DSTC7-AVSD: Scene-Aware Video Dialogue Systems with Dual Attention

Ramakanth Pasunuru

www.rama-kanth.com

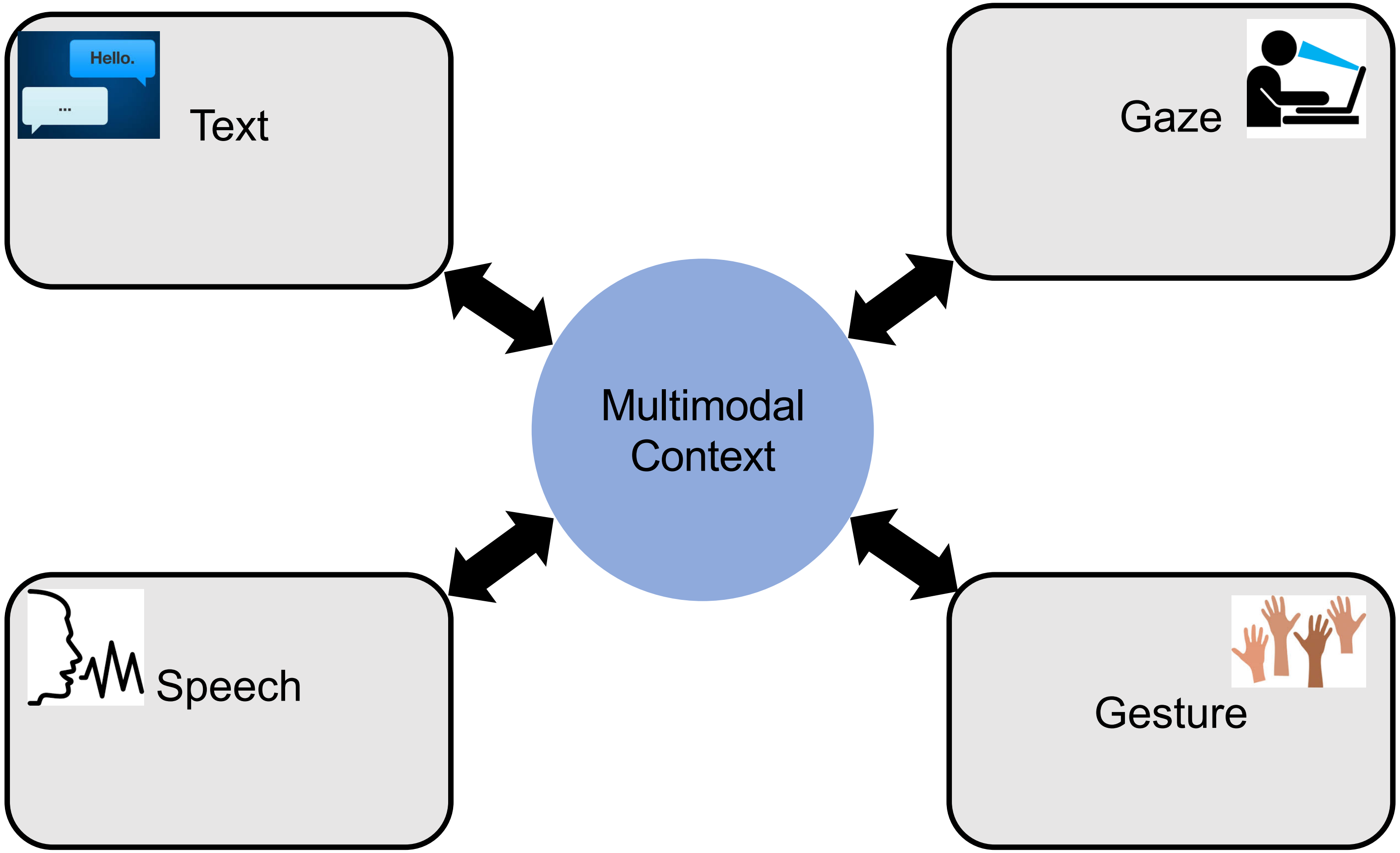
Mohit Bansal

www.cs.unc.edu/~mbansal/

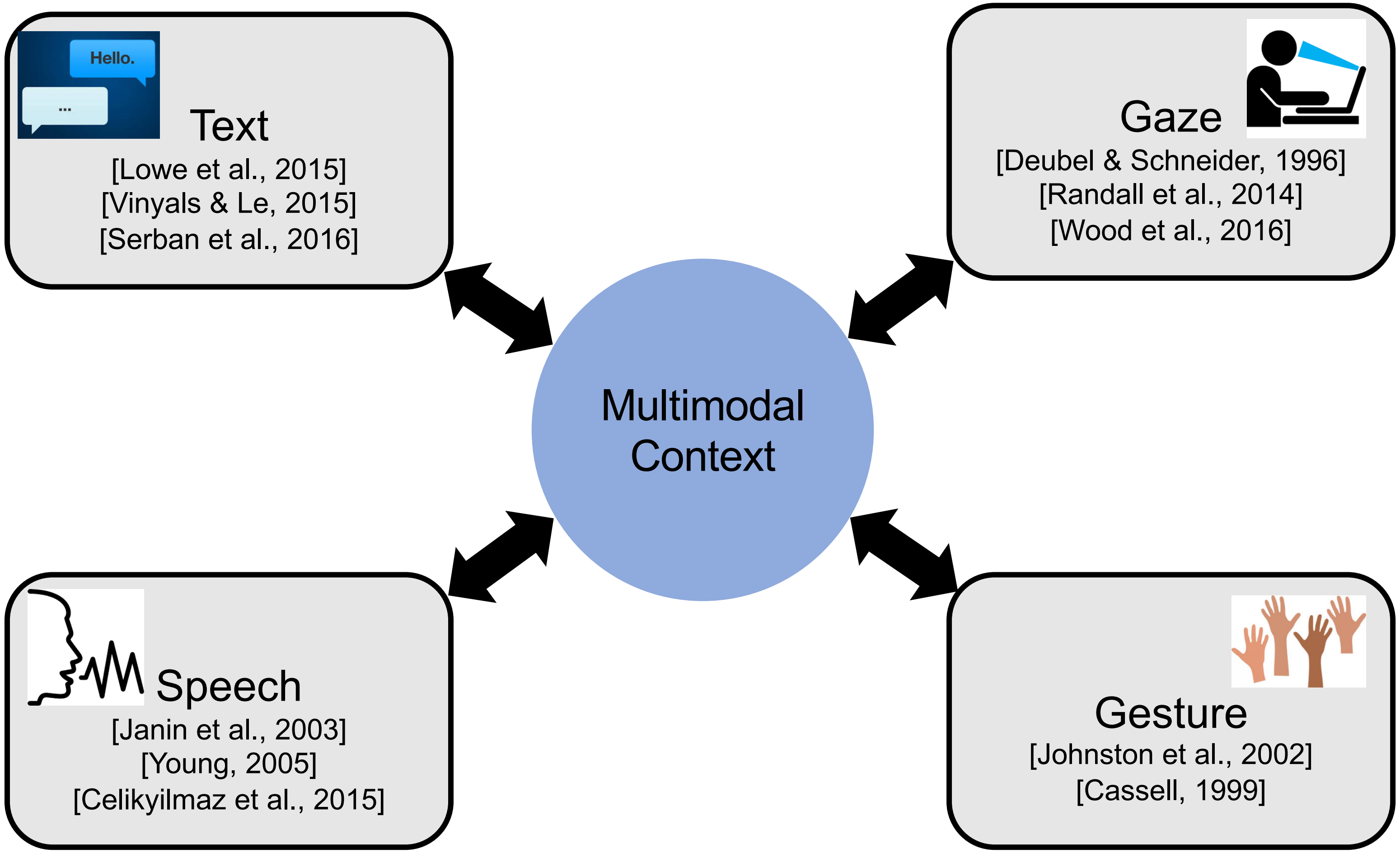
THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

