Active Scene Understanding for Video Labeling

Public Version

Team Moonshot:

John

Morgane

Andrei

Siyam

Andreas

Supervisors:

Fred

Keesiu

Co-supervisor:

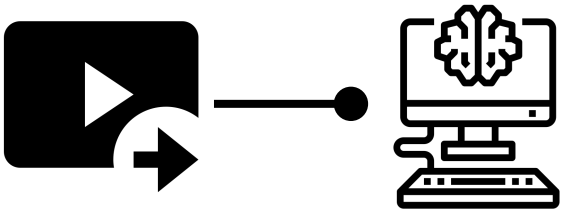
Marija Tepegjozova

TUM Data Innovation Lab
24 February 2021

Motivation


Annotated
video clips
Video streams
Human Labelling
Video storage

Motivation



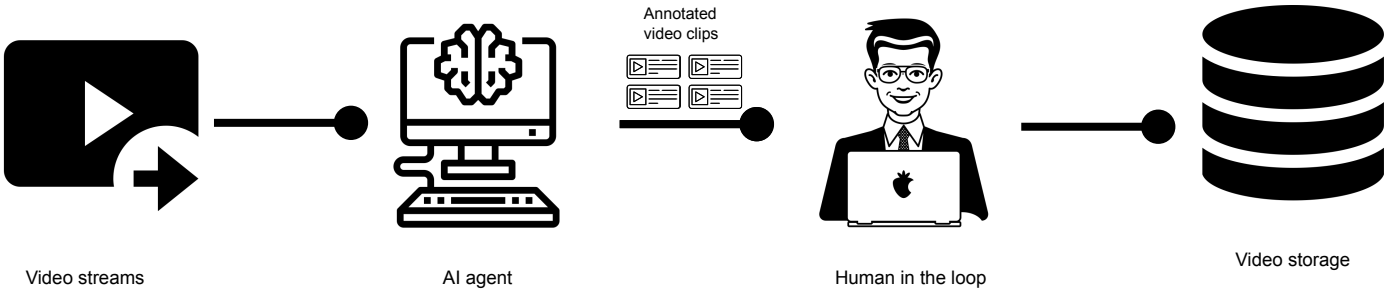
Video streams

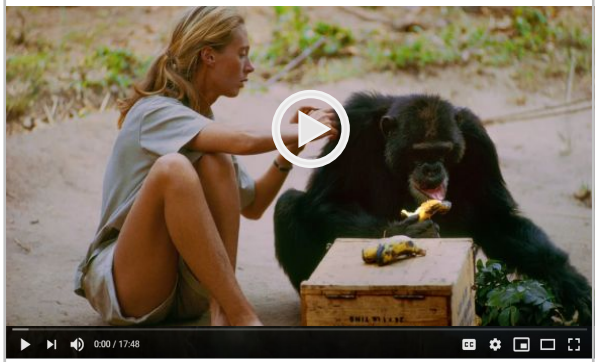
