Multimedia IE

New Frontiers of Information Extraction (Part V)

Manling Li

Department of Computer Science

University of Illinois Urbana-Champaign

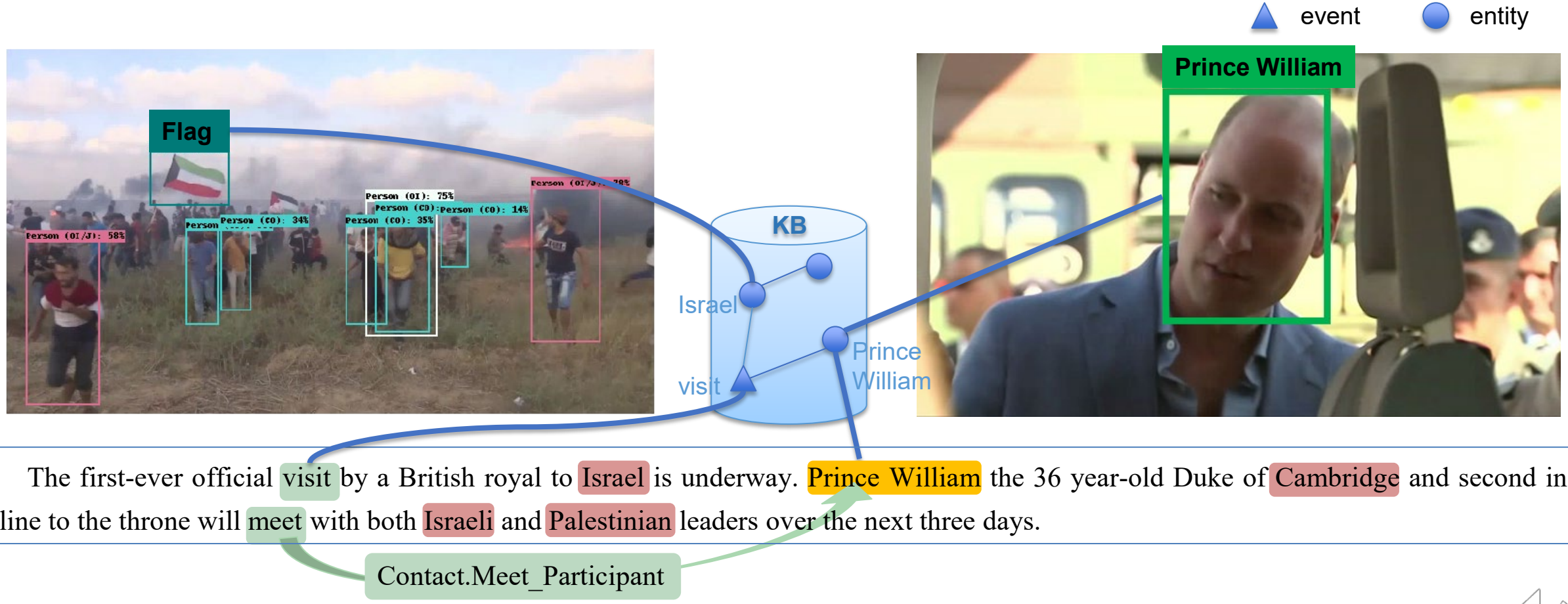
July 2022

NAACL Tutorials

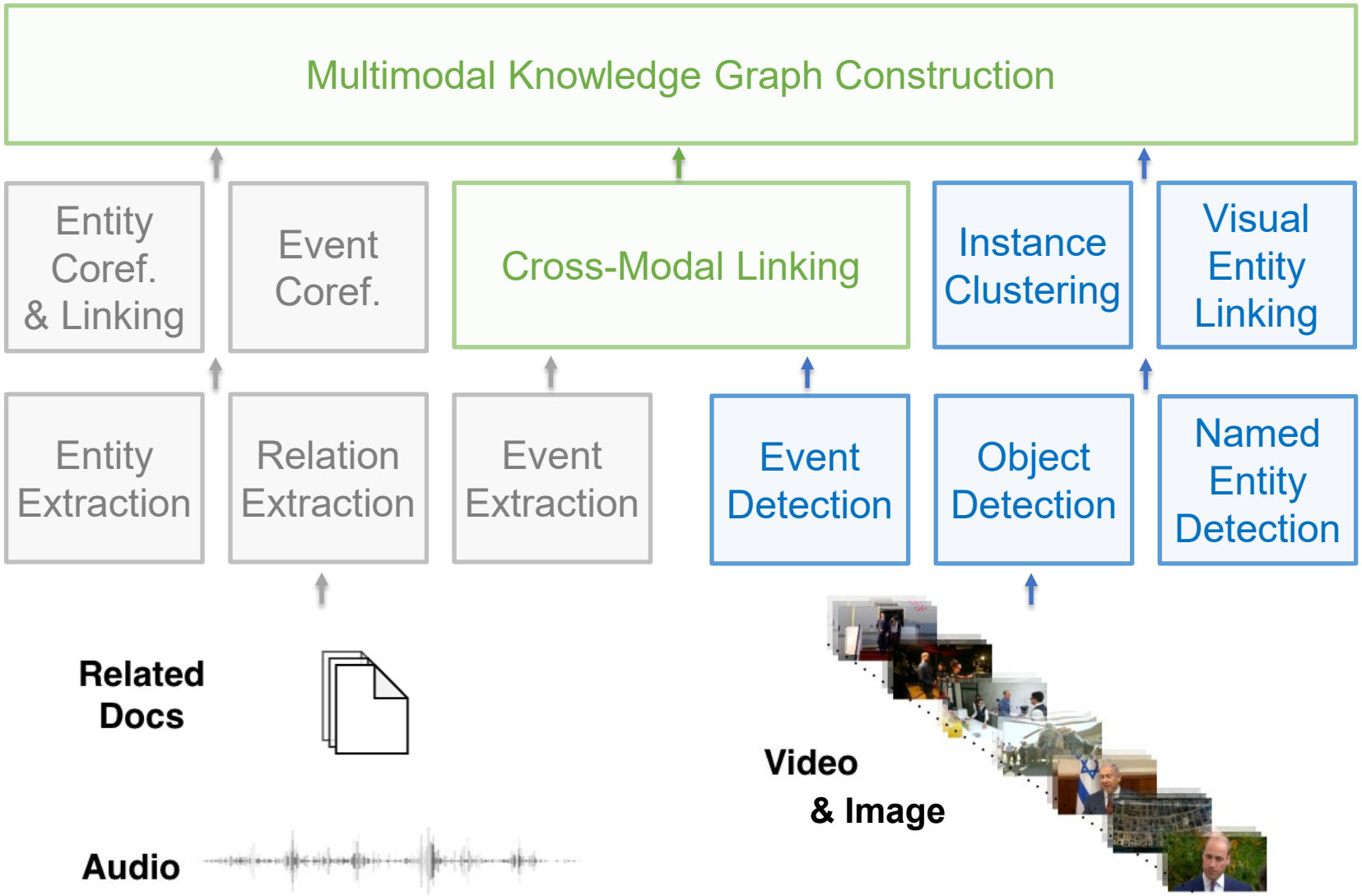
New Frontiers of Information Extraction

Multimedia Information Extraction

- Multimedia Knowledge Base with entities, relations and events.



Multimedia Information Extraction



Multimedia Information Extraction

- Treat Image/Video as a foreign language

Text	Image / Video Frame
Sentence	Image
Word	Image Region
Entity	Visual Object
Relation	Visual Relation
Entity-Relation Graph	Visual Scene Graph
Event Trigger	Visual Activity
Event Structure	**Image Event Graph**

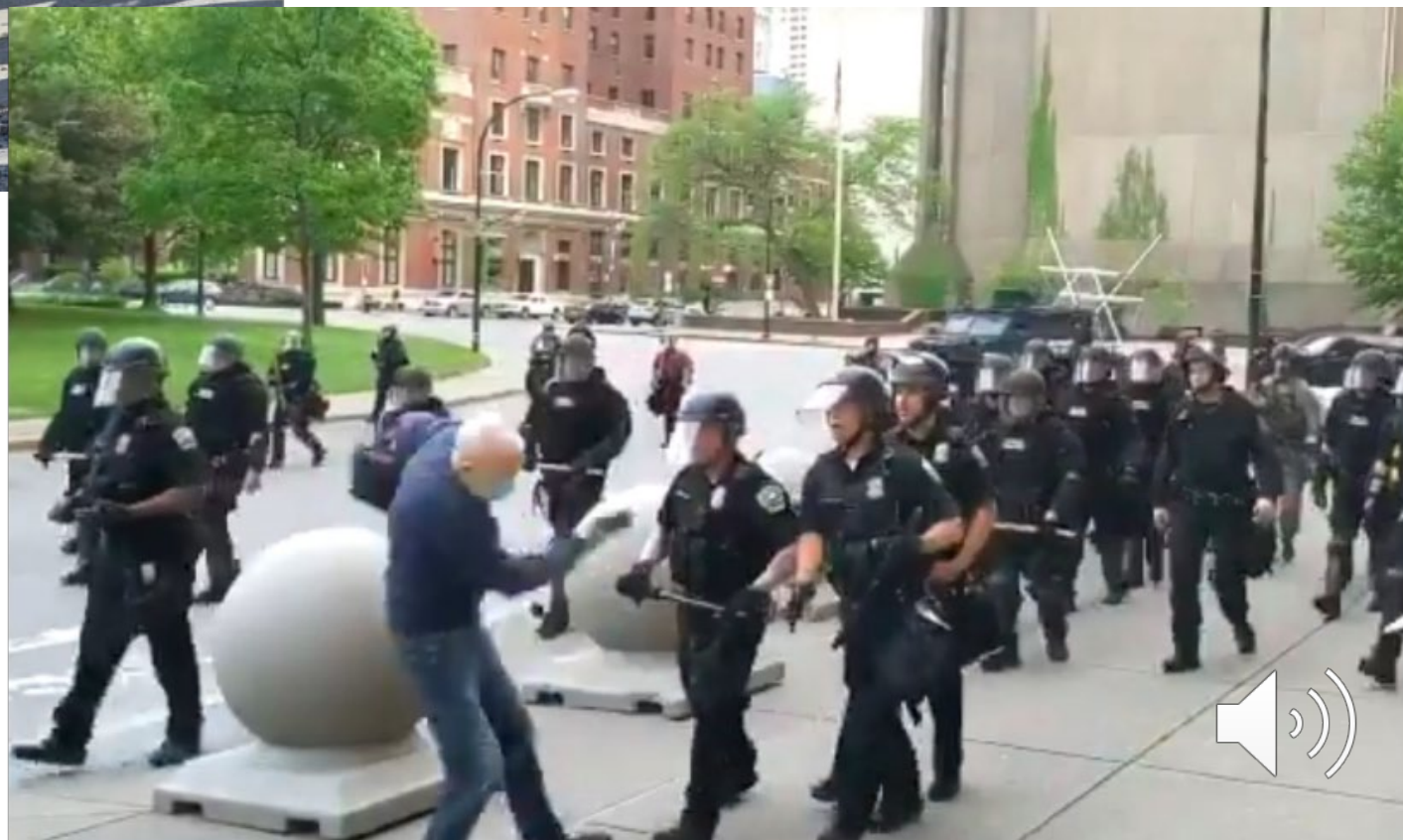
Object

Event

Event	Bombing
Item	Car

Argument

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

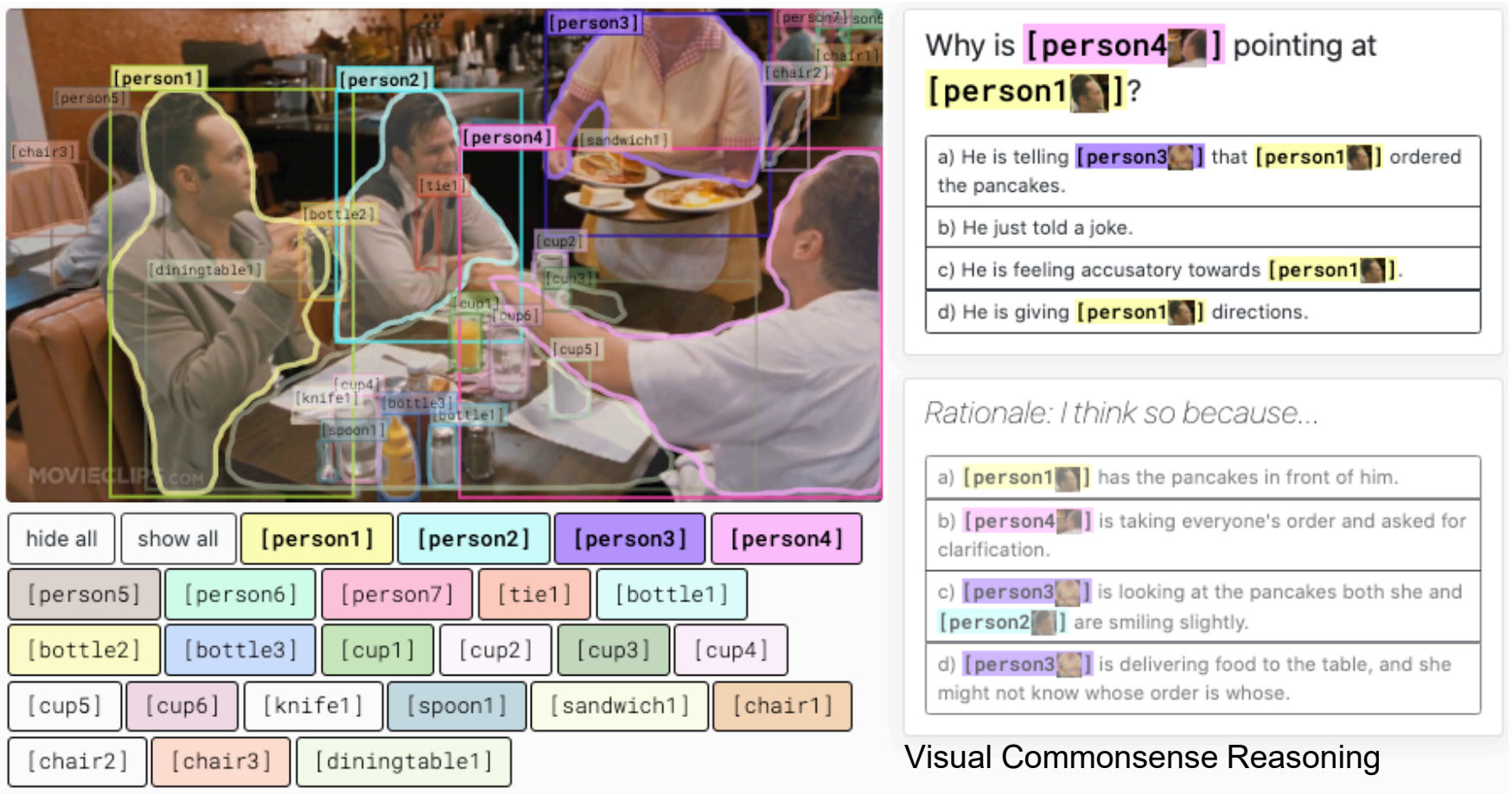
Event	Sanitizing
Agent	person
Tool	sprayer

Event	Testing
Agent	woman
place	car

Event	Vaccination
Agent	woman
Target	girl

Multimedia Applications

- Real-world multimedia applications requires image-language models to understand multiple levels of alignments such as **events**, **objects**, as well as **semantic structures**.

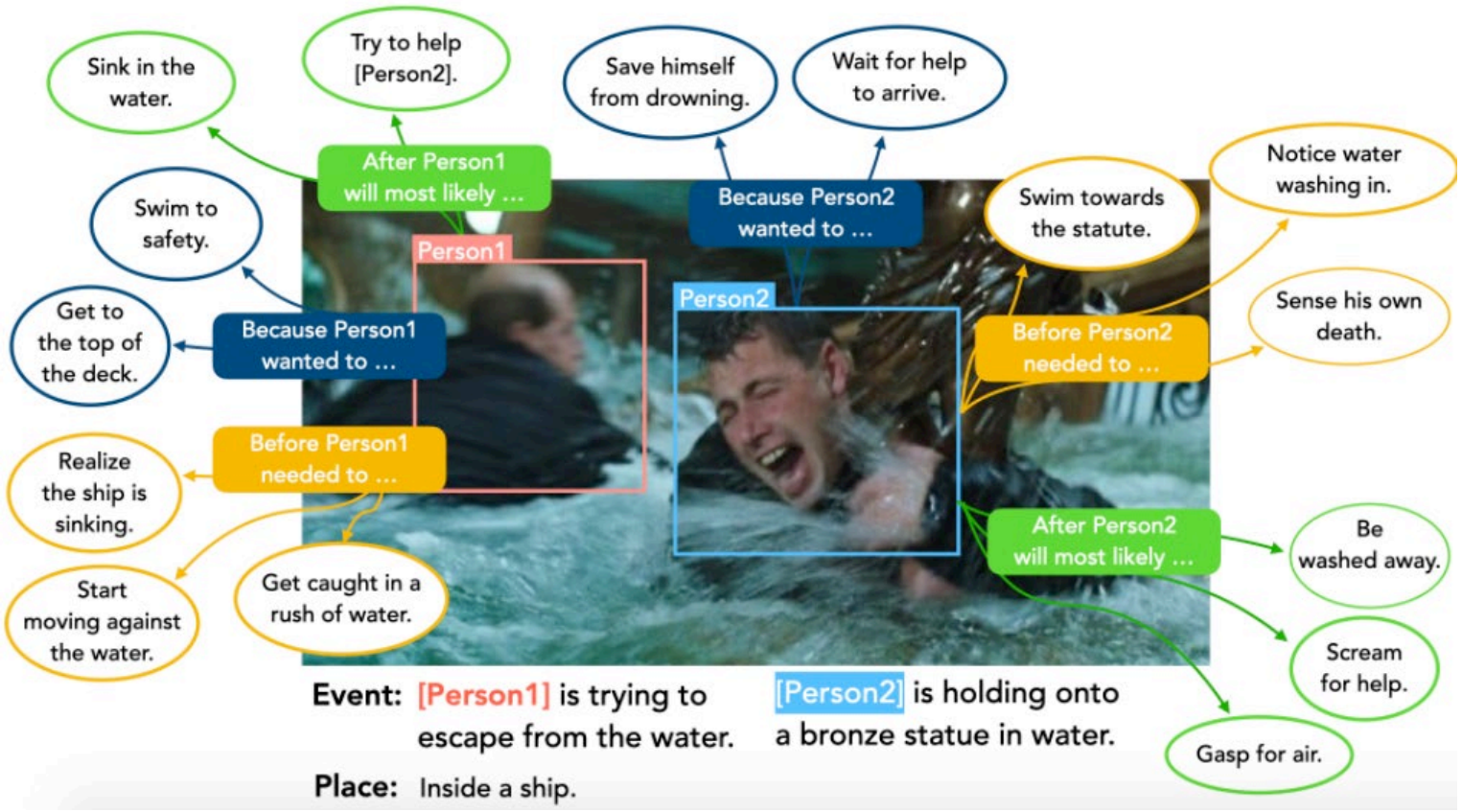


Visual Commonsense Reasoning

Zellers, Rowan, et al. "From recognition to cognition: Visual commonsense reasoning." *CVPR*. 2019.

Multimedia Applications

- Real-world multimedia applications requires image-language models to understand multiple levels of alignments such as **events**, **objects**, as well as **semantic structures**.

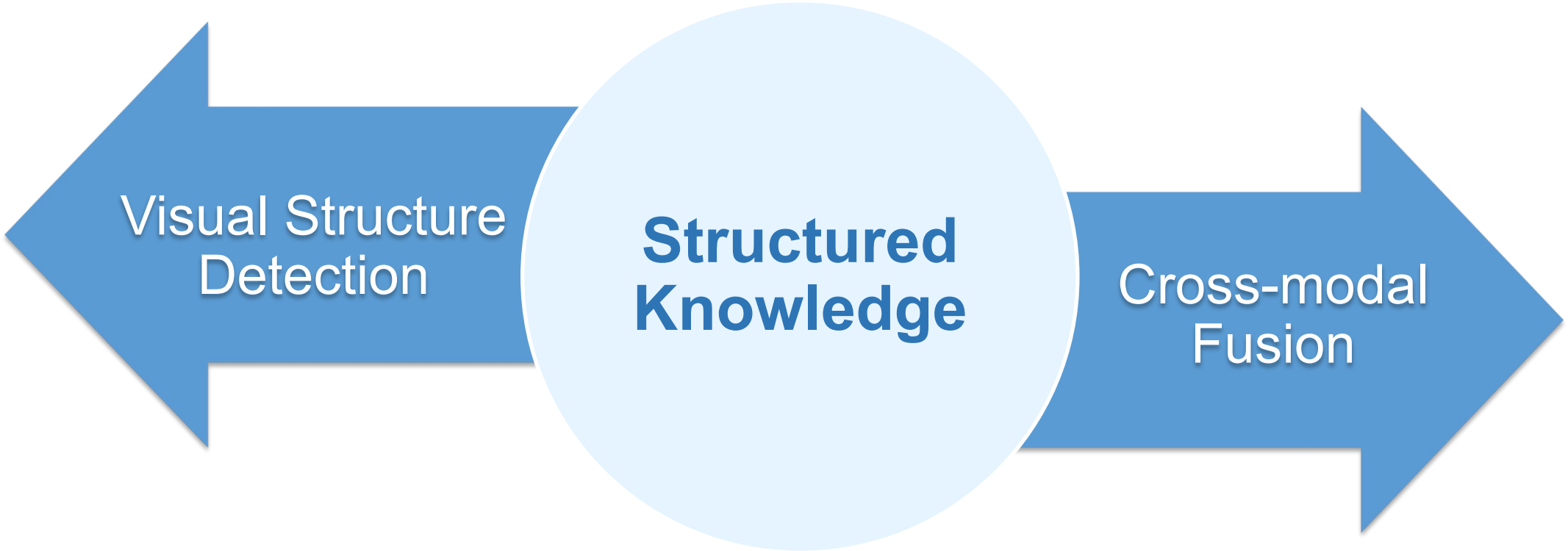


VisualCOMET (Visual Commonsense Reasoning in Time)

Park, Jae Sung, et al. "Visualcomet: Reasoning about the dynamic context of a still image." ECCV. Springer, Cham, 2020.

