sound." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

Language models as zero-shot planners: Extracting actionable knowledge for embodied agents

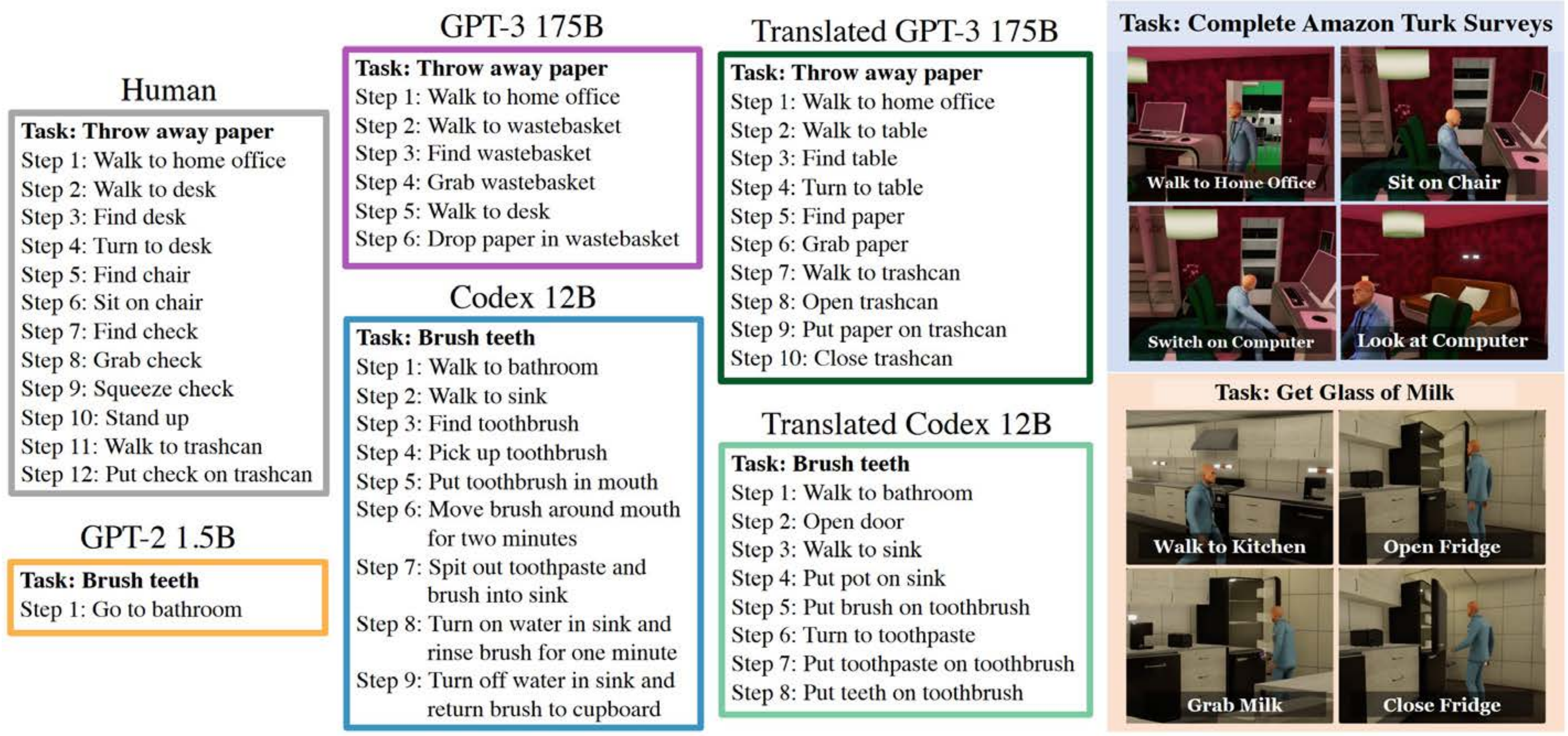
- Key Idea: Large language models learn rich procedural knowledge and such knowledge could be extracted.



Huang, Wenlong, et al. "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents." International Conference on Machine Learning. PMLR, 2022.

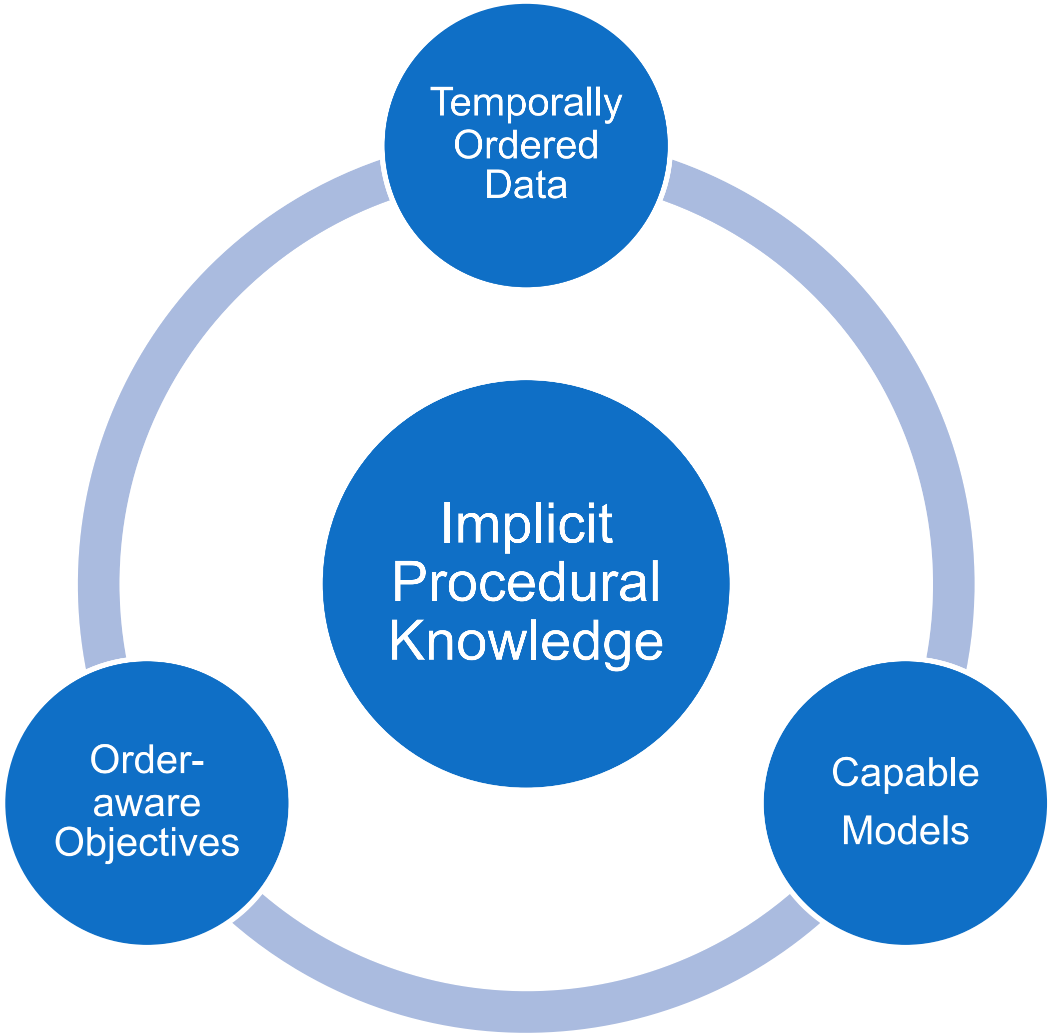
Language models as zero-shot planners: Extracting actionable knowledge for embodied agents

- Impressive results.
- Challenge: verification; groundability to real-world videos.



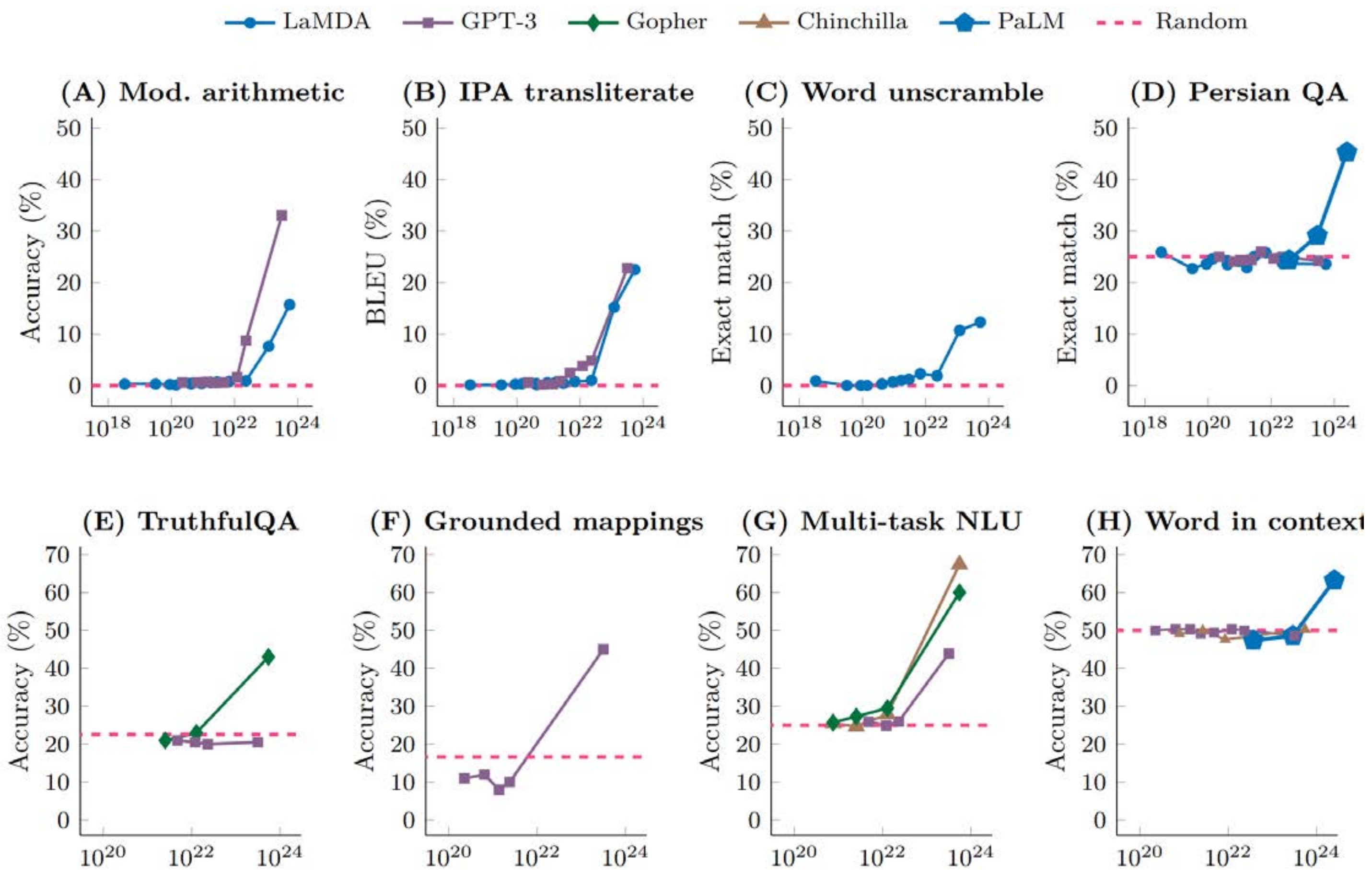
Huang, Wenlong, et al. "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents." International Conference on Machine Learning. PMLR, 2022.

Summary of Methods Learning Implicit Knowledge



Future Challenge: Is there a critical point on scale?

- Can models learn procedural knowledge with a limited scale?

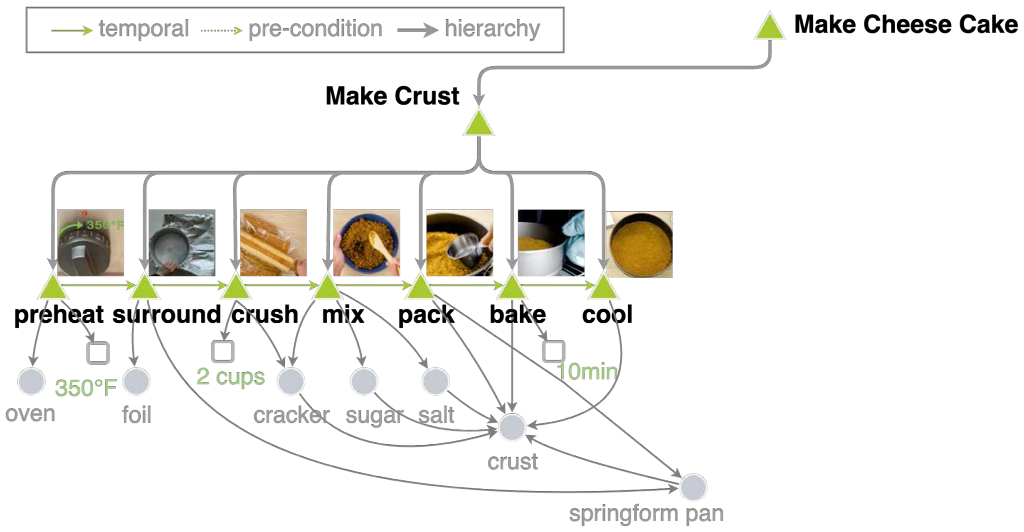


Many reasoning ability of large language models emerge when the model scale is larger than a critical point.

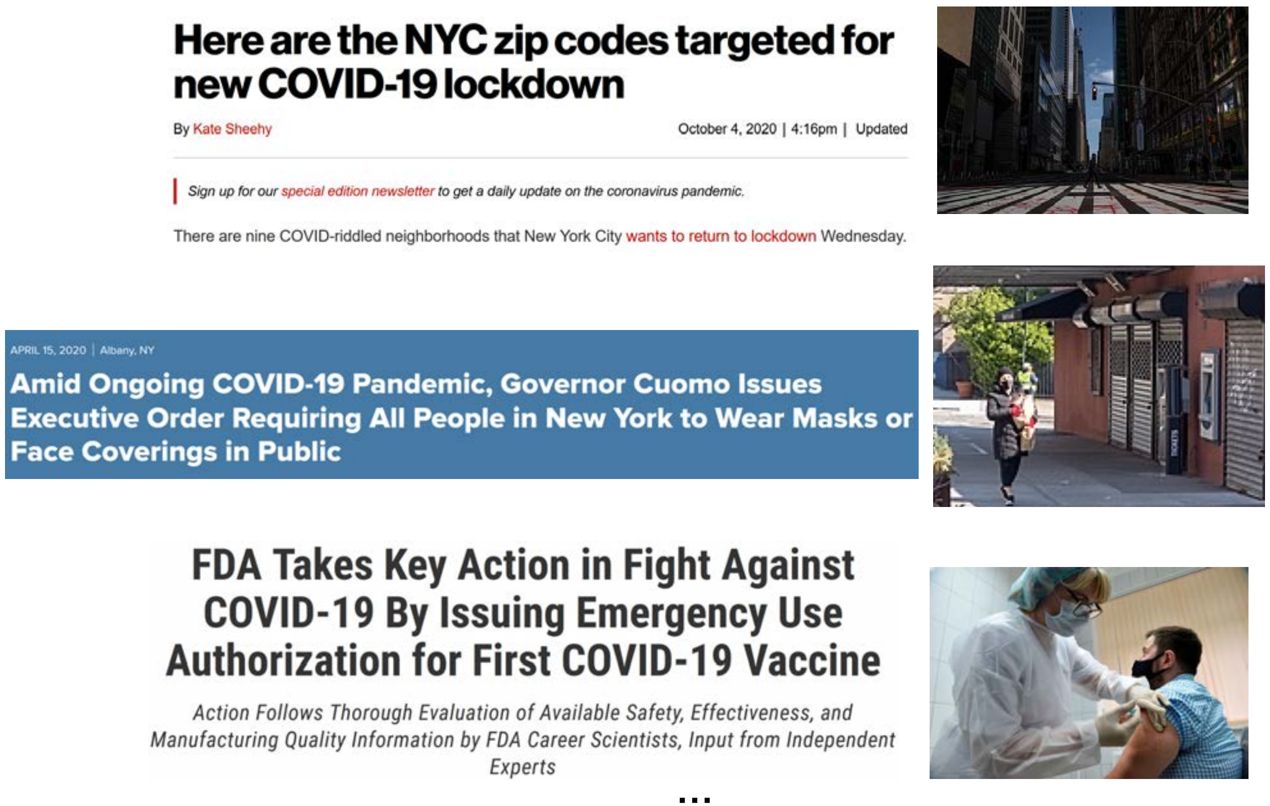
Wei, Jason, et al. "Emergent abilities of large language models." *arXiv preprint arXiv:2206.07682* (2022).

