NEURAL INFORMATION PROCESSING SYSTEMS

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

IIIT Hyderabad

Grounded Video Situation Recognition

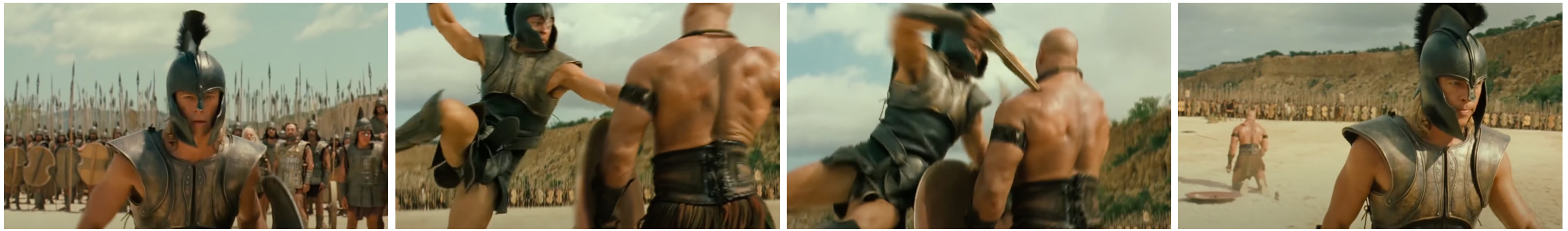
Zeeshan Khan

C.V. Jawahar

Makarand Tapaswi

https://zeeshank95.github.io/grvidsitu

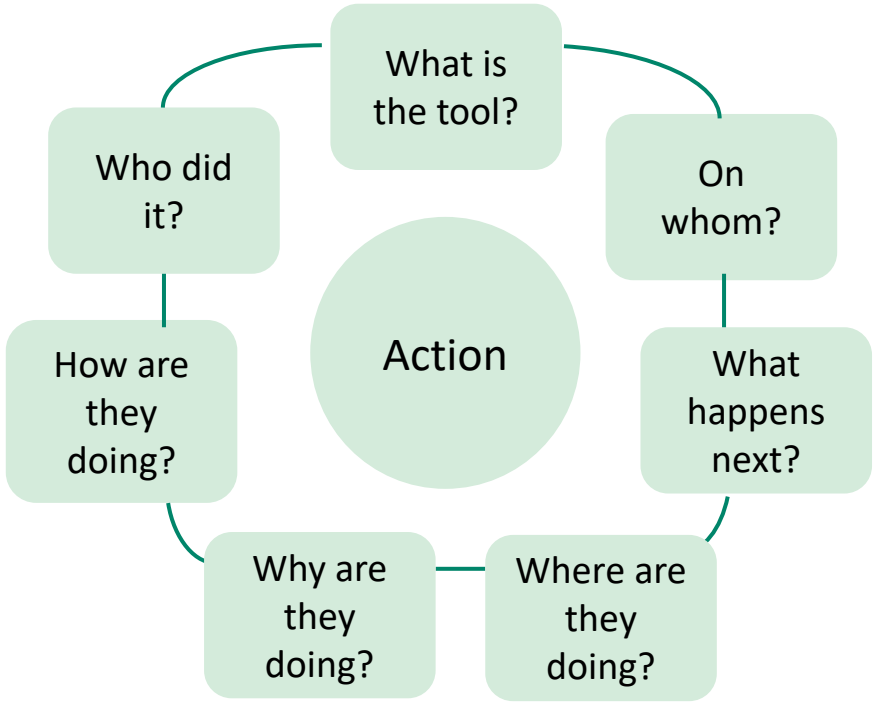
Video Understanding



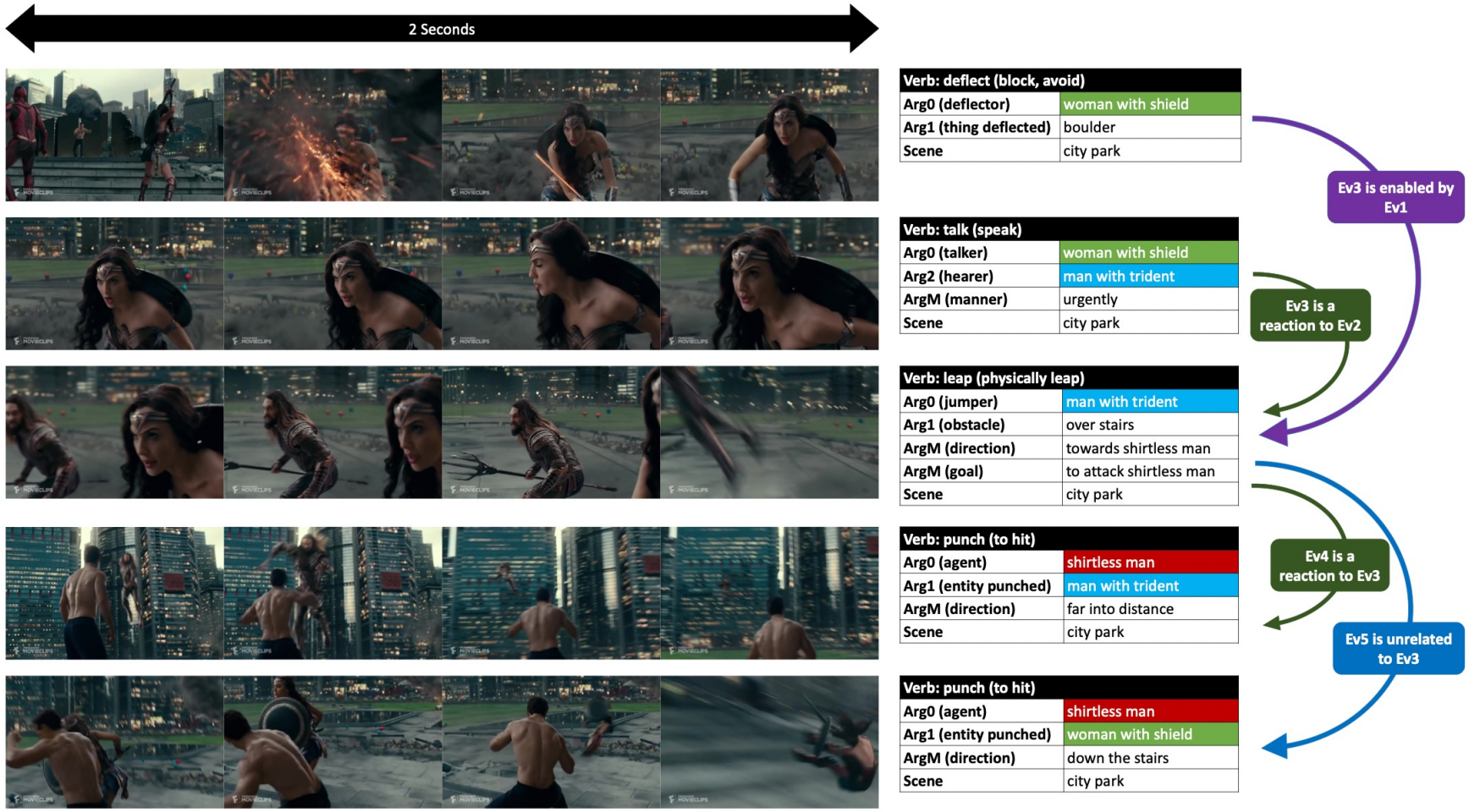
Sparse Uni-dimensional Understanding

- Action Recognition
- Captioning
- Spatio-Temporal Grounding
