

Active Scene Understanding for Video Labeling

Public Version

Team Moonshot:



John



Morgane



Andrei



Siyam



Andreas

Supervisors:



Fred



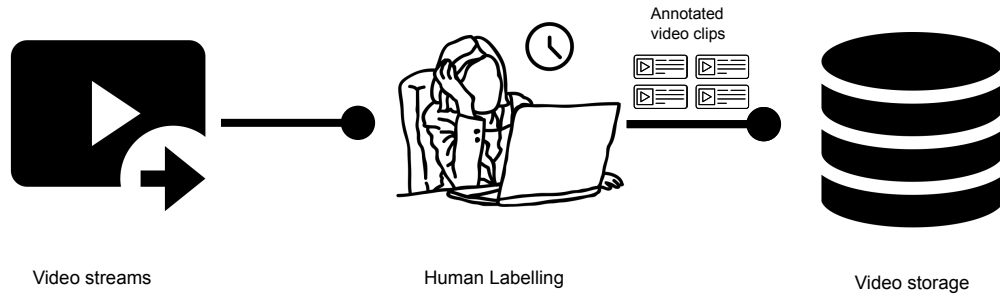
Keesiu

Co-supervisor:

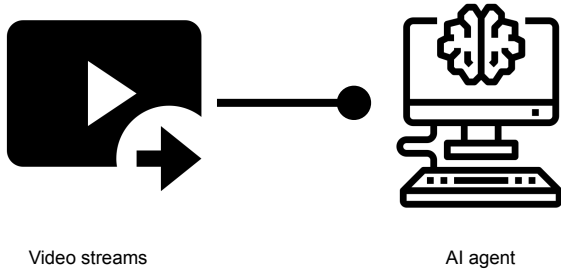


*Marija
Tepegjozova*

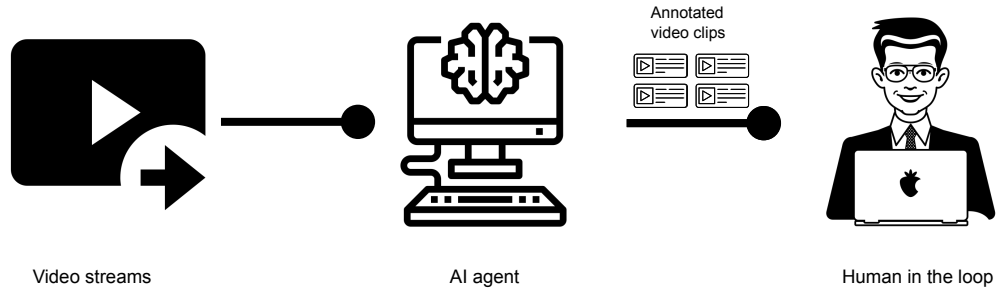
Motivation