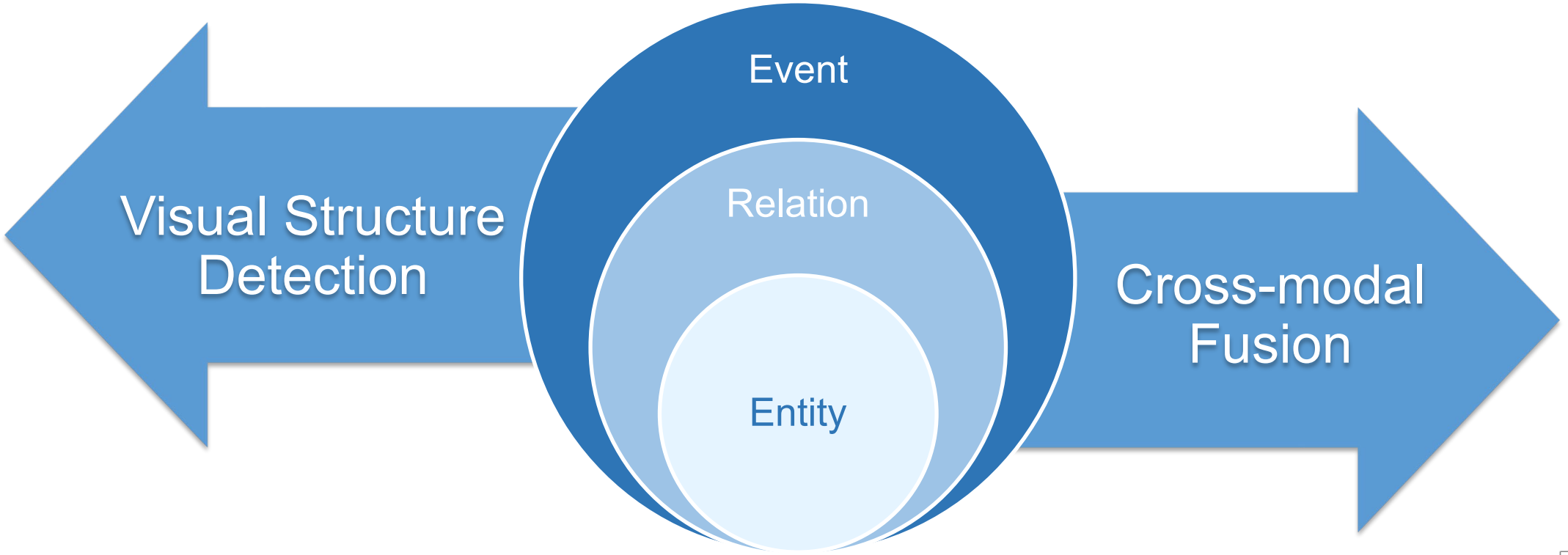
Challenges

- 1. How do we find structured knowledge in vision data?
- 2. How do we align structured across vision and text?



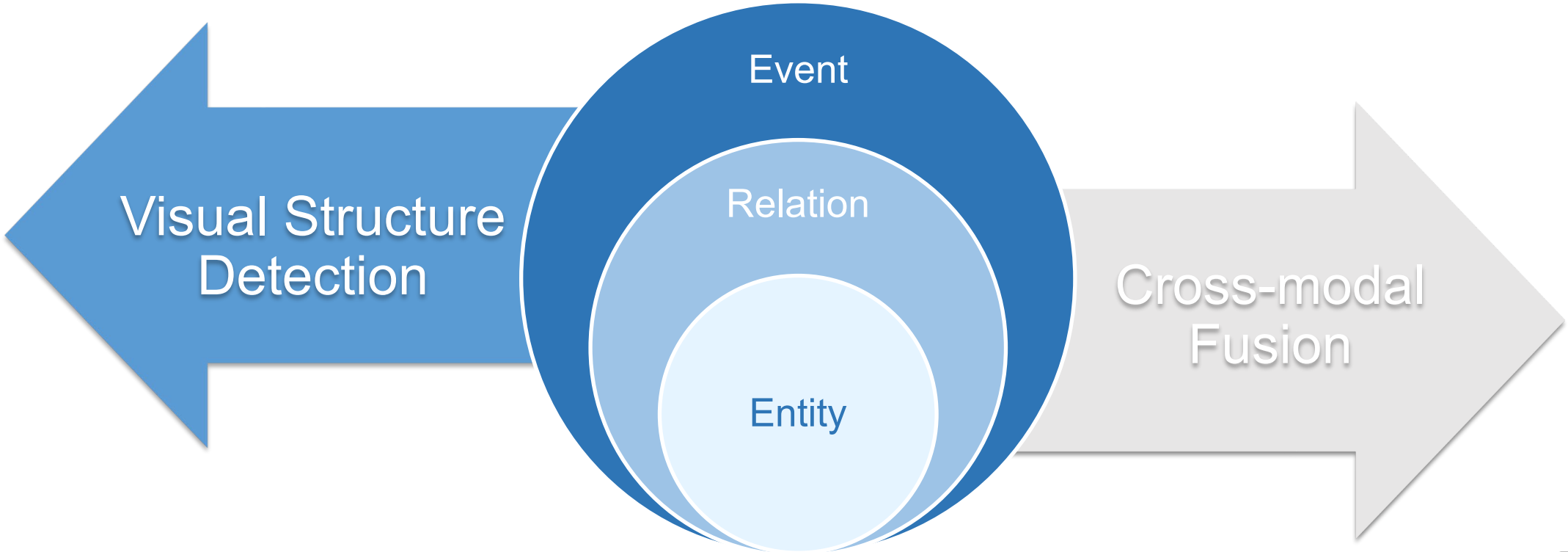
Challenges

- 1. How do we find structured knowledge in vision data?
- 2. How do we align structured across vision and text?

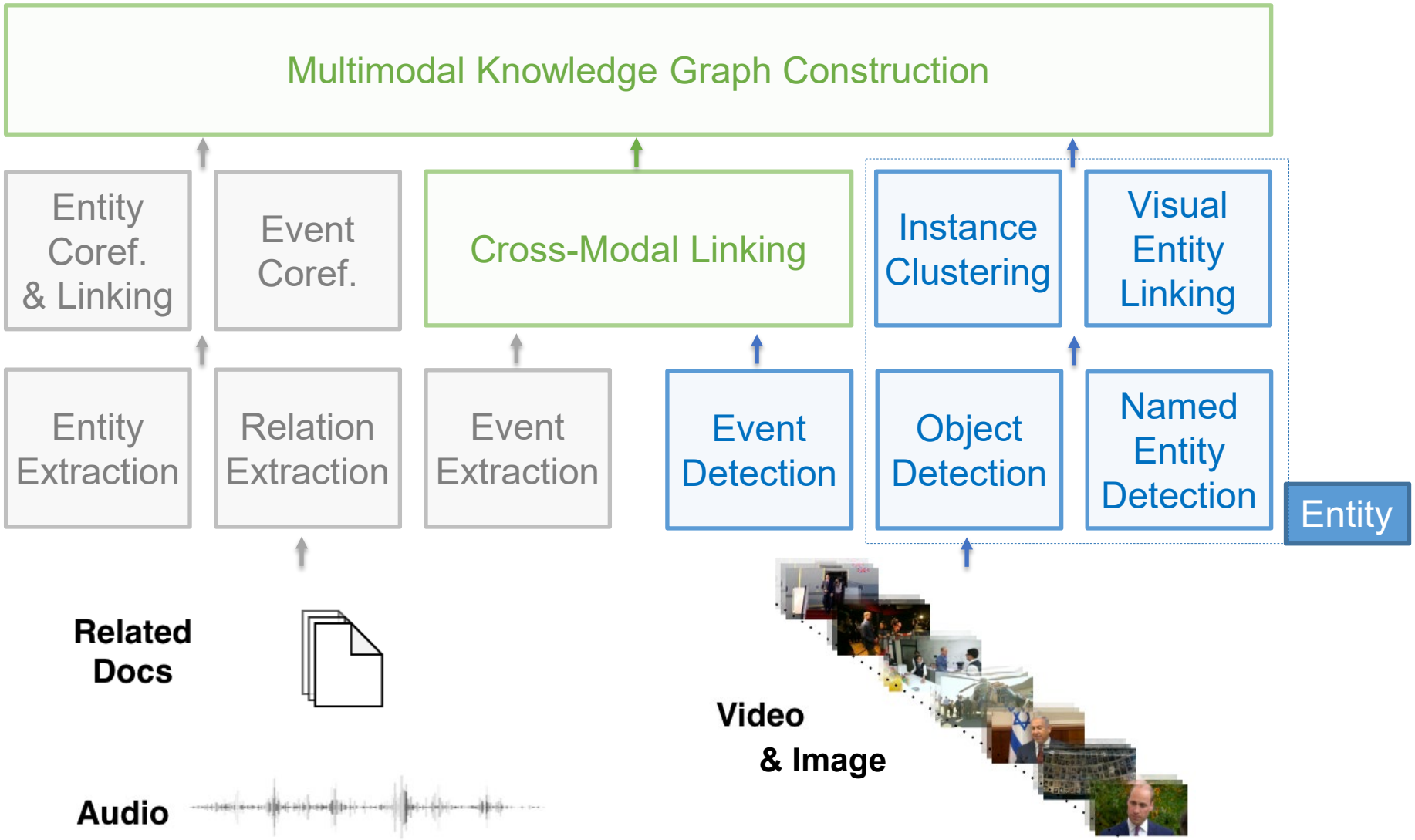


Challenges

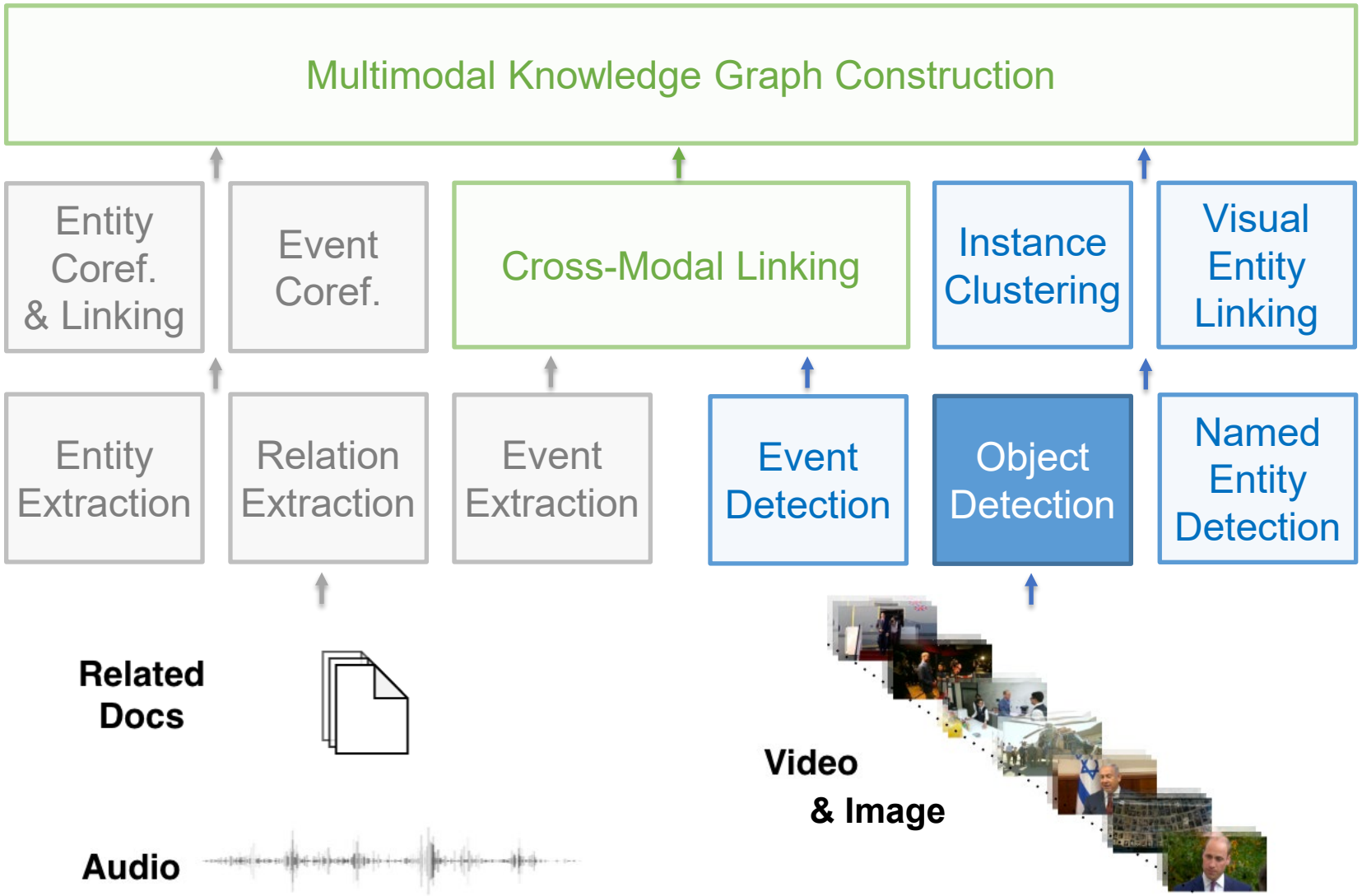
- 1. How do we find structured knowledge in vision data?
- 2. How do we align structured across vision and text?



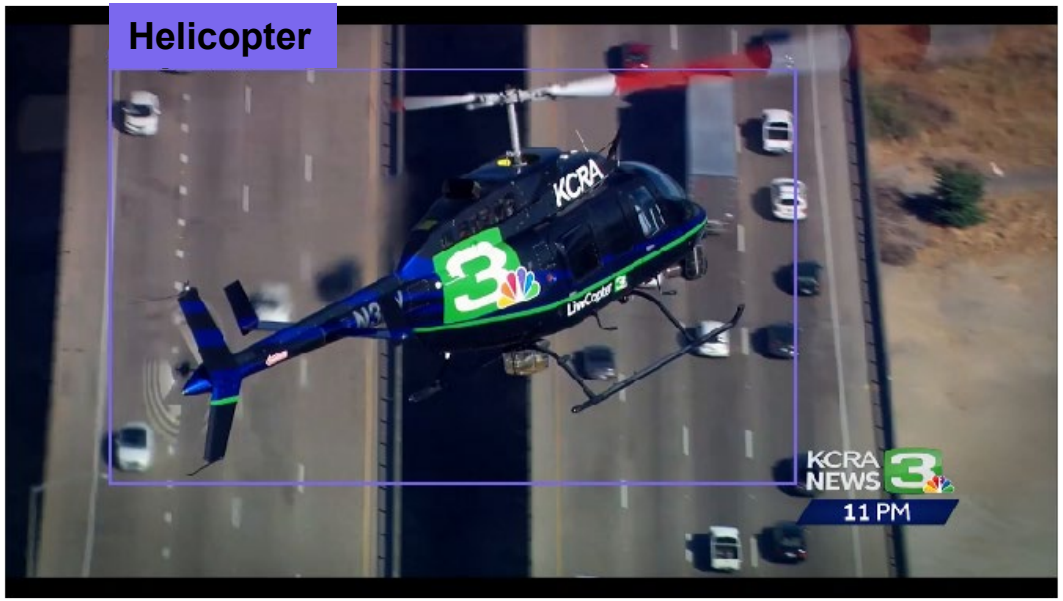
Multimedia Information Extraction

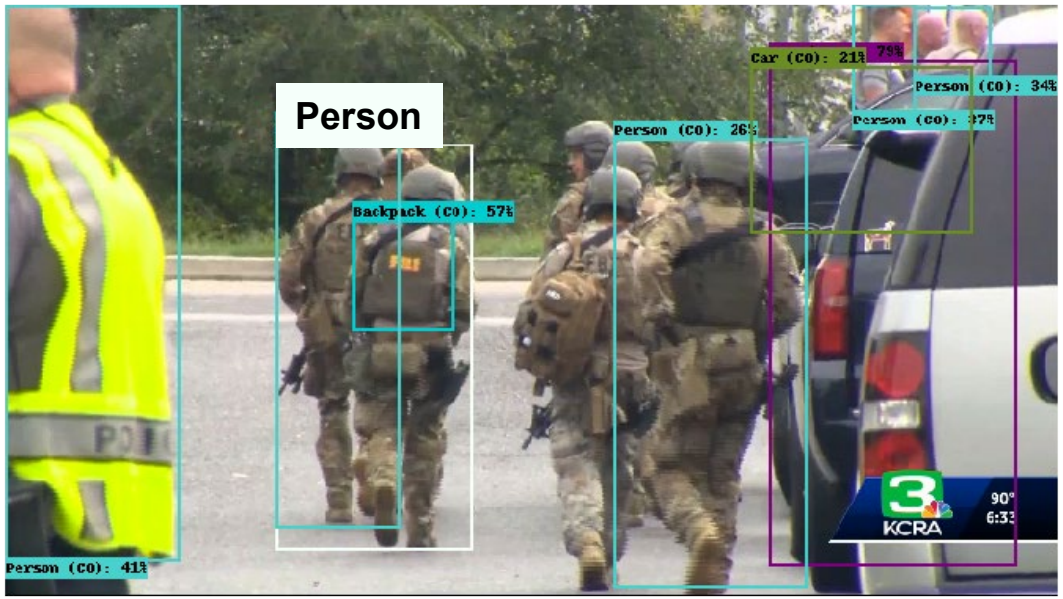


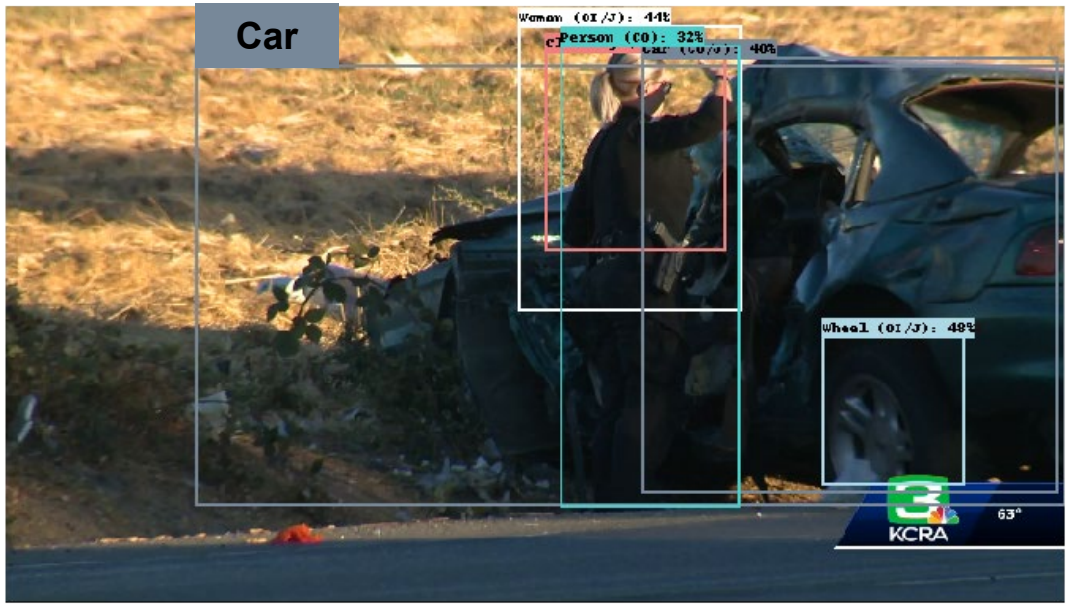
Multimedia Information Extraction



Visual Entity Extraction: Object Detection

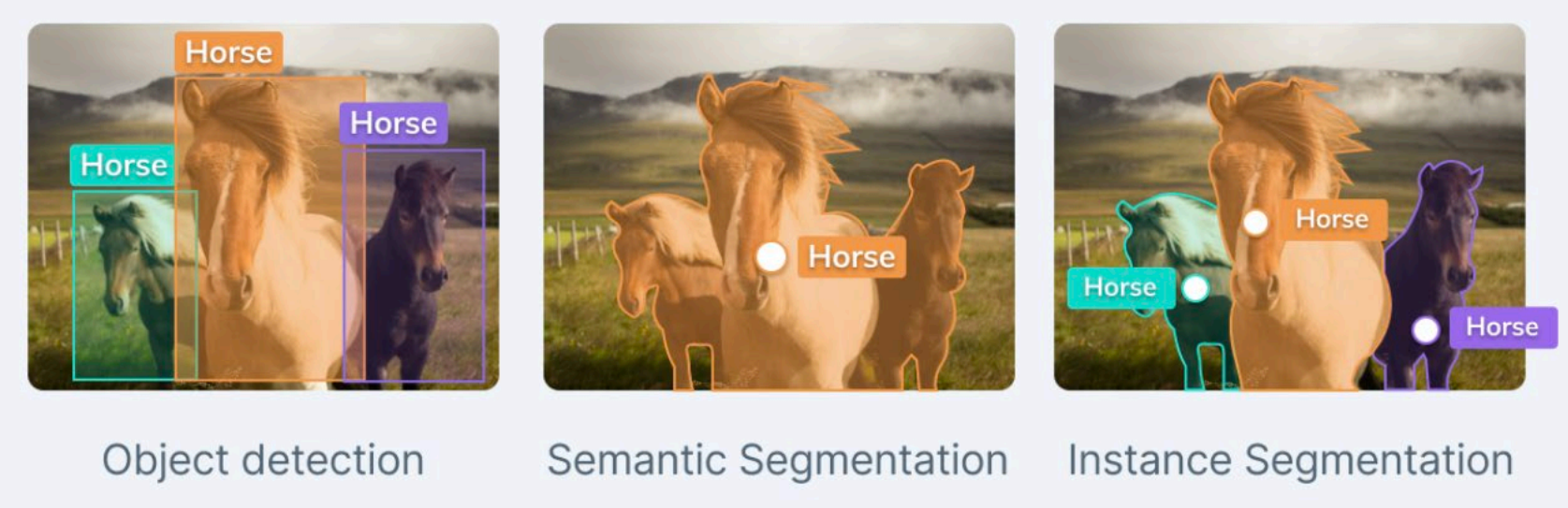






Visual Entity Extraction

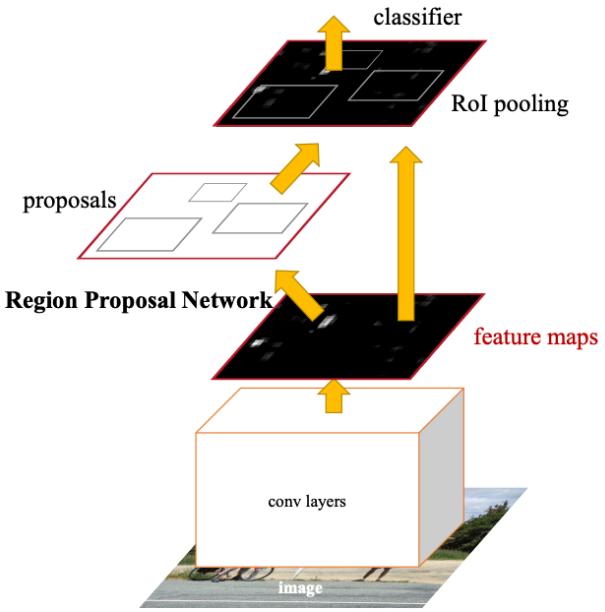
- Object Detection: Object instances at the bounding box level
- Semantic Segmentation: Object class at the pixel level
- Instance Segmentation: Object instances at the pixel level