Future Challenge: From an instance to a set

- Can models learn from temporally ordered sets of instances?



Current Data Instance



Here are the NYC zip codes targeted for new COVID-19 lockdown

By Kate Sheehy October 4, 2020 | 4:16pm | Updated

Sign up for our special edition newsletter to get a daily update on the coronavirus pandemic.

There are nine COVID-riddled neighborhoods that New York City wants to return to lockdown Wednesday.

APRIL 15, 2020 | Albany, NY

Amid Ongoing COVID-19 Pandemic, Governor Cuomo Issues Executive Order Requiring All People in New York to Wear Masks or Face Coverings in Public

FDA Takes Key Action in Fight Against COVID-19 By Issuing Emergency Use Authorization for First COVID-19 Vaccine

Action Follows Thorough Evaluation of Available Safety, Effectiveness, and Manufacturing Quality Information by FDA Career Scientists, Input from Independent Experts

...

Real-world complex task

https://en.wikipedia.org/wiki/COVID-19_pandemic_in_New_York_City#cite_note-48
https://nypost.com/2020/10/04/here-are-the-nyc-zip-codes-targeted-for-new-covid-19-lockdown/
https://www.governor.ny.gov/news/amid-ongoing-covid-19-pandemic-governor-cuomo-issues-executive-order-requiring-all-people-new
https://www.fda.gov/news-events/press-announcements/fda-takes-key-action-fight-against-covid-19-issuing-emergency-use-authorization-first-covid-19
https://www.nature.com/articles/d41586-020-02684-9

Take-away Messages

- Learning Procedural Knowledge
  - Explicit Knowledge Source
    - As data
    - As Supervision
    - In Models
  - Implicit Knowledge Source
    - Temporal Data + Temporal Objective + Capable Models

Future Challenges:

Interpret but not memorize;

Towards open-vocabulary;

Learning with limited scale;

From an instance to a set;

…

CVPR

Cross-Modal Knowledge Transfer

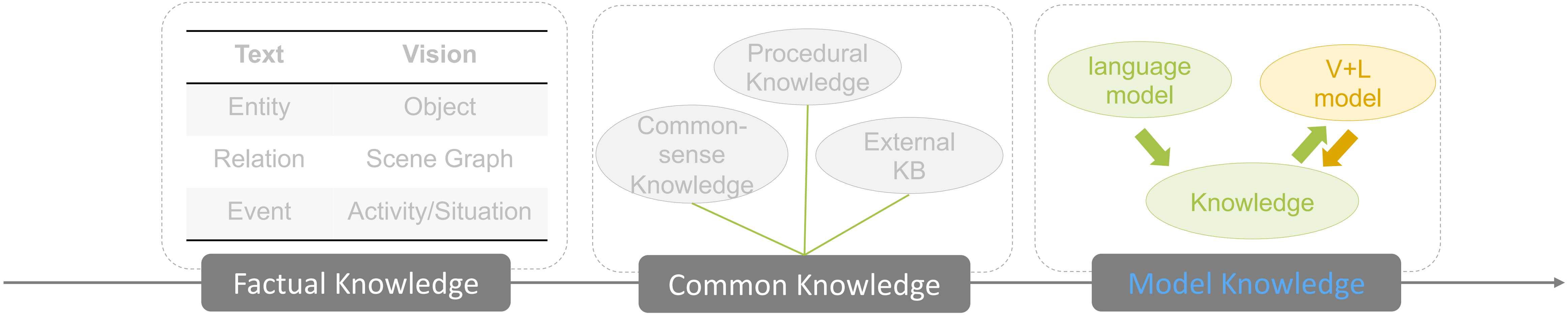
Knowledge-Driven Vision-Language Encoding (Part V)

Jie Lei
Meta AI
jielei@meta.com

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN
Northwestern University
COLUMBIA UNIVERSITY
Meta AI

Overview

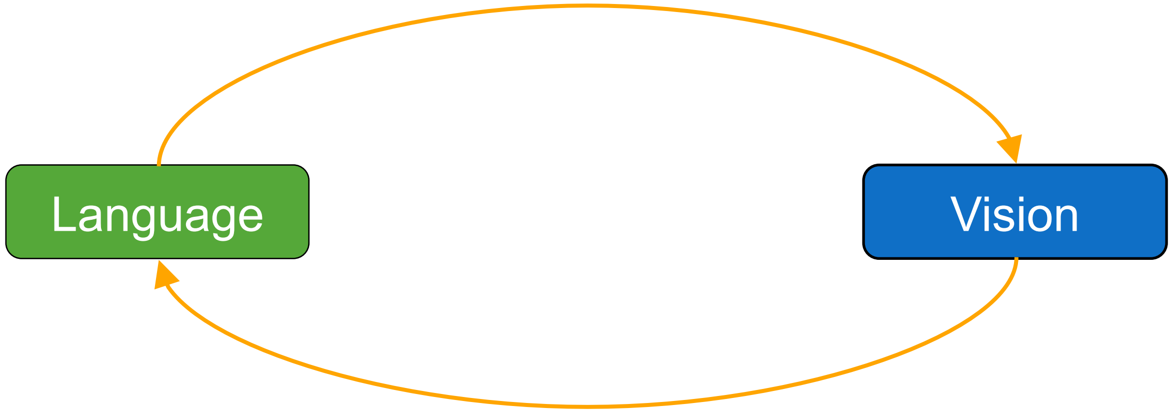
Compared to raw data, knowledge is **important and useful information.**



Agenda

Part 1. Language knowledge helps learn better vision models

- Pure vision tasks: object detection, image classification, etc.
- Multimodal tasks with vision signals: VQA, video captioning, etc.



Part 2. Vision knowledge helps learn better language models

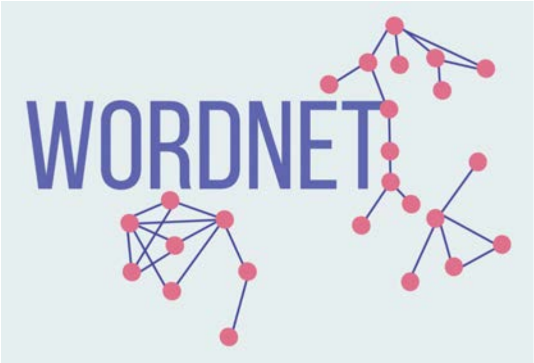
- Human learn language by connecting the words to their visual appearance in the surrounding world.

Part 1. Language → Vision

- Implicit knowledge from pre-trained Language Models (LM)

- Explicit knowledge from human curated sources (e.g., wiki) or model generated knowledge (e.g., GPT-3 generated category definitions)

Wiktionary
The free dictionary



❑ Concept name: snowberg

Def_wik: None

GPT3 Query:

Please explain the concept according to the context.

===

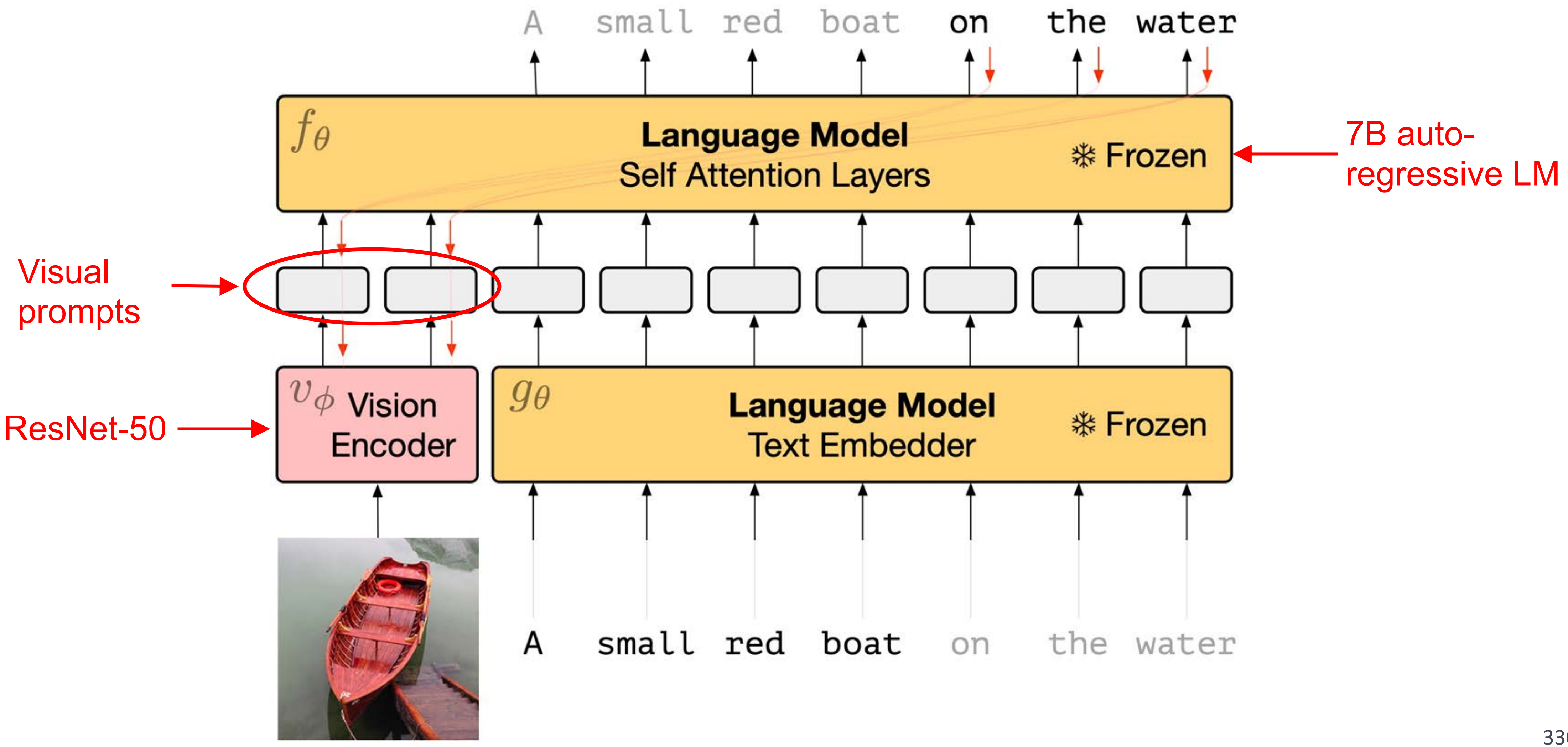
Q: ship

A: A water-borne vessel generally larger than a boat.

Part 1.1 Implicit Knowledge from Language Models

Frozen

- Preserve LM ability by freezing it during cross-modal model training.
- Gradient: frozen LM → vision encoder



Multimodal Few-Shot Learning with Frozen Language Models, Tsimpoukelli et al., NeurIPS 2021

Frozen

- Few-shot multimodal in-context learning after trained on 3M image-text pairs.

Wiki knowledge

- Reasonably good zero/few-shot performance, but still underperform SOTA: limited multimodal data? (3M); LM is relatively small? (7B)

VQAv2

n-shot Acc.	n=0	n=1	n=4	τ
Frozen	29.5	35.7	38.2	✗
Frozen scratch	0.0	0.0	0.0	✗
Frozen finetuned	24.0	28.2	29.2	✗
Frozen train-blind	26.2	33.5	33.3	✗
Frozen VQA	48.4	–	–	✓
Frozen VQA-blind	39.1	–	–	✓
Oscar [23]	73.8	–	–	✓

OKVQA

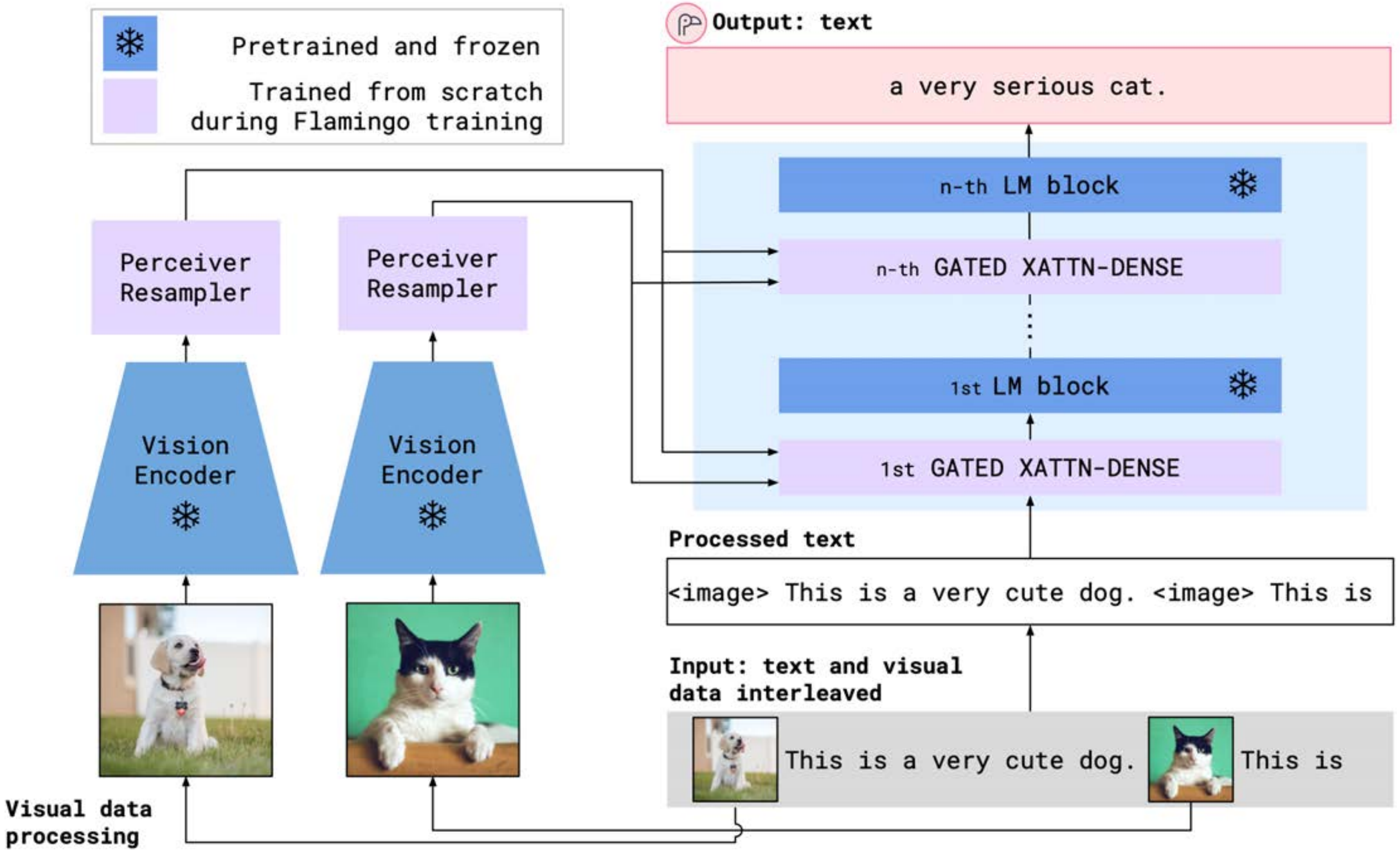
n-shot Acc.	n=0	n=1	n=4	τ
Frozen	5.9	9.7	12.6	✗
Frozen 400mLM	4.0	5.9	6.6	✗
Frozen finetuned	4.2	4.1	4.6	✗
Frozen train-blind	3.3	7.2	0.0	✗
Frozen VQA	19.6	–	–	✗
Frozen VQA-blind	12.5	–	–	✗
MAVEx [42]	39.4	–	–	✓

