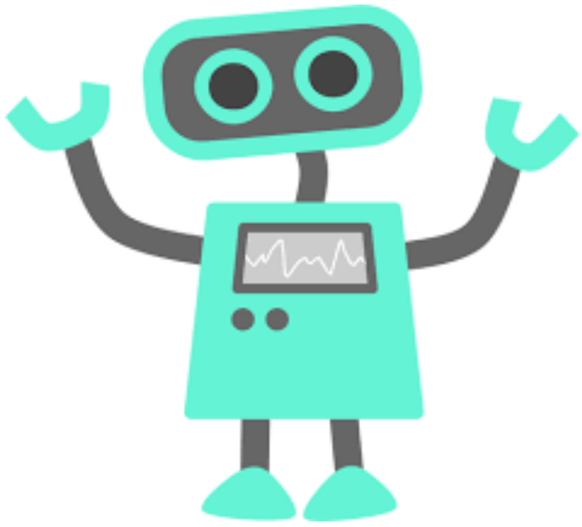
Knowledgeable and Dynamic Spatio-Temporal Language+Vision+Robotics

Mohit Bansal

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

(Lantern-EMNLP 2019 Workshop)

Beyond-Vision-Language's Diverse Requirements

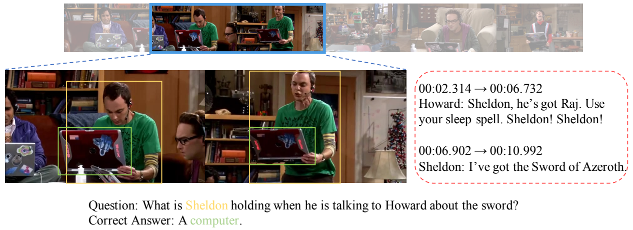


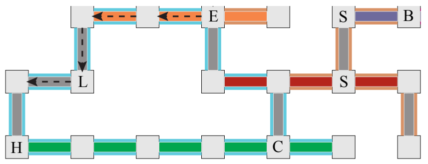
Commonsense & Auxiliary/ External Knowledge

Dynamic Video Context, not Static Images

Spatial-Temporal Localization & Referring Expressions

Language Understanding +Generation for Robotic Action Tasks





Theme

- Workshop theme: "work which goes beyond the task-specific integration of language and vision. That is, to leverage knowledge from external sources that are either provided by an environment or some fixed knowledge"
  - First we will talk about MTL and RL work that incorporates auxiliary knowledge such as entailment, video-generation, and saliency for video captioning style tasks (+AutoSeM)
  - Next, we will discuss our recent LXMERT framework that brings in external knowledge on both text and vision sides (as pretraining tasks) to do visual reasoning as new non-pretraining task
  - Spatial navigation w/ generalizable knowledge via unseen room+instruction data-augmentation
  - Commonsense reasoning for executing incomplete/ambiguous robotic instructions

  - 2nd part of the talk will briefly mention dynamic spatio-temporal knowledge for multimodal NLP:
    - Video- and subtitle-based multimodal QA task with spatial+temporal localization
    - Video-based dialogue dataset and task

External Knowledge and Commonsense

Auxiliary Knowledge via Multi-Task Learning

- MTL: Paradigm to improve generalization performance of a task using related tasks.

- The multiple tasks are learned in parallel (alternating optimization mini-batches) while using certain shared model representations/parameters.

- Each task benefits from extra information in the training signals of related tasks.

- Useful survey+blog by Sebastian Ruder for details of diverse MTL papers!

[Caruana, 1998; Argyriou et al., 2007; Kumar and Daume, 2012; Luong et al., 2016; Ruder, 2017]

Auxiliary Knowledge in Video Captioning

- Multi-Task & Reinforcement Learning for Entailment+Saliency Knowledge/Control in NLG (Video Captioning, Document Summarization, and Sentence Simplification)

Ground truth: A woman is slicing a red pepper.
SotA Baseline: A woman is slicing a carrot.
Our model: A woman is slicing a pepper.

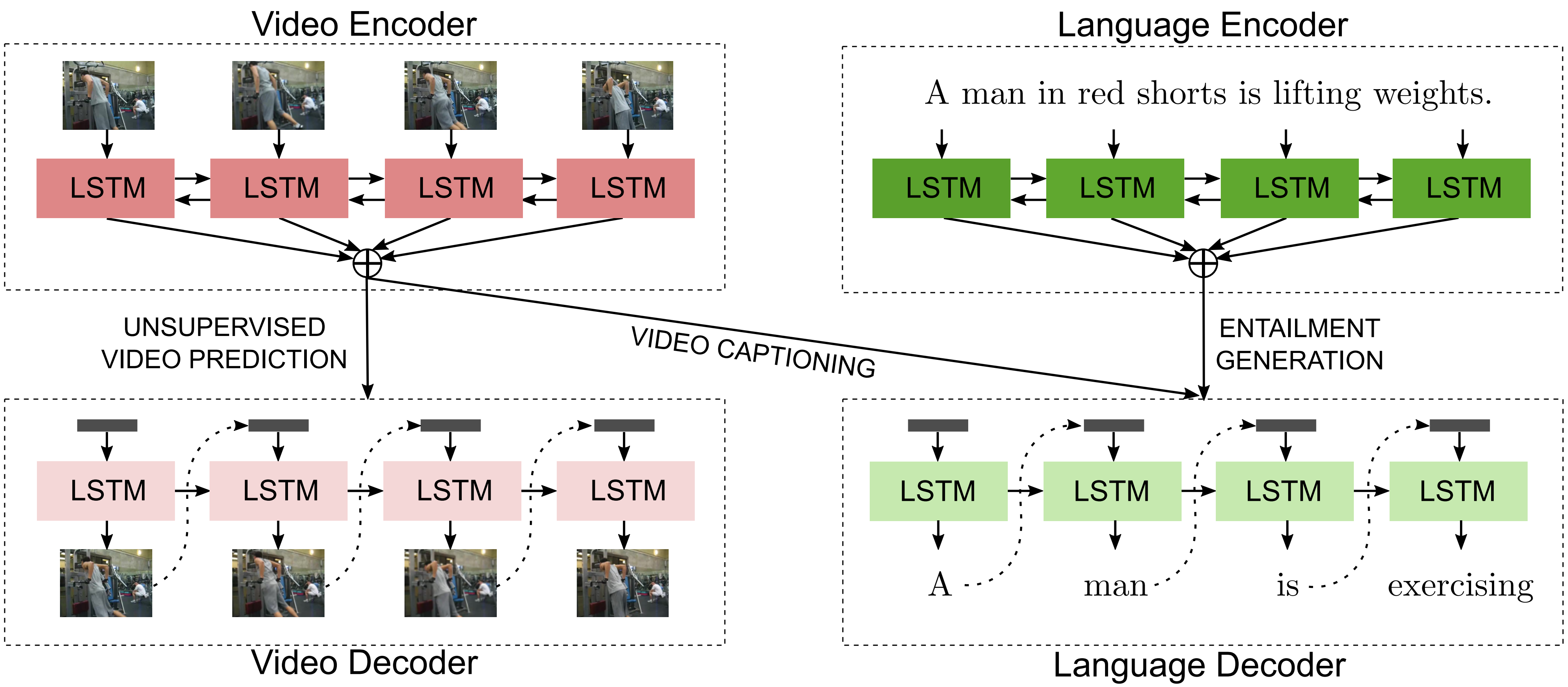
Ground truth: A group of boys are fighting.
SotA Baseline: A group of men are dancing.
Our model: Two men are fighting.

Document: *top activists arrested after last month 's anti-government rioting are in good condition , a red cross official said saturday .*
Ground-truth: *arrested activists in good condition says red cross*
SotA Baseline: *red cross says it is good condition after riots*
Our model: *red cross says detained activists in good condition*

Document: *canada 's prime minister has dined on seal meat in a gesture of support for the sealing industry .*
Ground-truth: *canadian pm has seal meat*
SotA Baseline: *canadian pm says seal meat is a matter of support*
Our model: *canada 's prime minister dines with seal meat*

Auxiliary Knowledge in Video Captioning

- Many-to-Many Multi-Task Learning for Video Captioning (with Video and Entailment Generation)



[Pasunuru and Bansal, ACL 2017 (*Outstanding Paper Award*)]

Results (YouTube2Text)

Models	METEOR	CIDEr-D	ROUGE-L	BLEU-4
PREVIOUS WORK				
LSTM-YT (Venugopalan et al., 2015b)	26.9	-	-	31.2
S2VT (Venugopalan et al., 2015a)	29.8	-	-	-
Temporal Attention (Yao et al., 2015)	29.6	51.7	-	41.9
LSTM-E (Pan et al., 2016b)	31.0	-	-	45.3
Glove + DeepFusion (Venugopalan et al., 2016)	31.4	-	-	42.1
p-RNN (Yu et al., 2016)	32.6	65.8	-	49.9
HNRE + Attention (Pan et al., 2016a)	33.9	-	-	46.7
OUR BASELINES				
Baseline (V)	31.4	63.9	68.0	43.6
Baseline (G)	31.7	64.8	68.6	44.1
Baseline (I)	33.3	75.6	69.7	46.3
Baseline + Attention (V)	32.6	72.2	69.0	47.5
Baseline + Attention (G)	33.0	69.4	68.3	44.9
Baseline + Attention (I)	33.8	77.2	70.3	49.9
Baseline + Attention (I) (E) ⊗	35.0	84.4	71.5	52.6
OUR MULTI-TASK LEARNING MODELS				
⊗ + Video Prediction (1-to-M)	35.6	88.1	72.9	54.1
⊗ + Entailment Generation (M-to-1)	35.9	88.0	72.7	54.4
⊗ + Video Prediction + Entailment Gener (M-to-M)	**36.0**	**92.4**	**72.8**	**54.5**

* All models (1-to-M, M-to-1 and M-to-M) stat. signif. better than strong SotA baseline.