AI agent

Motivation


Annotated
video clips
Video streams
AI agent
Human in the loop

Motivation


Annotated
video clips
Video streams
AI agent
Human in the loop
Video storage



Metadata

Scene 1#

Scene 2#

Scene 3#

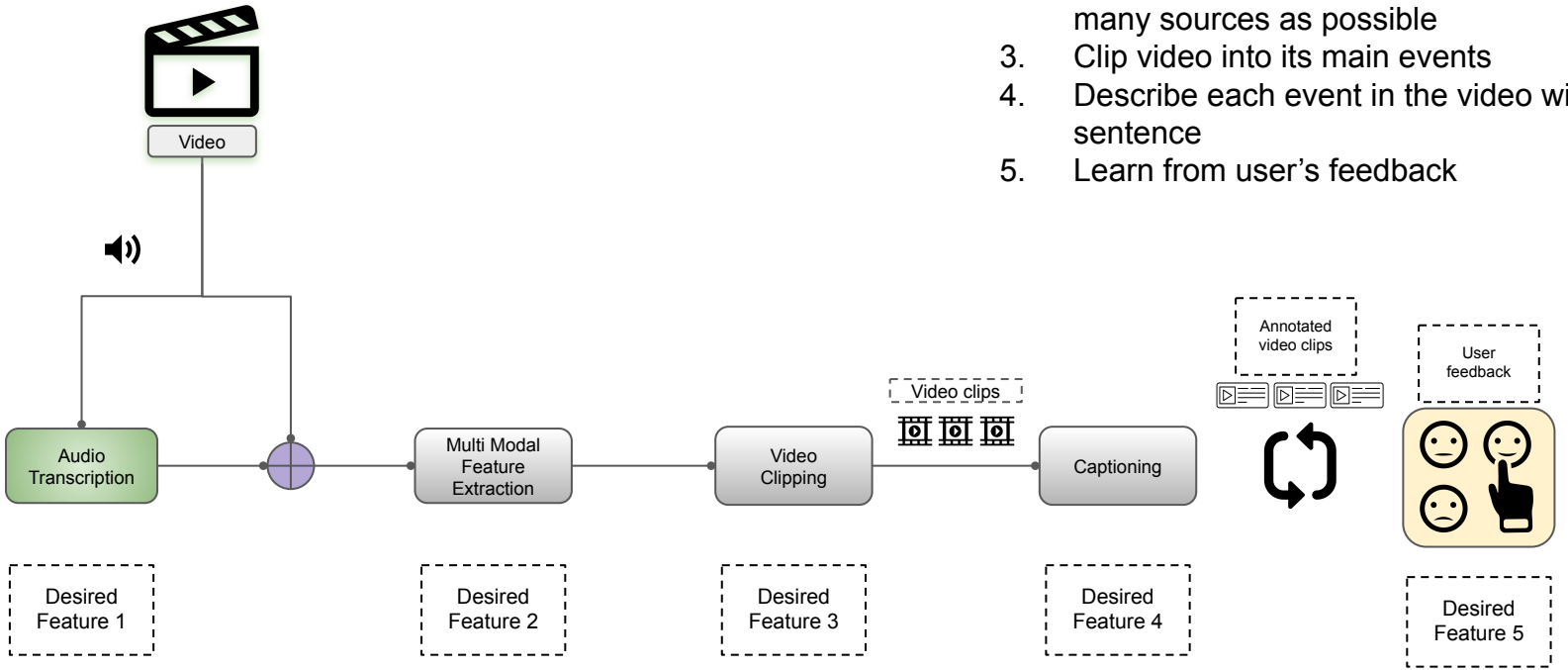
Scene 4#

TIME	PLOT SUMMARY	AUDIO TRANSCRIPT
	Describe Main Events	*Transcribe Audio to Text*

Task Overview

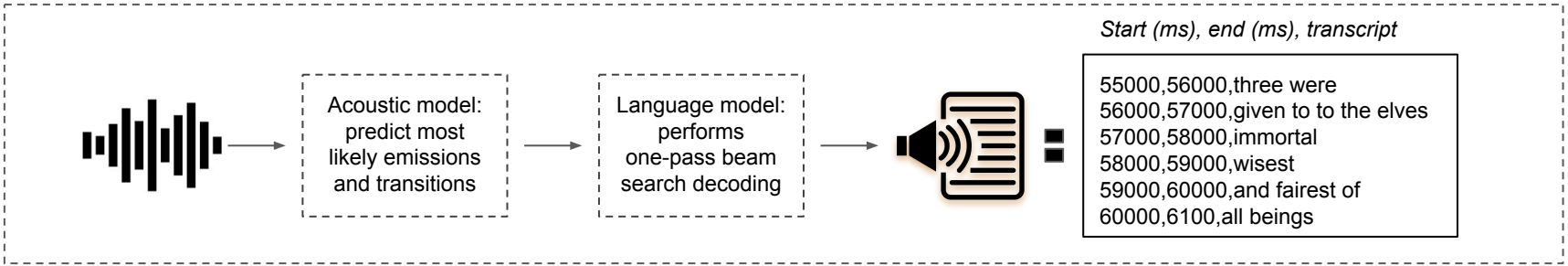
Why?