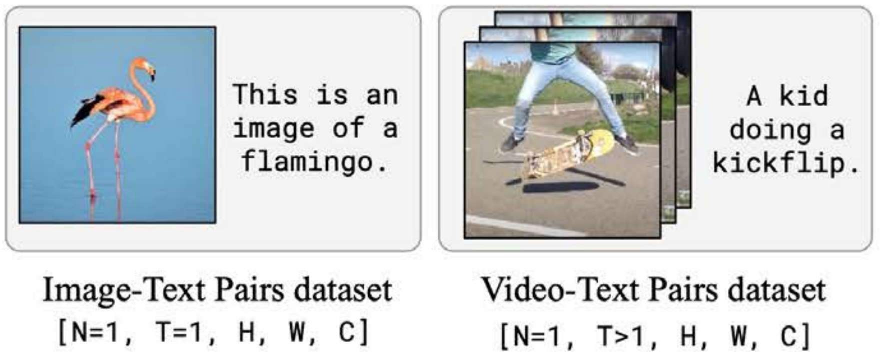
Large gap w/ SOTA

Multimodal Few-Shot Learning with Frozen Language Models, Tsimpoukelli et al., NeurIPS 2021

Flamingo

- A frozen 70B pre-trained LM + a frozen pre-trained ResNet.
- Trained w/ image/video-text pairs, along with interleaved image-text data (M3W), which is important for in-context learning.





ALIGN: 1.8B +
LTIP: 312M images

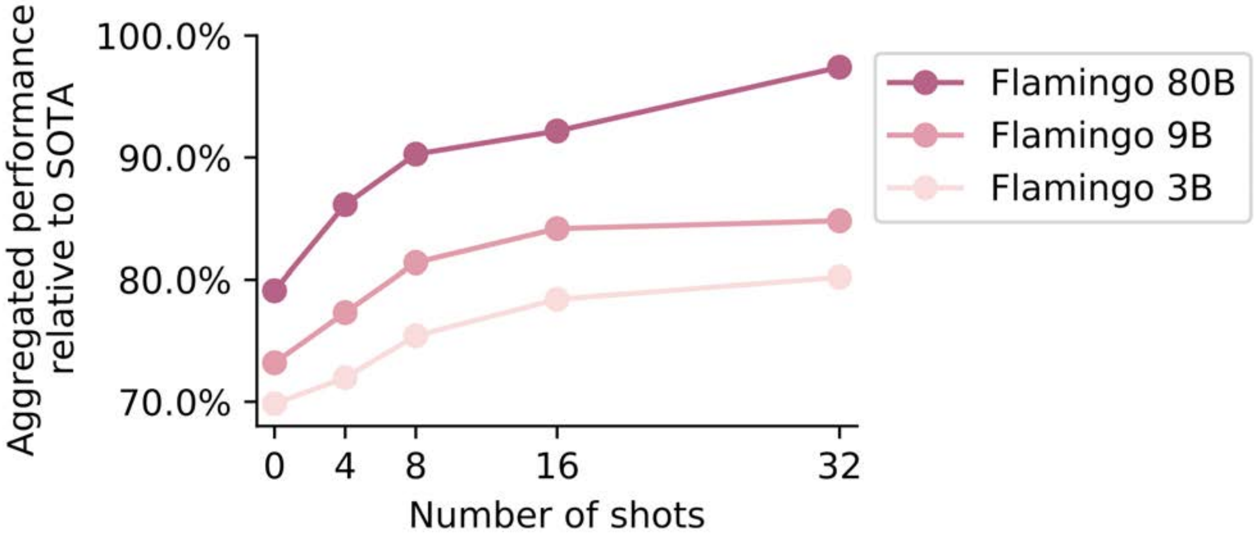
VTP: 27M videos

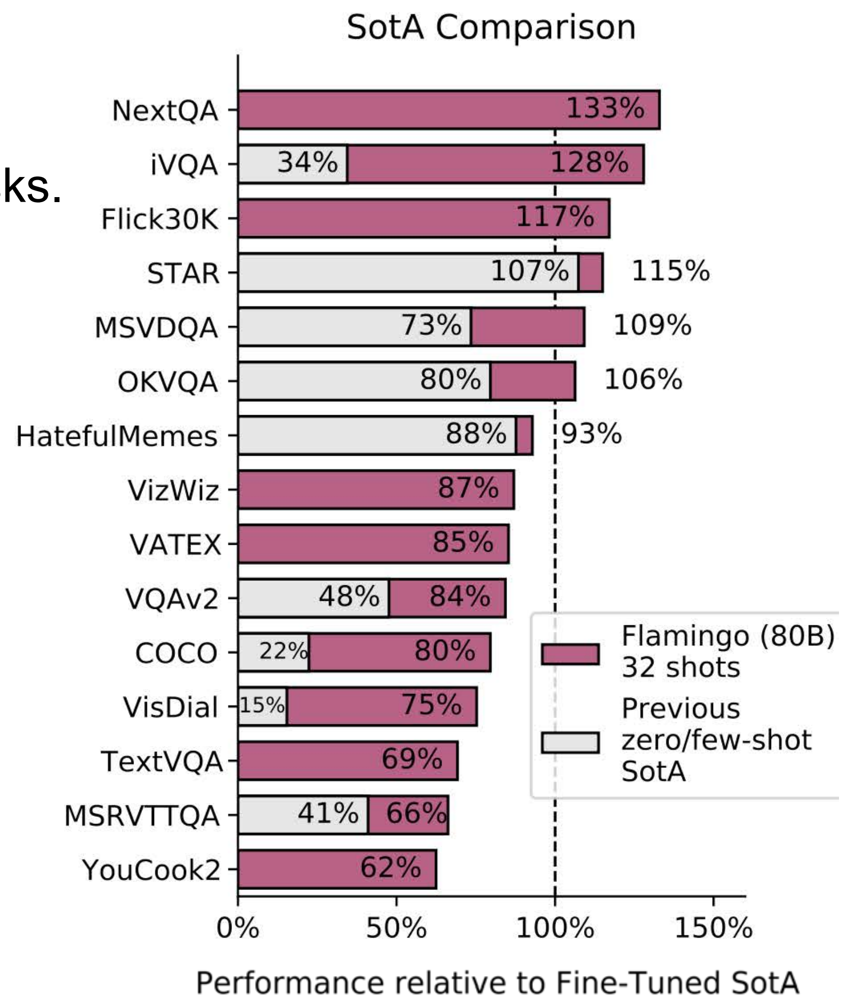
M3W: 43M webpages (185M images)

Flamingo: a Visual Language Model for Few-Shot Learning, Alayrac et al, NeurIPS 2022

Flamingo

- Left: larger model works better; more in-context examples helps.
- Right: thanks to larger model and more training data, he model achieves comparable or better results than SOTA on multiple tasks.

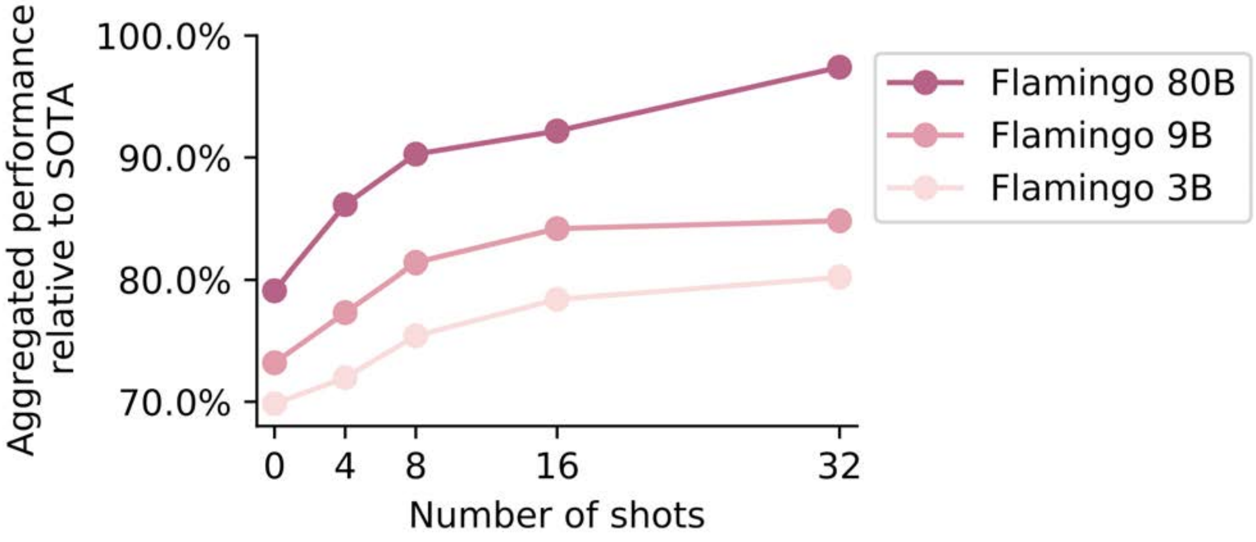


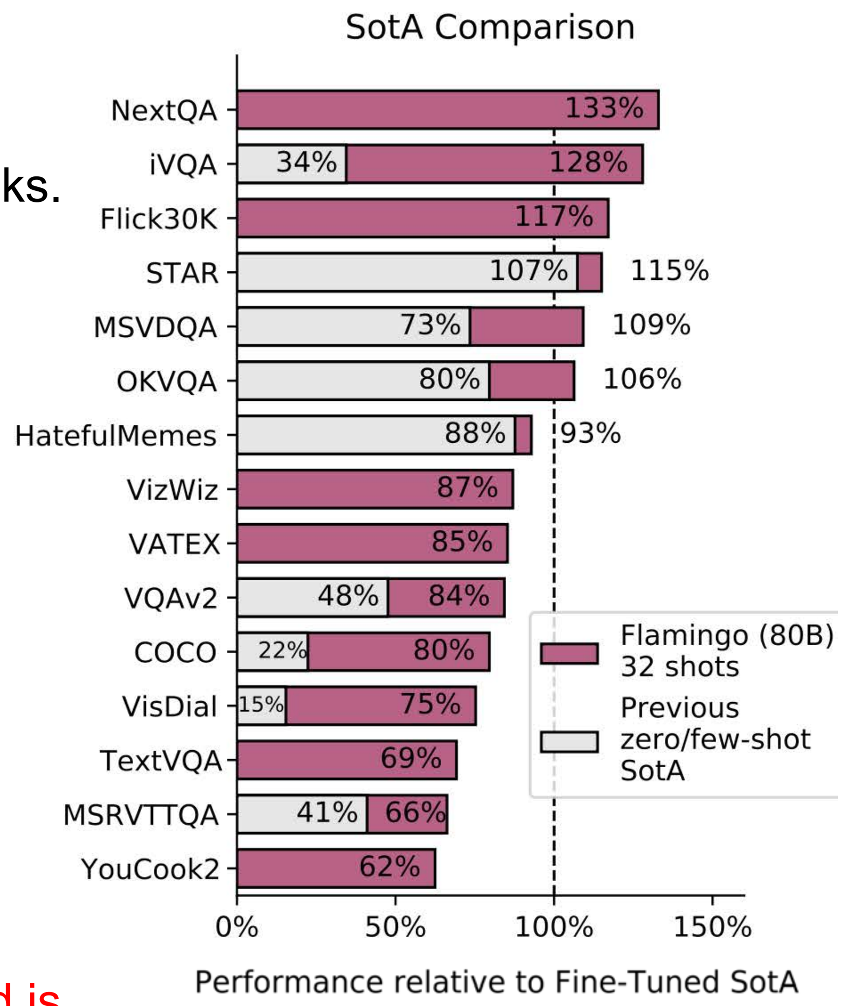


Flamingo: a Visual Language Model for Few-Shot Learning, Alayrac et al, NeurIPS 2022

Flamingo

- Left: larger model works better; more in-context examples helps.
- Right: thanks to larger model and more training data, he model achieves comparable or better results than SOTA on multiple tasks.



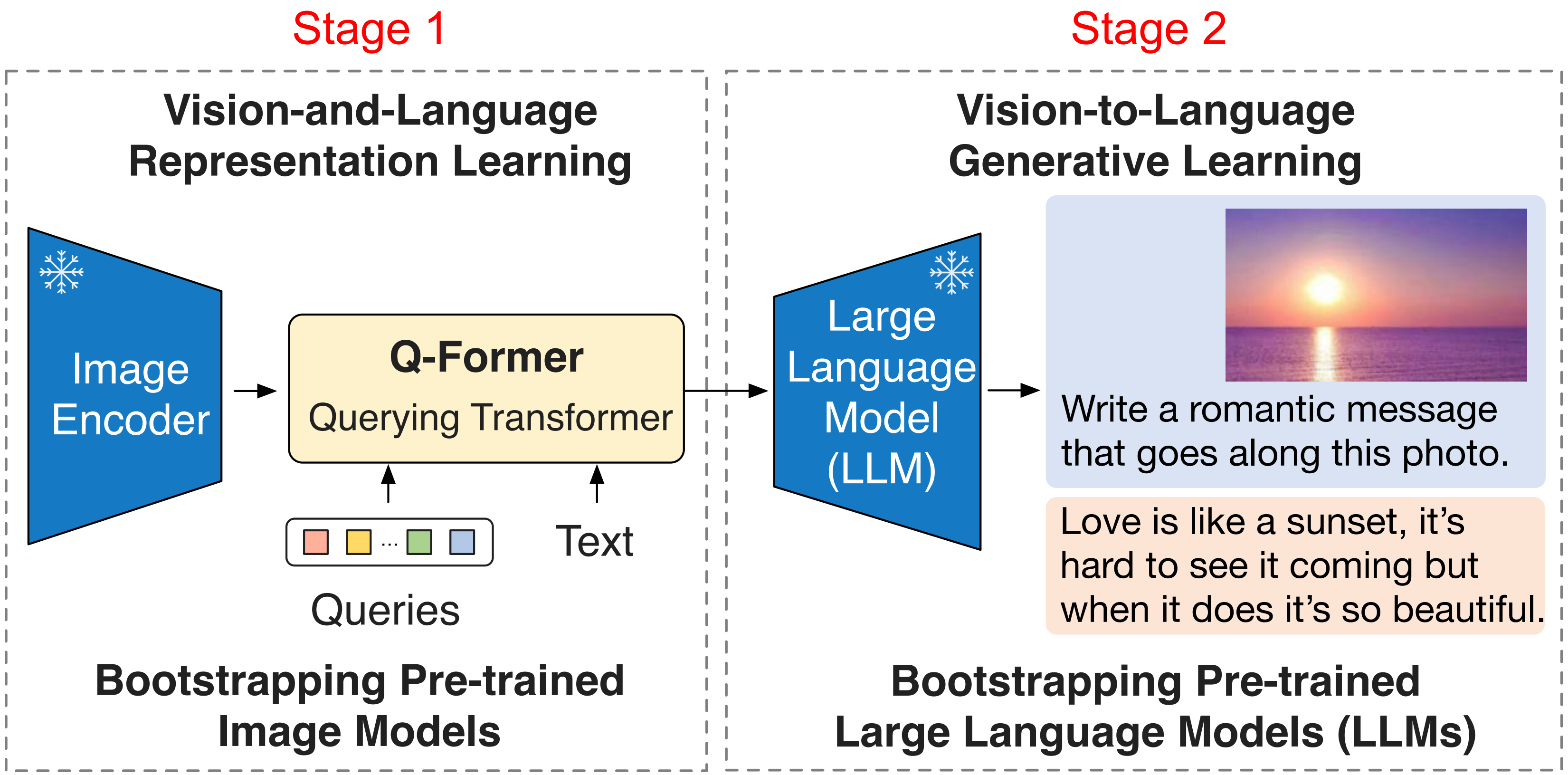


Expensive! The 80 model has 10B trainable parameters and is trained with 1536 TPUv4 chips for 15 days.