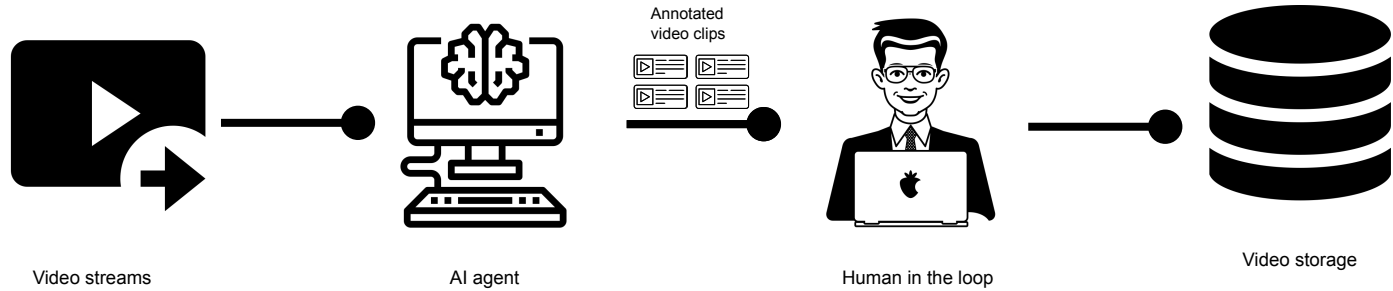
Motivation

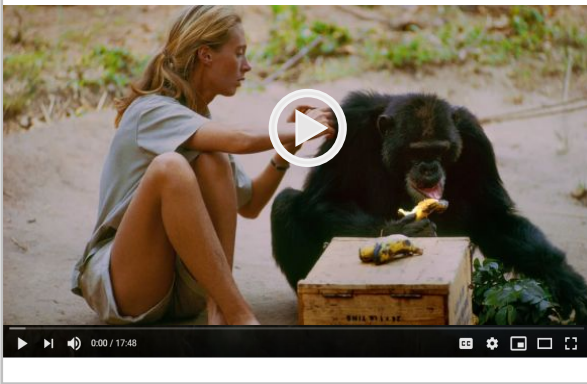


Motivation



Motivation





Metadata



TIME
Scene 1#

PLOT SUMMARY

AUDIO TRANSCRIPT



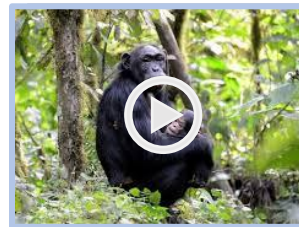
Scene 2#

Describe Main Events

Transcribe Audio to Text

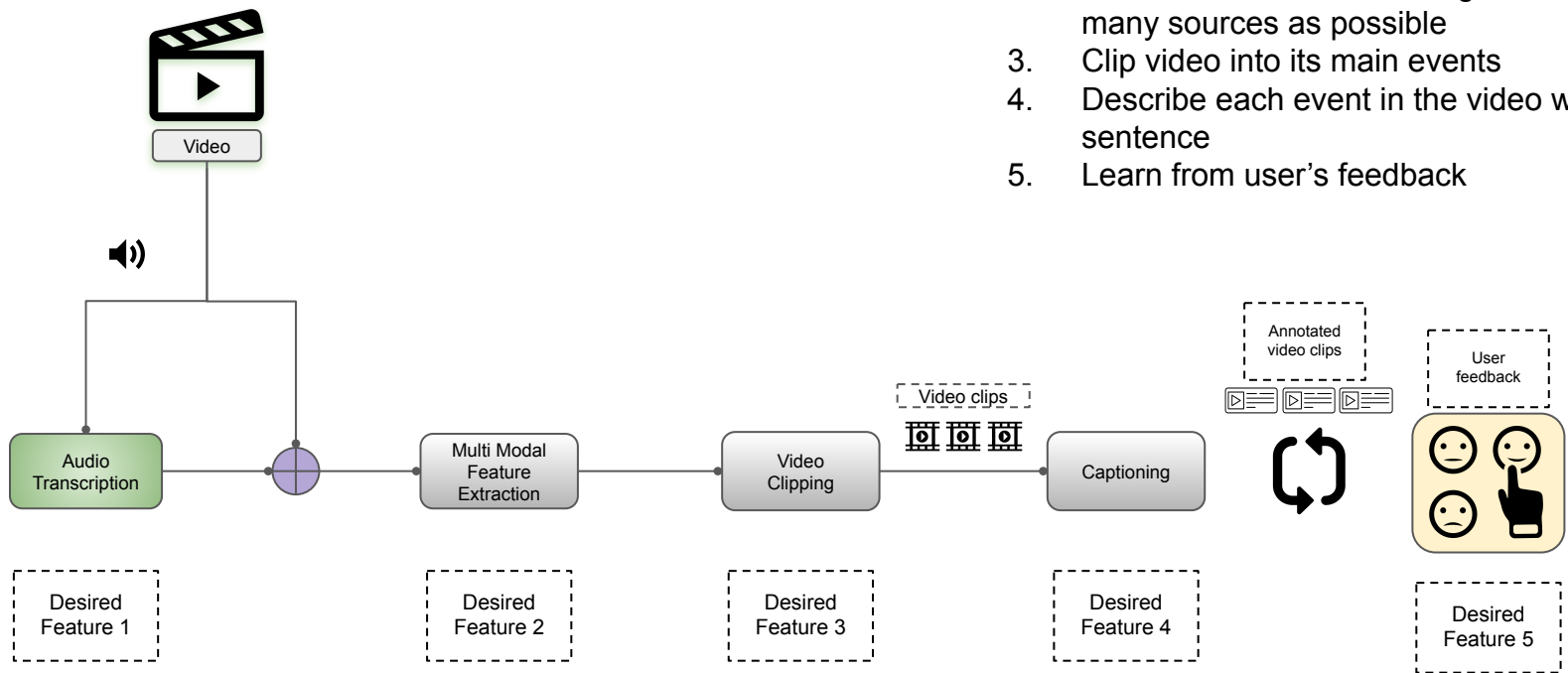


Scene 3#



Scene 4#

Task Overview



Why?