UNC
NLP

Dialogue Context



Dialogue Context



Text
[Lowe et al., 2015]
[Vinyals & Le, 2015]
[Serban et al., 2016]

Gaze
[Deubel & Schneider, 1996]
[Randall et al., 2014]
[Wood et al., 2016]

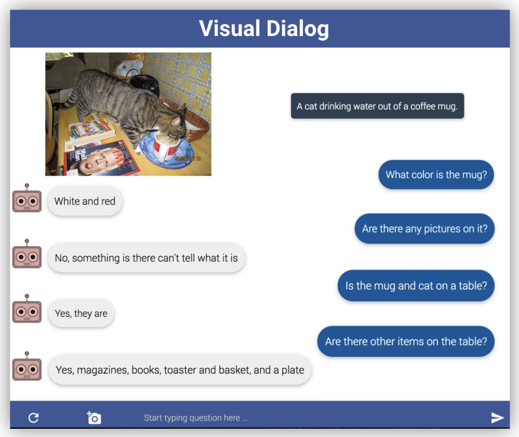
Multimodal Context

Speech
[Janin et al., 2003]
[Young, 2005]
[Celikyilmaz et al., 2015]

Gesture
[Johnston et al., 2002]
[Cassell, 1999]

Visual Context

Image-based Context



Visual Dialog

A cat drinking water out of a coffee mug.

What color is the mug?

White and red

Are there any pictures on it?

No, something is there can't tell what it is

Is the mug and cat on a table?

Yes, they are

Are there other items on the table?

Yes, magazines, books, toaster and basket, and a plate

Start typing question here ...

[Das et al., 2017]

Place near my house is getting ready for Halloween a little early.

Don't you think Halloween should be year-round, though?

That'd be fun since it's my favorite holiday!

It's my favorite holiday as well!

I never got around to carving a pumpkin last year even though I bought one.

Well, it's a good thing that they are starting to sell them early this year!

Is the photo in color?

Yes

Is the photo close up?

No

Is this at a farm?

Possibly

Do you think it's for Halloween?

That is possible

Do you see anyone?

No

Do you see trees?

No

Any huge pumpkins?

No

[Mostafazadeh et al., 2017]

Is it a person?	*No*
Is it an item being worn or held?	*Yes*
Is it a snowboard?	*Yes*
Is it the red one?	*No*
Is it the one being held by the person in blue?	*Yes*

Is it a cow?	*Yes*
Is it the big corn in the middle?	*No*
Is the cow on the left?	*No*
On the right ?	*Yes*
First cow near us?	*Yes*

[De Vries et al., 2017]

[System]: How can i help you today ?
[User]: *Find non-fiction books by Chomsky.*
[System]: (Fetches the following books from database)

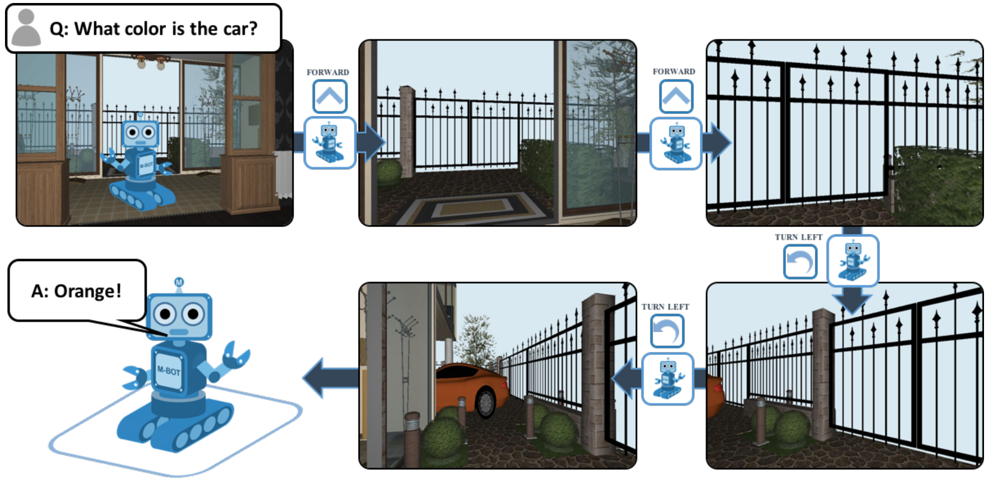
Hegemony or Survival 2003
Manufacturing Consent:The Pol 1988
Language and Thought 1993
Power Systems: Conversations w 2013
Aspects of the Theory of Syntax 1965

[User]: ***"show details for the oldest production"*** or
"details for the syntax book" or
"open the last one" or
"i want to see the one on linguistics" or
"bring me Jurafsky's text book"

[Celikyilmaz et al., 2014]

