Multimodal Knowledge Graphs

Generation Methods, Applications, and Challenges

Shih-Fu Chang

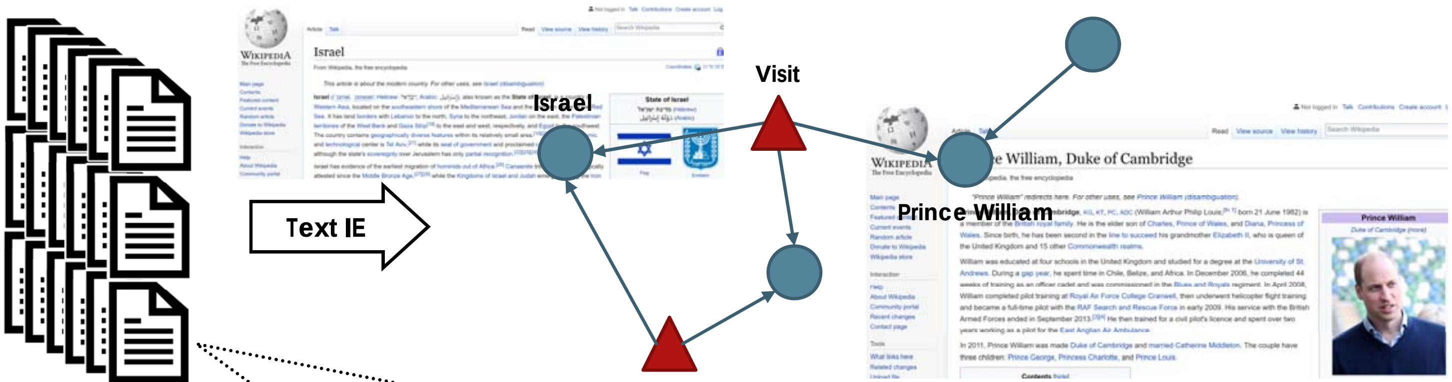
Alireza Zareian, Hassan Akbari, Brian Chen, Svebor Karaman,
Zhecan James Wang, and Haoxuan You
Columbia University

Prof. Heng Ji,
Manling Li, Di Lu, and Spencer Whitehead
University of Illinois, Urbana-Champaign

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Knowledge Graphs

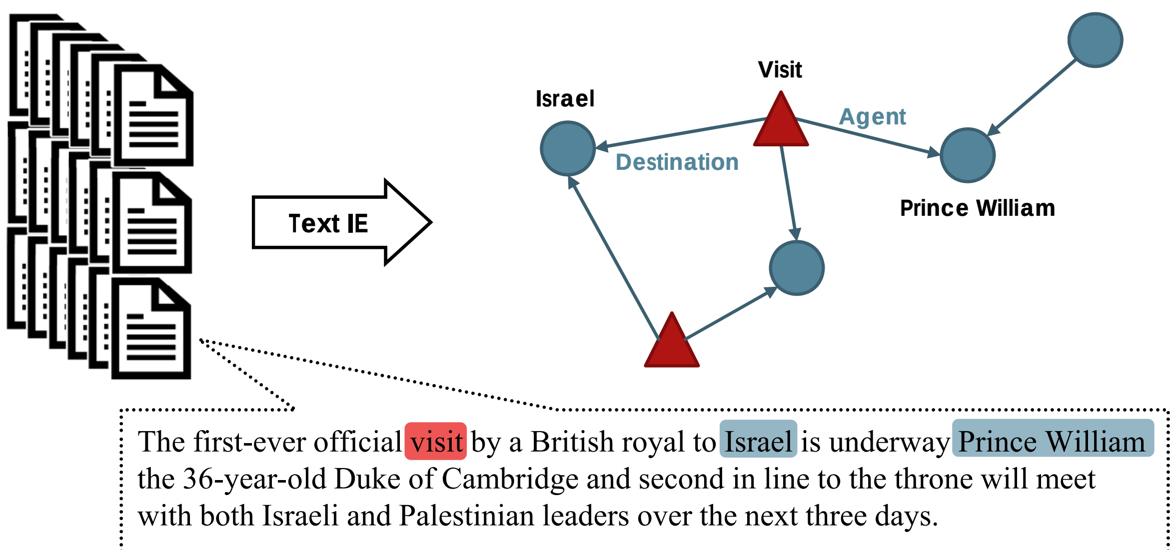
- Entities, events, relations, etc.



The first-ever official visit by a British royal to Israel is underway. Prince William the 36-year-old Duke of Cambridge and second in line to the throne will meet with both Israeli and Palestinian leaders over the next three days.

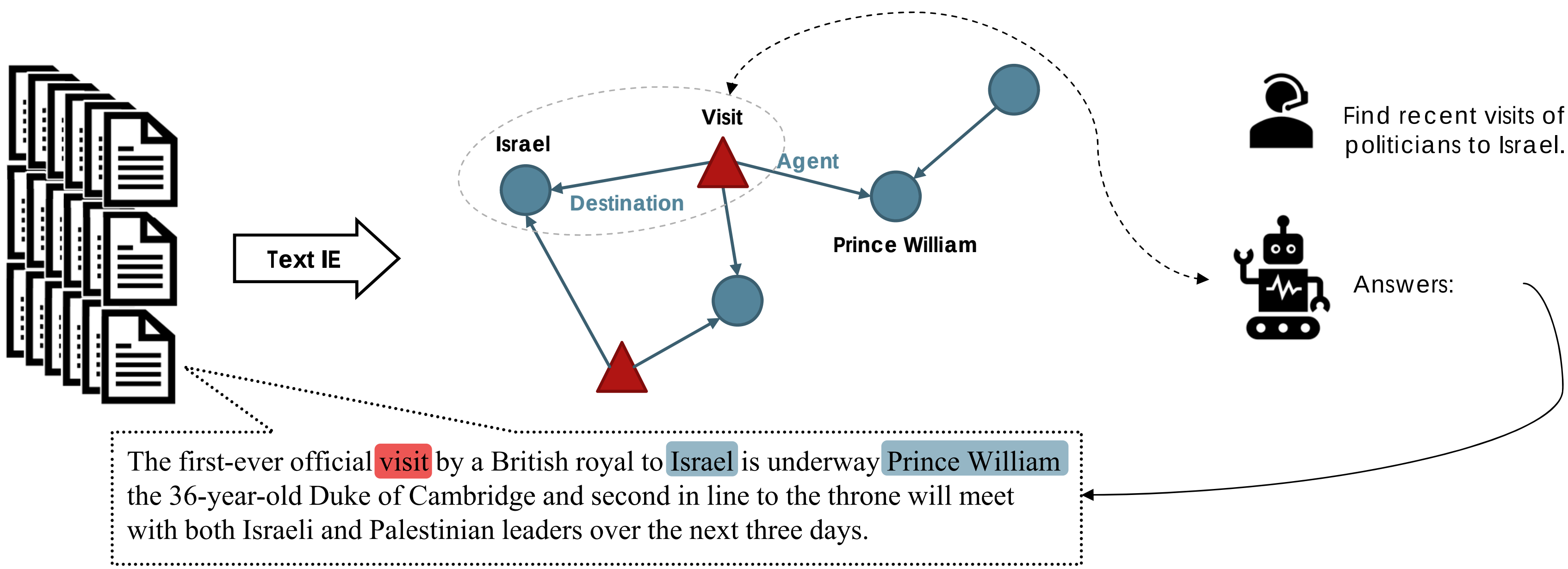
Knowledge Graphs

- Entities, events, relations, etc.
- Events describe what happens
  - Entities are characterized by the argument *role* they play in events

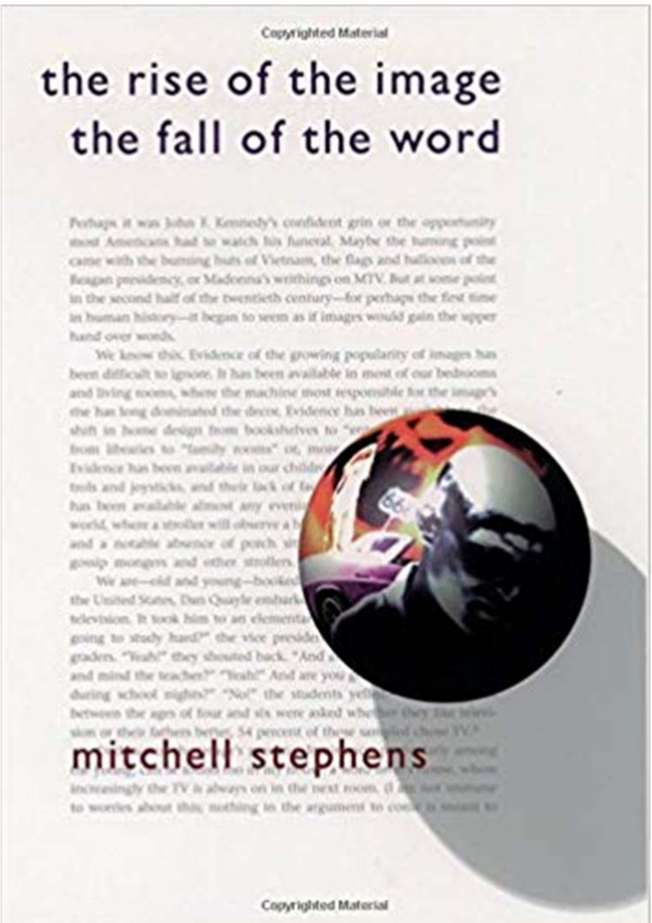


Knowledge Graphs

- Application: Question Answering, Reasoning, Hypothesis Verification and Discovery



Knowledge Beyond Text



- We communicate through **multi**media

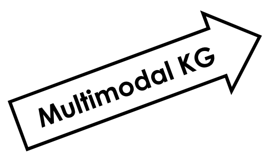
- Our experiment shows 34% of news images contain event arguments that are not mentioned in text

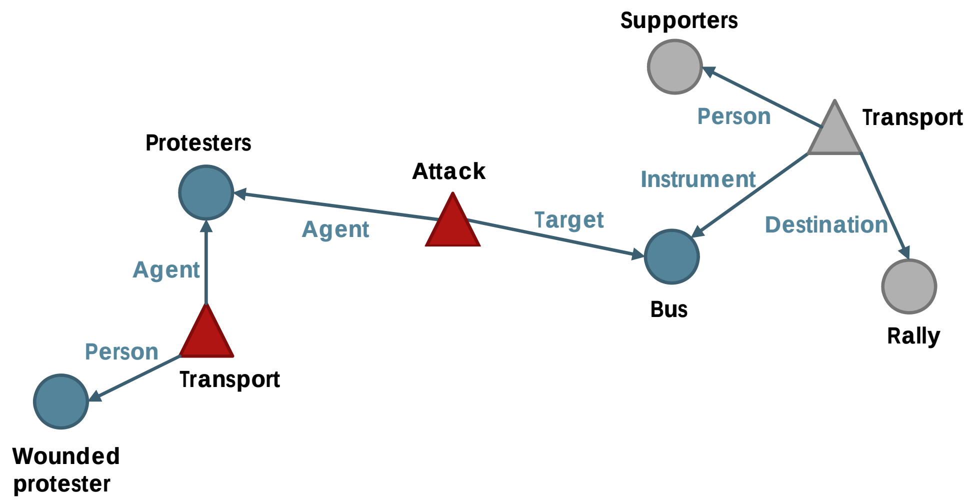
***TransportPerson_Instrument* = stretcher**

Why Multimodal?

- Visual data contains complementary data used for:
  - Visual Illustration
  - Disambiguation
  - Additional Details

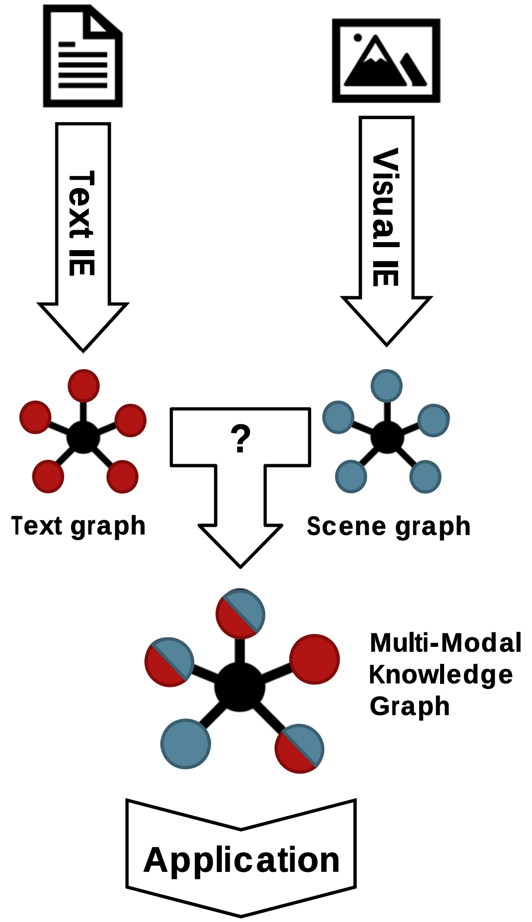
News Article: Thai opposition protesters[Attacker] attack[Attack] a bus[Target] carrying pro-government Red Shirt supporters on their way to a rally. Protesters[Agent] are carrying [TransportPerson] a wounded protester[Person] to . ...





