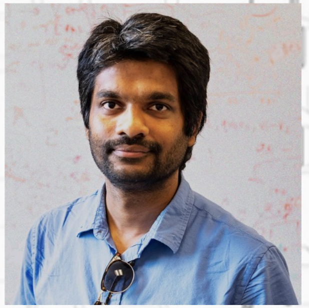
Knowledge-Driven Vision-Language Pretraining

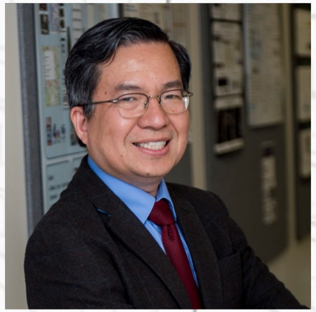
Manling Li
UIUC

Xudong Lin
Columbia

Jie Lei
Meta AI



Mohit Bansal
UNC



Shih-Fu Chang
Columbia

Heng Ji
UIUC

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

COLUMBIA UNIVERSITY

UNC

Meta AI

Feb 2023
AAAI Tutorials
Knowledge-Driven Vision-Language Pretraining

Factual Knowledge in V+L Pretraining: Information about Instances

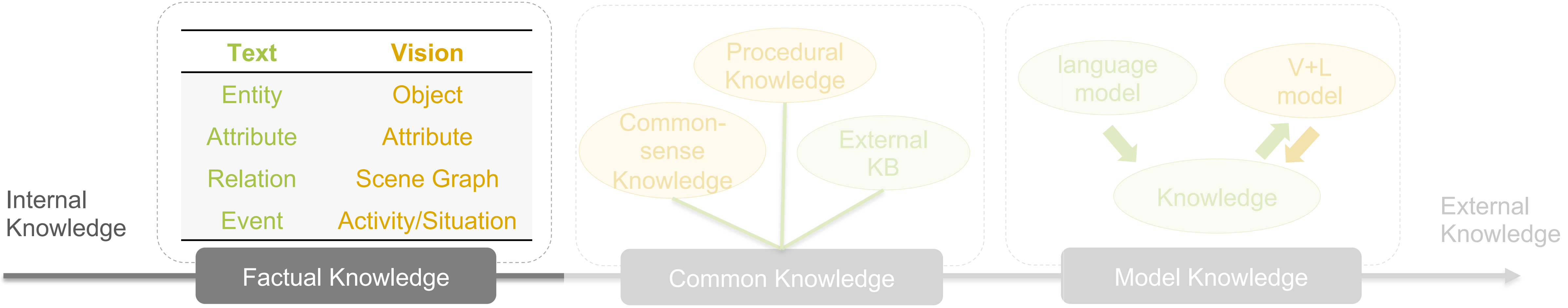
Knowledge-Driven Vision-Language Pretraining (Part II)

Manling Li
UIUC
manling2@illinois.edu

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN
COLUMBIA UNIVERSITY
UNC
Meta AI

Factual Knowledge

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

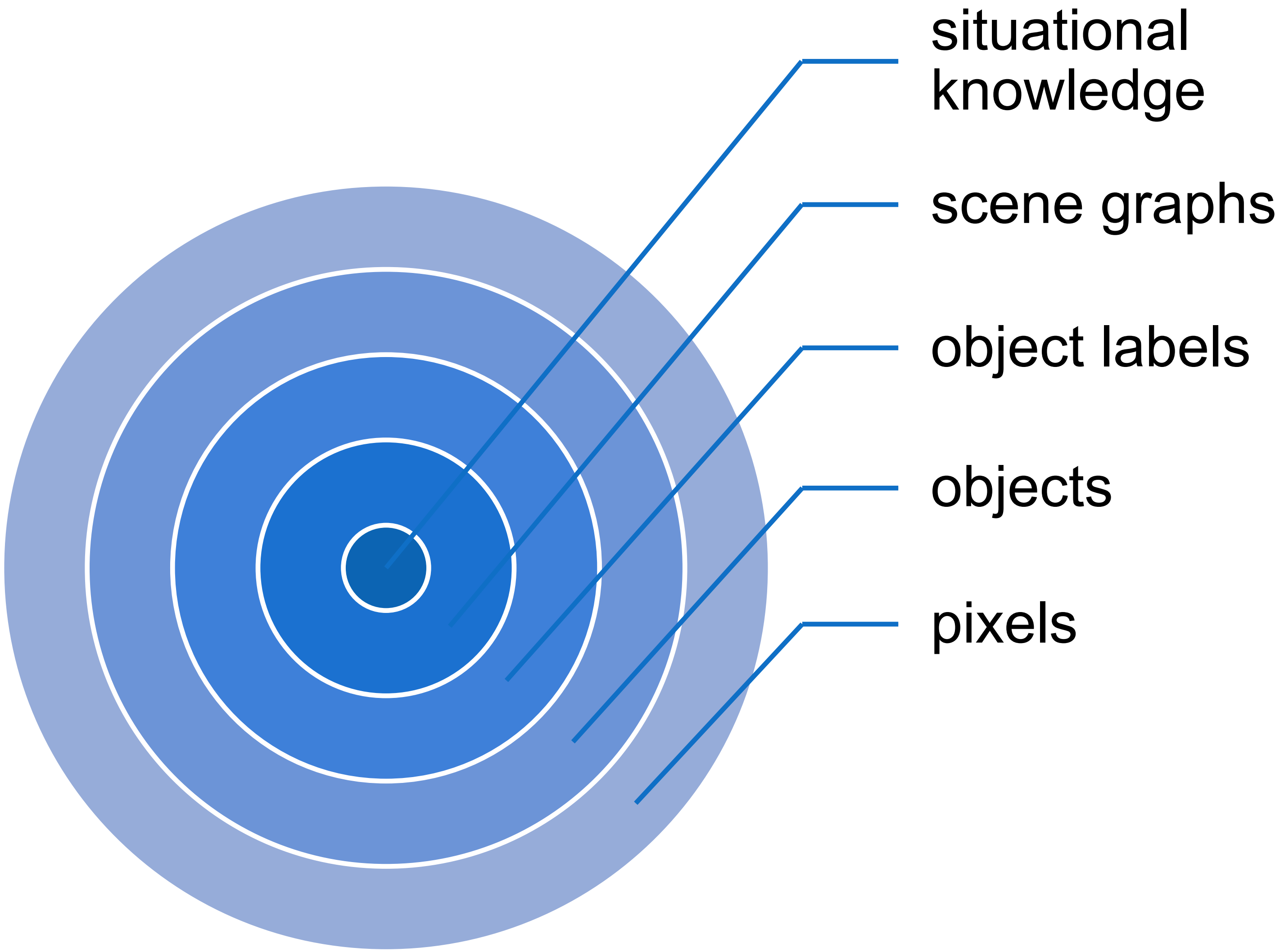


Text	Vision
Entity	Object
Attribute	Attribute
Relation	Scene Graph
Event	Activity/Situation

Adding knowledge to pretraining models

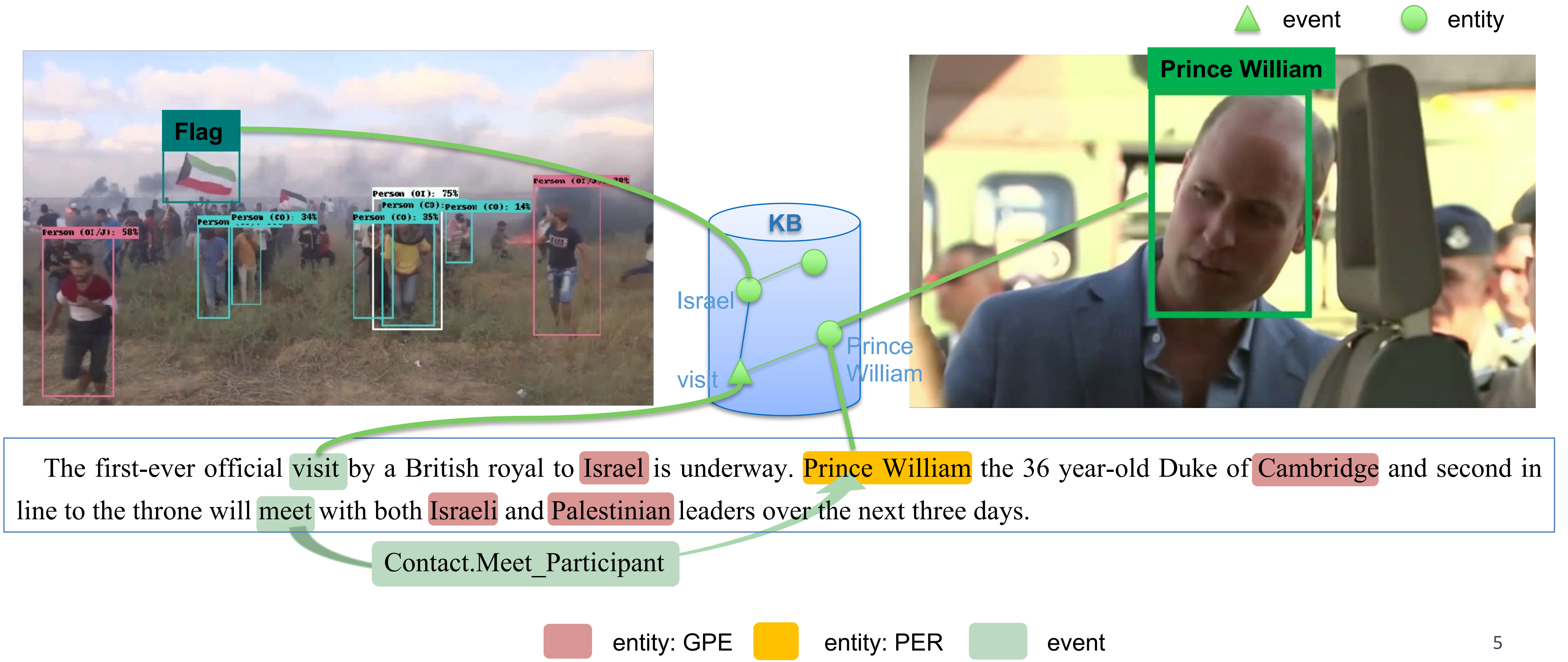
Knowledge for Model

Knowledge for data

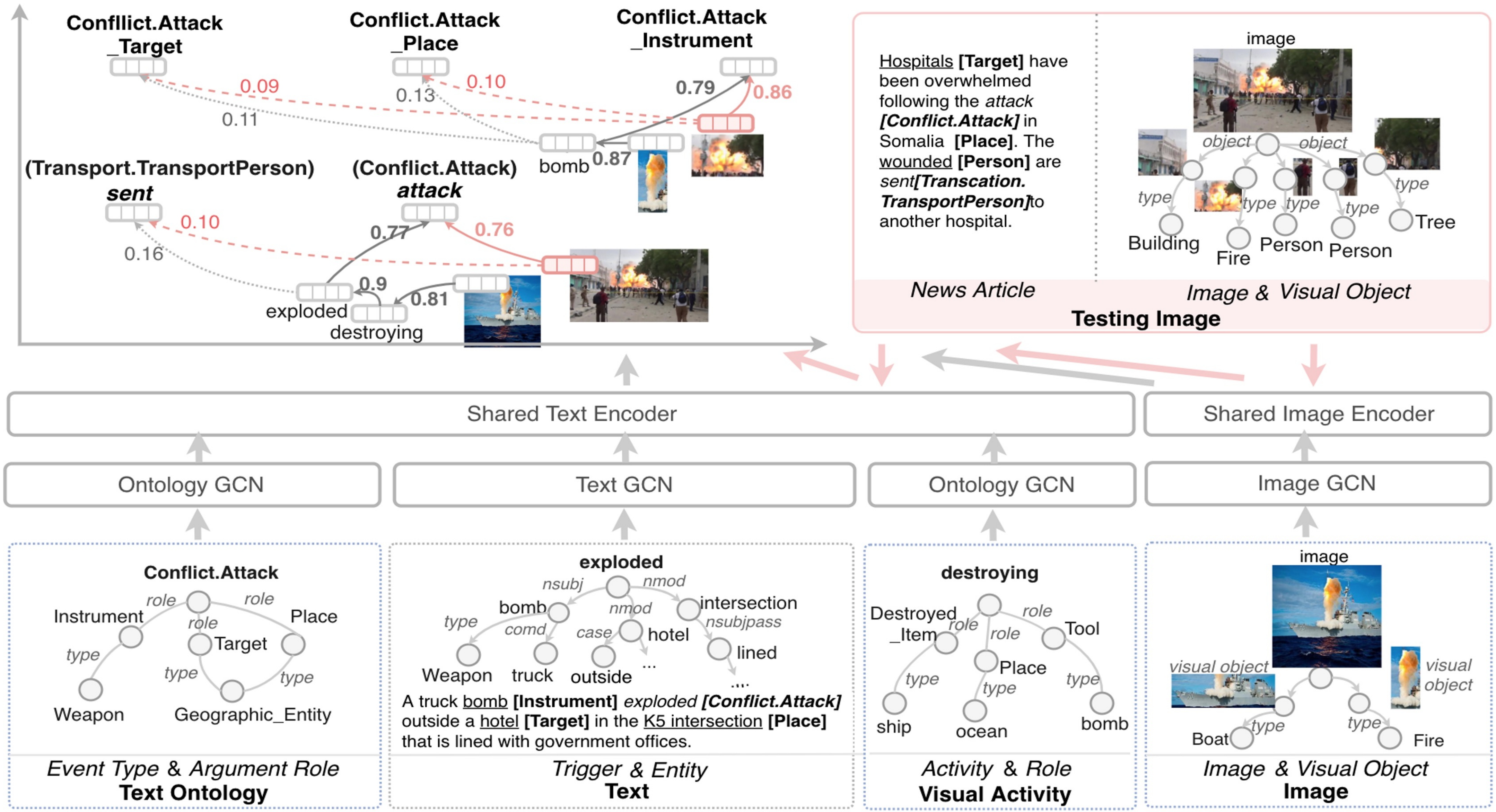


What is factual knowledge?

- Multimedia Knowledge Base with entities, relations and events.



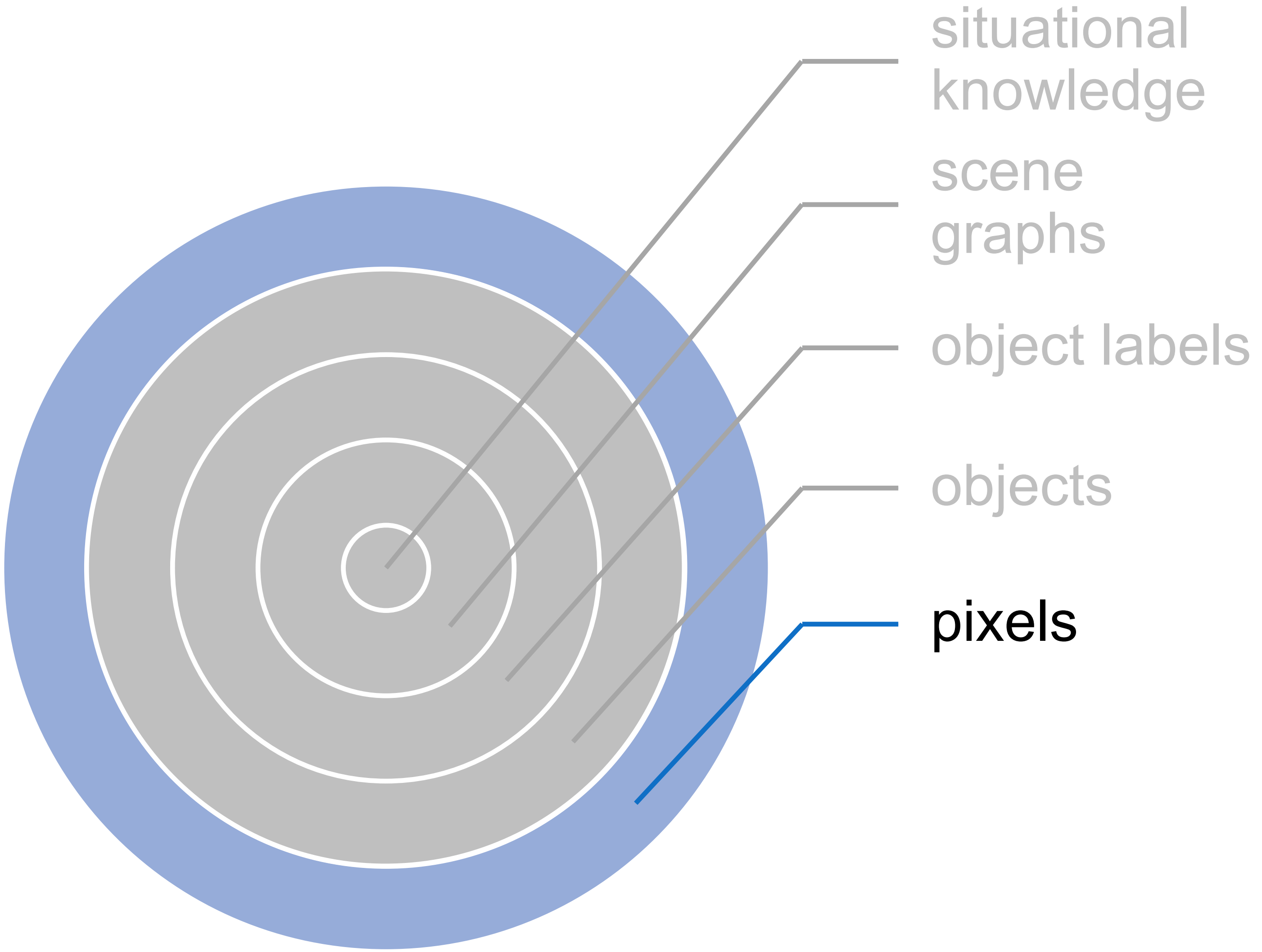
Goal: A joint representation of text and vision knowledge