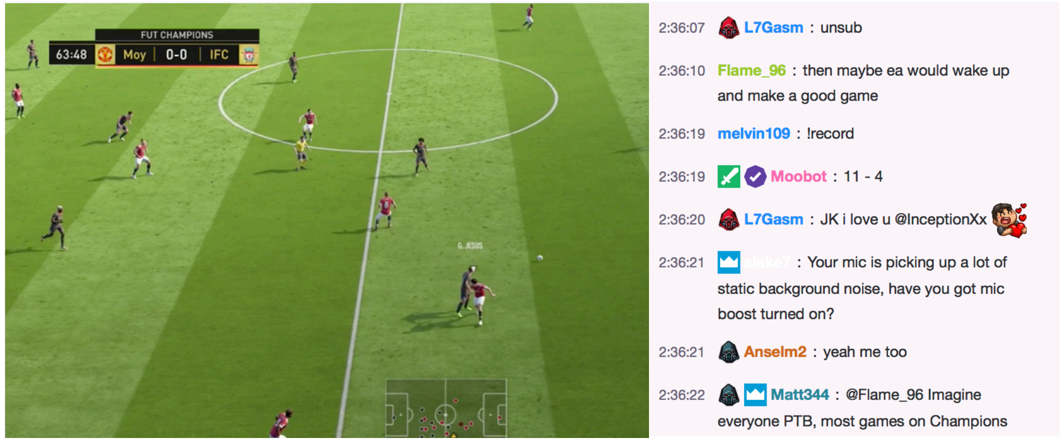
Visual Context

Dynamic-Visual Context



[Das et al., 2018]

Dynamic-Visual Context with Multi-Speaker



2:36:07 L7Gasm : unsub

2:36:10 Flame_96 : then maybe ea would wake up and make a good game

2:36:19 melvin109 : !record

2:36:19 Moobot : 11 - 4

2:36:20 L7Gasm : JK i love u @InceptionXx

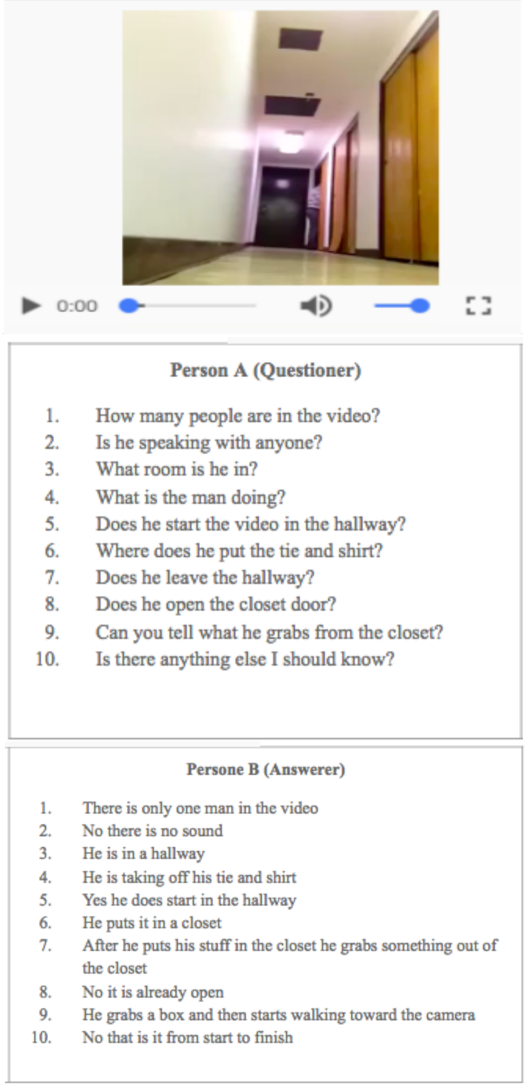
2:36:21 [redacted] : Your mic is picking up a lot of static background noise, have you got mic boost turned on?

2:36:21 Anselm2 : yeah me too

2:36:22 Matt344 : @Flame_96 Imagine everyone PTB, most games on Champions

[Pasunuru & Bansal, 2018]

Visual+Audio Context



Person A (Questioner)

1. How many people are in the video?
2. Is he speaking with anyone?
3. What room is he in?
4. What is the man doing?
5. Does he start the video in the hallway?
6. Where does he put the tie and shirt?
7. Does he leave the hallway?
8. Does he open the closet door?
9. Can you tell what he grabs from the closet?
10. Is there anything else I should know?

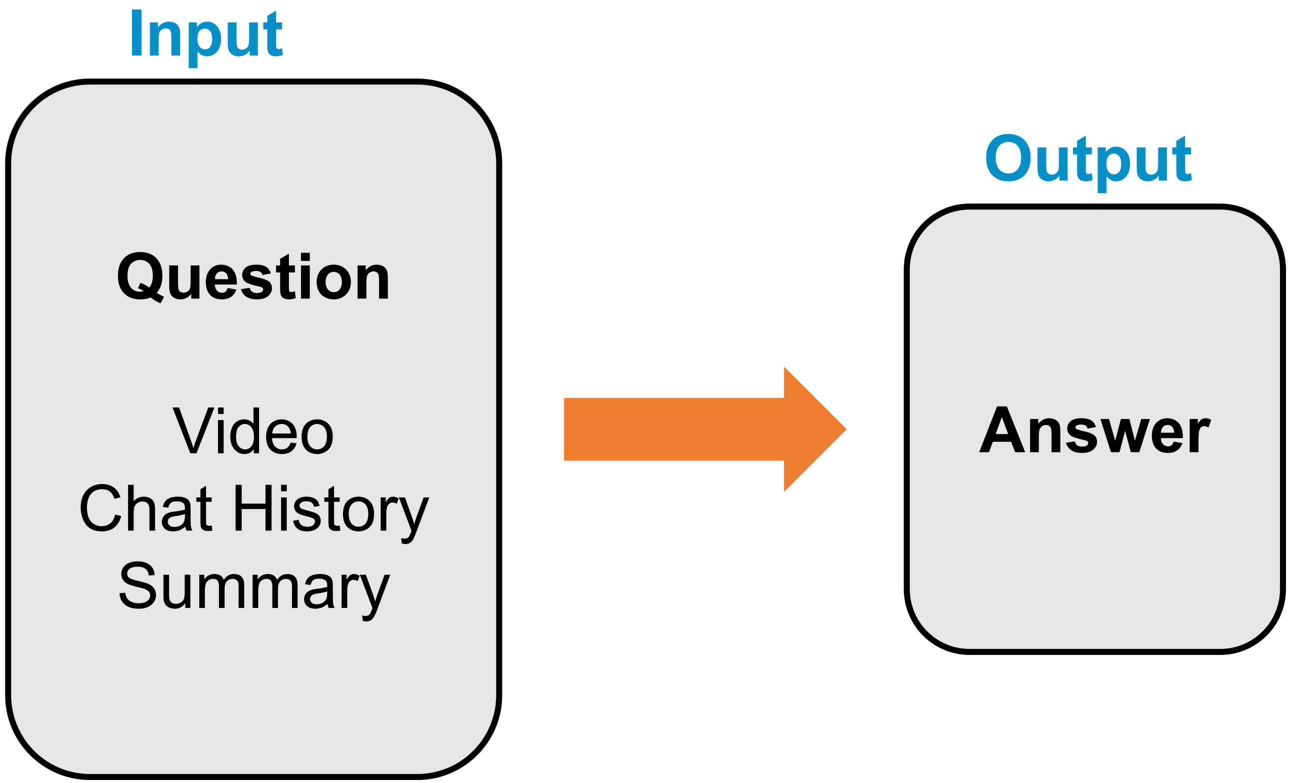
Persone B (Answerer)