1. Automatically show video conversations split by scene
2. Extract video understanding from as many sources as possible
3. Clip video into its main events
4. Describe each event in the video with one sentence
5. Learn from user's feedback

Desired Feature 1 - Audio Transcription

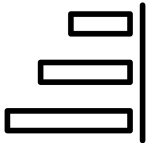
Requirements



Offline



Accurate



Speech-to-Text Alignment



Acoustic model:
predict most
likely emissions
and transitions



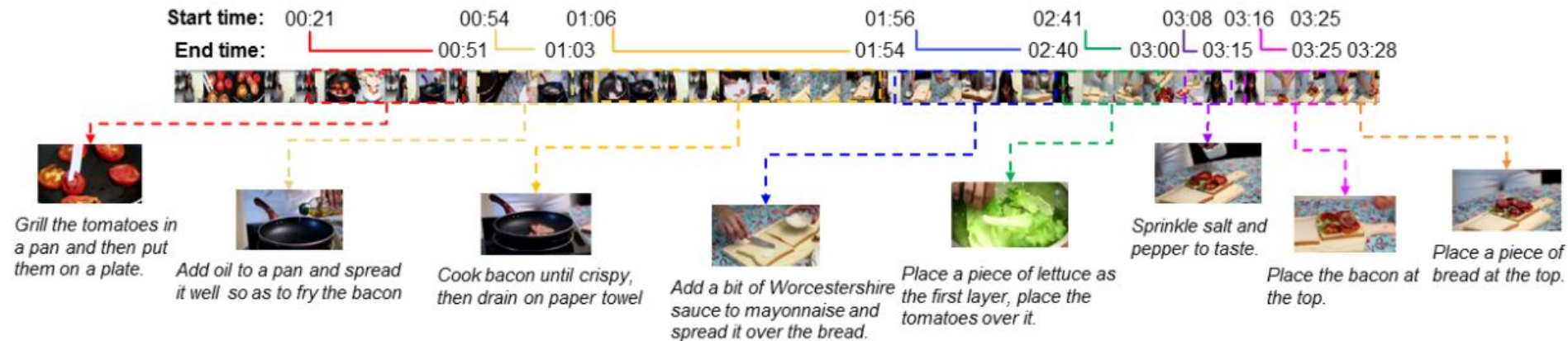
Language model:
performs
one-pass beam
search decoding



Start (ms), end (ms), transcript

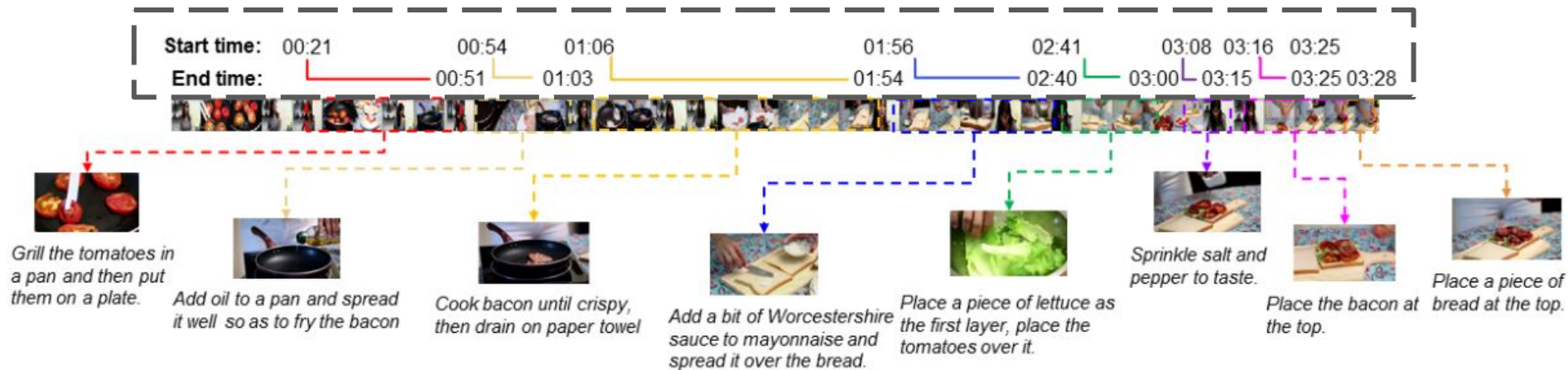
55000,56000,three were
56000,57000,given to to the elves
57000,58000,immortal
58000,59000,wisest
59000,60000,and fairest of
60000,61000,all beings