Adding knowledge to pretraining models

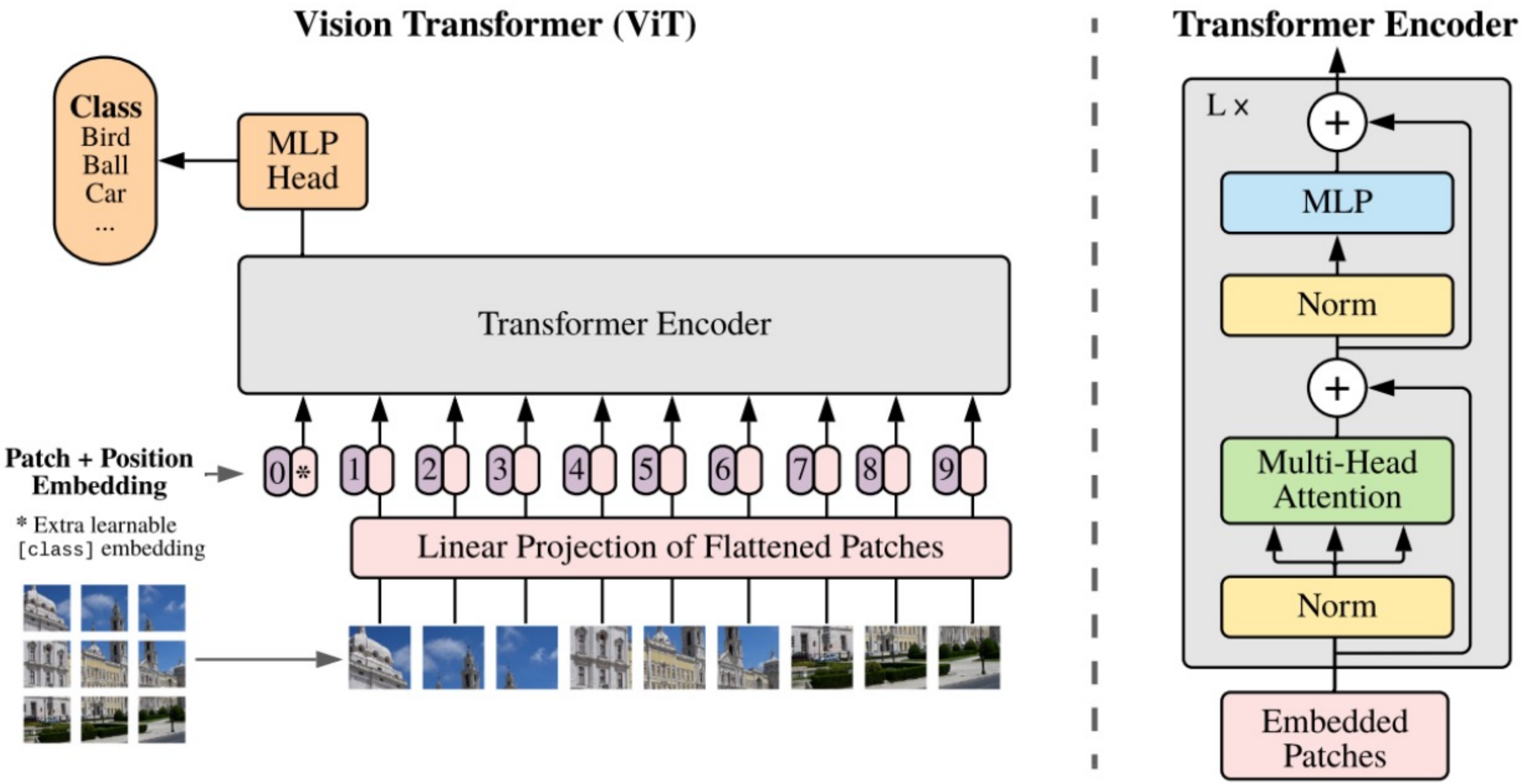
Knowledge for Model

Knowledge for data



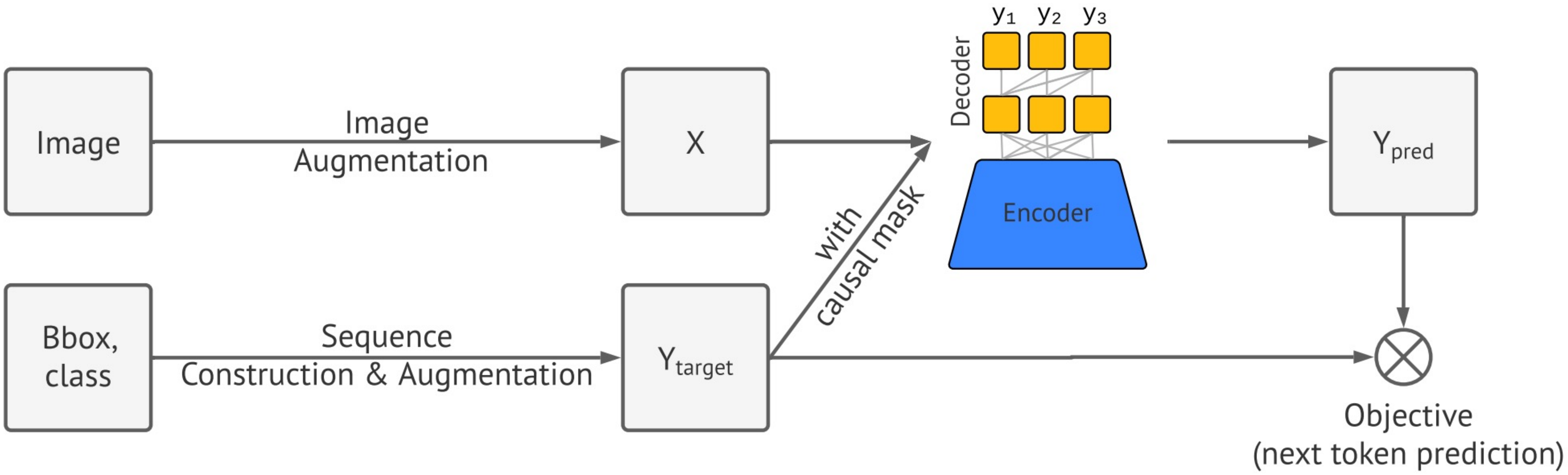
An Image is Worth 16x16 Words

The simplest way is to split an image into patches



Unified Model: Pix2Seq

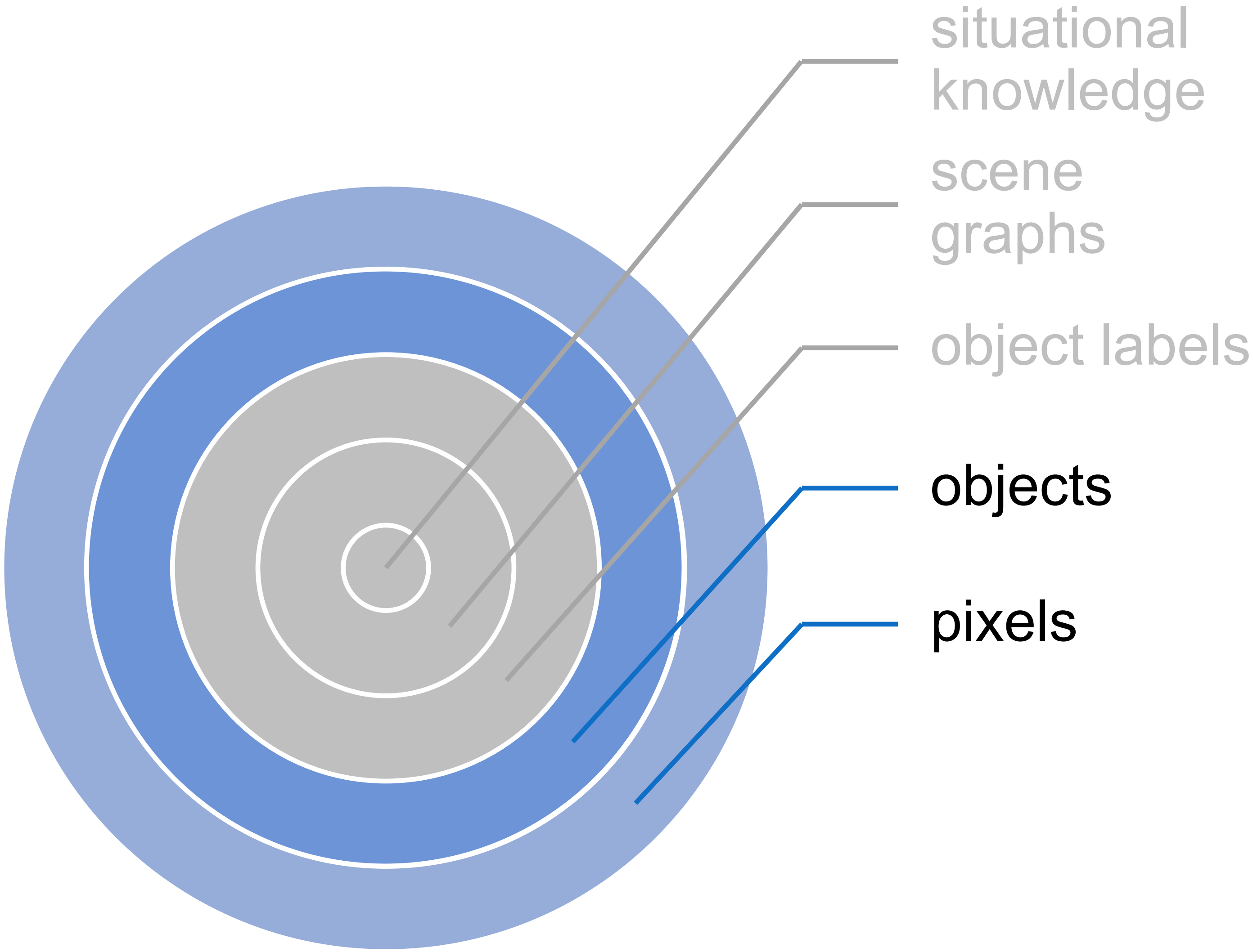
Another way is to treat pixels as tokens.



Adding knowledge to pretraining models

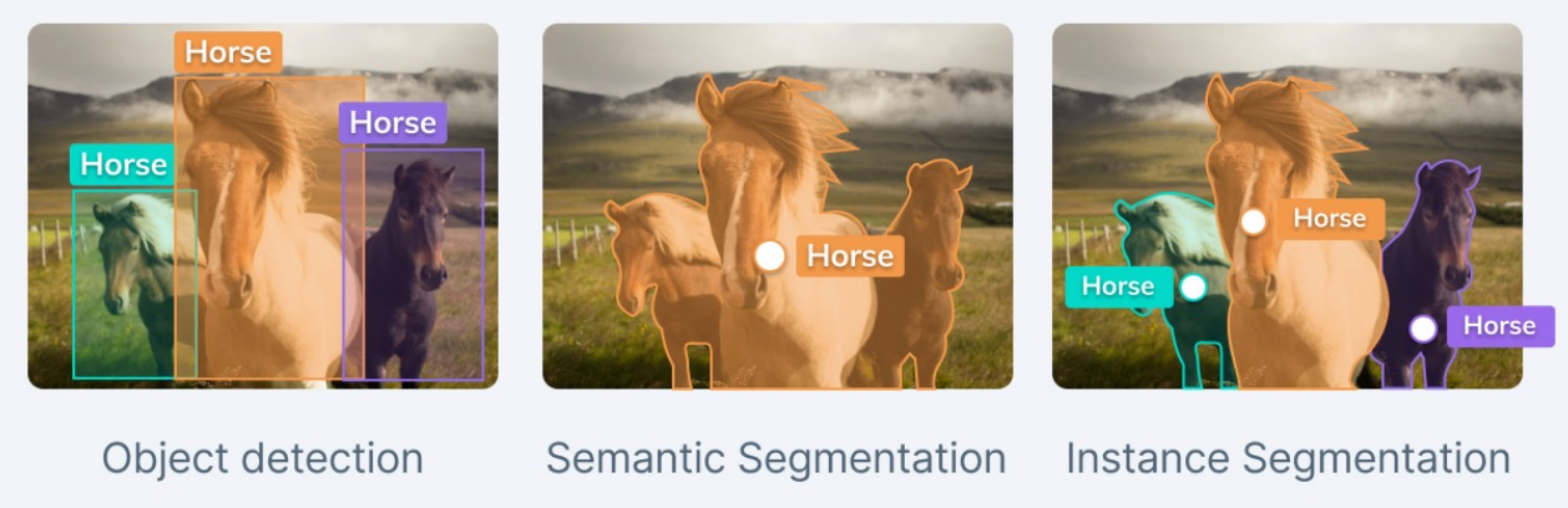
Knowledge for Model

Knowledge for data



Entity Knowledge

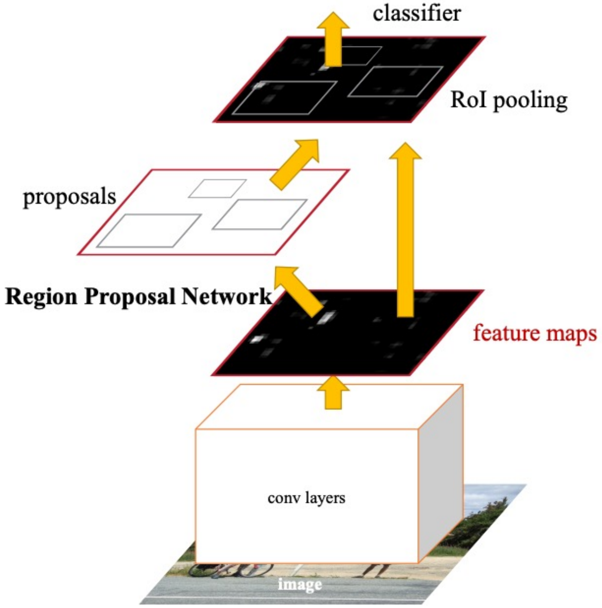
- Object Detection: Object instances at the bounding box level
- Semantic Segmentation: Object class at the pixel level
- Instance Segmentation: Object instances at the pixel level



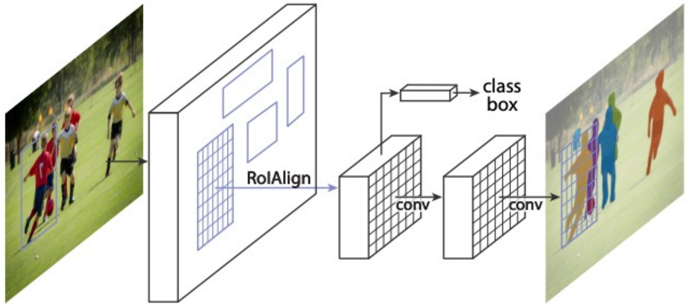
https://www.v7labs.com/blog/object-detection-guide

The way to obtain entity knowledge: Object Extraction

- Two-Stage (With Proposal)

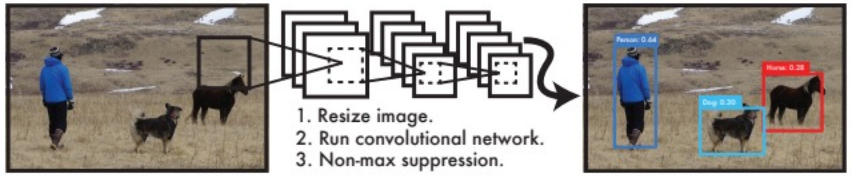


Faster RCNN

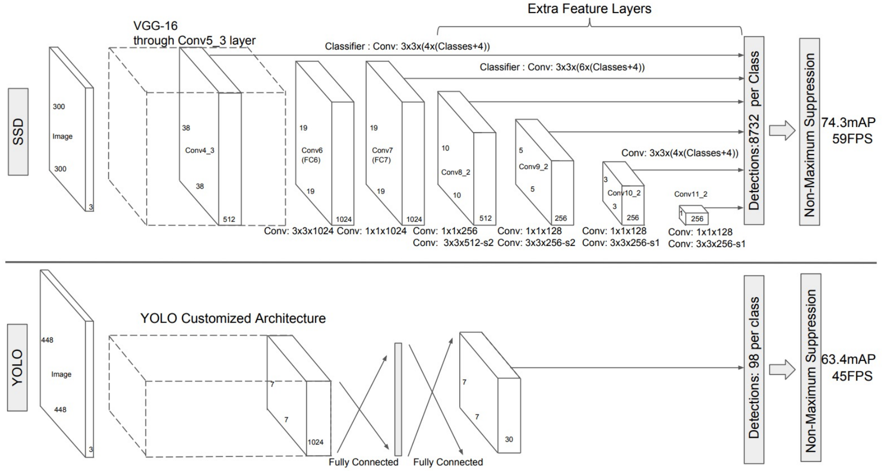


Mask RCNN

- One-Stage (Without Proposal)



YOLO

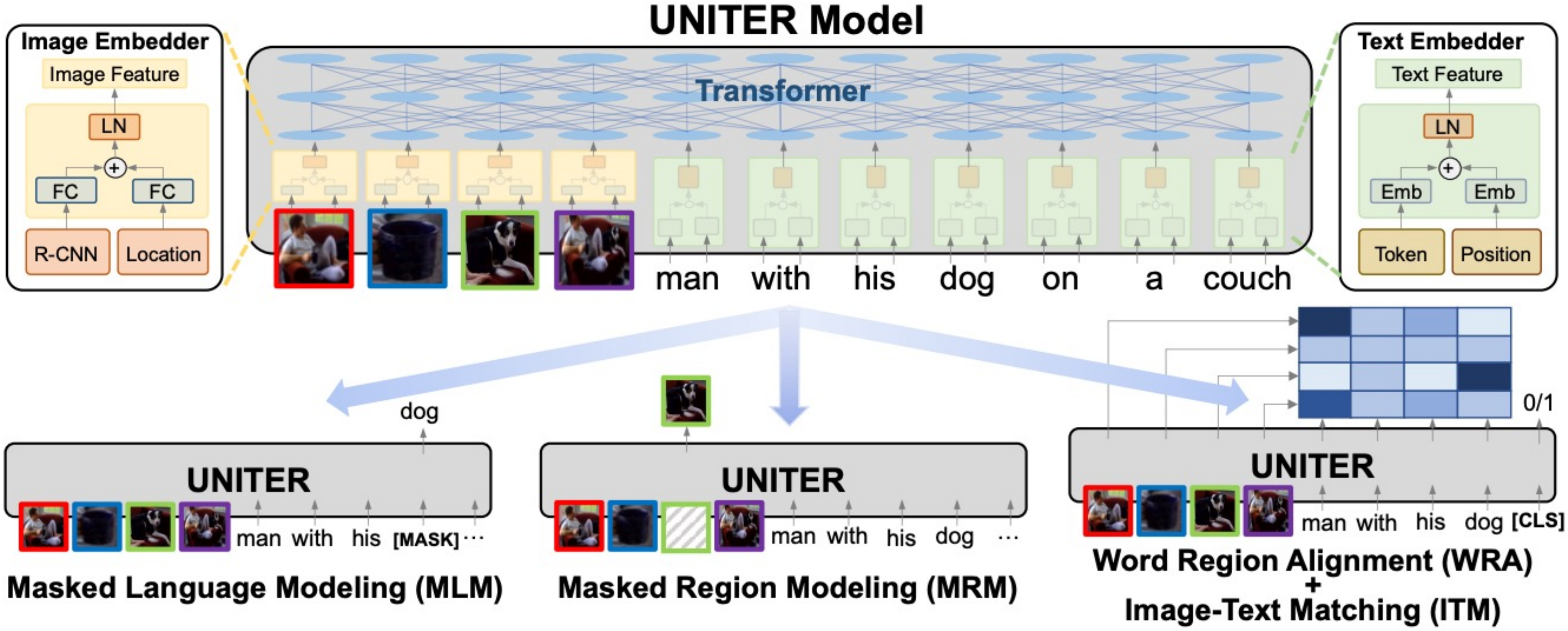


SSD

Ren, S., He, K., Girshick, R., & Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS 2015*.
He, Kaiming, et al. "Mask r-cnn." *CVPR* 2017.
Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." *CVPR* 2016.
Liu, Wei, et al. "Ssd: Single shot multibox detector." *ECCV* 2016.

Adding objects to V+L Pretraining

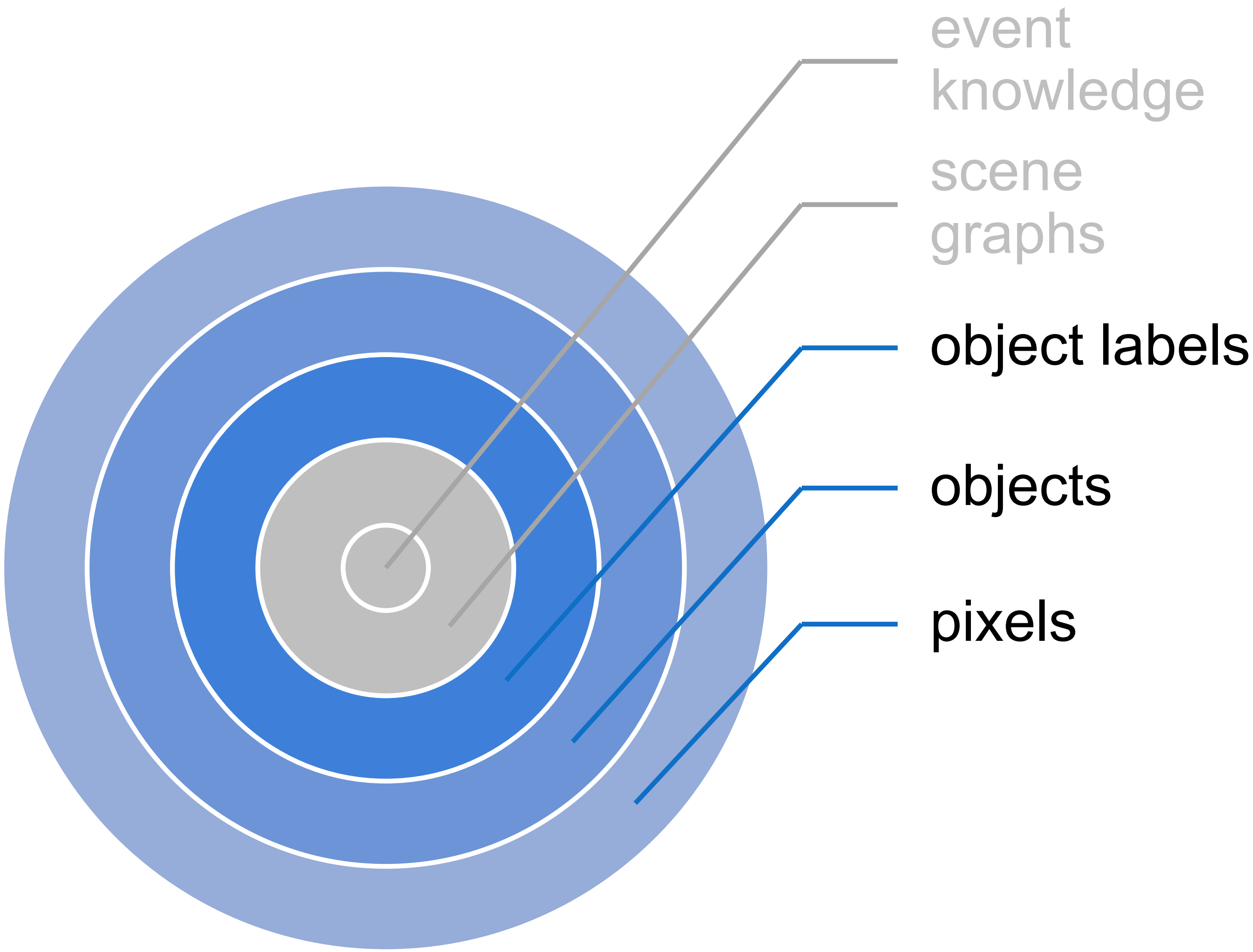
Objects are used to better mask the regions.