Flamingo: a Visual Language Model for Few-Shot Learning, Alayrac et al, NeurIPS 2022

BLIP-2, Learning w/ Frozen LLM

- Architecture: frozen image encoder + a light-weight Q-Former + frozen LLM
- Two stage training:
  - Stage 1 vision-language representation learning: image-text contrastive & matching, image captioning
  - Stage 2 vision-language generative learning: generate text conditioned on image
- Q-Former: BERT-base, using learned query vectors with cross-attention to extract visual info.



BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, Li et al, arXiv 2023

BLIP-2, Learning w/ Frozen LLM

- Architecture: frozen image encoder + a light-weight Q-Former + frozen LLM
- Two stage training:
  - Stage 1 vision-language representation learning: image-text contrastive & matching, image captioning
  - Stage 2 vision-language generative learning: generate text conditioned on image
- Q-Former: BERT-base, using learned query vectors with cross-attention to extract visual info.

- Trained on 129M image-text pairs

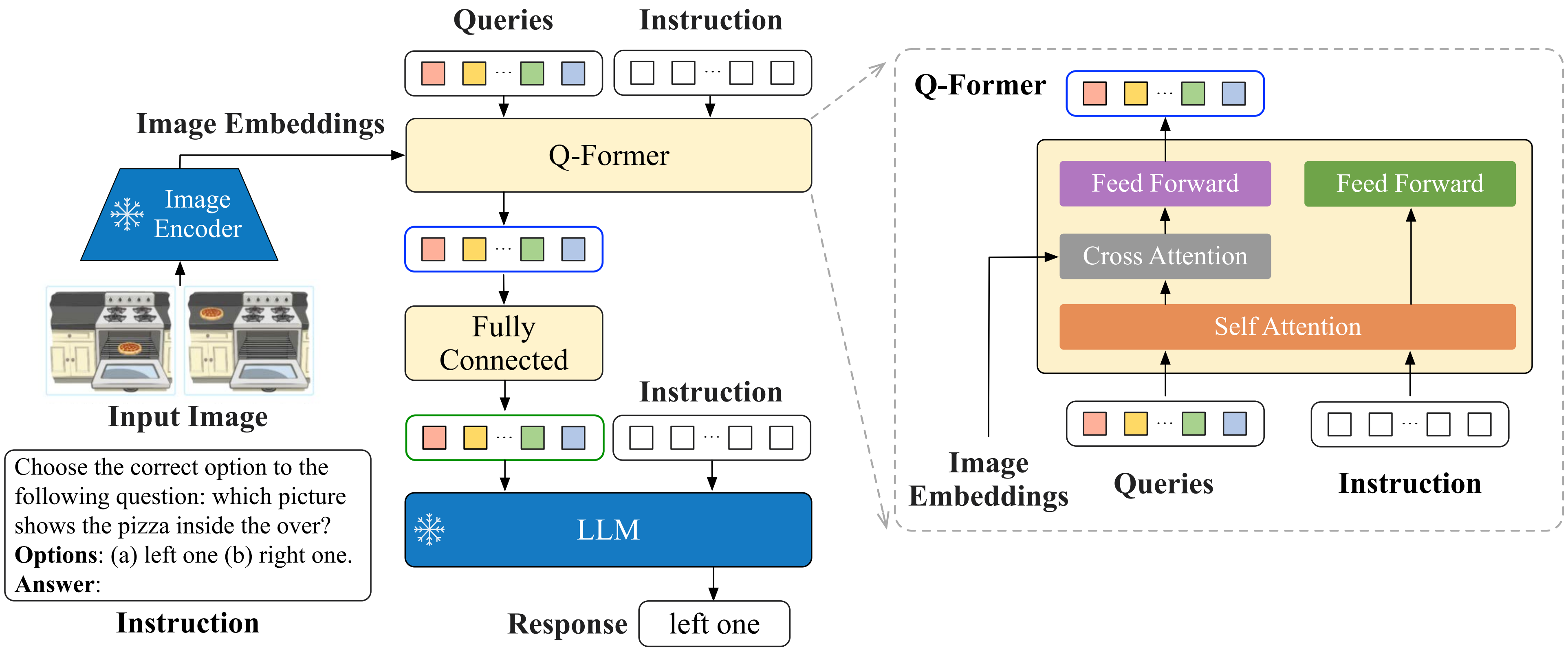
Models	#Trainable Params	Open-sourced?	**Visual Question Answering** VQAv2 (test-dev)	**Image Captioning** NoCaps (val)		**Image-Text Retrieval** Flickr (test)	
			VQA acc.	CIDEr	SPICE	TR@1	IR@1
BLIP (Li et al., 2022)	583M	✓	-	113.2	14.8	96.7	86.7
SimVLM (Wang et al., 2021b)	1.4B	✗	-	112.2	-	-	-
BEIT-3 (Wang et al., 2022b)	1.9B	✗	-	-	-	94.9	81.5
Flamingo (Alayrac et al., 2022)	10.2B	✗	56.3	-	-	-	-
BLIP-2	188M	✓	**65.0**	**121.6**	**15.8**	**97.6**	**89.7**

Results on zero-shot vision-language tasks.

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, Li et al, arXiv 2023

InstructBLIP

- Architecture: same as BLIP-2, except instruction text is added to Q-Former for instruction-aware visual feature extraction
- Training: BLIP-2 pre-training + Instruction Finetuning on 13 held-in datasets
- Evaluation: on both held-in and held-out datasets



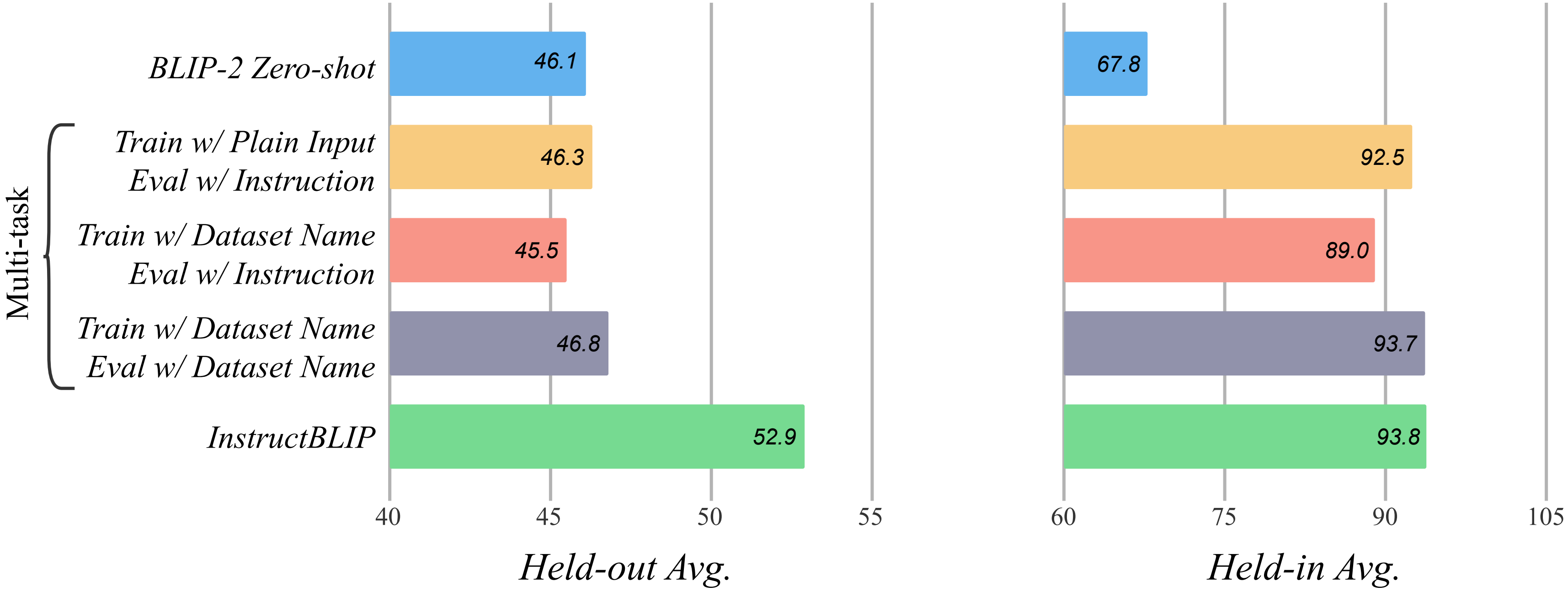
InstructBLIP Model Overview

InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning, Dai et al, arXiv 2023

InstructBLIP: Instruction Tuning vs. Multi-task

- Instruction tuned model excels in unseen datasets and tasks

Strategy	Template (use VQAv2 dataset as an example)
Instruction Tuning	• <image> Question: {question} Short answer: • <image> What is the answer to the following question? {question} • <image> Based on the image, respond to this question with a short answer: {Question}. Answer: • ...
Multi-task	• Plain text: {image} {question} → {answer} • Dataset Name: {image} [Visual question answering:VQAv2] {question} → {answer}

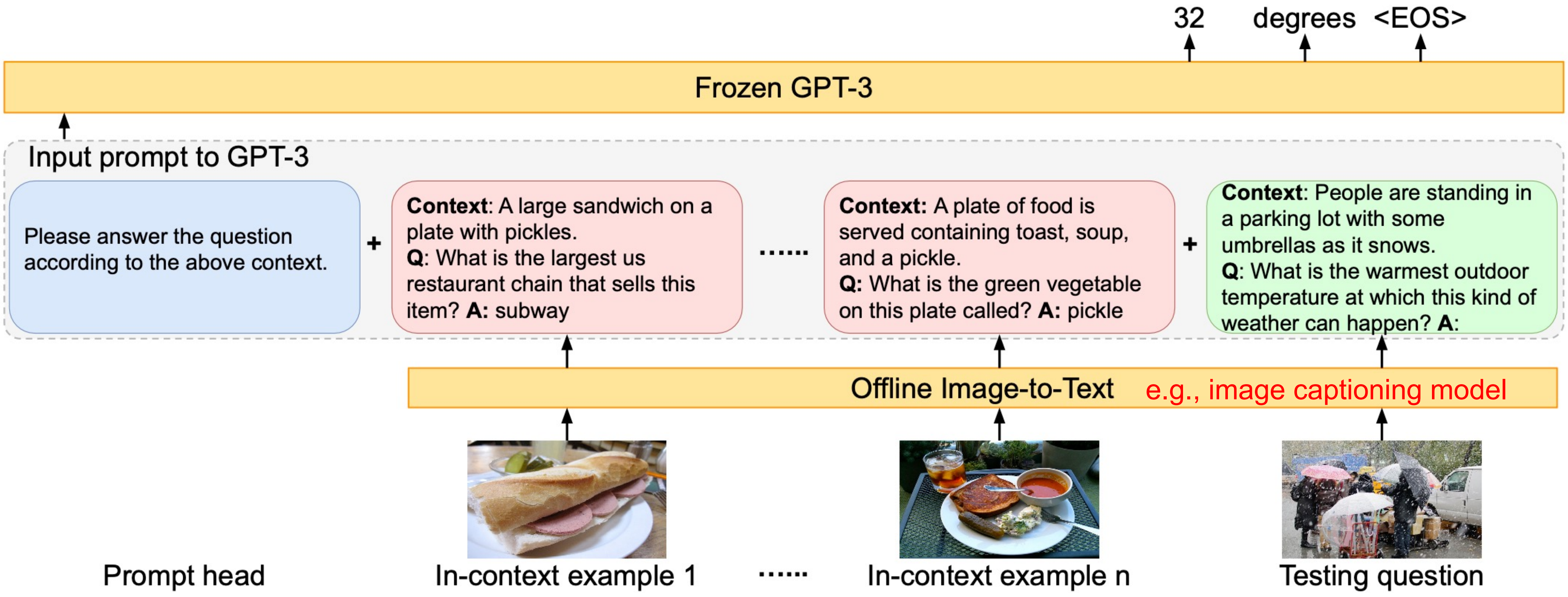


InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning, Dai et al, arXiv 2023

The method discussed above all require additional multi-modal pre-training, however, it is very expensive for LLMs. Is there an alternative way to utilize knowledge in LLMs?

PICa for few-shot knowledge-based VQA

- Summarize image info in text form with an image-to-text model, and prompt GPT-3 to get an answer.
    - Image QA problem is converted into a text QA problem.
    - Implicit GPT-3 knowledge <-> previous approaches explicitly query external knowledge
    - Few-shot w/o parameter update.



An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA, Yang et al, AAAI 2022

PICa for few-shot knowledge-based VQA

- Works better than fine-tuned models that use explicit wiki knowledge.

OKVQA

Method	Image Repr.	Knowledge Resources	Few-shot	Accuracy
MUTAN+AN (Ben-Younes et al. 2017)	Feature Emb.	Wikipedia	✗	27.8
Mucko (Zhu et al. 2020)	Feature Emb.	Dense Captions	✗	29.2
ConceptBert (Garderes et al. 2020)	Feature Emb.	ConceptNet	✗	33.7
ViLBERT (Lu et al. 2019)	Feature Emb.	None	✗	35.2
KRISP (Marino et al. 2021)	Feature Emb.	Wikipedia + ConceptNet	✗	38.9
MAVEx (Wu et al. 2021)	Feature Emb.	Wikipedia + ConceptNet + Google Images	✗	39.4
Frozen (Tsimpoukelli et al. 2021)	Feature Emb.	Language Model (7B)	✓	12.6
PICa-Base	Caption	GPT-3 (175B)	✓	42.0
PICa-Base	Caption+Tags	GPT-3 (175B)	✓	43.3
PICa-Full	Caption	GPT-3 (175B)	✓	46.9
PICa-Full	Caption+Tags	GPT-3 (175B)	✓	**48.0**

- A core issue: image-to-text models are not perfect, it will cause information loss.