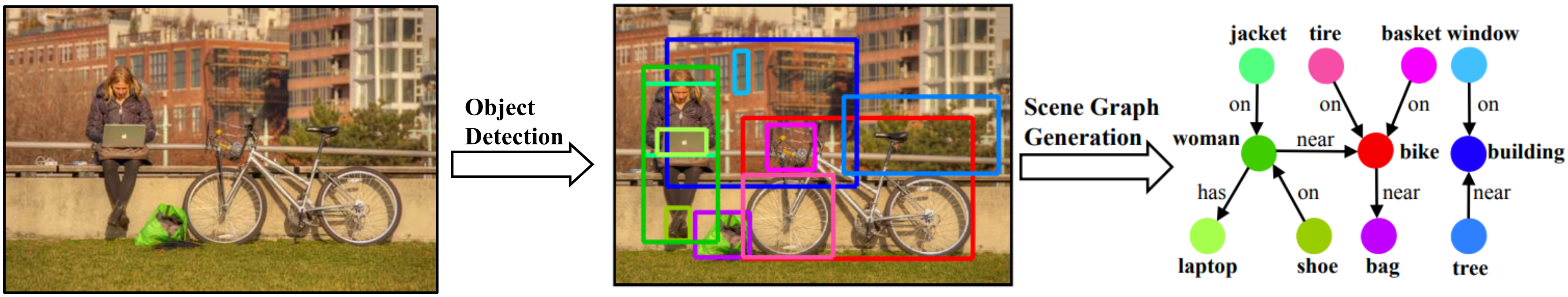
Challenges & Applications

- Challenges:
  - Parsing images/videos to structures
  - Grounding event/entities across modalities
  - Extracting complementary multimodal arguments



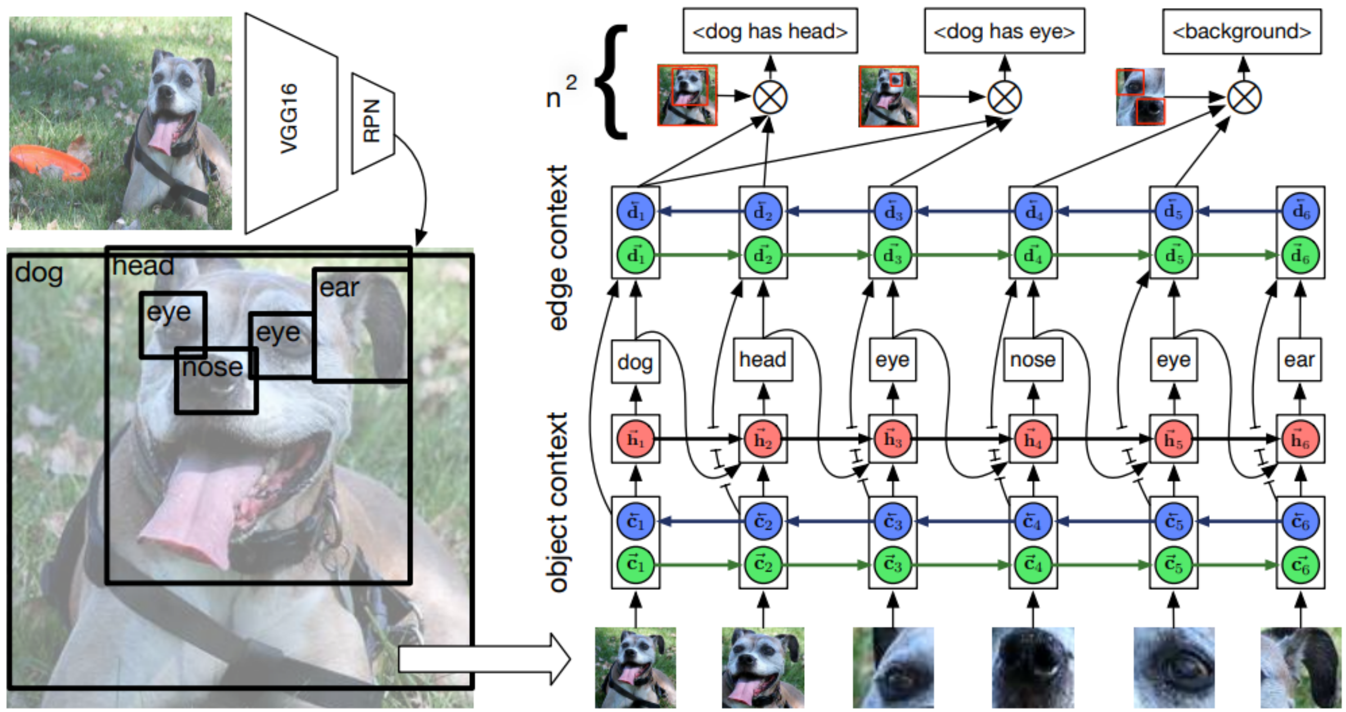
Challenge 1: Parsing Images to Scene Graphs

- Extract structured representation of a scene
  - Entities and their semantic relationships



Parsing Images to Scene Graphs

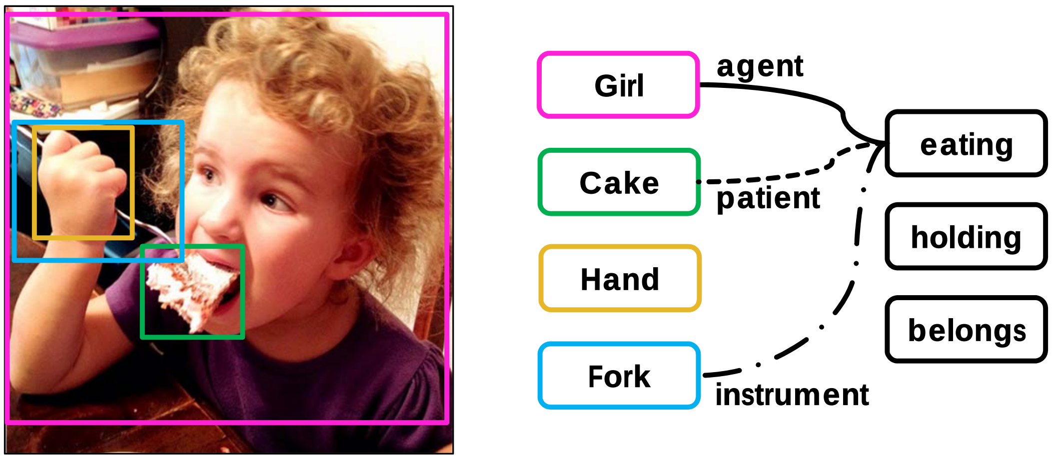
- Existing method
  - Extract object proposals
  - Contextualize features by RNN (or message passing)
  - Classify all nodes and pairs of nodes
- **Limitations**
  - Computationally exhaustive
    - $O(n^2)$ for $n \approx 100$ proposals
  - Difficult to model higher order relationships, e.g. *“girl eating cake with fork”*
  - Requires full supervision



Neural Motifs (Zellers, Yatskar, Thomson, Choi, CVPR 2018)
One of the SOTA methods for scene graph generation

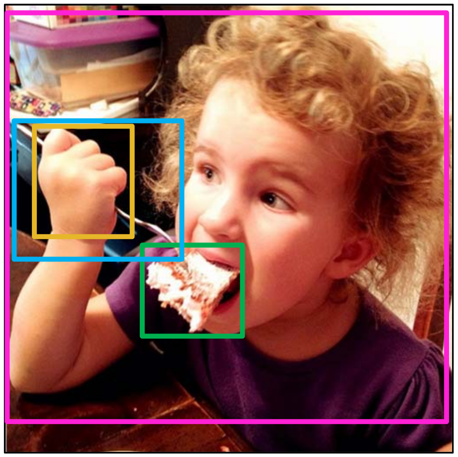
Reformulate as an Event-Centric Problem

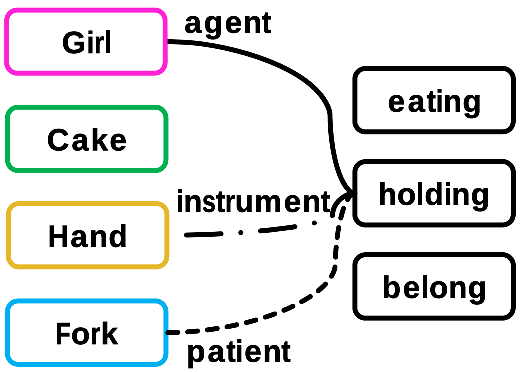
- Our work: **Visual Semantic Parsing Network (Zareian et al. CVPR19)**
  - Generalized formulation of scene graph generation
    - Entity-centric → bipartite representation of predicates & entities
    - Reduce computational complexity from $O(n^2)$ to sub-quadratic
    - Model argument role relations beyond (subject, object), (agent, patient) relations



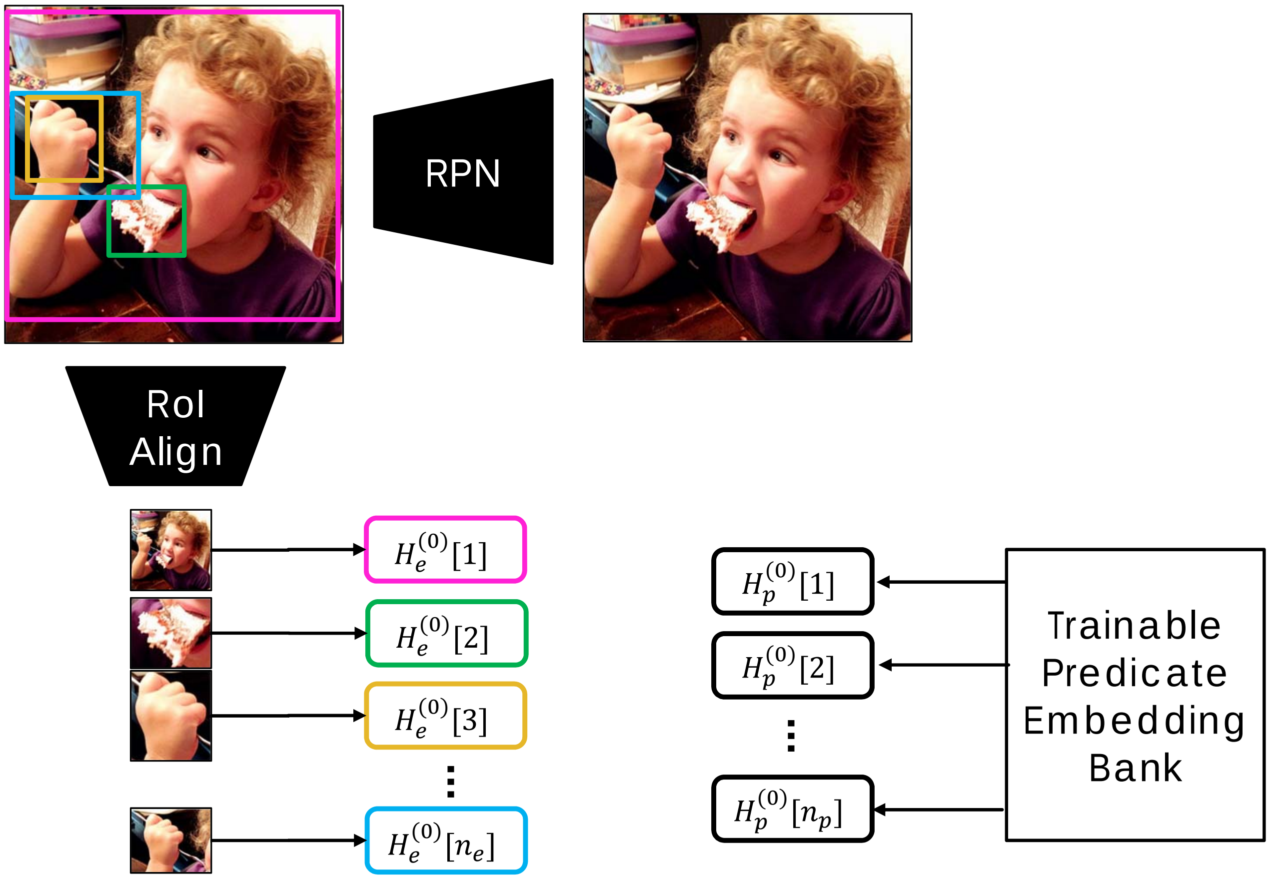
Reformulate as an Event-Centric Problem

- Our work: **Visual Semantic Parsing Network (Zareian et al. CVPR20)**
  - Generalized formulation of scene graph generation
    - Entity-centric → bipartite representation of predicates & entities
    - Reduce computational complexity from $O(n^2)$ to sub-quadratic
    - Model argument role relations beyond (subject, object), (agent, patient) relations



