Overview of the Medical Video Question Answering (MedVidQA) Task in TRECVID 2023

Deepak Gupta and Dina Demner-Fushman
LHNCBC, National Library of Medicine (NLM)
National Institutes of Health (NIH), MD, USA

Disclaimer

The views and opinions expressed do not necessarily state or reflect those of the U.S. Government, and they may not be used for advertising or product endorsement purposes.

Overview

- Introduction and Motivation
- Task Description
- Data Description
- Evaluation Metrics
- Participating Teams and Methods
- Results
- Analysis
- Conclusion

Introduction and Motivation

How can I ease my neck pain ?

Exercises for improving neck flexion

The following exercises build strength, relieve pain, and increase range of motion in your neck and upper back. You can do these exercises while sitting or standing.

Use slow, controlled movements and avoid forcing any movements. While moving your neck, keep the rest of your body still to maintain correct alignment and posture.

Neck flexion stretch

This exercise will help loosen your posterior neck muscles and reduce tightness.

1. Rest your arms alongside your body and engage your core muscles to stabilize your spine.
2. Draw your shoulder blades back and down.
3. Slowly draw your chin in toward your chest.
4. Hold for 15–30 seconds.
5. Do 2–4 repetitions.

The textual answer to this question will be hard to understand and act upon without visual aid.

Introduction and Motivation (cont'd…)

How can I ease my neck pain ?

The entire video can not be considered as the answer to the given question.

Instead, we want to refer to a particular temporal segment, or moment, from the video, where the answer is being shown, or the explanation is illustrated in the video.

Introduction and Motivation (cont'd...)

Question

How can I ease my neck pain ?

Textual Answer

Exercises for improving neck flexion

The following exercises build strength, relieve pain, and increase range of motion in your neck and upper back. You can do these exercises while sitting or standing.

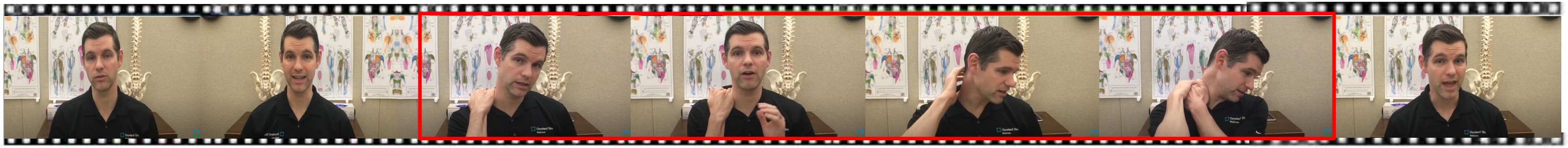
Use slow, controlled movements and avoid forcing any movements. While moving your neck, keep the rest of your body still to maintain correct alignment and posture.

Neck flexion stretch

This exercise will help loosen your posterior neck muscles and reduce tightness.

1. Rest your arms alongside your body and engage your core muscles to stabilize your spine.
2. Draw your shoulder blades back and down.
3. Slowly draw your chin in toward your chest.
4. Hold for 15–30 seconds.
5. Do 2–4 repetitions.