Adding knowledge to pretraining models

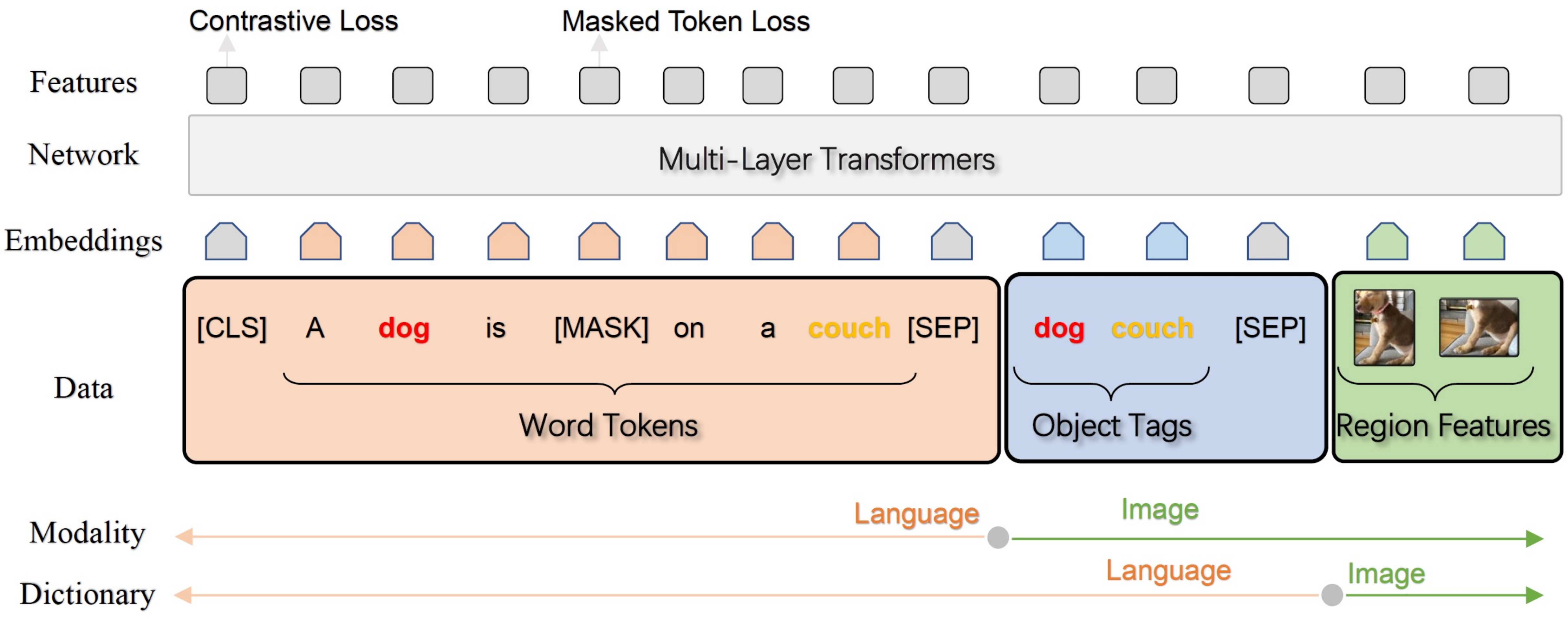
Knowledge for Model

Knowledge for data



Oscar [ECCV 2020] and VinVL [CVPR 2021]

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

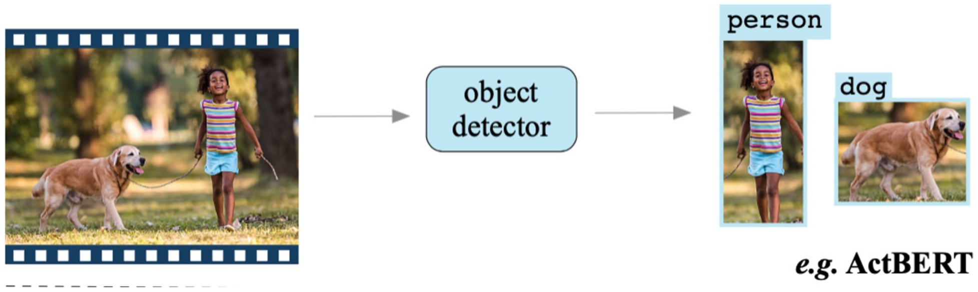


Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks, ECCV 2020
VinVL: Making Visual Representations Matter in Vision-Language Models, CVPR 2021

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

Individually pre-trained with Video-Text Contrastive loss



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

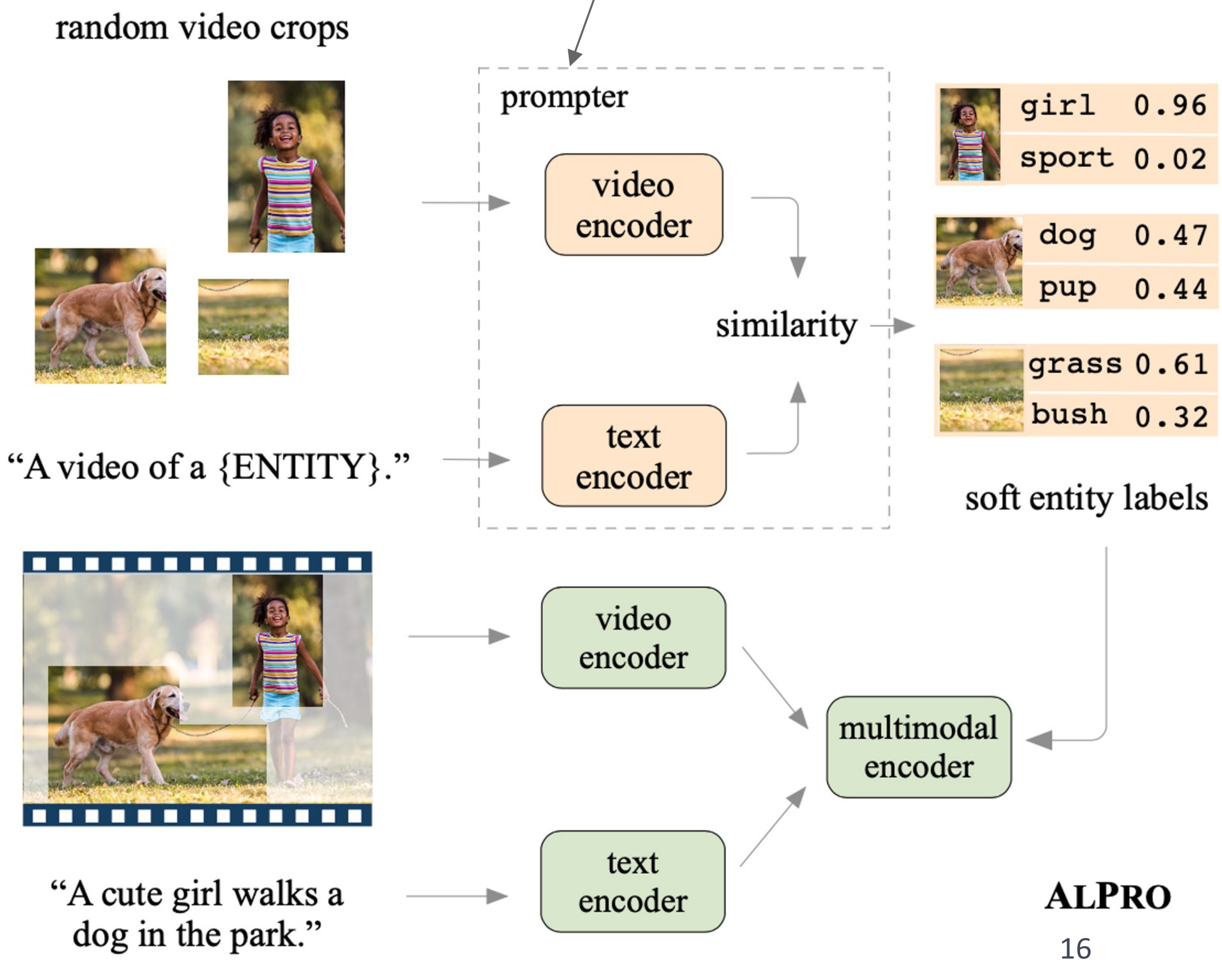
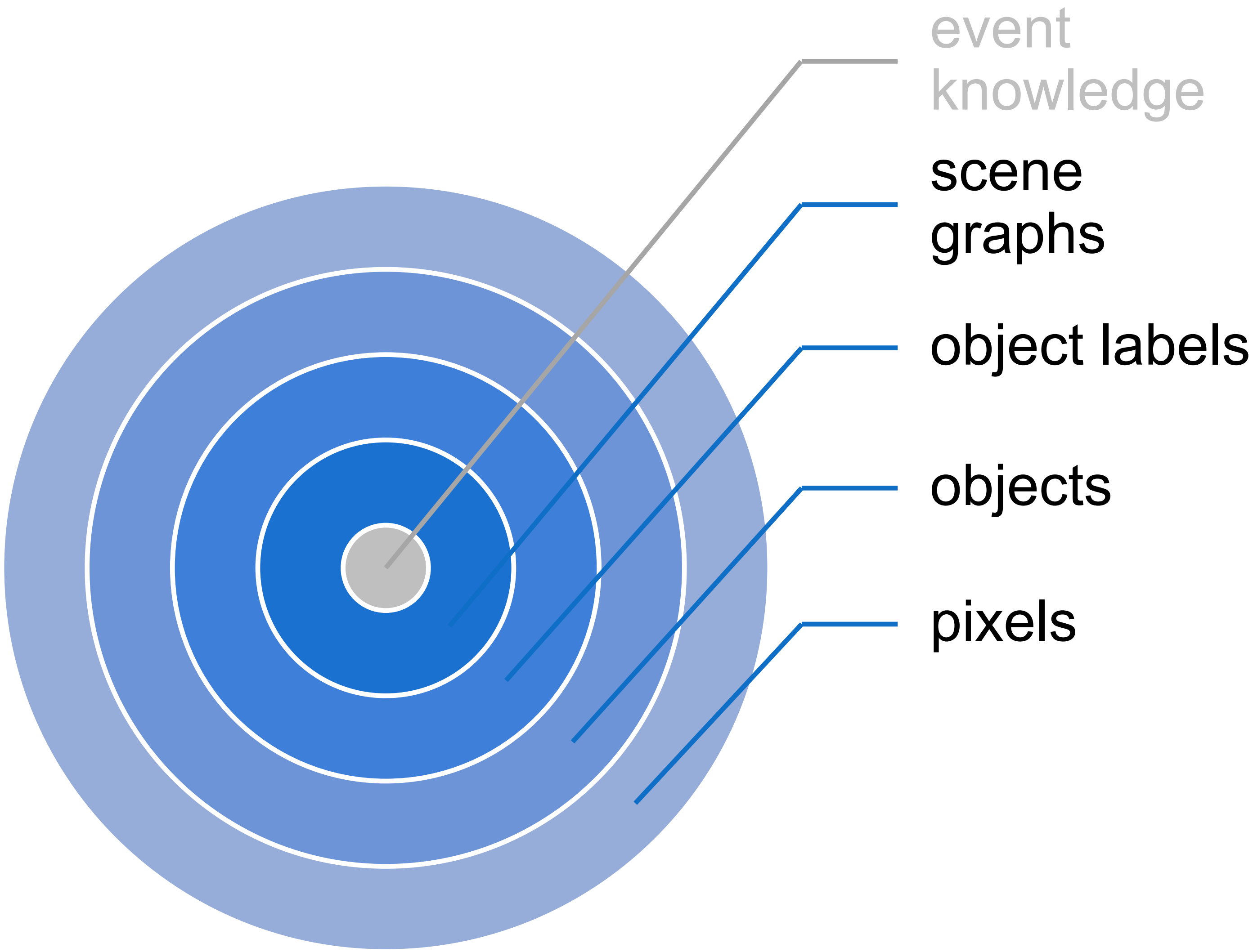


image source: Align and Prompt: Video-and-Language Pre-training with Entity Prompts

Adding knowledge to pretraining models

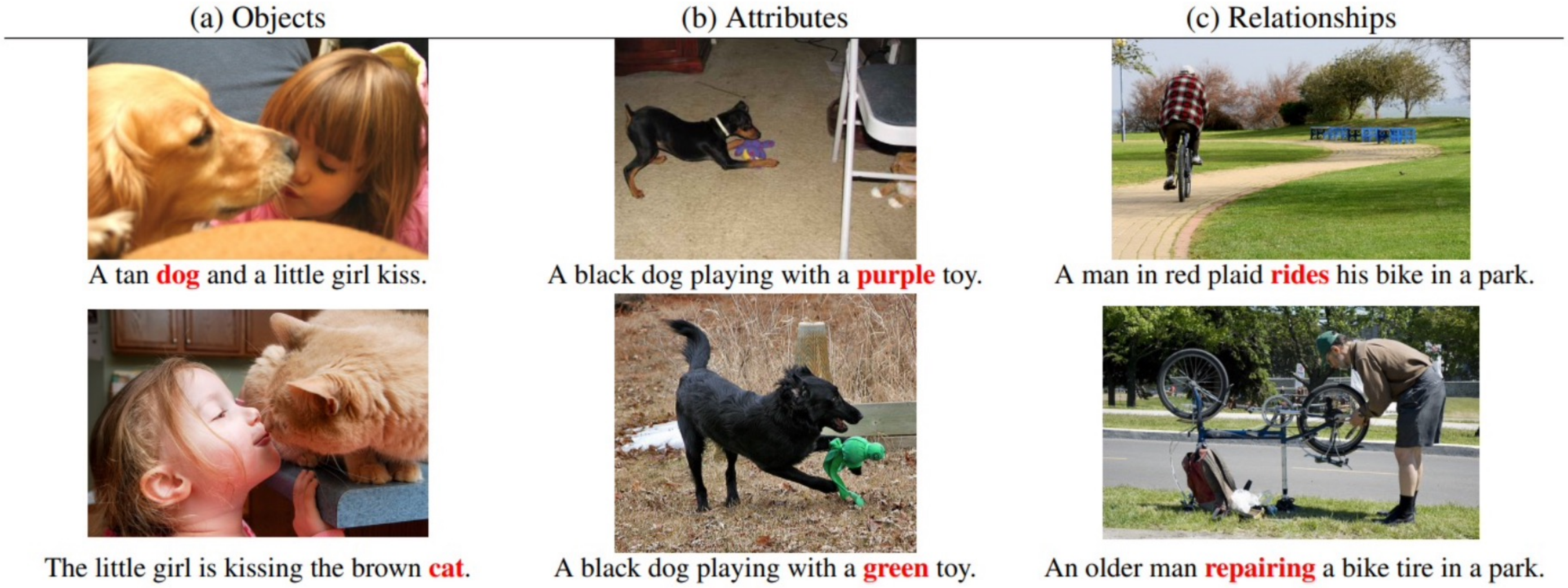
Knowledge for Model

Knowledge for data



ERINE-ViL [AAAI2021]

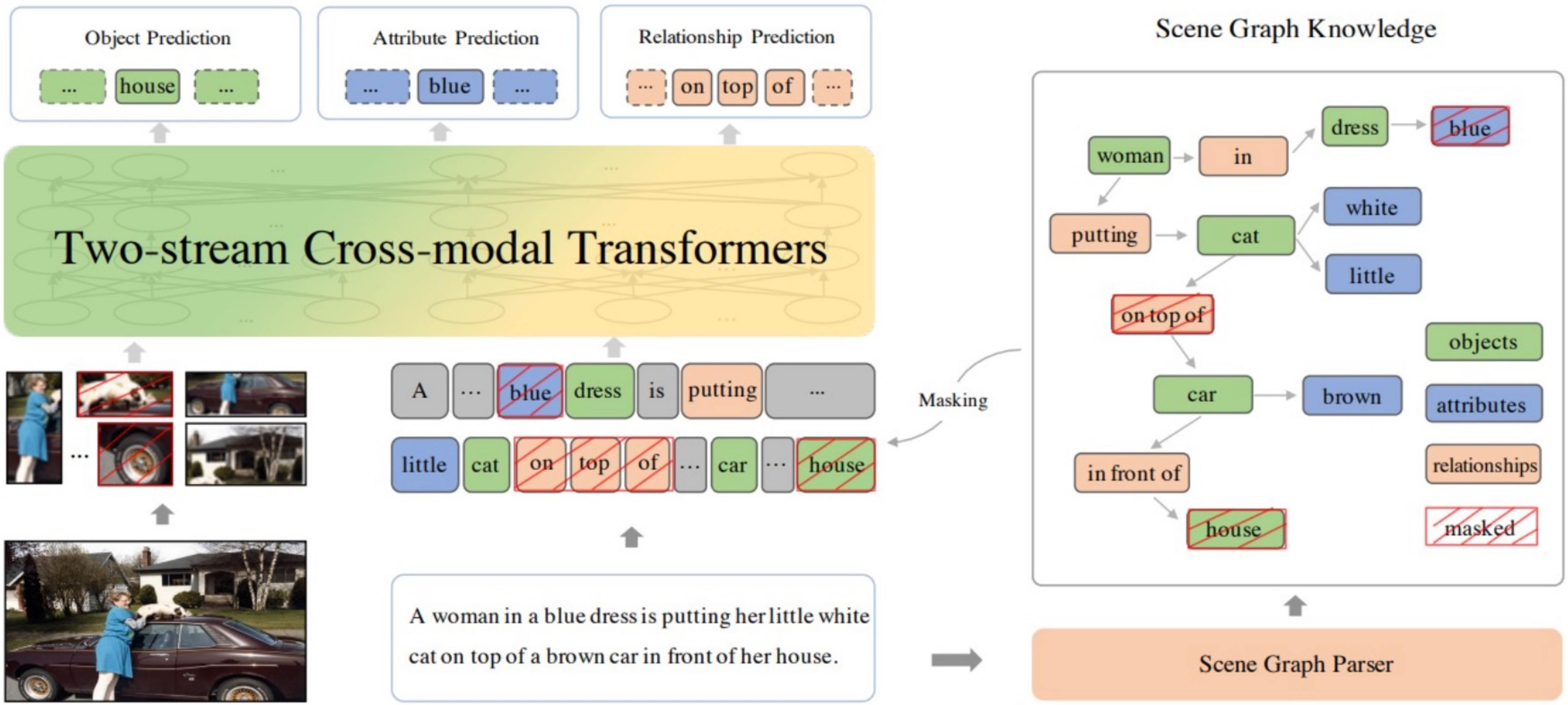
- Add scene graph knowledge as downstream tasks
  - Object prediction
  - Attribute prediction
  - Relationship prediction



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

ERINE-ViL [AAAI2021]

- Add scene graph knowledge as downstream tasks

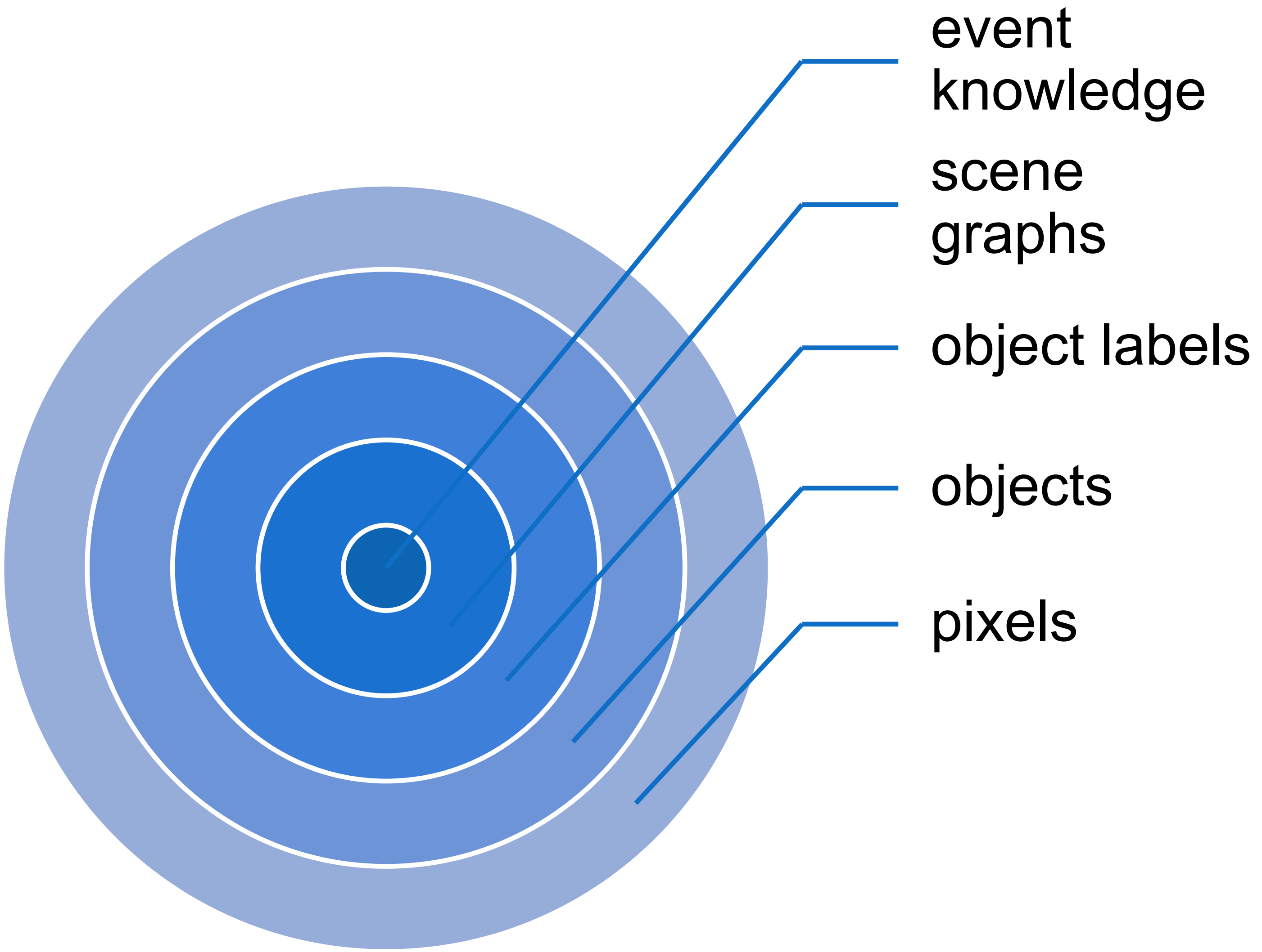


ERNIE-ViL: Knowledge Enhanced Vision-Language Representations Through Scene Graph, AAAI 2021

Adding knowledge to pretraining models

Knowledge for Model

Knowledge for data



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

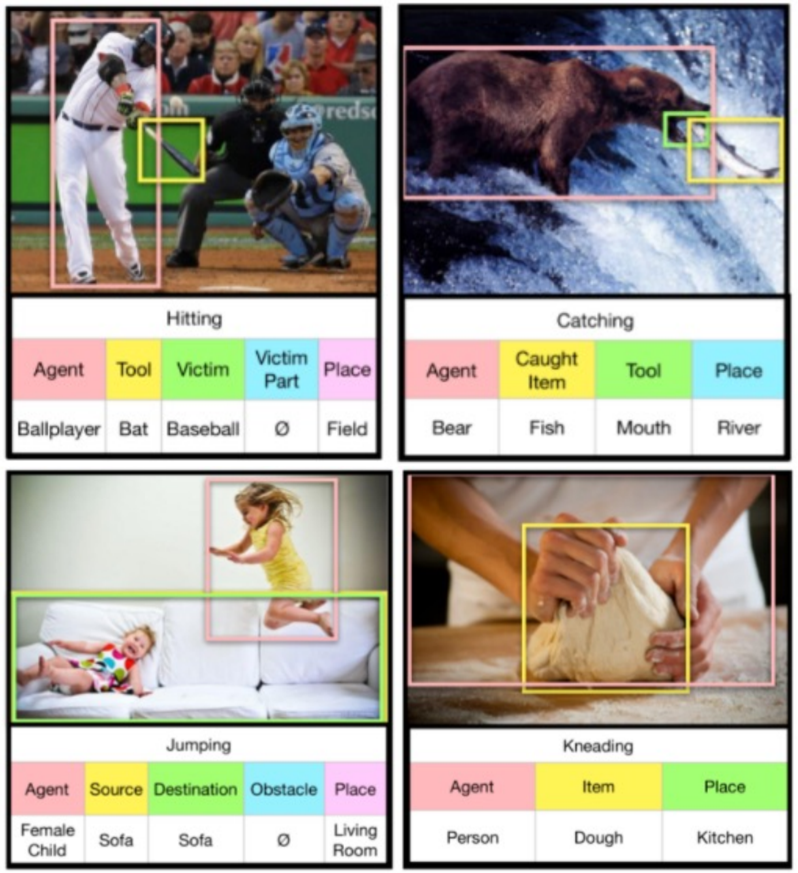
ROLE	VALUE	ROLE	VALUE
AGENT	BOY	AGENT	BEAR
SOURCE	CLIFF	SOURCE	ICEBERG
OBSTACLE	-	OBSTACLE	WATER
DESTINATION	WATER	DESTINATION	ICEBERG
PLACE	LAKE	PLACE	OUTDOOR