(e) What color is the man's jacket?
Context: A man flying through the air while riding a snowboard.
Answer: black
GT Answer: ['red', 'red', 'red', 'orange', 'red', 'red', 'red', 'red', 'red', 'red']

Acc.: 0.0

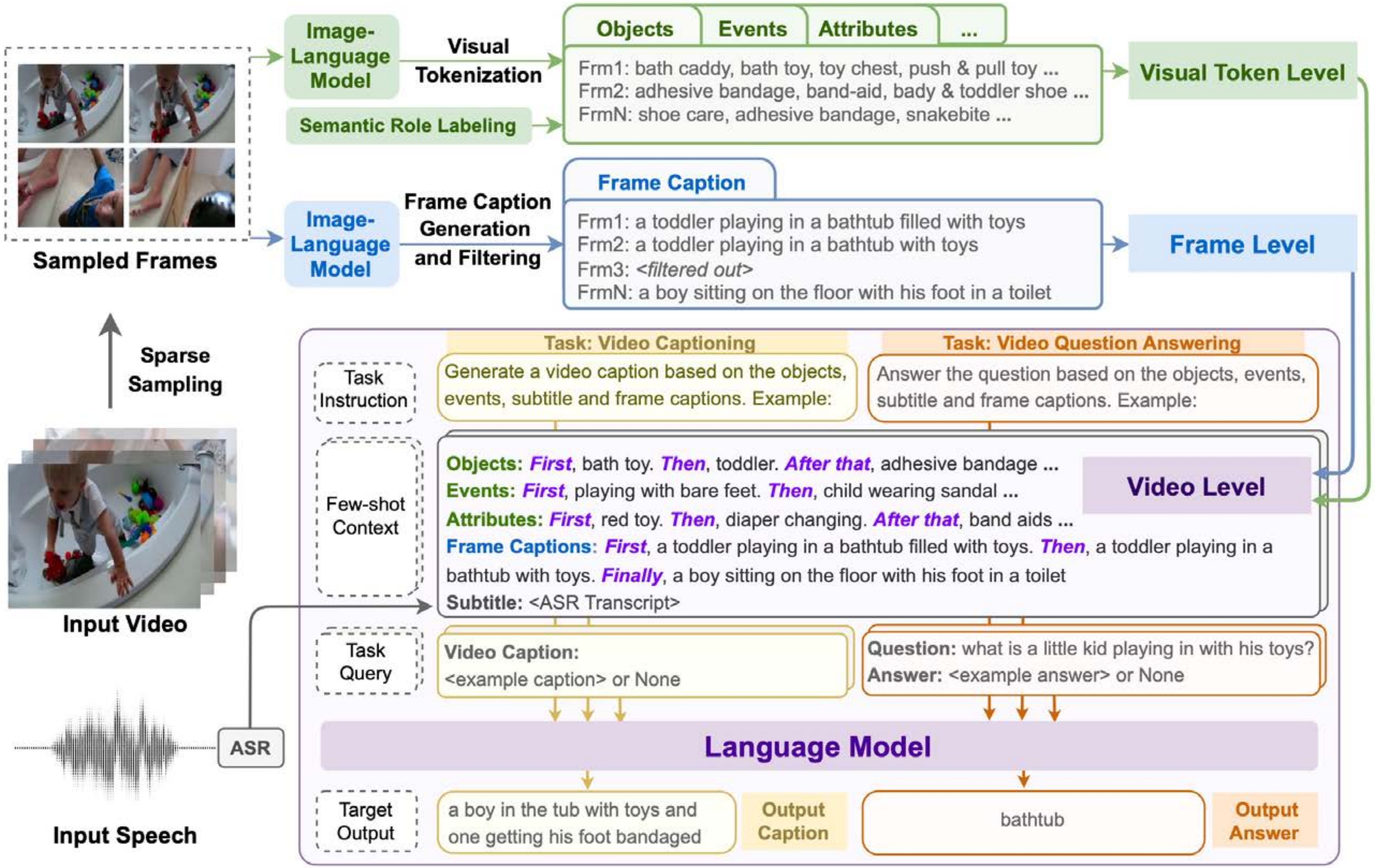
(f) How many giraffes are there?
Context: A herd of giraffe standing next to a wooden fence.
Answer: 3
GT Answer: ['6', '6', '8', '6', '8', '6', '6', '7', '8', '7']

Acc.: 0.0

An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA, Yang et al, AAAI 2022

VidIL: LLM video + language learning

- Generate frame-level info at various granularity, and put them in a temporal aware prompt for LLM.

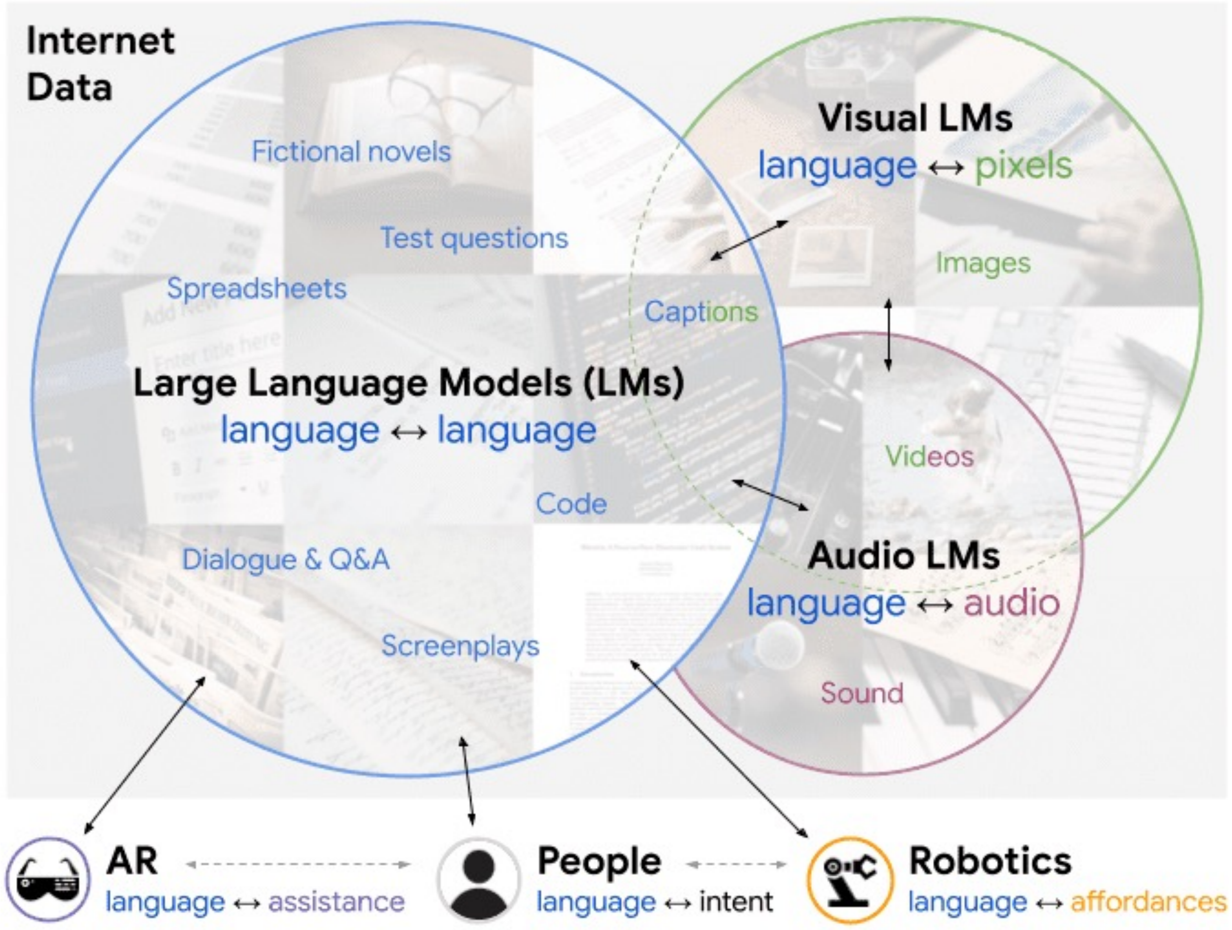


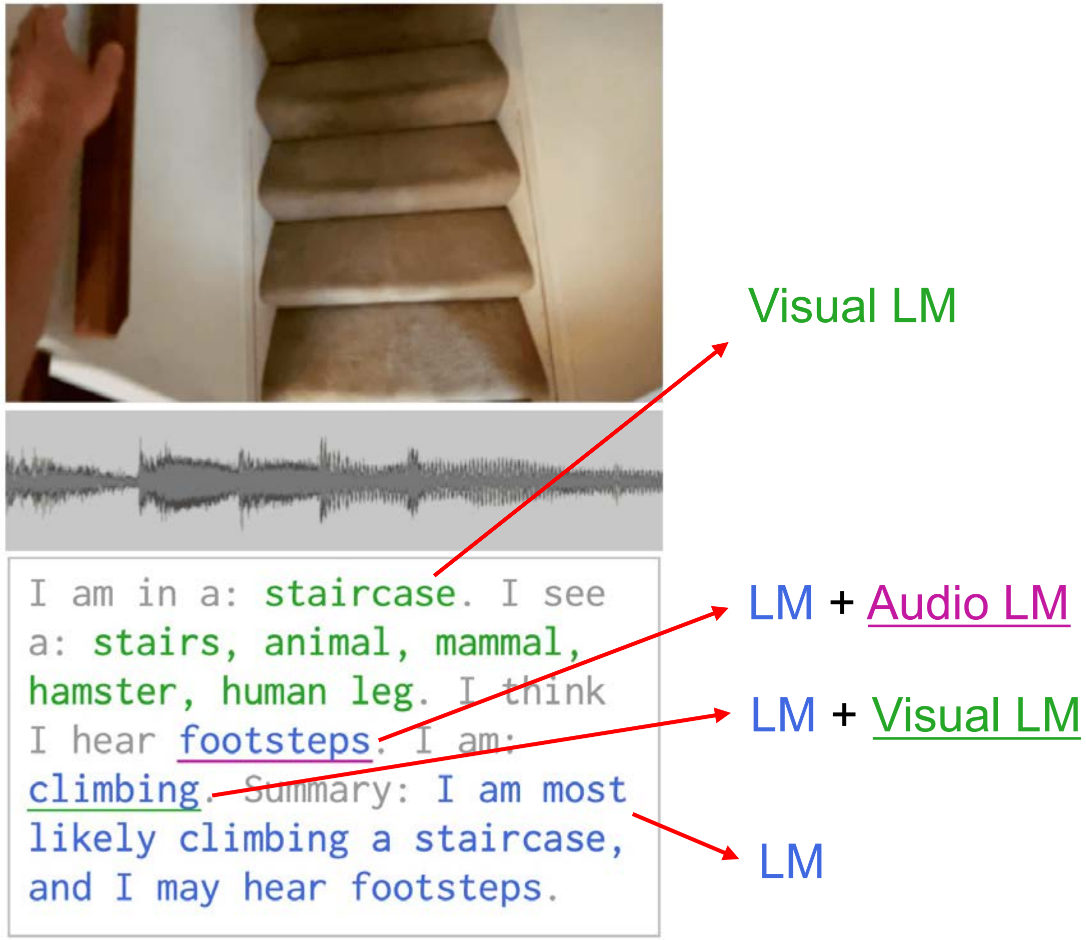
Temporal-Aware Few-shot Prompt

Language Models with Image Descriptors are Strong Few-Shot Video-Language Learners, Wang et al, NeurIPS 2022

Socratic: Composing Multi-modality w/ LLM

- A modular framework in which multiple pretrained models may be composed zero-shot through language without training.





Summarize ego-centric videos.

Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language, Zeng et al, ICLR 2023

Socratic: Composing Multi-modality w/ LLM

- The model works well on vision-language tasks such as image captioning, it can also parse & generate robot instructions from free form human language.

Human defines when to use which model.

Method	BLEU-4	METEOR	CIDEr	SPICE	ROUGE-L
*ClipCap [45]	40.7	30.4	152.4	25.2	60.9
†MAGIC [61]	11.4	16.4	56.2	11.3	39.0
ZeroCap [62]	0.0	8.8	18.0	5.6	18.3
SMs 0-shot (ours)	6.9	15.0	44.5	10.1	34.1
SMs 3-shot (ours)	**18.3**	**18.8**	**76.3**	**14.8**	**43.7**

* finetuned on full training set with image-text pairs.
† finetuned on unpaired training set, zero-shot on image-text pairs.

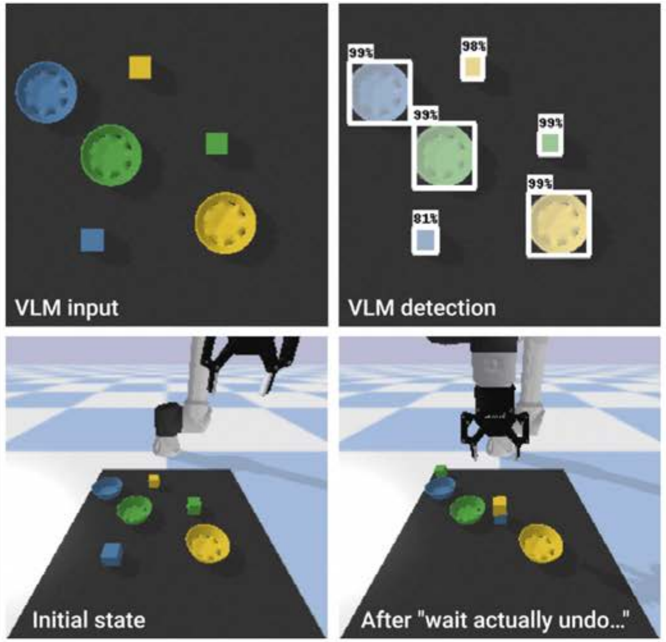
COCO Captions

Visual LM

User Instruction

LM

```
objects = ["green block", "blue block", "yellow block", "green
bowl", "blue bowl", "yellow bowl"]
# move all the blocks to different corners.
Step 1. robot.pick_and_place("green block", "top left corner")
Step 2. robot.pick_and_place("blue block", "top right corner")
Step 3. robot.pick_and_place("yellow block", "bottom left corner")
# now move the blue block to the middle.
Step 1. robot.pick_and_place("blue block", "middle")
# stack the blocks on top of each other.
Step 1. robot.pick_and_place("yellow block", "blue block")
Step 2. robot.pick_and_place("green block", "yellow block")
# wait actually undo that last step.
Step 1. robot.pick_and_place("green block", "top left corner")
# put the yellow block in the bowl you think it best fits.
Step 1. robot.pick_and_place("yellow block", "yellow bowl")
# ok now sort the remaining blocks in the same way.
Step 1. robot.pick_and_place("blue block", "blue bowl")
Step 2. robot.pick_and_place("green block", "green bowl")
```



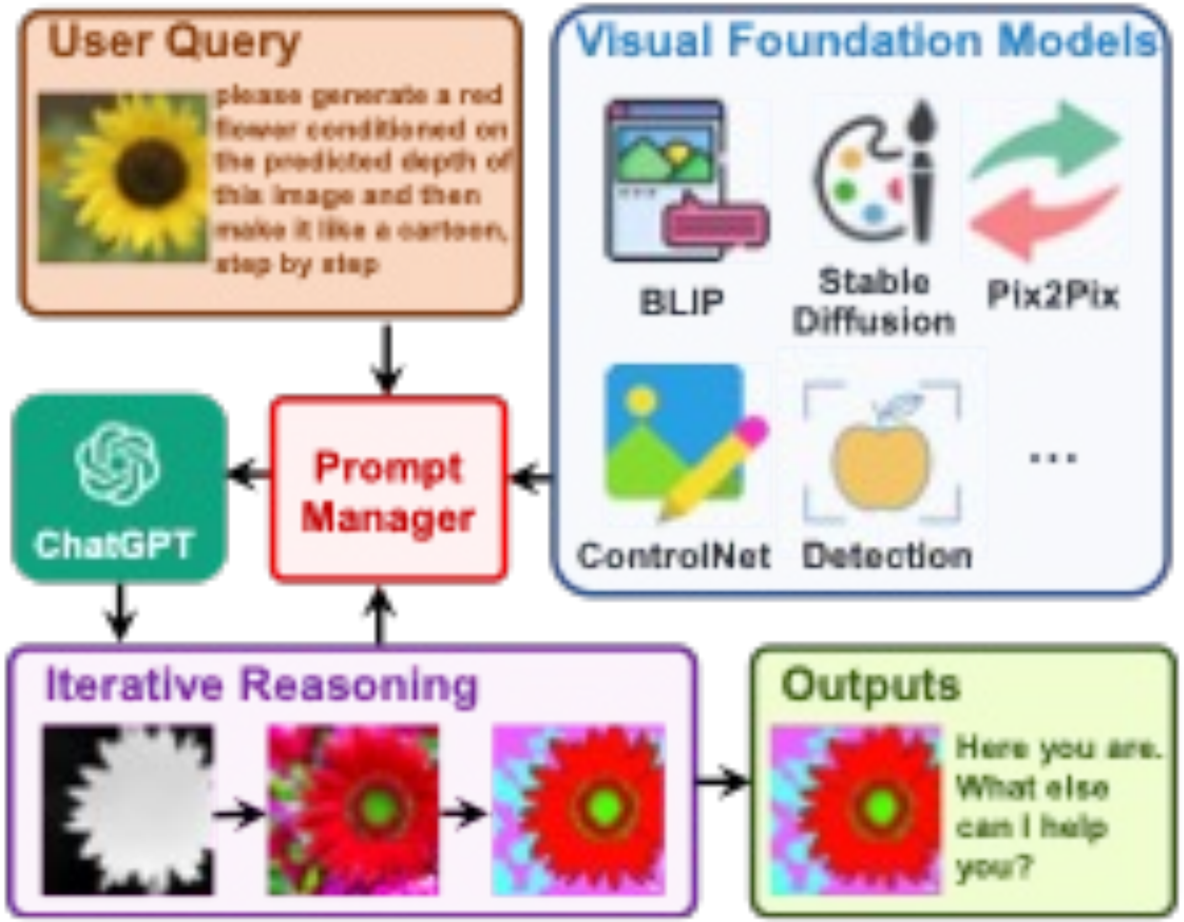
Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language, Zeng et al, ICLR 2023

Visual ChatGPT, ViperGPT, …

- Visual ChatGPT defines a few system principles, and give ChatGPT the autonomy execute actions:
  - System definition.
  - Define name & usage of vision models.
  - Chain-of-Thought.
  - Be strict about filename.
  - Regex to parse executable actions from language.
  - …

Visual ChatGPT is very flexible, as the LLM controls when to use which foundation models, instead of human.