Human Evaluation

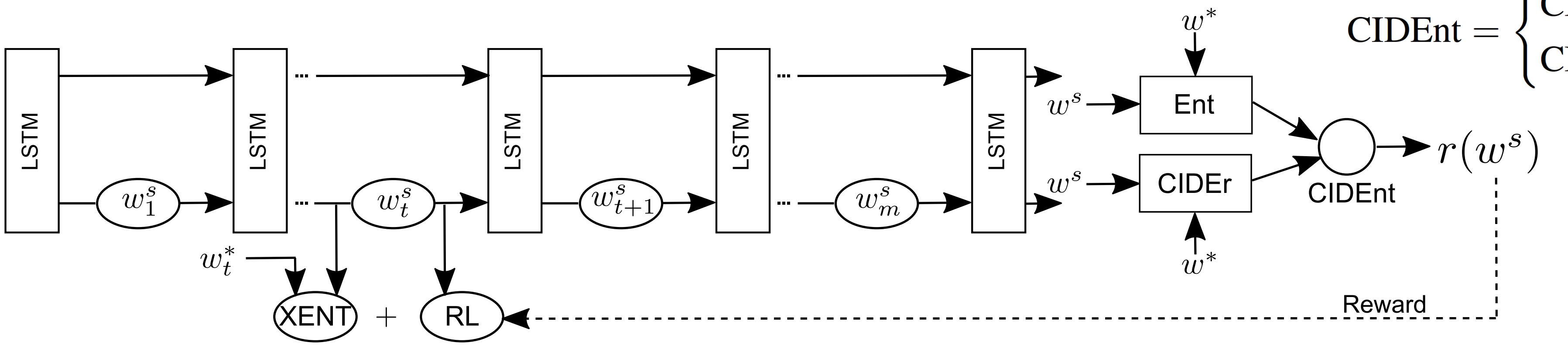
- Pilot human evaluations on 300-sized samples
- Multi-task model > strong non-multitask baseline on relevance and coherence/fluency (for both video captioning and entailment generation)

	YouTube2Text		Entailment	
	Relev.	Coher.	Relev.	Coher.
Not Distinguish.	70.7%	92.6%	84.6%	98.3%
SotA Baseline Wins	12.3%	1.7%	6.7%	0.7%
Multi-Task Wins	**17.0%**	**5.7%**	**8.7%**	**1.0%**

Auxiliary Knowledge via RL

- RL Reward = Entailment-corrected phrase-matching metrics such as CIDEr → CIDEnt

Ground-truth caption	Generated (sampled) caption	CIDEr	Ent
a man is spreading some butter in a pan	puppies is melting butter on the pan	140.5	0.07
a panda is eating some bamboo	a panda is eating some fried	256.8	0.14
a monkey pulls a dogs tail	a monkey pulls a woman	116.4	0.04
a man is cutting the meat	a man is cutting meat into potato	114.3	0.08
the dog is jumping in the snow	a dog is jumping in cucumbers	126.2	0.03
a man and a woman is swimming in the pool	a man and a whale are swimming in a pool	192.5	0.02

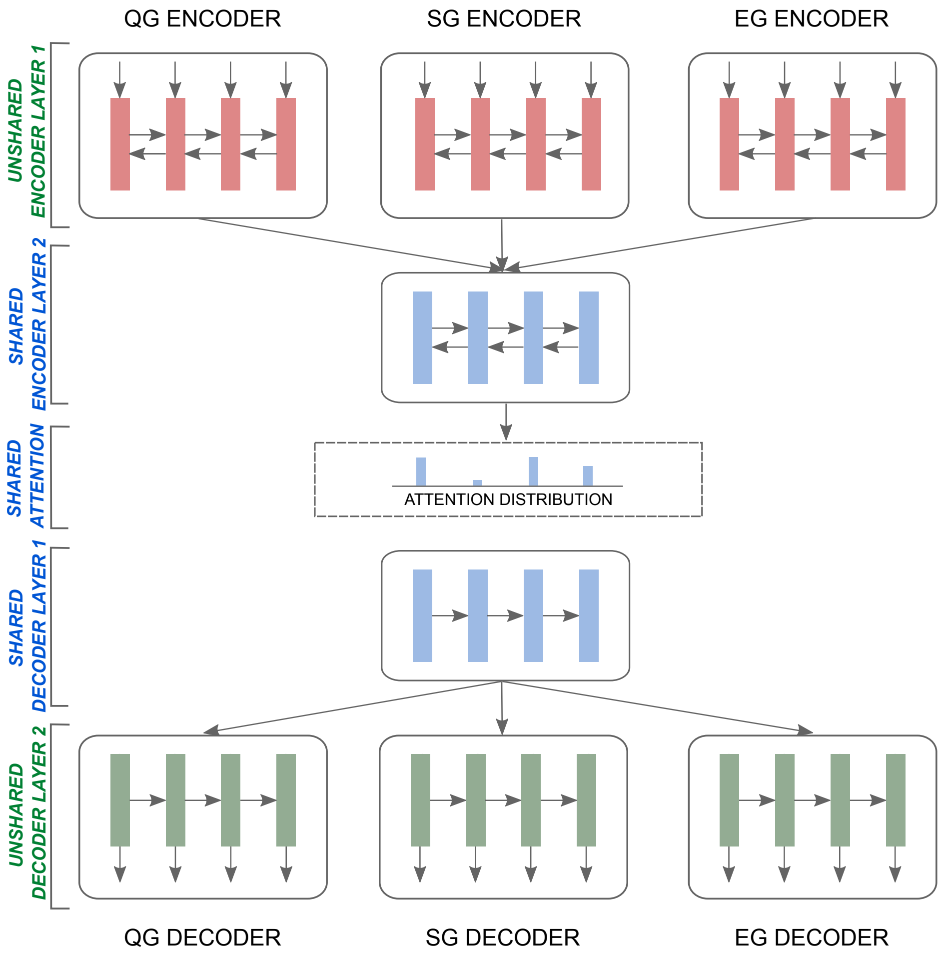


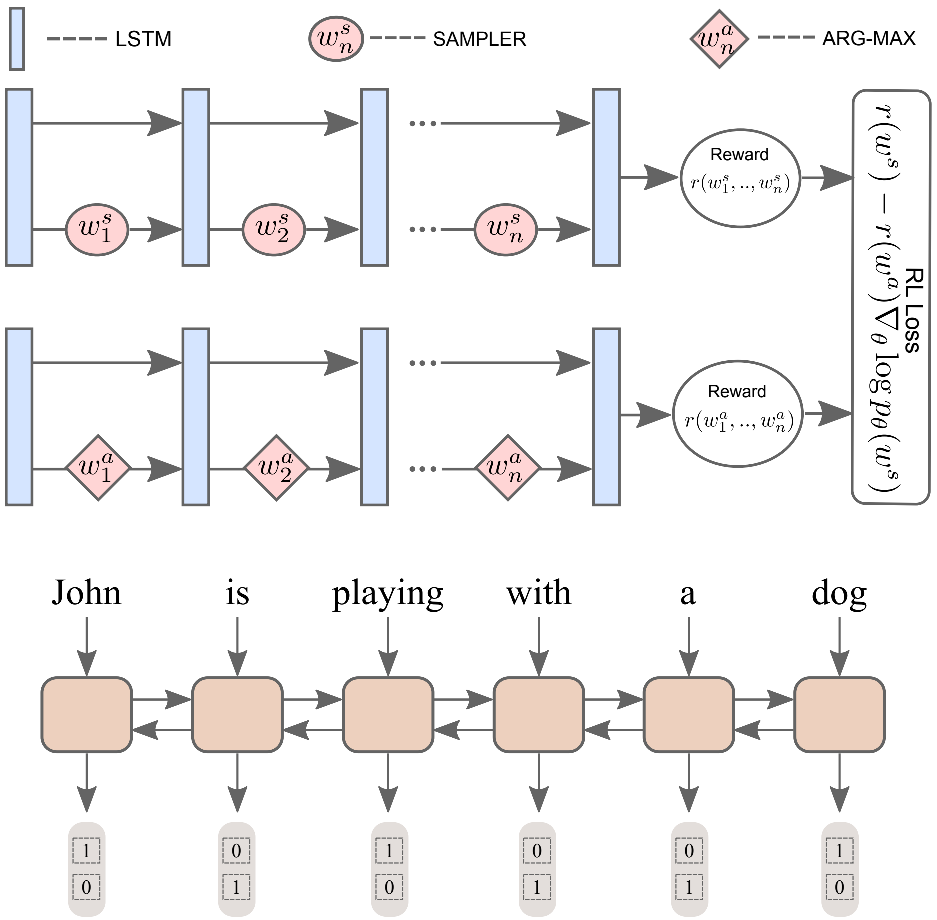
$$\text{CIDEnt} = \begin{cases} \text{CIDEr} - \lambda, & \text{if Ent} < \beta \\ \text{CIDEr}, & \text{otherwise} \end{cases}$$

[Pasunuru and Bansal, EMNLP 2017]

Auxiliary Knowledge in Language Generation

- Multi-Task & Reinforcement Learning with Entailment+Saliency Knowledge for Summarization





[Guo, Pasunuru, and Bansal, ACL 2018; Pasunuru and Bansal, NAACL 2018]

Auxiliary Knowledge in Language Generation

Input Document: celtic have written to the scottish football association in order to gain an 'understanding' of the refereeing decisions during their scottish cup semi-final defeat by inverness on sunday . the hoops were left outraged by referee steven mclean 's failure to award a penalty or red card for a clear handball in the box by josh meekings to deny leigh griffith 's goal-bound shot during the first-half . caley thistle went on to win the game 3-2 after extra-time and denied rory delia 's men the chance to secure a domestic treble this season . celtic striker leigh griffiths has a goal-bound shot blocked by the outstretched arm of josh meekings . ……after the restart for scything down marley watkins in the area . greg tansey duly converted the resulting penalty . edward ofere then put caley thistle ahead , only for john guidetti to draw level for the bhoys . with the game seemingly heading for penalties , david raven scored the winner on 117 minutes , breaking thousands of celtic hearts . celtic captain scott brown -lrb- left -rrb- protests to referee steven mclean but the handball goes unpunished . griffiths shows off his acrobatic skills during celtic 's eventual surprise defeat by inverness . celtic pair aleksandar tonev -lrb- left -rrb- and john guidetti look dejected as their hopes of a domestic treble end .

Ground-truth Summary: celtic were defeated 3-2 after extra-time in the scottish cup semi-final . leigh griffiths had a goal-bound shot blocked by a clear handball. however, no action was taken against offender josh meekings. the hoops have written the sfa for an 'understanding' of the decision .

See et al. (2017): **john hartson** was once on the end of a major **hampden injustice** while playing for celtic . but he can not see any point in his old club writing to the scottish football association over the latest controversy at the national stadium . hartson had a goal wrongly disallowed for offside while celtic were leading 1-0 at the time but went on to lose 3-2 .

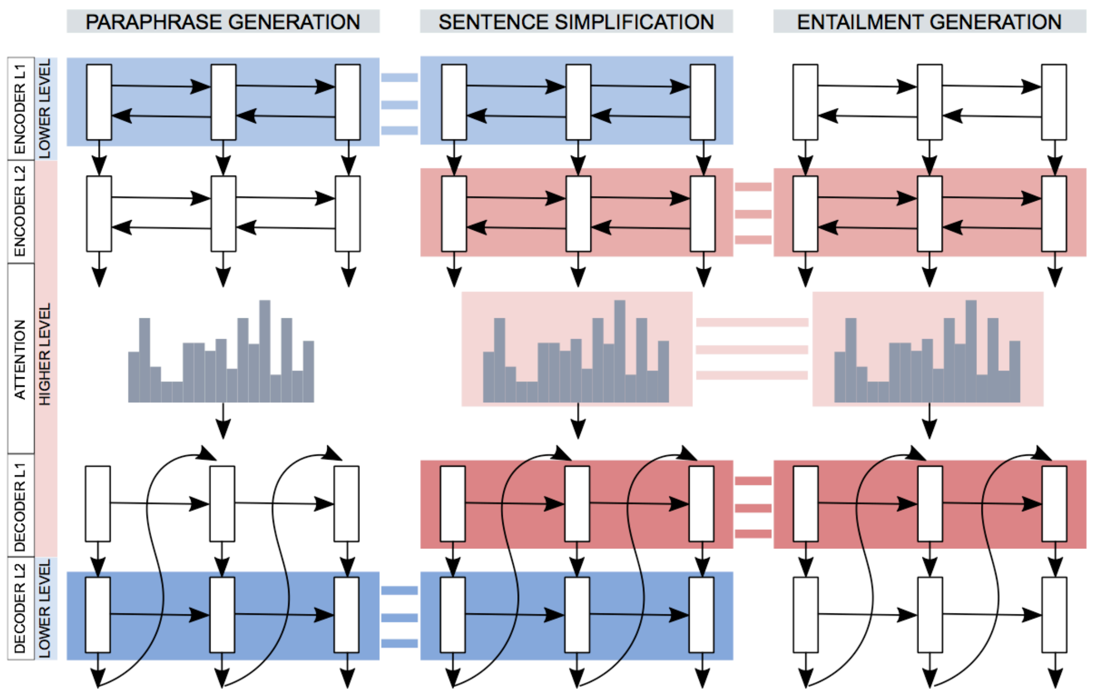
Our Baseline: **john hartson** scored the late winner in 3-2 win against celtic . celtic were leading 1-0 at the time but went on to lose 3-2 . some fans have questioned how referee steven mclean and **additional assistant alan muir** could have missed the infringement .