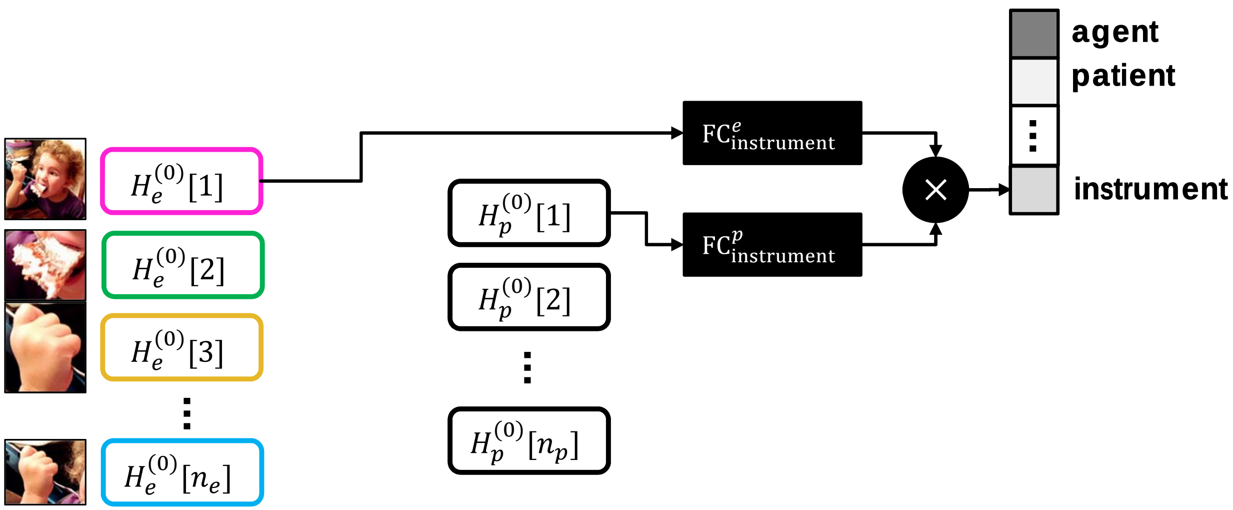


Bipartite Embeddings for Entity & Predicate



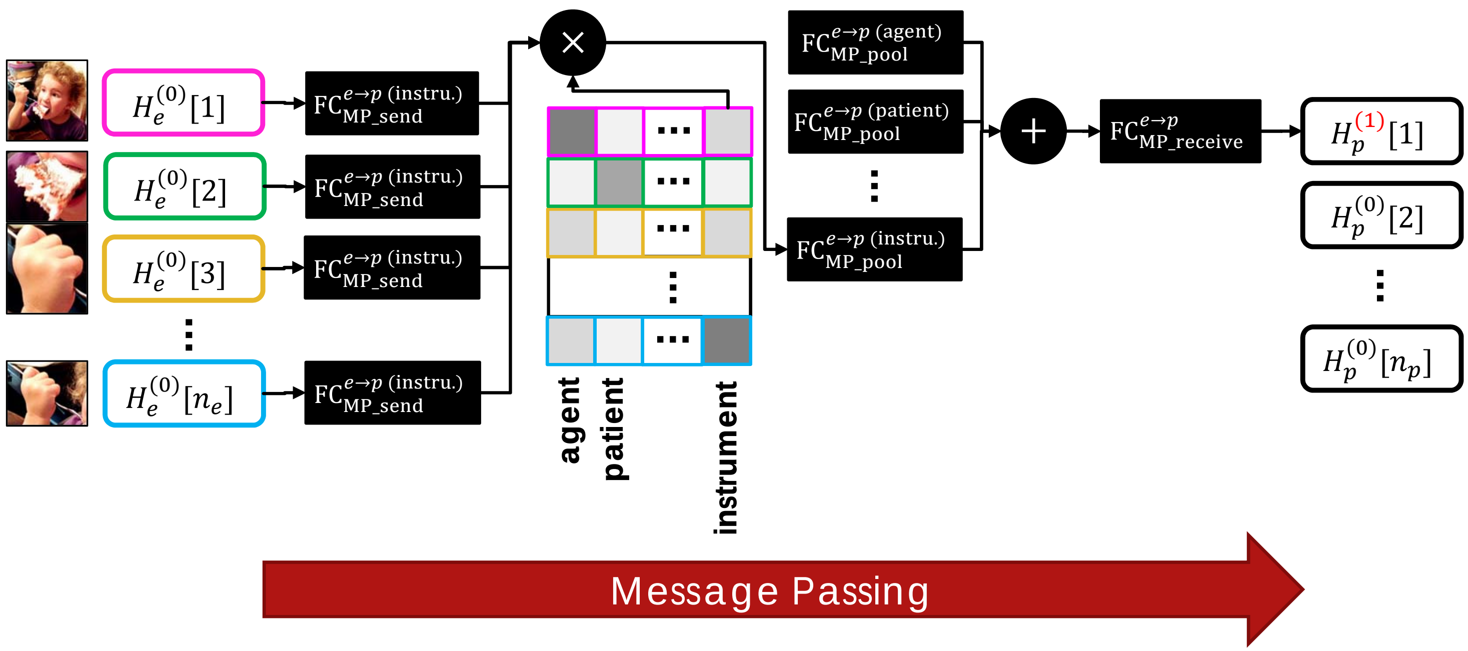
Argument Role Prediction

- Initialize entity and predicate nodes
- Compute role-specific attention scores
  - Input: entity-predicate feature pairs
  - Output: scalar for each thematic role