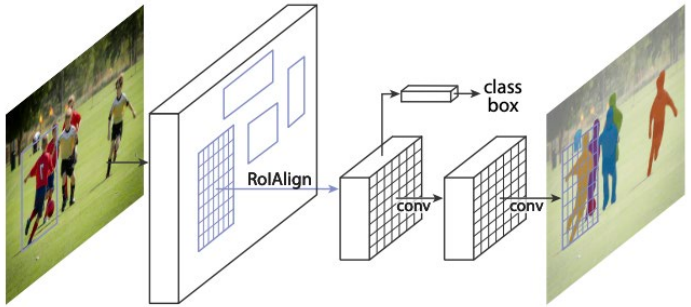
https://www.v7labs.com/blog/object-detection-guide

Visual Entity Extraction

Two-Stage (With Proposal)

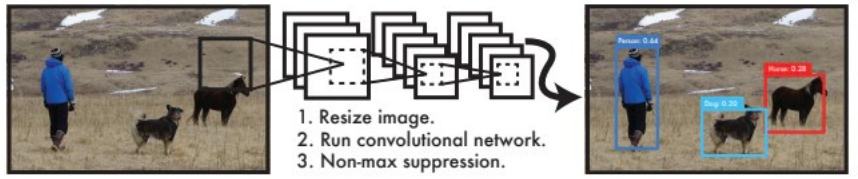


Faster RCNN

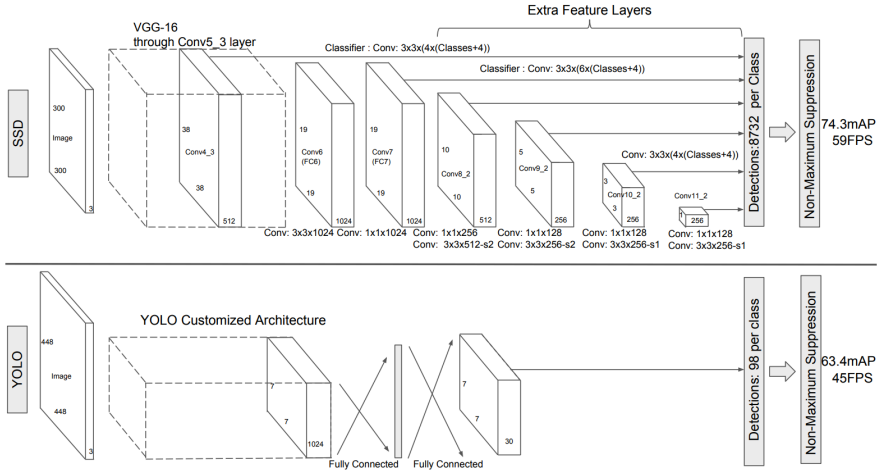


Mask RCNN

One-Stage (Without Proposal)



YOLO



SSD

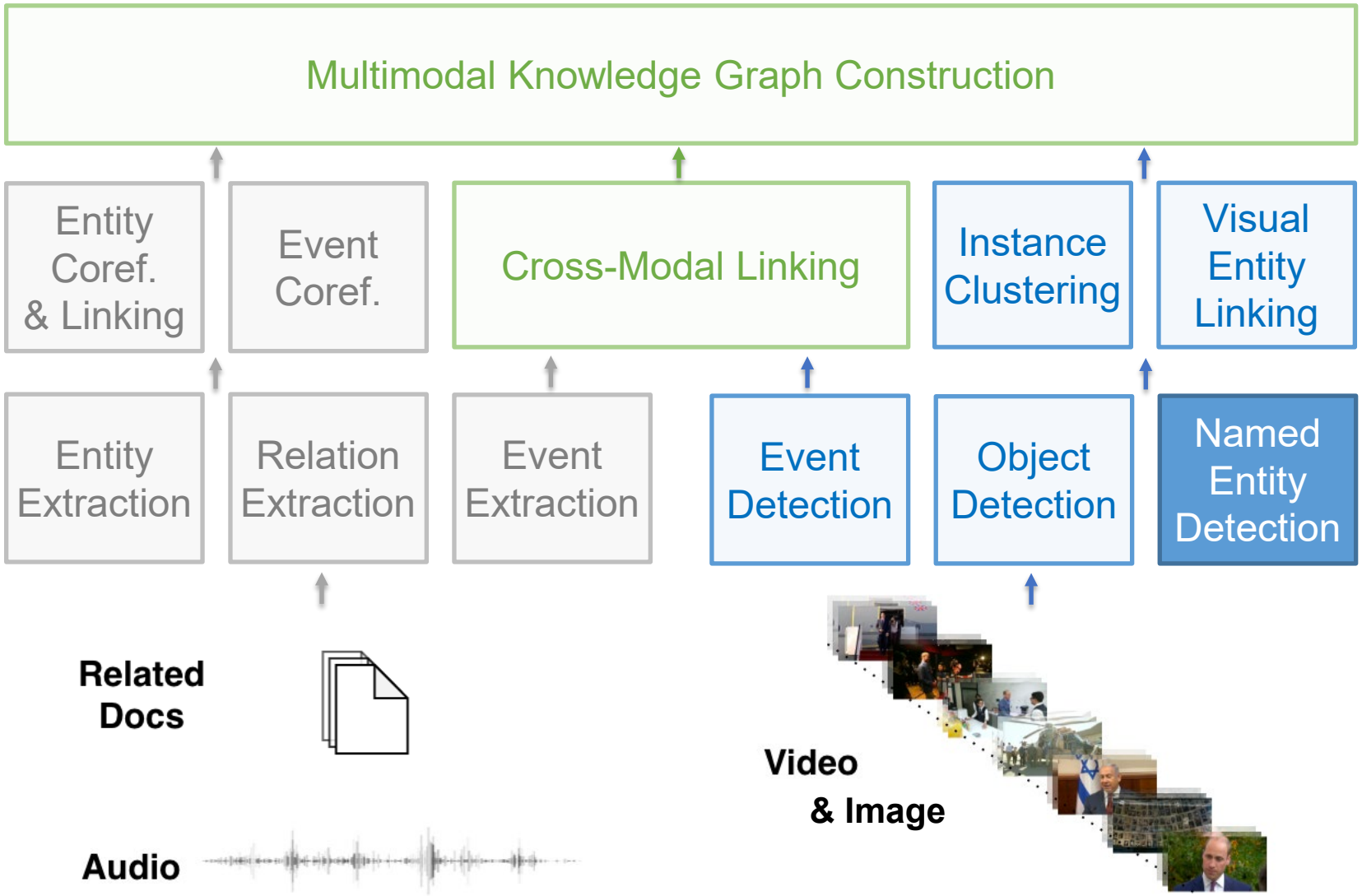
Ren, S., He, K., Girshick, R., & Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS 2015*.

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

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

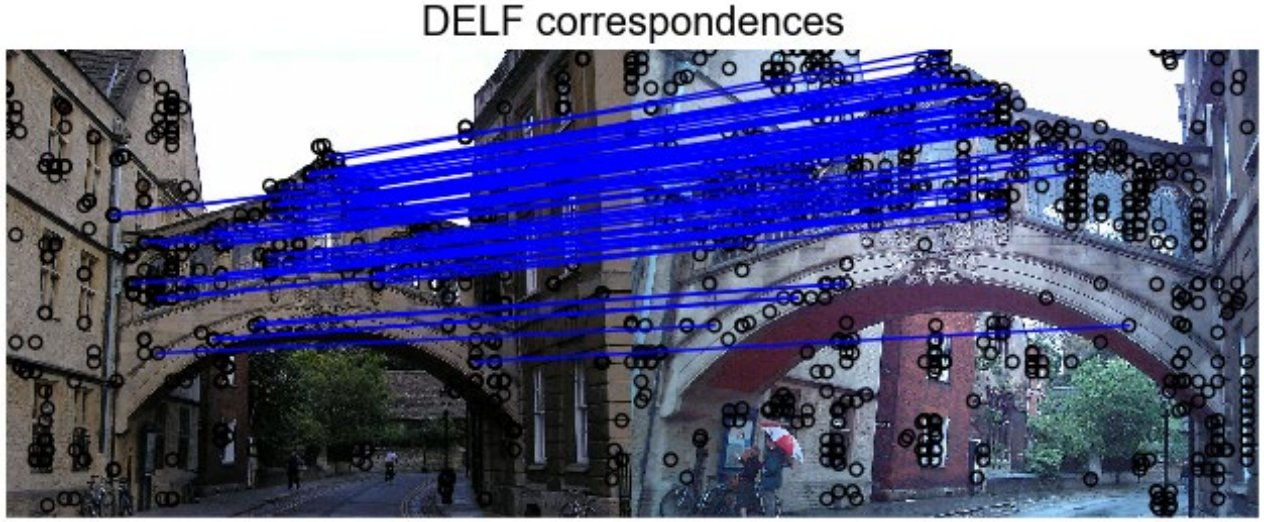
Liu, Wei, et al. "Ssd: Single shot multibox detector." *ECCV* 2016.

Multimedia Information Extraction



Visual Entity Linking: Named Entity Recognition

- Named Entity Recognition
  - Face Recognition
  - Landmark Recognition
  - Flag Recognition
  - Logo Recognition
  - …



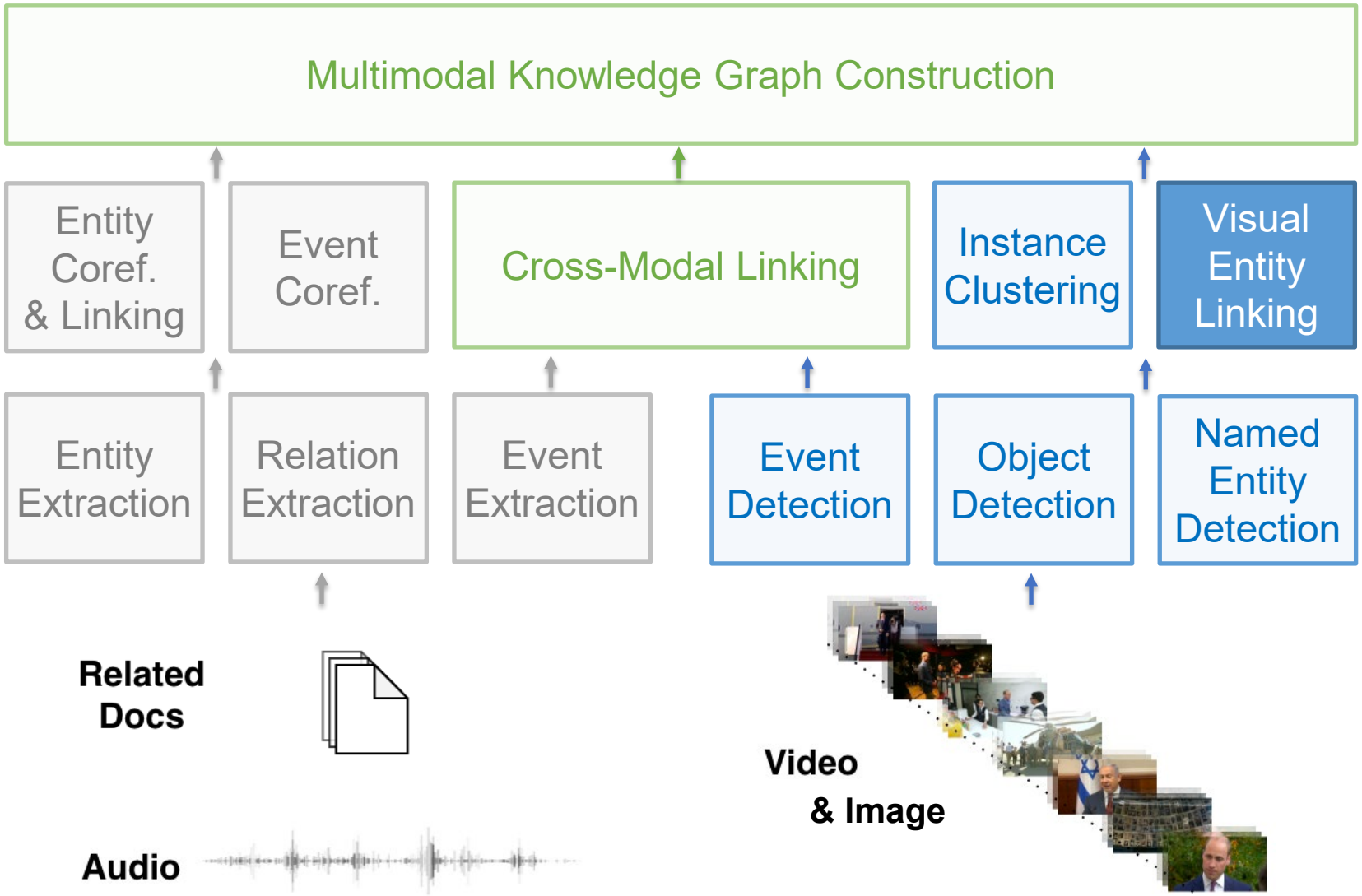
Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." *CVPR* 2015.

Weyand, Tobias, et al. "Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval." *CVPR* 2020.

Wu, Shou-Fang, et al. "FlagDetSeg: Multi-Nation Flag Detection and Segmentation in the Wild." *AVSS* 2021.

Bianco, Simone, et al. "Deep learning for logo recognition." *Neurocomputing* 245 (2017): 23-30.

Multimedia Information Extraction



Visual Entity Linking and Coreference

- Landmark recognition

Visual Entity Linking and Coreference

- Flag recognition

US Flag

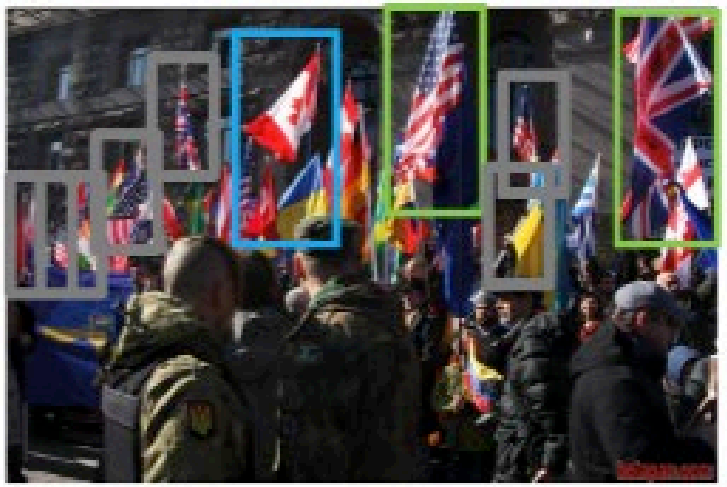
Missed detection

Different shape and angle
Partial observe

Ukraine Flag

Low resolution

Russia Flag (X) US Flag UK Flag



Visual Entity Linking and Coreference

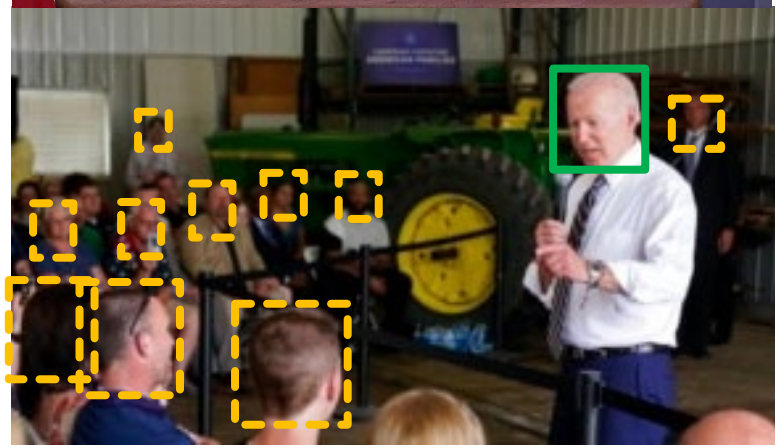
- Face detection
- Entity linking and coreference