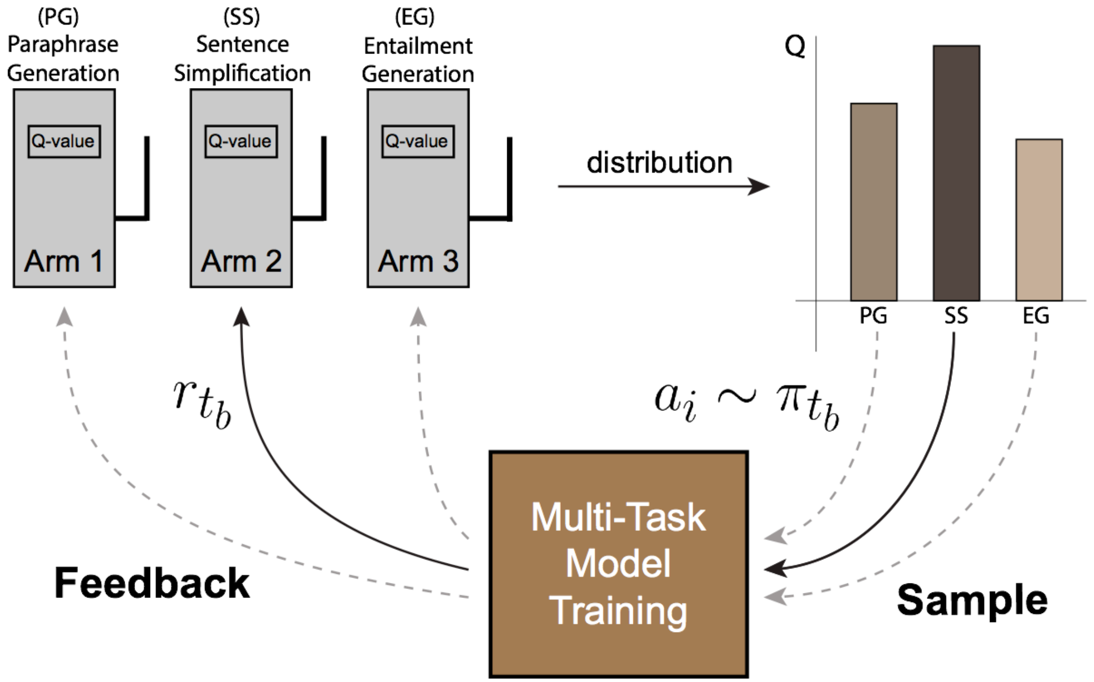
Our Multi-task Summary: celtic have written to the scottish football association in order to gain an ' understanding ' of the refereeing decisions . the hoops were left outraged by referee steven mclean 's failure to award a penalty or red card for a clear handball in the box by josh meekings . celtic striker leigh griffiths has a goal-bound shot blocked by the outstretched arm of josh meekings .

[Guo, Pasunuru, and Bansal, ACL 2018; Pasunuru and Bansal, NAACL 2018]

Auxiliary Knowledge in Language Generation

- Dynamic-Curriculum MTL with Entailment+Paraphrase Knowledge for Sentence Simplification

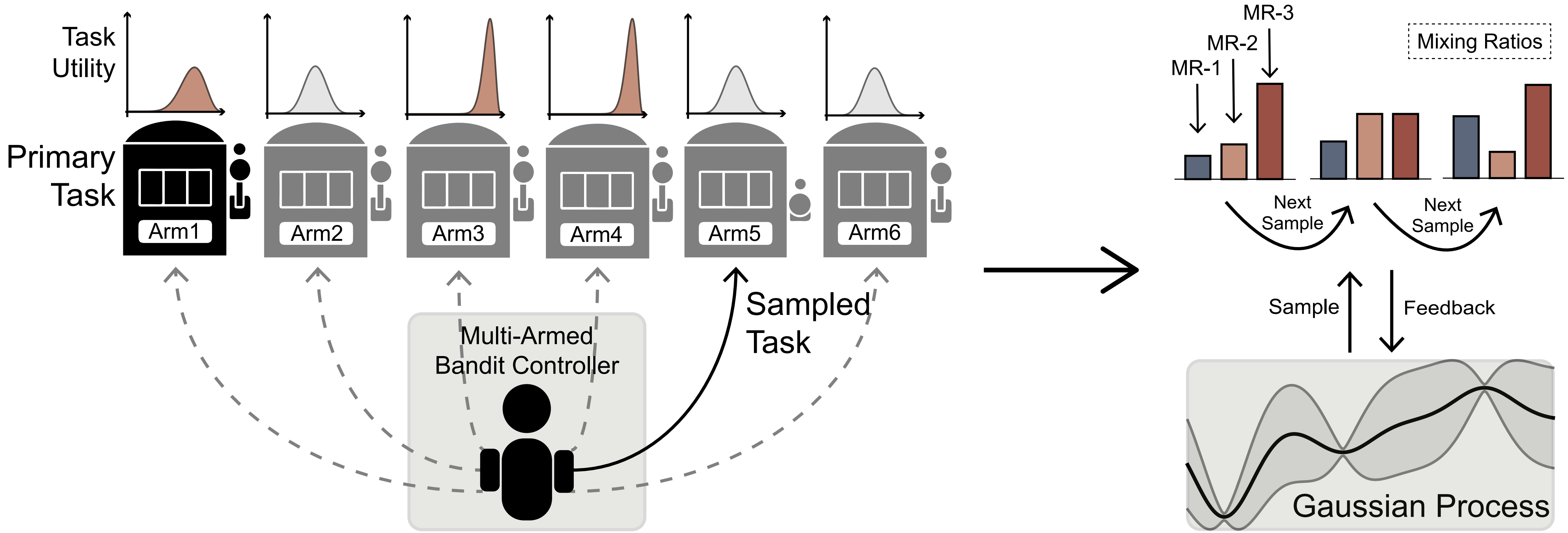




Code: https://github.com/HanGuo97/MultitaskSimplification

[Guo, Pasunuru, and Bansal, COLING 2018 (*Area Chair Favorites*)]

AutoSeM: Automatic Auxiliary Task Selection+Mixing



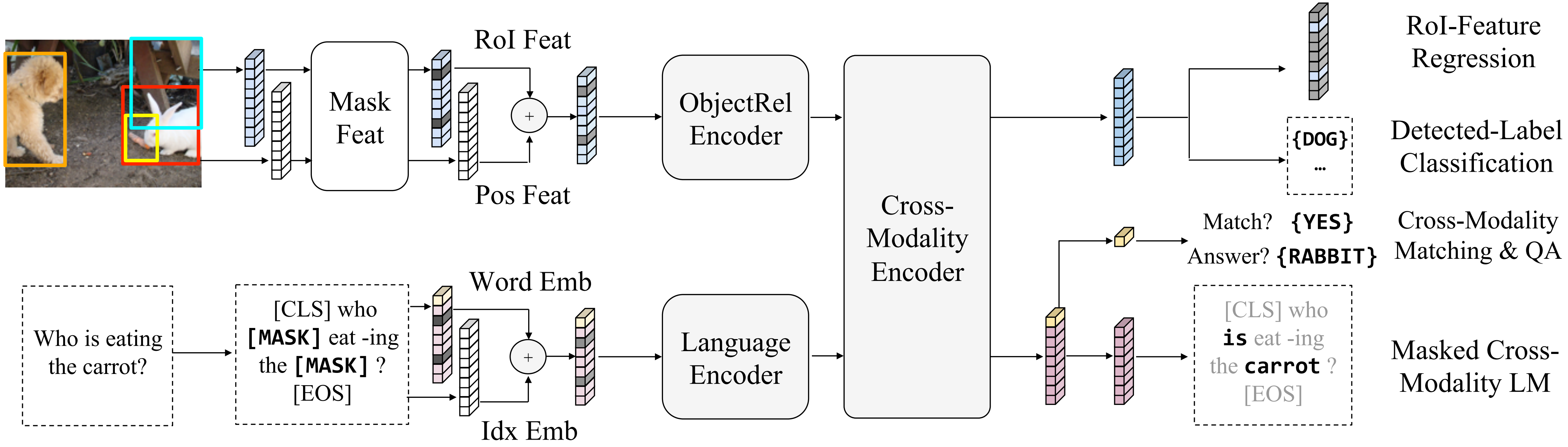
Left: the multi-armed bandit controller used for task selection, where each arm represents a candidate auxiliary task. The agent iteratively pulls an arm, observes a reward, updates its estimates of the arm parameters, and samples the next arm. Right: the Gaussian Process controller used for automatic mixing ratio (MR) learning. The GP controller sequentially makes a choice of mixing ratio, observes a reward, updates its estimates, and selects the next mixing ratio to try, based on the full history of past observations.

Code: https://github.com/HanGuo97/AutoSeM

[Guo, Pasunuru, and Bansal, NAACL 2019]

Large-Scale XModal Pretraining MTL Knowledge: LXMERT

- LXMERT brings in external knowledge on text, vision and cross-modal matching sides for MTL (as pretraining tasks in MTL setup): vision-lang transformers with 3 encoders: (object relations, language, cross-modal) & 5 pretraining tasks: masked-LM, masked-Object-Prediction (feature regression+label classification), cross-modality matching, image-QA (SotA on several vision-language tasks!)



[Tan and Bansal, EMNLP 2019]

Large-Scale XModal Pretraining MTL Knowledge: LXMERT

NLVR2 Leaderboard

NLVR2 presents the task of determining whether a natural language sentence is true about a pair of photographs.

Rank	Model	Dev. (Acc)	Test-P (Acc)	Test-U (Acc)	Test-U (Cons)
	Human Performance *Cornell University* (Suhr et al. 2019)	96.2	96.3	96.1	-
1 Aug 20, 2019	LXMERT *UNC* (Tan and Bansal 2019)	**74.9**	**74.5**	**76.2**	**42.1**
2 Aug 11, 2019	VisualBERT *UCLA & AI2 & PKU* (Li et al. 2019)	67.4	67.0	67.3	26.9
3 Nov 1, 2018	MaxEnt *Cornell University* (Suhr et al. 2019)	54.1	54.8	53.5	12.0
4 Nov 1, 2018	CNN+RNN *Cornell University* (Suhr et al. 2019)	53.			
5 Nov 1, 2018	FiLM *MILA, ran by Cornell University* (Perez et al. 2018)	5			
6 Nov 1, 2018	Image Only (CNN) *Cornell University* (Suhr et al. 2019)	51.			
7 Nov 1, 2018	N2NMN, policy search from scratch *UC Berkeley, ran by Cornell University* (Hu et al. 2017)	51.			