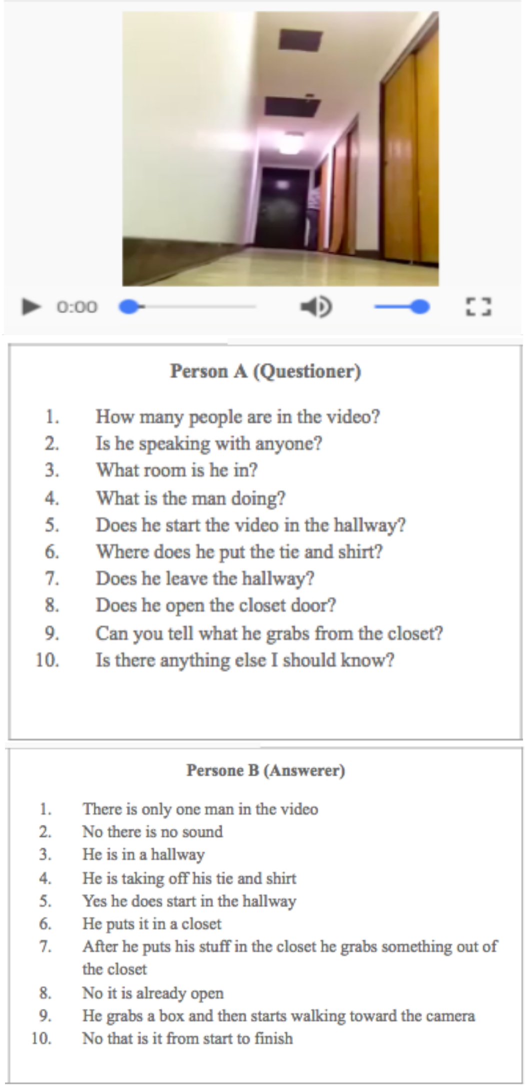
1. There is only one man in the video
2. No there is no sound
3. He is in a hallway
4. He is taking off his tie and shirt
5. Yes he does start in the hallway
6. He puts it in a closet
7. After he puts his stuff in the closet he grabs something out of the closet
8. No it is already open
9. He grabs a box and then starts walking toward the camera
10. No that is it from start to finish

[Alamri et al., 2018]

Visual+Audio Context

Task:





Person A (Questioner)

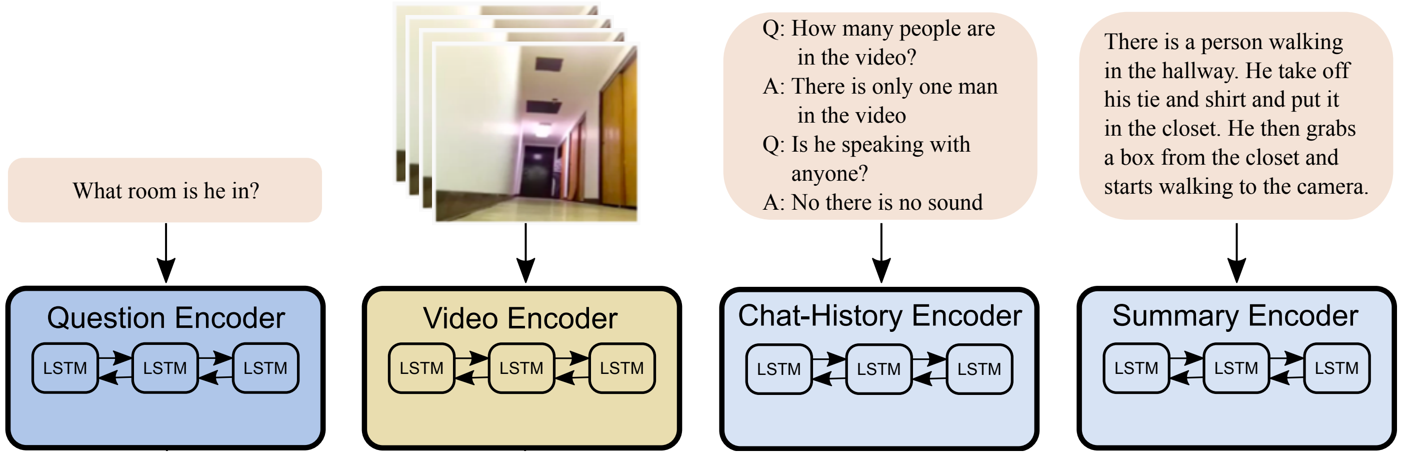
1. How many people are in the video?
2. Is he speaking with anyone?
3. What room is he in?
4. What is the man doing?
5. Does he start the video in the hallway?
6. Where does he put the tie and shirt?
7. Does he leave the hallway?
8. Does he open the closet door?
9. Can you tell what he grabs from the closet?
10. Is there anything else I should know?