Unknown people

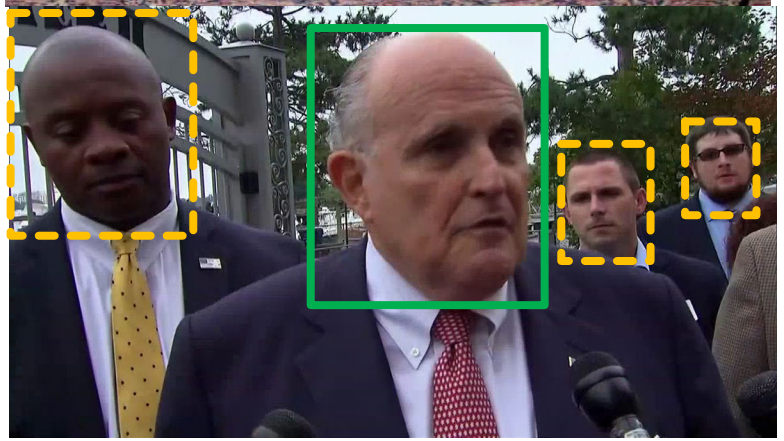
Kirstjen Nielsen

Joe Biden

Vladimir Putin

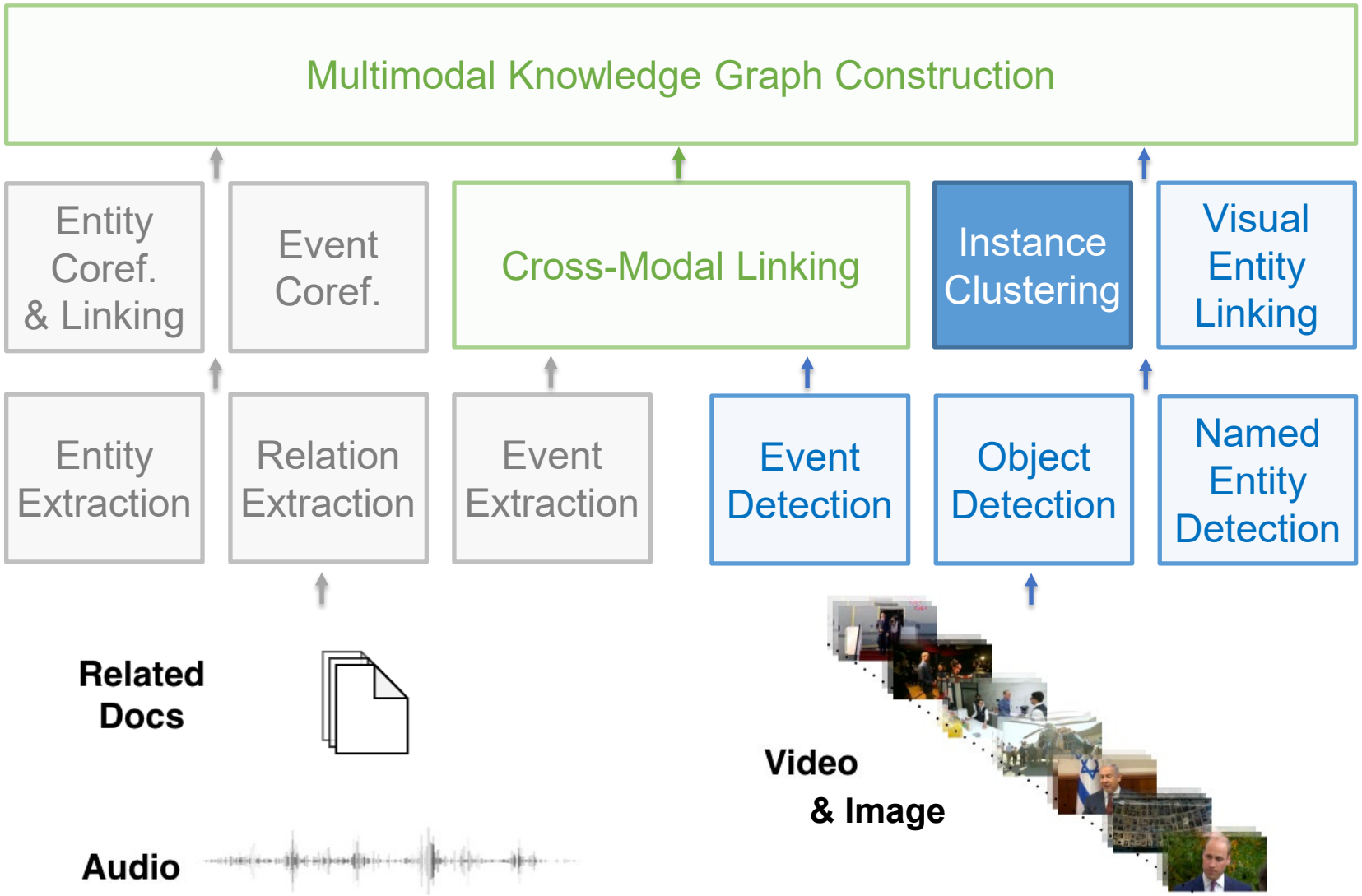


Joe Biden



Rudy Giuliani

Multimedia Information Extraction

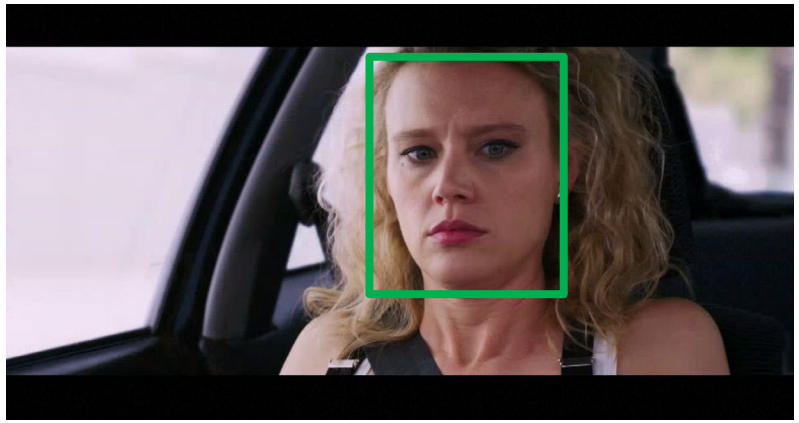


Visual Entity Linking and Coreference

- Face detection
- Entity linking and coreference

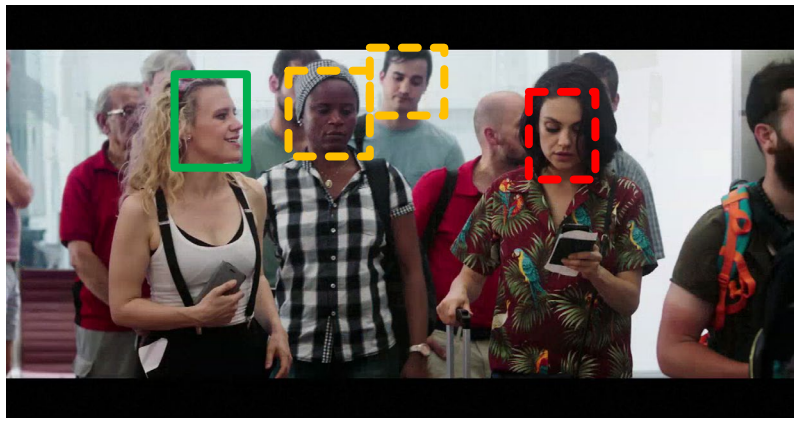
Face cluster

Unknown people



Kirstjen Nielsen

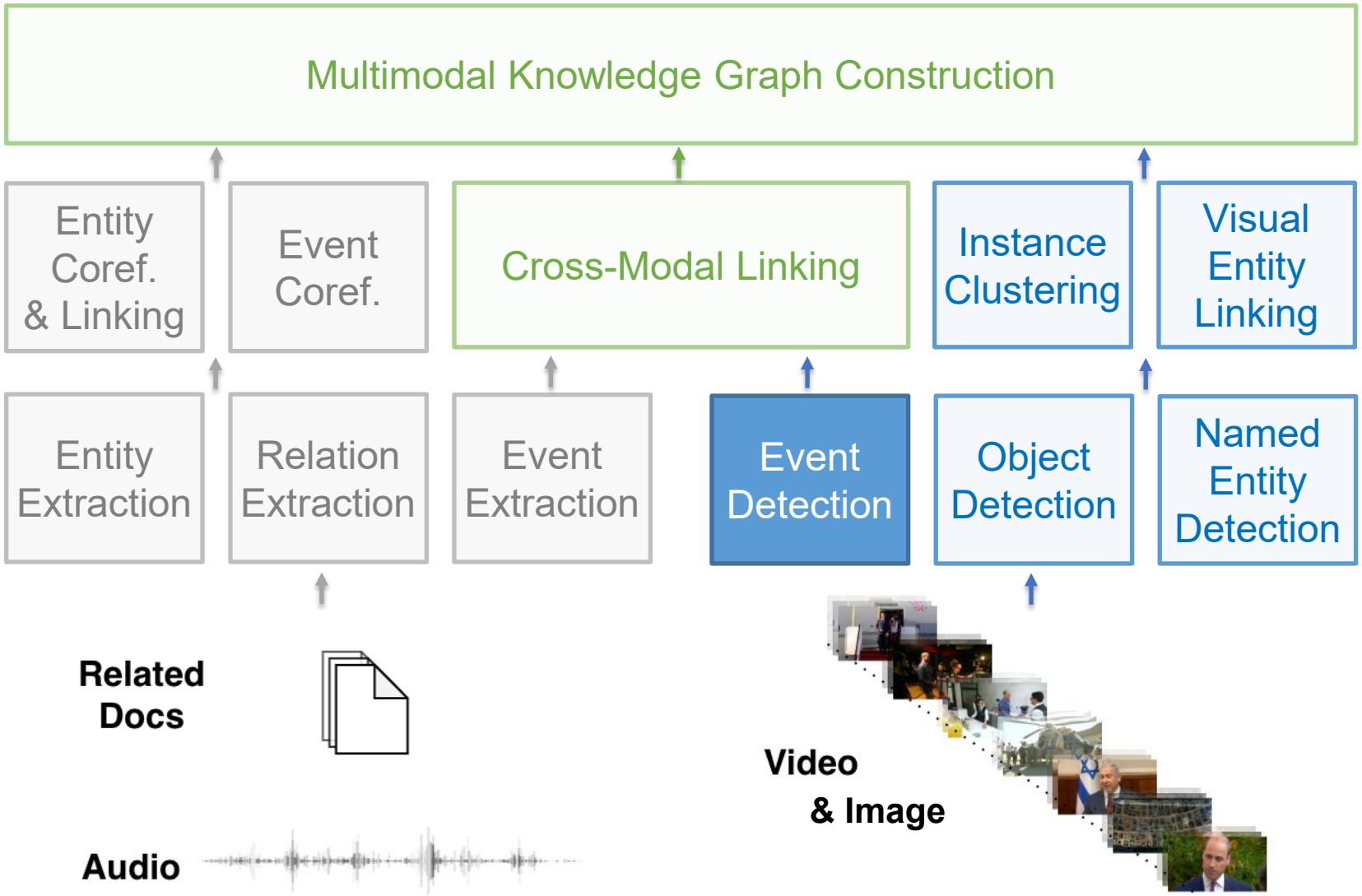
Kirstjen Nielsen



Kirstjen Nielsen

Kirstjen Nielsen

Multimedia Information Extraction



Visual Event Extraction: Situation Recognition

- Situation Recognition (image → text + boundingbox)

RIDING							
ROLE	AGENT	VEHICLE	PLACE	ROLE	AGENT	VEHICLE	PLACE
VALUE	MAN	HORSE	OUTSIDE	VALUE	DOG	SKATEBOARD	SIDEWALK

Situation Recognition [Yatskar et al. 2016]

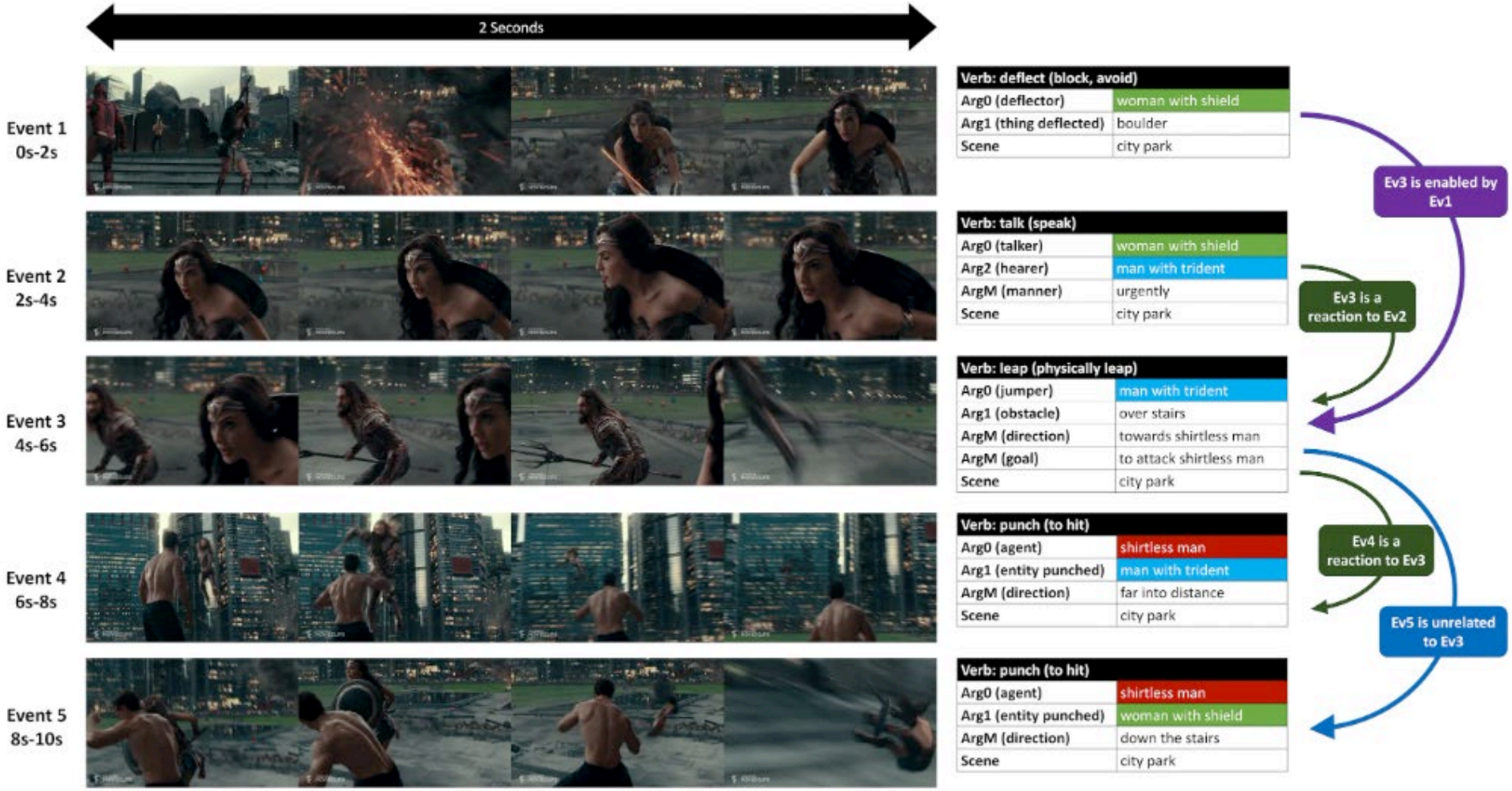
Hitting				
Agent	Tool	Victim	Victim Part	Place
Ballplayer	Bat	Baseball	Ø	Field

Catching			
Agent	Caught Item	Tool	Place
Bear	Fish	Mouth	River

Grounded Situation Recognition [Pratt et al, 2020]

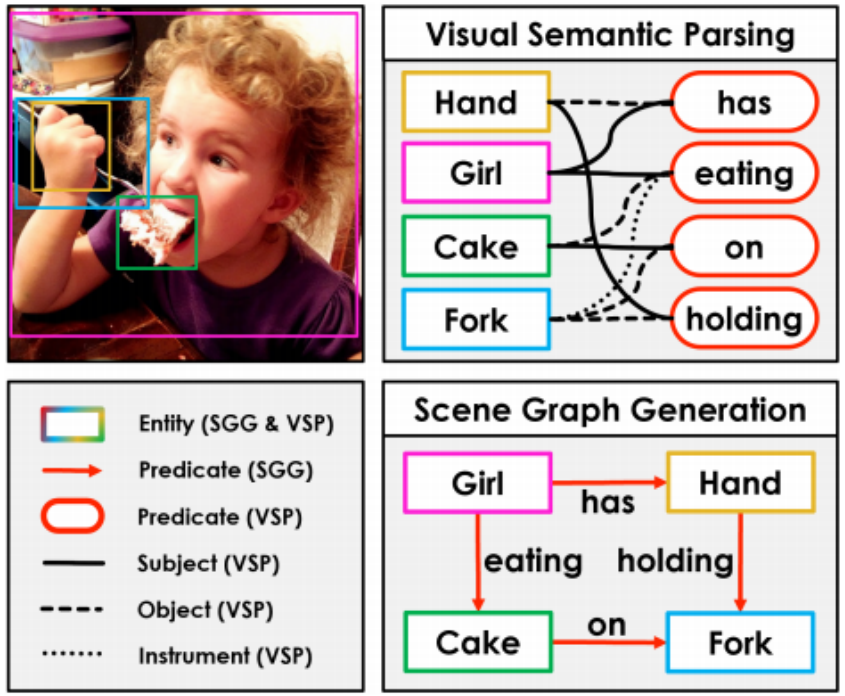
Visual Event Extraction: Video Situation Recognition

- Visual Features can provide more details for event interactions

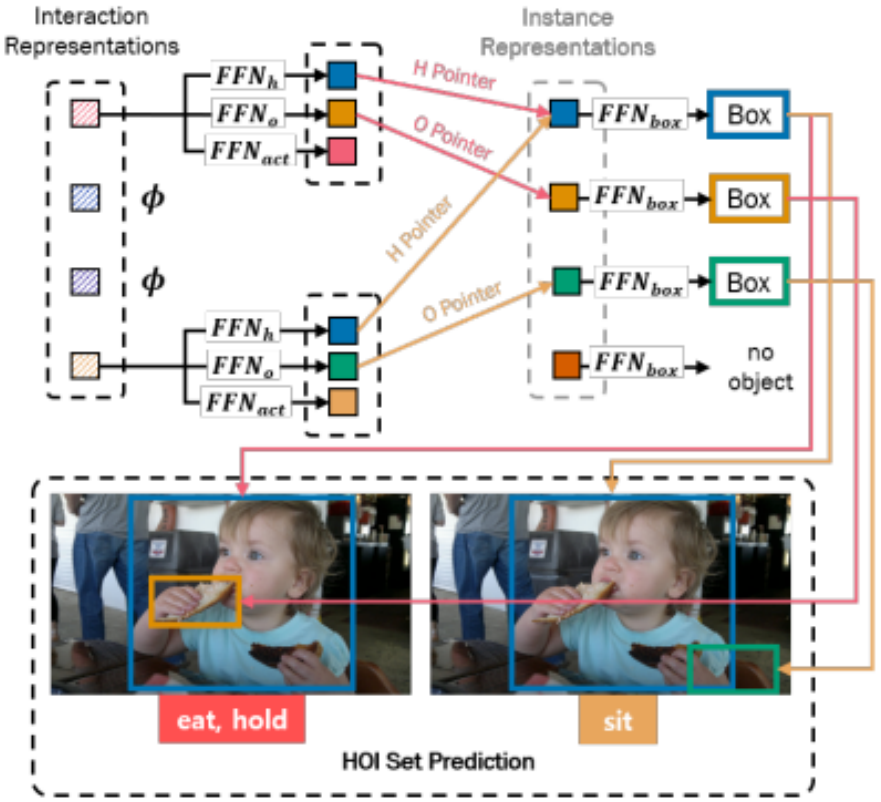


Visual Event Extraction: Visual Semantic Parsing

- Scene Graph Based (image → text + boundingbox)
  - Trained and evaluated on Scene Graph annotation (Visual Genome)
  - Added bounding box alignment



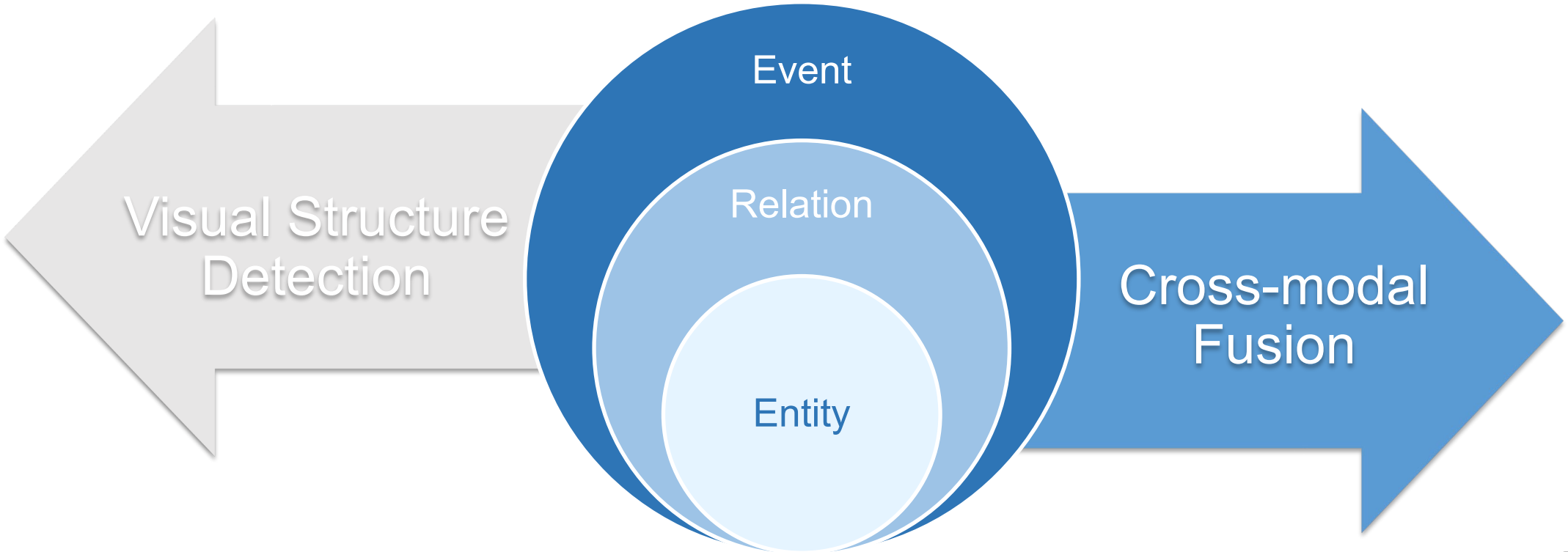
Visual Semantic Parsing [Zareian el al, 2020]



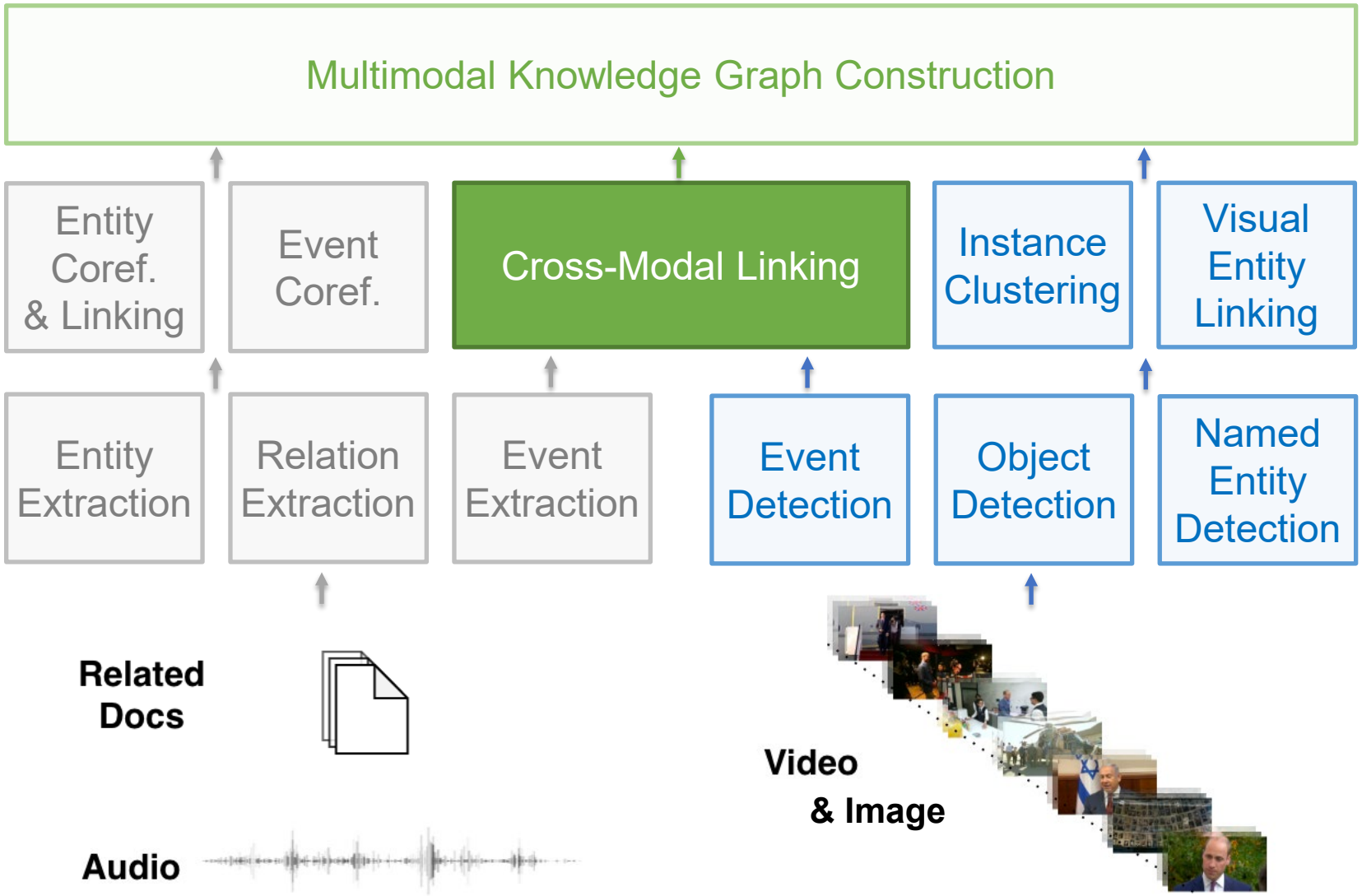
Human-Object Interaction [Kim el al, 2021]

Challenges

- 1. How do we find structured knowledge in vision data?
- 2. How do we align structured across vision and text?