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

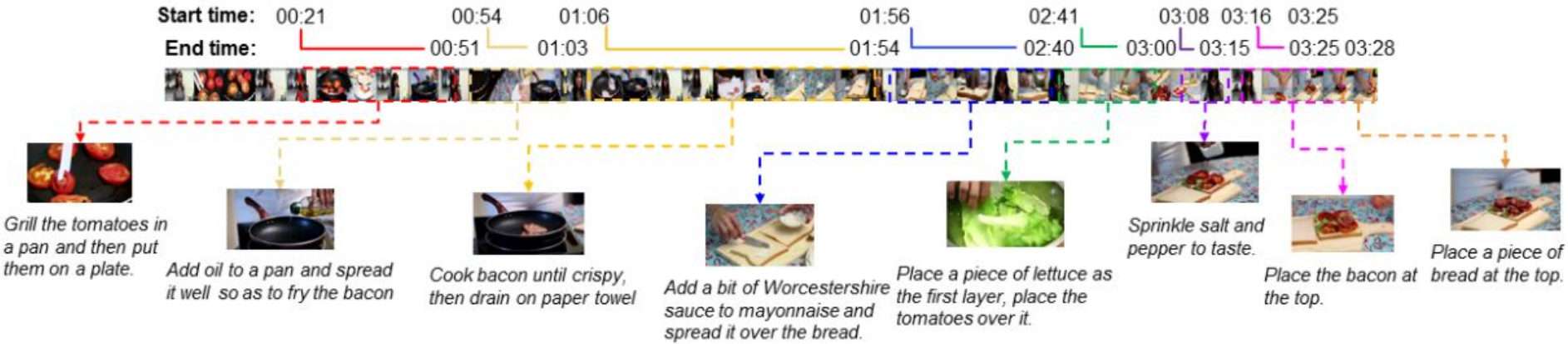


Desired Feature 1 - Audio Transcription

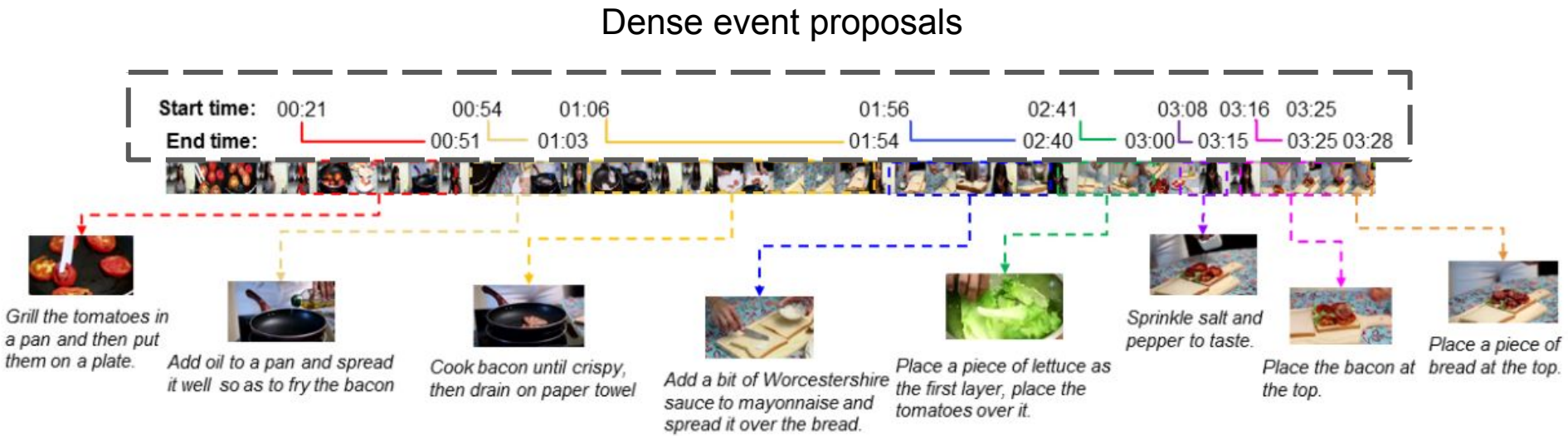
Requirements



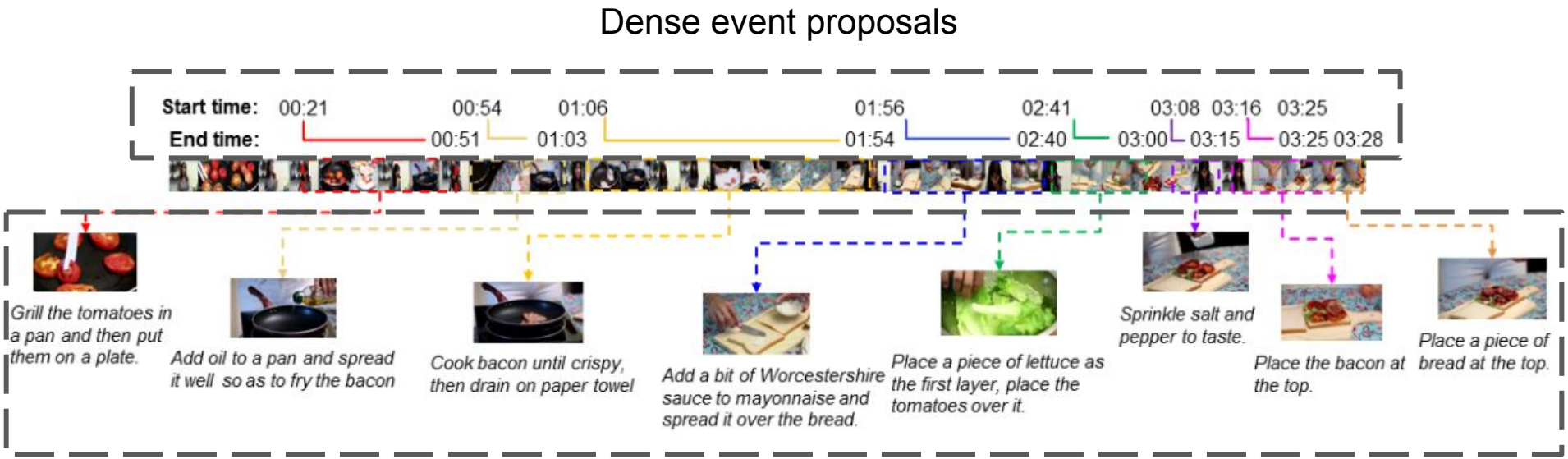
Dense Video Captioning



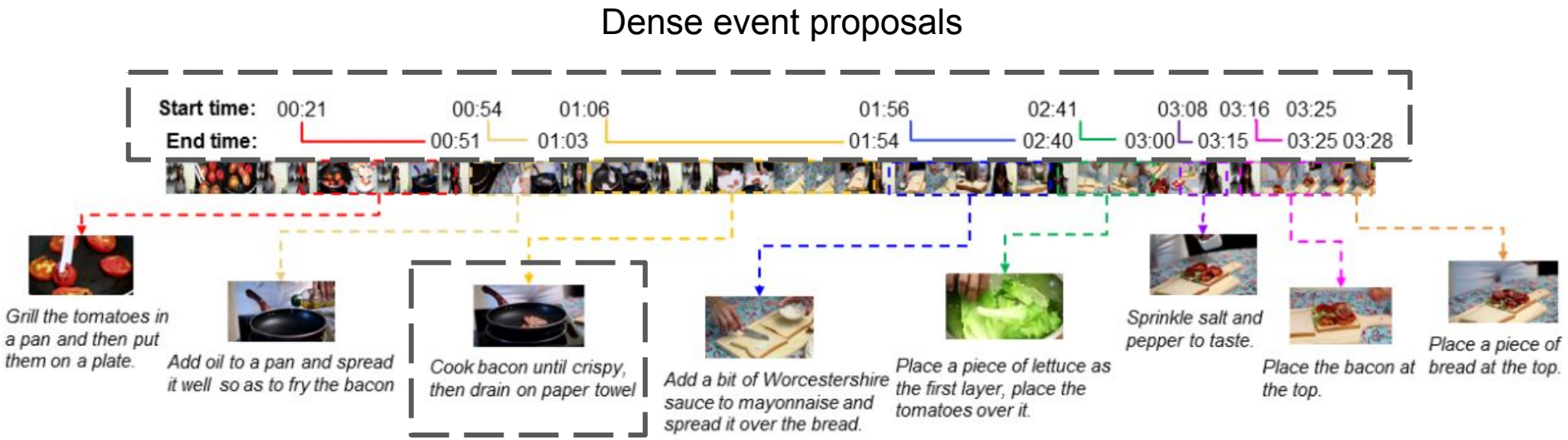
Dense Video Captioning



Dense Video Captioning



Dense Video Captioning



Each proposal represents one main event

Datasets

Public Datasets

YouCookII

ActivityNet Captions

Datasets

Too application specific!

Public Datasets

YouCookII

ActivityNet Captions

Datasets

Too application specific!