- Question Answering

Holistic Understanding



VidSitu: Video Situation Recognition

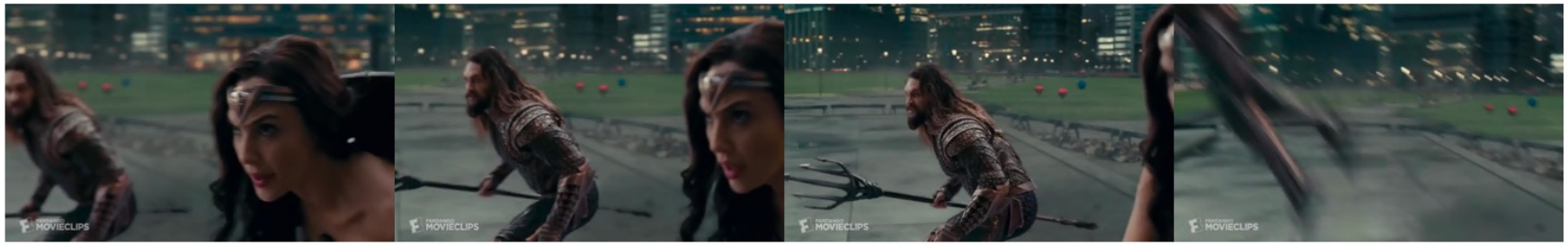


Holistic understanding through *Semantic Role Labelling:*

- Recognize the salient action verb in every event
- Predict roles and their entities that are part of this action
- Model simple event relations such as *enable* or *cause*

Arka Sadhu, et al. Visual Semantic Role Labeling for Video Understanding. In *CVPR*, 2021.