SPRAYING

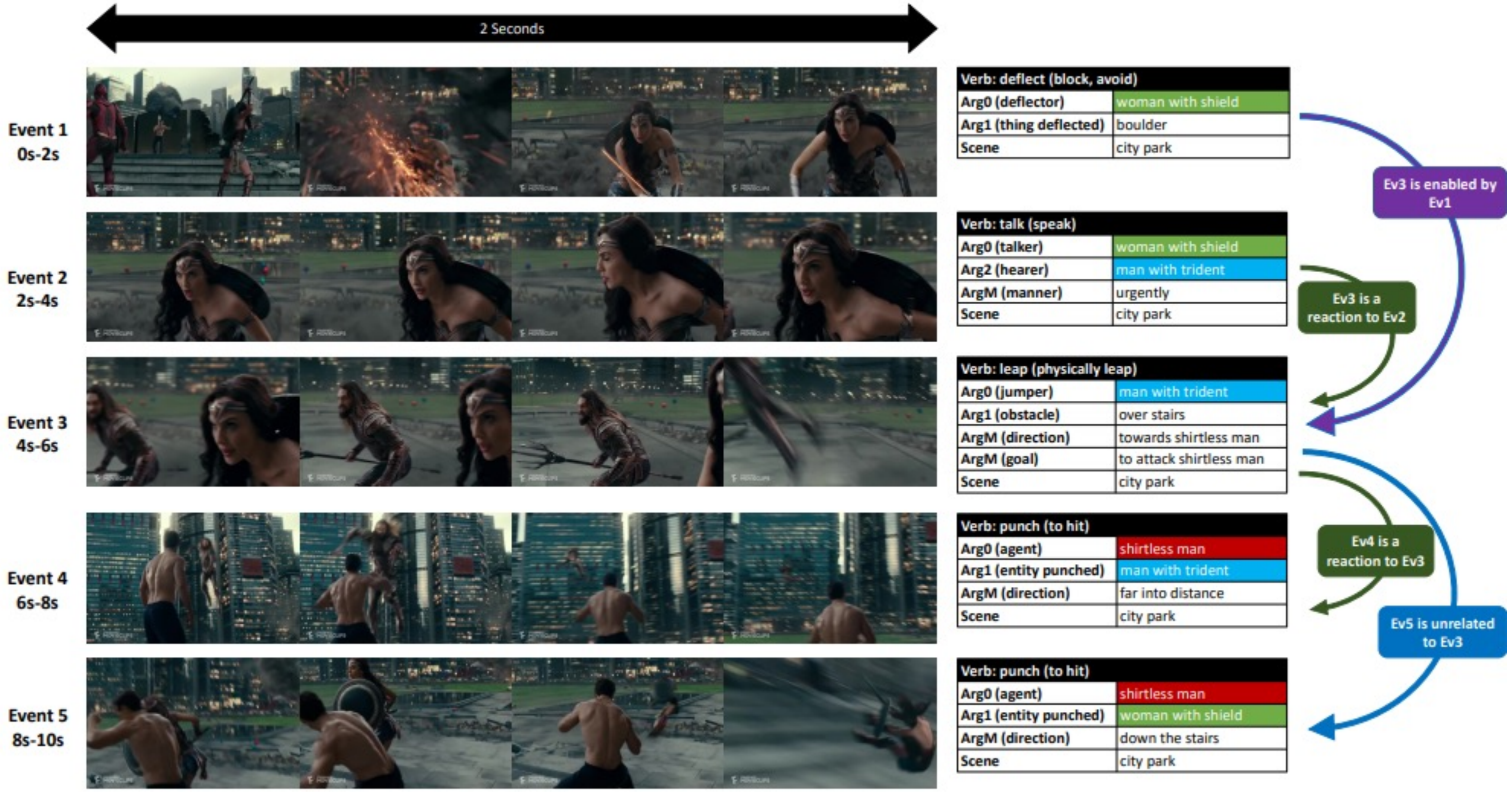
ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	FIREMAN
SOURCE	SPRAY CAN	SOURCE	HOSE
SUBSTANCE	PAINT	SUBSTANCE	WATER
DESTINATION	WALL	DESTINATION	FIRE
PLACE	ALLEYWAY	PLACE	OUTSIDE

Vision-only Event and Argument Extraction

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

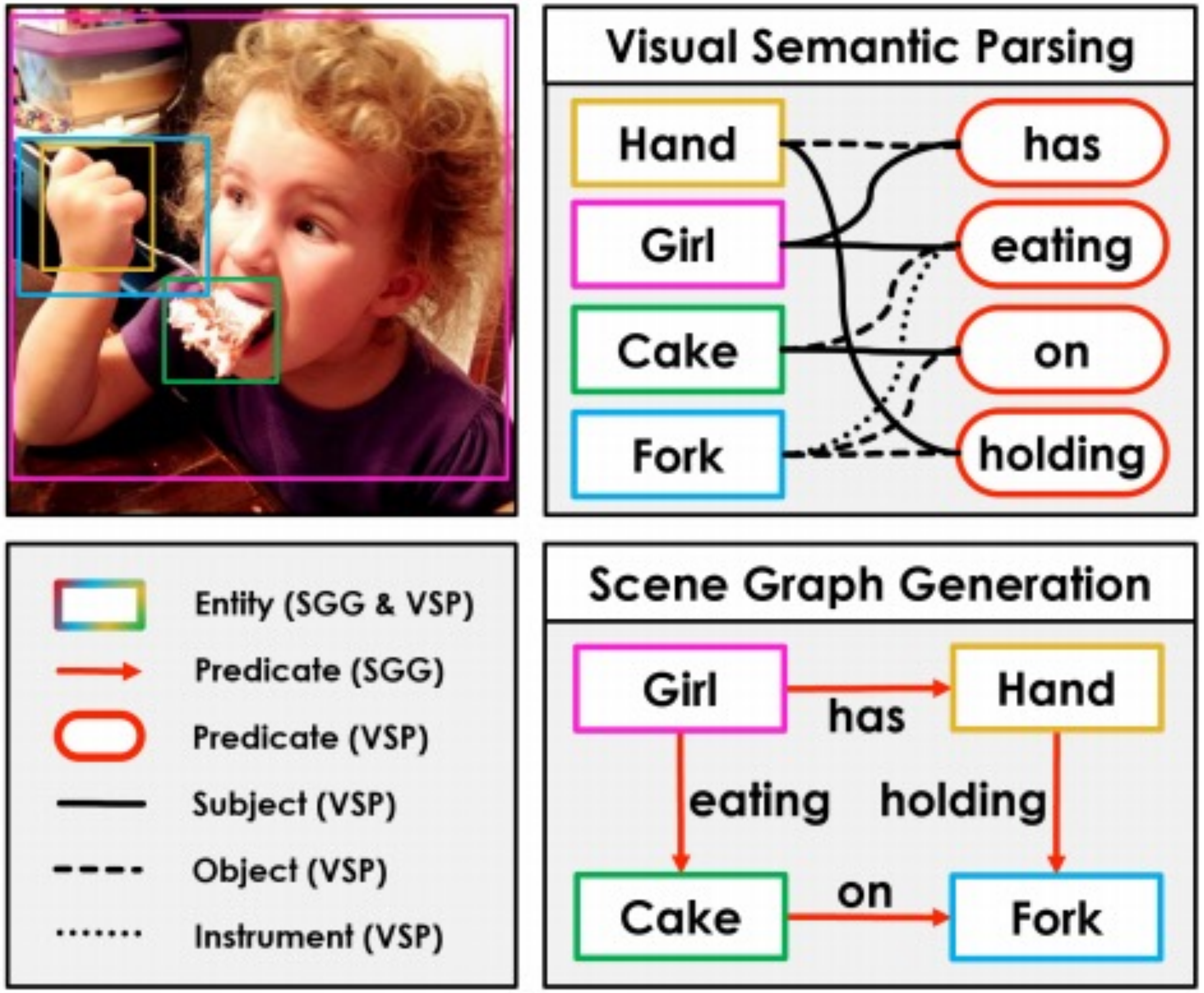


- Video Situation Recognition extends the work to videos [Sadhu et al, 2021]



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

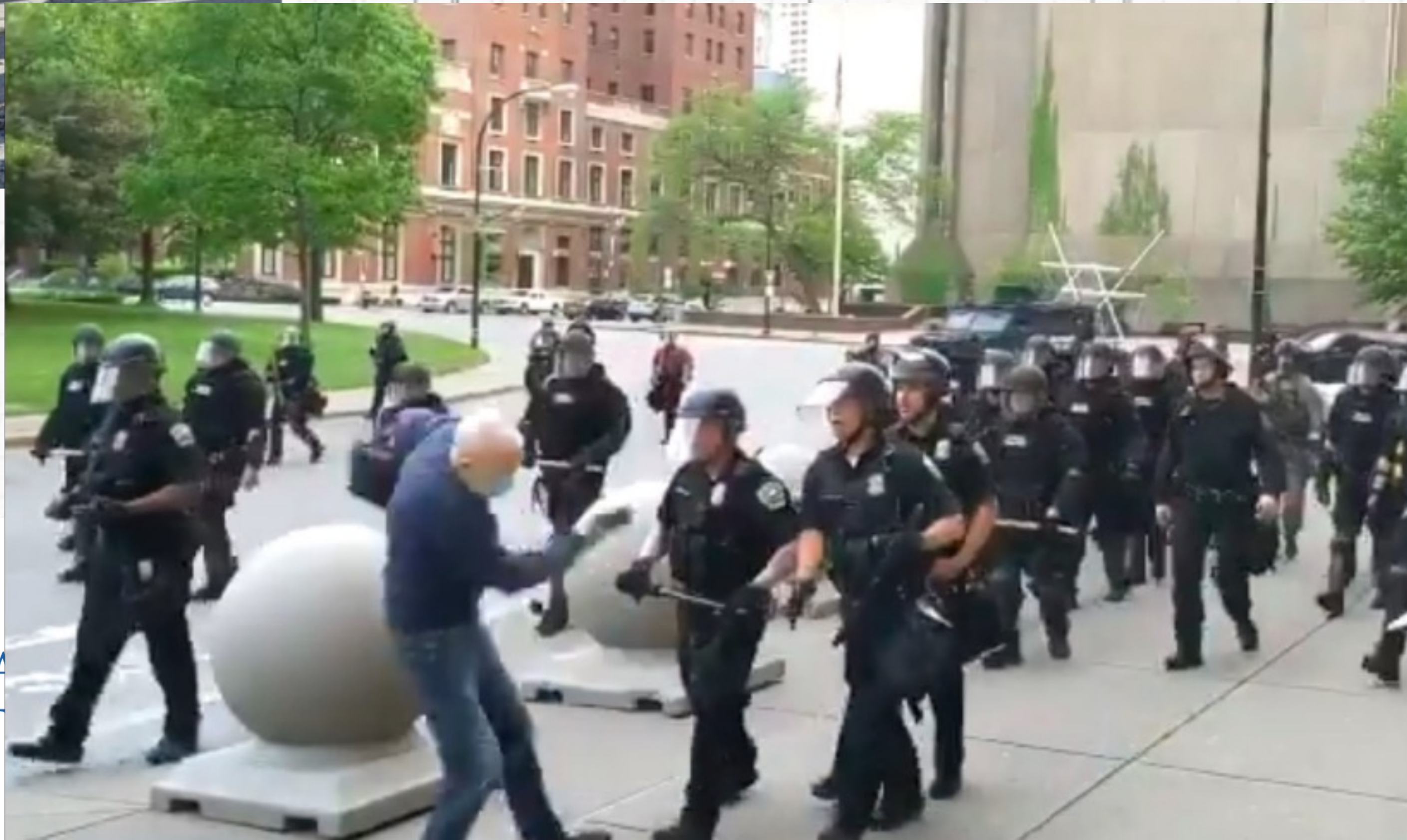


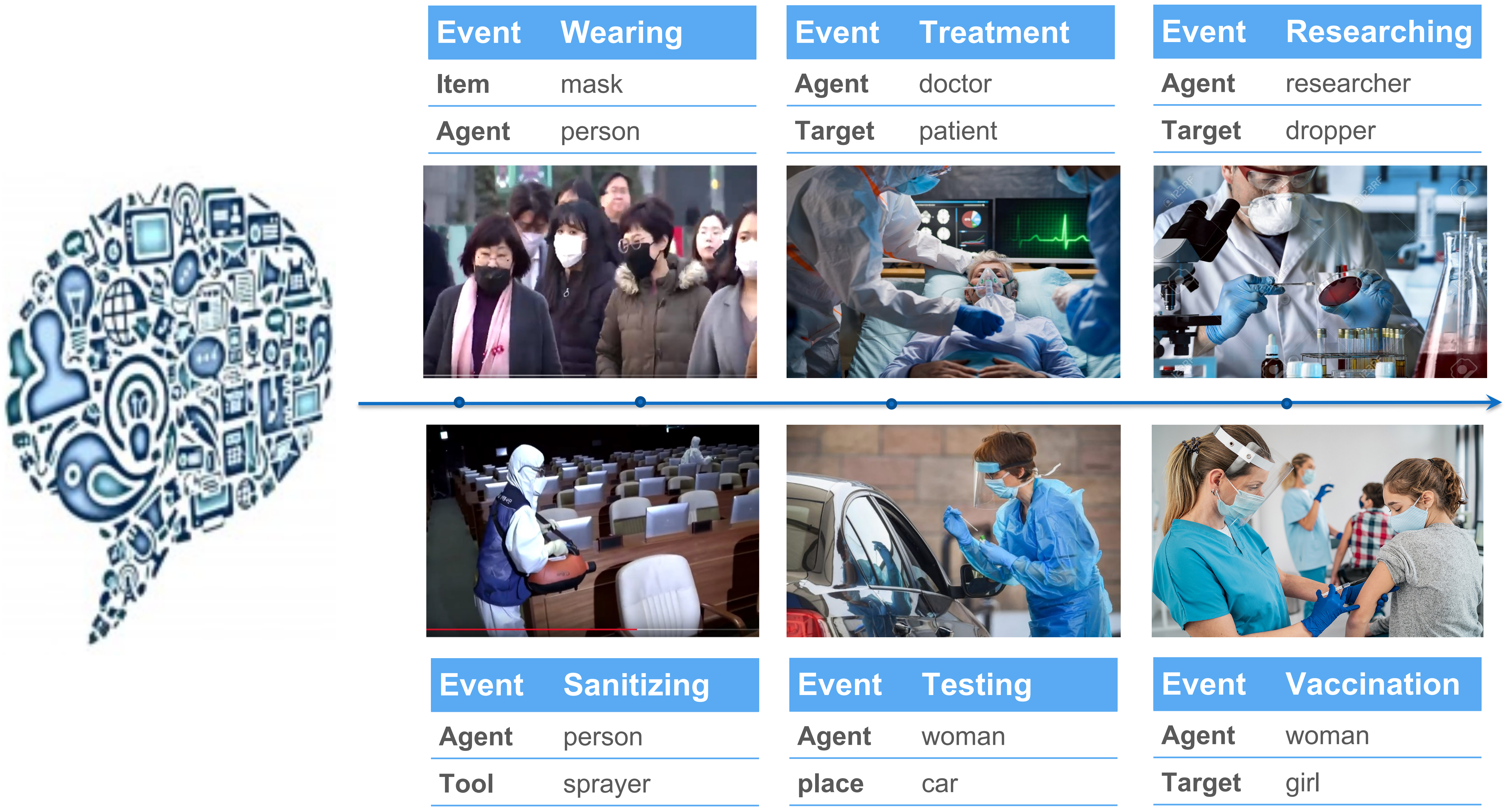

Car


Car
Event Bombing
Item Car
Witness People

Event	Attacking
Attacker	protesters
Target	police

Event	Attacking
Attacker	police
Target	protester





Event Wearing
Item mask
Agent person
Event Treatment
Agent doctor
Target patient
Event Researching
Agent researcher
Target dropper
Event Sanitizing
Agent person
Tool sprayer
Event Testing
Agent woman
place car
Event Vaccination
Agent woman
Target girl

CLIP-Event: Event-Driven Vision-Language Pretraining

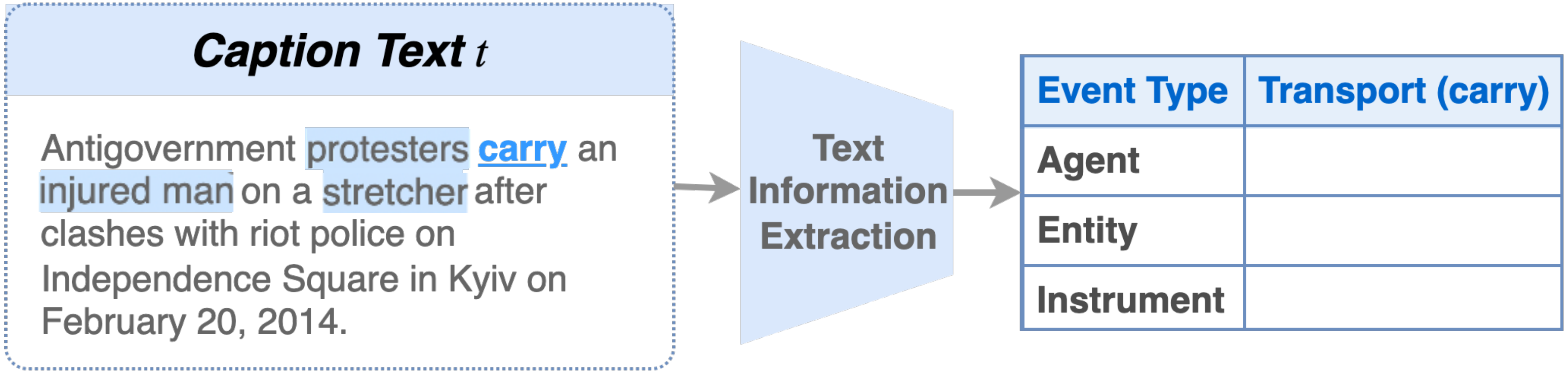
Caption Text t

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

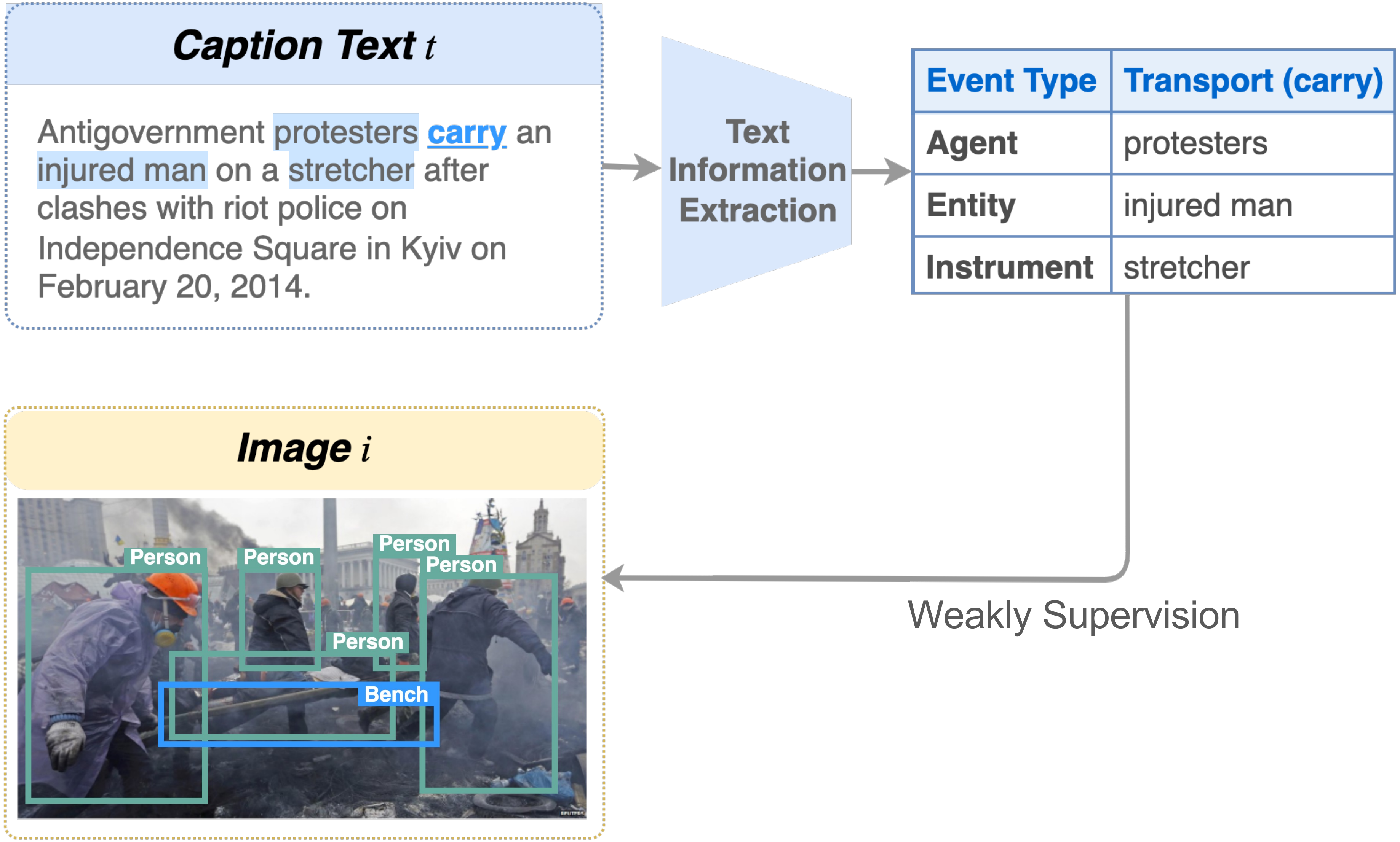
Text Information Extraction

Event Type	Transport ( )

CLIP-Event: Event-Driven Vision-Language Pretraining



- Transfer text event knowledge to images: Using text event structures as a distant supervision



- Construct **hard negatives** by manipulating event structures.