Public Datasets

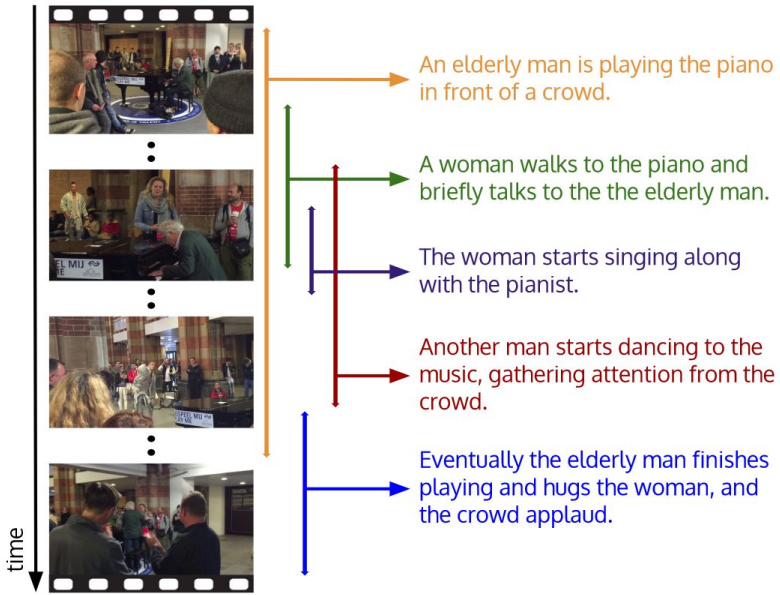
YouCookII

ActivityNet Captions

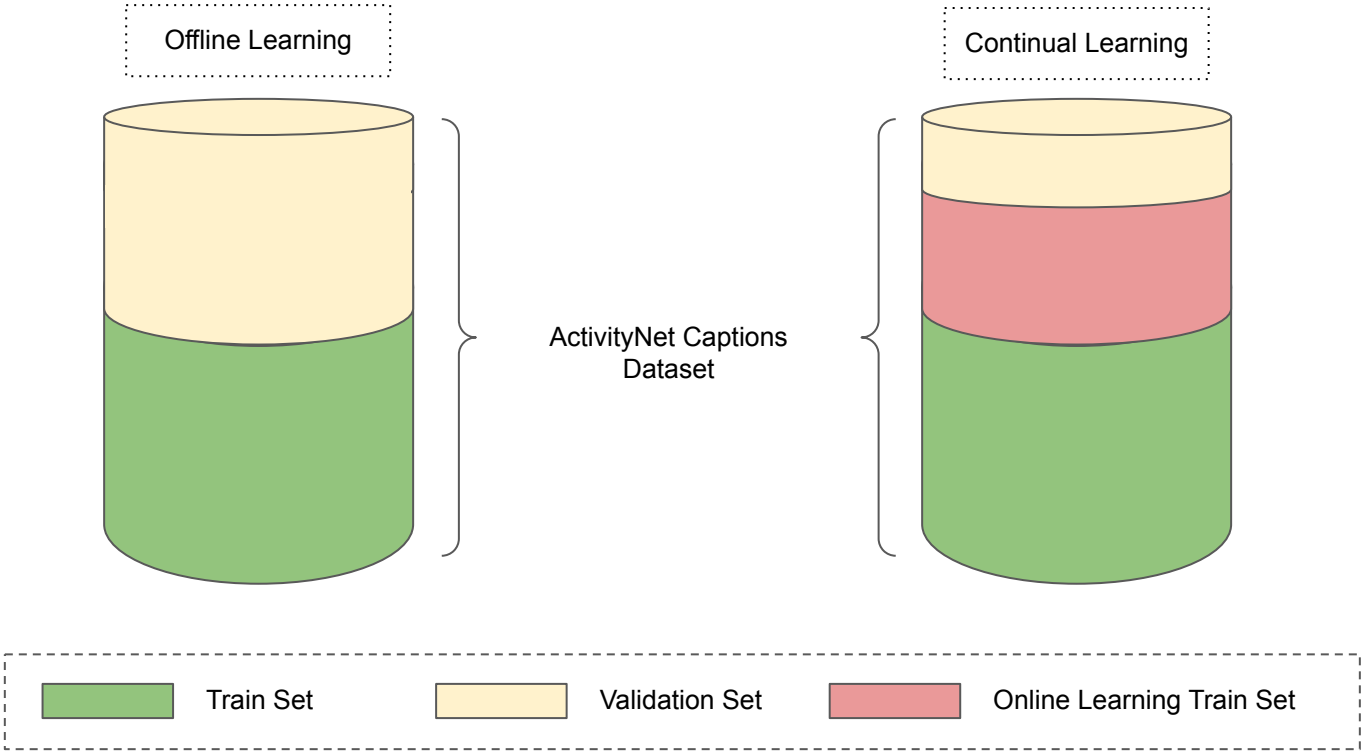
Encompasses many domains

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
Mun et al.	Yes	-	-	-	-	-	6.92
Krishna et al.	Yes	4.09	1.60	8.88	1.90	0.71	5.69
Li et al.	Yes	4.51	1.71	9.31	2.05	0.74	6.14
Zhou et al.	Yes	5.78	2.71	11.16	2.91	1.44	6.91
Wang et al.	Yes	-	-	10.89	2.27	1.13	6.10
Teng et al.	No	-	1.98	11.49	2.53	-	7.65
Iashin et al.	No	4.52	1.46	11.07	1.85	1.01	7.46
Rahman et al.	No	3.04	1.99	7.23	3.84	0.90	4.93
BMT	No	4.63	1.99	10.90	3.84	1.88	8.44
MMT [Ours]	**No**	**5.83**	**2.86**	**11.72**	**4.00**	**2.01**	**9.43**