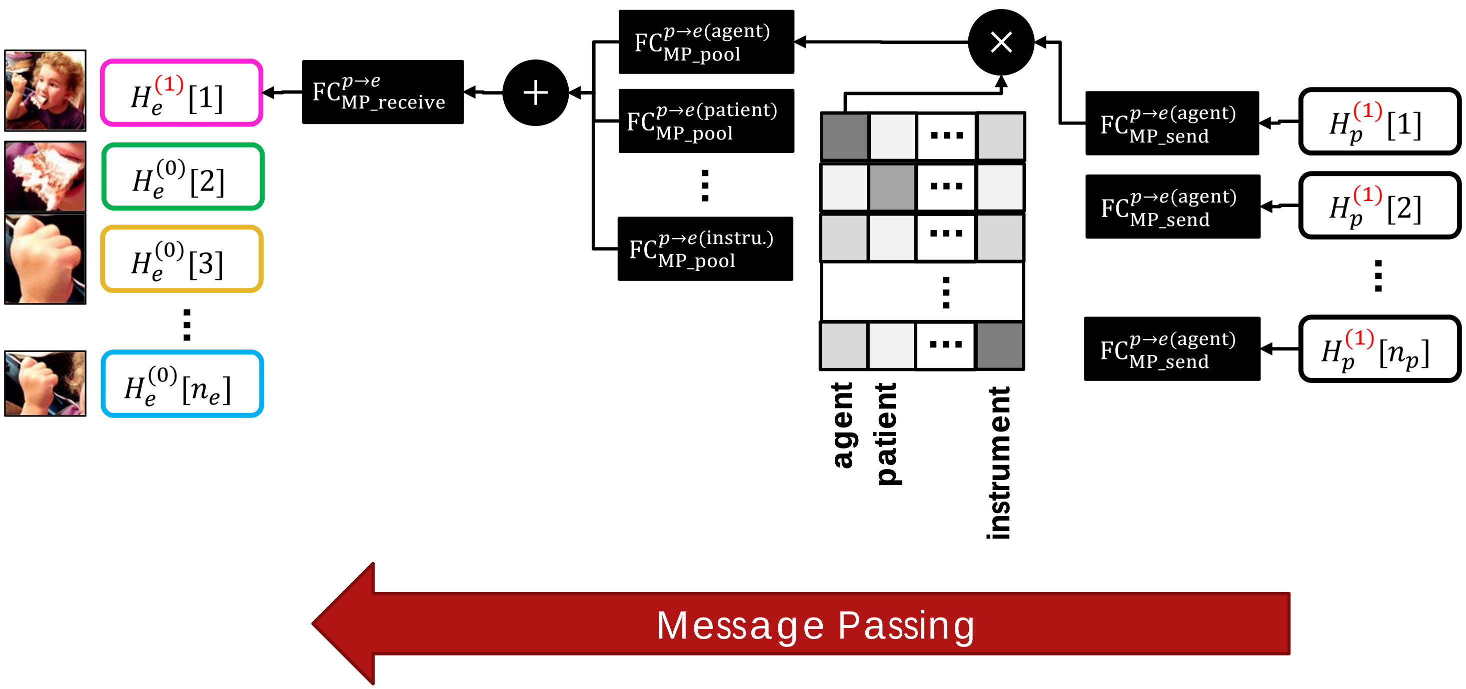
Role-Dependent Message Passing

- **Bi-directional Message passing**
- **Entities → Roles → Predicates**



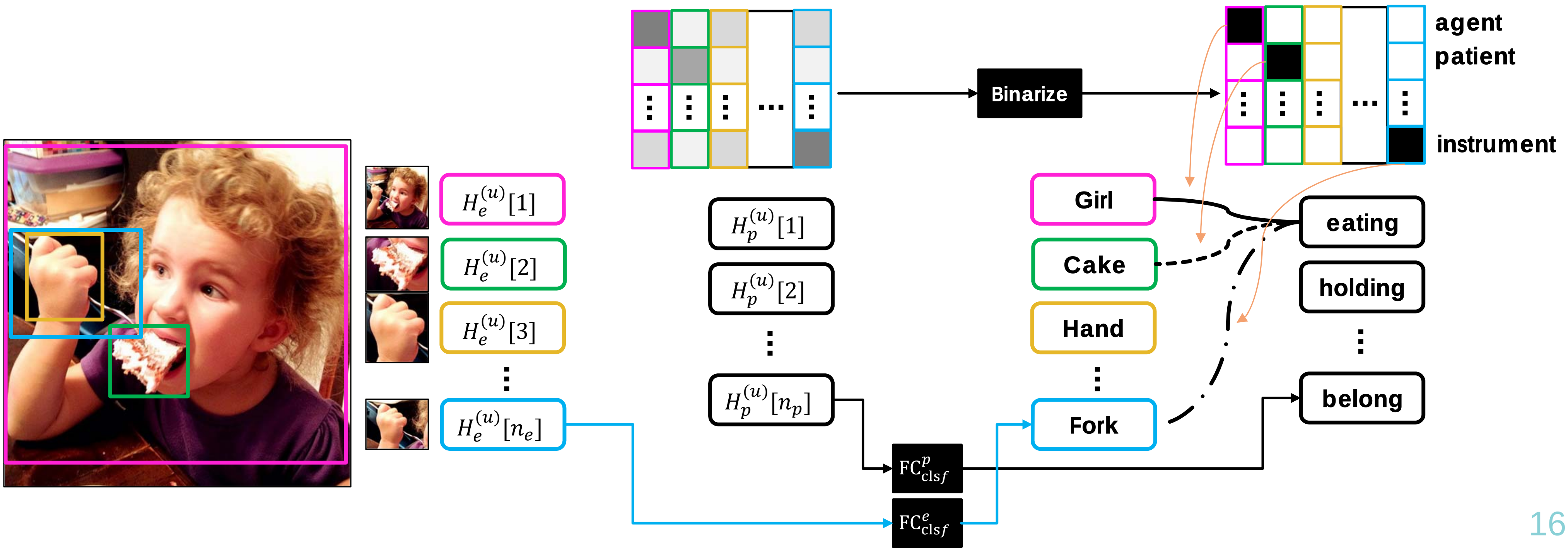
Role-Dependent Message Passing

- **Bi-directional Message passing**
- **Entities ← Roles ← Predicates**



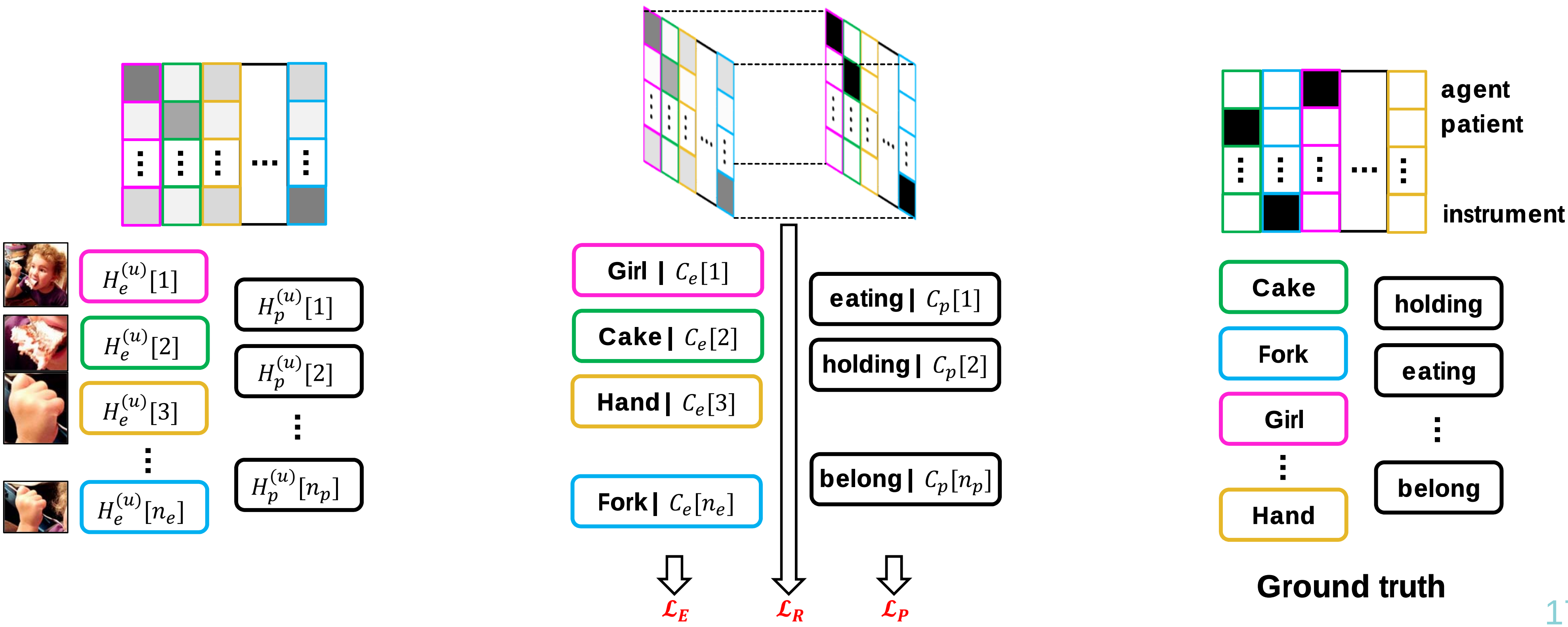
Visual Semantic Parsing Network

- Bi-directional Message passing
- Repeat for u iterations
- Classify nodes and edges

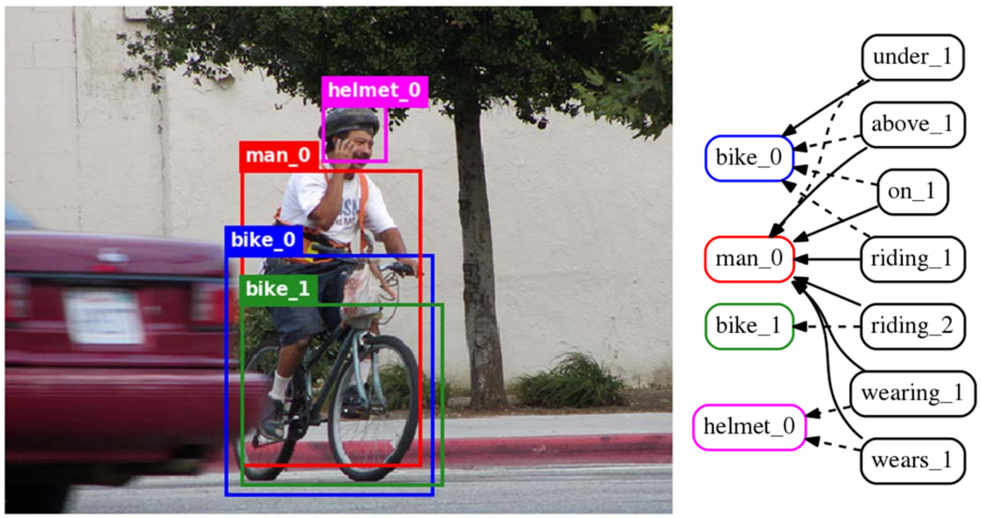


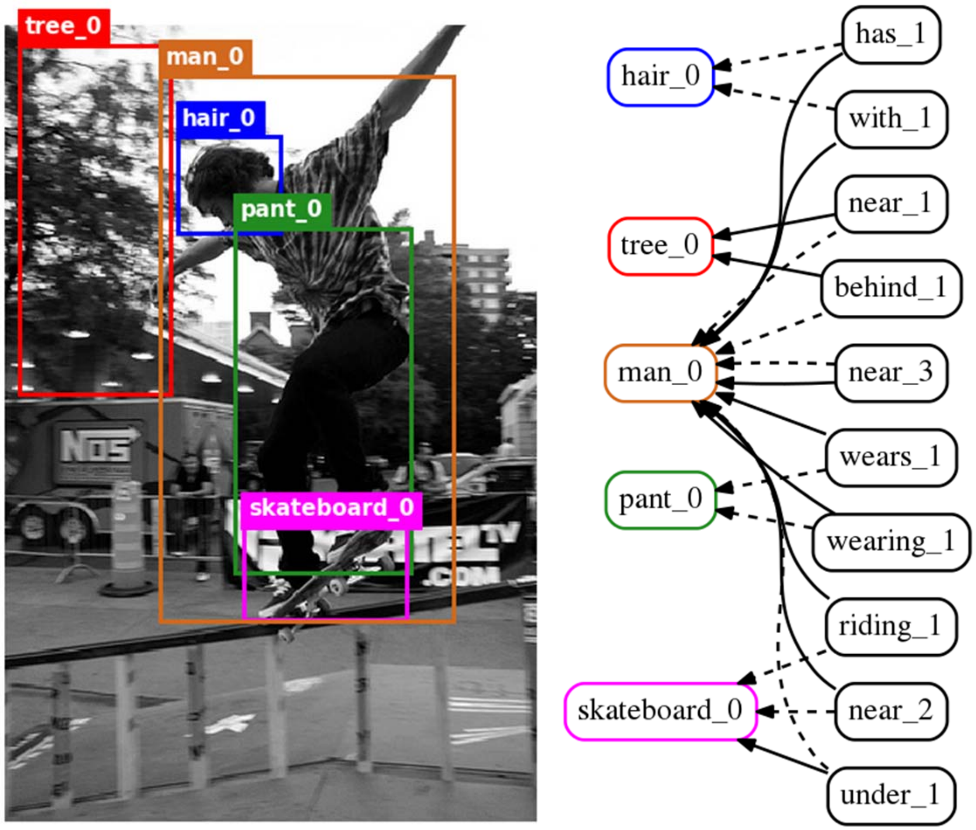
Visual Semantic Parsing Network

- **Weakly supervised training**
  - Unknown alignment between output and ground truth graphs



Visual Semantic Parsing Network

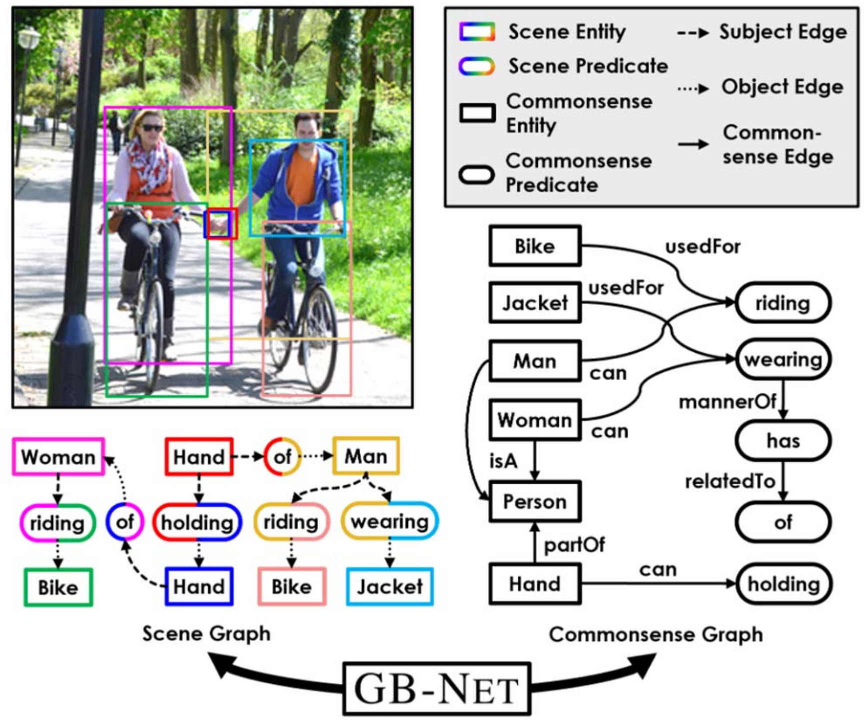




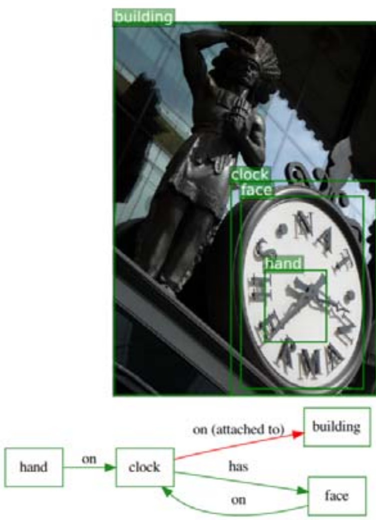
Incorporate External KB (Zareian, et al, ECCV20)

- Link concepts in scene graphs to external knowledge bases such as ConceptNet
- Pass messages over bridges between scene graphs and external graphs
- Refine bridges between graphs

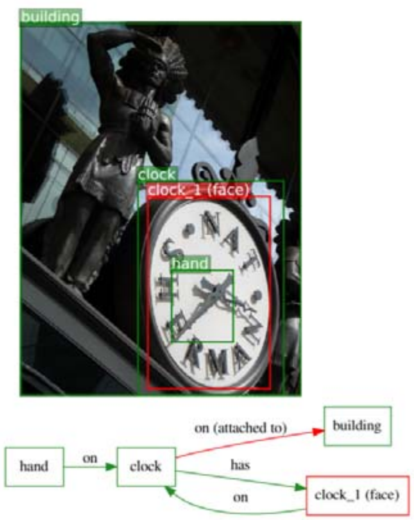
Task	Metric	GC	IMP+	FREQ	SMN	KERN	GB-NET	GB-NET-β
			Method					
SGGEN	mR@50	Y	3.8	4.3	5.3	*6.4*	6.1	**7.1**
		N	5.4	5.9	9.3	*11.7*	9.8	**11.7**
	mR@100	Y	4.8	5.6	6.1	7.3	*7.3*	**8.5**
		N	8.0	8.9	12.9	*16.0*	14.0	**16.6**
	R@50	Y	20.7	23.5	**27.2**	*27.1*	26.4	26.3
		N	22.0	25.3	*30.5*	**30.9**	29.4	29.3
	R@100	Y	24.5	27.6	**30.3**	29.8	*30.0*	29.9
		N	27.4	30.9	*35.8*	**35.8**	35.1	35.0