Persone B (Answerer)

1. There is only one man in the video
2. No there is no sound
3. He is in a hallway
4. He is taking off his tie and shirt
5. Yes he does start in the hallway
6. He puts it in a closet
7. After he puts his stuff in the closet he grabs something out of the closet
8. No it is already open
9. He grabs a box and then starts walking toward the camera
10. No that is it from start to finish

[Alamri et al., 2018]

Model

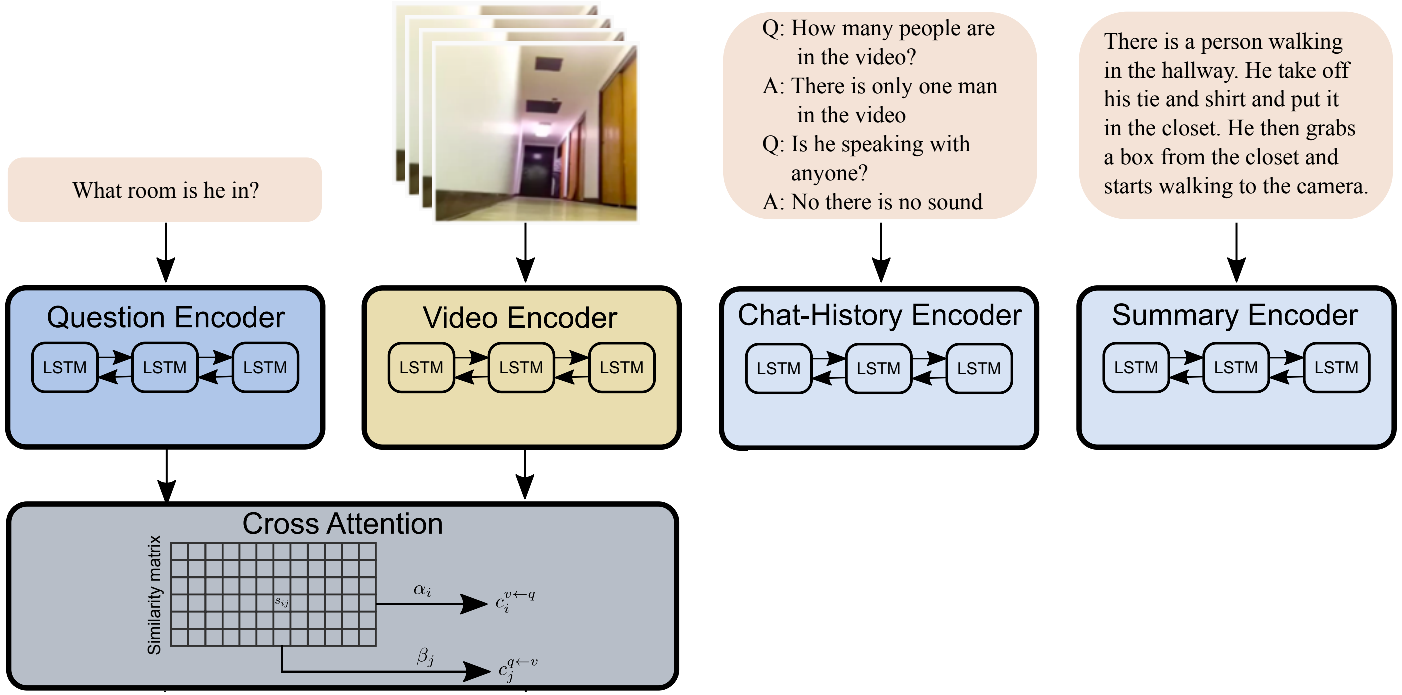


What room is he in?

Q: How many people are in the video?
A: There is only one man in the video
Q: Is he speaking with anyone?
A: No there is no sound

There is a person walking in the hallway. He take off his tie and shirt and put it in the closet. He then grabs a box from the closet and starts walking to the camera.

Model



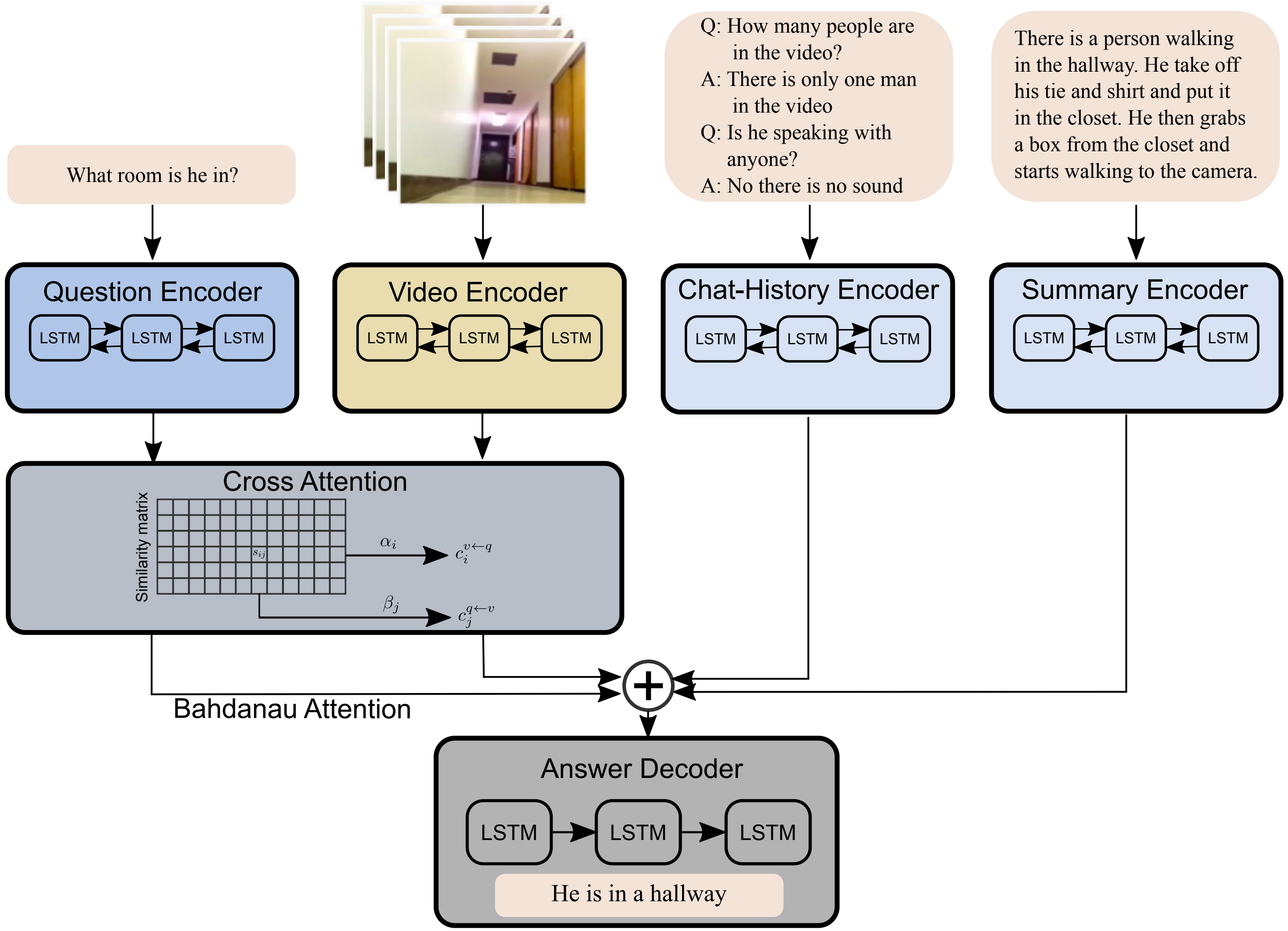
What room is he in?