VQA	Yes/No	Number	Other	Overall
MIL@HDU[11]	90.33	58.91	65.91	75.26
MSM@MSRA[15]	89.74	59.01	65.89	75.01
LXRT[13]	89.33	57.29	65.32	74.38
XFZ[23]	87.86	57.87	64.3	73.35
AIOZ[4]	87.99	56.16	63.93	73.04
ks_vqa[32]	87.97	55.17	63.97	72.94
THEQS[20]	87.99	54.67	63.87	72.84
MS D365 AI[14]	87.87	56.23	63.08	72.59
HappyTeam[12]	87.95	54.49	62.21	72.03
BAN[5]	86.86	54.92	62.36	71.69
fm[28]	86.44	54.47	62.7	71.62
vqa_team_tsuh[36]	86.5	54.2	62.7	71.62
		53.7	62.62	71.52
		53.7	62.62	71.51
		53.6	62.52	71.39
		53.52	62.49	71.39
		53.66	62.49	71.37
		53.51	62.51	71.36
		54.76	61.63	71.15
		52.89	62.12	70.98

VizWiz

	Participant team	yes/no	number	other
1	LXMERT (LXR955, No Ensemble)	74.00	24.76	39.00
2	fw_vqa_	71.18	26.90	37.60
3	Gail-VisQA-Ultra (B-Ultra)	68.12	28.81	35.41
4	BAN (BAN)	68.12	17.86	31.50
5	ss	58.91	20.24	26.97
6	hdhs	56.61	22.14	26.83
7	Colin	59.85	20.24	23.70

Large-Scale XModal Pretraining MTL Knowledge: LXMERT

NLVR² Leaderboard

NLVR² presents the task of determining whether a natural language sentence is true about a pair of photographs.

Rank	Model	Dev. (Acc)	Test-P (Acc)	Test-U (Acc)	Test-U (Cons)
	Human Performance *Cornell University* (Suhr et al. 2019)	96.2	96.3	96.1	-
1 Aug 20, 2019	LXMERT *UNC* (Tan and Bansal 2019)	**74.9**	**74.5**	**76.2**	**42.1**
2 Aug 11, 2019	VisualBERT *UCLA & AI2 & PKU* (Li et al. 2019)	67.4	67.0	67.3	26.9
3 Nov 1, 2018	MaxEnt *Cornell University* (Suhr et al. 2019)	54.1	54.8	53.5	12.0
4 Nov 1, 2018	CNN+RNN *Cornell University* (Suhr et al. 2019)	53.			
5 Nov 1, 2018	FiLM *MILA, ran by Cornell University* (Perez et al. 2018)	5			
6 Nov 1, 2018	Image Only (CNN) *Cornell University* (Suhr et al. 2019)	51.			
7 Nov 1, 2018	N2NMN, policy search from scratch *UC Berkeley, ran by Cornell University* (Hu et al. 2017)	51.			

VQA

	By Answer Type Yes/No	Number	Other	Overall
MIL@HDU[11]	90.33	58.91	65.91	75.26
MSM@MSRA[15]	89.74	59.01	65.89	75.01
LXRT[13]	89.33	57.29	65.32	74.38
XFZ[23]	87.86	57.87	64.3	73.35
AIOZ[4]	87.99	56.16	63.93	73.04
ks_vqa[32]	87.97	55.17	63.97	72.94
THEQS[20]	87.99	54.67	63.87	72.84
MS D365 AI[14]	87.87	56.23	63.08	72.59
HappyTeam[12]	87.95	54.49	62.21	72.03
BAN[5]	86.86	54.92	62.36	71.69
fm[28]	86.44	54.47	62.7	71.62
vqa_team_tsuh[36]	86.5	54.2	62.7	71.62
		53.7	62.62	71.52
		53.7	62.62	71.51
		53.6	62.52	71.39
		53.52	62.49	71.39
		53.66	62.49	71.37
		53.51	62.51	71.36

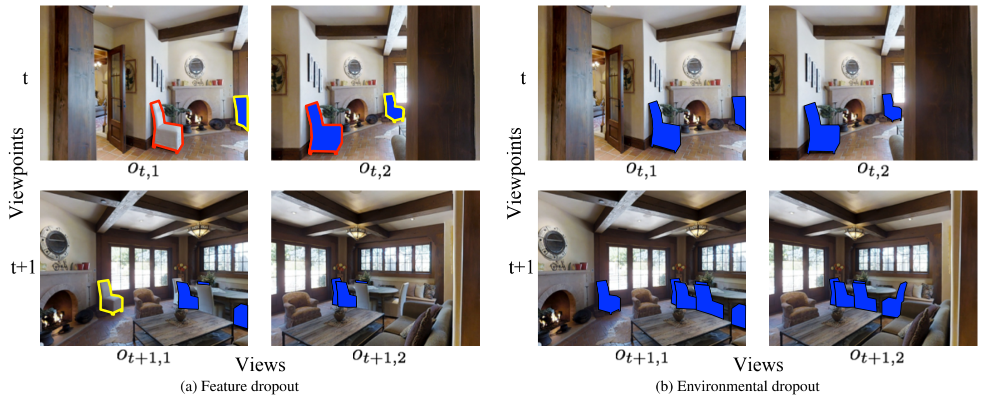
VizWiz

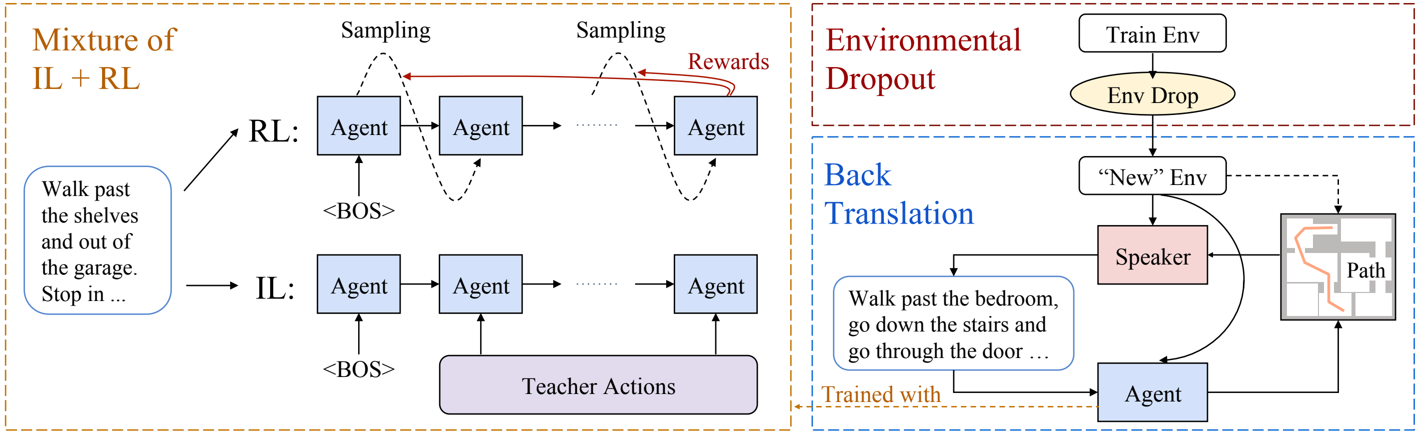
	Participant team	yes/no	number	other
1	LXMERT (LXR955, No Ensemble)	74.00	24.76	39.00
2	fw_vqa_	71.18	26.90	37.60
3	Gail-VisQA-Ultra (B-Ultra)	68.12	28.81	35.41
4	BAN (BAN)	68.12	17.86	31.50
5	ss			
6	hdhs			
7	Colin			

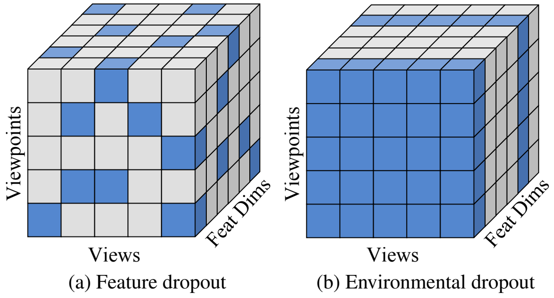
See Hao's talk on Nov7 1.30pm! (with several visualizations and ablations and challenges)!

Spatial Navigation w/ Generalizable Knowledge

- Learning to Navigate Unseen Environments: Back Translation with Environmental Dropout (to create new rooms with view and viewpoint consistency; generate instructions for new rooms; use generated room-instruction data in semi-supervised setup)







[Tan, Yu, Bansal. NAACL 2019]

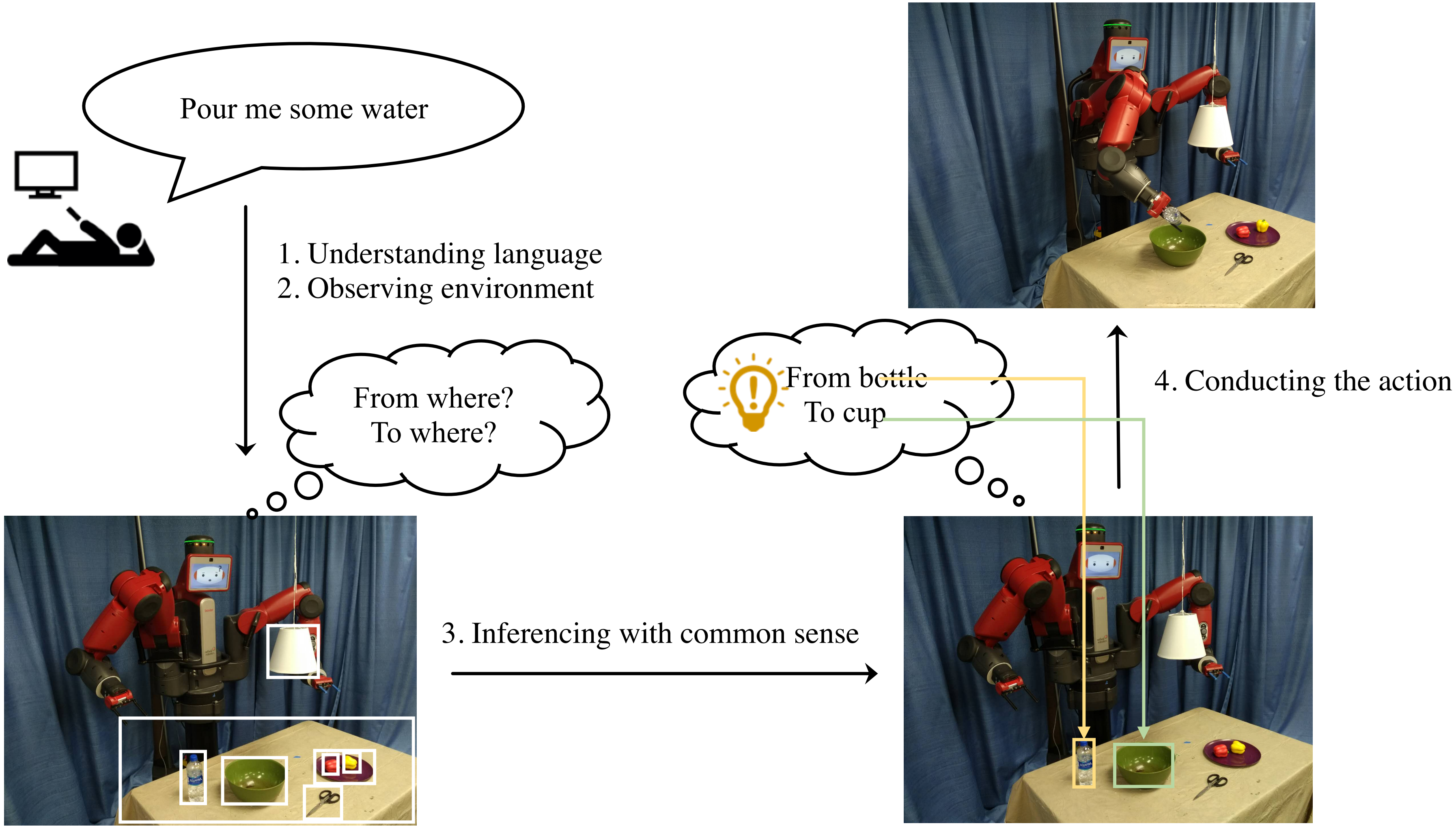
Spatial Navigation w/ Generalizable Knowledge