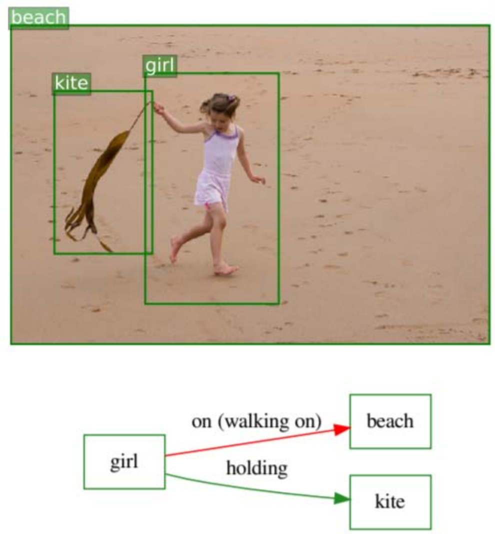
Scene Graph Examples of GB-NET



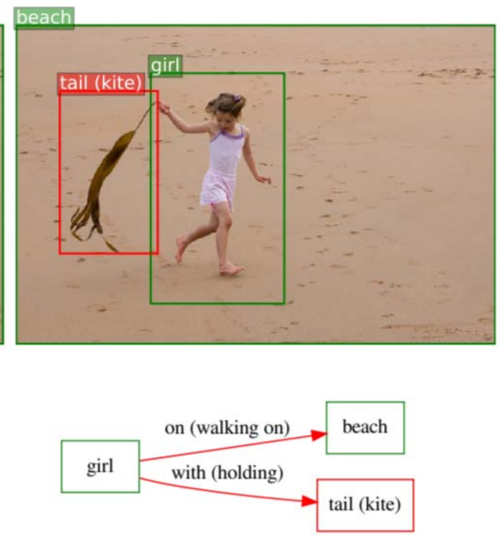
Ours (GB-Net)



Baseline (KERN)



Ours (GB-Net)



Baseline (KERN)

Challenge 2: Text-Visual Grounding (Akbari et al CVPR19)

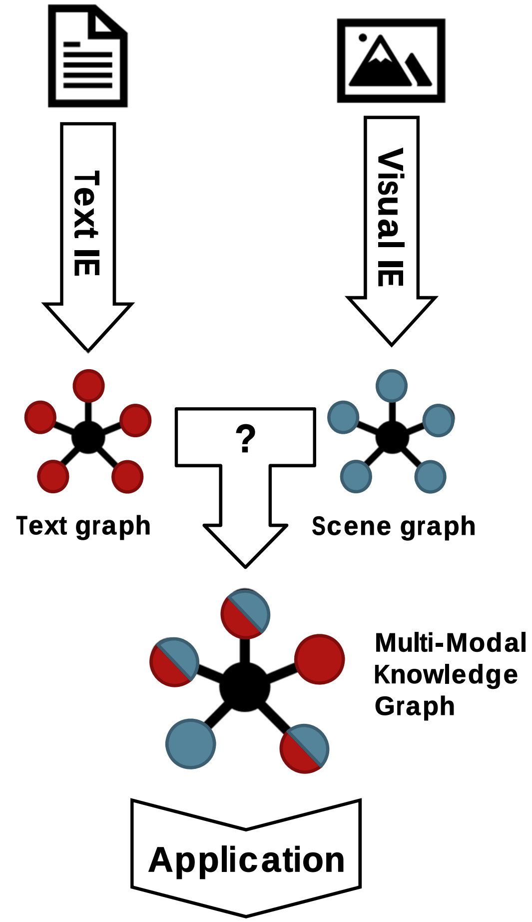
- Localize text query in image
  - Bridge visual and text knowledge graphs
  - Without using predefined classifiers

Figure 5. Image-sentence pair from Flickr30k with four queries (colored text) and corresponding heatmaps and selected max value (stars).

- Challenges
  - Sensitive to domain variations
  - Abstract concept not groundable

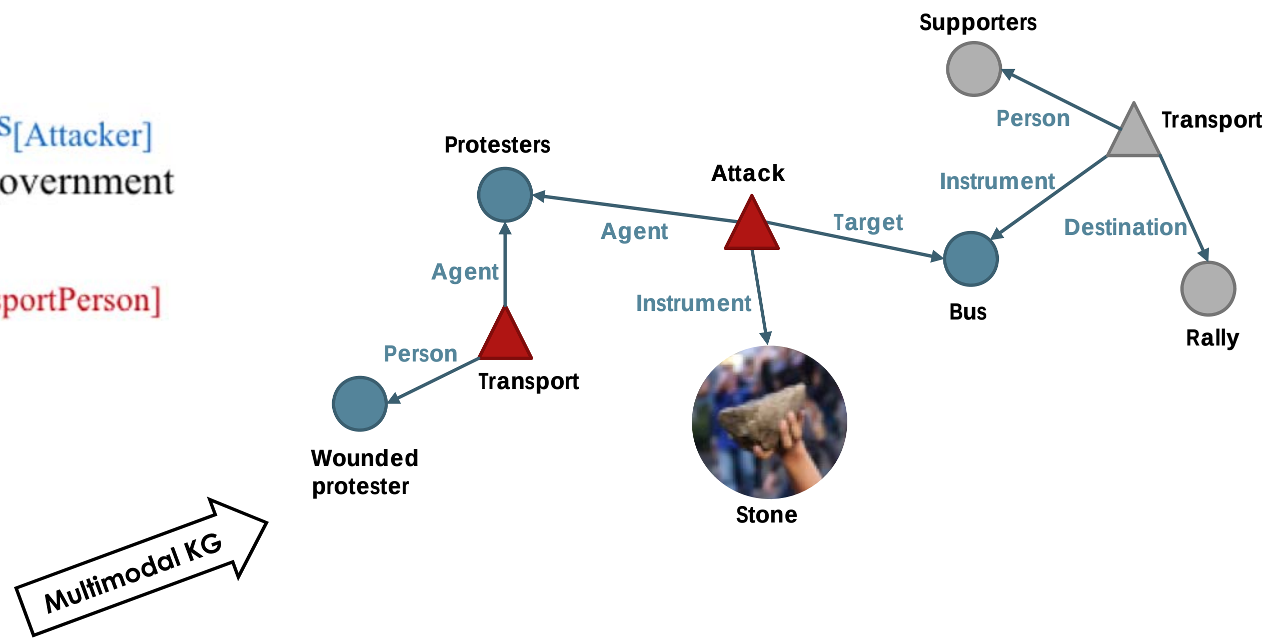
Challenge 3: Multimodal Event & Argument Extraction

- Challenges:
  - Parsing images/videos to structures
  - Grounding entities across modalities
  - Joint extraction of multimodal argument



Multimodal KG Example

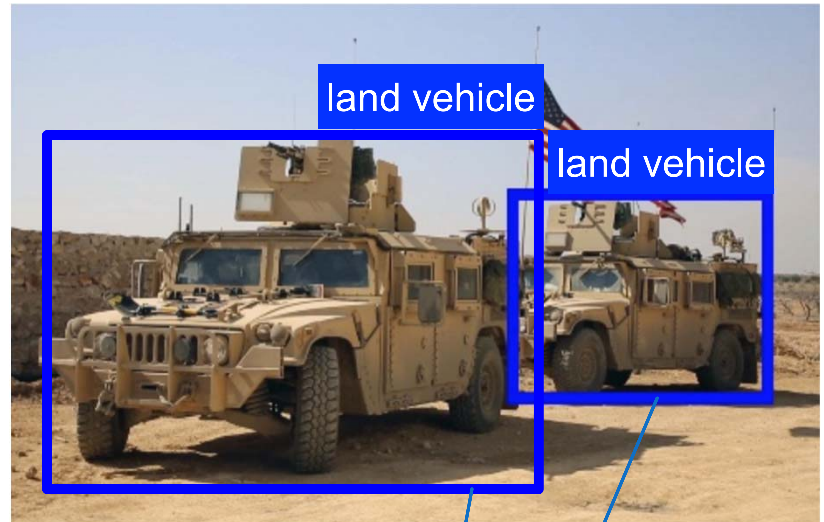
News Article: Thai opposition protesters[Attacker] attack[Attack] a bus[Target] carrying pro-government Red Shirt supporters on their way to a rally. Protesters[Agent] are carrying [TransportPerson] a wounded protester[Person] to . ...



A New Task: 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: Image-related Events & Visual Argument Roles

Event	Movement.TransportPerson	deploy
Arguments	Transporter	United States
	Destination	outskirts
	Passenger	soldiers
	Vehicle	land vehicle
	Vehicle	land vehicle

A New Task: Multimedia Event Extraction (M^2E^2)

Input: News article text and image

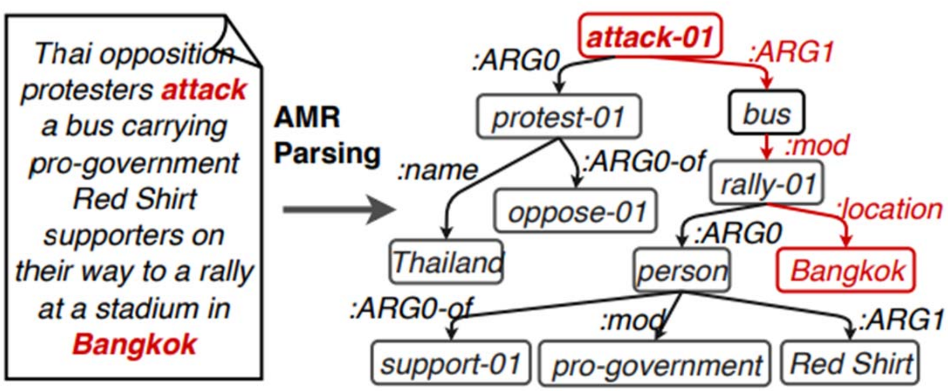
In March , Turkish forces escalated attacks on the YPG in northern Syria , forcing U.S. to deploy a small number of forces in and around the town of Manbij to the northwest of Raqqa to “deter” Turkish - SDF clashes and ensure the focus remains on Islamic State. Meanwhile, Raqqa is being pummeled by **airstrikes** mounted by **U.S.-led coalition forces** and Syrian warplanes. Local anti-IS activists say the air raids fail to distinguish between military and non-military targets …

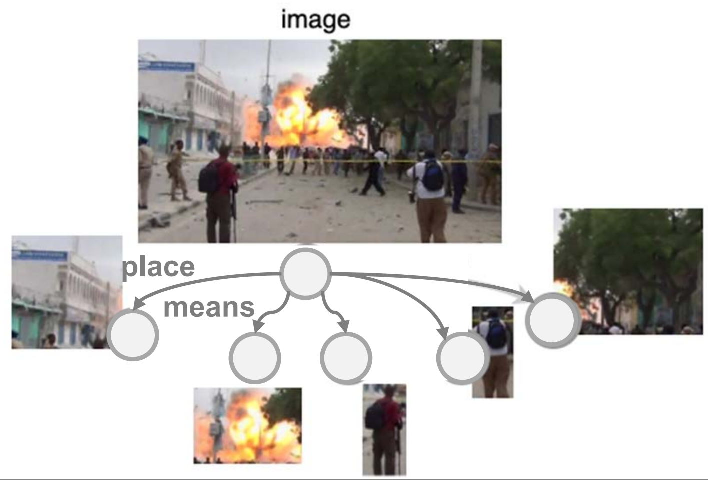
Output: Image-related Events & Visual Argument Roles

Event	Conflict.Attack	airstrikes
Arguments	Attacker	U.S.-led coalition forces
	Target	airplane
	Target	vehicle

Cross-media Structured Common Space

- Treat image as another language
- Represent it with a structure that is similar to AMR in text
- <u>Can we find a common representation?</u>





Linguistic Structure (Abstract Meaning Representation (AMR) / Dependency Tree)	**Visual Semantic Graph** [Zareian et al. CVPR20]

Image to Event Graph

- ImSitu dataset: situation recognition (Yatskar et al., 2016)
  - Classify an image as one of 500+ FrameNet verbs (sharing part of ACE)
  - Identify 192 generic semantic roles