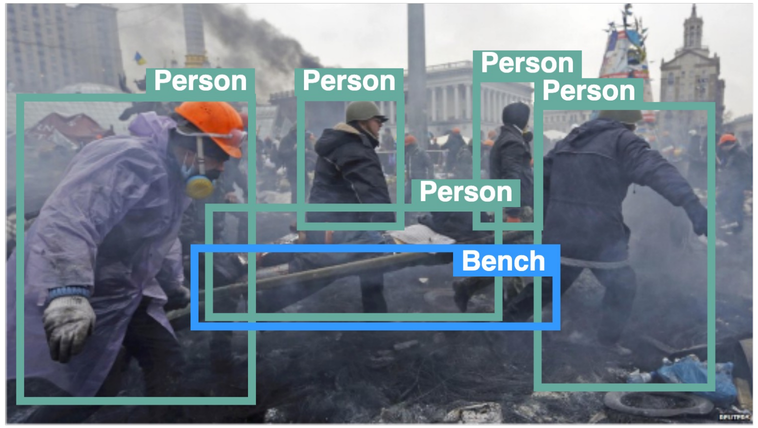
Caption Text t

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
Entity	injured man
Instrument	stretcher

Image i



Negative Labels (events)

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

Negative Labels (arguments)

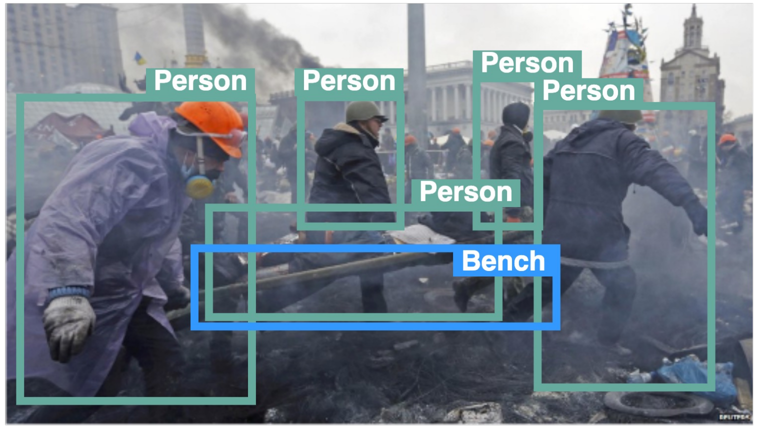
Event Type	Transport (carry)
Agent	protesters
Entity	injured man
Instrument	stretcher

- Construct **hard negatives** by manipulating event structures.

Caption Text t

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

Image i



Positive Labels

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

prompt → Protesters transported injured man using a stretcher.

Negative Labels (events)

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

prompt → Protesters arrested injured man using a stretcher.

Negative Labels (arguments)

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

prompt → Injured man transported a stretcher with protesters.

- Construct **hard negatives** by 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.

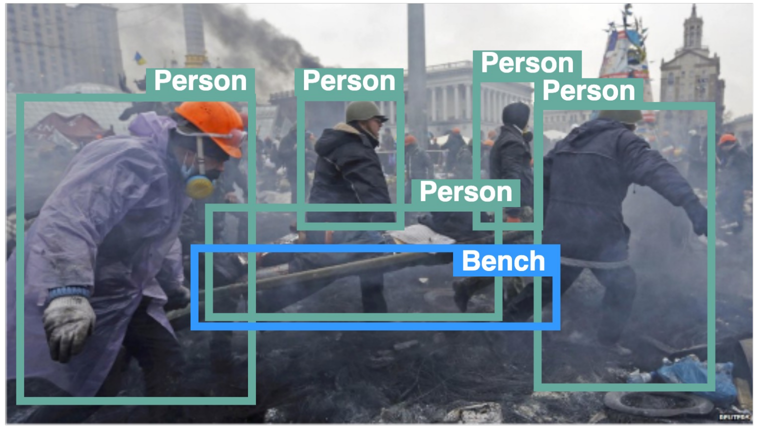
Positive Labels

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

prompt

Protesters transported injured man using a stretcher.

Image i



Negative Labels (events)

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

prompt

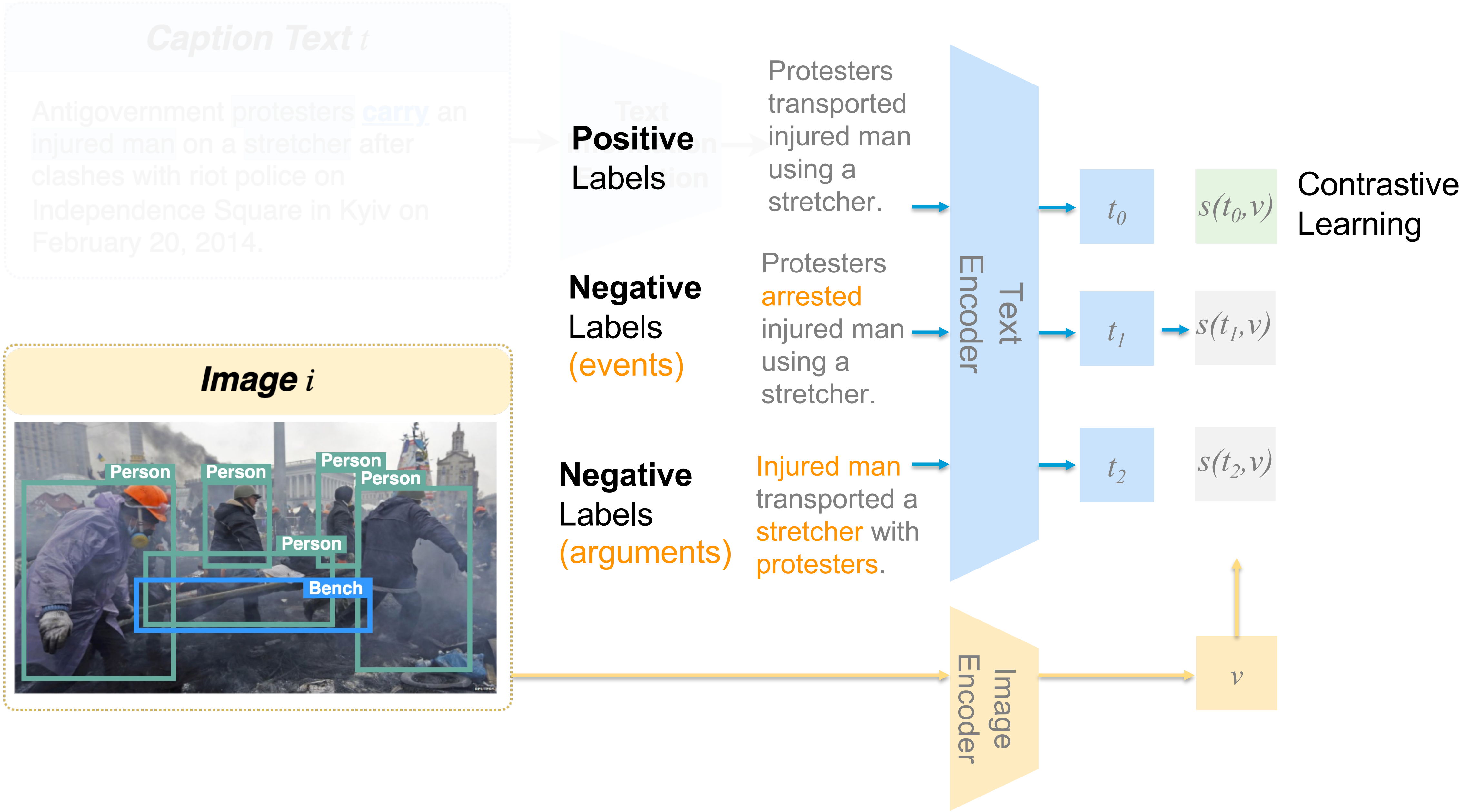
Protesters arrested injured man using a stretcher.

Negative Labels (arguments)

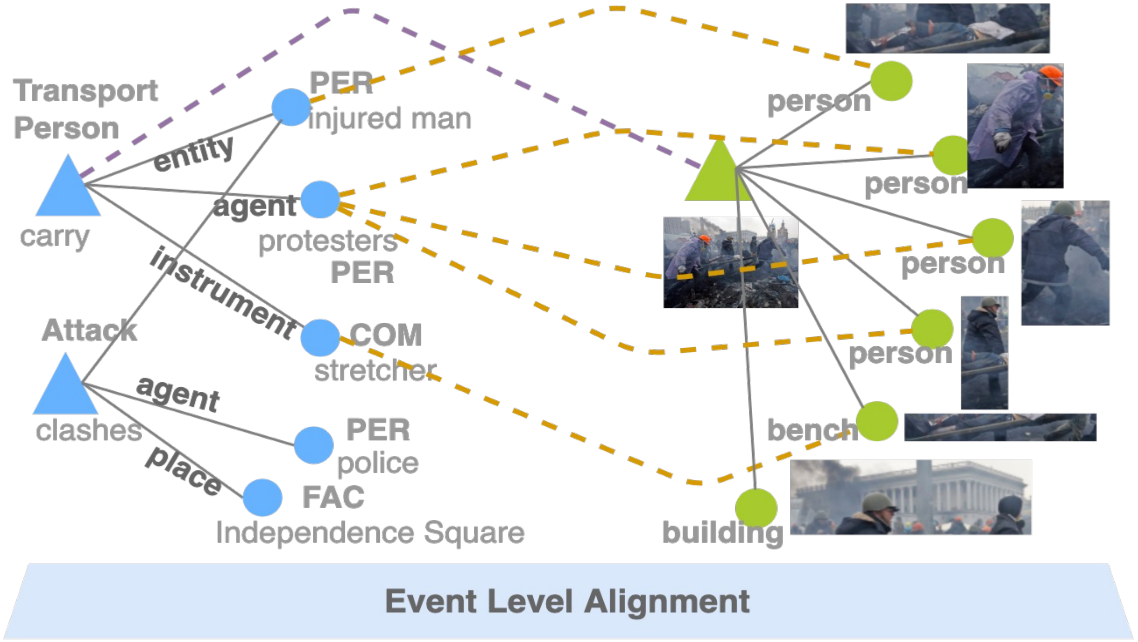
Event Type	Transport (carry)
Agent	injured man
Entity	stretcher
Instrument	protesters

prompt

Injured man transported a stretcher with protesters.


Caption Text t
Antigovernment protesters carry an injured man on a stretcher after clashes with riot police on Independence Square in Kyiv on February 20, 2014.
Image i
Person
Person
Person
Person
Person
Bench
Positive Labels
Protesters transported injured man using a stretcher.
Negative Labels (events)
Protesters arrested injured man using a stretcher.
Negative Labels (arguments)
Injured man transported a stretcher with protesters.
Text Encoder
Image Encoder
t_0
t_1
t_2
v
$s(t_0,v)$
$s(t_1,v)$
$s(t_2,v)$
Contrastive Learning

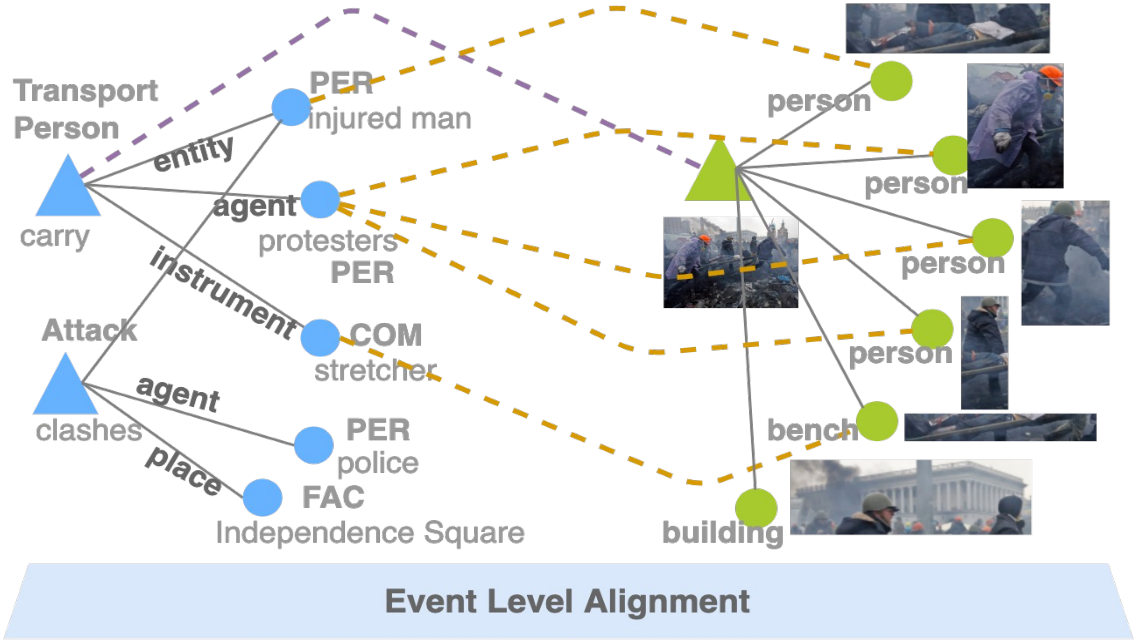
Event-Driven Vision-Language Pretraining



Event-Driven Vision-Language Pretraining

Structured Alignment via Optimal Transport

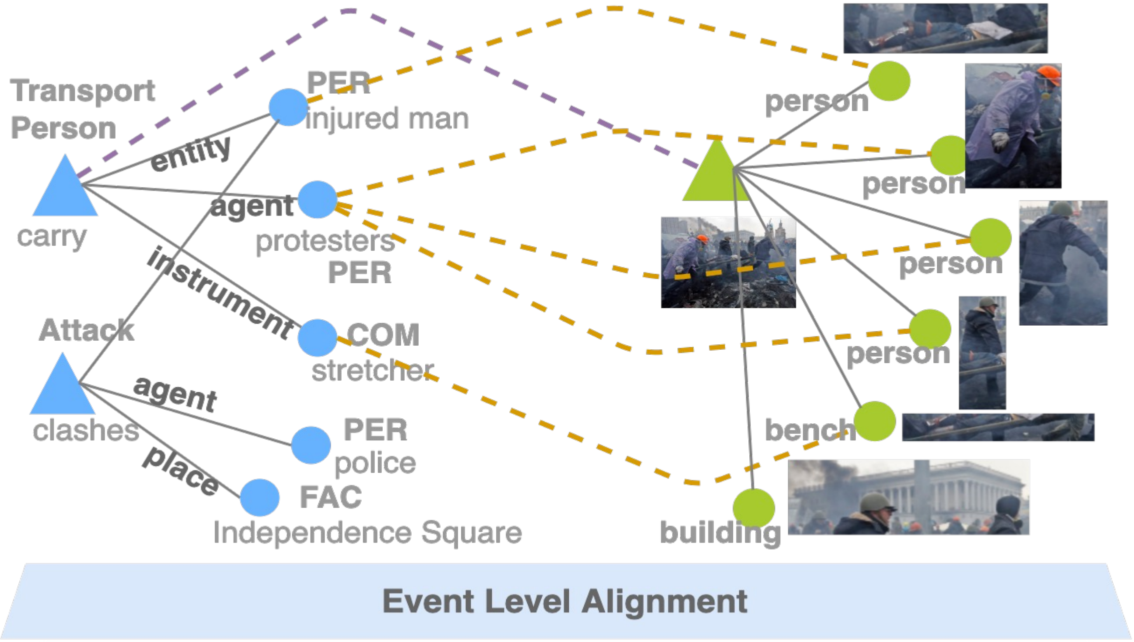
Text Event Graph ←→ Image Event Graph



Event-Driven Vision-Language Pretraining

Structured Alignment via Optimal Transport

Text Event Graph ←→ Image Event Graph



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

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

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

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

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

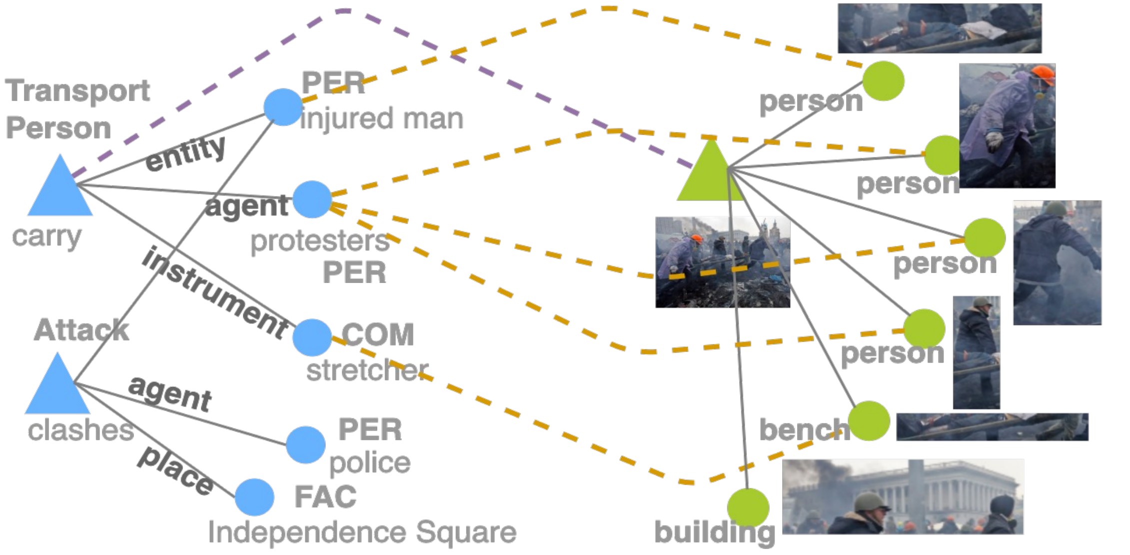
$$\boldsymbol{T}^{k} := \operatorname{diag}(\boldsymbol{p}^{k})\boldsymbol{K}\operatorname{diag}(\boldsymbol{q}^{k})$$

Event-Driven Vision-Language Pretraining

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 $\boldsymbol{T}$ within k iterations



Event Level Alignment

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

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

$$\text{for } i = 0, 1, 2, \ldots \text{ until convergence,}$$
$$\boldsymbol{p}^{i+1} = \mathbf{1} \oslash (\boldsymbol{K}\boldsymbol{q}^{i}),$$
$$\boldsymbol{q}^{i+1} = \mathbf{1} \oslash (\boldsymbol{K}^{\top}\boldsymbol{p}^{i+1}),$$
$$\boldsymbol{T}^{k} := \operatorname{diag}(\boldsymbol{p}^{k}) \boldsymbol{K} \operatorname{diag}(\boldsymbol{q}^{k})$$

Event-rich Image-Caption Dataset

- We collect 106,875 image-captions that are rich in events from VOA news website.

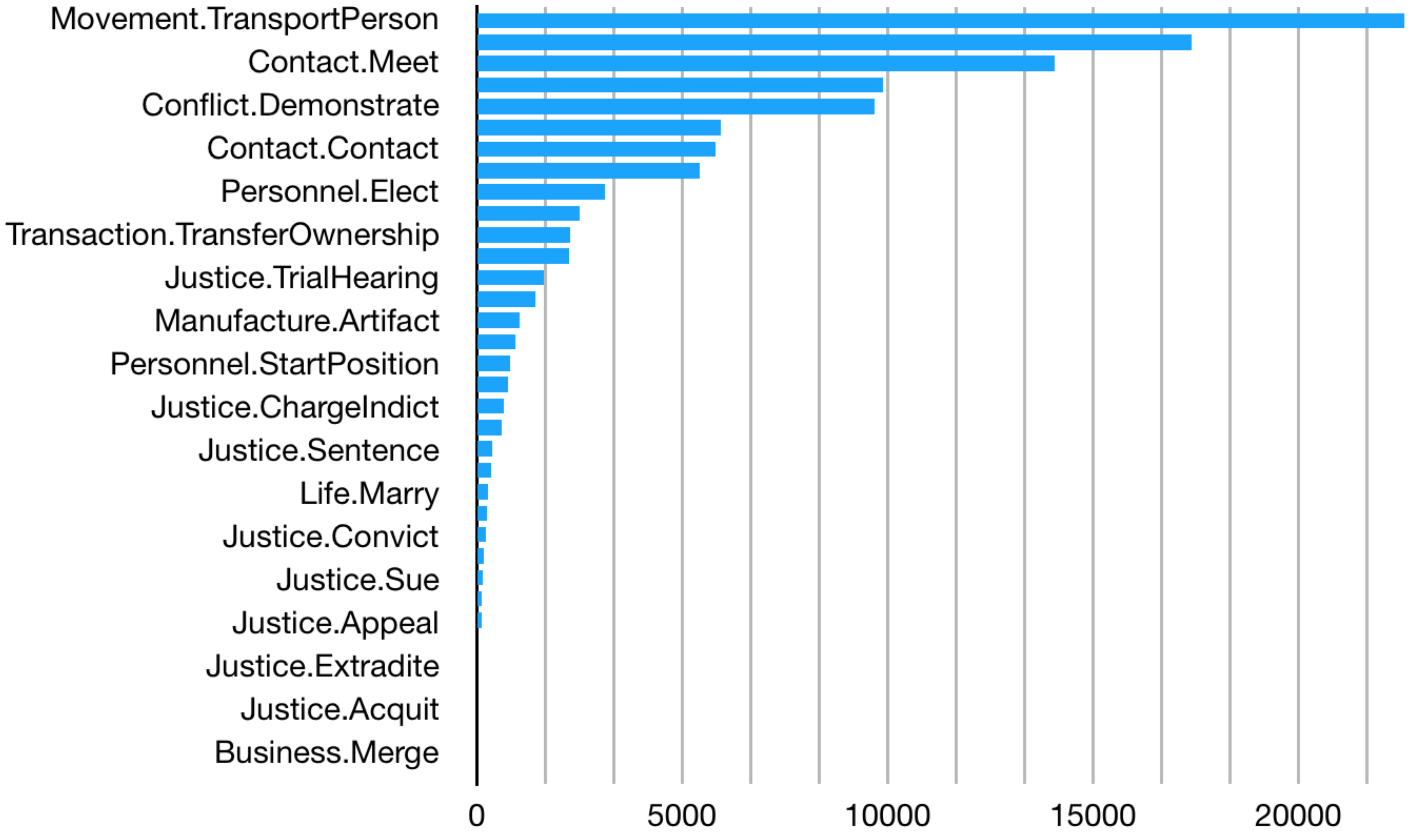
Split	# image	# event	# arg	# ent
Train	76,256	84,120	148,262	573,016
Test	18,310	21,211	39,375	87,671
No-event	12,309	-	-	-

- It is a challenging image-retrieval benchmark, aiming to understand long sentenc

	Flickr30k	MS COCO	VOANews
Average sentence length	13.4	11.3	28.2

Text Event Extraction Results

- State-of-the-art IE (149 event types, Lin et al, 2020)
- 108,693 captions
- 84,120 events
- 0.8 events in average (we filter the captions without events during training)



CLIP-Event on Visual Event Extraction

Supporting Zero-shot Vision Event Extraction the first time.

Event Type	Arrest
Agent	person
Detainee	person

Event Type	protesting
Agent	people
Place	outdoors

CLIP-Event on General Vision-Language Tasks

Injecting event knowledge benefits various generic tasks.

Question: Why is Person1 attacking Person2?