Challenges of VidSitu



Verb: talk (speak)	
Arg0 (talker)	woman with shield
Arg2 (hearer)	man with trident
ArgM (manner)	urgently
Scene	city park

Verb: leap (physically leap)	
Arg0 (jumper)	man with trident
Arg1 (obstacle)	over stairs
ArgM (direction)	towards shirtless man
ArgM (goal)	to attack shirtless man
Scene	city park

Challenges of VidSitu:

- Long-tailed distribution of actions and caption nouns
- Rare concepts such as shield, boulder, trident
- Ambiguity in captions: “Woman with shield” or “Woman in a costume” or “Woman with black hair”

Challenges of evaluating VidSitu:

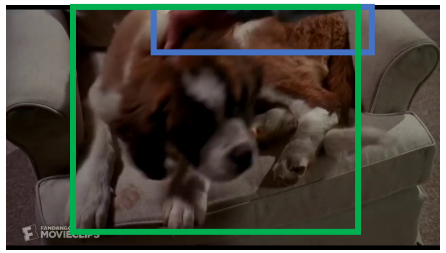
- Correct role disambiguation but incorrect caption
- Incorrect role disambiguation but correct caption
- Semantically correct caption but syntactically different

GVSR: Grounded Video Situation Recognition

Video 1

Event 1

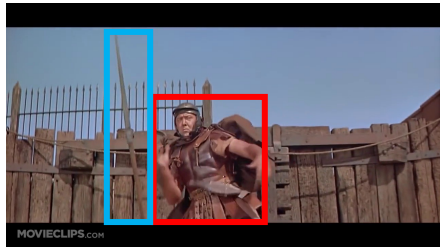
Event N



Event 1, Verb: ROLL	
Arg0 (Roller)	Boy in striped shirt
Arg1 (Thing rolled)	Himself
ArgM (Direction)	Back and forth
Arg Scene	Backyard

Event N, Verb: RUB	
Arg0 (Rubber)	Person in blue shirt
Arg1 (Thing rubbed)	Dog
Arg2 (Surface)	Hand
Arg Scene	Backyard

Video 2



Event 1, Verb: HIT	
Arg0 (Hitter)	Man in armor
Arg1 (Thing hit)	Bald man
Arg2 (Instrument)	Spear
Arg Scene	Arena

Event N, Verb: WINCE	
Arg0 (Wincer)	Bald man
Arg Scene	Arena

Video 3

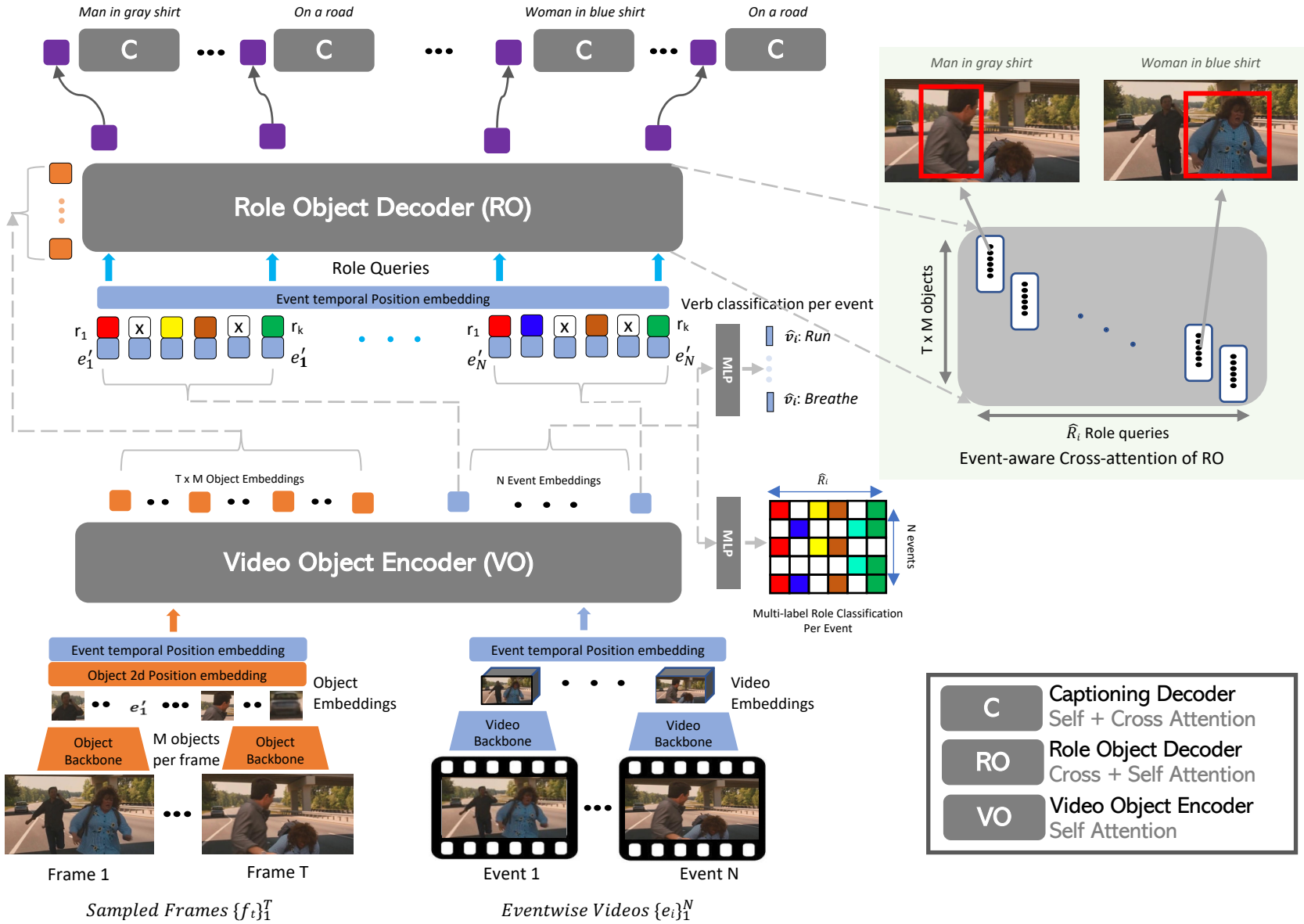
Event 1, Verb: LIFT	
Arg0 (Elevator)	Blonde woman
Arg1 (Thing lift)	Her phone
ArgM (Direction)	Up
ArgM (Manner)	Quickly
Arg Scene	An open field