Video Containing Answer

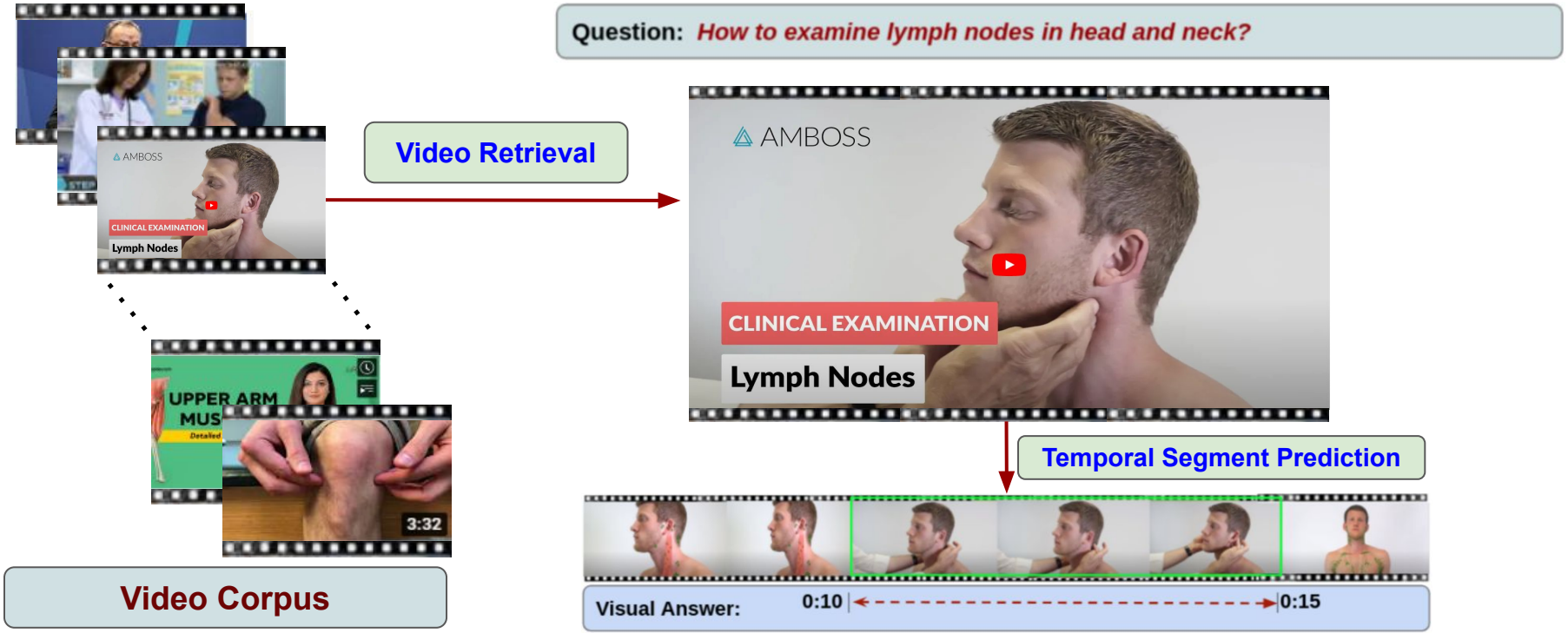


Visual Answer: 00:37 |◄- -►| 01:16

Task Description

Task A: Video Corpus Visual Answer Localization (VCVAL)

retrieve the appropriate video from the video collection and then locate the temporal segments where the answer to the medical query is being shown or the explanation is illustrated in the video.



Task Description (cont'd…)

Task B: Medical Instructional Question Generation (MIQG)

generate the instructional question for which the given video segment is the visual answer

Video segment

Question Generation

Instructional Question

How do I check lymph nodes in neck and head?

Carefully palpate the individual lymph node stations. To facilitate differentiation between lymph nodes and muscles, the area that is palpated should be as relaxed as possible. Every palpable lymph node is considered enlarged. If there is enlargement, pay attention to consistency, tenderness, mobility, the number of enlarged lymph nodes, and any erythema in the affected area.

Subtitle/Transcript

Applications

Video Corpus Visual Answer Localization

- First aid
- Medical emergency, and
- Medical education

Medical Instructional Question Generation

- To generate additional visual answer localization dataset
- Creating an automatic human-computer dialogue system
- Developing intelligent tutor systems in a multimodal environment

Datasets

VCVAL Task

- Video Retrieval
  - Developed a video corpus considering the videos from the '*Personal Care and Style*,' '*Health*,' and '*Sports and Fitness*' categories within the HowTo100M [1] dataset.
  - Follow the strategy discussed in [2] to select the medical instructional videos from the HowTo100M dataset.
  - A total of **12,657** medical instructional videos, which we considered as video corpus to retrieve the relevant videos against the query.