Multimedia Information Extraction

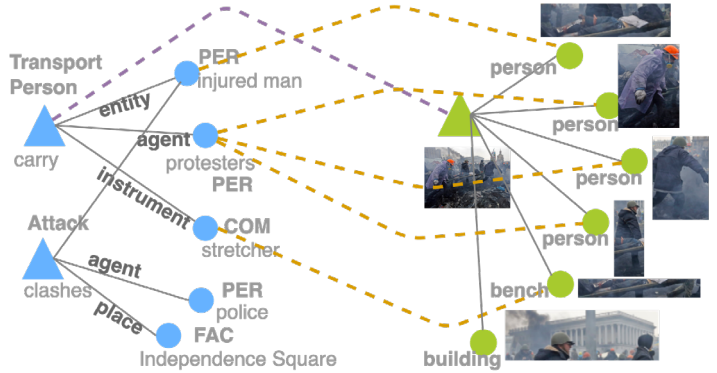


Cross-Media Fusion: Outline

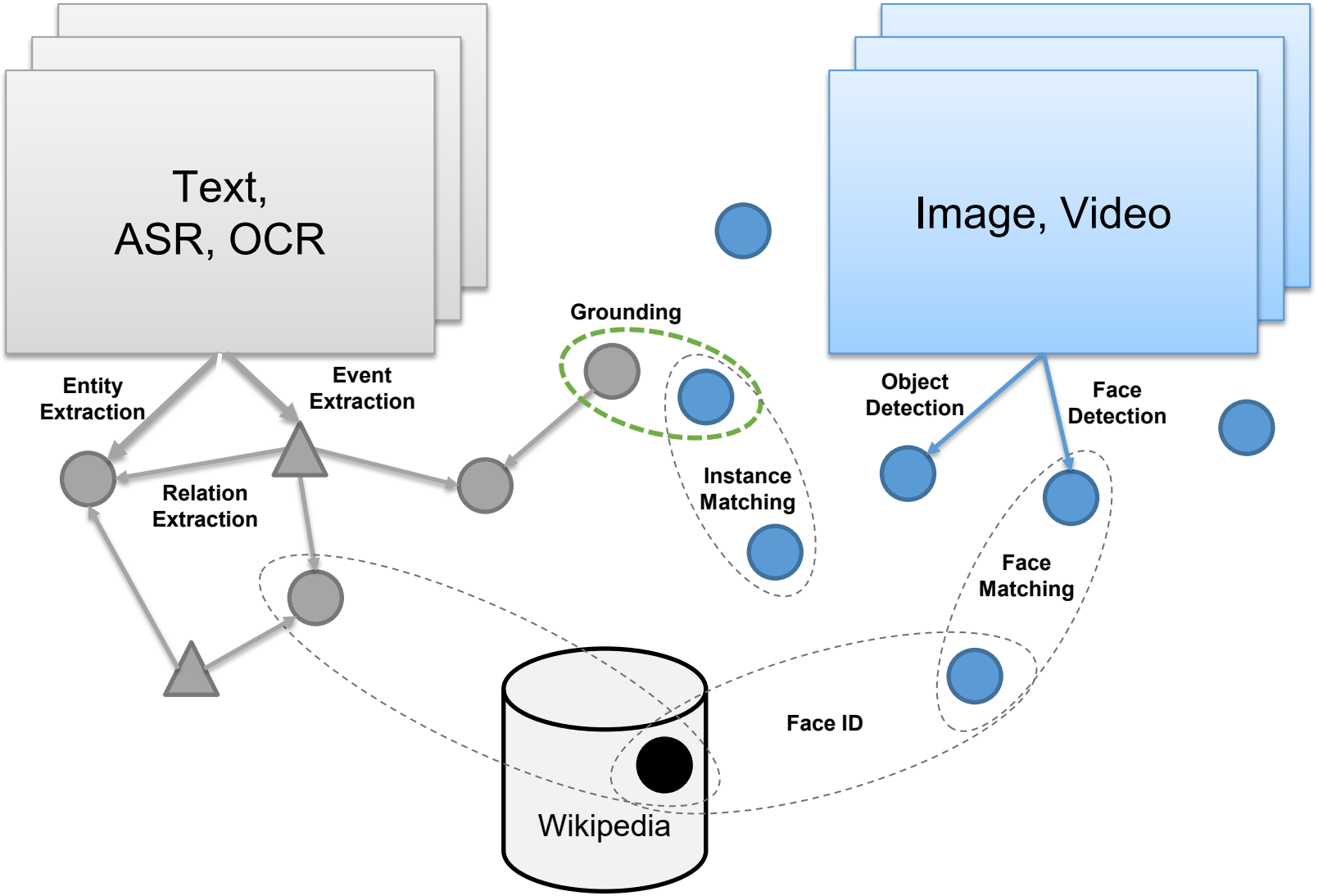
Linking based

Grounding based

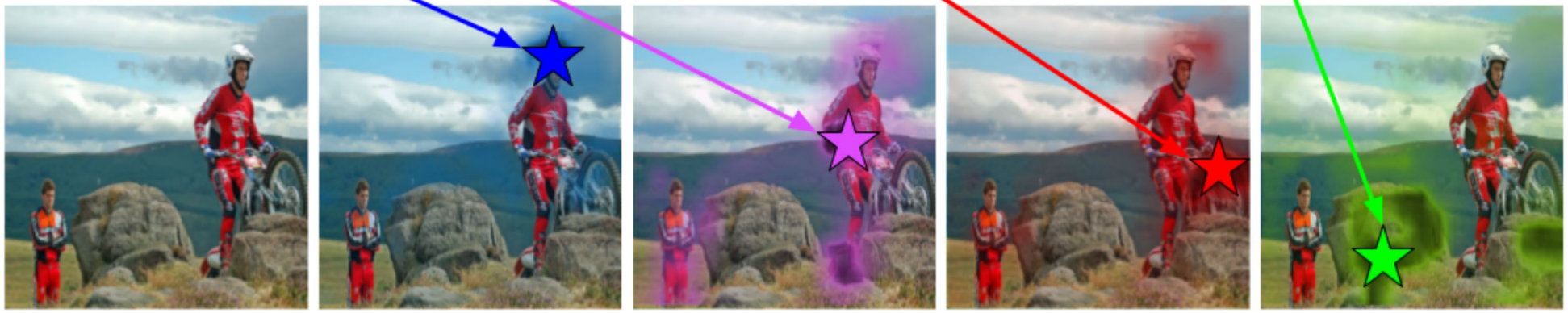
Structure based

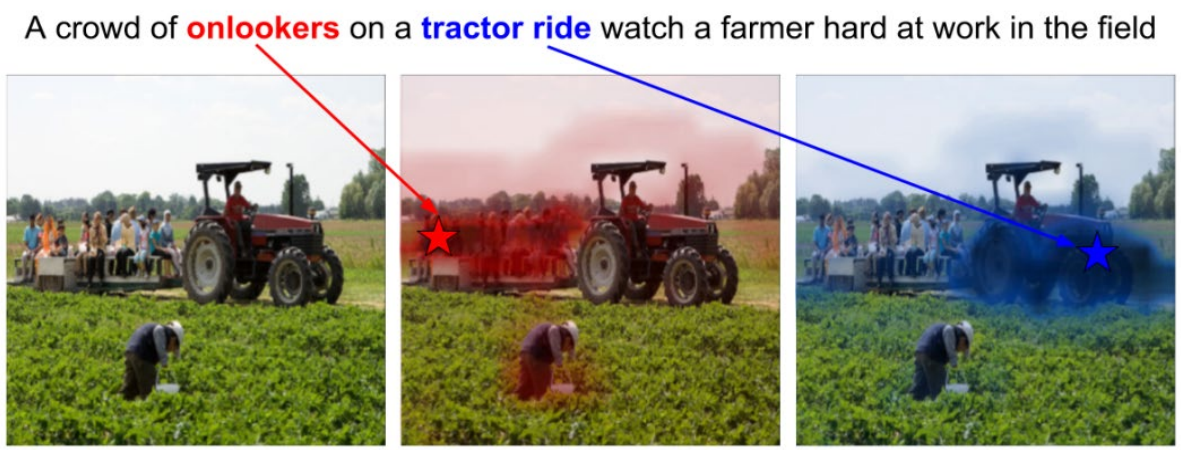


Cross-Media Fusion: Linking-based



Cross-Media Fusion: Grounding-based

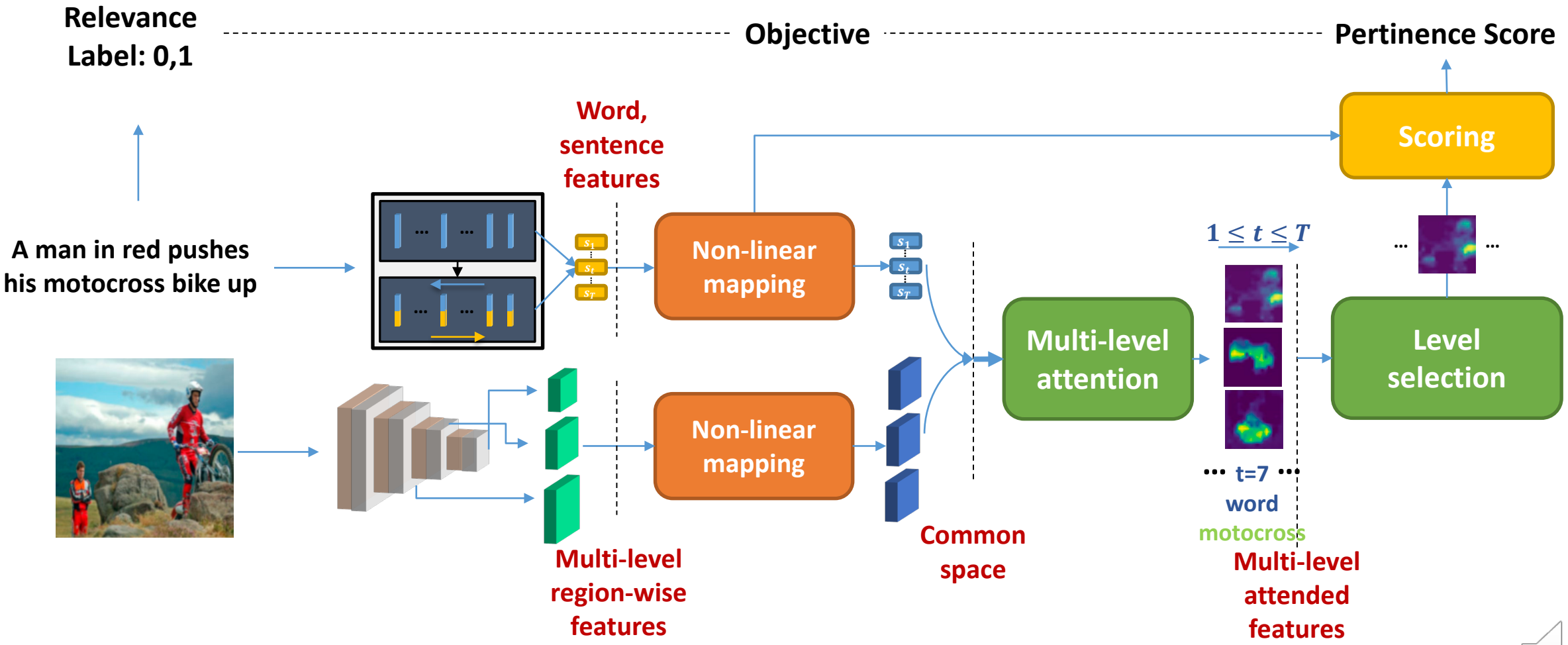




Akbari, Hassan, et al. "Multi-level Multimodal Common Semantic Space for Image-Phrase Grounding." *CVPR* 2019.

Cross-Media Fusion: Grounding-based

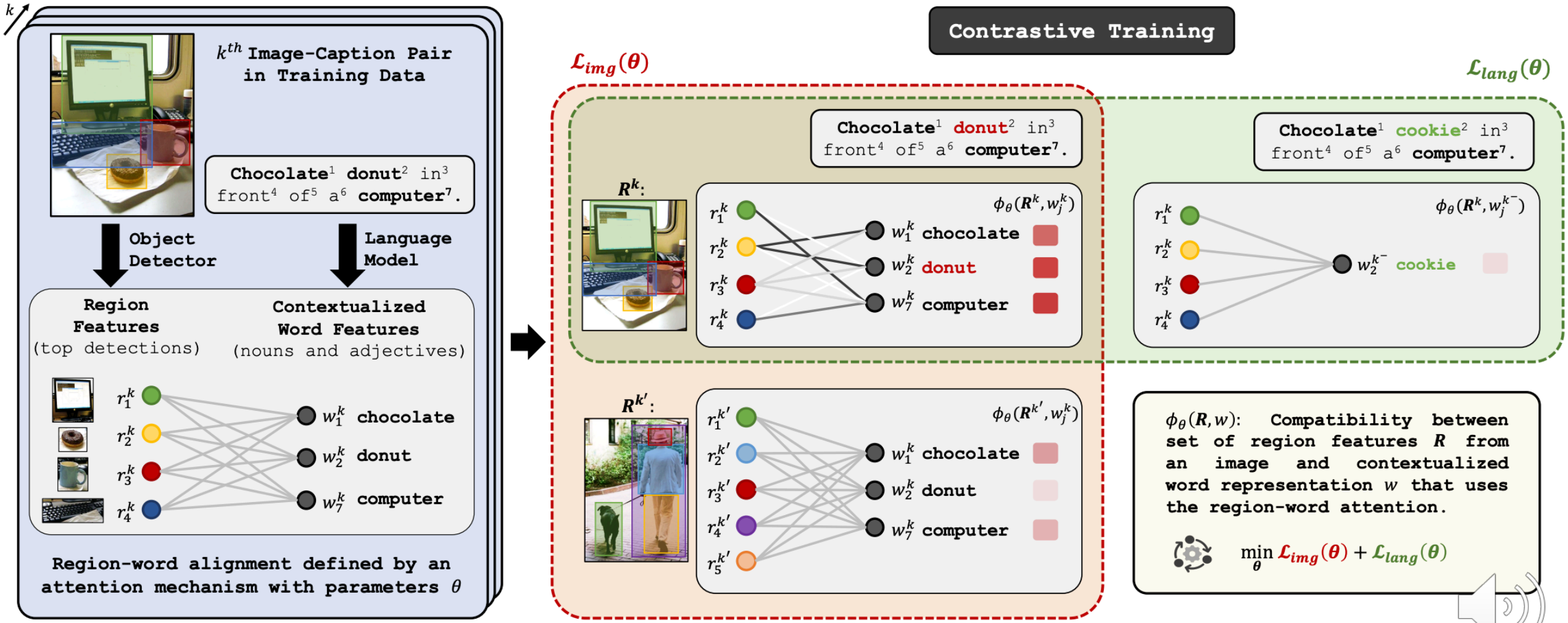
Cross-Attention between Word and Regions:



Akbari, Hassan, et al. "Multi-level Multimodal Common Semantic Space for Image-Phrase Grounding." *CVPR* 2019.

Cross-Media Fusion: Grounding-based

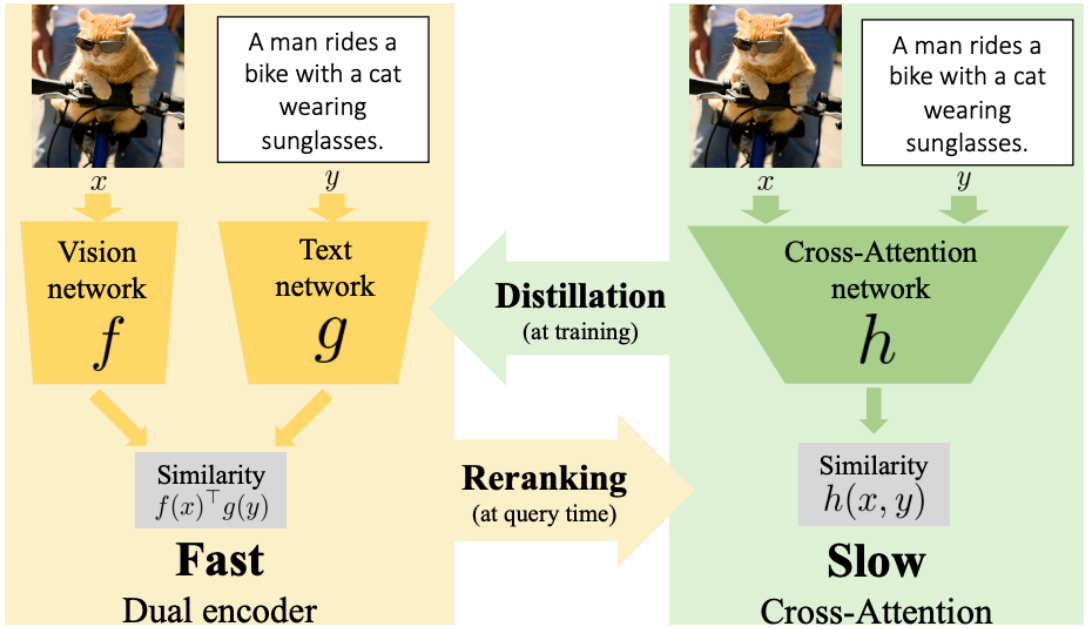
Add additional supervision: Contrastive learning trains a word-region attention mechanism



Gupta, Tanmay, et al. "Contrastive learning for weakly supervised phrase grounding." *ECCV* 2020.

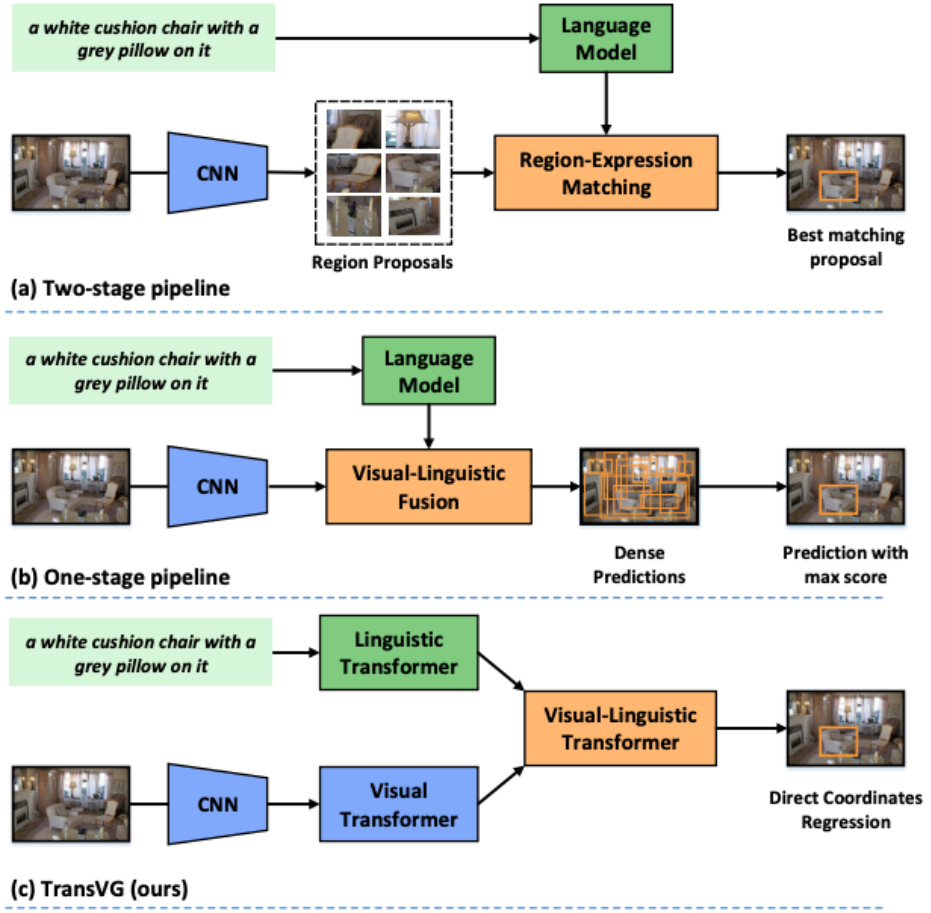
Cross-Media Fusion: Grounding-based

- Cross-attention in Transformers



Fast and Slow

Miech, Antoine, et al. “Thinking fast and slow: Efficient text-to-visual retrieval with transformers.” *CVPR* 2021.