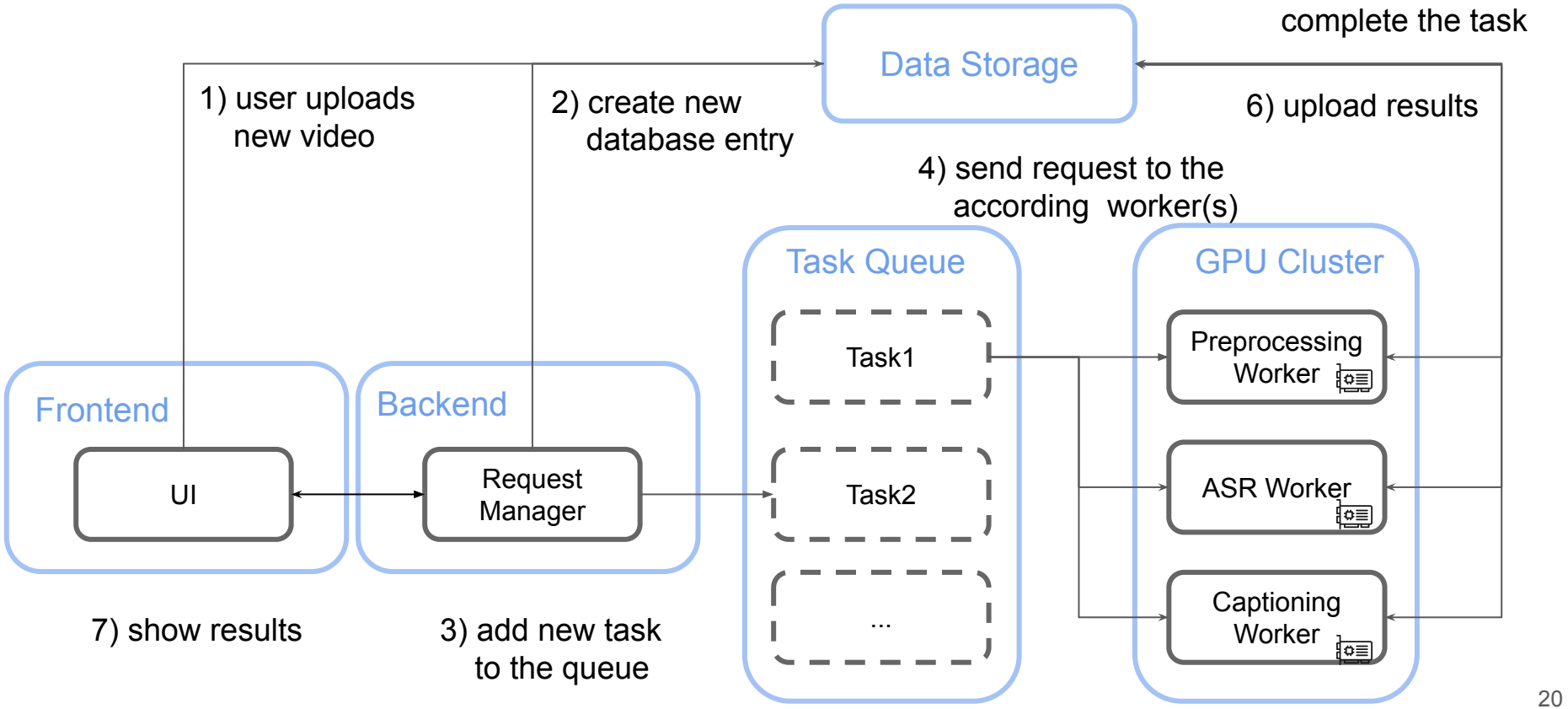
Multi-Modal Transformer (MMT) surpasses the state of the art in both GT and Learned Proposals **on all metrics**