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

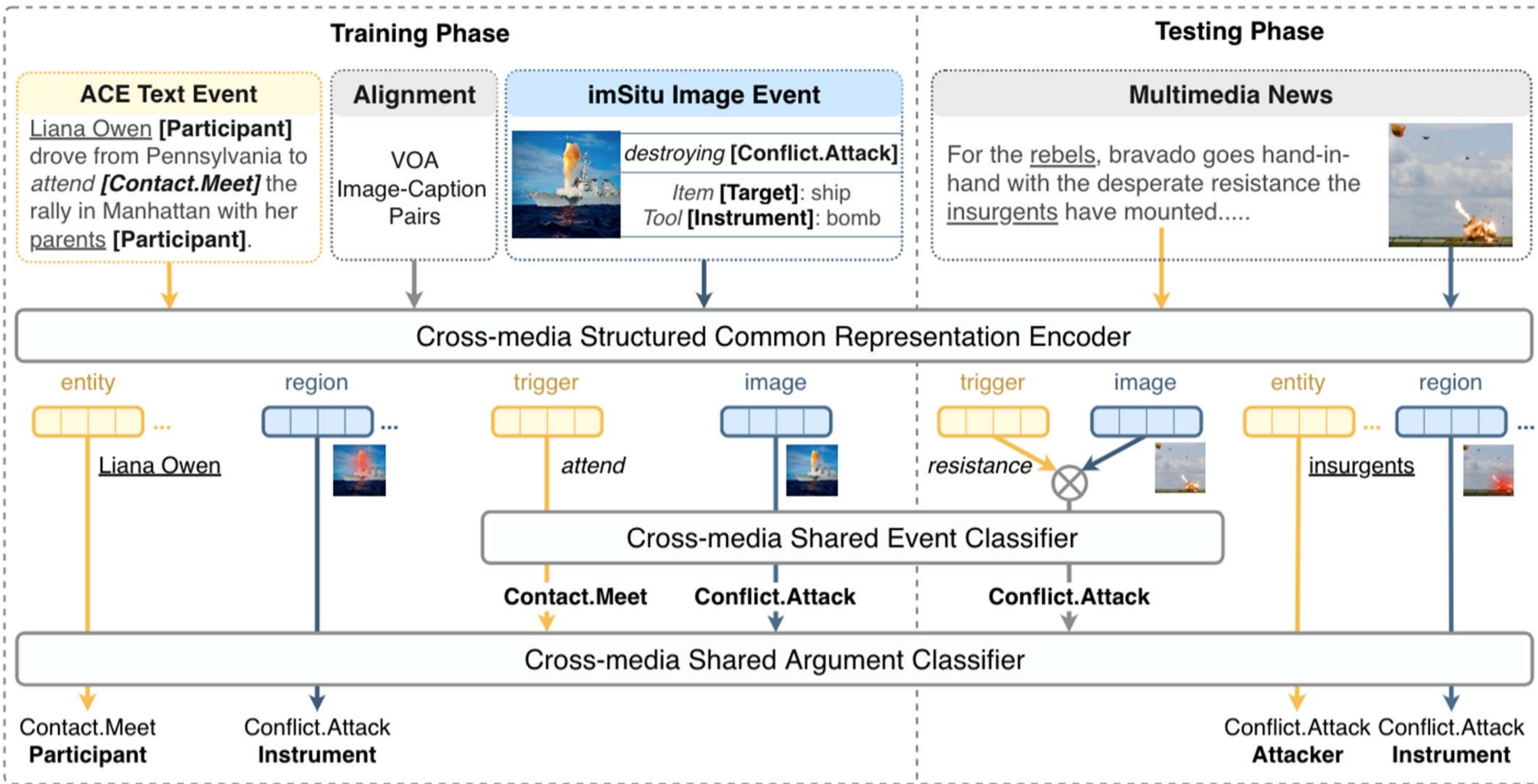
ROLE	VALUE	ROLE	VALUE
AGENT	BOY	AGENT	BEAR
SOURCE	CLIFF	SOURCE	ICEBERG
OBSTACLE	-	OBSTACLE	WATER
DESTINATION	WATER	DESTINATION	ICEBERG
PLACE	LAKE	PLACE	OUTDOOR

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

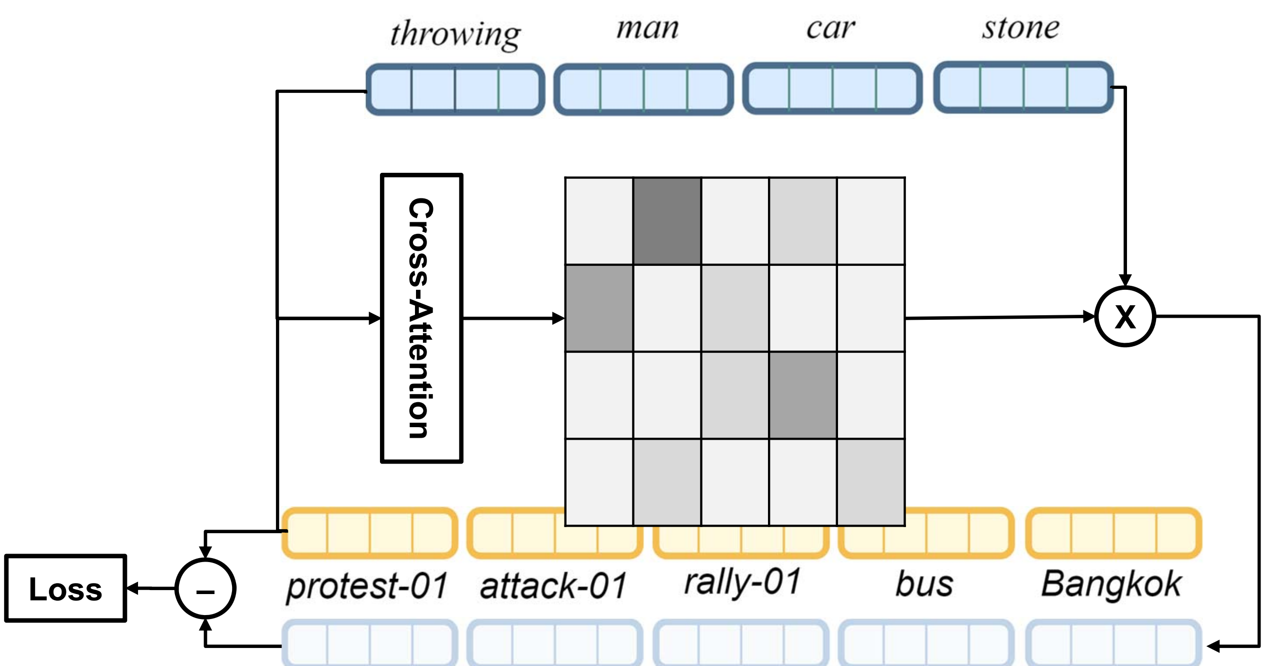
Weakly Aligned Structured Embedding (WASE)

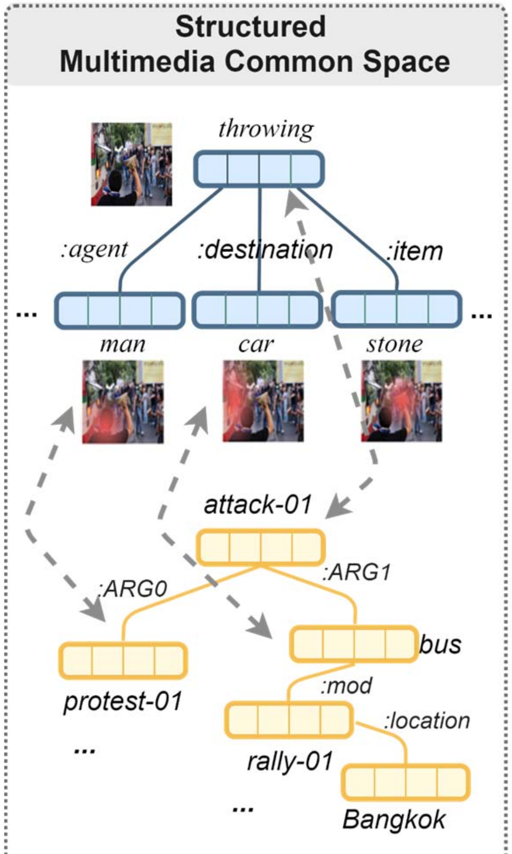
-- Cross-media shared representation and classifiers **(Li, Zareian, et al, ACL20)**



Use image-caption data for graph alignment

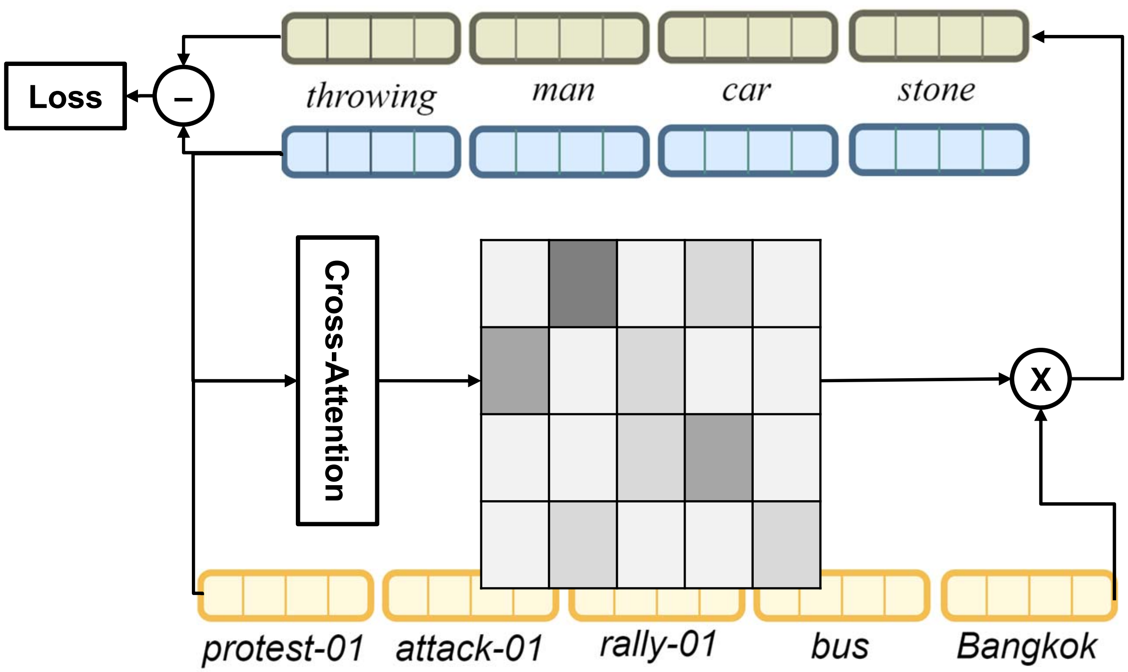
- Prior work aligns image-caption vectors by triplet loss.
- We want to align two graphs, not just single vectors.

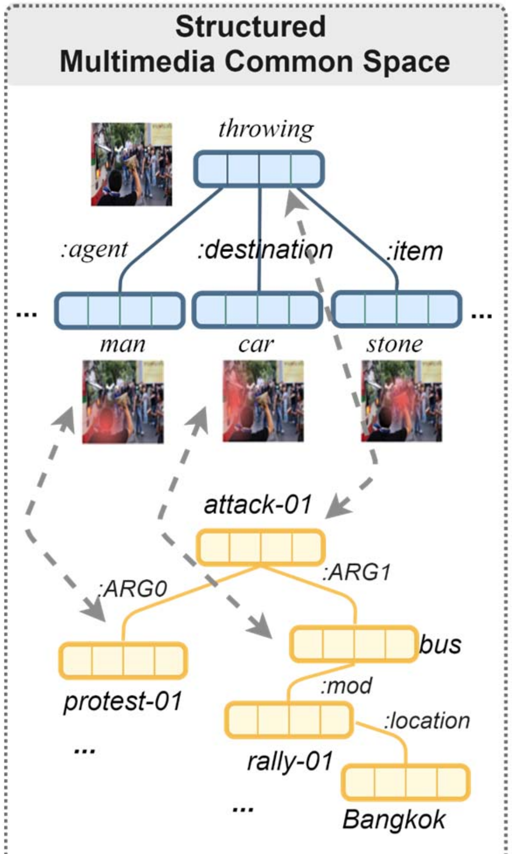




Use image-caption data for graph alignment

- Prior work aligns image-caption vectors by triplet loss.
- We want to align two graphs, not just single vectors.