Event N, Verb: TALK	
Arg0 (Talker)	The kneeling man
Arg1 (Hearer)	Blonde woman
Arg2 (Manner)	Confidently
Arg Scene	An open field

GVSR:

- Removes bias of the captioning module through spatio-temporal grounding of roles.
- Features joint prediction of three tasks:
  (i) Verb classification,
  (ii) Semantic Role Labelling, and
  (iii) Grounding for SRL.
- Grounding is performed in a weakly-supervised setting, without requirement for ground-truth bounding boxes during training!

VideoWhisperer



Three-stage Transformer model:

- **VO** encoder: Video-object encoder to capture fine-grained actions and entities across multiple events.
- **RO** decoder: Role-Object cross-attention decoder to identify and localize the role-entities.
- **C** decoder: Generate captions for each role in parallel.

Quantitative Results

Semantic role labeling

SoTA comparison, results for SRL and grounding with GT verb and role pairs

Method	CIDEr	C-Vb	C-Arg	R-L	Lea	IoU@0.3	IoU@0.5
SlowFast+TxE+TxD [28]	46.01	56.37	43.58	43.04	**50.89**	-	-
Slow-D+TxE+TxD [38]	60.34 $\pm$ 0.75	69.12 $\pm$ 1.43	53.87 $\pm$ 0.97	43.77 $\pm$ 0.38	46.77 $\pm$ 0.61	-	-
VideoWhisperer (Ours)	**68.54** $\pm$ 0.48	**77.48** $\pm$ 1.52	**61.55** $\pm$ 0.79	**45.70** $\pm$ 0.30	47.54 $\pm$ 0.55	**0.29** $\pm$ 0.013	**0.12** $\pm$ 0.01
Human Level	84.85	91.7	80.15	39.77	70.33	-	-

Grounded Video Situation Recognition

Results for end-to-end situation recognition. Model architecture is VO + RO + C

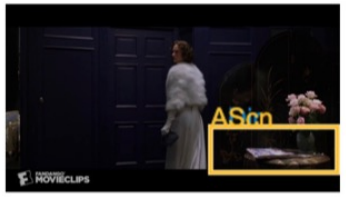
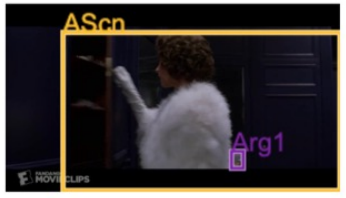
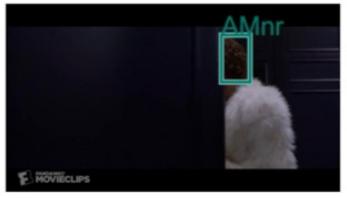
Model	Prediction			Verb	CIDEr	C-Vb	C-Arg	R-L	Lea	IoU	
	Verb	Role	SRL	Acc@1						0.3	0.5
VidSitu [28]	✓	✓	✓	46.79	30.33	39.56	23.97	29.98	35.92	-	-
VideoWhisperer	✓	✓	✓	44.06	52.30	61.77	38.18	35.84	38.00	0.13	0.05
	✓	GT	✓	45.06	68.54	77.48	61.55	45.70	47.54	0.29	0.12