EvalAI All Challenges Forum Sign Up Log In

Rank	Participant team	length	error	oracle success	success	spl	Last submission at
1	human	11.85	1.61	0.90	0.86	0.76	1 year ago
2	Back Translation with Environmental Dropout (with Beam Search) (null)	686.82	3.26	0.99	0.69	0.01	10 months ago
3	vBot (Greedy)	10.24	3.76	0.71	0.65	0.62	3 months ago
4	Back Translation with Environmental Dropout (exploring unseen environments before testing)	9.79	3.97	0.70	0.64	0.61	10 months ago
5	Reinforced Cross-Modal Matching (optimized for SR; with beam search)	357.62	4.03	0.96	0.63	0.02	10 months ago
6	sjtu_test (null)	1,228.45	3.98	0.97	0.62	0.01	10 months ago
7	Self-Monitoring Navigation Agent (with beam search) (Self-Aware Co-Grounded Model)	373.09	4.48	0.97	0.61	0.02	1 year ago
8	Tactical Rewind - long	196.53	4.29	0.90	0.61	0.03	9 months ago
9	Reinforced Cross-Modal Matching + SIL (exploring unseen environments before testing) (SIL-R2)	9.48	4.21	0.67	0.60	0.59	10 months ago
10	AAEI-Agent	13.16	4.61	0.65	0.57	0.50	2 months ago
11	test-sf	10.99	4.57	0.65	0.57	0.50	5 months ago
12	PreSS (Greedy)	10.52	4.53	0.63	0.57	0.53	4 months ago
13	tourist (null)	1,214.94	4.57	0.96	0.56	0.01	11 months ago
14	Tactical Rewind - short	22.08	5.14	0.64	0.54	0.41	10 months ago
15	Speaker-Follower (optimized for success rate) (Speaker-Follower)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
16	Kjtest-sp	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
17	licr19	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

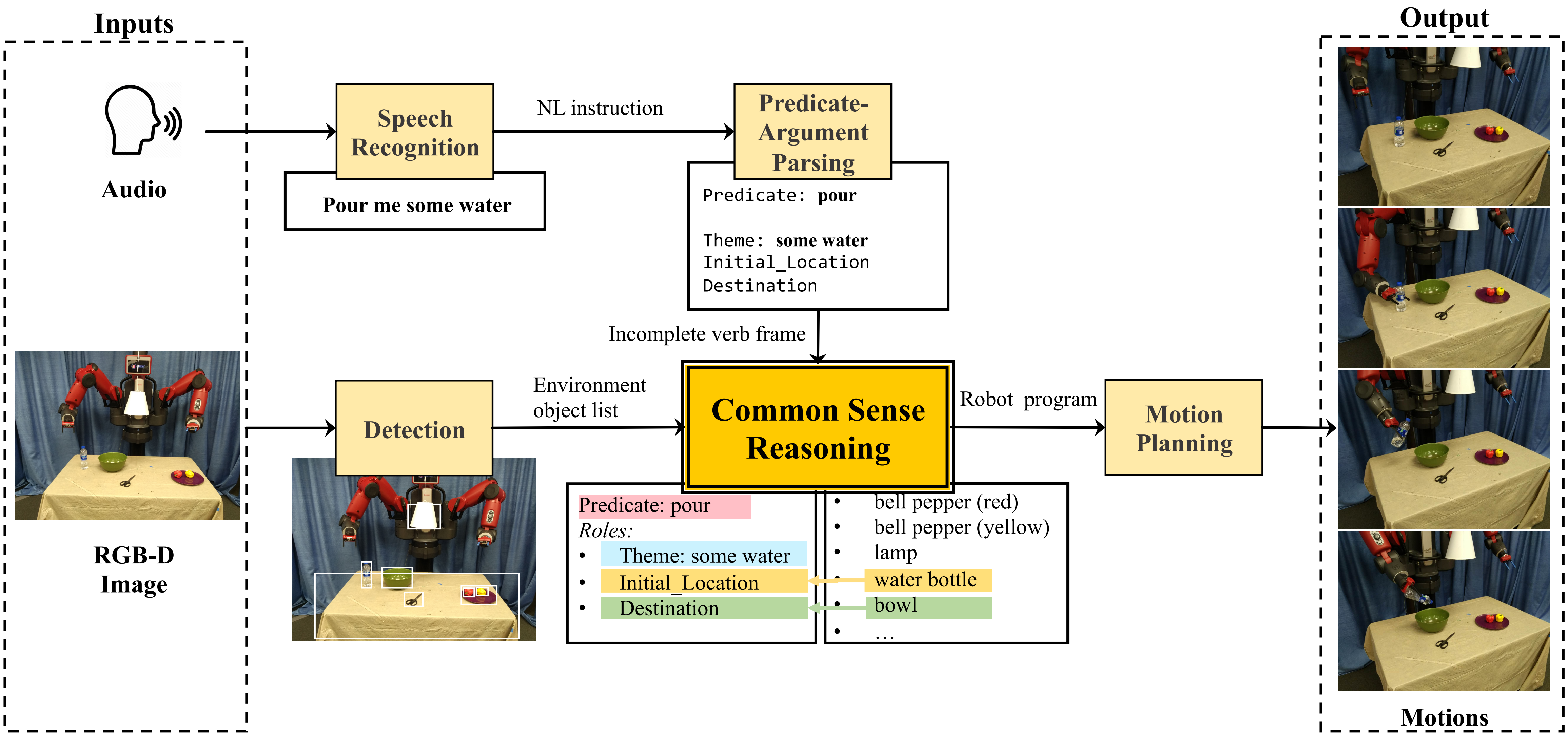
Still several challenges/ long way to go, e.g., better object detectors, diverse language, etc.!

Commonsense in Robotic Instructions



[Chen et al., 2019]

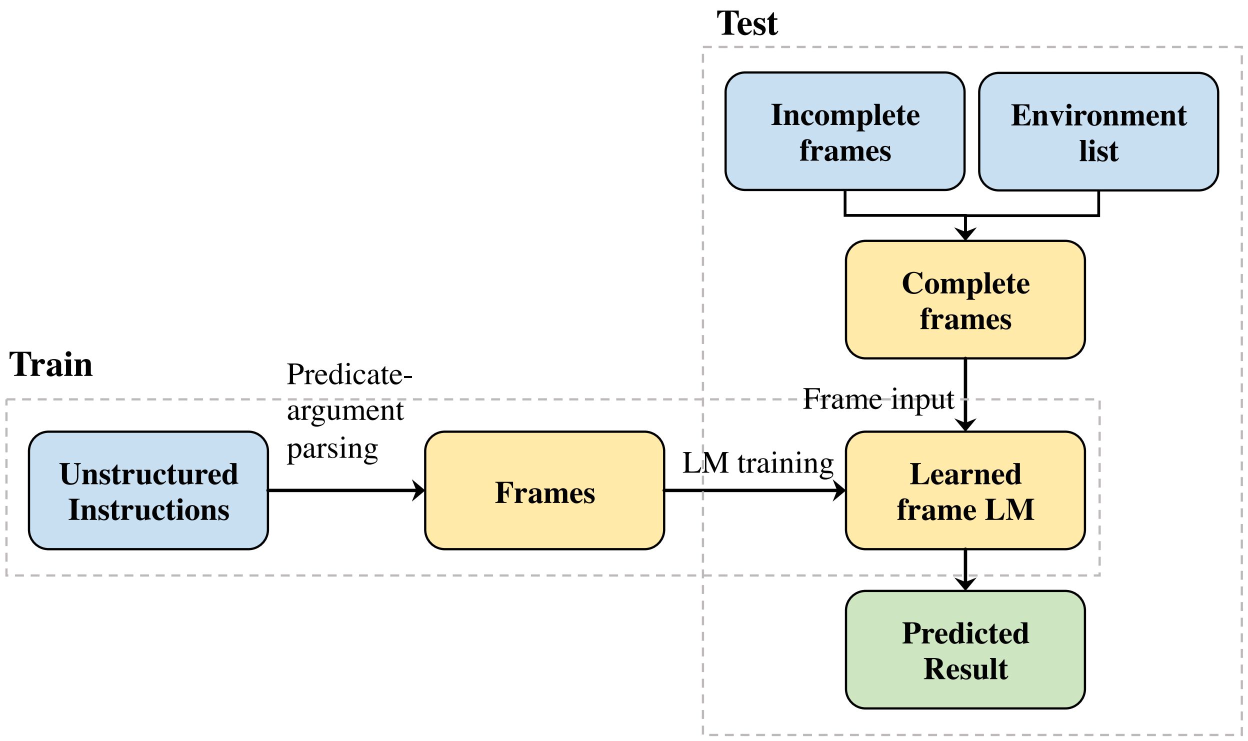
Commonsense in Robotic Instructions



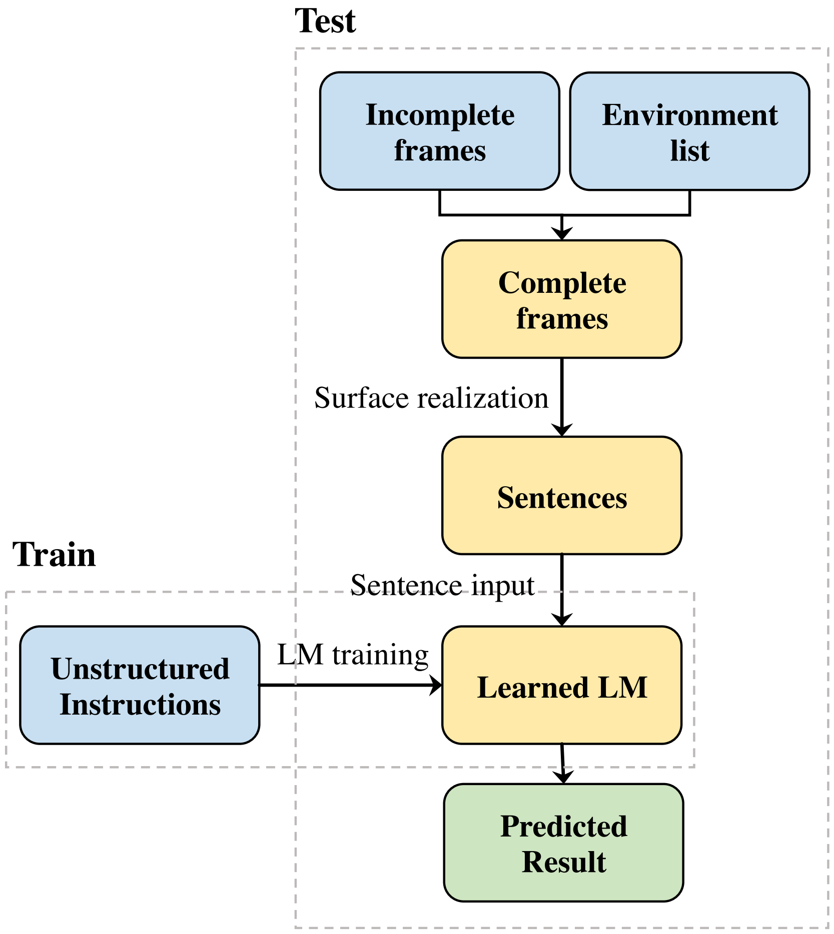
[Chen et al., 2019]