A New Multimodal Dataset for M2E2 Evaluation

(Li, Zareian, et al, ACL20)

- Ontology: shared between ACE and imSitu
  - **Event Types**: cover 52% of ACE event types
  - **Argument Roles**: Based on ACE argument roles, add additional detectable visual roles (marked in red)

Event Type	Argument Roles
Life.Die	Agent, Victim, Instrument, Place, Time
Transaction.TransferMoney	Giver, Recipient, Beneficiary, Money, Instrument, Place, Time
Conflict.Attack	Attacker, Instrument, Place, Target, Time
Conflict.Demonstrate	Demonstrator, Instrument, Police, Place, Time
Contact.Phone-Write	Participant, Instrument, Place, Time
Contact.Meet	Participant, Place, Time
Justice.ArrestJail	Agent, Person, Instrument, Place, Time
Movement.Transport	Agent, Artifact/Person, Instrument, Destination, Origin, Time

Experiment Results

Training	Model	Text-Only Evaluation						Image-Only Evaluation						Multimedia Evaluation					
		Event Mention			Argument Role			Event Mention			Argument Role			Event Mention			Argument Role		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Text	JMEE	42.5	58.2	48.7	22.9	28.3	25.3	-	-	-	-	-	-	42.1	34.6	38.1	21.1	12.6	15.8
	GAIL	43.4	53.5	47.9	23.6	29.2	26.1	-	-	-	-	-	-	44.0	32.4	37.3	22.7	12.8	16.4
	$\text{WASE}^{\mathbb{T}}$	42.3	58.4	48.2	21.4	30.1	24.9	-	-	-	-	-	-	41.2	33.1	36.7	20.1	13.0	15.7
Image	$\text{WASE}^{\mathbb{I}}_{\text{att}}$	-	-	-	-	-	-	29.7	61.9	40.1	9.1	10.2	9.6	28.3	23.0	25.4	2.9	6.1	3.8
	$\text{WASE}^{\mathbb{I}}_{\text{obj}}$	-	-	-	-	-	-	28.6	59.2	38.7	13.3	9.8	11.2	26.1	22.4	24.1	4.7	5.0	4.9
Multimedia	VSE-C	33.5	47.8	39.4	16.6	24.7	19.8	30.3	48.9	26.4	5.6	6.1	5.7	33.3	48.2	39.3	11.1	14.9	12.8
	Flat_{att}	34.2	63.2	44.4	20.1	27.1	23.1	27.1	57.3	36.7	4.3	8.9	5.8	33.9	59.8	42.2	12.9	17.6	14.9
	Flat_{obj}	38.3	57.9	46.1	21.8	26.6	24.0	26.4	55.8	35.8	9.1	6.5	7.6	34.1	56.4	42.5	16.3	15.9	16.1
	WASE_{att}	37.6	66.8	48.1	27.5	33.2	**30.1**	32.3	63.4	42.8	9.7	11.1	10.3	38.2	67.1	49.1	18.6	21.6	**19.9**
	WASE_{obj}	42.8	61.9	**50.6**	23.5	30.3	26.4	43.1	59.2	**49.9**	14.5	10.1	**11.9**	43.0	62.1	**50.8**	19.5	18.9	19.2

Training with MM

Multimodal Task

Compare to Single Modality Extraction

- Image helps textual event extraction, and surrounding sentence helps visual event extraction

Missed by text-only model

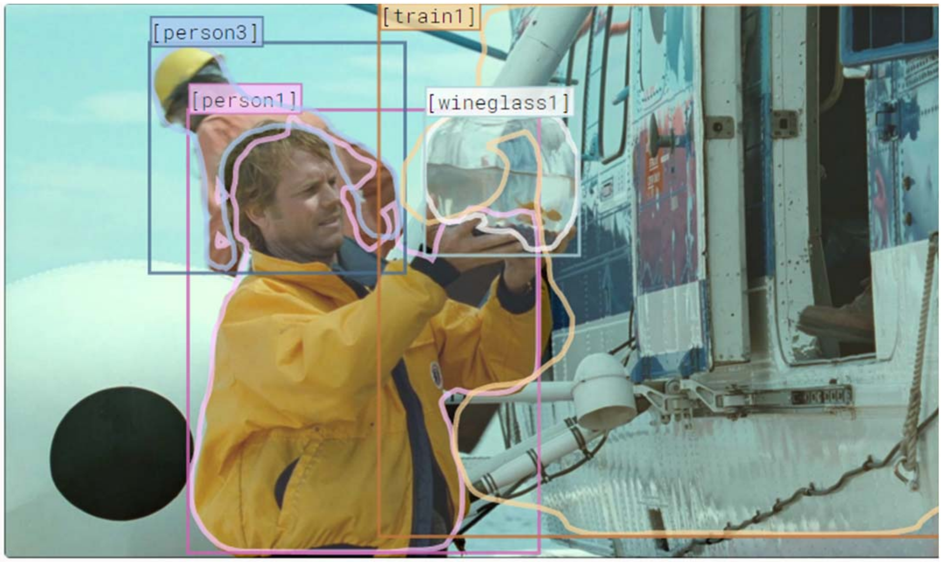
Iraqi security forces <u>*search*</u> ***[Justice.Arrest]*** a civilian in the city of Mosul.

People celebrate Supreme Court ruling on Same Sex Marriage in front of the Supreme Court in Washington.

Misclassified by image-only model as "Demonstration"

Application 1: Visual Commonsense Reasoning (VCR)

- Understand semantics in images and language, explore commonsense
- Provide to-the-point answer



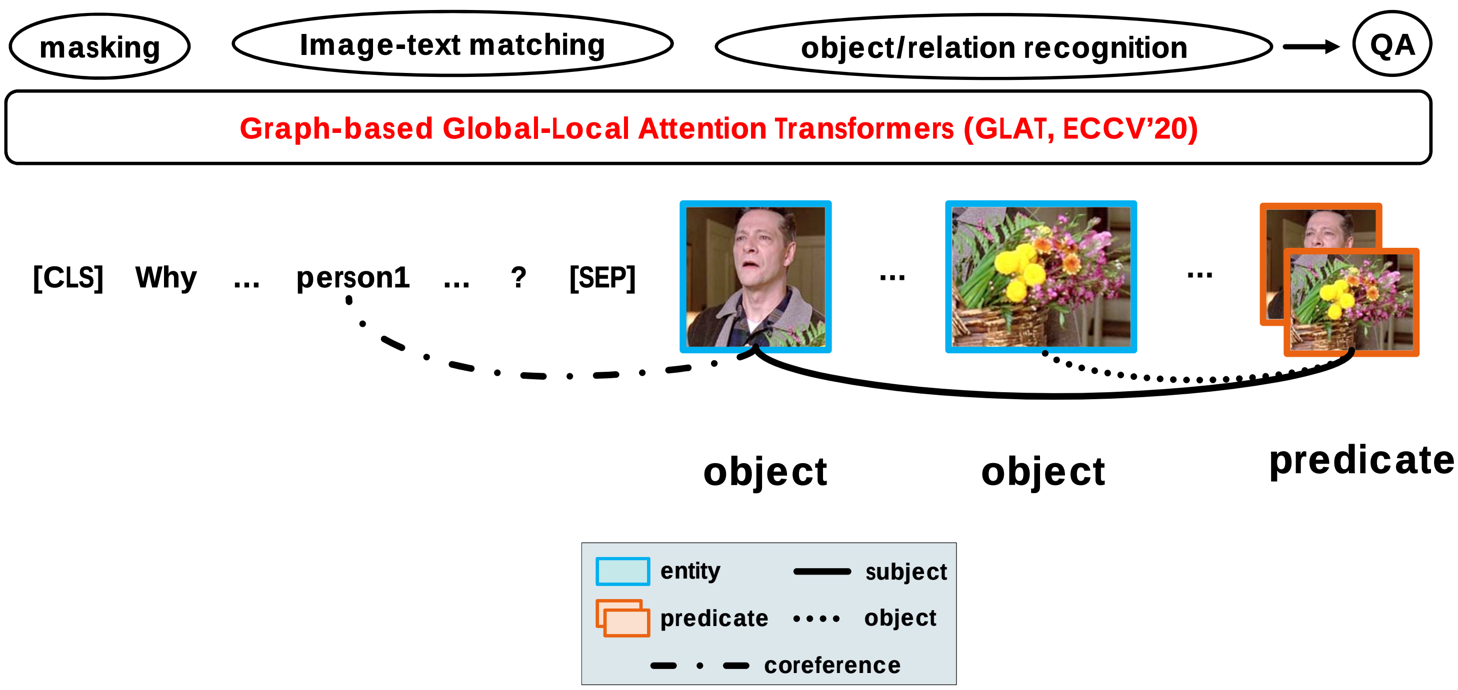
2. What is [person1] going to do next?

a) [person1] is going to put his hand in his pocket. **23.9%**
b) He is going to throw the paper on the ground and rant and rave at [person3] and [person2]. **11.2%**
c) [person2] is going to decide to start following a person who is out of camera's range but in his view. **0.0%**
d) He's going to put the fishbowl in the helicopter. **64.8%**

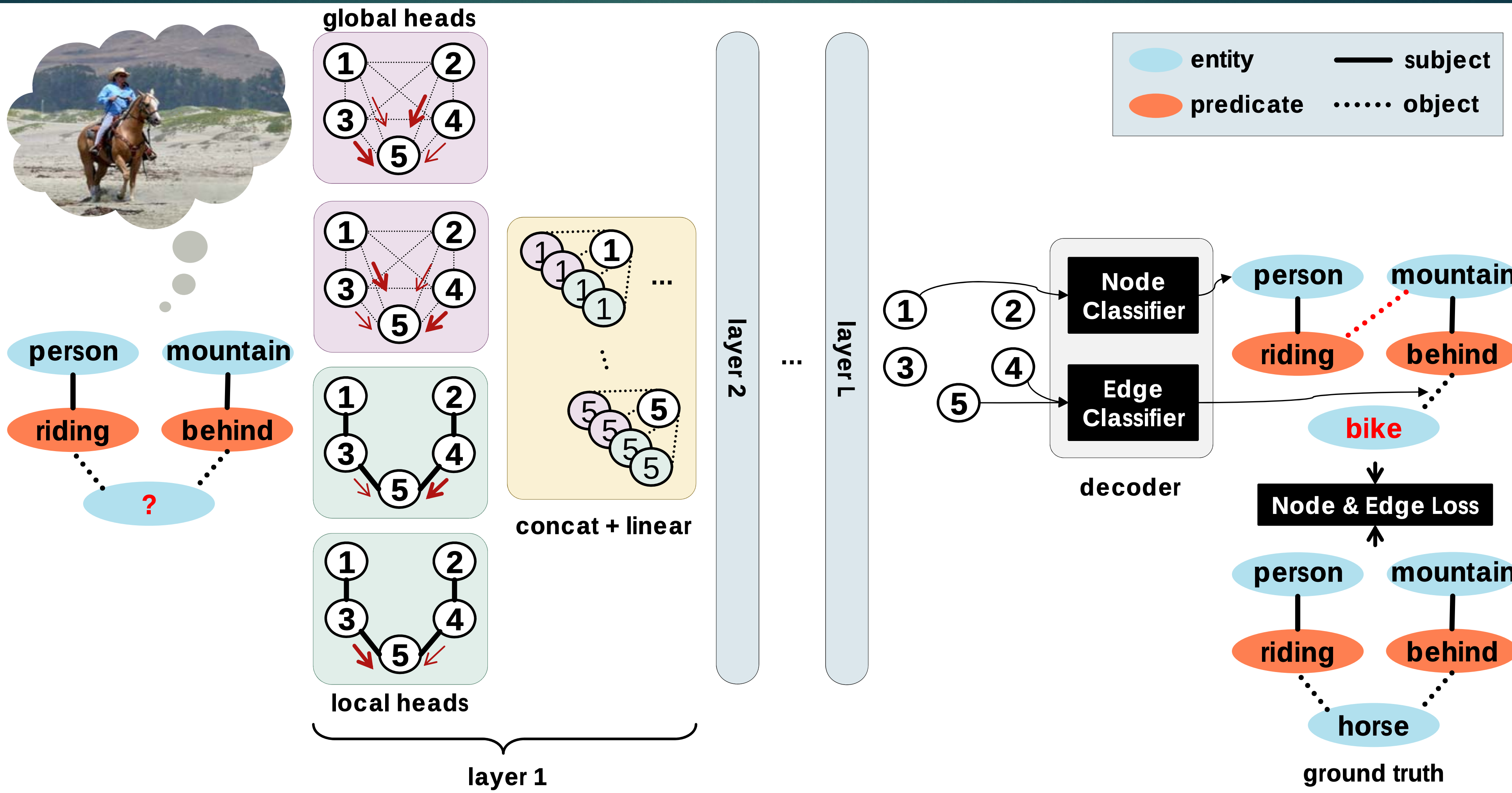
Zellers *et al.* CVPR 2019

Combine Visual Scene Graphs with VCR

- Expand input to include objects and predicate relations in graph
- Attention transformers limited to sparse connections in scene graphs