- More general tool learning framework
  - ViperGPT uses generated Python code to compose pre-defined APIs.
  - AutoGPT, New Bing, Bard, …



Architecture of Visual ChatGPT

Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models, Wu et al, arXiv 2023
ViperGPT: Visual Inference via Python Execution for Reasoning, Suris et al, arXiv 2023

LLM for ZS multi-modal learning: Pros/Cons

Pros

- It provides an efficient way to utilize foundation models of different modalities, no extra training required.
- The approaches are modular: new modules can be seamlessly plugged into the framework.

Cons

- Modality specific models are not perfect, there will be info loss when converted into text.
  - The lower performance vs. e2e trained Flamingo model might partly due to this info loss.

(e) What color is the man's jacket?
Context: A man flying through the air while riding a snowboard.
Answer: black
GT Answer: ['red', 'red', 'red', 'orange', 'red', 'red', 'red', 'red', 'red', 'red']

Acc.: 0.0

(f) How many giraffes are there?
Context: A herd of giraffe standing next to a wooden fence.
Answer: 3
GT Answer: ['6', '6', '8', '6', '8', '6', '6', '7', '8', '7']

Acc.: 0.0

Failure cases from the PICa model.

An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA, Yang et al, AAAI 2022

The use of implicit knowledge from pre-trained LMs shows strong zero-shot performance for multi-modal tasks, however, they are hard to interpret. Is there a more interpretable way of using language knowledge?

Part 1.2 Explicit Knowledge from Language

K-LITE

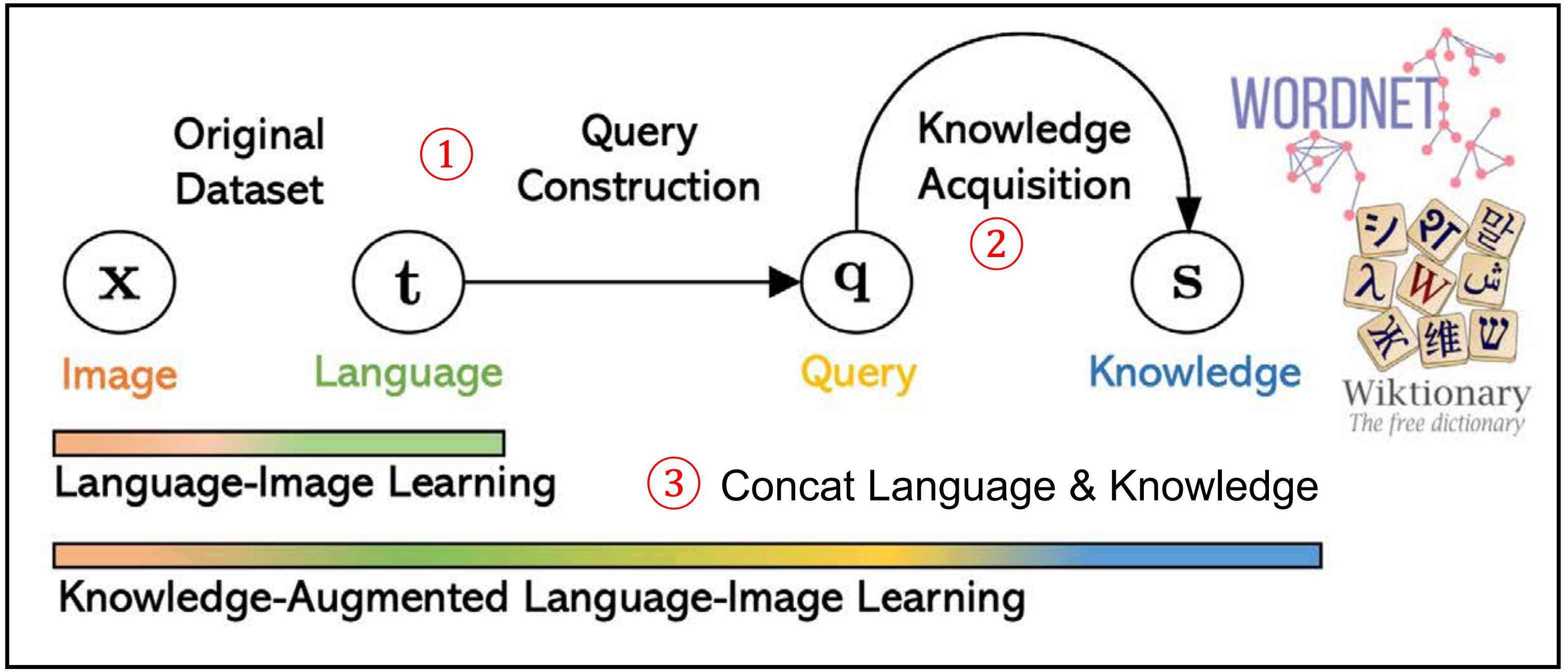
- External knowledge is useful to help the model understand rare concepts.

Sashimi

A dish consisting of ***thin slices*** or pieces of ***raw fish or meat***.

Takoyaki

A ***ball-shaped*** Japanese ***dumpling*** made of batter, filled with diced octopus, ***tempura scraps***, pickled ginger, and ***green onion***.



- a photo of sashimi,
- a photo of takoyaki,

…

Add knowledge

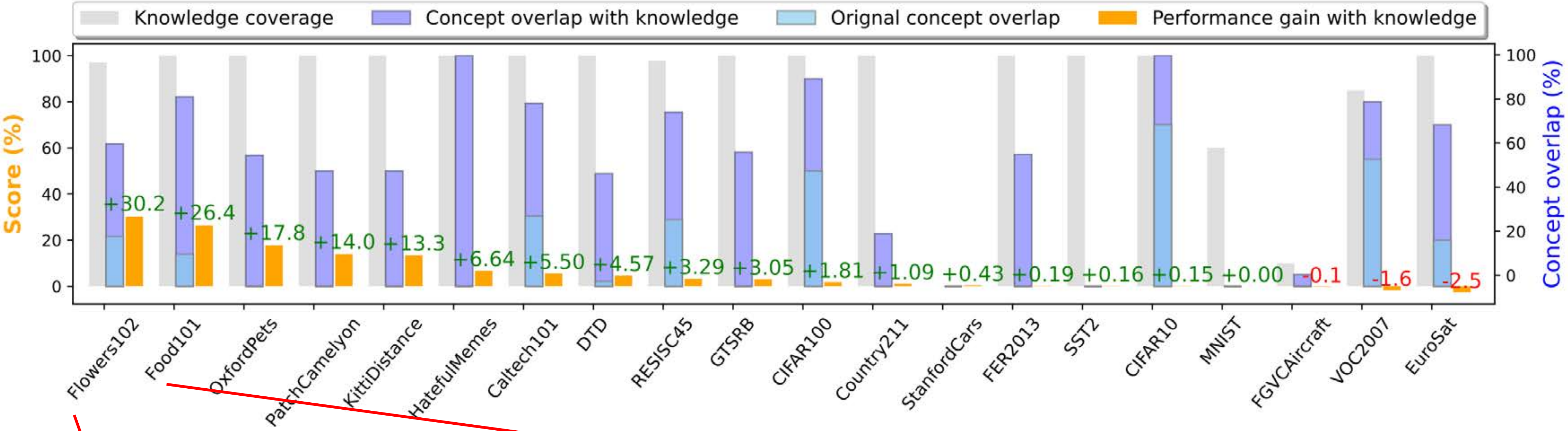
- a photo of sashimi. A dish consisting of slices or ….
- a photo of takoyaki, a ball-shaped Japanese dumpling…

…

K-LITE: Learning Transferable Visual Models with External Knowledge, Shen et al., NeurIPS 2022

K-LITE

- Orange: knowledge improves zero-shot performance on 16/20 image classification datasets.



English marigold: Any of the Old World plants, of the genus Calendula, with orange, yellow or reddish flowers.

Wallflower: Any of several short-lived herbs or shrubs of the Erysimum genus with bright yellow to red flowers.

Lobster bisque: A thick creamy soup made from fish, shellfish, meat or vegetables.

Hot and sour soup: Any one of several soups, served in various Asian cuisines, which are both spicy and sour

ELEVATER

- Same K-LITE model, but with GPT-3 knowledge
- GPT-3 knowledge improves ZS image classification and object detection. More is better.
- GPT-3 + wiki is often better for image classification, but not for object detection.

Concept name: snowberg

Def_wik: None

GPT3 Query:

Please explain the concept according to the context.

===

Q: ship

A: A water-borne vessel generally larger than a boat.

===

Q: storage tank

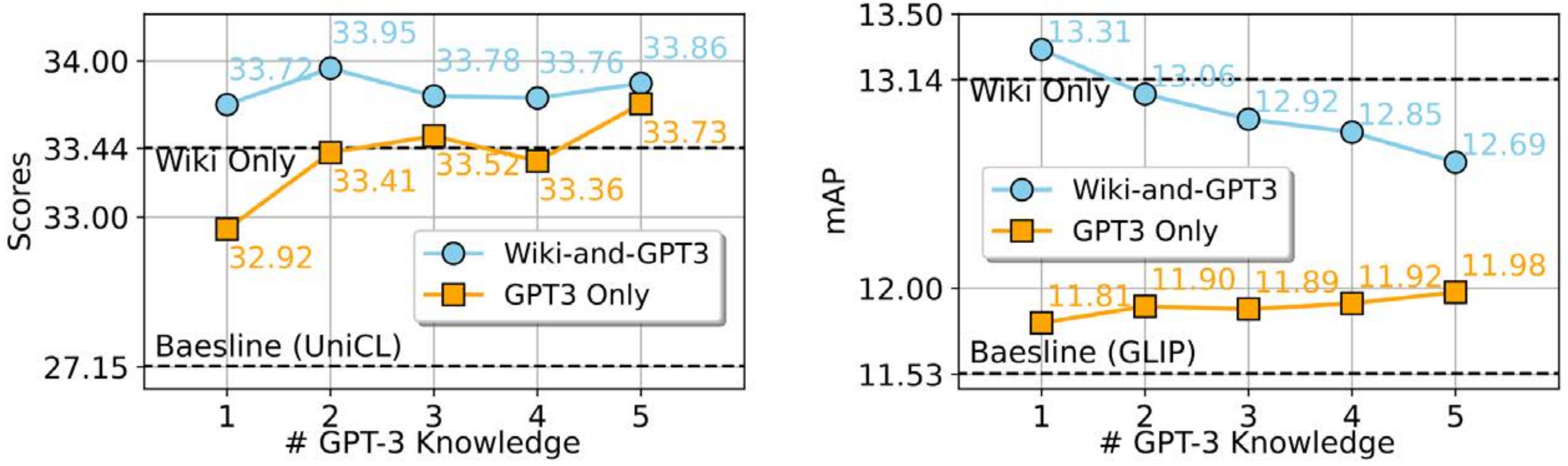
A: A closed container for liquids or gases.

===

Q: snowberg

A:

GPT3 Answer: A large mass of ice floating in the sea.



(a) Image classification (b) Object detection

Zero-shot performance

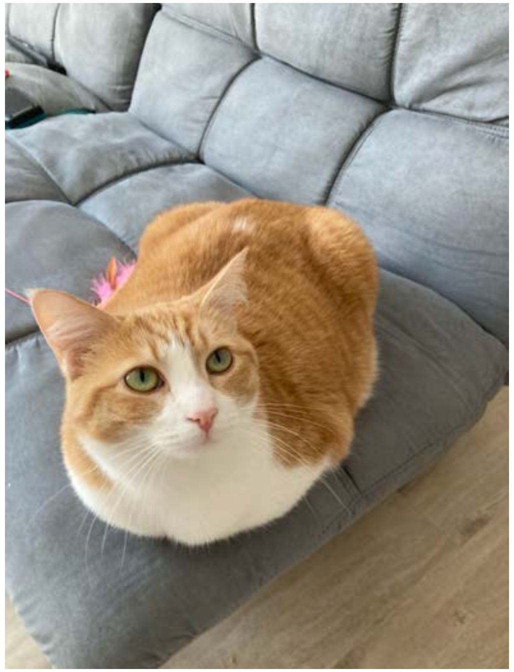
ELEVATER: A Benchmark and Toolkit for Evaluating Language-Augmented Visual Models, NeurIPS 2022

Vision → Language

Could vision knowledge help learn language?

Could vision knowledge help learn language?

- Visual pointing is an essential step for most children to learn meanings of words [Bloom 2002].