Dense Video Captioning



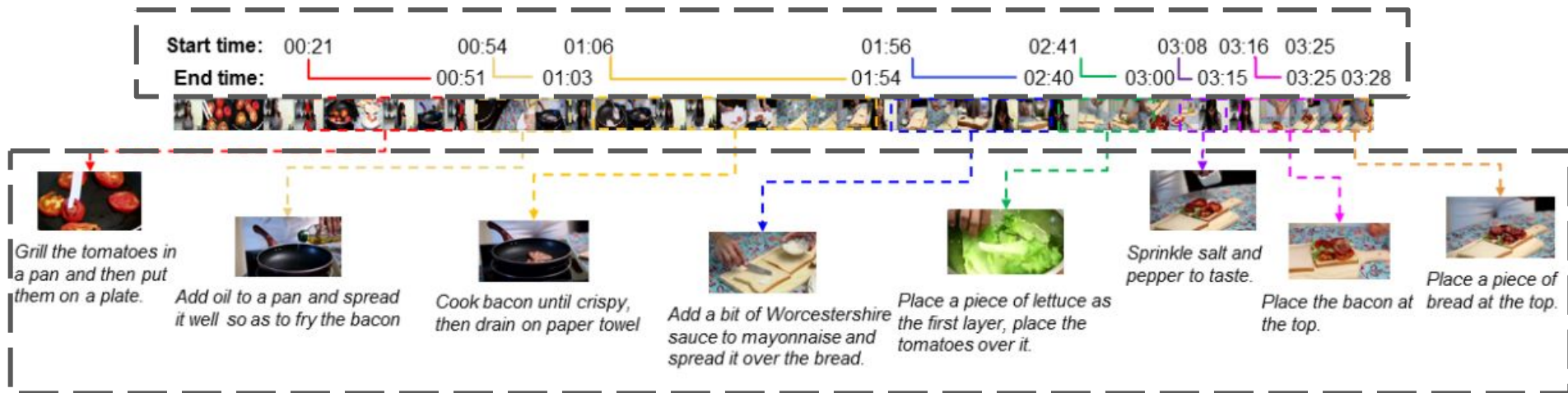
Dense Video Captioning

Dense event proposals



Dense Video Captioning

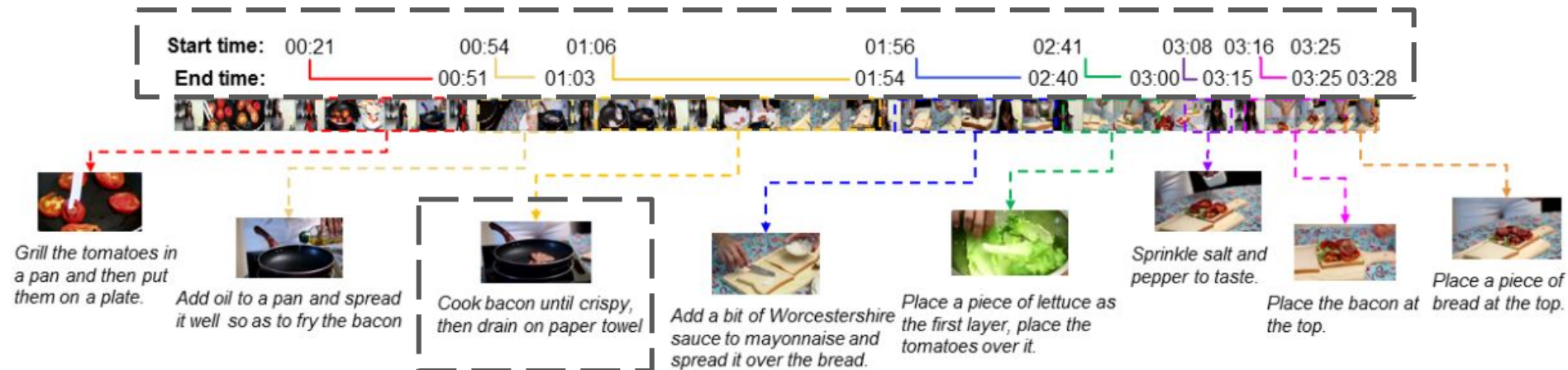
Dense event proposals



One-sentence captioning
of each proposal

Dense Video Captioning

Dense event proposals



Each proposal
represents one main
event

Datasets

YouCookII



Public Datasets

*ActivityNet
Captions*



Datasets

Too application specific!

YouCookII



Public Datasets

*ActivityNet
Captions*



Datasets

Too application specific!