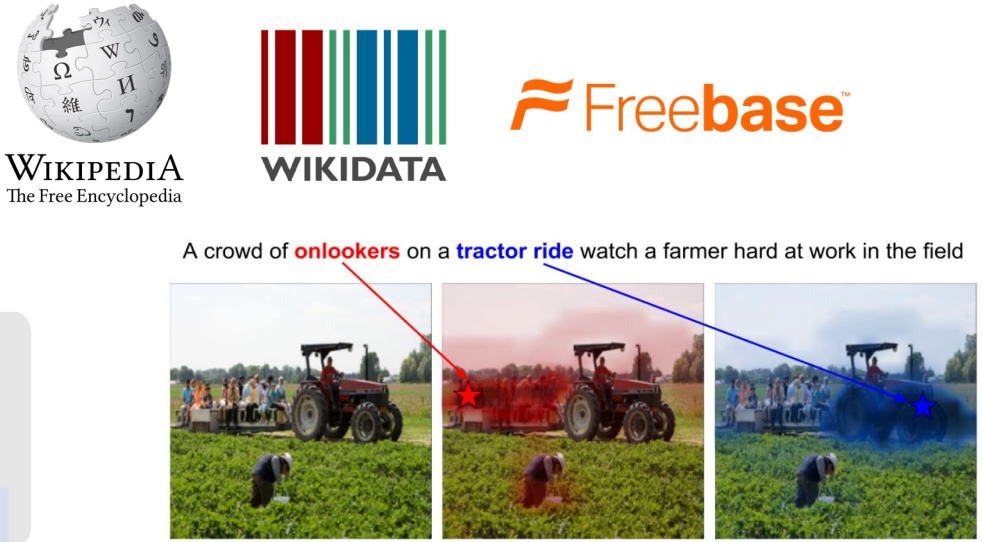
Deng, Jiajun, et al. "Transvg: End-to-end visual grounding with transformers." *CVPR* 2021.

Cross-Media Fusion: Outline

Linking based

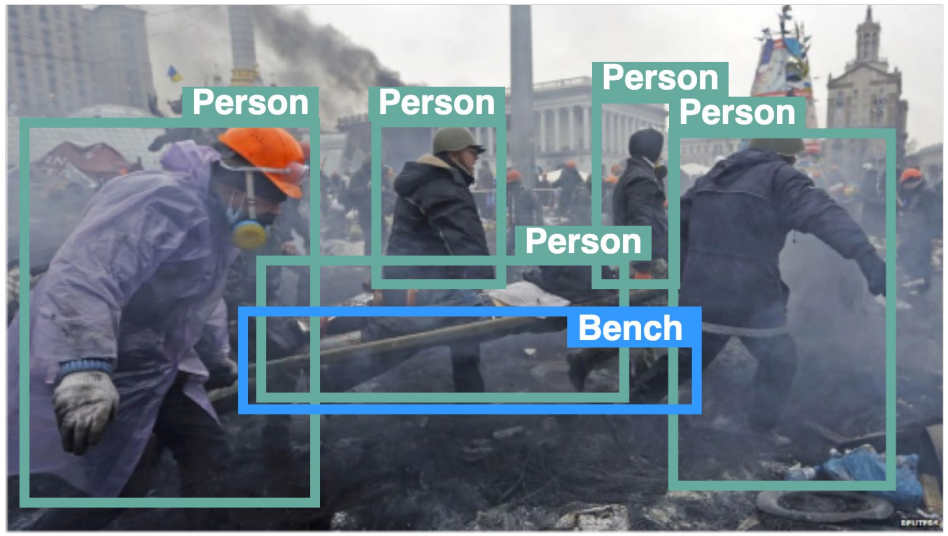
Grounding based

Structure based



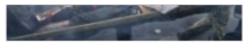
Cross-Media Fusion: Structure-based

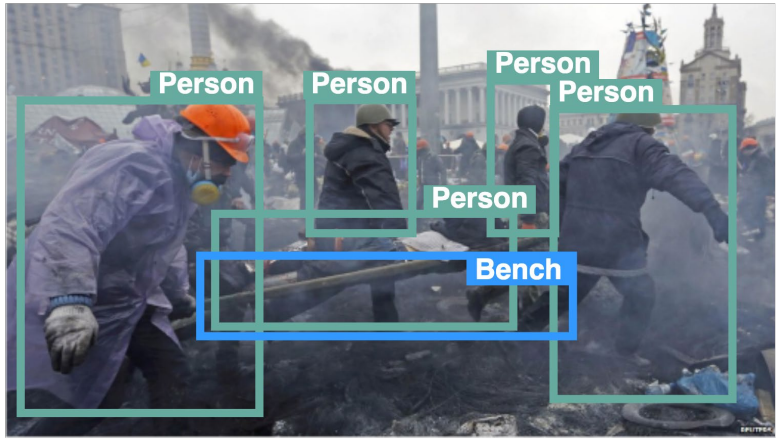
Antigovernment protesters **carry** an injured man on a stretcher after clashes with riot police on Independence Square in Kyiv on February 20, 2014.



Cross-Media Fusion: Structure-based

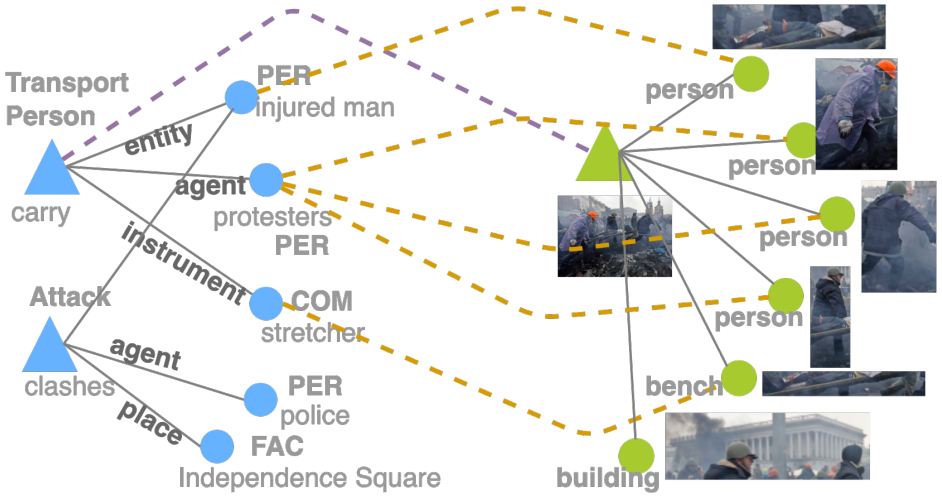
Antigovernment protesters **carry** an injured man on a stretcher after clashes with riot police on Independence Square in Kyiv on February 20, 2014.

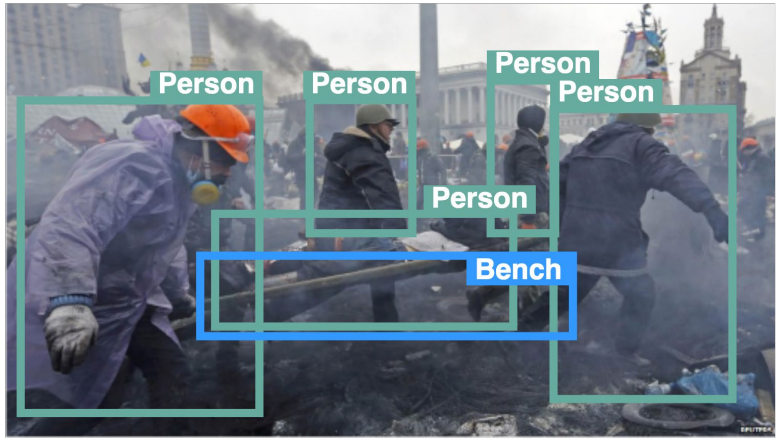
Event	Transport	
Agent	protesters	
Entity	injured man	
Instrument	stretcher	



Cross-Media Fusion: Structure-based

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

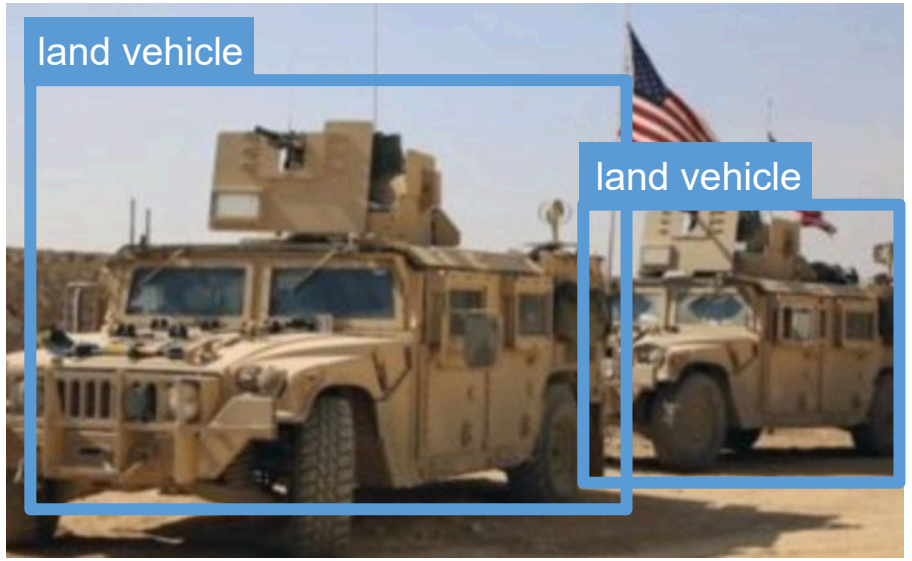




Multimedia Event Extraction (M^2E^2)

Input: News Article Text and Image

Last week , U.S . Secretary of State Rex Tillerson visited Ankara, the first senior administration official to visit Turkey, to try to seal a deal about the battle for Raqqa and to overcome President Recep Tayyip Erdogan's strong objections to Washington's backing of the Kurdish Democratic Union Party (PYD) militias. Turkish forces have attacked SDF forces in the past around Manbij, west of Raqqa, forcing the **United States** to **deploy** dozens of **soldiers** on the **outskirts** of the town in a mission to prevent a repeat of clashes, which risk derailing an assault on Raqqa.



Output: Events & Argument Roles

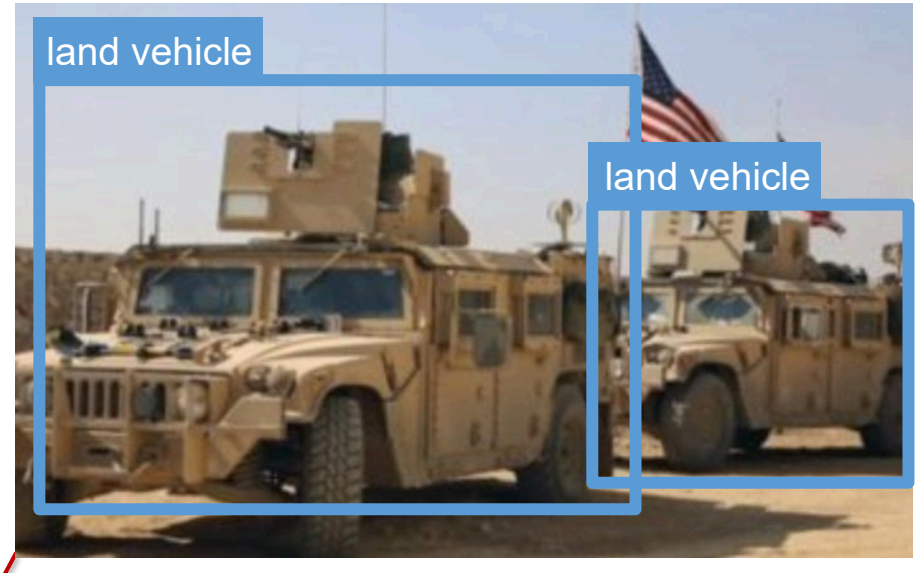
Event Type	Movement.Transport	
Event	**Text Trigger**	deploy
	Image	

Arguments	**Agent**	United States
	Destination	outskirts
	Artifact	soldiers
	Vehicle	
	Vehicle	

Multimedia Event Extraction (M^2E^2)

Input: News Article Text and Image

Last week , U.S . Secretary of State Rex Tillerson visited Ankara, the first senior administration official to visit Turkey, to try to seal a deal about the battle for Raqqa and to overcome President Recep Tayyip Erdogan's strong objections to Washington's backing of the Kurdish Democratic Union Party (PYD) militias. Turkish forces have attacked SDF forces in the past around Manbij, west of Raqqa, forcing the **United States** to **deploy** dozens of **soldiers** on the **outskirts** of the town in a mission to prevent a repeat of clashes, which risk derailing an assault on Raqqa.



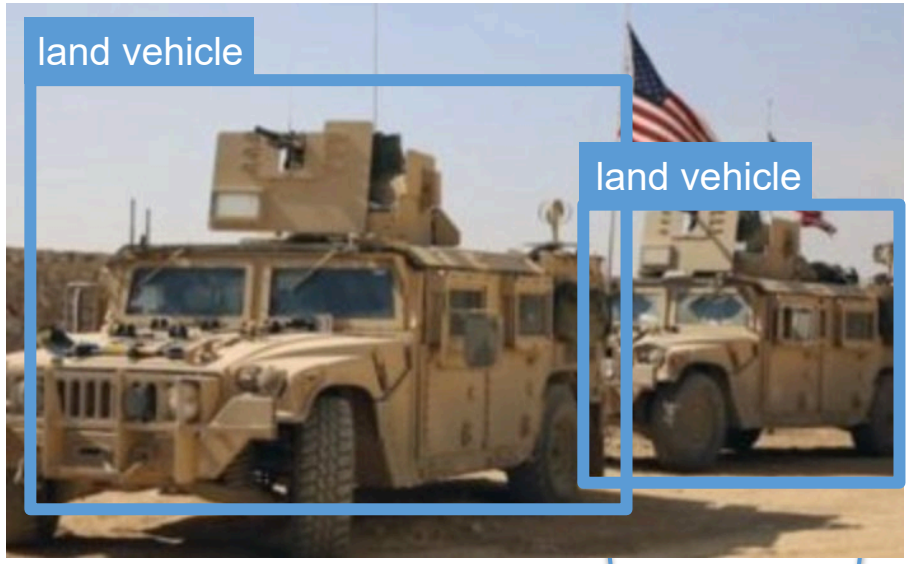
Output: Multimedia Events & Argument Roles

Event Type	Movement.Transport			
Event	**Text Trigger**	deploy	**Arguments**	**Agent** United States
				Destination outskirts
				Artifact soldiers
	Image			**Vehicle** 
				Vehicle 

Multimedia Event Extraction (M^2E^2)

Input: News Article Text and Image

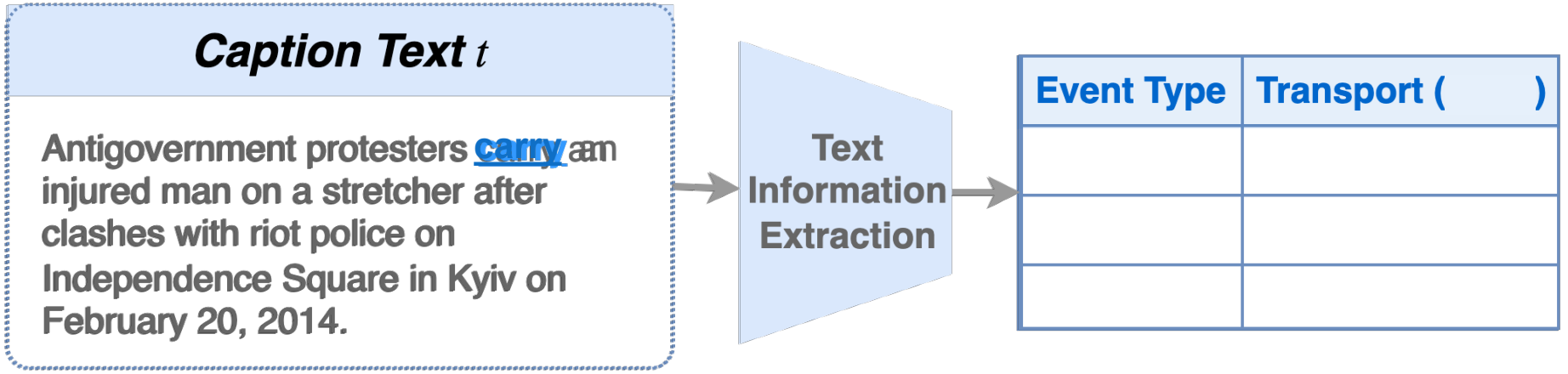
Last week , U.S . Secretary of State Rex Tillerson visited Ankara, the first senior administration official to visit Turkey, to try to seal a deal about the battle for Raqqa and to overcome President Recep Tayyip Erdogan's strong objections to Washington's backing of the Kurdish Democratic Union Party (PYD) militias. Turkish forces have attacked SDF forces in the past around Manbij, west of Raqqa, forcing the **United States** to **deploy** dozens of **soldiers** on the **outskirts** of the town in a mission to prevent a repeat of clashes, which risk derailing an assault on Raqqa.



Output: Multimedia Events & Argument Roles

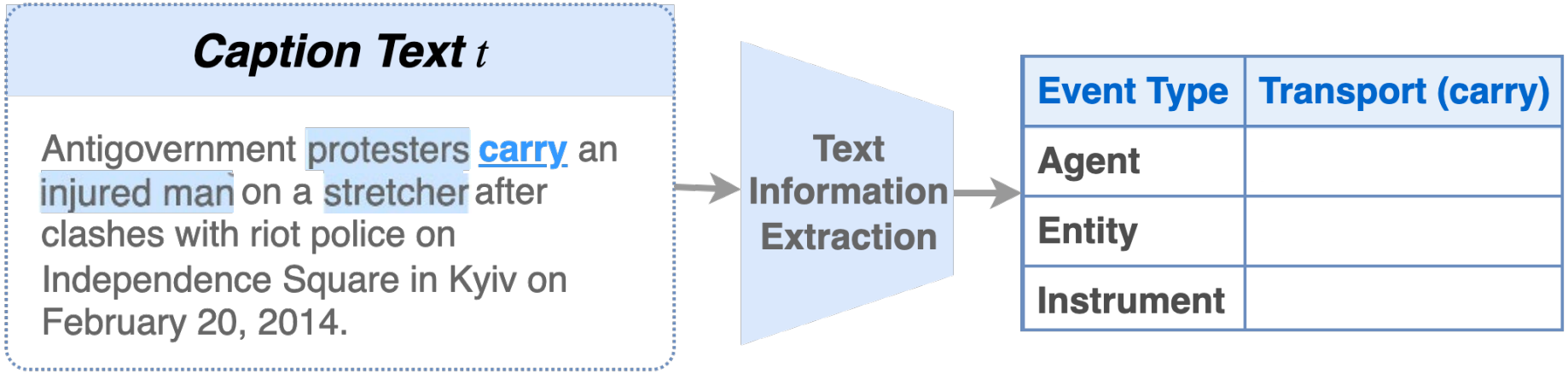
Event Type	Movement.Transport			
Event	**Text Trigger**	deploy	**Arguments**	**Agent** United States
	Image			**Destination** outskirts
				Artifact soldiers
				Vehicle 
				Vehicle 

Transfer Structured Knowledge Across Modalities



Natural Language Processing | Event Extraction { Trigger Word

Transfer Structured Knowledge Across Modalities



Natural Language Processing | Event Extraction
- Trigger Word
- Argument (Participant)

Transfer Structured Knowledge Across Modalities

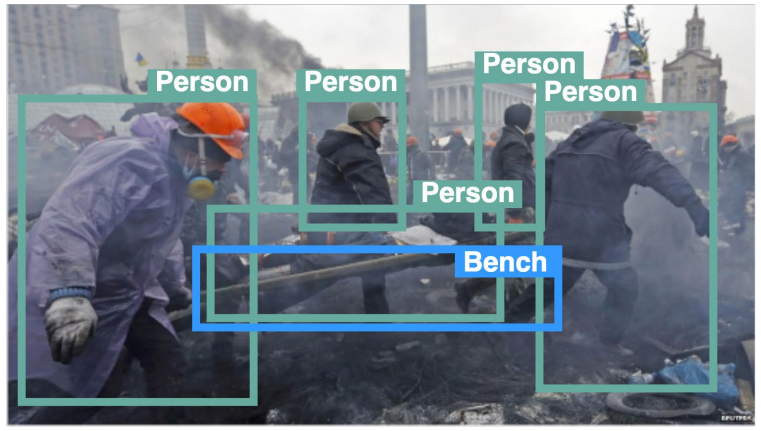
Caption Text t

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 (carry)
Agent	protesters
Entity	injured man
Instrument	stretcher

Image i



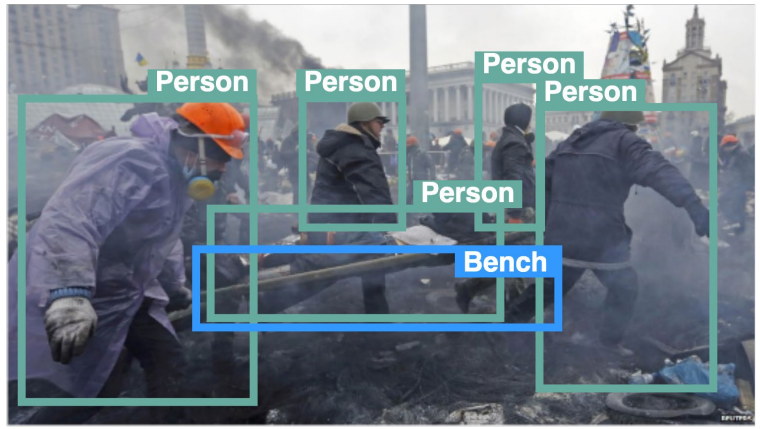
Weakly Supervision

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.