Q: How many people are in the video?
A: There is only one man in the video
Q: Is he speaking with anyone?
A: No there is no sound

There is a person walking in the hallway. He take off his tie and shirt and put it in the closet. He then grabs a box from the closet and starts walking to the camera.

[Seo et al., 2017]

Model



What room is he in?

Q: How many people are in the video?
A: There is only one man in the video
Q: Is he speaking with anyone?
A: No there is no sound

There is a person walking in the hallway. He take off his tie and shirt and put it in the closet. He then grabs a box from the closet and starts walking to the camera.

Question Encoder

Video Encoder

Chat-History Encoder

Summary Encoder

LSTM LSTM LSTM

Cross Attention

Similarity matrix

s_{ij}

α_i $c_i^{v\leftarrow q}$

β_j $c_j^{q\leftarrow v}$

Bahdanau Attention

Answer Decoder

LSTM LSTM LSTM

He is in a hallway

Note that we do not use audio features in our models

[Bahdanau et al., 2015; Seo et al., 2017]

Results

Model	METEOR	CIDEr	BLEU-4	ROUGE-L
Video Only	12.43	95.54	8.83	34.23
Video + Chat History	14.13	105.39	10.58	36.54
Video + Chat History + Summary	14.94	112.80	11.22	37.53
Video + Chat History + Summary + Cross-attention	14.95	115.82	11.38	37.87

Our models' performance on AVSD dataset's public test set. All of these models use the question information.

Results

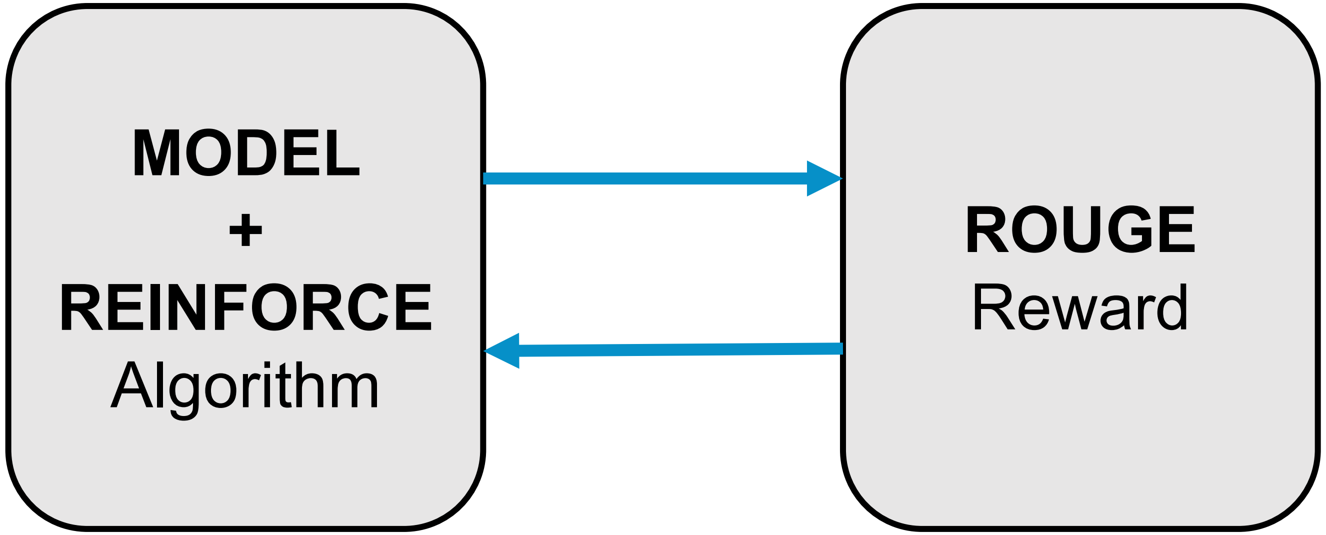
Model	METEOR	CIDEr	BLEU-4	ROUGE-L
Video Only	12.43	95.54	8.83	34.23
Video + Chat History	14.13	105.39	10.58	36.54
Video + Chat History + Summary	14.94	112.80	11.22	37.53
Video + Chat History + Summary + Cross-attention	14.95	115.82	11.38	37.87

Our models' performance on AVSD dataset's public test set. All of these models use the question information (no audio information).

Other Methods

- Policy gradient based reinforcement learning
- Contextualized ELMo word embeddings
- Using external data
- Pointer-generator copy model