YouCookII



Public Datasets

*ActivityNet
Captions*

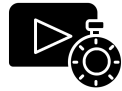


Encompasses many domains

Datasets: ActivityNet Captions



20k videos



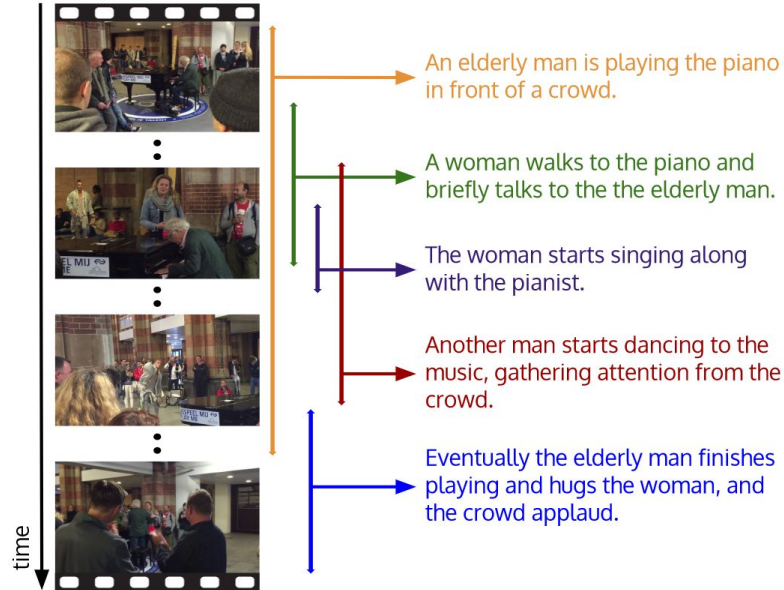
180s on average



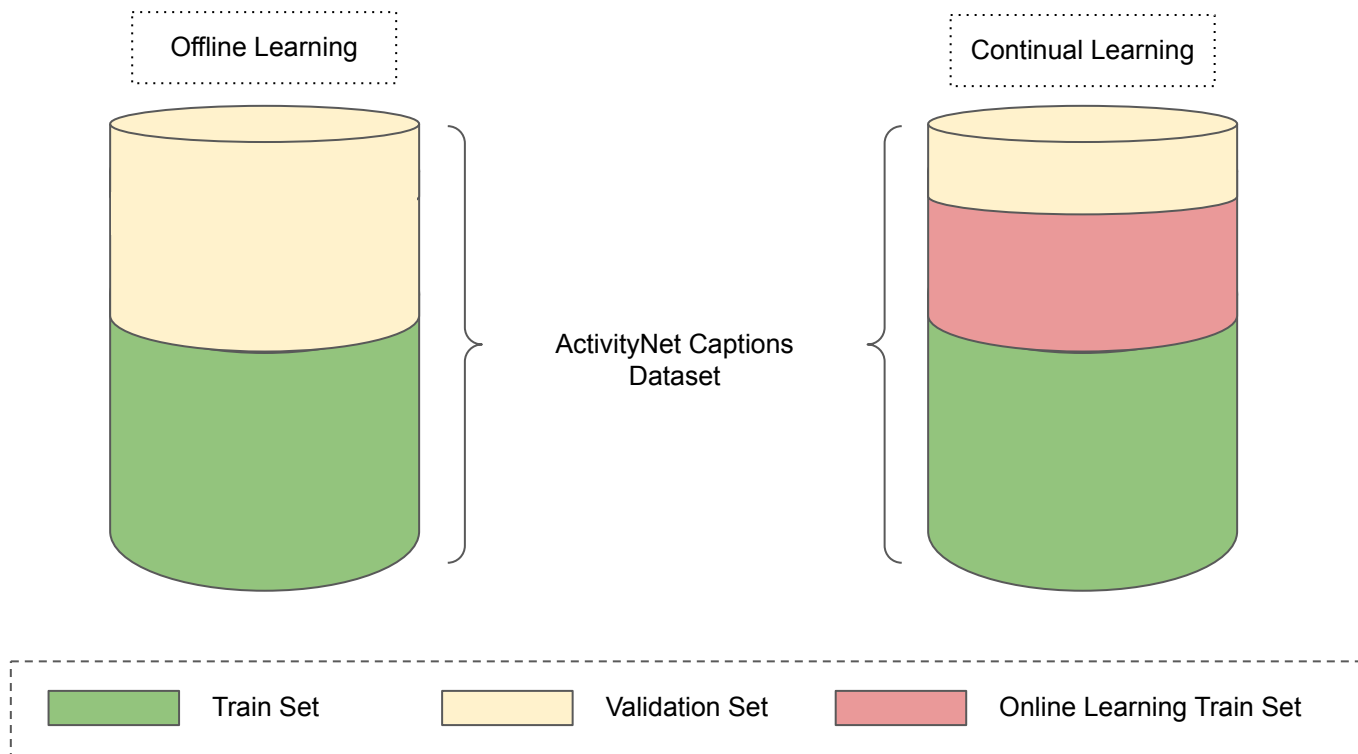
100k sentences



13.48 WPS



Results - Evaluation Procedure



Results - Proposal Generation & Captioning

	Full Dataset Available	GT Proposals			Learned Proposals		
		B@3	B@4	METEOR	B@3	B@4	METEOR
<i>Mun et al.</i>	Yes	-	-	-	-	-	6.92
<i>Krishna et al.</i>	Yes	4.09	1.60	8.88	1.90	0.71	5.69
<i>Li et al.</i>	Yes	4.51	1.71	9.31	2.05	0.74	6.14
<i>Zhou et al.</i>	Yes	5.78	2.71	11.16	2.91	1.44	6.91
<i>Wang et al.</i>	Yes	-	-	10.89	2.27	1.13	6.10

<i>Teng et al.</i>	No	-	1.98	11.49	2.53	-	7.65
<i>lashin et al.</i>	No	4.52	1.46	11.07	1.85	1.01	7.46
<i>Rahman et al.</i>	No	3.04	1.99	7.23	3.84	0.90	4.93
<i>BMT</i>	No	4.63	1.99	10.90	3.84	1.88	8.44
<i>MMT [Ours]</i>	No	5.83	2.86	11.72	4.00	2.01	9.43