Image i



Positive Labels

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

Negative Labels (events)

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

Negative Labels (arguments)

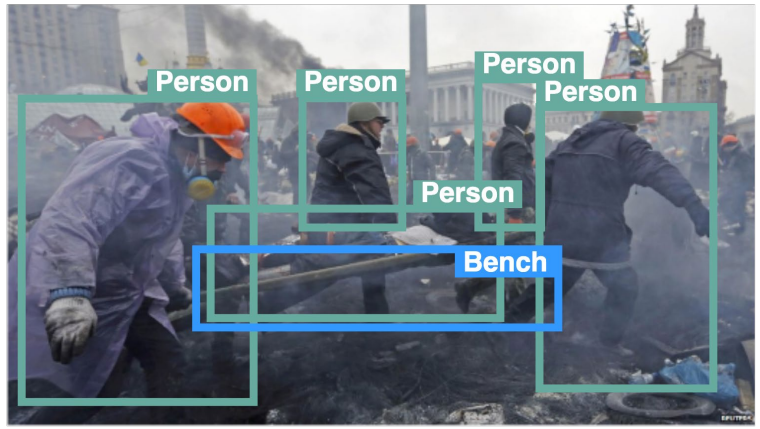
Event Type	Transport (carry)
Agent	protesters
Entity	injured man
Instrument	stretcher

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

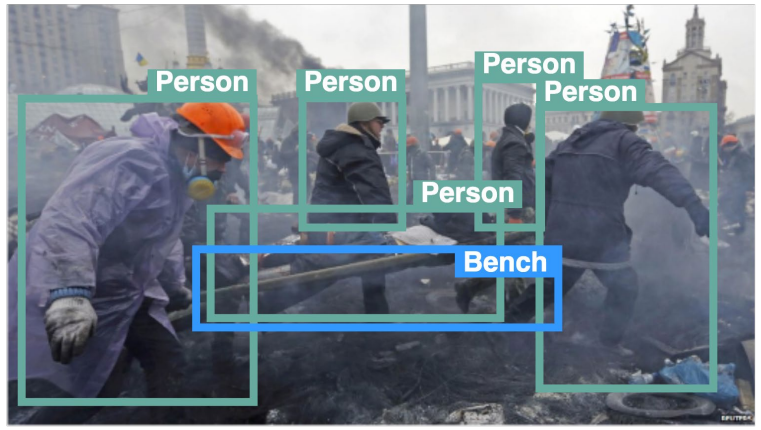
prompt → Injured man transported a stretcher with protesters.

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.

Image i



Positive Labels

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

Protesters transported injured man using a stretcher.

Negative Labels (events)

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

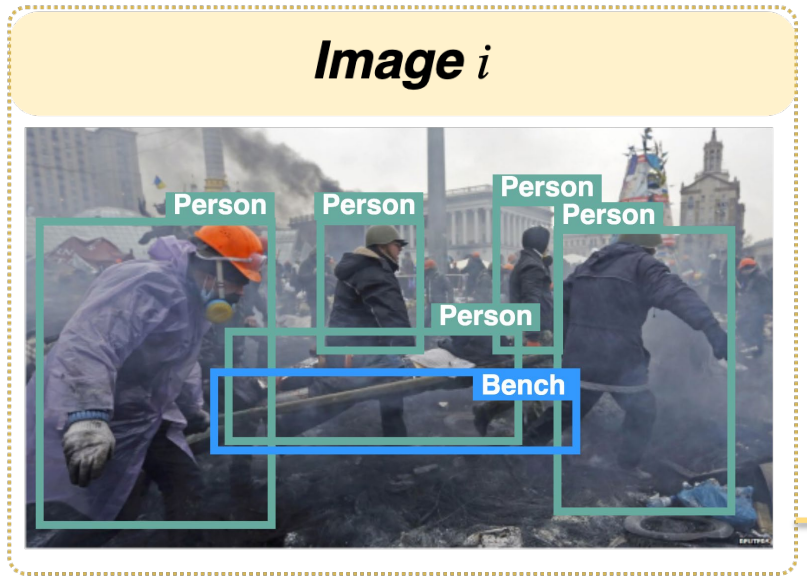
Protesters arrested injured man using a stretcher.

Negative Labels (arguments)

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

Injured man transported a stretcher with protesters.

Construct hard negatives by manipulating event structures.



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

t_0 t_1 t_2

$s(t_0, v)$ $s(t_1, v)$ $s(t_2, v)$

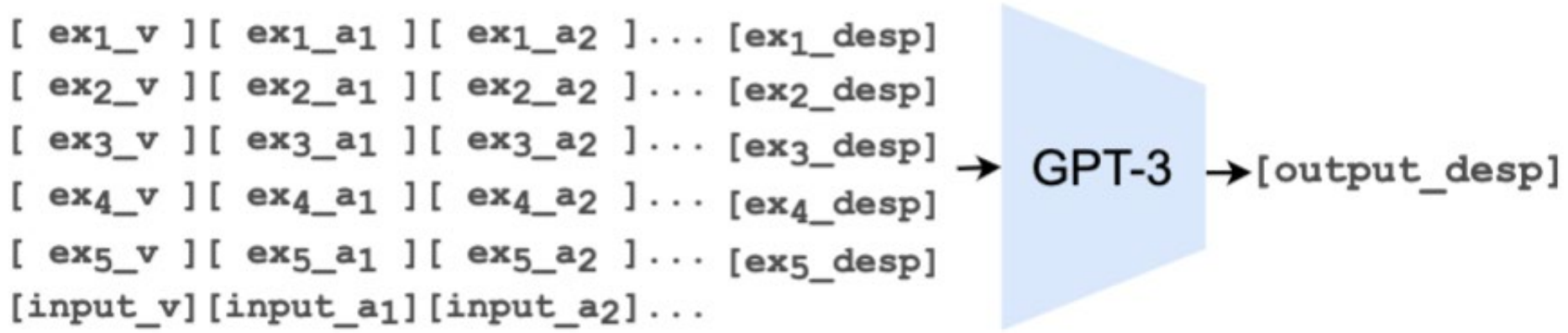
Contrastive Learning

Image Encoder

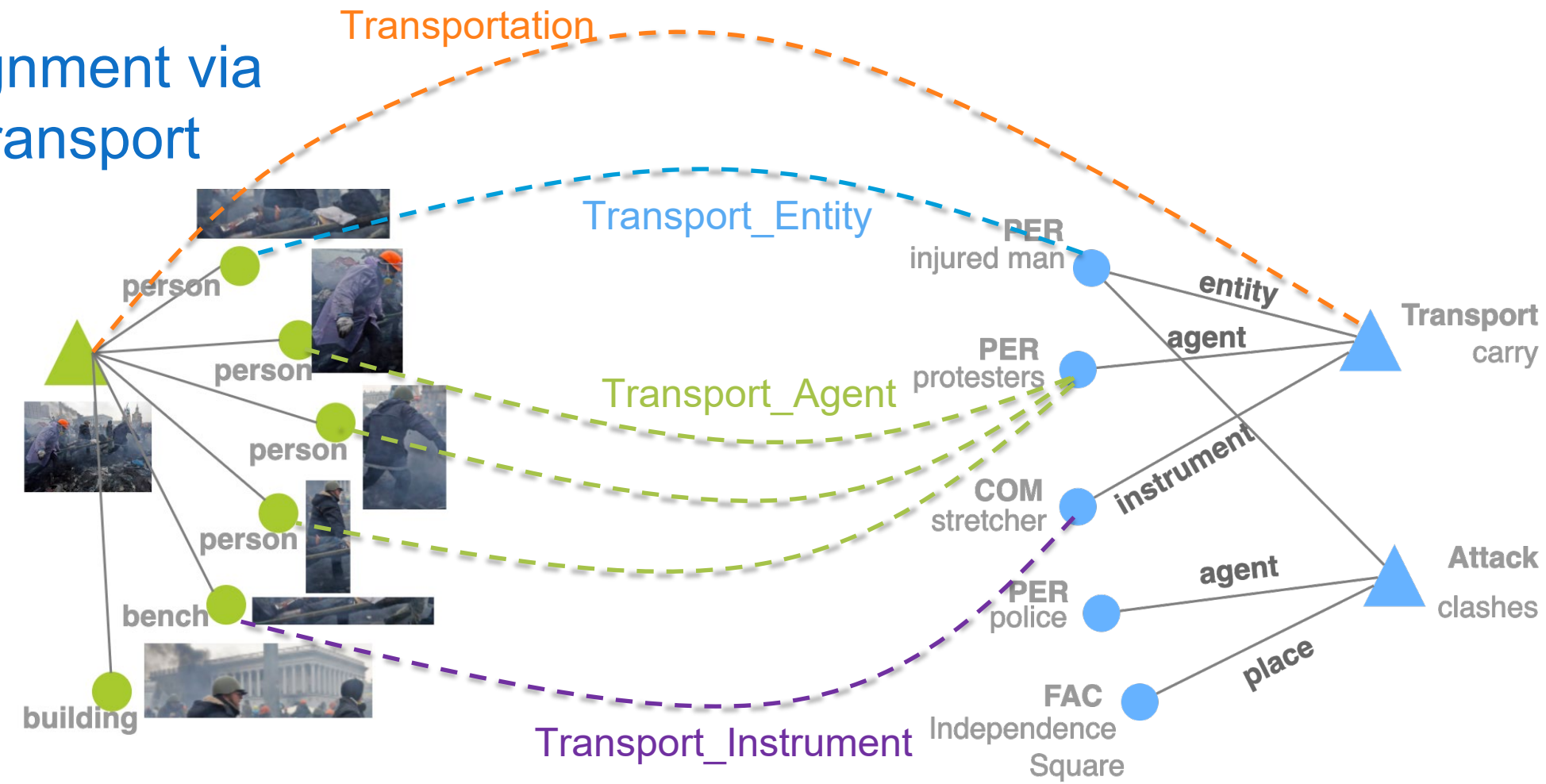
v

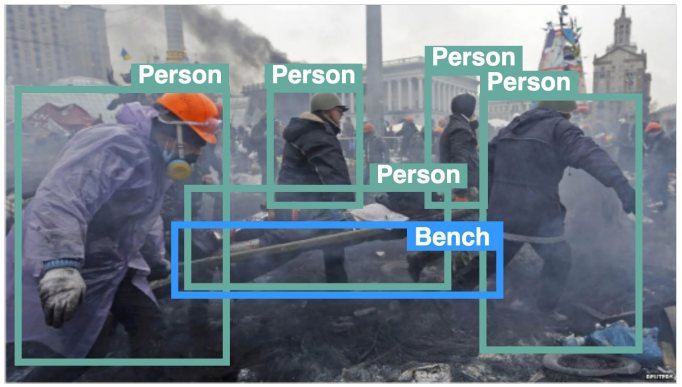
GPT-3 Based Prompt

- From Event Structure to Natural Language Description:
  - We use five manual event description examples as few-shot prompts to control the generation.



Graph Alignment via Optimal Transport

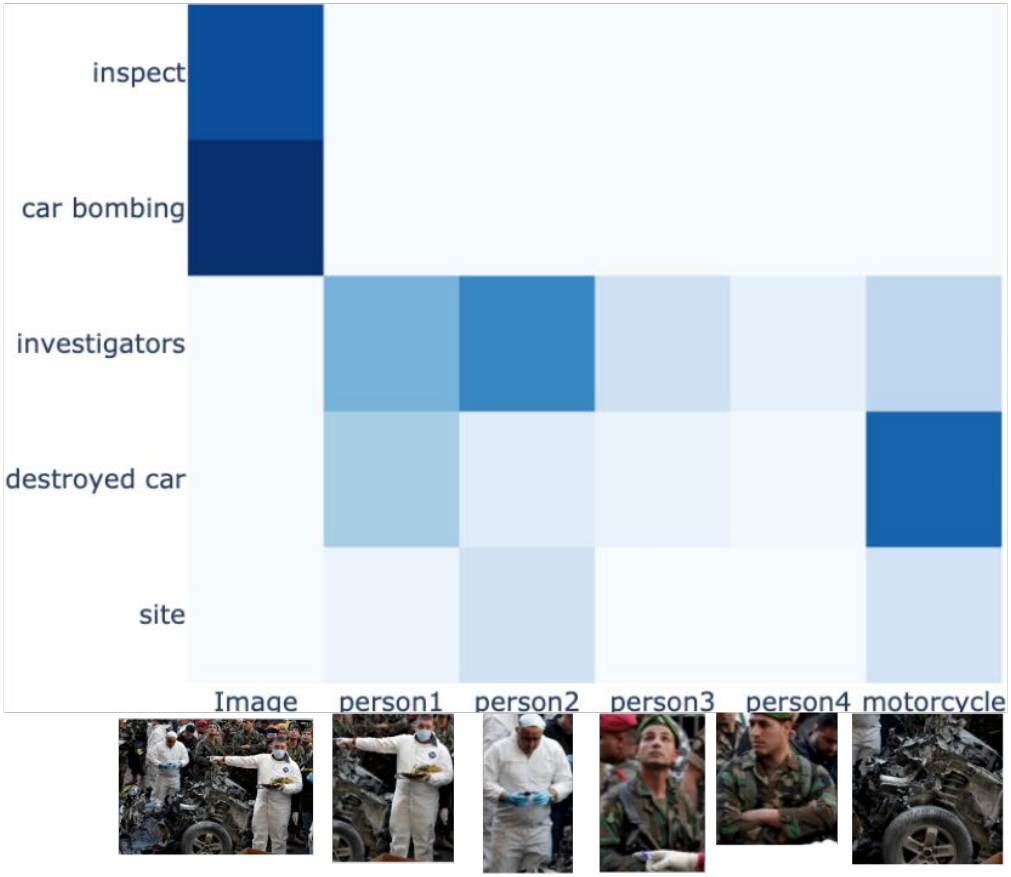




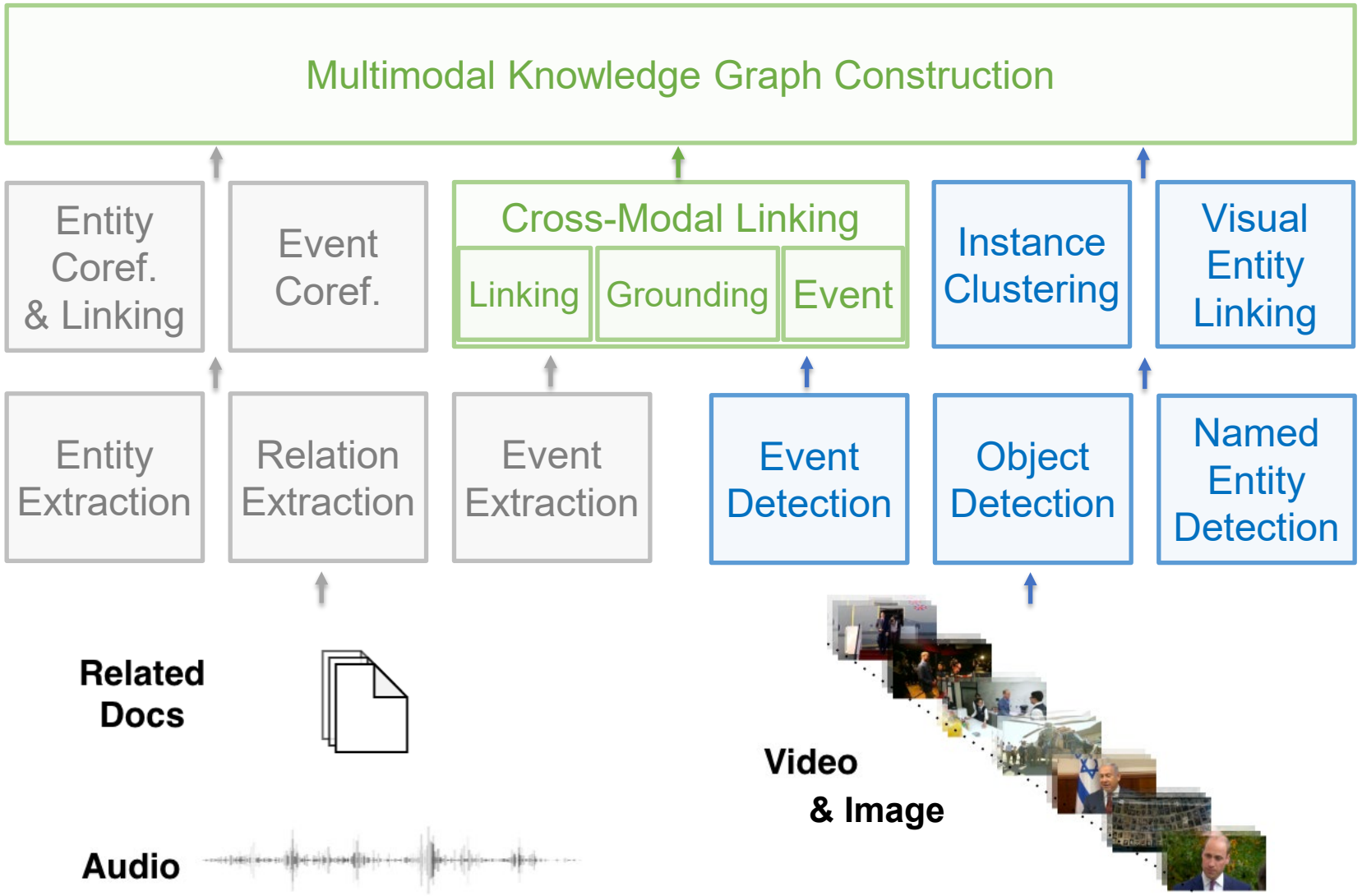
Antigovernment protesters carry an injured man on a stretcher after clashes with riot police on Independence Square in Kyiv on February 20, 2014.

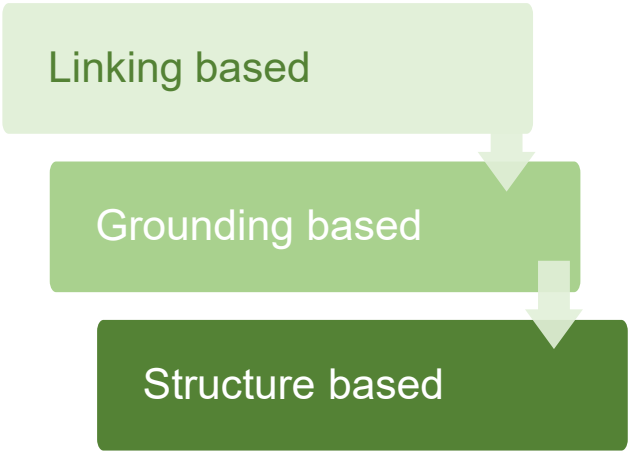
Event-level Cross-media Alignment

Investigators **inspect** parts of a destroyed car at the site of a car bombing in Beirut, Jan. 21, 2014.