- Visual Answer Localization
  - **Training and Validation:** MedVidQA collections [2] consisting of 3,010 human-annotated instructional questions and visual answers from 899 health-related videos.
  - **Test dataset:** Sampled a total of 60 videos from the video corpus and created forty (40) medical instructional questions.
    - **Basic**: 20 questions
      - Formulated according to the annotation guidelines discussed in [2]. (video subtitles were visible to the annotators while creating the questions.)

[1] Miech, Antoine, et al. "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips." *Proceedings of the IEEE/CVF international conference on computer vision*. 2019.
[2] *Deepak Gupta, Kush Attal, and Dina Demner-Fushman. A Dataset for Medical Instructional Video Classification and Question Answering, Sci Data 10, 158 (2023)*

Datasets (cont'd…)

VCVAL Task (cont'd…)

- Visual Answer Localization (cont'd…)
  - **Training and Validation:** MedVidQA collections [2] consisting of 3,010 human-annotated instructional questions and visual answers from 899 health-related videos.
  - **Test dataset:** Sampled a total of 60 videos from the video corpus and created forty (40) medical instructional questions.
    - **Basic**: 20 questions
      - Formulated according to the annotation guidelines discussed in [2]. (video subtitles were visible to the annotators while creating the questions.)
    - **Visual Information Required (VIR):** 20 questions
      - With the following annotation guidelines
        - Formulate such a question that cannot be answered with just the subtitles or captions available within the video (i.e., just listening to the video alone and not watching should not be enough to answer the question).
        - The question should not be answered by reading the embodied text in the video.

[2] *Deepak Gupta, Kush Attal, and Dina Demner-Fushman. A Dataset for Medical Instructional Video Classification and Question Answering, Sci Data 10, 158 (2023)*

Datasets (cont'd…)

VCVAL Task (cont'd…)

- Judgement
  - Participants needed to retrieve up to 1000 relevant videos and their timestamps from a pool of **12,657** videos.
  - Performed the manual judgments of all the submitted videos (**943**) and visual answers by the participants.
  - A total of **eight** assessors were performed the judgement.
  - **Evaluate videos for relevance:**
    - **Definitely Relevant**
      - if it contains a visual segment that can be considered a complete visual answer to the question.
    - **Possibly Relevant**
      - if it contains a visual segment that can be considered a partial/incomplete visual answer to the question
    - **Not Relevant**
      - if the visual segments from the videos do not provide any visual answers to the question, the video can be marked as not relevant.

Datasets (cont'd…)

VCVAL Task (cont'd…)

- Judgement
  - The assessors were asked to provide the judgment with the following instructions:
    - Only provide the time stamps for definitely relevant and possibly relevant videos.
    - For each definitely relevant and possibly relevant video, provide the time stamps from the video that can be considered a visual answer.
    - The time stamps should be the shortest span in the video, which can be considered as a complete (for definitely relevant video) or partially complete (for possibly relevant video) visual answer to the question.
    - In case a video has multiple visual answers to the same question, assessors were asked to provide all the visual answers.

Datasets (cont'd…)

MIQG Task

- **Training and Validation:** MedVidQA collections [2] consisting of 3,010 human-annotated instructional questions and time stamps as visual answers from 899 health-related videos.
- **Test dataset:** Sampled a total of 100 videos from the video corpus and created forty (80) medical instructional questions.
  - **Basic**: 52 questions
    - Formulated according to the annotation guidelines discussed in [2]. (video subtitles were visible to the annotators while creating the questions.)
  - **Visual Information Required (VIR):** 28 questions
    - With the following annotation guidelines
      - Formulate such a question that cannot be answered with just the subtitles or captions available within the video (i.e., just listening to the video alone and not watching should not be enough to answer the question).
      - The question should not be answered by reading the embodied text in the video.

[2] *Deepak Gupta, Kush Attal, and Dina Demner-Fushman. A Dataset for Medical Instructional Video Classification and Question Answering, Sci Data 10, 158 (2023)*

Evaluation of the Systems

Video Corpus Visual Answer Localization

1. **Video retrieval**
   a. Mean Average Precision (MAP)
   b. Recall@k (k=5,10)
   c. Precision@k, (k=5,10)
   d. nDCG

   We follow the trec_eval evaluation library to report the performance of participating systems.