Commonsense in Robotic Instructions

Frame LM v.s. sentence LM



Frame LM



Sentence LM

[Chen et al., 2019]

Commonsense in Robotic Instructions

https://drive.google.com/file/d/1C9xsuyW1bVBzLimvVFbBfOcKCzV5ueHs/view

[Chen et al., 2019]

Commonsense in Robotic Instructions

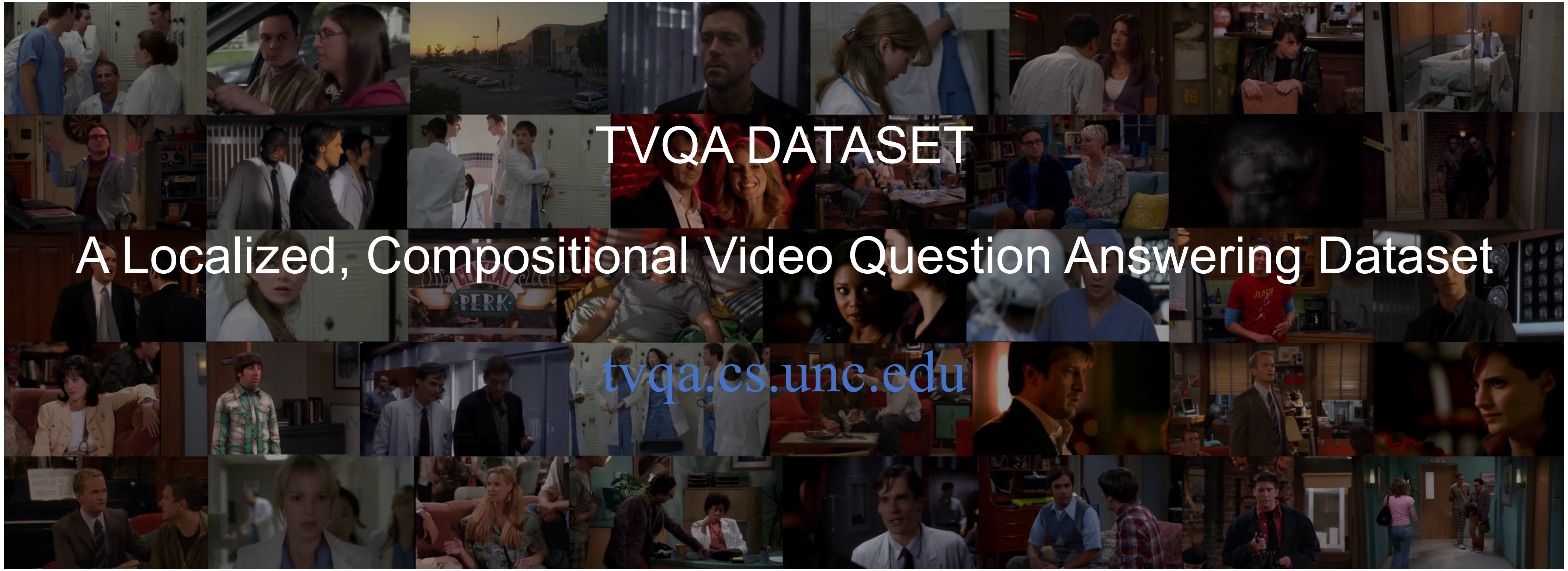
Still several challenges/ long way to go, e.g., longer ambiguities and more structured knowledge for robotic tasks!

https://drive.google.com/file/d/1C9xsuyW1bVBzLimvVFbBfOcKCzV5ueHs/view

[Chen et al., 2019]

Video-based Dynamic Context and Spatial-Temporal Localization

TVQA (videos with audio and subtitles)



Large-Scale

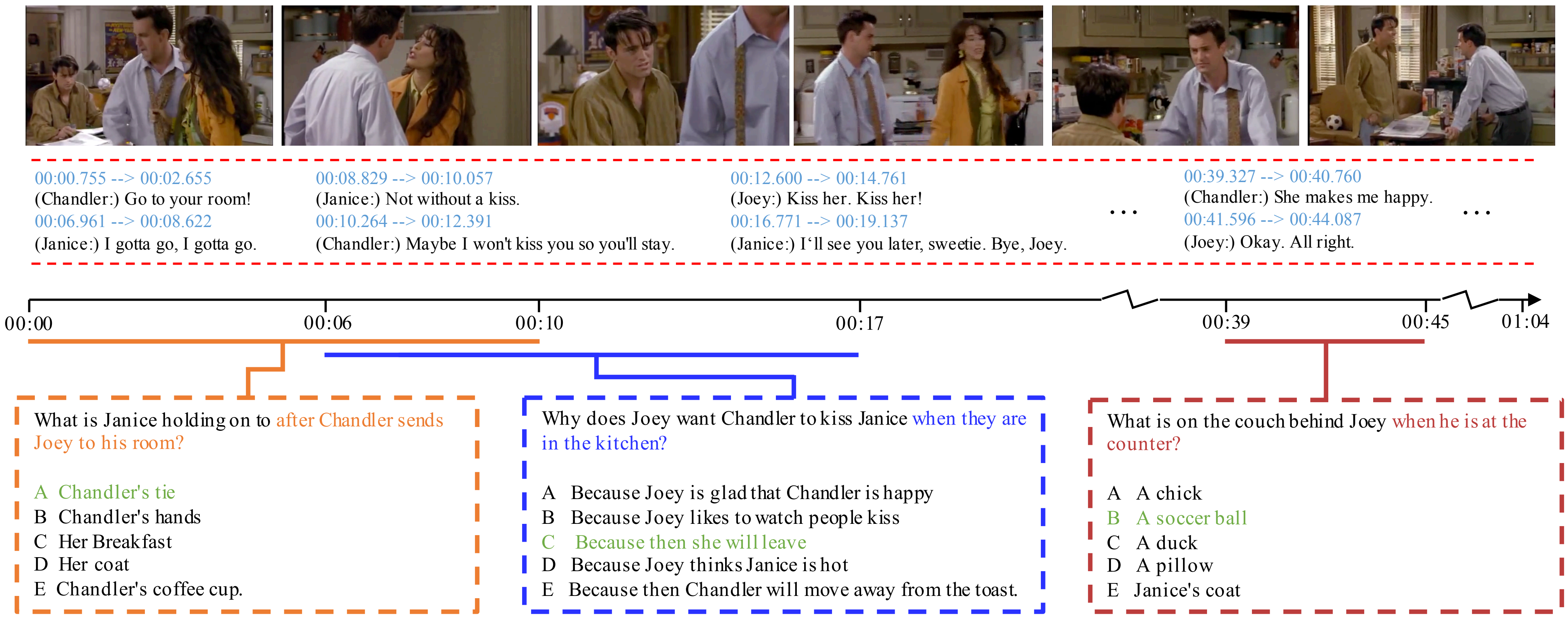
Compositional

Localized

Fun!

TVQA (videos with audio and subtitles)

- Largest video-QA dataset with 6 video categories/genres, videos+subtitles QA, compositional, spatio-temporal localization (timestamps + bounding boxes)



[Lei et al., EMNLP 2018, ArXiv 2019]

TVQA Compositionality (Localization + VQA)

0s 62s

Write a question:

[What/Why/...] ___ [when/before/after] ___

Question + Localization

What is Sheldon holding when
he is talking to Howard about swords?

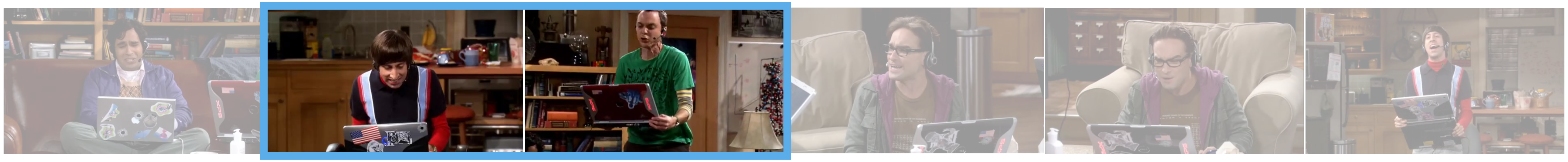
TVQA Data Collection

What is Sheldon holding when
he is talking to Howard about swords?

A computer

1) A comic book
2) A sword
3) A toy train
4) A drink

Mark the START and END timestamps:

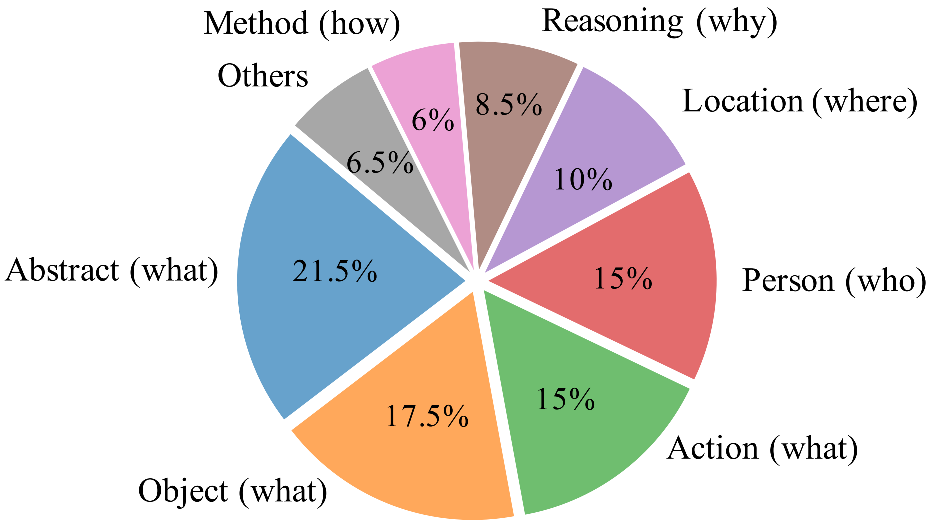


2s | ⟷ | 10s

START END

TVQA Data Statistics

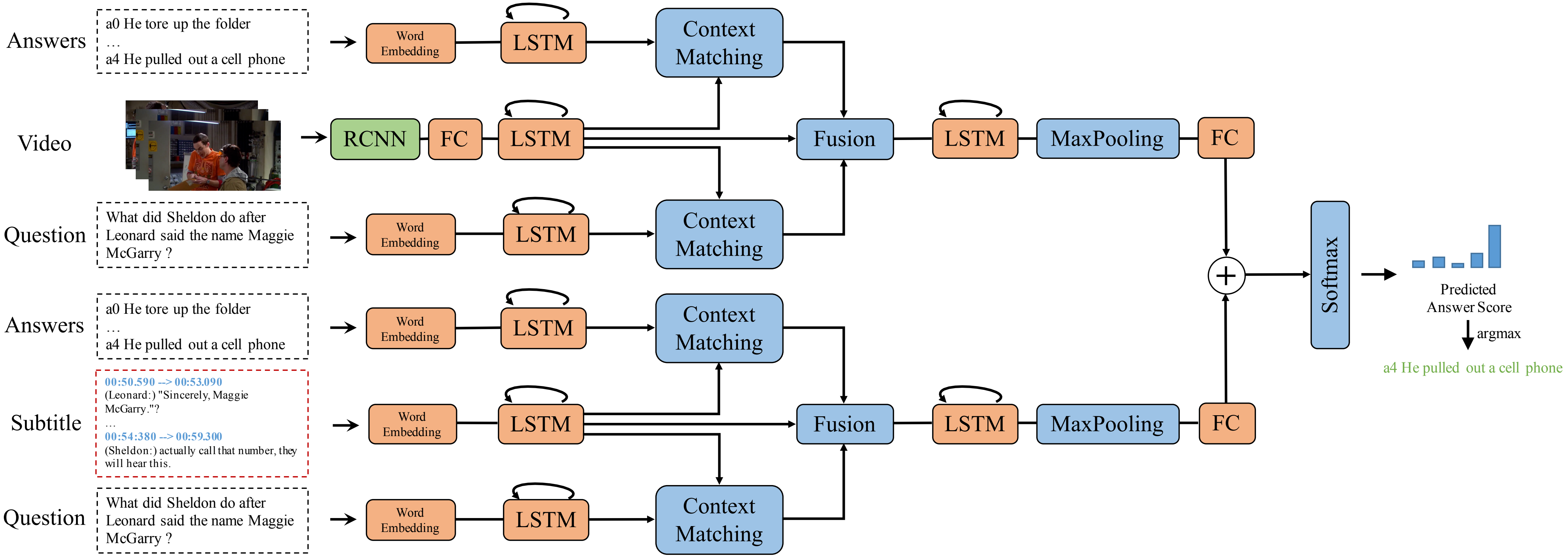
Show	Genre	#Sea.	#Epi.	#Clip	#QA
BBT	sitcom	10	220	4,198	29,384
Friends	sitcom	10	226	5,337	37,357
HIMYM	sitcom	5	72	1,512	10,584
Grey	medical	3	58	1,427	9,989
House	medical	8	176	4,621	32,345
Castle	crime	8	173	4,698	32,886
Total	—	44	925	21,793	152,545



Dataset	V. Src.	QType	#Clips / #QAs	Avg. Len.(s)	Total Len.(h)	Q. Src. text	Q. Src. video	Timestamp annotation
MovieFIB	Movie	OE	118.5k / 349k	4.1	135	✓	-	-
Movie-QA	Movie	MC	6.8k / 6.5k	202.7	381	✓	-	✓
TGIF-QA	Tumblr	OE&MC	71.7k / 165.2k	3.1	61.8	✓	✓	-
Pororo-QA	Cartoon	MC	16.1k / 8.9k	1.4	6.3	✓	✓	-
TVQA (our)	TV show	MC	21.8k / 152.5k	76.2	461.2	✓	✓	✓

TVQA Models

Multiple streams (video, subtitle), each stream deals with different contextual input



TVQA Results

	Method	Video Feature	Test Accuracy w/o ts	Test Accuracy w/ ts
0	Random	-	20.00	20.00
1	Longest Answer	-	30.41	30.41
2	Retrieval-Glove	-	22.48	22.48
3	Retrieval-SkipThought	-	24.24	24.24
4	Retrieval-TFIDF	-	20.88	20.88
5	NNS-Glove Q	-	22.40	22.40
6	NNS-SkipThought Q	-	23.79	23.79
7	NNS-TFIDF Q	-	20.33	20.33
8	NNS-Glove S	-	23.73	29.66
9	NNS-SkipThought S	-	26.81	37.87
10	NNS-TFIDF S	-	49.94	51.23
11	Our Q	-	43.34	43.34
12	Our V+Q	img	42.67	43.69
13	Our V+Q	reg	42.75	44.85
14	Our V+Q	cpt	43.38	45.41
15	Our S+Q	-	63.14	66.23
16	Our S+V+Q	img	63.57	66.97
17	Our S+V+Q	reg	63.19	67.82
18	Our S+V+Q	cpt	**65.46**	**68.60**

Question only

Add Video

Add Subtitle

Add Video, Subtitle

Accuracy for different methods on TVQA test set. Q = Question, S = Subtitle, V = Video, img = ImageNet features, reg = regional visual features, cpt = visual concept features, ts = timestamp annotation.

Both visual and textual information are important!

TVQA Results

	Method	Video Feature	Test Accuracy	
			w/o ts	w/ ts
0	Random	-	20.00	20.00
1	Longest Answer	-	30.41	30.41
2	Retrieval-Glove	-	22.48	22.48
3	Retrieval-SkipThought	-	24.24	24.24
4	Retrieval-TFIDF	-	20.88	20.88
5	NNS-Glove Q	-	22.40	22.40
6	NNS-SkipThought Q	-	23.79	23.79
7	NNS-TFIDF Q	-	20.33	20.33
8	NNS-Glove S	-	23.73	29.66
9	NNS-SkipThought S	-	26.81	37.87
10	NNS-TFIDF S	-	49.94	51.23
11	Our Q	-	43.34	43.34
12	Our V+Q	img	42.67	43.69
13	Our V+Q	reg	42.75	44.85
14	Our V+Q	cpt	43.38	45.41
15	Our S+Q	-	63.14	66.23
16	Our S+V+Q	img	63.57	66.97
17	Our S+V+Q	reg	63.19	67.82
18	Our S+V+Q	cpt	**65.46**	**68.60**

Accuracy for different methods on TVQA test set. Q = Question, S = Subtitle, V = Video, img = ImageNet features, reg = regional visual features, cpt = visual concept features, ts = timestamp annotation.

Timestamp information is helpful! But still several challenges/ long way to go from human performance 90%!

TVQA Leaderboard

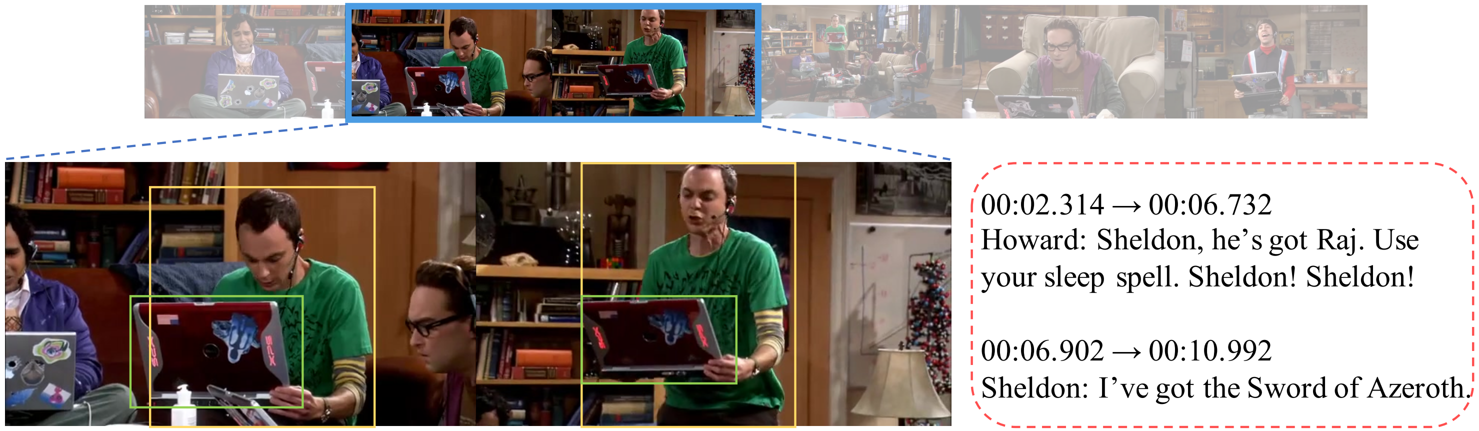
With Timestamp Annotation

Rank	Date	Model	Val	Test-Public
-	Aug 27, 2018	Human Performance	93.44	91.95
1	Mar 22, 2019	ZGF (sin		
2	Aug 27, 2018	multi-stream mo		
3	Aug 27, 2018	NNS-T		
4	Aug 27, 2018	NNS-Skip		
5	Aug 27, 2018	Longes		
6	Aug 27, 2018	Retrieval-S		
7	Aug 27, 2018	NNS-Skip		
8	Aug 27, 2018	Rar		

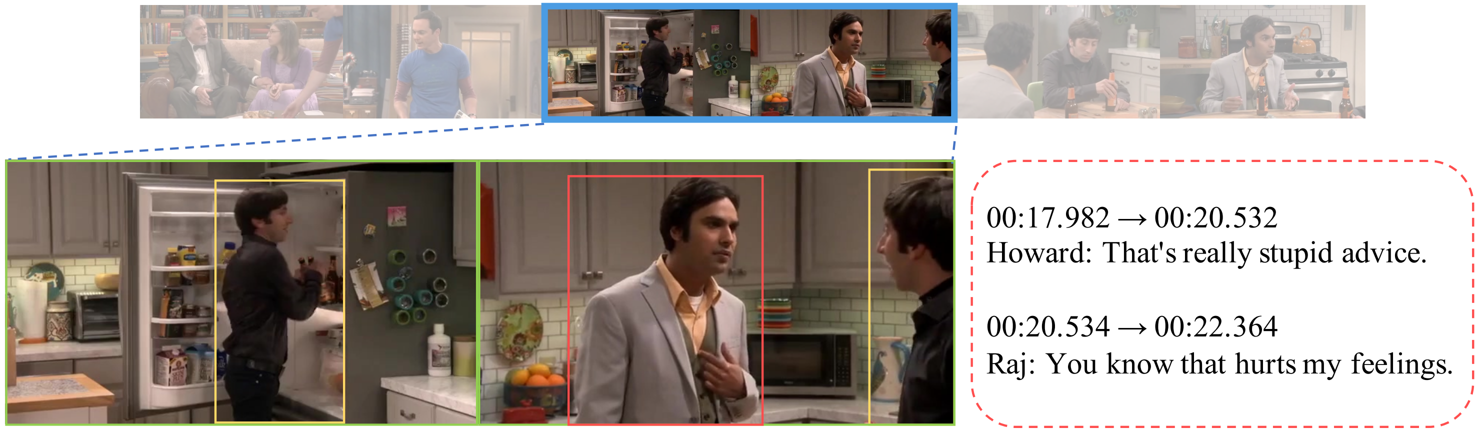
Without Timestamp Annotation

Rank	Date	Model	Val	Test-Public
-	Aug 27, 2018	Human Performance	89.61	89.41
1	Mar 22, 2019	STAGE (span) (single model)	70.50	70.23
2	Mar 22, 2019	ZGF (single model)	68.90	68.77
3	Dec 14, 2018	Multi-task learning, sub+vcpt (single model)	66.22	67.05
4	Apr 3, 2019	PAMN_subvcpt (single model)	66.38	66.77
5	Aug 27, 2018	multi-stream model (single model)	65.85	66.46
6	Aug 27, 2018	NNS-TFIDF-S	50.33	49.59
7	Aug 27, 2018	Longest Answer	29.59	30.22
8	Aug 27, 2018	NNS-SkipThought-S	27.50	26.93
9	Aug 27, 2018	Retrieval-SkipThought	22.95	24.27
10	Aug 27, 2018	NNS-SkipThought-Q	23.87	23.39
11	Aug 27, 2018	Random	20.00	20.00

TVQA+ (spatial localization: bounding box annotations)



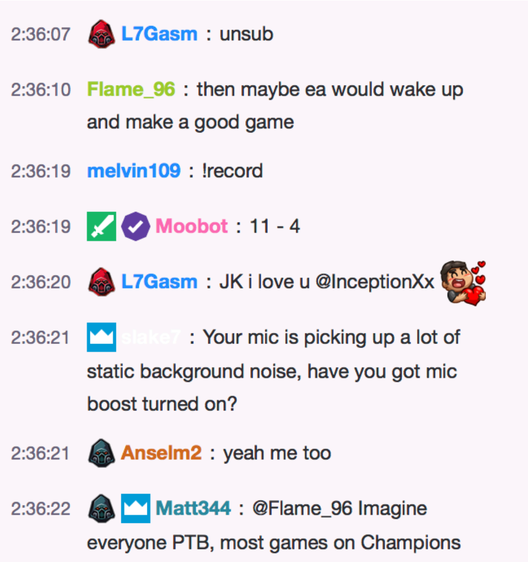
Question: What is Sheldon holding when he is talking to Howard about the sword?
Correct Answer: A computer.