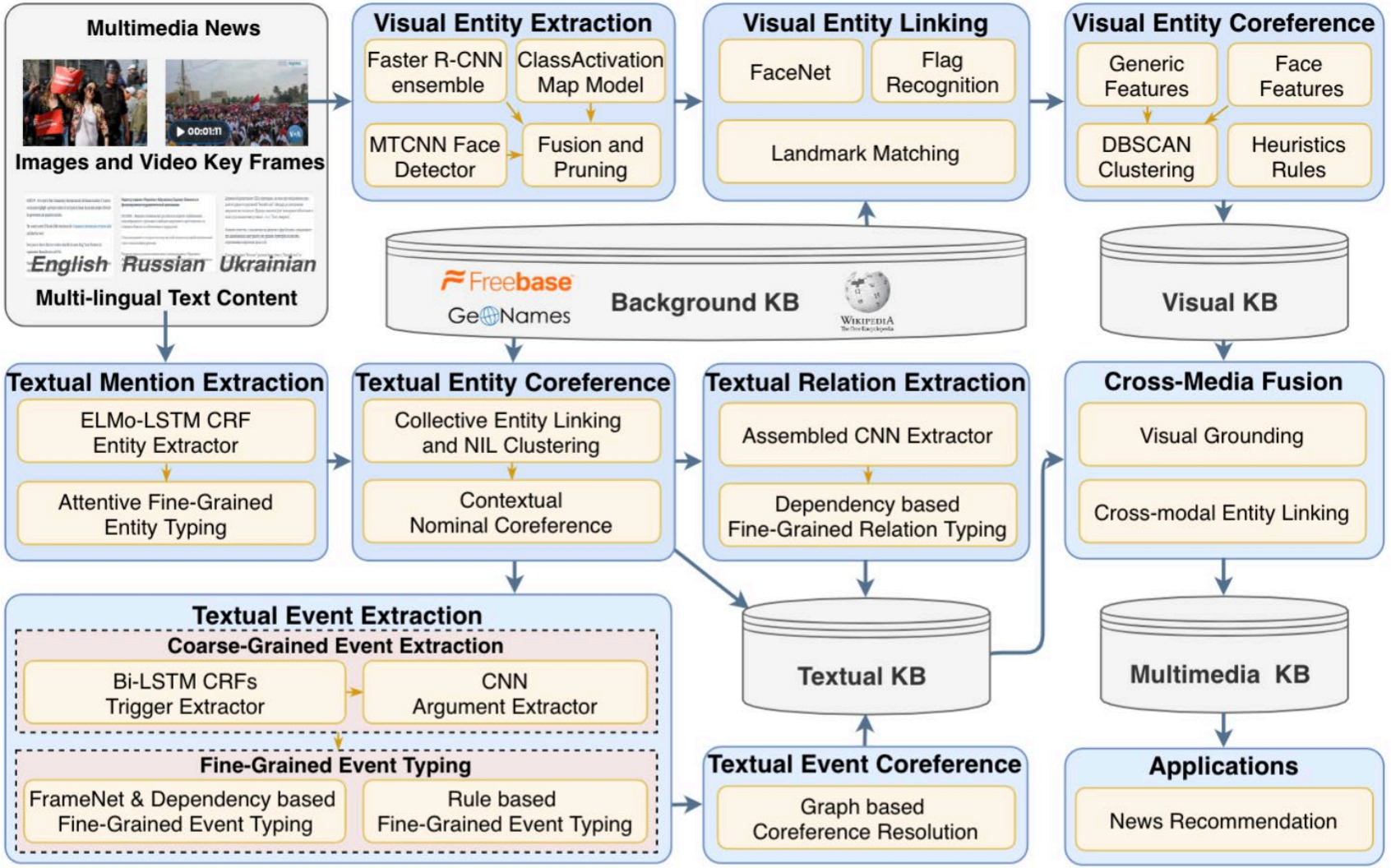


Conclusions





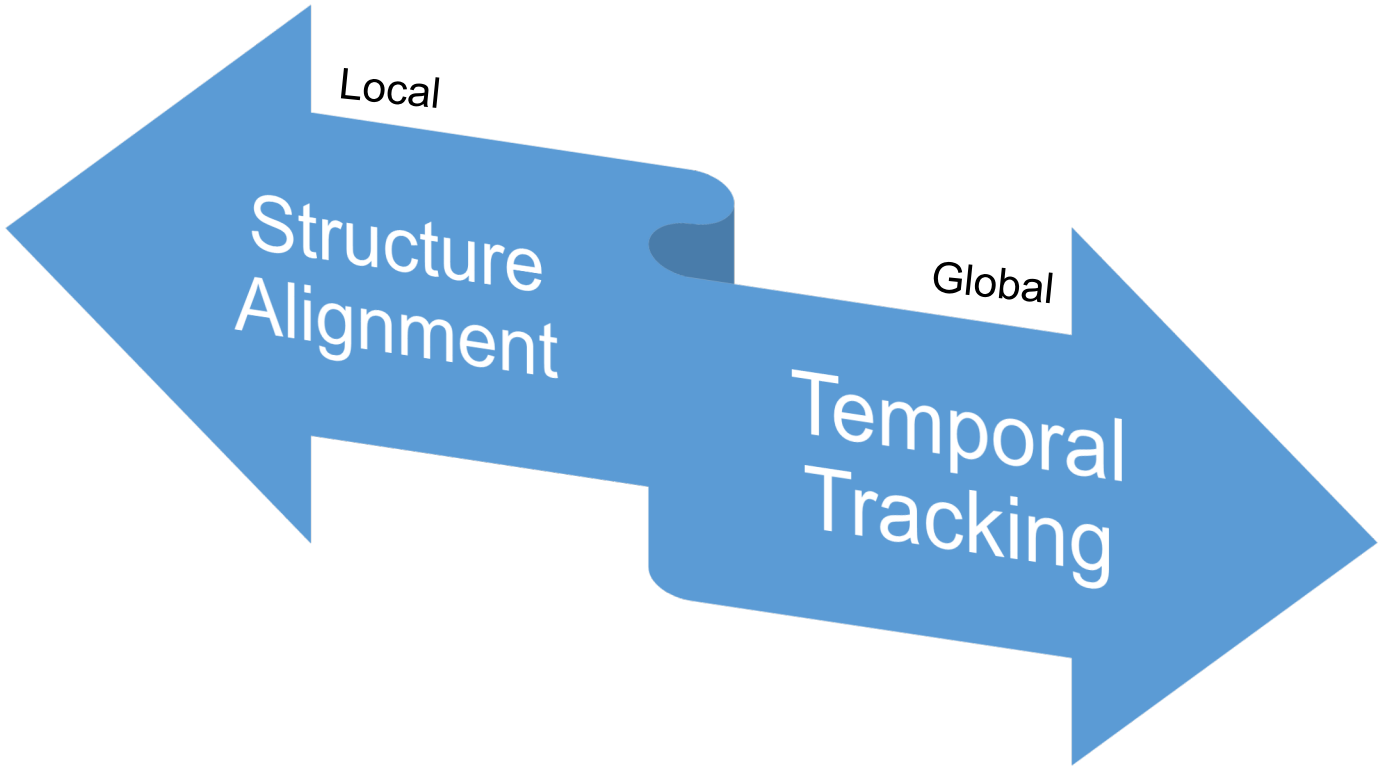
GAIA: Open-source Multimedia Knowledge Extraction System



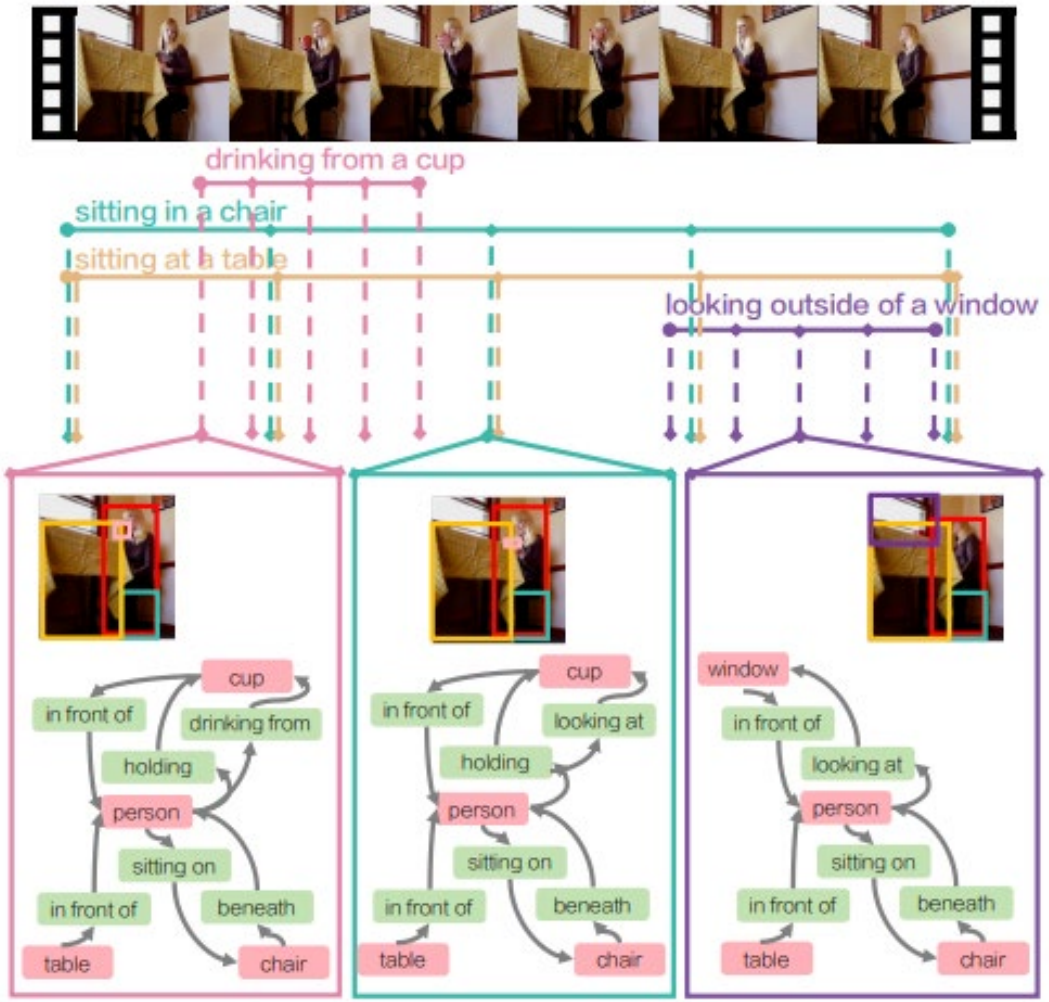
- Available at http://blender.cs.illinois.edu/software/gaia-ie
- Achieved best performance at TAC SM-KBP 2019 and 2020 Evaluation (10% higher than the second ranked team in TAC SM-KBP 2019)

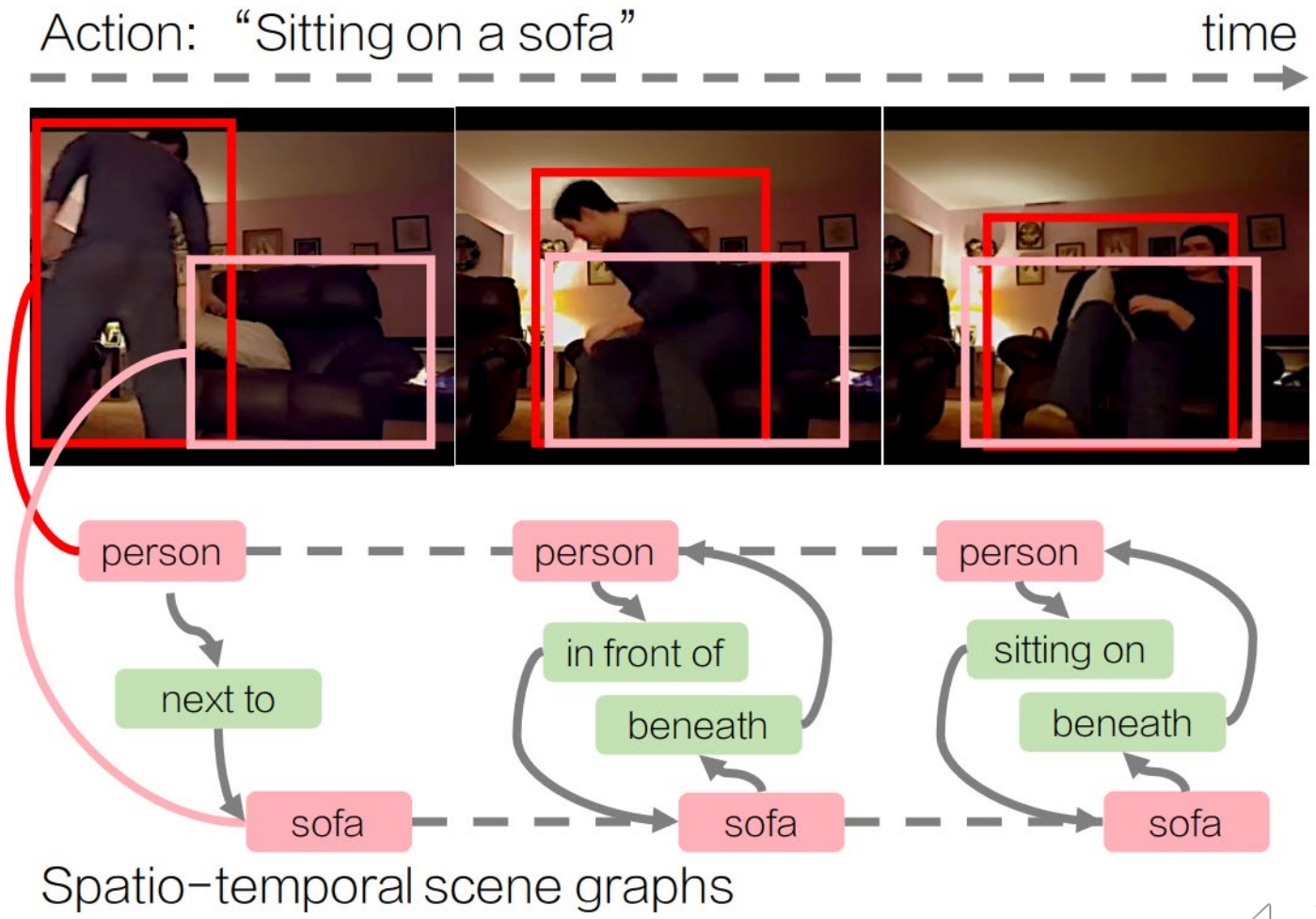
Future Challenges

- Local: Capturing semantic structure
- Global: Understanding a more global context of multiple entities and events



Future Direction: Event Tracking





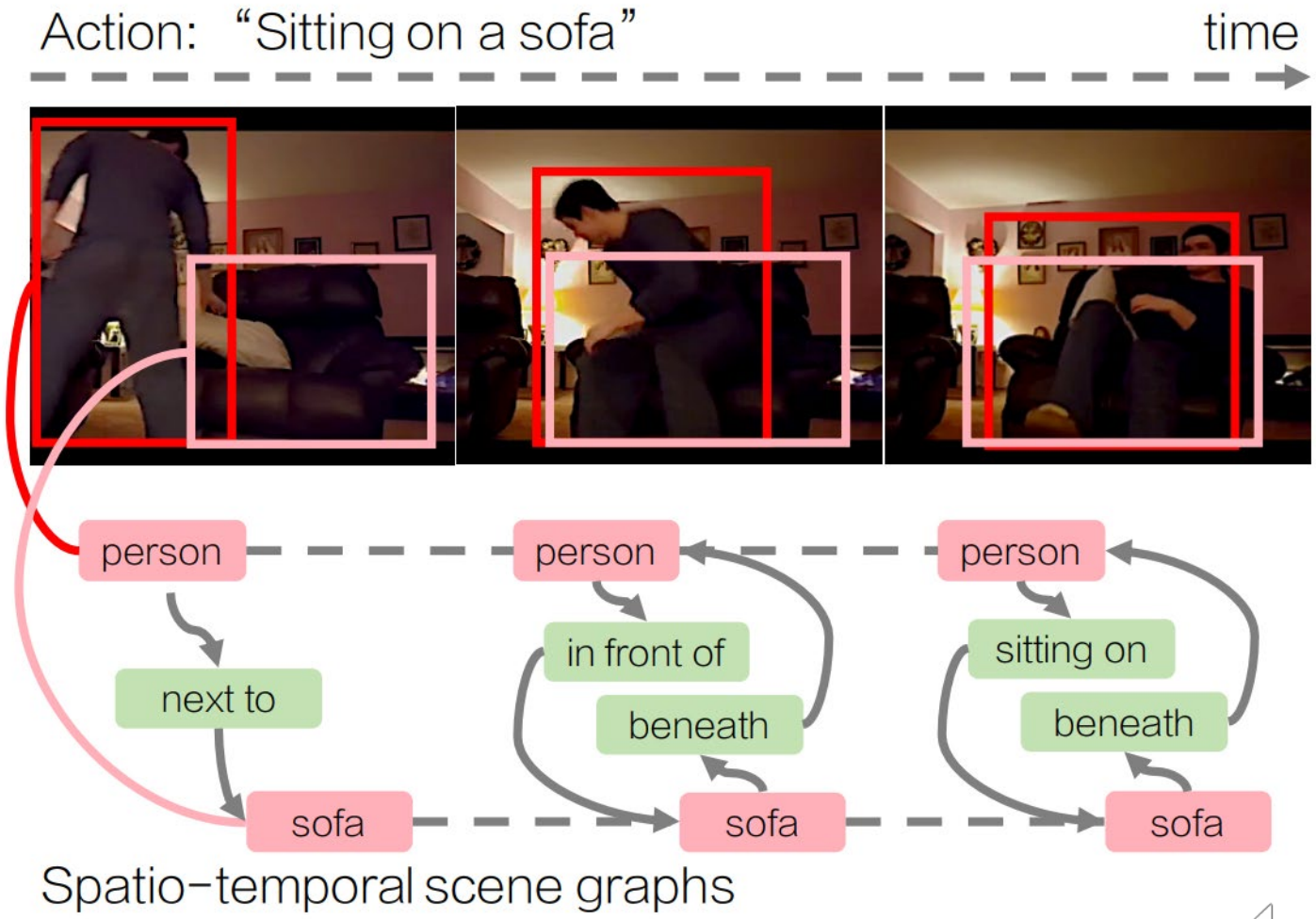
[Ji el al, 2019]

Future Direction: Text and Vison are Complementary

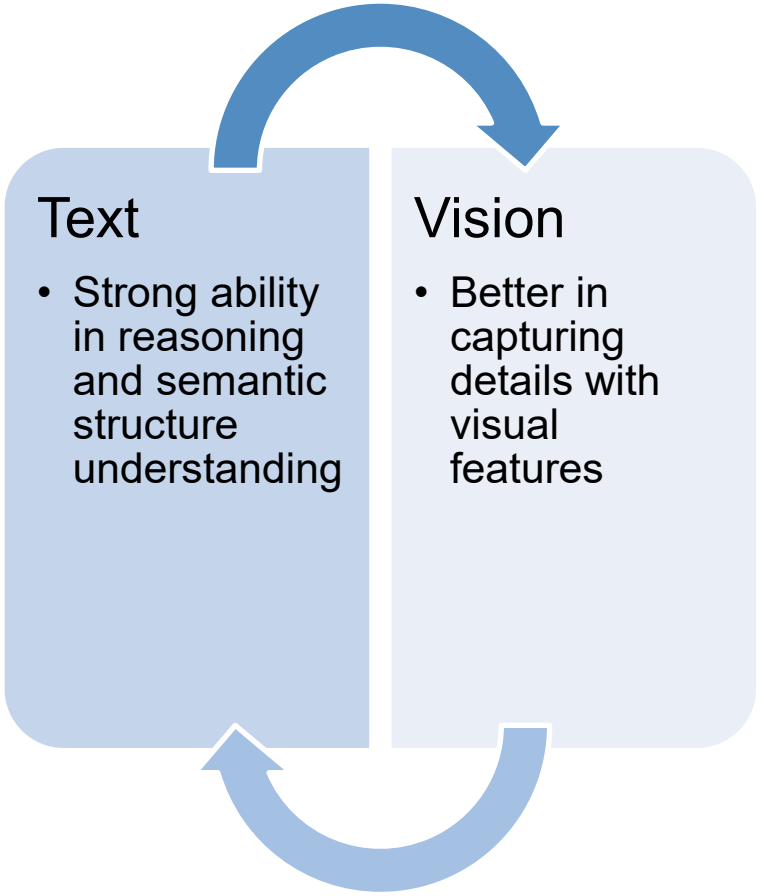
Text

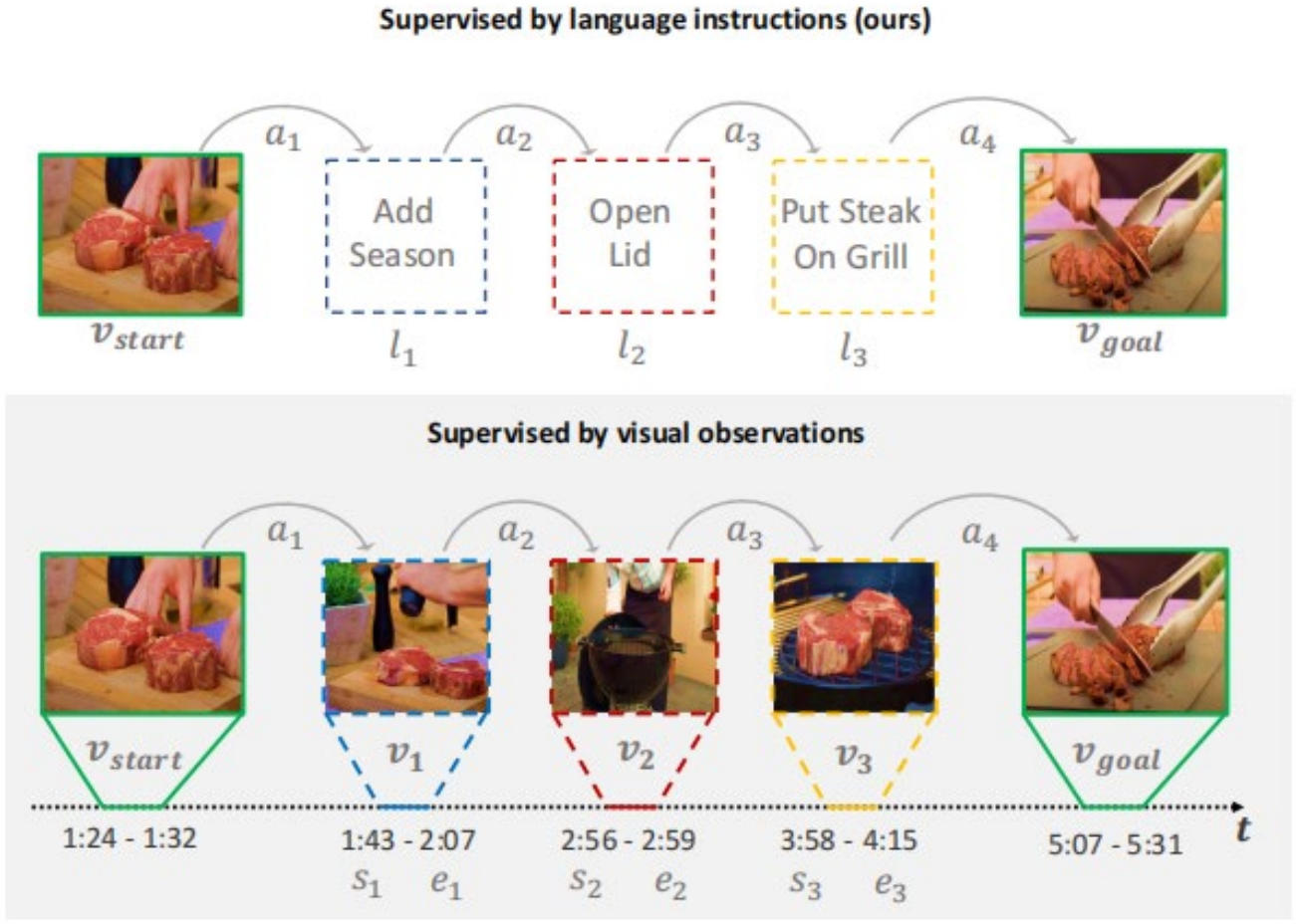
Vision

- Better in capturing details with visual features



Future Direction: Text and Vison are Complementary





P3IV: Probabilistic Procedure Planning from Instructional Videoswith Weak Supervision

Thank You