2. **Visual Answer Localization**

   a. A model prediction is considered correct if:
      i. at least one video out of n-predicted videos belongs to the ground-truth videos, and
      ii. the predicted temporal segment overlaps with the segment from the ground-truth video
   b. Intersection over Union (IoU) metric.

$$< R@n, IoU = \mu > = \frac{1}{N} \sum_{i=1}^{i=N} s(q_i, \mu), \text{ and}$$

$$s(q_i, \mu) = \begin{cases} 1, & \text{if } IoU(q_i) \geq \mu \\ 0, & \text{otherwise} \end{cases}$$

μ={0.3, 0.5, 0.7}
n={1,3,5,10}

Evaluation of the Systems (cont'd…)

Medical Instructional Question Generation (MIQG)

- Language generation metrics
  - BLEU [1]
  - ROUGE [2]
  - BERTScore [3]

[1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311–318, 2002.
[2] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pages 74–81, 2004.
[3] Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. "Bertscore: Evaluating text generation with bert." arXiv preprint arXiv:1904.09675 (2019).

Participants

Team Name	Team Affiliations	VCVAL	MIQG
ML_TJU	Tianjin University	✓	✗
VPAI	Hunan University/CAS	✓	✓
UNCWAI	University of North Carolina Wilmington	✓	✗
UMBCVQA	University of Maryland Baltimore County	✗	✓
doshisha_uzl	Doshisha University and University of Lubeck	✗	✓

Participating teams and their task participation in the MedVidQA.

Approaches

VPAI

VCVAL

- Cross-modal fusion method to address video retrieval.
  - Combining text features (question and subtitles) extracted from pre-training language models with visual features from image frames.
- Jointly trained the video corpus retrieval and visual answer localization subtasks using the global-span matrix.
  - A knowledge transfer strategy is adopted for enhancing the results.

MIQG

- Multi-modal video understanding approach.
- BLIP-2 to translate each frame in the videos.
- LLAMA-2 is used for question generation.

Approaches

MI_TJU

VCVAL

- Proposed a scoring mechanism to compute the relatenes of the video transcripts and question.
- Used a pre-trained visual encoder to extract video features, as well as a pre-trained text encoder to extract questions and subtitle features for each video.
- The extracted video, subtitles, and question features are fused into multimodal representations through the cross-modal attention mechanism.

Approaches

UNCWAI

VCVAL

- Three staged pipelined approach
  - **First stage**: text similarity is calculated between questions and video subtitles to select the most related videos.
  - **Second stage**: T5 model is fine-tuned to generate textual answers for the question based on video subtitles inputs.
  - **Three stage**: Embedding cosine similarity is calculated to locate the subtitle fragment.

Approaches

doshisha_uzl

MIQG

- Crossmodal vision-language foundation model mPLUG-Owl that generates summaries of the video clip contents using the video, its associated text transcript and a prompt as input.
- Generated summaries was passed to the Flan-T5 transformer model to generate the question.

UMBCQA

MIQG

- DEEP-CAM a multimodal approach.
- Uses video frames and video subtitles as inputs.
- Proposed the cross-attention multimodal encoder-decoder for question generation.

