MI_TJU at TRECVID 2023: Medical Video Question Answering

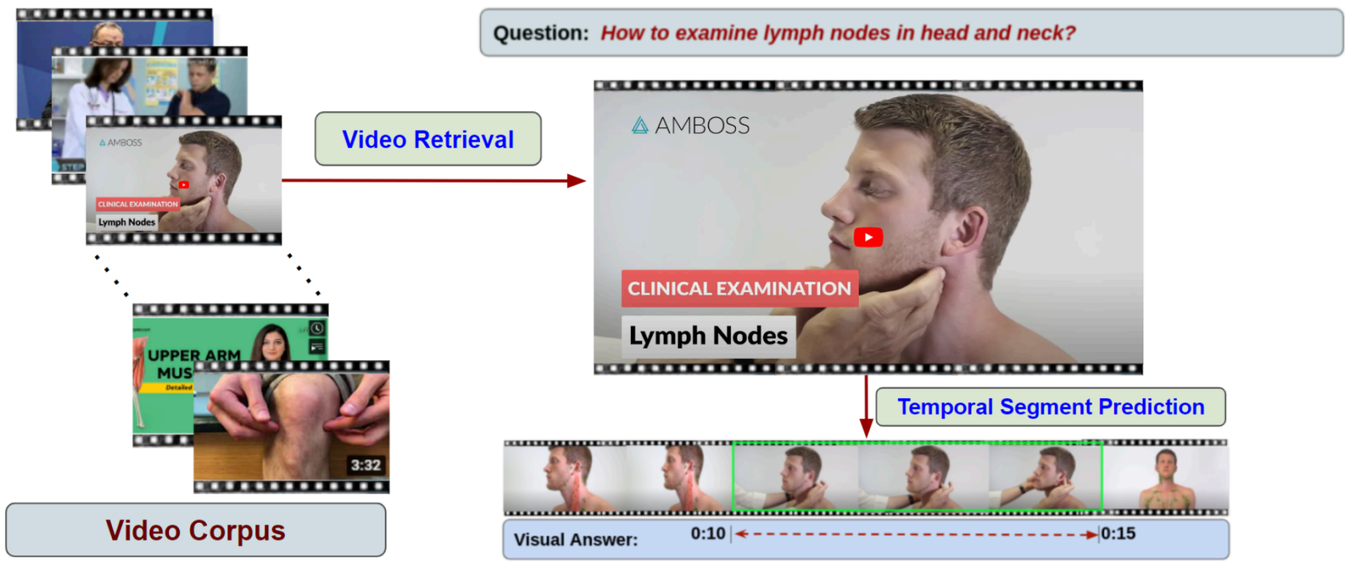
Zibo Xu, Weizhi nie, Qiang Li, Ning Xu, Yingchen Zhai, Zimu Lu, Anan Liu.

Multimedia Lab, Tianjin University.