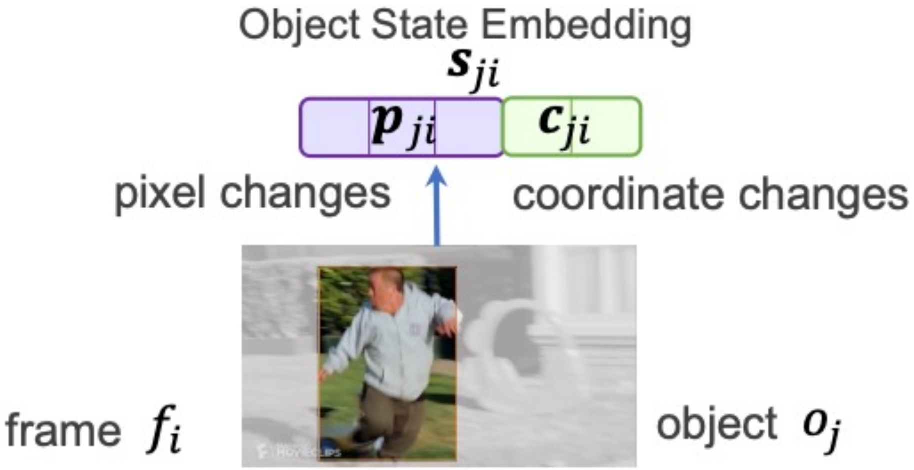
Answer:
(1) Person1 is trying to defeat Person2 so that he can help Person1 escape .
(2) Person2 does not want to be having the conversation , and Person1 has cornered him into it.
(3) Because he is angry at him.
(4) Person1 is a bully and is beating him up .

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)



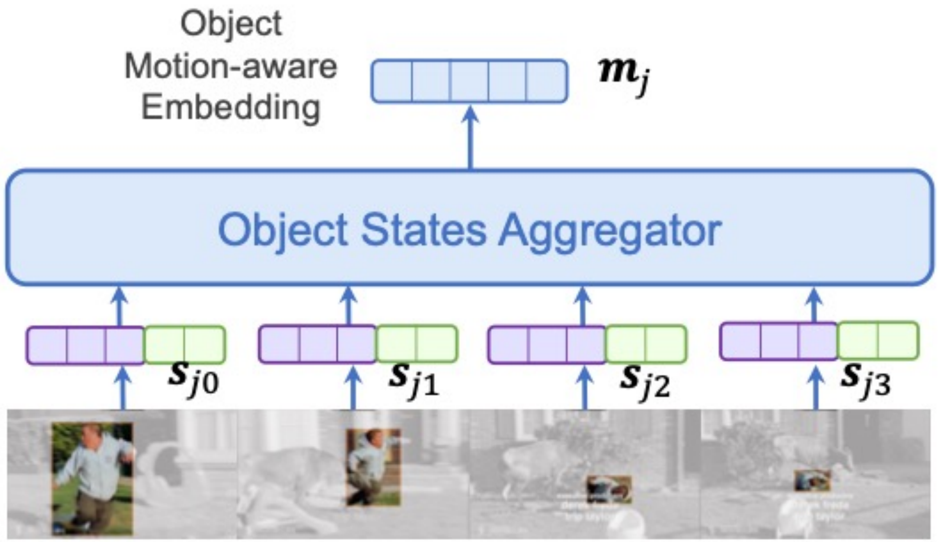
Object State Embedding of the man

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)

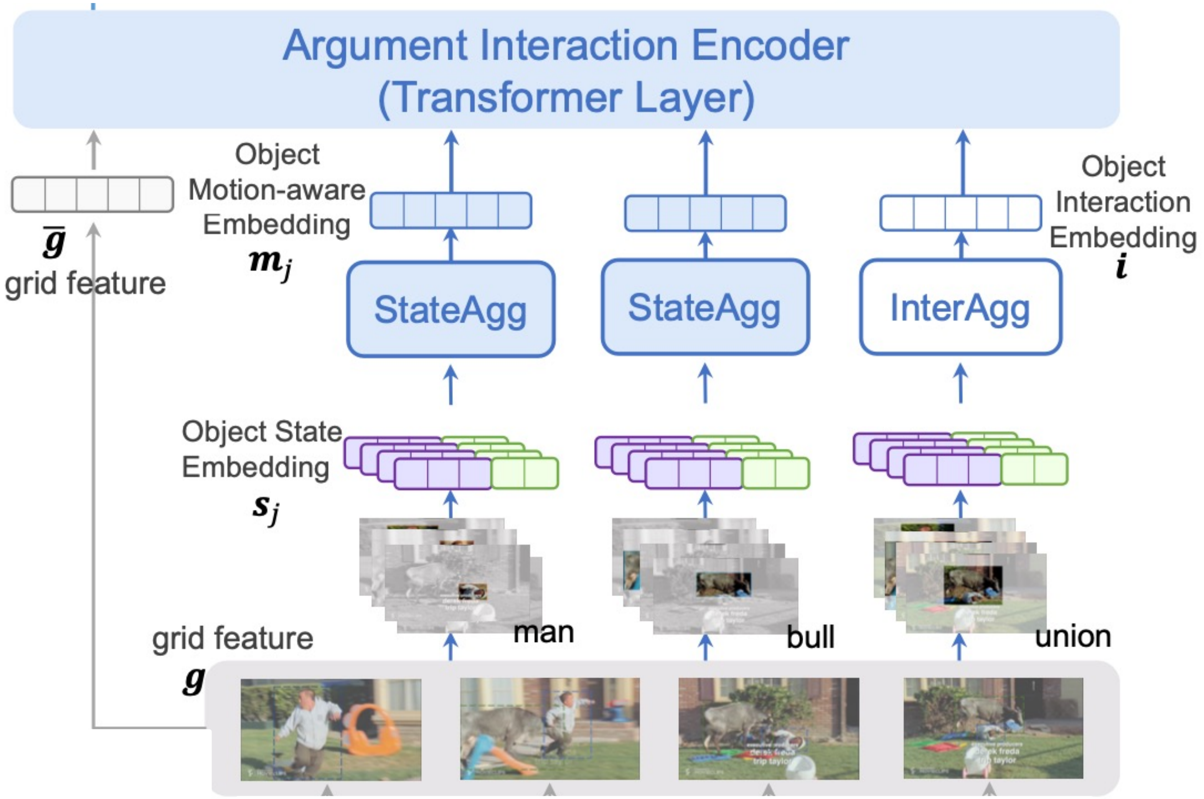


Video Events as Argument State Changes

Video Event =

Status Changes of Arguments

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

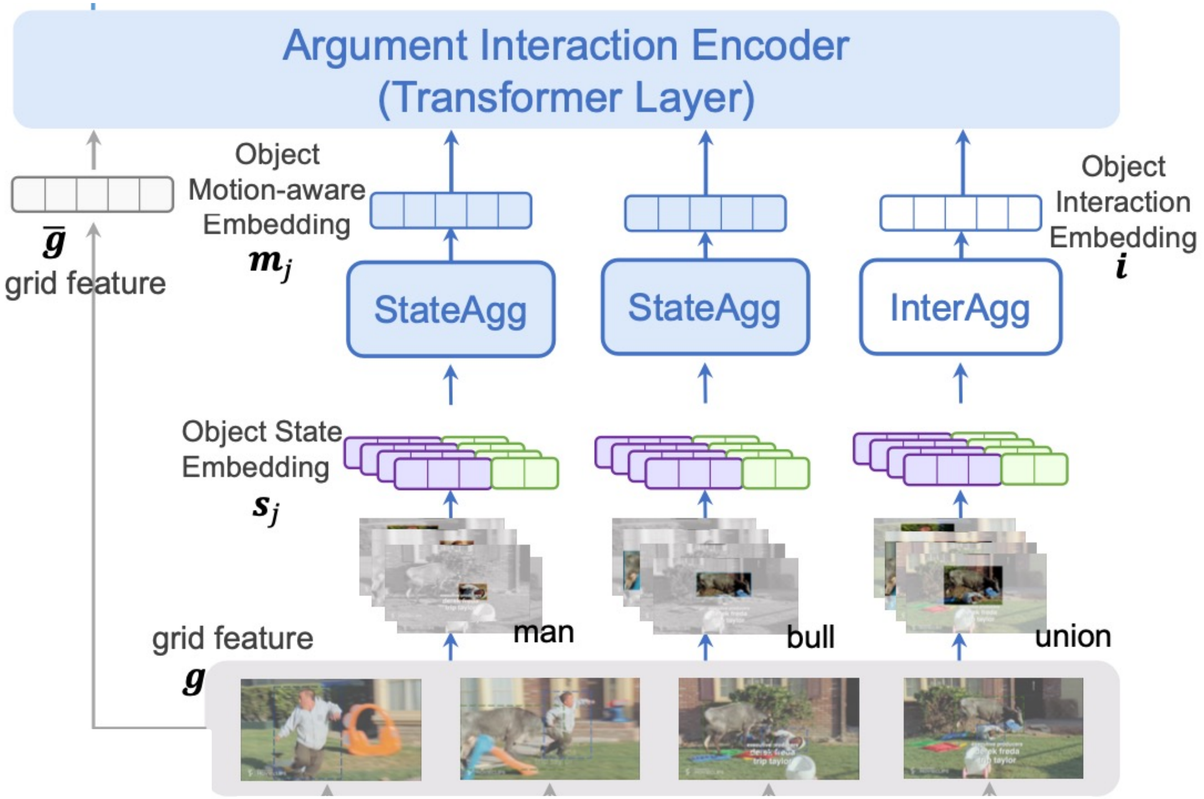


Video Events as Argument State Changes

Video Event =

Status Changes of Arguments

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



Results – Verb & Semantic Role Prediction

Model	Kinetics	Val Acc@1	Val Acc@5	Val Rec@5	Val F_1@5	Test Acc@1	Test Acc@5	Test Rec@5	Test F_1@5
TimeSformer	✓	45.91	79.97	23.61	18.23	-	-	-	-
I3D†	✗	30.17	66.83	4.88	4.56	31.43	67.70	5.02	4.67
SlowFast†	✗	32.64	69.22	6.11	5.61	33.94	70.54	6.56	6.00
I3D†	✓	29.65	60.77	18.21	14.01	29.87	59.10	19.54	14.68
SlowFast†	✓	46.79	75.90	23.38	17.87	46.37	75.28	25.78	19.20
Ours (OSE-pixel + OME)	✓	52.75	83.88	28.44	21.24	52.14	**83.84**	30.66	22.45
Ours (OSE-pixel/disp + OME)	✓	53.32	**84.00**	28.61	21.34	51.88	83.55	**30.83**	**22.52**
Ours (OSE-pixel/disp + OME + OIE)	✓	**53.36**	83.94	**28.72**	**21.40**	**52.39**	83.47	30.74	22.47

Results on Verb Classification

Model	CIDEr Avg	CIDEr Std	CIDEr-Verb Avg	CIDEr-Verb Std	CIDEr-Arg Avg	CIDEr-Arg Std	ROUGE-L Avg	ROUGE-L Std
GPT2†	34.67		42.97		34.45		40.08	
I3D†	47.06		51.67		42.76		42.41	
SlowFast†	45.52		55.47		42.82		42.66	
SlowFast	44.49	±2.30	51.73	±2.70	40.93	±2.42	40.83	±1.27
Ours (OSE-pixel + OME)	47.82	±2.12	54.51	±3.00	44.32	±2.45	40.91	±1.32
Ours (OSE-pixel/disp + OME)	**48.46**	**±1.84**	**56.04**	**±2.12**	**44.60**	**±2.33**	**41.89**	**±1.12**
Ours (OSE-pixel/disp + OME + OIE)	47.16	±1.71	53.96	±1.32	42.78	±2.74	40.86	±2.54

Results on Semantic Role Prediction

Understanding videos via Objects, Events, Attributes

➤ **Unique challenges for video-language tasks**

Multiple levels of semantics: a video may contain visual features with different granularity

Solution: Hierarchical textual representation of videos by leveraging *image-language foundation models* and *semantic role labeling guidance*

The temporal dimension: objects and events in videos are dynamically related

Solution: Temporal-aware few-shot prompt

How to make GPT-3 understand videos?

Input Video

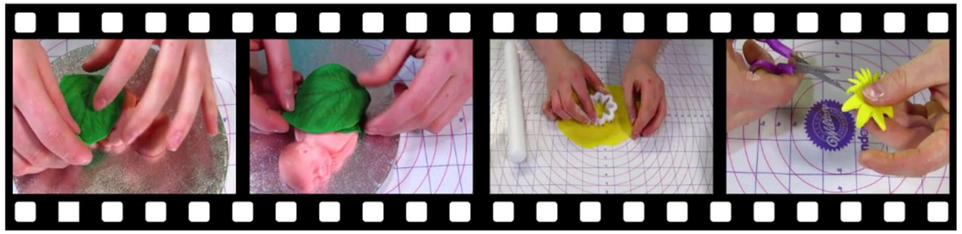


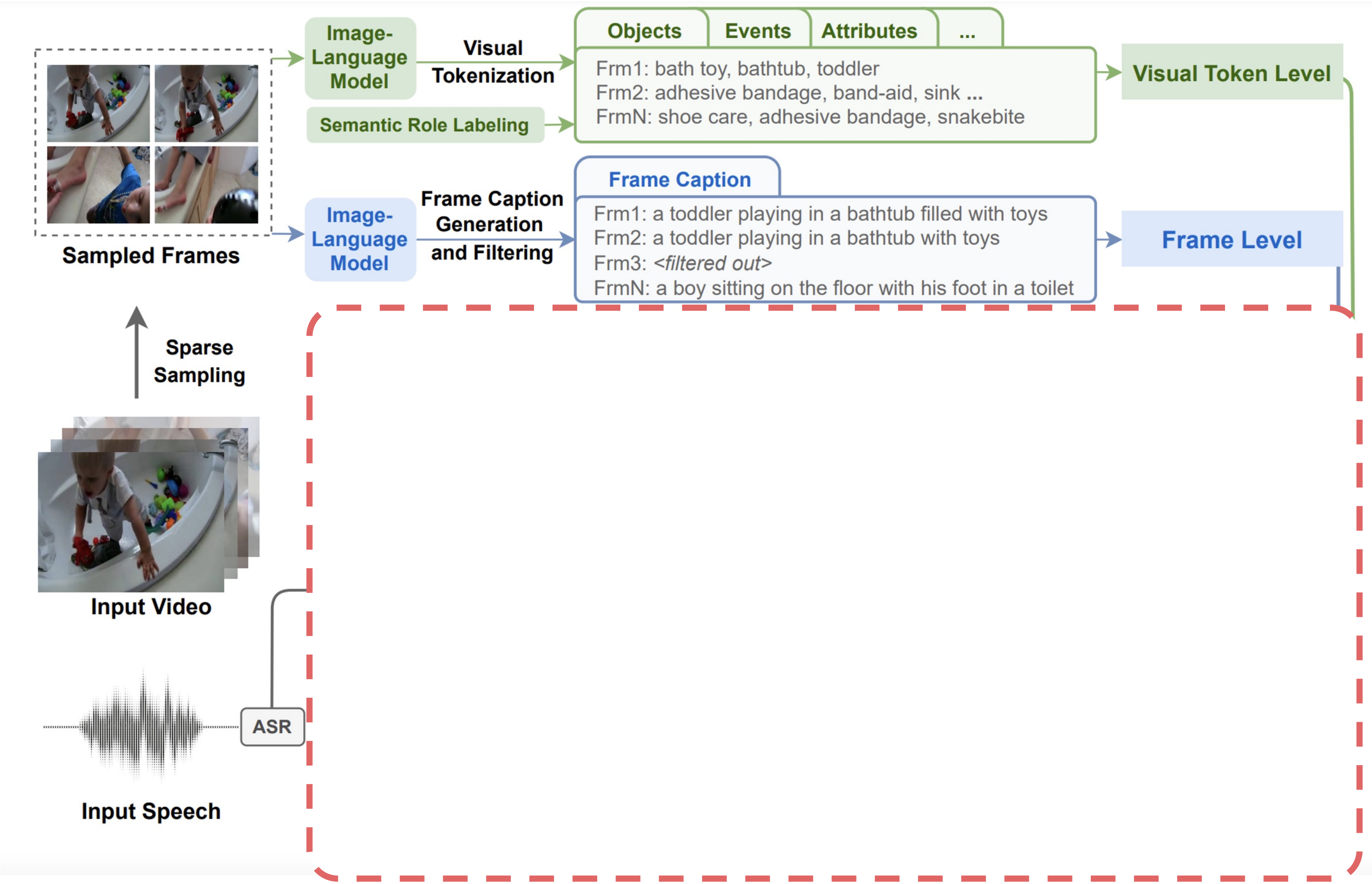
Image-Language Model

Visual Token Level	**Objects**	cake decorating, sugar paste, clay animation, play-doh
	Events	cutting mat, woman shaped cake, cake is made, flowered design
	Attributes	made of fondant, edging, rubbing, paper doilies, green goo
Frame Level	**Frame Captions**	a person holding a green object in their hand; a person is putting a green leaf on a baby's head; a person cutting a piece of paper with a pair of scissors

Language Model

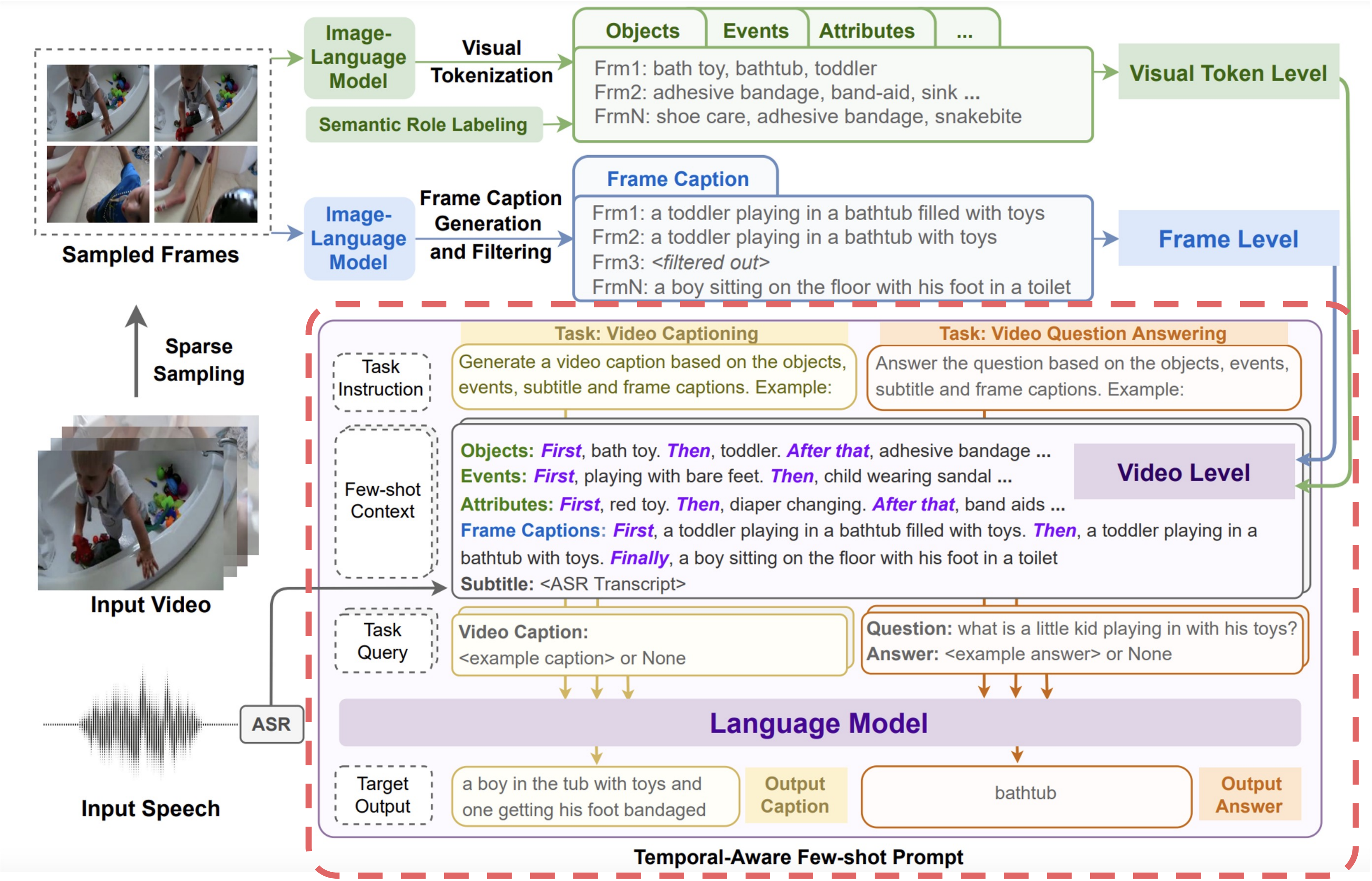
Video Level	**Next Event Prediction**	**Question: What will happen next?** **Answer:** the person puts the flower on top of the baby-shaped cake