Results

VCVAL Task

Video Retrieval Subtask

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

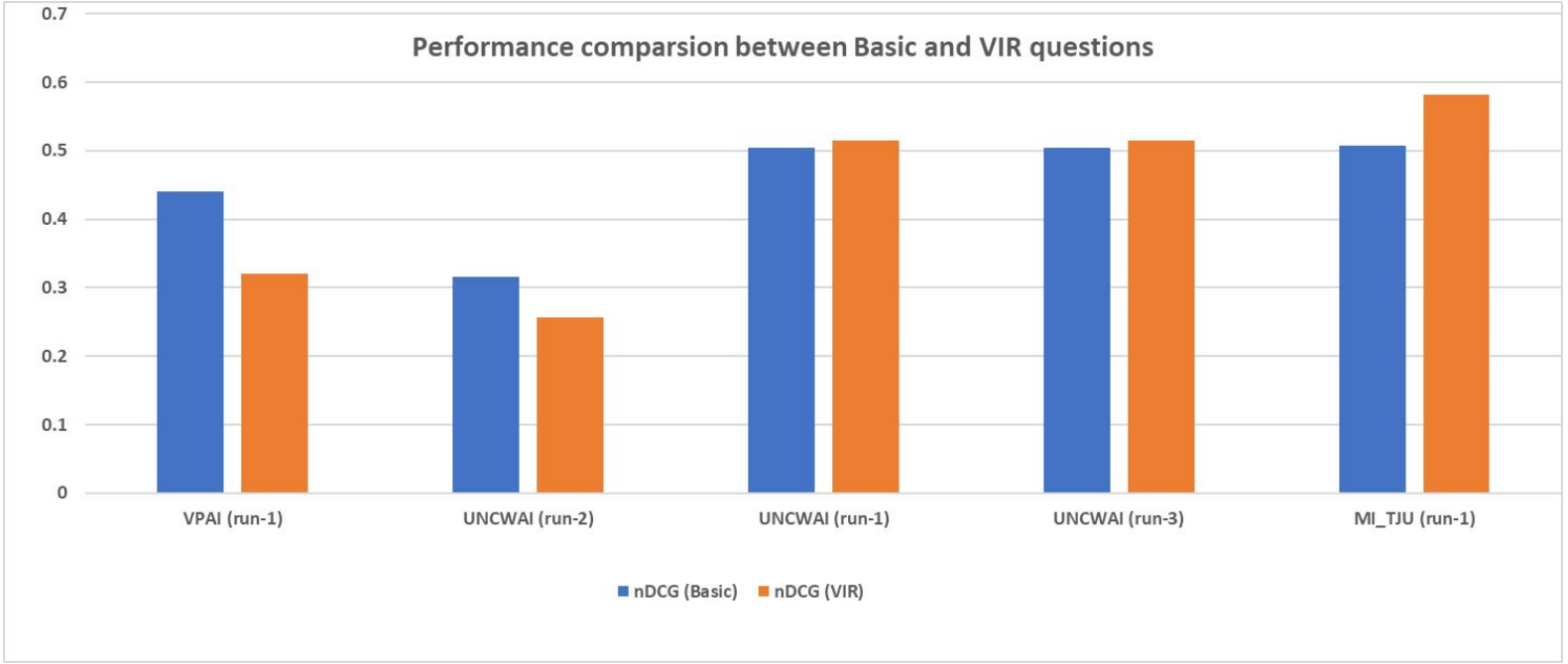
Visual Answer Localization Subtask

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

Results (cont'd…)