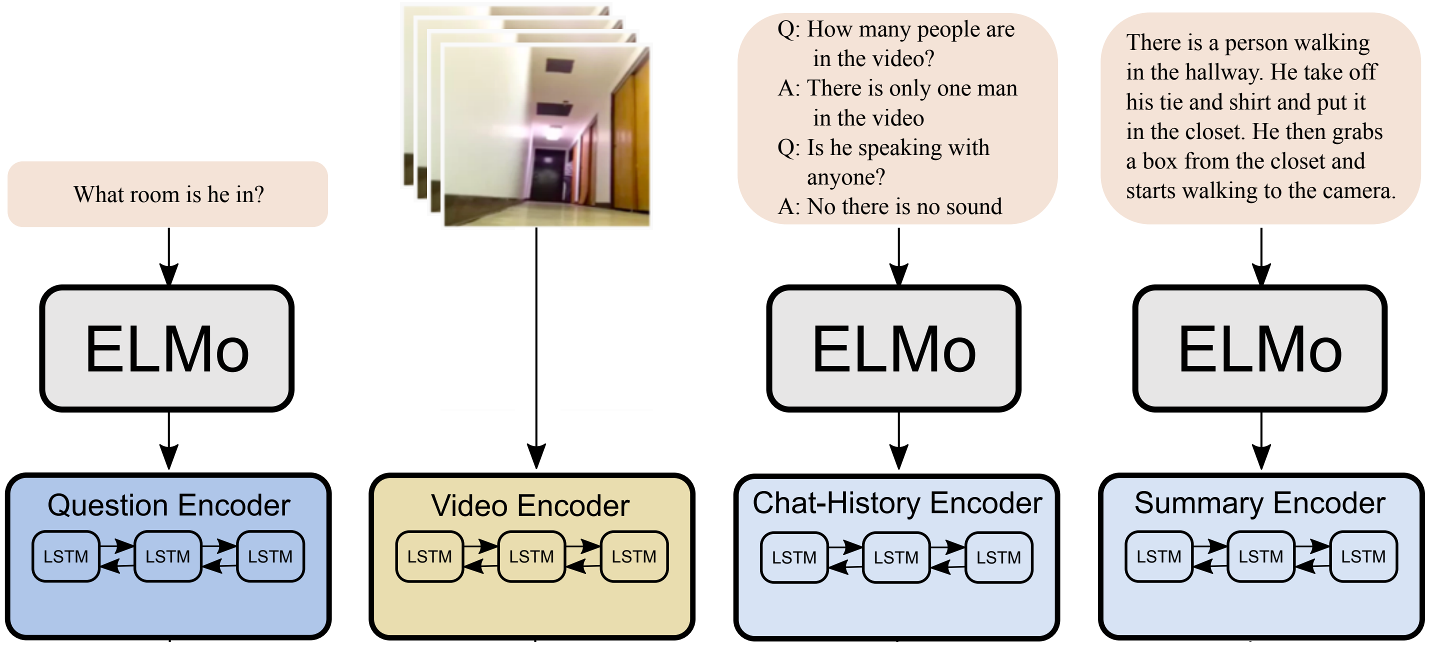
Policy Gradients



$$\nabla_\theta L(\theta) = -\mathbb{E}_{w^s \sim p_\theta}[r(w^s) \cdot \nabla_\theta \log p_\theta(w^s)]$$

[Williams, 1992]

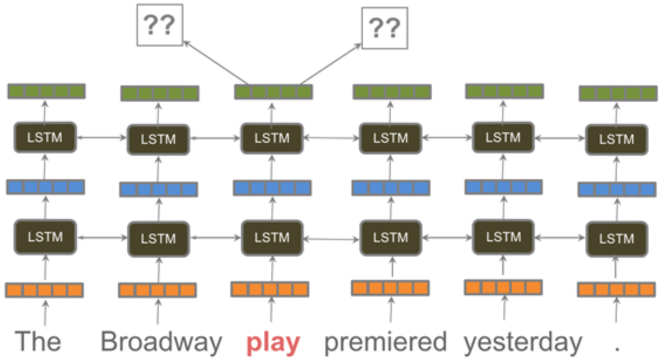
Contextualized ELMo Word Embeddings



What room is he in?

Q: How many people are in the video?
A: There is only one man in the video
Q: Is he speaking with anyone?
A: No there is no sound

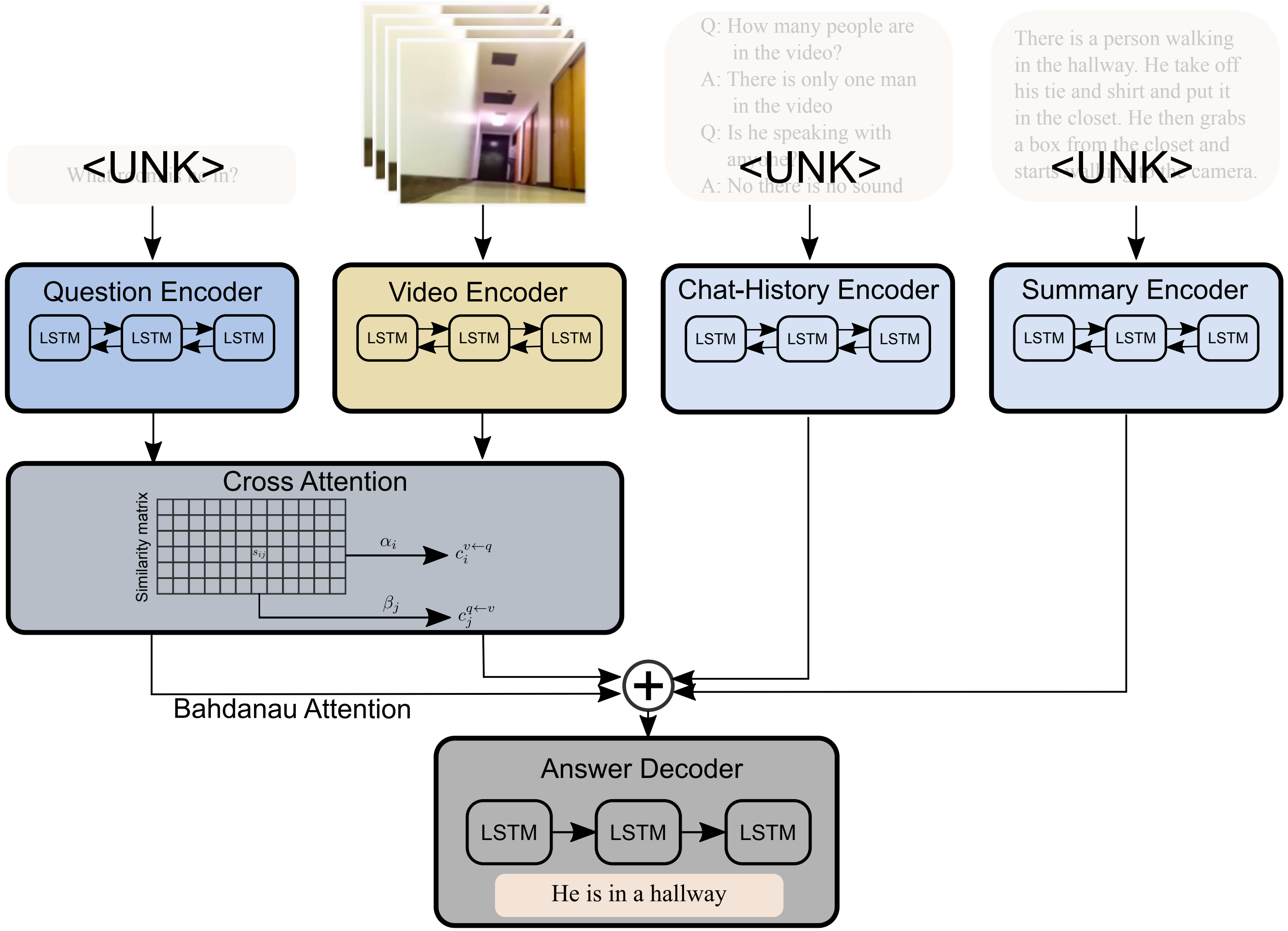
There is a person walking in the hallway. He take off his tie and shirt and put it in the closet. He then grabs a box from the closet and starts walking to the camera.