Understanding videos via Objects, Events, Attributes



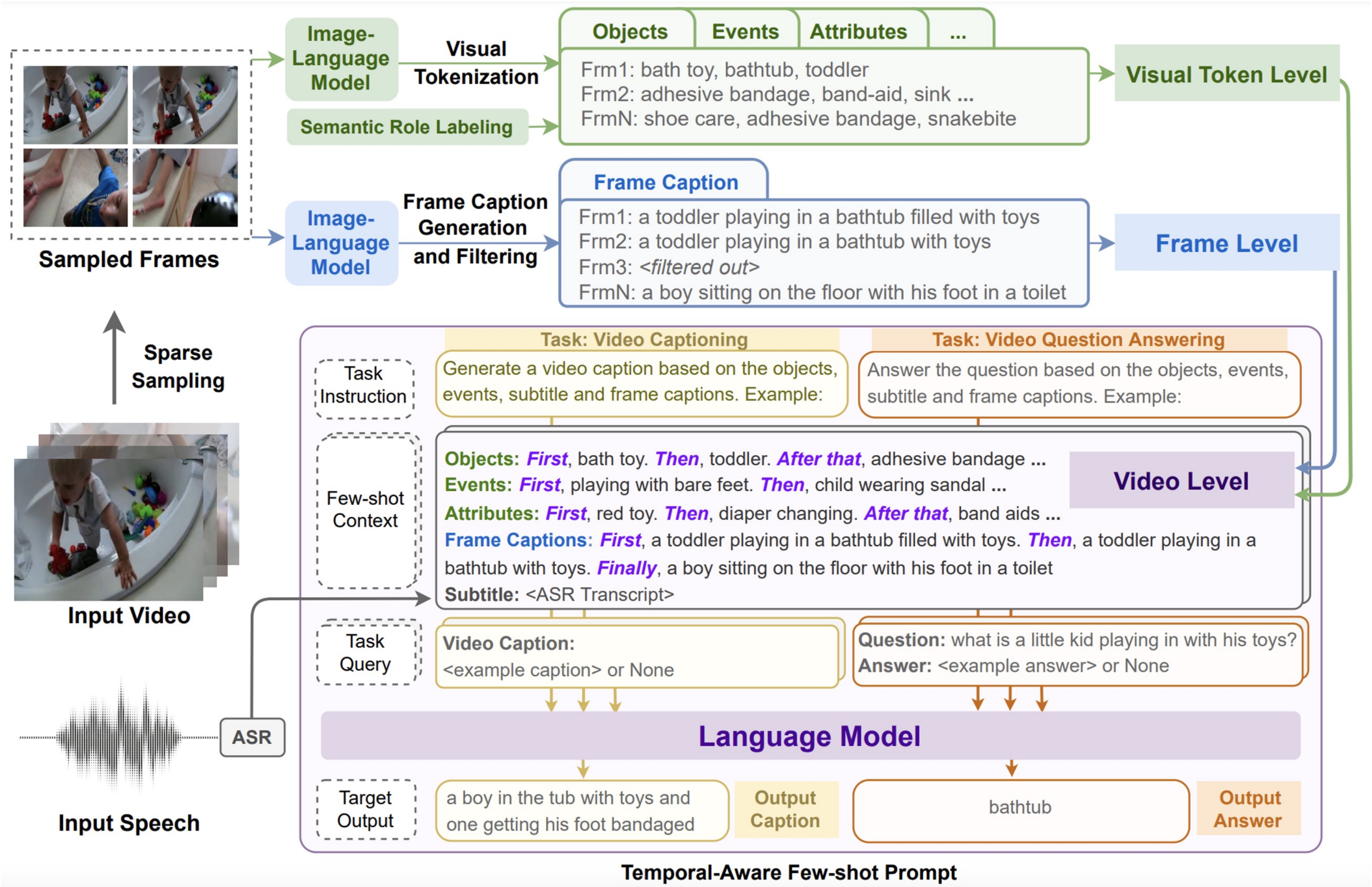
VidIL (Ours) Framework

Understanding videos via Objects, Events, Attributes



VidIL (Ours) Framework

Understanding videos via Objects, Events, Attributes



Does **NOT** require **ANY** video data for pretraining

Flexibility in adding **additional modalities**, e.g., ASR

Understanding videos via Objects, Events, Attributes

Method	ASR	MSR-VTT Caption B-4	R-L	M	C	YouCook2 Caption B-4	R-L	M	C	VaTex Caption B-4	R-L	M	C	Avg C
Few-shot														
UniVL	No	2.1	22.5	9.5	3.6	**3.3**	**25.3**	**11.6**	**34.1**	1.7	15.7	8.0	2.1	13.3
BLIP	No	**27.7**	43.0	23.0	**39.5**	0.7	9.0	3.4	11.5	13.5	39.5	15.4	20.7	23.9
BLIP$_{cap}$	No	21.6	48.0	22.7	30.2	3.7	8.6	3.8	9.4	20.7	41.5	17.4	28.9	22.8
VidIL(ours)	No	26.0	**51.7**	**24.7**	36.3	2.6	22.9	9.5	27.0	**22.2**	**43.6**	**20.0**	**36.7**	**33.3**
UniVL	Yes	-	-	-	-	4.3	26.4	12.2	48.6	2.7	17.7	10.2	3.4	26.0
VidIL(ours)	Yes	-	-	-	-	**10.7**	**35.9**	**19.4**	**111.6**	**23.2**	**44.2**	**20.6**	**38.9**	**75.3**
Fine-tuning														
UniVL	No	42.0	61.0	29.0	50.1	11.2	40.1	17.6	127.0	22.8	38.6	22.3	33.4	70.2
UniVL	Yes	-	-	-	-	16.6	45.7	21.6	176.8	23.7	39.3	22.7	35.6	106.2

Video Captioning

Method	#video$_{PT}$	#video$_{FT}$	MSR-VTT	MSVD
BLIP	0	0-shot	0.55	0.45
BLIP	0	5-shot	0.84	0.53
BLIP$_{VQA}$ [26]	0	0-shot	19.2	35.2
VidIL(ours)	0	5-shot	**21.2**	**39.1**
♠Flamingo-3B [2]	27M	4-shot	14.9	33.0
♠Flamingo-3B [2]	27M	8-shot	19.6	37.0
♠Flamingo-80B [2]	27M	4-shot	23.9	41.7
♠Flamingo-80B [2]	27M	8-shot	27.6	45.5
ALPRO [25]	2M	full-shot	42.1	45.9

Video Question Answering

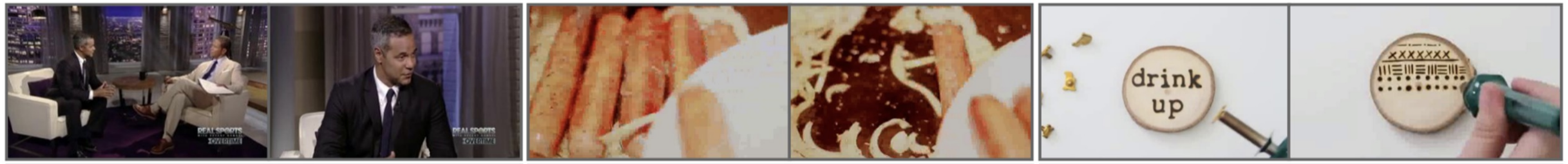
Method	#video$_{FT}$	Acc
VLEP [23]	20142	67.5
MERLOT [67]	20142	68.4
VidIL(ours)	10-shot	**72.0**
Human	-	90.5

supervised

Video-Language Future Event Prediction (VLEP)

Understanding videos via Objects, Events, Attributes

Video Captioning



MSR-VTT Caption	YouCook2 Caption	VaTex Caption
Objects: First, interview. Then, cable television. After that, television program. Finally, sports commentator. **Events: ... Attributes: ... Frame Captions: ...**	**Objects:... Events:... Attributes:... Captions:... Subtitle:** Now our sausages are pretty much cooks going to take those out all the time. And we're going to now, my cat gravy as source.	**Objects: ... Events: ... Attributes:** First, tagging. Then, woodburning. After that, wood burning. Finally, turning on dial. **Frame Captions:** First, a piece of wood with words drink up written on it ...
UniVL: a man is playing a man with a man . **BLIP:** a man in a suit and tie sitting on a couch **Ours: an interview with a sports commentator**	**UniVL:** add the sausages to the pan **Ours: take the sausages out of the pan and add some gravy to the plate**	**UniVL:** you ' re ready to decorate your cake **BLIP:** a person holding a string with a small object in front of them **Ours: A person is making a sign that says "Drink Up" with a wood burning kit.**
Ground Truths: • 2 men are discussing sports on a talk show • a man being interviewed on a tv show	**Ground Truth:** • remove sausages from pan	**Ground Truth:** Someone uses a wood burning tool to burn a design into a slice of wood and then begins to brush polyurethane unto it.

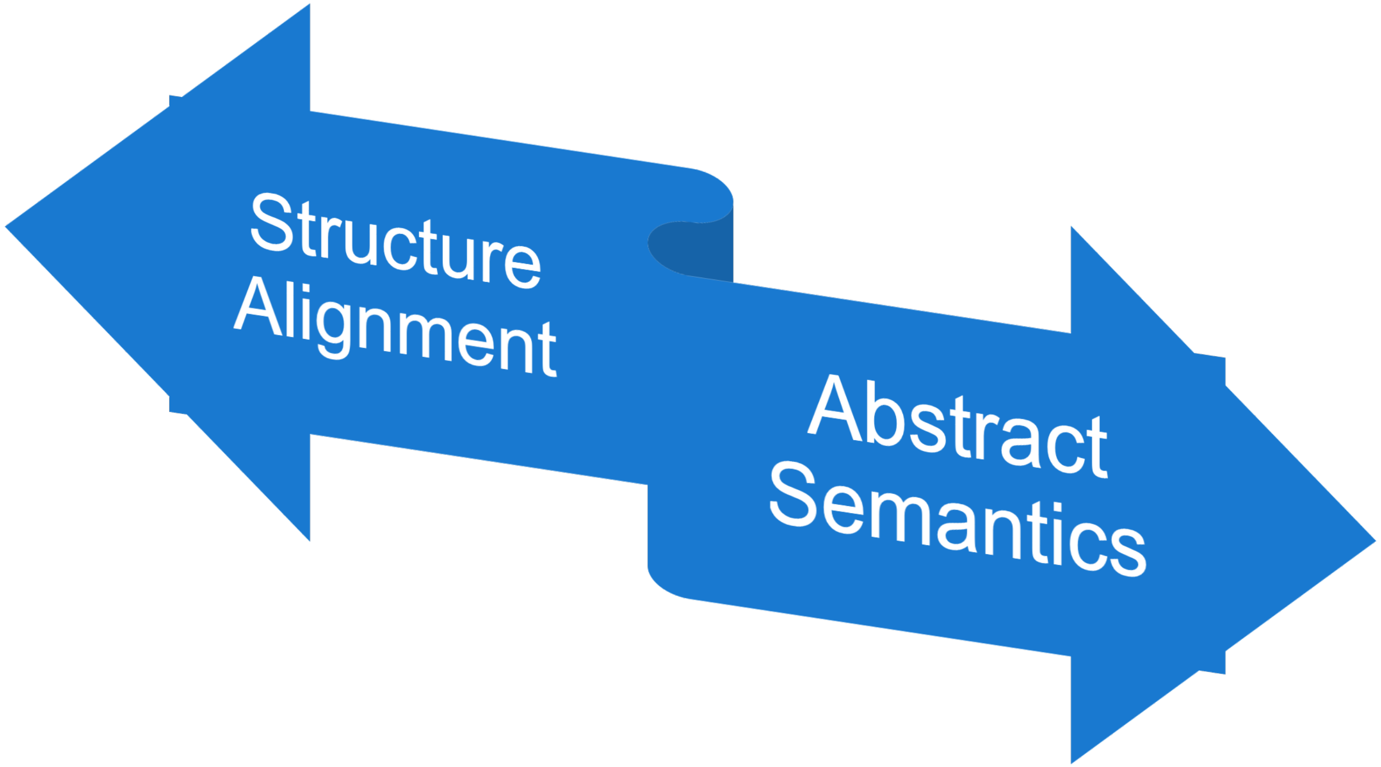
Video-language Future Event Prediction

Frame Captions: First, a woman holding a plate in a kitchen. Then, a man sitting at a table with two mugs. Finally, a woman holding a pizza in a kitchen.
Dialogue: Bernadette : I don't think you are. Raj : You didn't think I was gonna be in your kitchen this morning, Raj : yet here I am.
Question: What is more likely to happen next? **A:**Bernadette will drop the dishes and break them. **B:**Bernadette will put the dishes in the sink
Answer:

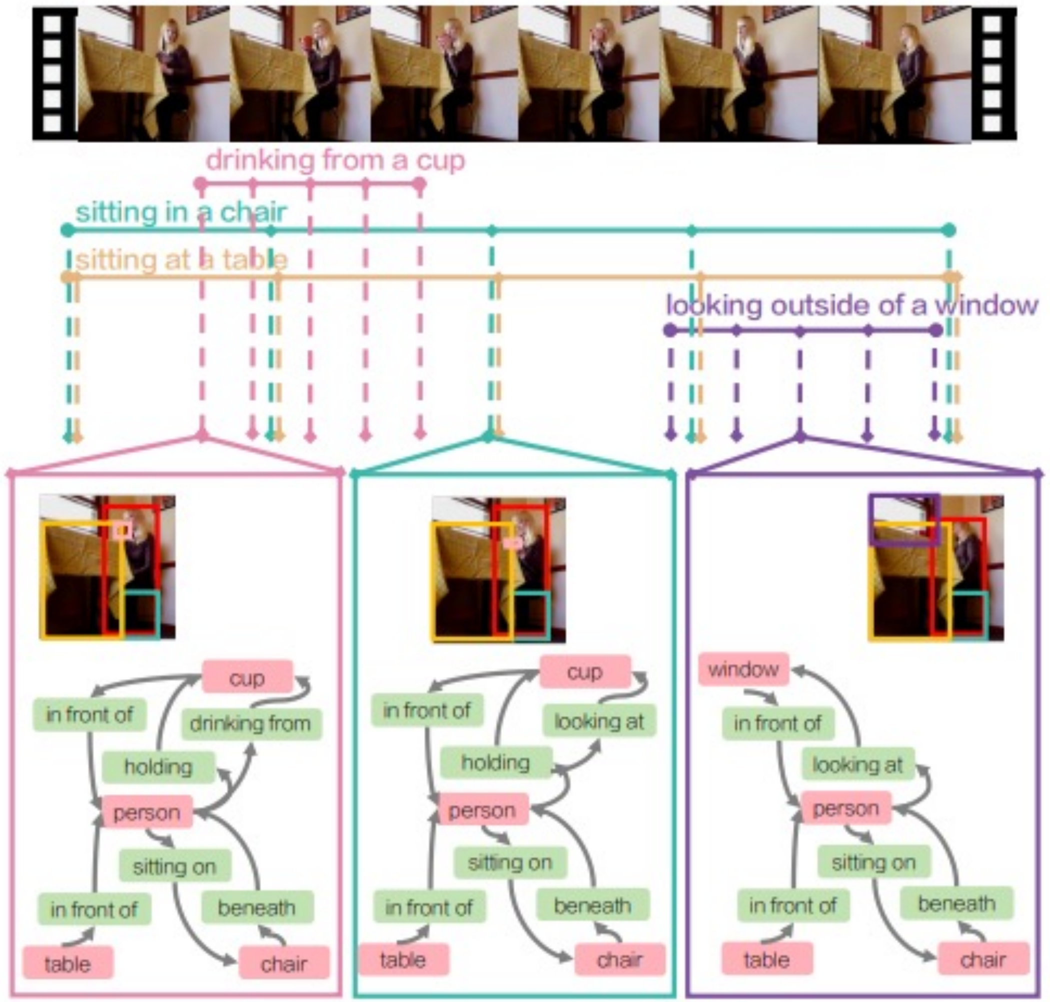
VidIL Prediction: Bernadette will put the dishes in the sink

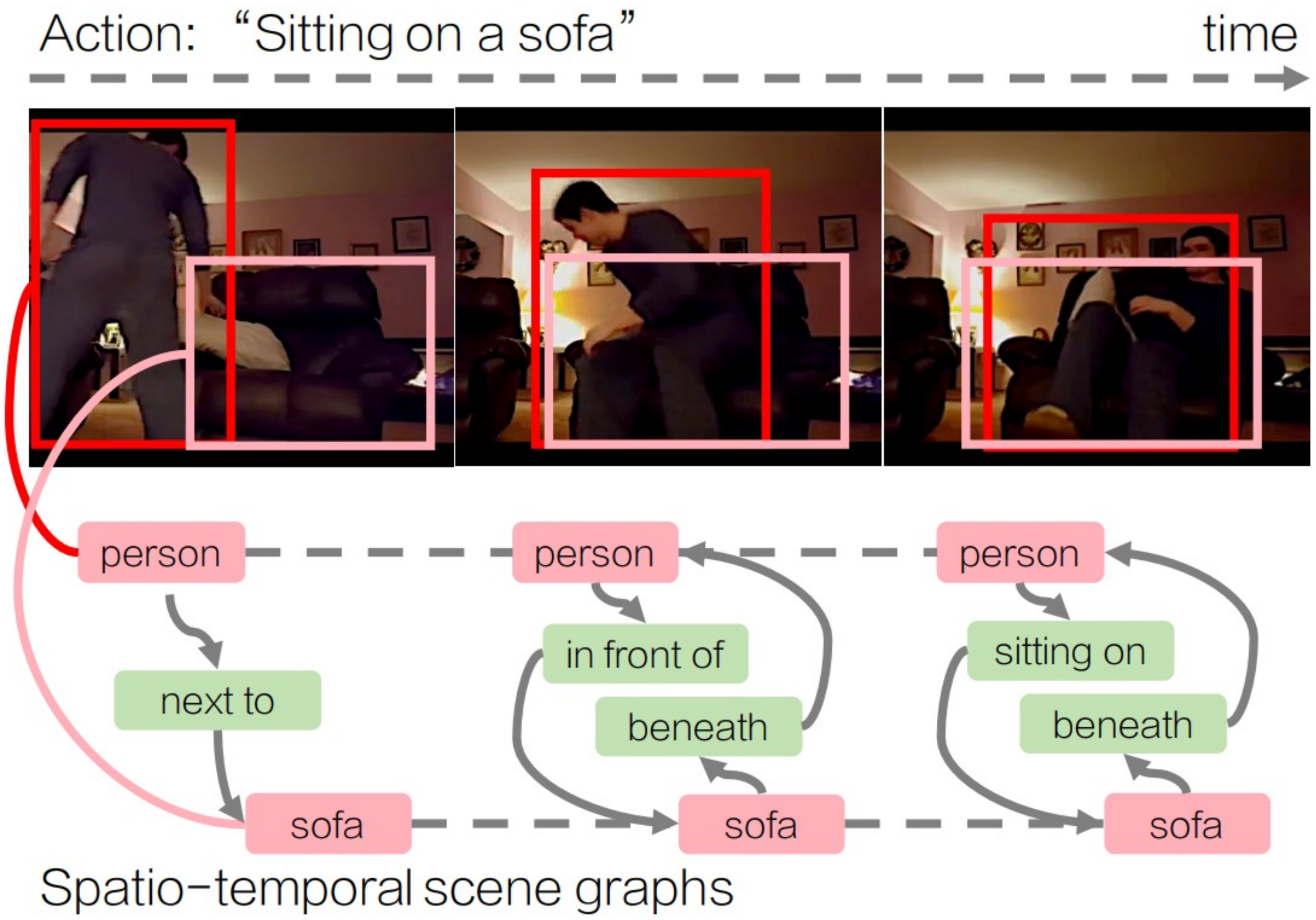
Future Challenges

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



Future Direction 1: Structured Encoding





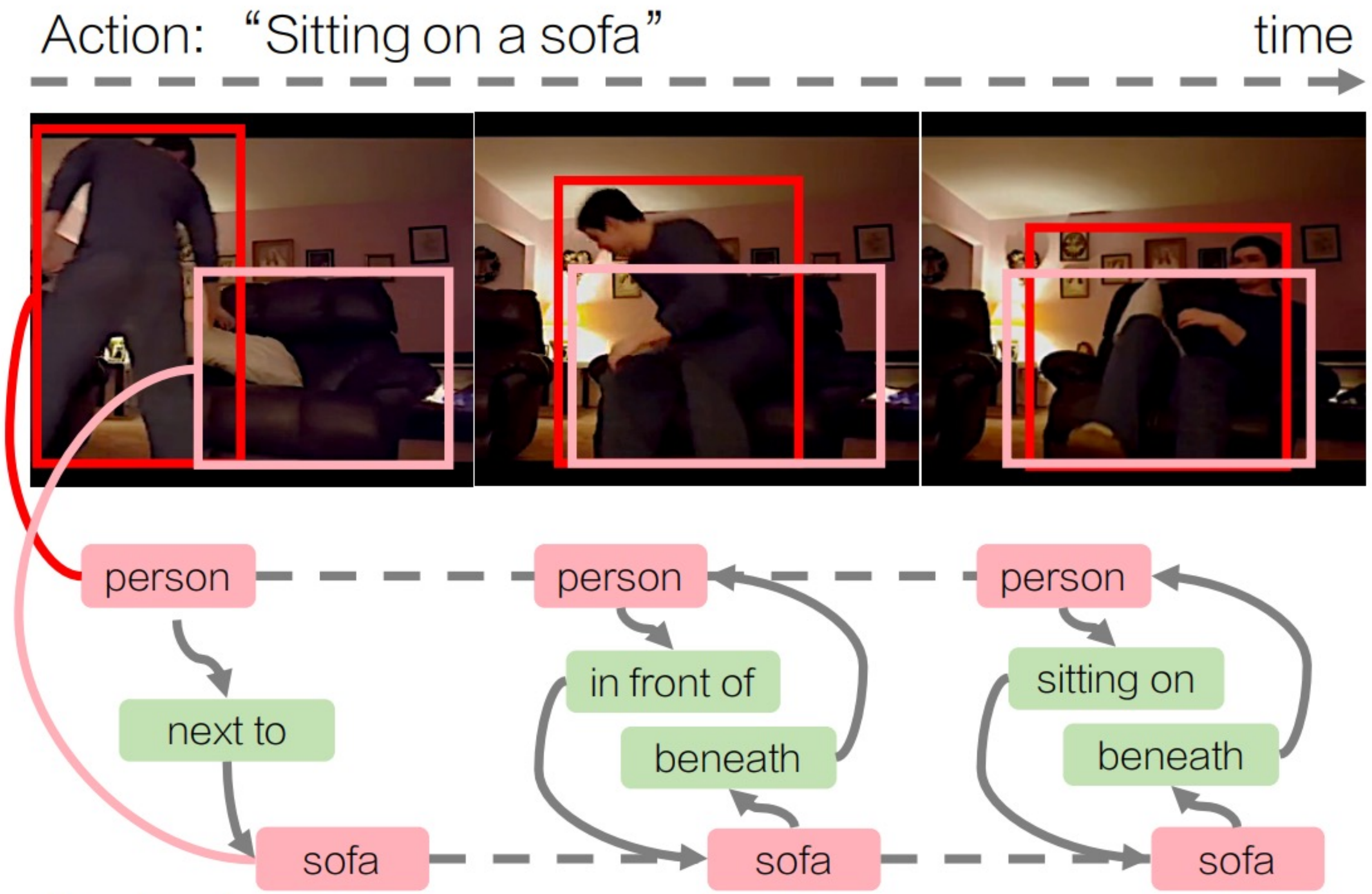
[Ji el al, 2019]