Question: Who is talking to Howard when he is in the kitchen upset?
Correct Answer: Raj is talking to Howard.

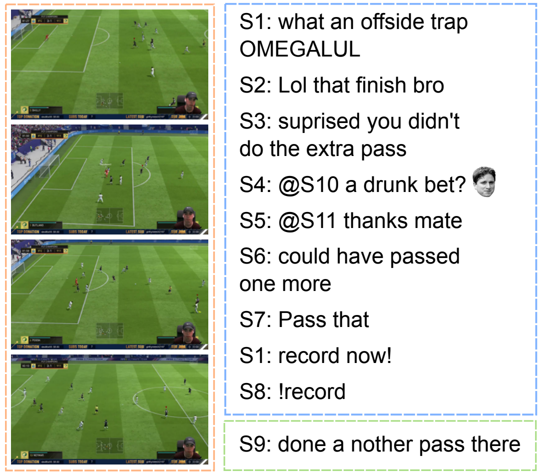
Video-based Dialogue

- Generating chat responses given both video and previous dialogue history:
- Unique Twitch language:
  - Time-constrained, not just space
  - Lots of special vocab, symbols, emoticons
  - Multi-user with several interleaving turns
  - Multi-lingual

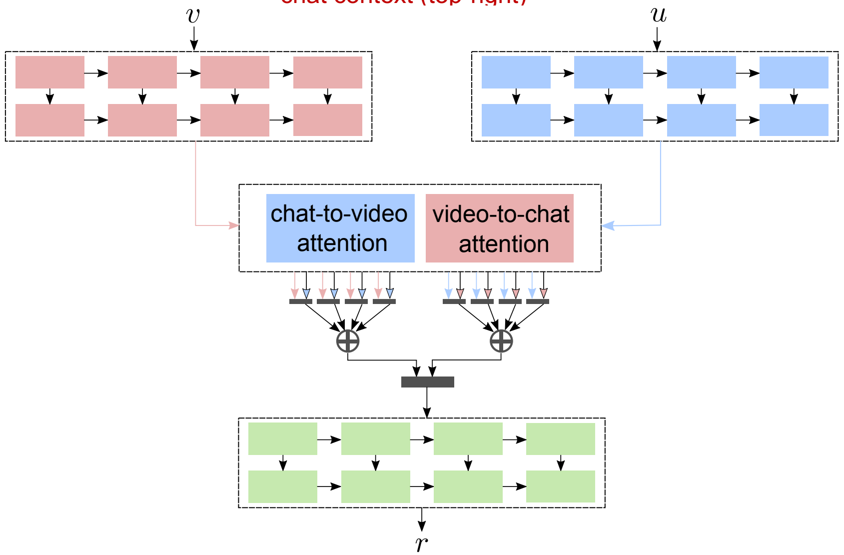


Video + Chat based Context

Multiple speakers



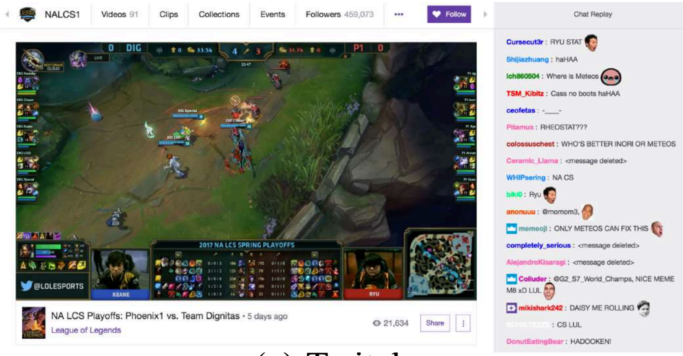
The task is to predict the response (bottom-right) using the video context (left) and the chat context (top-right)



[Pasunuru and Bansal, EMNLP 2018]

Multilingual Video Summary/Highlight Prediction

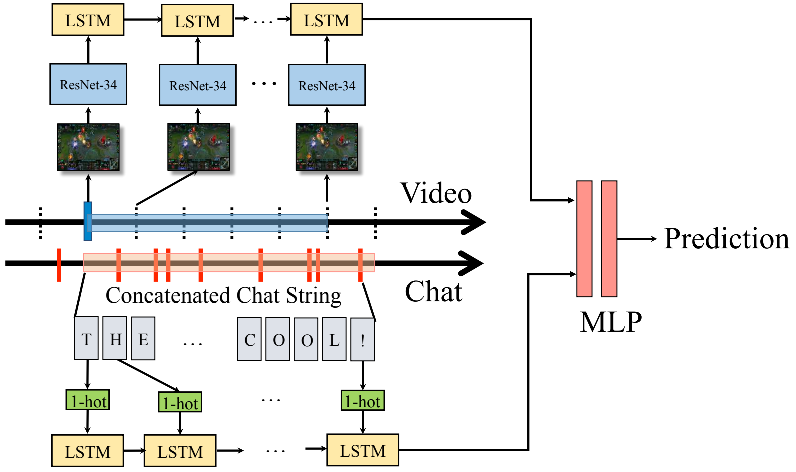
- Sports video portals offer an exciting domain for research on multimodal, multilingual analysis.
- Automatic video highlight prediction based on joint video and textual chat features from the real-world audience discourse with complex slang, in both English and Chinese.



(a) Twitch

(b) Youtube

(c) Facebook



Method	Data	NALCS	LMS
L-Char-LSTM	chat	43.2	39.7
V-CNN-LSTM	video	72.2	69.2
lv-LSTM	chat+video	**74.7**	**70.0**

Table 3: Test Results on the NALCS (English) and LMS (Traditional Chinese) datasets.

[Fu, Lee, Bansal, Berg, EMNLP 2017]

Thoughts/Challenges/Future Work

- Longer ambiguities and more structured knowledge for robotic tasks
- Strengths vs limitations of large-scale BERT/LXMERT pretraining
- Contrasting structured knowledge versus large-scale BERT/LXMERT pretraining?
- Multilingual extensions of TVQA and Video-Dialogue
- Multilingual+Multimodal LXMERT
- Adding other modalities such as speech and non-verbal cues

UNC NLP

HOME PEOPLE PUBLICATIONS SOFTWARE/DATA SPONSORS TEACHING COLLOQUIA JOIN PHOTOS






Welcome to the UNC-NLP Research Group

Our lab has research interests in statistical natural language processing and machine learning, with a focus on multimodal, grounded, and embodied semantics (i.e., language with vision and speech, for robotics), human-like language generation and Q&A/dialogue, and interpretable and structured deep learning. We are a group of PhD, MS, BS, and visiting students who work with Prof. Mohit Bansal and collaborators in the Computer Science department (lab located in Brooks Building FB-241C) at the University of North Carolina (UNC) Chapel Hill.

News

Aug 2019 Congrats to Peter Hase for the Royster Society PhD Fellowship!

Aug 2019 5 new papers in EMNLP 2019.

July 2019 Congrats to Hyounghun for ACL 2019 Best Short Paper Nomination!

July 2019 We have a Postdoc opening - please apply!

July 2019 Thanks for the NSF-CAREER Award (details).

July 2019 Thanks for the Google Focused Research Award (details).

May 2019 6 new papers in ACL 2019.

Apr 2019 Congrats to Darryl Hannan for the 3-year NSF PhD Fellowship!

Mar 2019 Congrats to Hao Tan for 1st Rank on the Room-to-Room Vision-Language-Navigation Leaderboard!

Feb 2019 5 new papers: 3 in NAACL 2019, 1 in CVPR 2019, 1 in ICRA 2019.

Jan 2019. Congrats to Ramakanth Pasunuru for being awarded the 2-year Microsoft Research PhD Fellowship!

Mar 2018. Thanks to Adobe for the Adobe Research Award.

Feb 2018. 9 new 2018 papers in NAACL, CVPR, AAAI, WACV.

Sept 2017. Thanks to DARPA for the DARPA Young Faculty Award (link).

Sept 2017. Thanks to Facebook for the Facebook ParlAI Research Award.

July 2017. 3 papers at EMNLP 2017 and 2 papers at the Summarization-Frontiers and RepEval workshops.

June 2017. Top single model results on the RepEval-NLI Shared Task at EMNLP 2017 (congrats Yixin!).

June 2017. Outstanding Paper Award at ACL 2017 (congrats Ram!).

Feb 2017. Thanks to Google for a Google Faculty Research Award (link).

Nov 2016. 3 papers on navigational instruction generation, coherent dialogue w/ attn-LMs, and on context-RNN-GAN models to appear at AAAI 2017 and HRI 2017.

July 2016. 5 papers to appear at EMNLP 2016: visual story sorting, visual question relevance, neural network interpretation (for

Tweets by @uncnlp

UNC NLP Retweeted

emnlp2019
@emnlp2019
Registration for EMNLP 2019 will open in a few days. In the meantime, you can have a look at the registration fees for the conference.emnlp-ijcnlp2019.org/registration/

EMNLP-IJCNLP 2019 Registration Fees

Type	Register	Full package	Main conference	Main + 1 Day	1 Day	2 Days
Regular	Early	$995	$685	$825	$220	$330
	Late	$1120	$800	$1075	$275	$415
	Onsite	$1315	$905	$1235	$330	$495

PhD Students


Lisa Bauer
PhD at UNC


Darryl Hannan
PhD at UNC


Peter Hase
PhD at UNC


Yichen Jiang
PhD at UNC


Hyounghun Kim
PhD at UNC
(co-advised w/ H. Fuchs)


Jie Lei
PhD at UNC
(co-advised w/ T. Berg)


Adyasha Maharana
PhD at UNC


Yixin Nie
PhD at UNC


Ramakanth Pasunuru
PhD at UNC


Swarnadeep Saha
PhD at UNC


Hao Tan
PhD at UNC


Shiyue Zhang
PhD at UNC


Yubo Zhang
PhD at UNC
(co-advised w/ A. Tropsha)


Xiang Zhou
PhD at UNC

Undergraduate Students


Tsion Coulter
UG at UNC


Han Guo
UG at UNC


Akshay Jain
UG at UNC


Sweta Karlekar
UG at UNC


Antonio Mendoza
UG at UNC


Yicheng Wang
UG at UNC


Songhe Wang
UG at UNC

Thank you!

Webpage: http://www.cs.unc.edu/~mbansal/

Email: mbansal@cs.unc.edu

UNC-NLP Lab: http://nlp.cs.unc.edu/

Postdoc Openings!!: ~mbansal/postdoc-advt-unc-nlp.pdf