Medical Video Question Answering

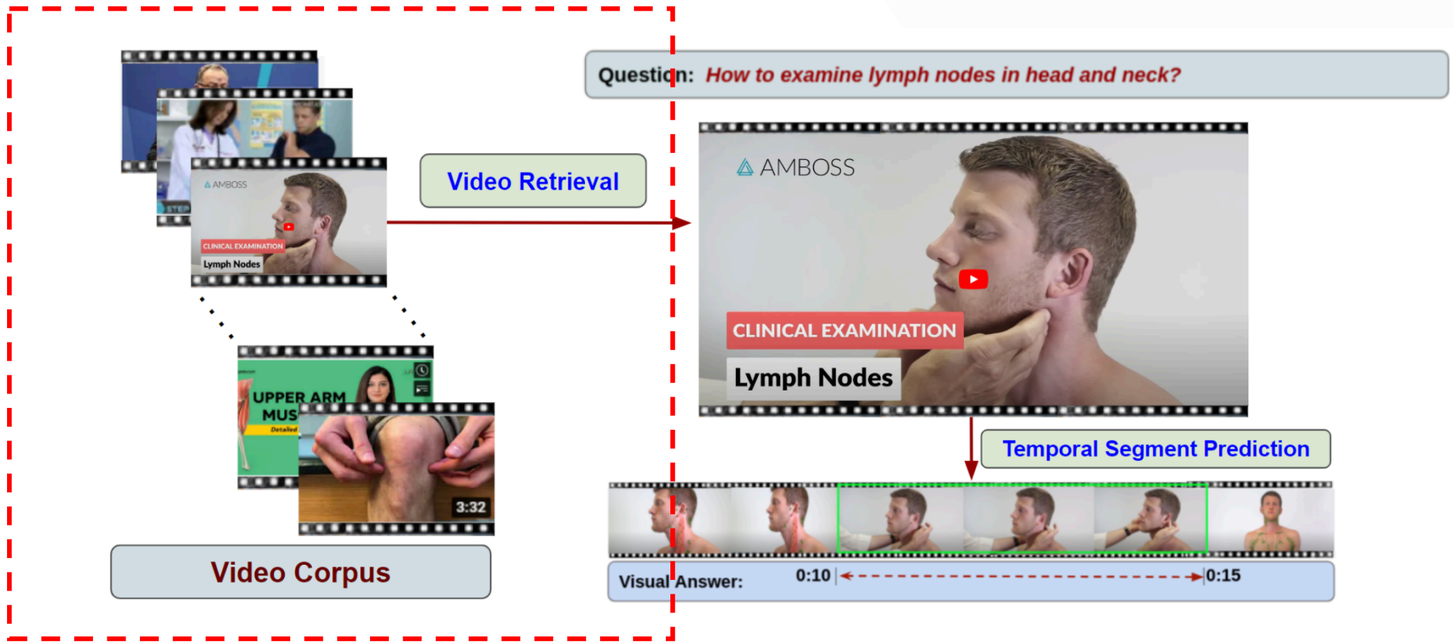
Video Corpus Visual Answer Localization VCVAL

- Video Retrieval
- Temporal Segment Prediction



Video Retrieval

- Goal: Identify the relevant videos according to the questions
- Video corpus: 12657 videos from Youtube.

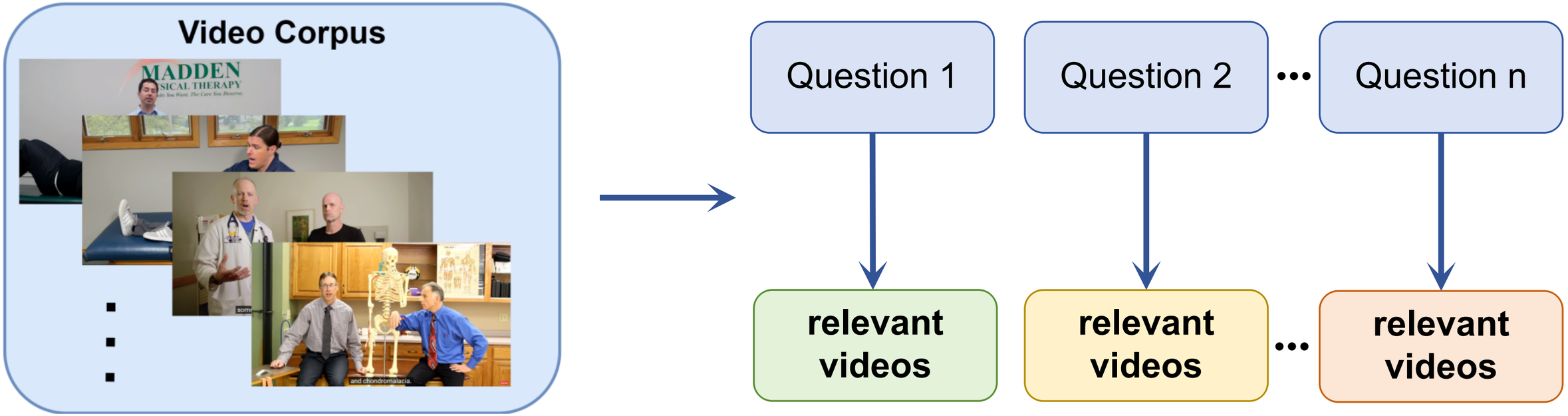


Video corpus

- Video corpus: 12657 videos from Youtube.
- Main Video type: medical instructional videos
- eg. the use of medical instruments, handling injuries, and providing care...
- Extensive subtitles
- Slower pace
- ...