Qualitative Results

https://www.youtube.com/watch?v=q6j_0vS_NNM&t=175s

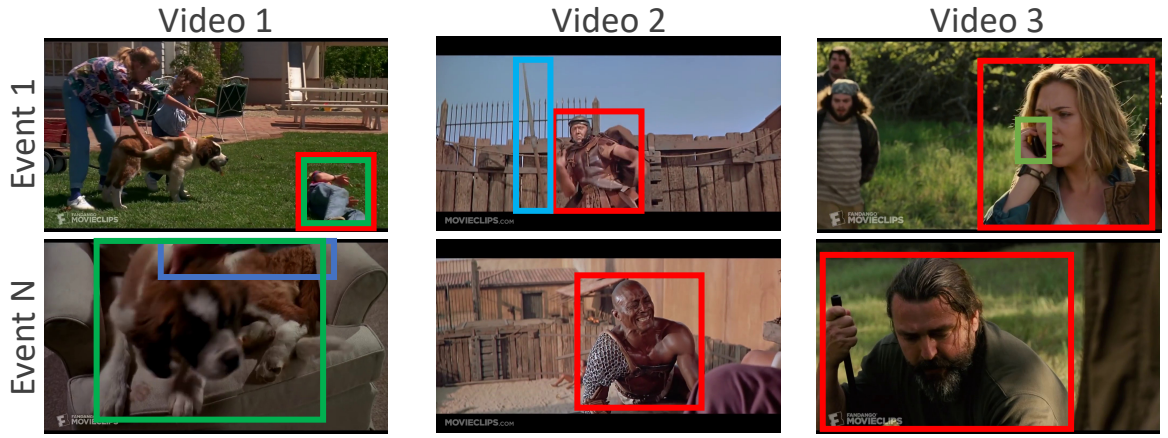
Qualitative Results

Video ID: v_q6j_0vS_NNM_seg_175_185

Event	Frame 1	Frame 2	Frame 3	Verb	Arg0	Arg1	Arg2	ADir	AMnr	ALoc	AScn
Ev1				walk.01	woman in white dress			towards the door	slowly		in a house
				walk.01	woman wearing white			towards a door	slowly		inside of a room with purple walls
Ev2				walk.01	woman in white dress	herself		around			in a house
				turn.01	woman wearing white	herself		back			inside of a room with purple walls
Ev3				walk.01	woman in white dress	herself	to get to the door	towards the door			in a house
				reach.03	woman wearing white	her arm	to open a cabinet	in front of her			inside of a room with purple walls
Ev4				open.01	woman in white dress	door			quickly		in a house
				open.01	woman wearing white	a cabinet			one at a time		inside of a room with purple walls
Ev5				write.01	woman in white dress	the door					in a house
				rummage.01	woman wearing white	shelves					inside of a room with purple walls

Thanks for listening!

Data and code at the project page
https://zeeshank95.github.io/grvidsitu



Event 1, Verb: ROLL	
Arg0 (Roller)	Boy in striped shirt
Arg1 (Thing rolled)	Himself
ArgM (Direction)	Back and forth
Arg Scene	Backyard

Event N, Verb: RUB	
Arg0 (Rubber)	Person in blue shirt
Arg1 (Thing rubbed)	Dog
Arg2 (Surface)	Hand
Arg Scene	Backyard

Event 1, Verb: HIT	
Arg0 (Hitter)	Man in armor
Arg1 (Thing hit)	Bald man
Arg2 (Instrument)	Spear
Arg Scene	Arena

Event N, Verb: WINCE	
Arg0 (Wincer)	Bald man
Arg Scene	Arena

Event 1, Verb: LIFT	
Arg0 (Elevator)	Blonde woman
Arg1 (Thing lift)	Her phone
ArgM (Direction)	Up
ArgM (Manner)	Quickly
Arg Scene	An open field

Event N, Verb: TALK	
Arg0 (Talker)	The kneeling man
Arg1 (Hearer)	Blonde woman
Arg2 (Manner)	Confidently
Arg Scene	An open field