VCVAL Task

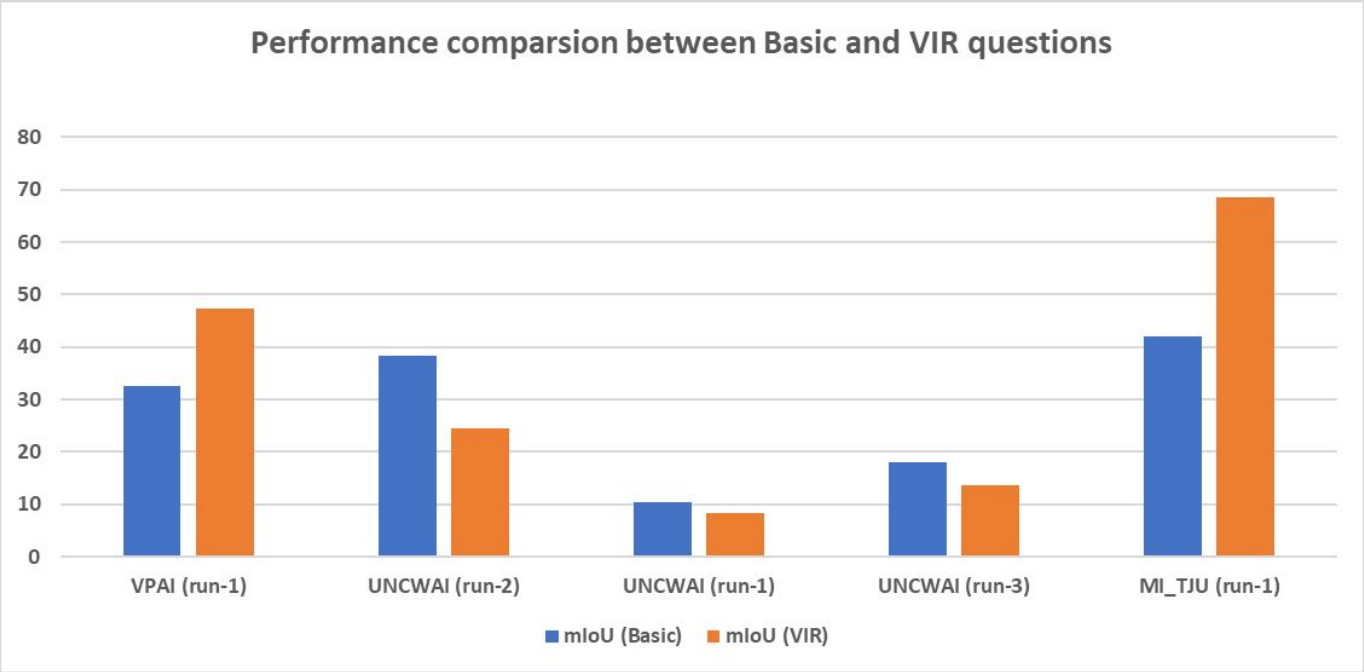
Video Retrieval Subtask



Results (cont'd…)

VCVAL Task

Visual Answer Localization Subtask



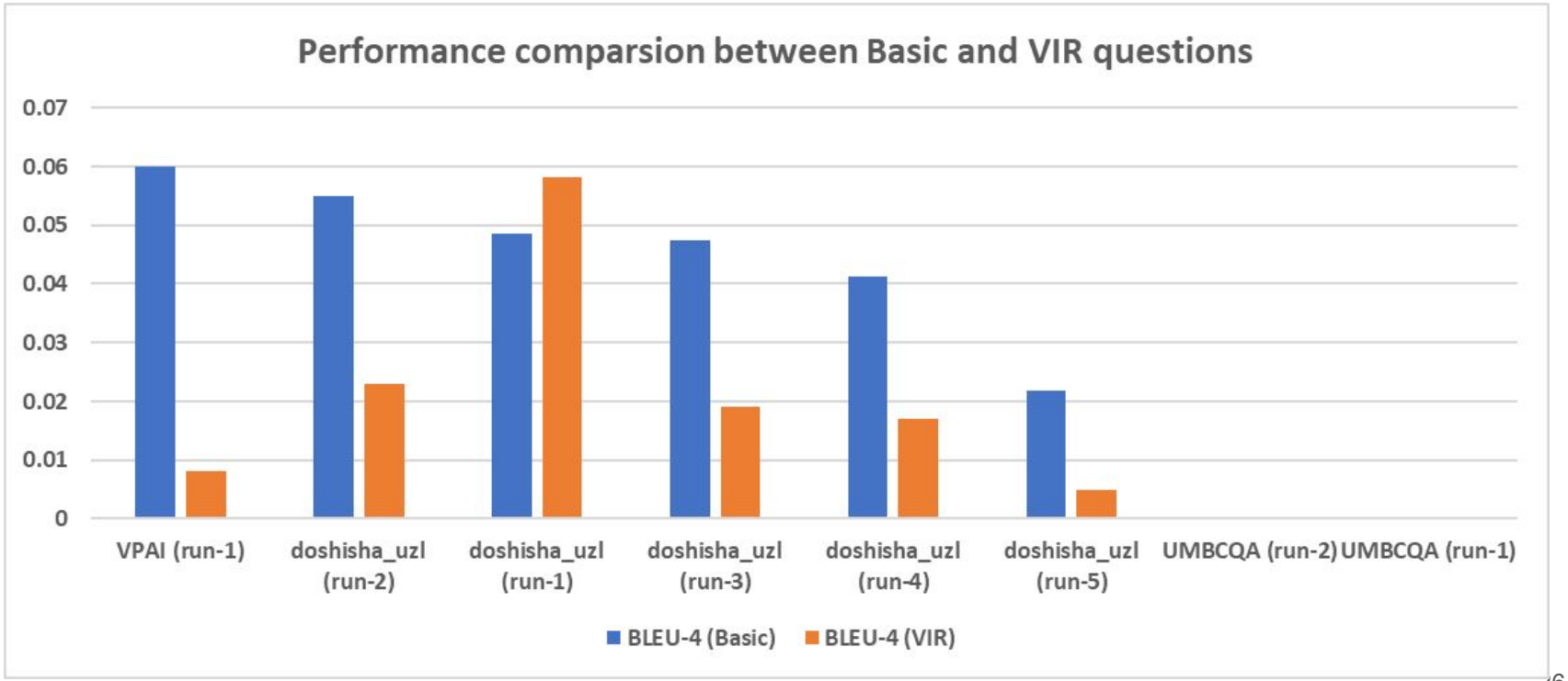
Results (cont'd…)