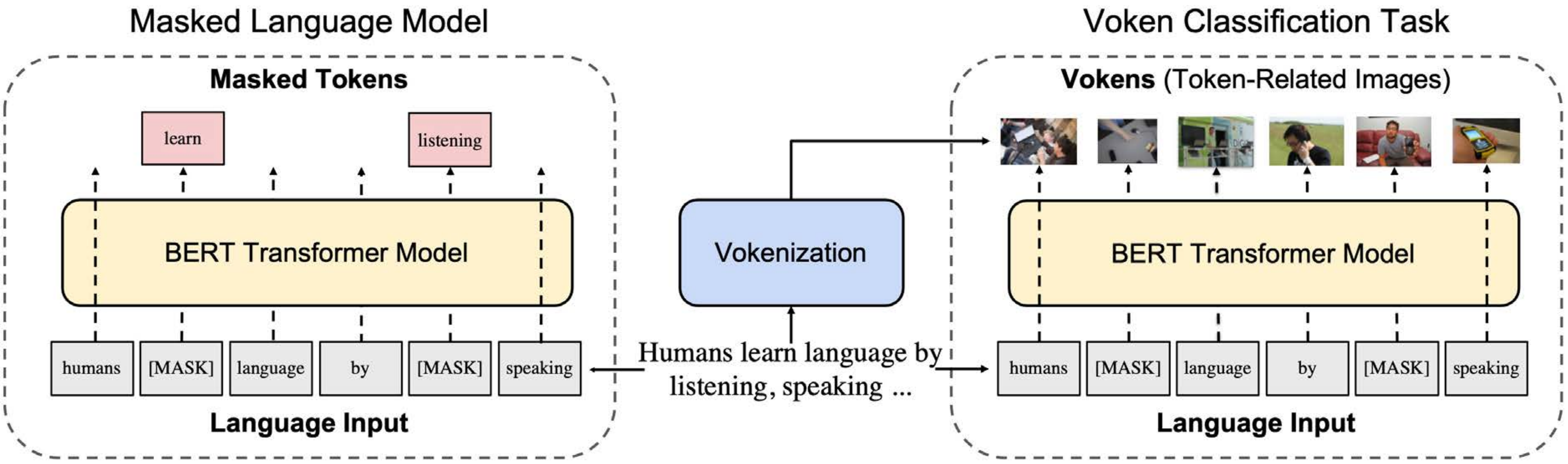
Look! This is a “cat”!

“ButterCup”, cat photo credit to Xiaoyu Xiang
How children learn the meanings of words. Paul Bloom. 2002. MIT press.
Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision, Tan et al., EMNLP 2020

Vokenization: LM w/ Vision Supervision

- Besides standard Masked Language Modeling (MLM), the LM is also trained w/ a voken classification task, by assigning each text token into one of the images (vokens) in the pool.

- Vokens are pre-defined, and are obtained by using a pre-trained image-text retrieval model



Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision, Tan et al., EMNLP 2020

Vokenization: LM w/ Vision Supervision

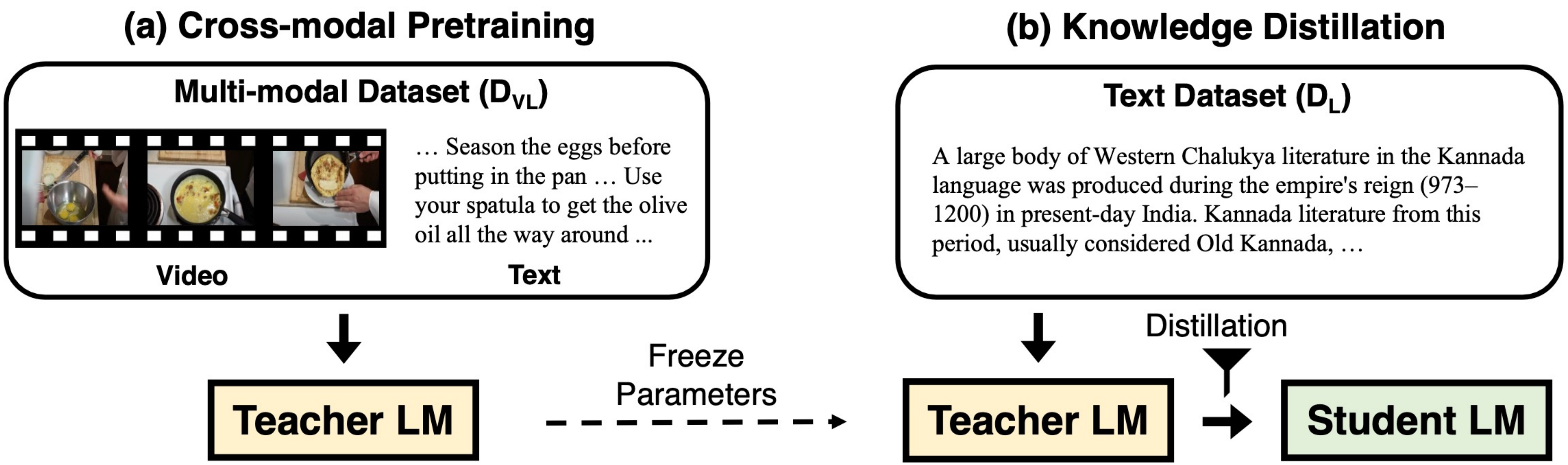
- Voken classification task improves LM performance on a wide range of pure-language tasks.
- This conclusion holds for both BERT and RoBERTa.

Method	SST-2	QNLI	QQP	MNLI	SQuAD v1.1	SQuAD v2.0	SWAG	Avg.
$\text{BERT}_{6L/512H}$	88.0	85.2	87.1	77.9	71.3/80.2	57.2/60.8	56.2	75.6
$\text{BERT}_{6L/512H}$ + Voken-cls	89.7	85.0	87.3	78.6	71.5/80.2	61.3/64.6	58.2	76.8
$\text{BERT}_{12L/768H}$	89.3	87.9	83.2	79.4	77.0/85.3	67.7/71.1	65.7	79.4
$\text{BERT}_{12L/768H}$ + Voken-cls	**92.2**	**88.6**	**88.6**	**82.6**	**78.8/86.7**	68.1/71.2	**70.6**	**82.1**
$\text{RoBERTa}_{6L/512H}$	87.8	82.4	85.2	73.1	50.9/61.9	49.6/52.7	55.1	70.2
$\text{RoBERTa}_{6L/512H}$ + Voken-cls	87.8	85.1	85.3	76.5	55.0/66.4	50.9/54.1	60.0	72.6
$\text{RoBERTa}_{12L/768H}$	89.2	87.5	86.2	79.0	70.2/79.9	59.2/63.1	65.2	77.6
$\text{RoBERTa}_{12L/768H}$ + Voken-cls	**90.5**	**89.2**	**87.8**	**81.0**	**73.0/82.5**	**65.9/69.3**	**70.4**	**80.6**

Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision, Tan et al., EMNLP 2020

VidLanKD: LM w/ Video-Distilled Knowledge

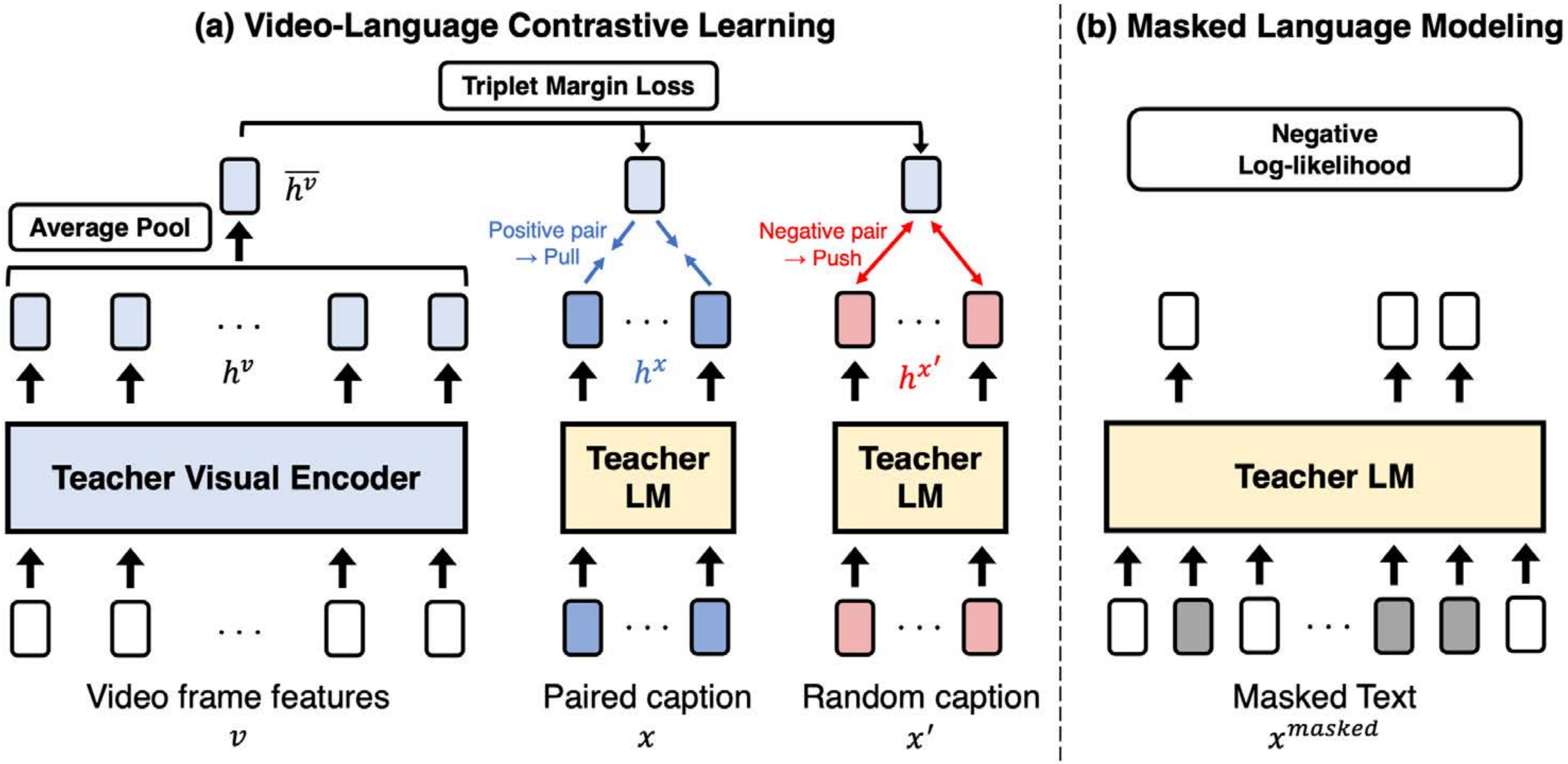
- Vokenization suffers from approximation error of using finite image labels + the lack of vocabulary diversity of a small image-text dataset (COCO).
- VidLanKD improves it by (1) using knowledge distillation instead of discrete vokenization to avoid approximation error; (2) using a large-scale video-language dataset HowTo100M.



VIDLANKD: Improving Language Understanding via Video-Distilled Knowledge Transfer, Tang et al., NeurIPS 2021

VidLanKD: LM w/ Video-Distilled Knowledge

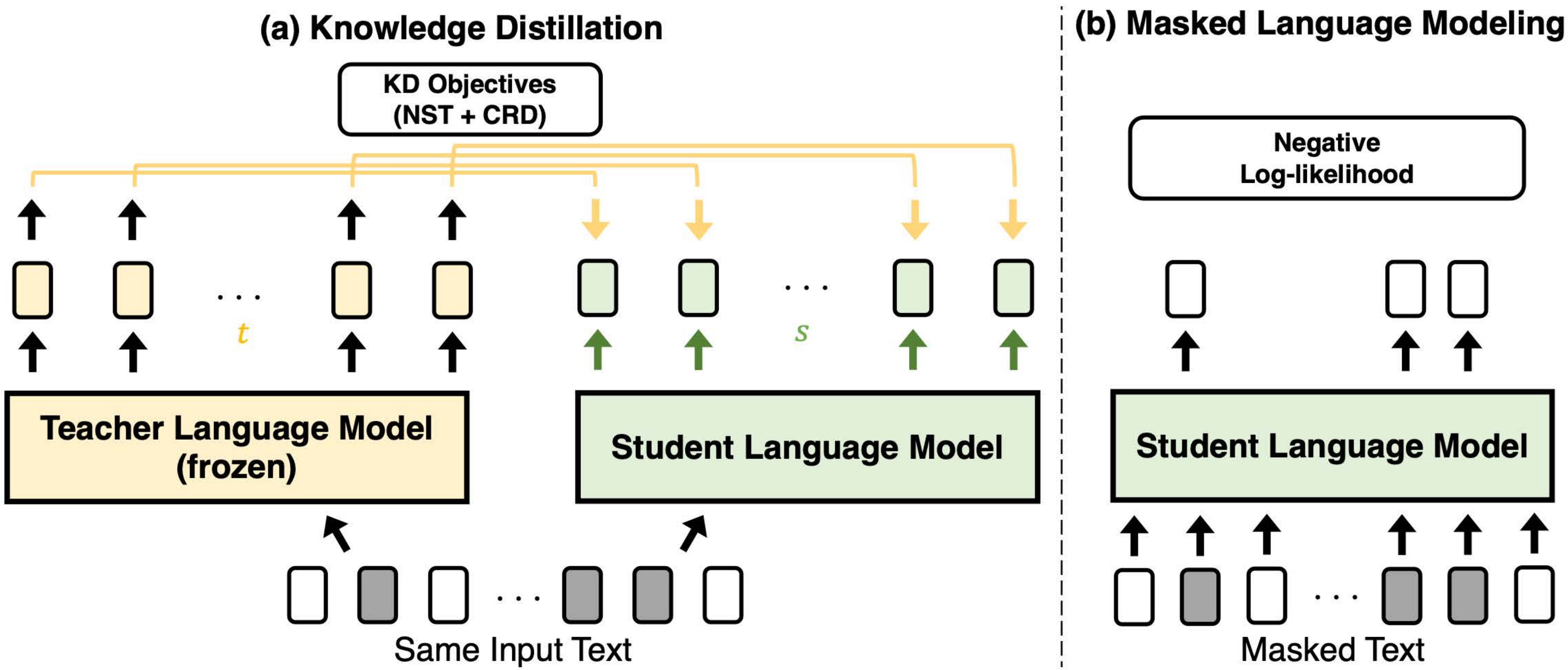
- The teacher LM is trained with (a) video-language triplet loss; + (b) masked language modeling



VIDLANKD: Improving Language Understanding via Video-Distilled Knowledge Transfer, Tang et al., NeurIPS 2021

VidLanKD: LM w/ Video-Distilled Knowledge

- The student LM is trained with (a) knowledge distillation; + (b) masked language modeling



VIDLANKD: Improving Language Understanding via Video-Distilled Knowledge Transfer, Tang et al., NeurIPS 2021

VidLanKD: LM w/ Video-Distilled Knowledge

- Cross-modal KD (last 2 rows) achieves better performance than image vokenization.