Future Direction 1: Structured Encoding

Text

Vision

- Better in capturing details with visual features

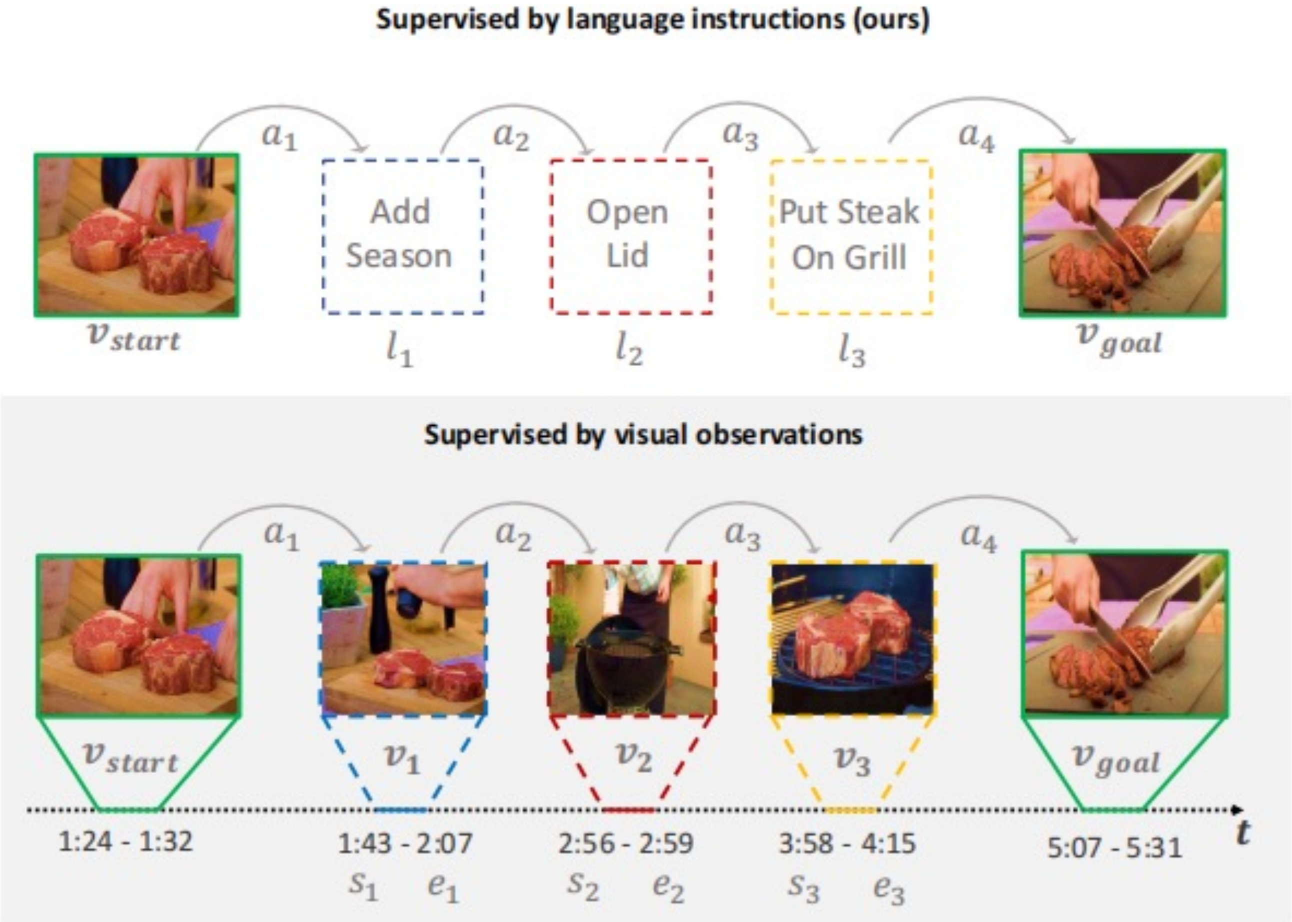


Spatio-temporal scene graphs

Future Direction 1: Structured 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**

Complex Situation + **Long Horizon** →

Surface → Deep

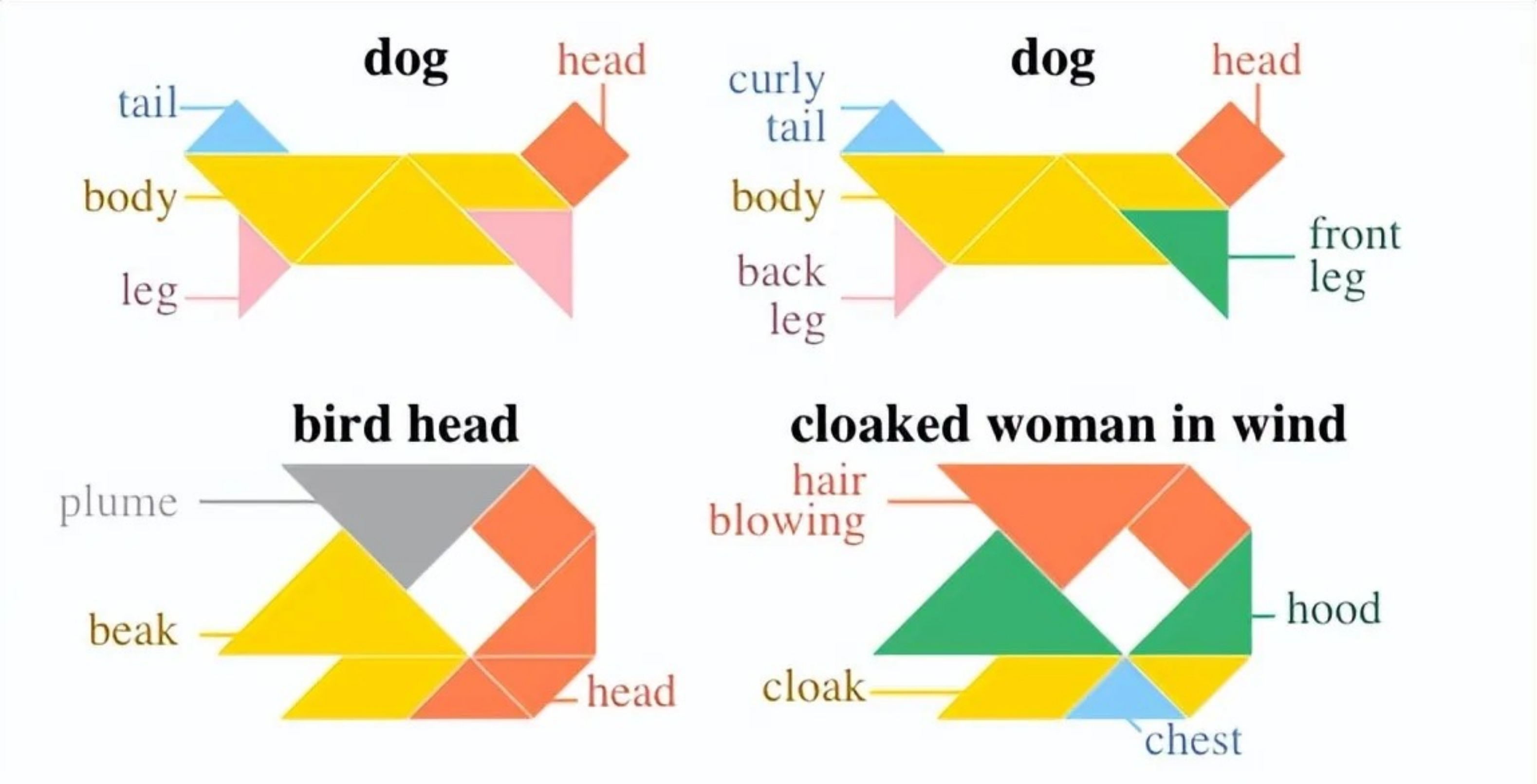
Concrete → Abstract

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

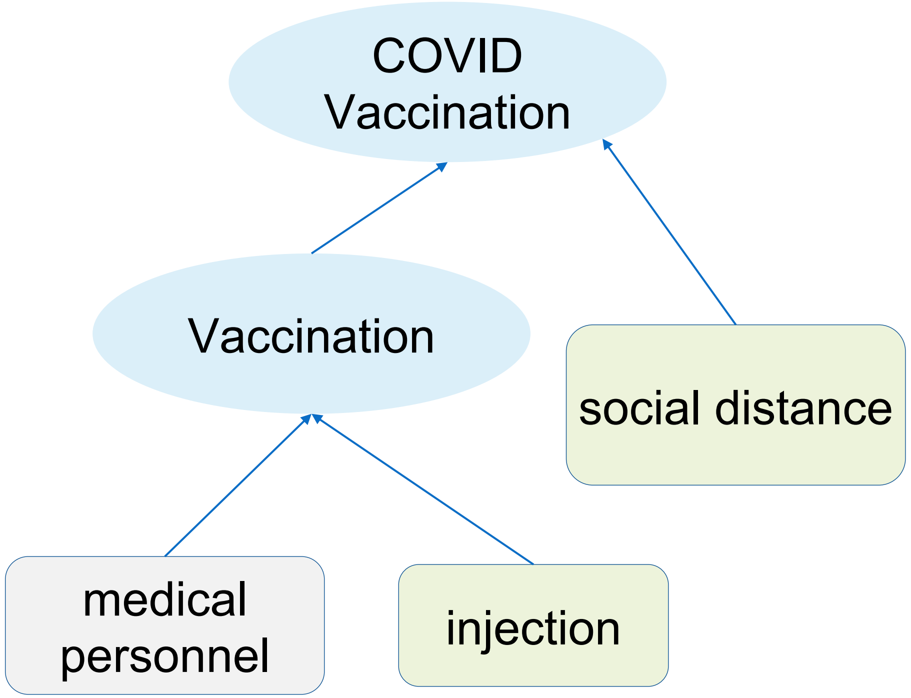
Love

Happiness

Emotion <-> Music

Future Direction 2: Compositional Semantics

Compositional



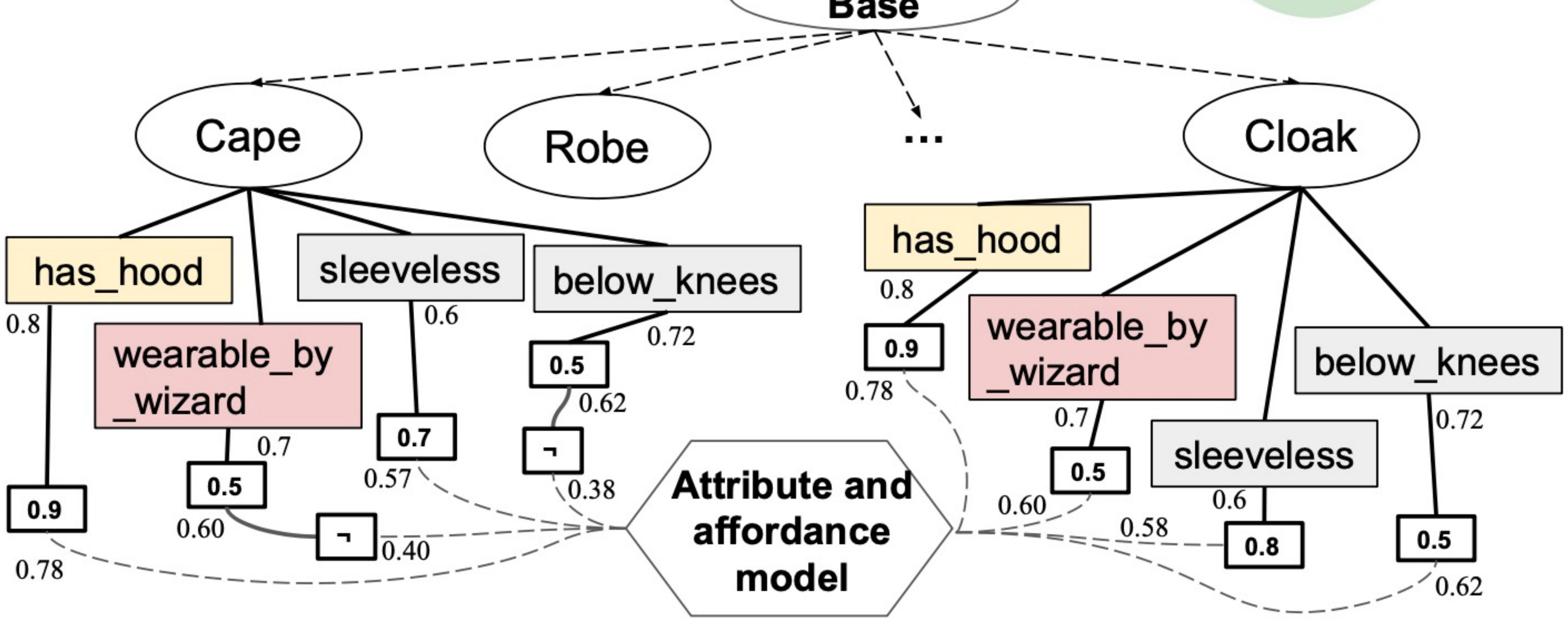
Future Direction 2: Compositional Semantics

Reasoning

Image and object bounding box

cape(x) = [0.9:has_hood](x) ∧ [0.7:sleeveless](x) ∧ (¬ [0.5:below_knees](x)) ∧ (¬ [0.5:wearable_by_wizard](x))

cloak(x) = [0.9:has_hood](x) ∧ [0.8:sleeveless](x) ∧ [0.5:below_knees](x) ∧ [0.5:wearable_by_wizard](x)



Future Research: From Surface to Deep Semantics

Text

Surface

Named Entity Recognition

Relation Extraction

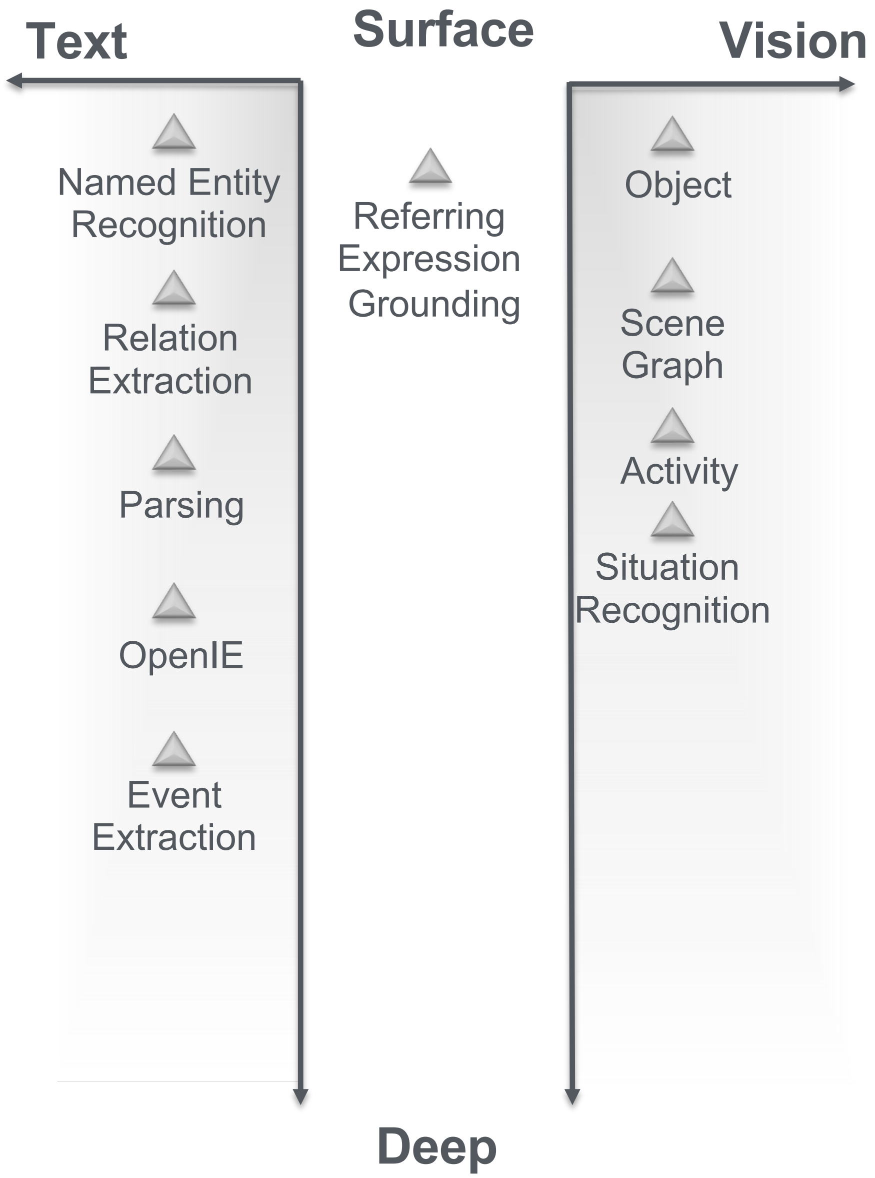
Parsing

OpenIE

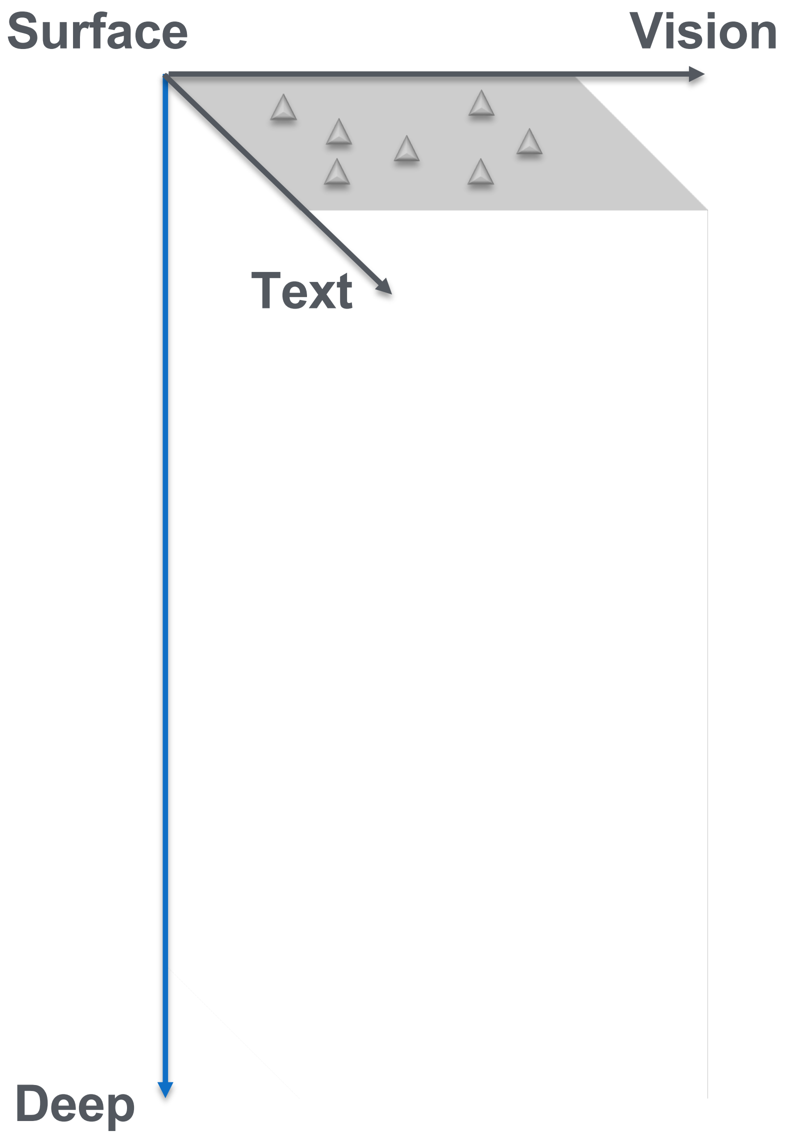
Event Extraction

Deep

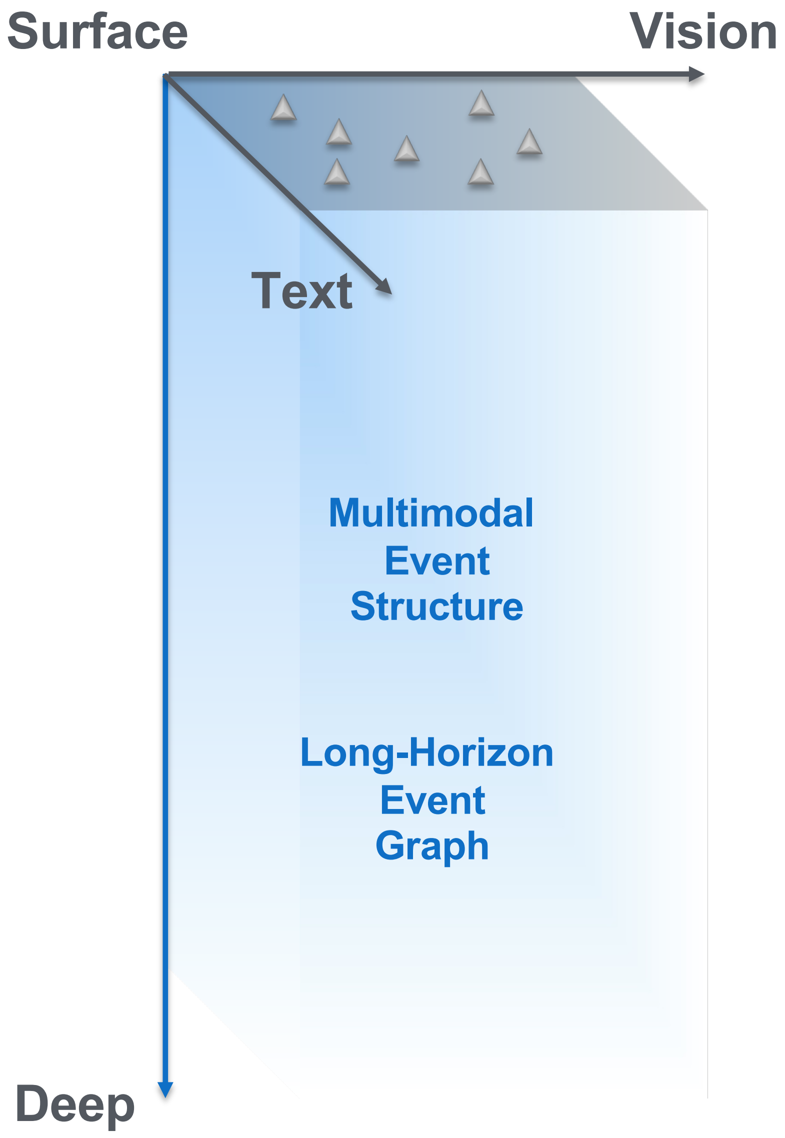
Future Direction: From Surface to Deep Semantics



Future Direction: From Surface to Deep Semantics



Future Direction: From Surface to Deep Semantics



Future Direction: From Surface to Deep Semantics