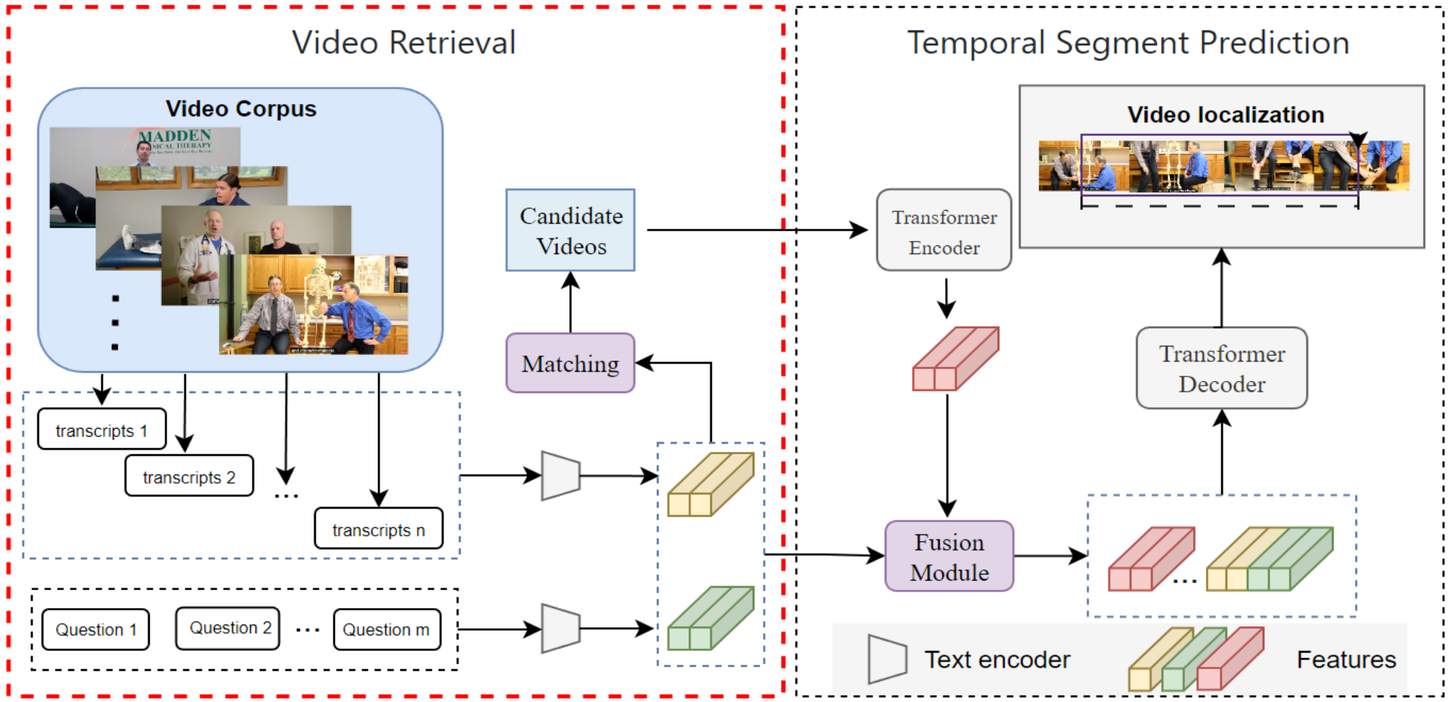
Video Retrieval

- How to summarize the main information of the video?
- Video transcripts / Video substitles