ELMo

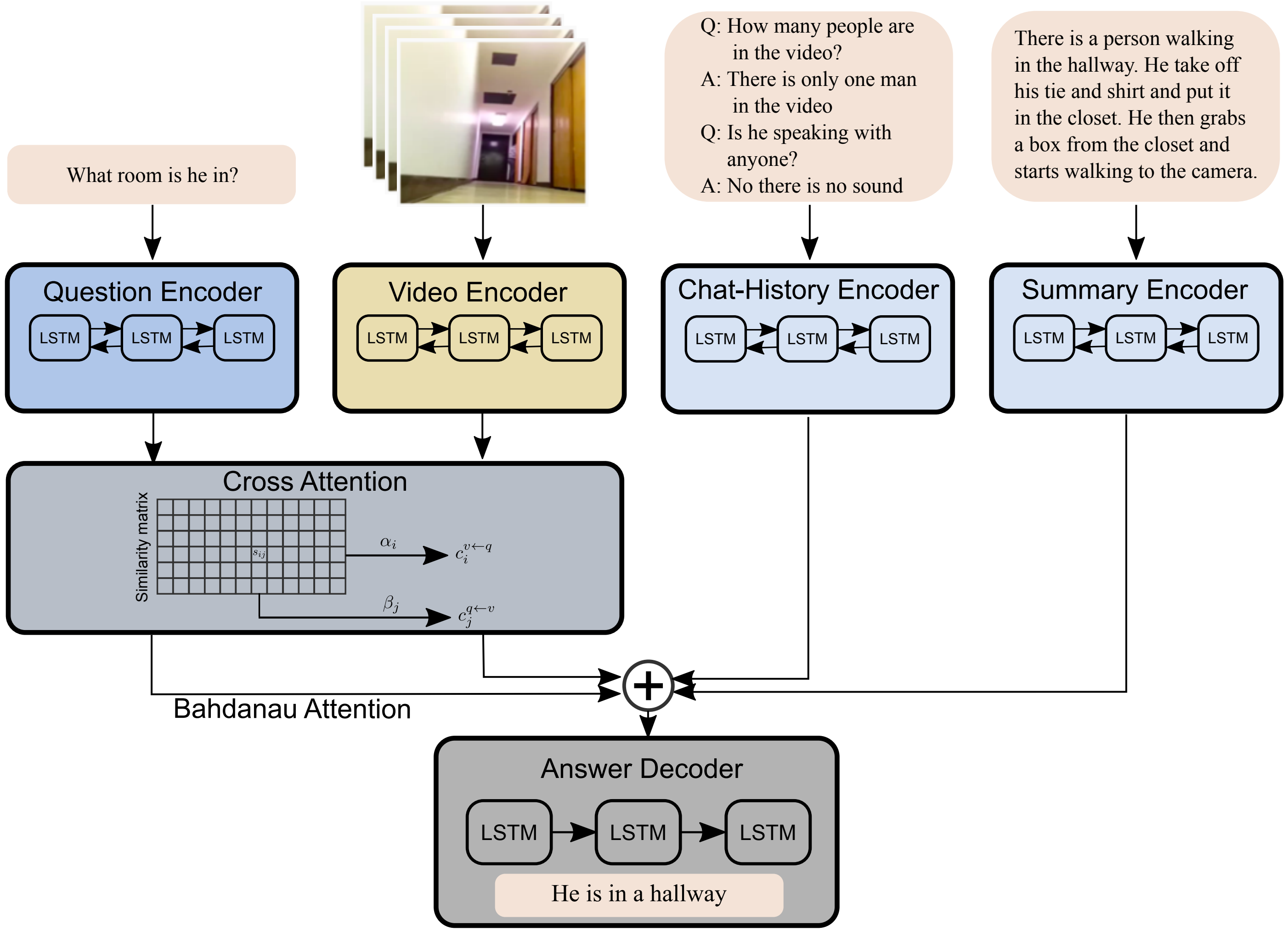
[Peters et al., 2018]

Using External Data (MSR-VTT)



[Xu et al., 2016]

Using External Data (MSR-VTT)



What room is he in?

Q: How many people are in the video?
A: There is only one man in the video
Q: Is he speaking with anyone?
A: No there is no sound

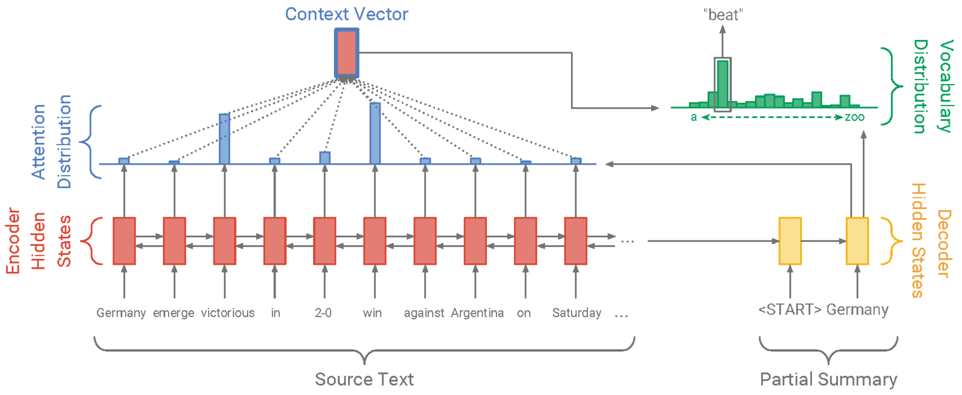
There is a person walking in the hallway. He take off his tie and shirt and put it in the closet. He then grabs a box from the closet and starts walking to the camera.

He is in a hallway

[Xu et al., 2016]

Pointer-generator Copy Model

- Lot of words in the question can also be present in the answer
- The final word distribution is a weighted combination of the vocab distribution and attention distribution
- Question-based pointer
- Joint question- and summary-based pointer



[See et al., 2017]

Future Work

- Further analyze and improve these promising approaches with specific RL rewards, contextualized large language models, and joint copy models

- We will add Audio features to our final model

- Effective ways of extending cross-attention to multiple modalities (question+summary; question+chat-history)

Thanks!

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

UNC
NLP

Acknowledgment: DARPA YFA17-D17AP00022, Faculty awards from Google, Facebook, and Salesforce