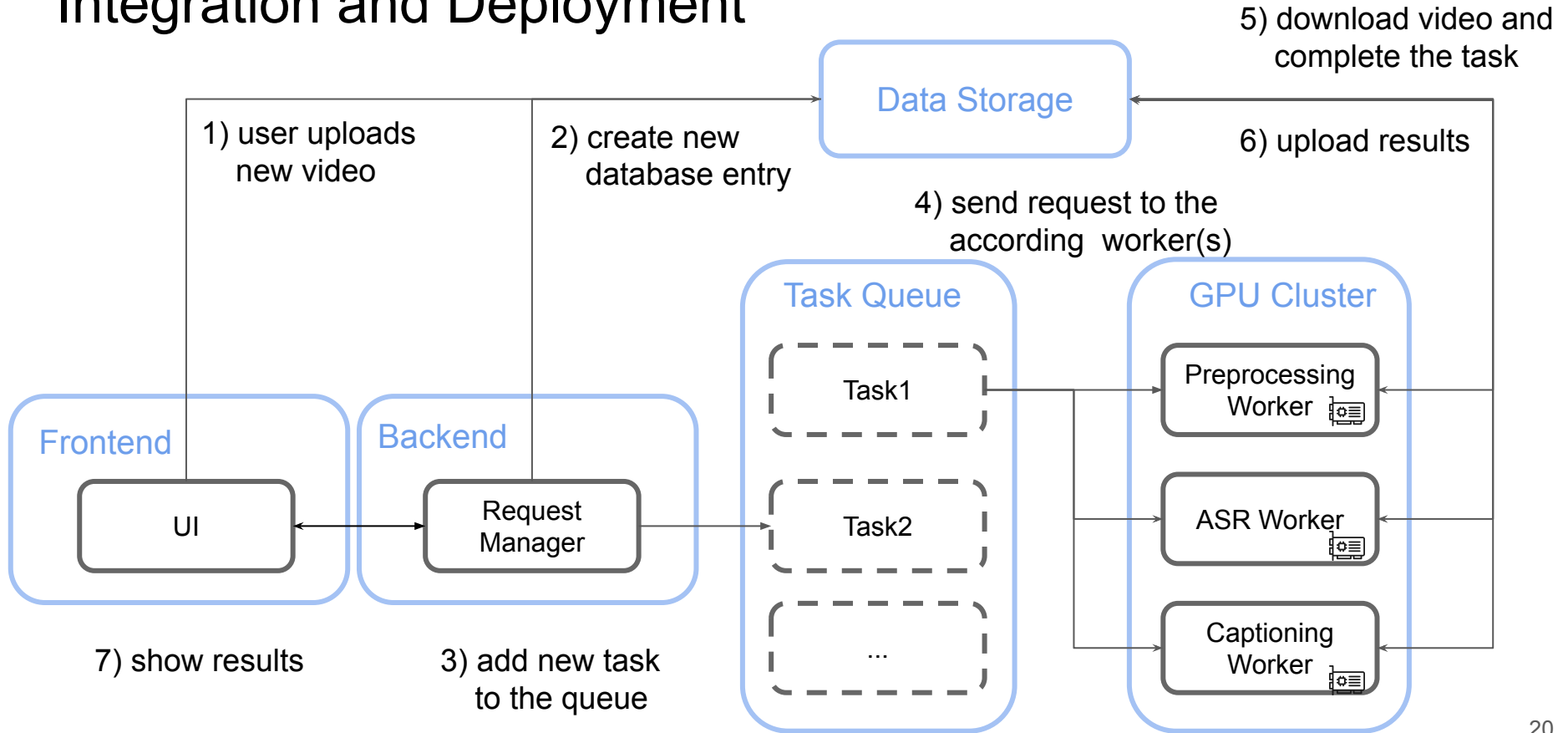
Multi-Modal Transformer (MMT) surpasses the state of the art in both GT and Learned Proposals on all metrics

Results - Continual Learning


Method	B@3	B@4	METEOR
BMT(offline + Train Set)	4.9	2.3	10.51
BMT(offline + Train Set + 80% Val Set)	5.66	2.71	11.12
ALTC (ours)	5.7	2.8	11.23



- Bi-Modal Transformer (BMT) used as lower and upper bound offline baseline
- ALTC applied on BMT **surpasses** the offline upper bound

Integration and Deployment




Demo - Application

DESIGN 

 UPLOAD VIDEO  LOGOUT


Videos > Snowboarding in Alps




Processed 00:03:58 LANDSCAPE 854x480 25 FPS

29.77 MB


Title

 Snowboarding in Alps


Date

21.01.2015 

Author

 GuGa

Director


 Robert B. Weide

Scenes

[+ ADD SCENE](#)


Summary

A camera pans around a snowy area and leads into a man riding down a snowy hill


Scene #2 Edited 


Start End


00:01:12:20 - 00:03:37:20



Audio Transcripts, - 00:01:16:00


 00:01:46:00 - 00:01:50:00
night i go to back i don't do that we be good for the night i go to
everything with smoke about his you know me you know tell me

 00:01:52:00 - 00:01:54:00
i you and i get back

 00:01:57:00 - 00:02:01:00
so as i that turn over so that we

Summary

He continues riding down the hill while moving his arms and ends by holding a selfie-stick

Scene #3 Edited 

Start End

Conclusion



Leveraged Multiple Modalities



Beat the State-of-the-Art in DVC



Adapted Continual Learning to DVC

Our Moonshot Goal



The Multi Modal
Transformer + ALTC

“**John, Morgane, Andreas and Keesiu** are playing an **exciting** game of kicker during **COVID-19** times while **Andrei ignores them** and **Siyam watches.**”