	SST-2 Acc	QNLI Acc	QQP Acc	MNLI Acc	SQuAD v1.1 EM†	SQuAD v2.0 EM	SWAG Acc	Avg.
$\text{BERT}_{12L/768H}$ [68]	89.3	87.9	83.2	79.4	77.0	67.7	65.7	78.6
+ KD (Img-Voken) [68]	92.2	88.6	88.6	82.6	78.8	68.1	70.6	81.4
$\text{BERT}_{12L/768H}$	89.0	88.0	86.2	79.2	77.2	68.0	65.0	78.9
+ KD (Vid-Voken) w/ ResNet	93.4	89.2	88.7	83.0	78.9	68.7	70.0	81.7
+ KD (Vid-Voken) w/ CLIP	94.1	**89.8**	89.0	83.9	79.2	68.6	71.6	82.3
+ KD (NST+CRD) w/ ResNet	94.2	89.3	89.7	84.0	79.0	**68.9**	71.8	82.4
+ KD (NST+CRD) w/ CLIP	**94.5**	89.6	**89.8**	**84.2**	**79.6**	68.7	**72.0**	**82.6**

- Performance gain is mostly from knowledge, physical interaction, & temporal reasoning

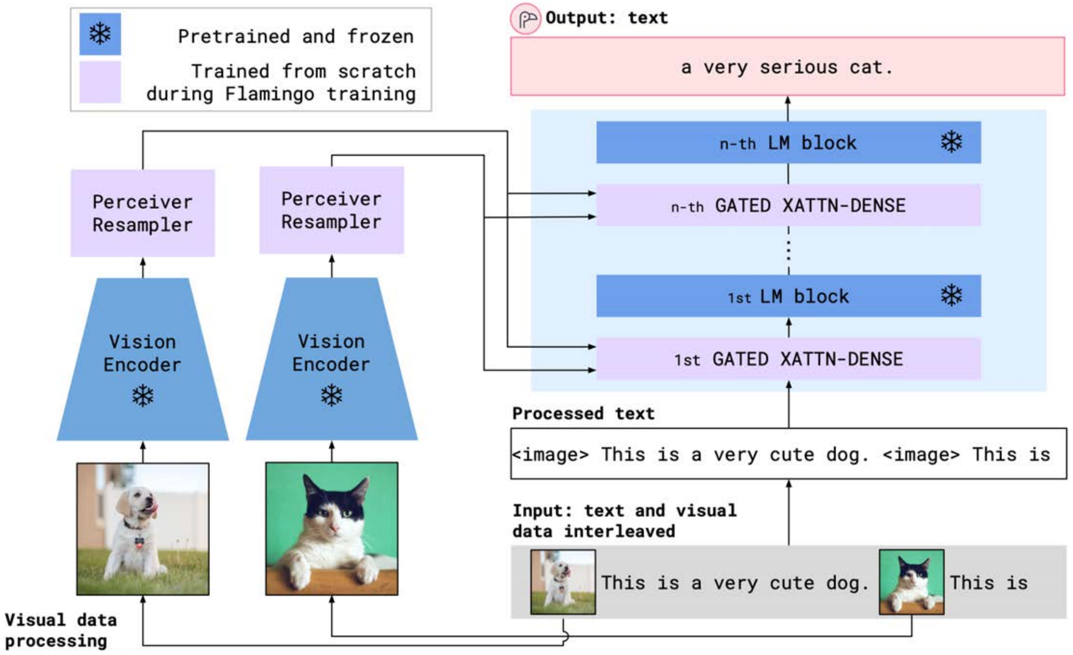
	GLUE diagnostics				PIQA	TRACIE
	Lexicon	Predicate	Logic	Knowledge		
$\text{BERT}_{6L/512H}$	53.0	64.2	44.5	44.0	56.9	63.4
+ KD-NST	53.3 (+0.3)	63.7 (-0.5)	44.8 (+0.3)	48.6 (**+4.6**)	60.0 (**+3.1**)	66.7 (**+3.3**)

PIQA: QA w/ physical interactions + commonsense reasoning
TRACIE: a temporal reasoning benchmark

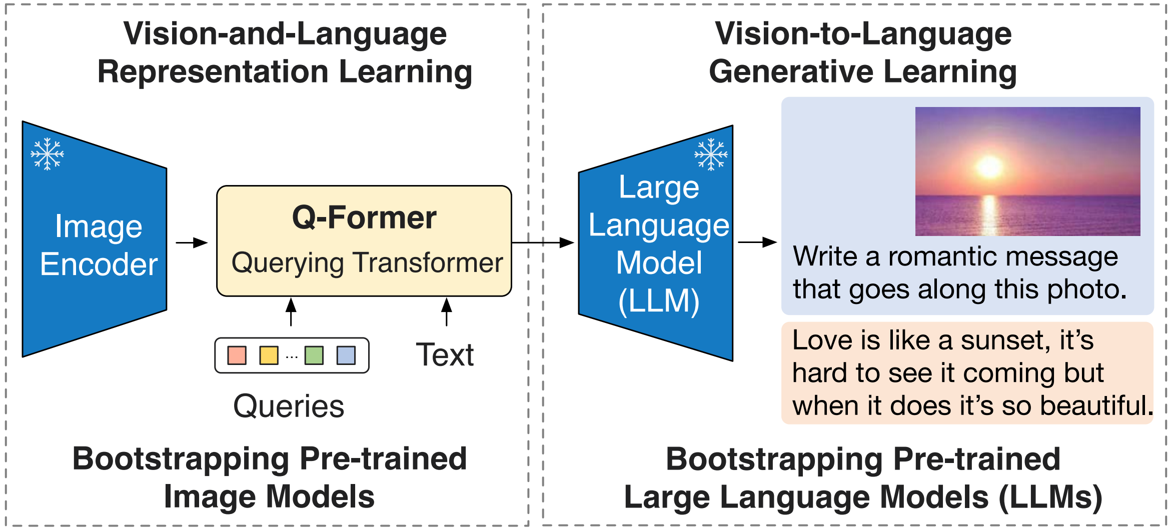
VIDLANKD: Improving Language Understanding via Video-Distilled Knowledge Transfer, Tang et al., NeurIPS 2021

Future Work

- How to better bridge LLM and other modalities?
    - Is frozen LLM the best approach?
    -



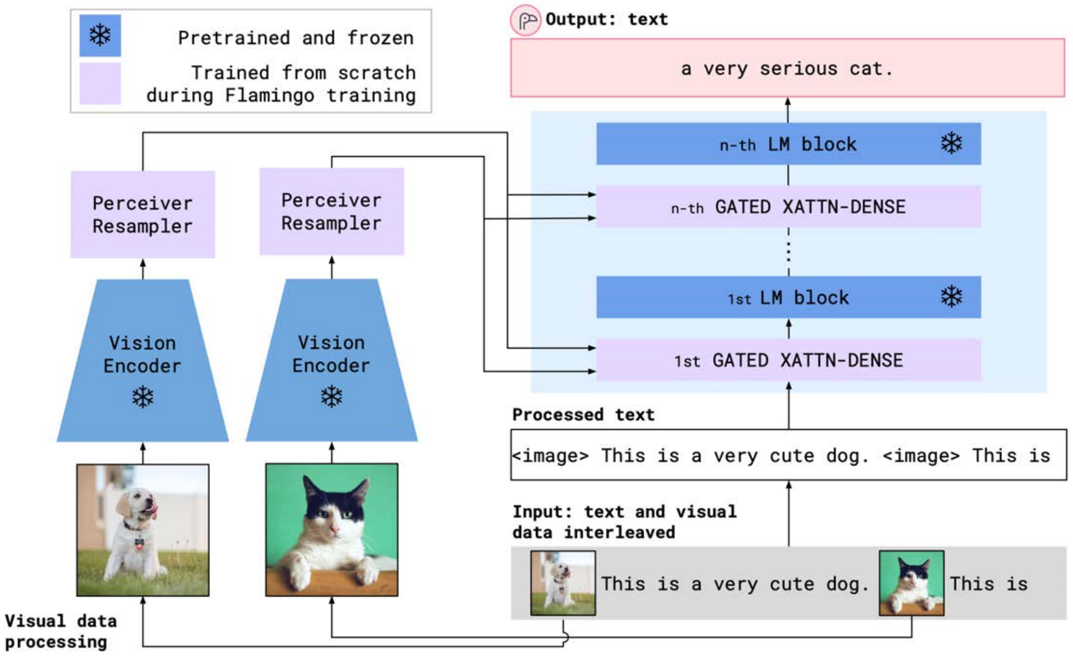
Flamingo



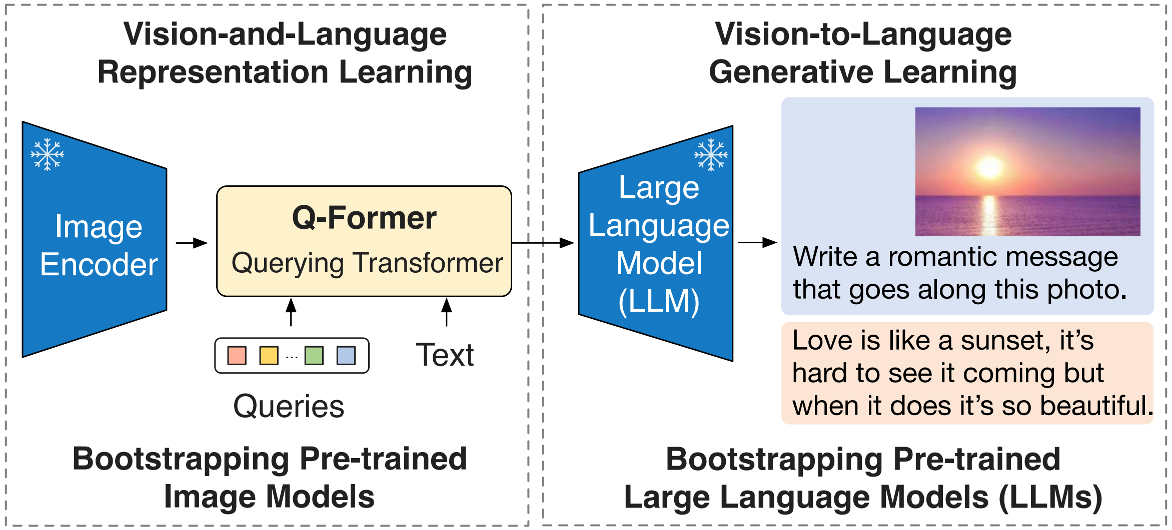
BLIP-2

Future Work

- How to better bridge LLM and other modalities?
  - Is frozen LLM the best approach?
  - If more than one modalities are needed, how to better model them together?



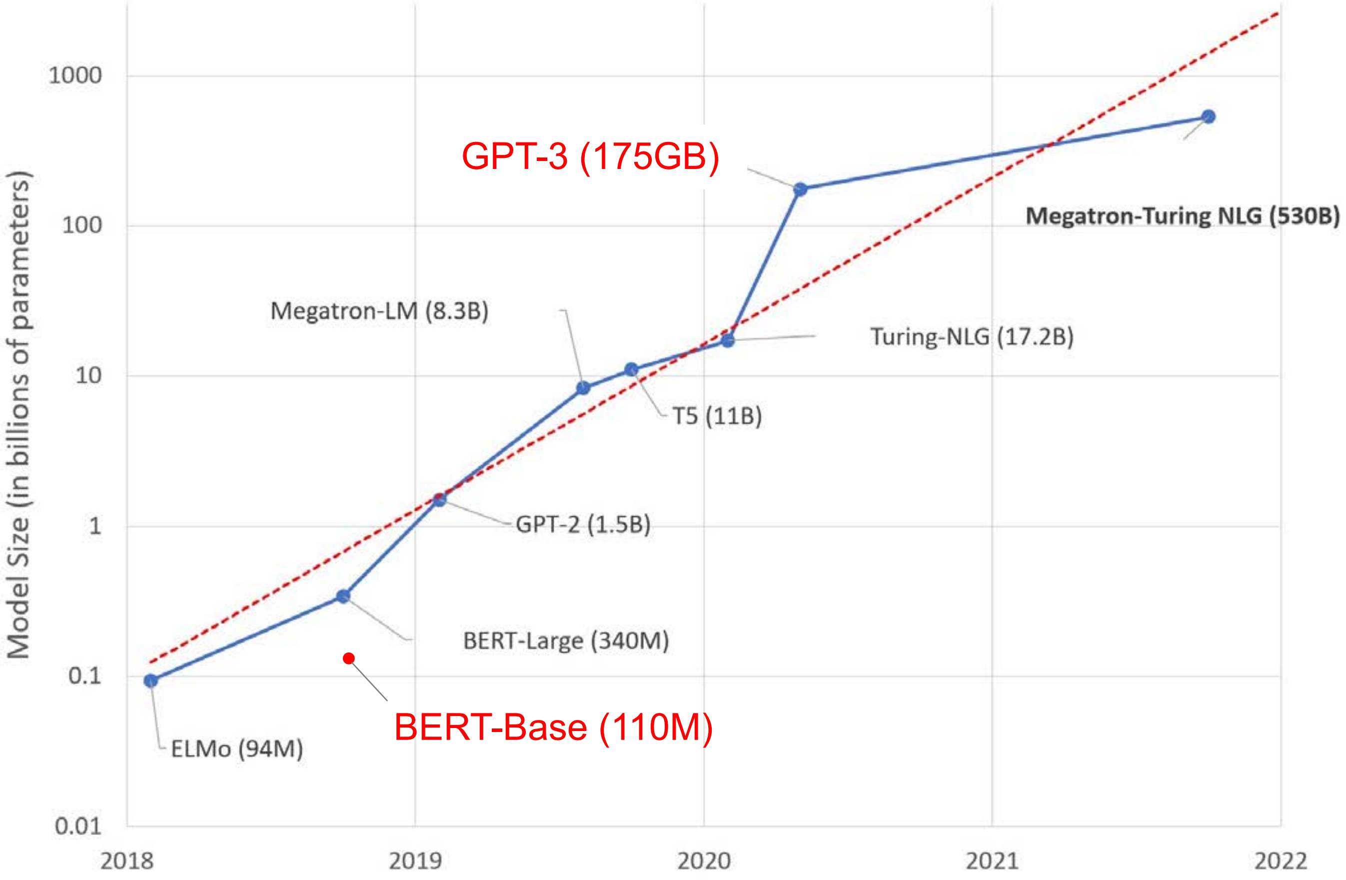
Flamingo



BLIP-2

Future Work

- Using vision (image or video) supervision has shown some early success.



- Bidirectional LM only, casual LM is not explored.
- Small model (up to 110M BERT-base), *vs.*, 175B GPT-3

- How about using other modalities (audio) as supervision?

Figure credit to Julien Simon, https://huggingface.co/blog/large-language-models

Take-way Messages

L → V: Implicit Knowledge

Training vision model w/ frozen LM

- Preserves the in-context learning ability of LM.
- Larger LM is better, the same as pure language tasks.
- They are quite general and are applicable to a wide range of tasks.

Convert multimodal task as text task for LM

- All of above.
- Computation efficient: no finetuning is required.
- Inherently modular, easy to update individual modules.
- May suffer info loss when during the conversion to text.

L → V: Explicit Knowledge

- Human curated (e.g., wordnet) or LLM (GPT-3) improves image classification and object detection, especially these with rare concepts.

V → L

- Vision knowledge via vokenization or distillation improves LMs, especially for physical and commonsense knowledge, and temporal reasoning.

Future work

- Efficient learning with LLM
- Scaling-up vision supervised LMs
- …

Thanks!

CVPR

Jun 2023 9:00-12:30
CVPR Tutorials
East 8, Vancouver Convention Center

Knowledge-Driven Vision-Language Encoding

Panel 1: Explicit Knowledge vs Implicit Knowledge

What is the appropriate format of knowledge representation?

Does explicit knowledge still have value in the era of large models?

Mohit Bansal
UNC

Carl Vondrick
Columbia

Xudong Lin
Columbia

Jie Lei
Meta AI

Manling Li
UIUC

CVPR

Jun 2023 9:00-12:30
CVPR Tutorials
East 8, Vancouver Convention Center

Knowledge-Driven Vision-Language Encoding

Panel 2: LLMs for Multimodality

What can we borrow from Large Language Models (LLMs)?

Mohit Bansal
UNC

Carl Vondrick
Columbia

Xudong Lin
Columbia

Jie Lei
Meta AI

Manling Li
UIUC

CVPR

Jun 2023 9:00-12:30
CVPR Tutorials
East 8, Vancouver Convention Center

Knowledge-Driven Vision-Language Encoding

Panel 3: Image vs Video vs Audio vs Embodied AI

What is the bottleneck for each single modality?

What is the bottleneck to bring multiple modalities together?

Mohit Bansal
UNC

Carl Vondrick
Columbia

Xudong Lin
Columbia

Jie Lei
Meta AI

Manling Li
UIUC

CVPR

Jun 2023 9:00-12:30
CVPR Tutorials
East 8, Vancouver Convention Center

Knowledge-Driven Vision-Language Encoding

Panel 4: Open Challenges

What is the recommended thesis topic for next few years?

Mohit Bansal
UNC

Carl Vondrick
Columbia

Xudong Lin
Columbia

Jie Lei
Meta AI

Manling Li
UIUC