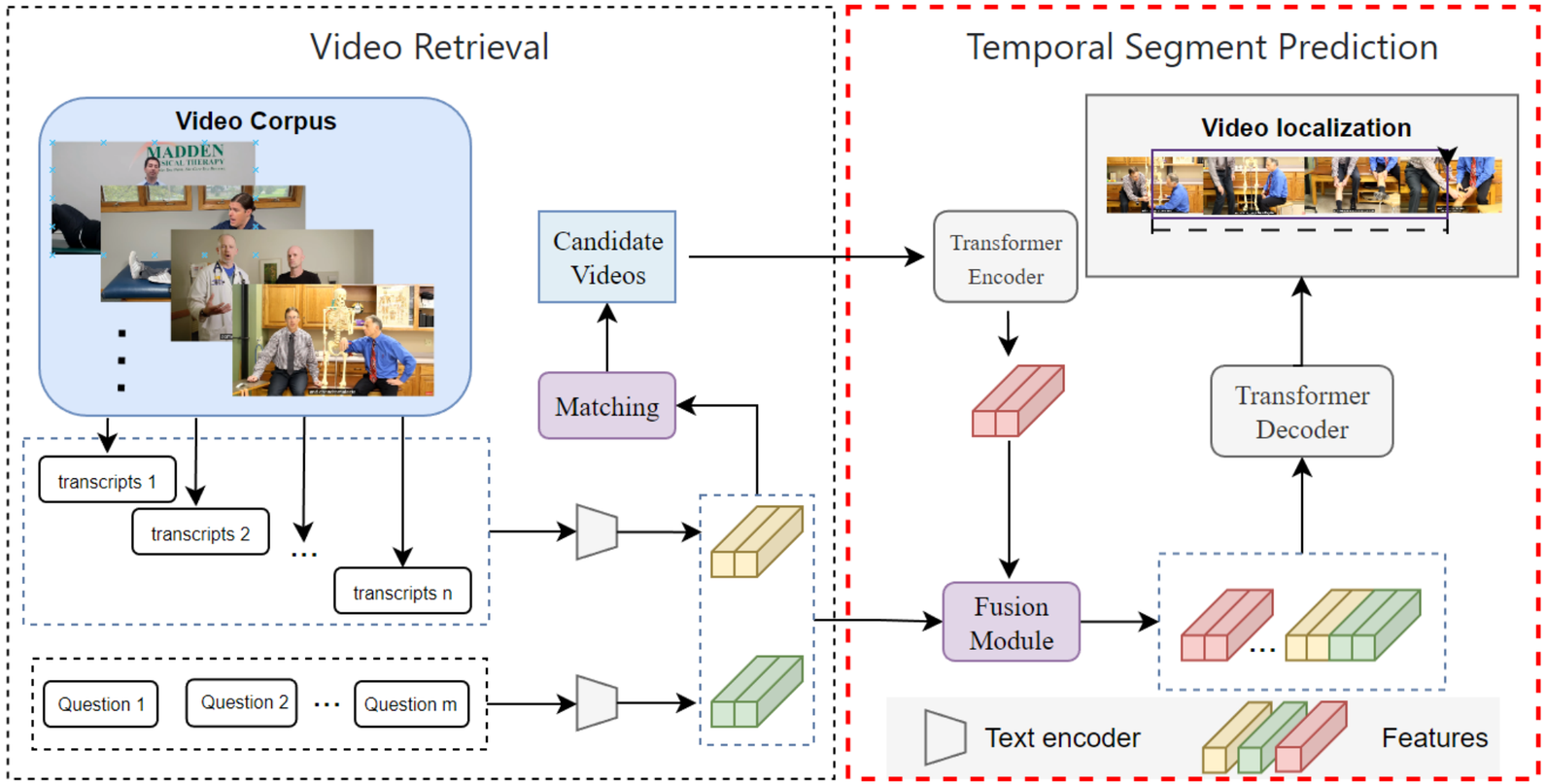
Video Retrieval

- A scoring task between video transcripts and questions
- Visual features are not imperative during video retrieval



Temporal Segment Prediction

- Locate the segments that provides the answer
- or relevant medical information is visually displayed



Results

Video Retrieval :

Table 1. The results of the video retrieval task.

Team	Run ID	MAP	R@5	R@10	P@5	P@10	nDCG
UNCWAI	run-2.json	0.1839	0.1903	0.1903	0.29	0.145	0.2858
VPAI	run-1.json	0.2427	0.2489	0.2489	0.31	0.155	0.3804
UNCWAI	run-1.json	0.3669	0.2221	0.3654	0.395	0.3575	0.5094
UNCWAI	run-3.json	0.3669	0.2221	0.3654	0.395	0.3575	0.5094
MI_TJU	run-1.json	**0.404**	**0.3549**	**0.4132**	**0.545**	**0.3625**	**0.5448**

Results

Temporal Segment Prediction :

Table 2. The results of the temporal segment prediction.

Team	Run ID	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
UNCWAI	run-1.json	10	7.5	0	9.32
UNCWAI	run-3.json	25	10	5	15.78
UNCWAI	run-2.json	42.5	32.5	22.5	31.37
VPAI	run-1.json	57.5	35	25	39.97
MI_TJU	run-1.json	**67.5**	**62.5**	**50**	**55.24**

Advantages

- Use the information of different modals
- Reduce the cost
- Guarantee the accuracy
- ...

Conclusion

- An efficient video retrieval method
- A localization method (cross-modal representations)
- Highest results in two subtasks

For the future

- From localization to question-answering?
- Some videos in corpus may not be publicly visible

Thank you for your attention!