MIQG Task

Team	RunID	BLEU	BLEU-4	ROUGE-2	ROUGE-L	BERTScore
doshisha_uzl	run-1	0.15828	0.05153	0.27845	0.47822	0.91092
doshisha_uzl	run-2	0.14352	0.04546	0.24372	0.44667	0.90523
VPAI	run-1	0.12969	0.04331	0.27329	0.47979	0.90981
doshisha_uzl	run-3	0.14593	0.03875	0.27379	0.47418	0.91099
doshisha_uzl	run-4	0.13289	0.03404	0.24227	0.45566	0.9078
doshisha_uzl	run-5	0.093	0.01627	0.20113	0.4085	0.90248
UMBCQA	run-2	0	0	0.12253	0.26042	0.85332
UMBCQA	run-1	0	0	0.1317	0.31554	0.87683

Results (cont'd…)

MIQG Task



Analysis

- For the video retrieval task, team MI_TJU achieved the best result by proposing a scoring mechanism that considers mono-modal video and textual encoder.
- The maximum nDCG of 0.5448 signifies the challenges of instructional video retrieval for the medical domain.
- Three out of five runs submitted for video retrieval subtask performed better on VIR questions compared to the Basic questions.
- The team MI_TJU achieved the best performance (55.24 mIoU) on the VAL subtask with a multimodal approach.
- The team UNCWAI utilized only the textual modality for the VCVAL task and reported a performance of 31.37 (mIoU).
- Two out of five runs submitted for the VAL subtask performed better on Basic questions compared to the VIR questions.
- The team doshisha_uzl achieved the best performance (0.05153 BLEU-4) on the MIQG task with a combination of mono and multimodal approaches.
- One out of eight runs submitted for the MIQG task performed better on VIR questions compared to the Basic questions.

Conclusion

- Introduced two new tasks in multimodal understanding and generation in the medical domain.
- Discussed the MedVidQA@TRECVID 2023 task, datasets, evaluation metrics, and key results.
- A total of five teams participated in the MedVidQA@TRECVID 2023 and submitted 5 and 8 individual runs for the VCVAL and MIQG tasks, respectively.
- The system's performance shows room for improvement for both tasks.
- Human evaluation is needed to evaluate the performance of question-generation approaches.
- The introduced datasets and manual judgement will foster research toward designing systems that can understand medical videos and provide assistance for first aid and medical emergency questions.

Thank you for Your Attention

Questions and feedback: deepak.gupta@nih.gov