Results - Continual Learning

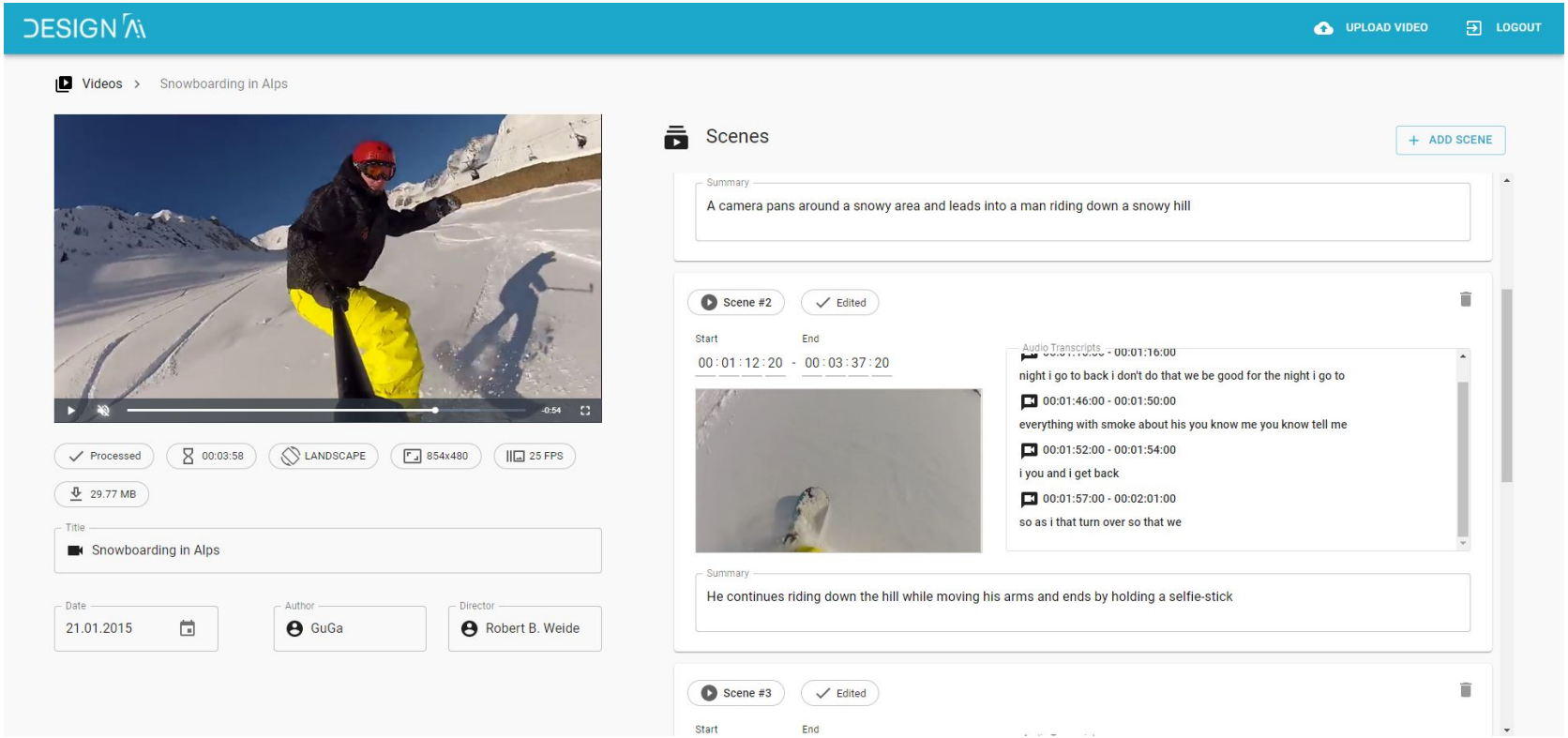
Method	B@3	B@4	METEOR
BMT(offline + Train Set)	4.9	2.3	10.51
BMT(offline + Train Set + 80% Val Set)	5.66	2.71	11.12
ALTC (ours)	**5.7**	**2.8**	**11.23**

- 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



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.”