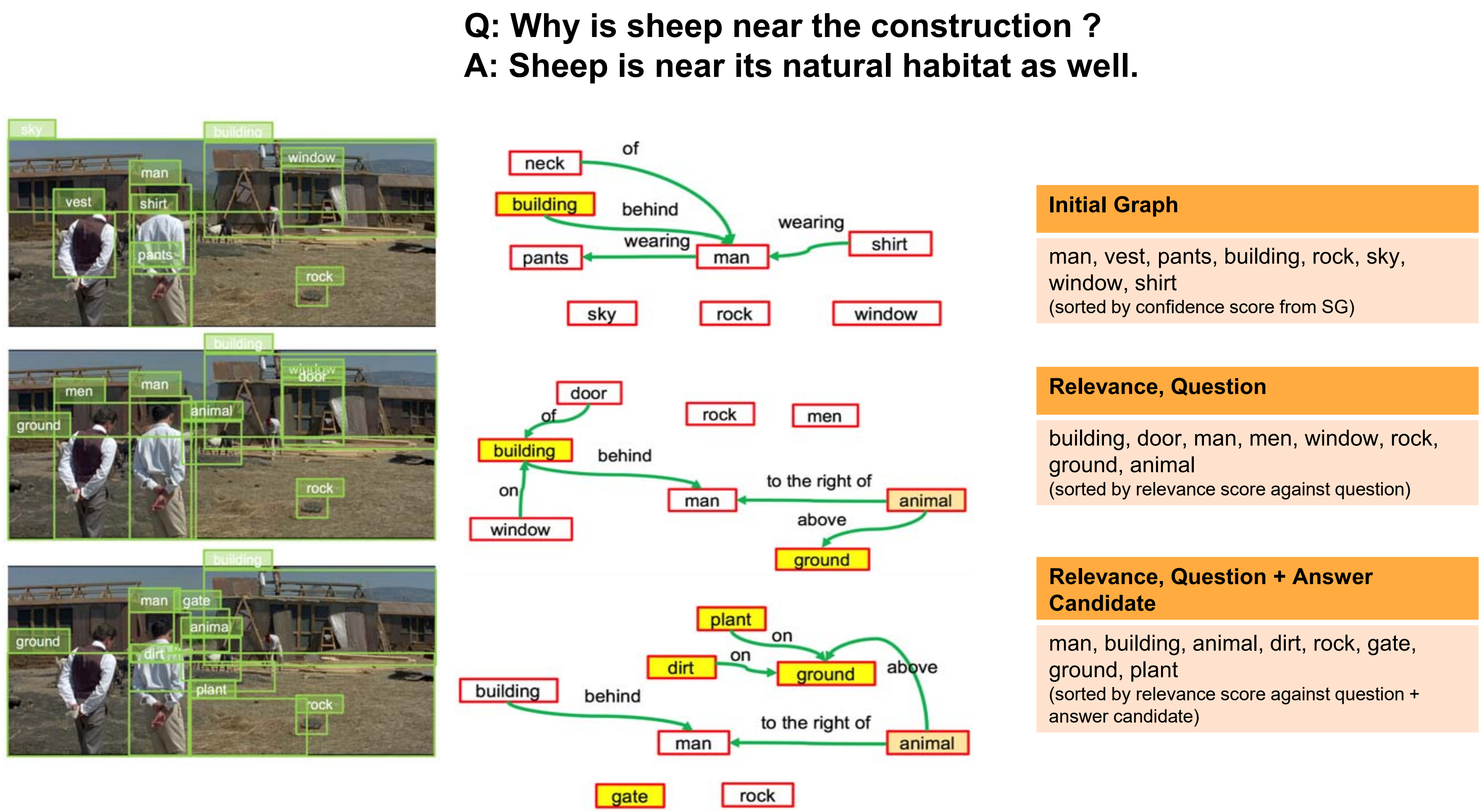
Graph-based Global-Local Attention Transformers (Zareian, et al ECCV20)



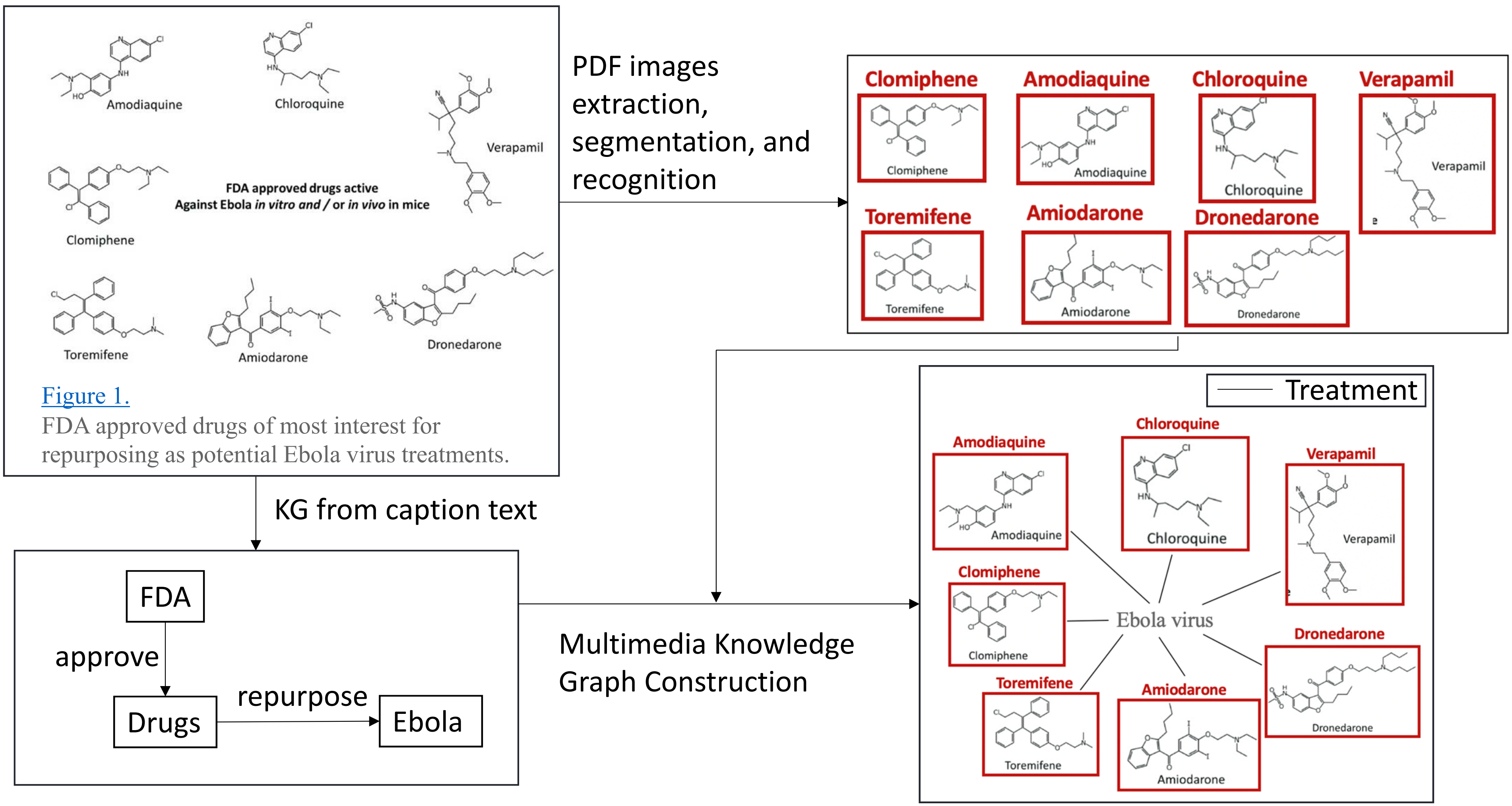
Scene Graph + Query-Adaptive Concept Selection

- For each question, select most relevant nodes on the scene graph

Model	Type	(Entity #, Predicate #)	Q -> A
LXMERT	Initial Graph	(36,18)	65.09 (baseline)
	Relevance Sel.	(8, x)	**74.04 (+8.95)**
GLAT (LXMERT)	Initial Graph	(36, 18)	65.24 (baseline)
	Relevance Sel.	(26, x)	69.57 (+4.33)
	Relevance Sel.	(18, x)	72.33 (+7.09)
	Relevance Sel.	(8, x)	**74.45 (+9.21)**

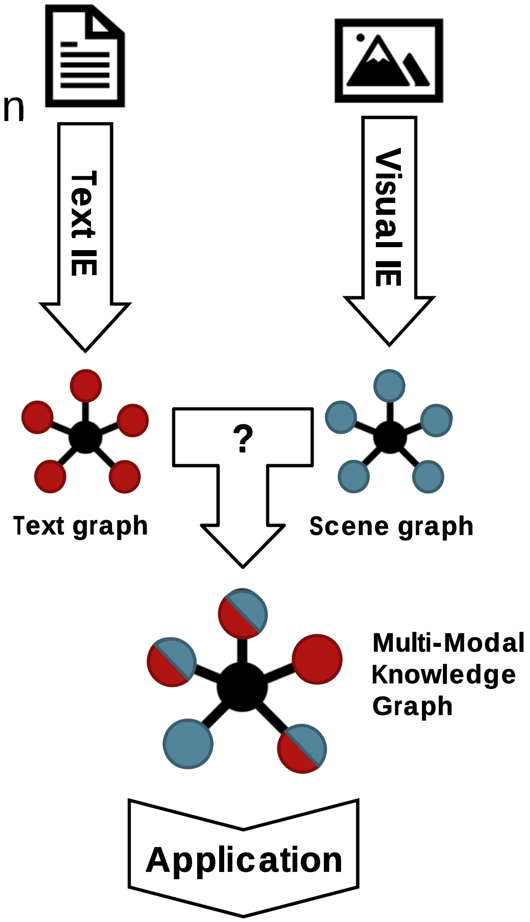

Q: Why is sheep near the construction ?
A: Sheep is near its natural habitat as well.
neck
of
building
behind
wearing
shirt
pants
wearing
man
sky
rock
window
door
rock
men
of
building
behind
on
man
to the right of
animal
window
above
ground
plant
on
on
dirt
ground
above
building
behind
man
to the right of
animal
gate
rock
Initial Graph
man, vest, pants, building, rock, sky, window, shirt
(sorted by confidence score from SG)
Relevance, Question
building, door, man, men, window, rock, ground, animal
(sorted by relevance score against question)
Relevance, Question + Answer Candidate
man, building, animal, dirt, rock, gate, ground, plant
(sorted by relevance score against question + answer candidate)

Application 2: Multimodal KG Extraction from COVID-19 Medical Papers



Conclusions

- Multimodal Knowledge Graphs
  - Understanding semantic structures in both language and vision
  - Joint representation and models
- Applications
  - Reasoning (VCR)
  - Discovery (COVID-19)
- Challenges
  - Open-vocabulary and Self-Supervised models
  - Knowledge graphs for video
  - Commonsense Extraction from MM KG
    physics, behavior, causal/temporal



References

- Zareian, Alireza, Svebor Karaman, and Shih-Fu Chang. "Weakly Supervised Visual Semantic Parsing." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR 2020.
- Zareian, Alireza, Svebor Karaman, and Shih-Fu Chang. "Bridging knowledge graphs to generate scene graphs." arXiv preprint arXiv:2001.02314 (2020). ECCV 2020.
- Akbari, Hassan, Svebor Karaman, Surabhi Bhargava, Brian Chen, Carl Vondrick, and Shih-Fu Chang. "Multi-level multimodal common semantic space for image-phrase grounding." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.
- Li, Manling, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. "Cross-media Structured Common Space for Multimedia Event Extraction." *arXiv preprint arXiv:2005.02472* (2020). ACL 2020.
- Zareian, Alireza, Haoxuan You, Zhecan Wang, and Shih-Fu Chang. "Learning Visual Commonsense for Robust Scene Graph Generation." *arXiv preprint arXiv:2006.09623* (2020). ECCV 2020.