Overview of TREC 2024 Medical Video Question Answering (MedVidQA) Track

Deepak Gupta and Dina Demner-Fushman

National Library of Medicine, NIH

1 Overview

One of the key goals of artificial intelligence (AI) is the development of a multimodal system that facilitates communication with the visual world (image and video) using a natural language query. Earlier works [1, 2, 3, 4, 5, 6], on medical question answering primarily focused on textual and visual (image) modalities, which may be inefficient in answering questions requiring demonstration. In recent years, significant progress has been achieved due to the introduction of large-scale language-vision datasets and the development of efficient deep neural techniques that bridge the gap between language and visual understanding. Improvements have been made in numerous vision-and-language tasks, such as visual captioning [7, 8], visual question answering [9], and natural language video localization [10]. Most of the existing work on language vision focused on creating datasets and developing solutions for open-domain applications. We believe medical videos may provide the best possible answers to many first aid, medical emergency, and medical education questions. With increasing interest in AI to support clinical decision-making and improve patient engagement [11], there is a need to explore such challenges and develop efficient algorithms for medical language-video understanding and generation. Toward this, we introduced new tasks to foster research toward designing systems that can understand medical videos to provide visual answers to natural language questions, and are equipped with multimodal capability to generate instruction steps from the medical video. These tasks have the potential to support the development of sophisticated downstream applications that can benefit the public and medical professionals.

2 Tasks

- **Task A: Video Corpus Visual Answer Localization (VCVAL).** Given a medical query and a collection of videos, the task aims to retrieve the appropriate video from the video collection and then locate the temporal segments (start and end timestamps) in the video where the answer to the medical query is shown or the explanation is illustrated in the video. The proposed VCVAL task can be considered as video retrieval and then find a series of "*medical instructional activity-based frame localization*" where a potential solution first searches for all medical instructional activities for a given medical query and then locates the activities in an untrimmed medical instructional video. This task is the extension of the MVAL task introduced in MedVidQA-2022 [12], where we only focused on locating the segment from a given video. In contrast, the VCVAL task deals with relevant video retrieval followed by the visual answer segment localization (*cf.* Figure 1). The video retrieval system requires the ability to identify the medical instructional video and retrieve the most relevant video for the health-related query.
- **Task B: Query-Focused Instructional Step Captioning (QFISC).** Given a medical query and a video, this task aims to generate step-by-step textual summaries of the visual instructional segment that can be considered as the answer to the medical query. The proposed QFISC task can be considered an extension of the visual answer localization task, where the system needs to locate a series of instructional segments that serve as the answer to the query. The QFISC requires identifying the boundaries of instructional steps and generating a caption for each step. This task comes under multimodal generation, where the system has

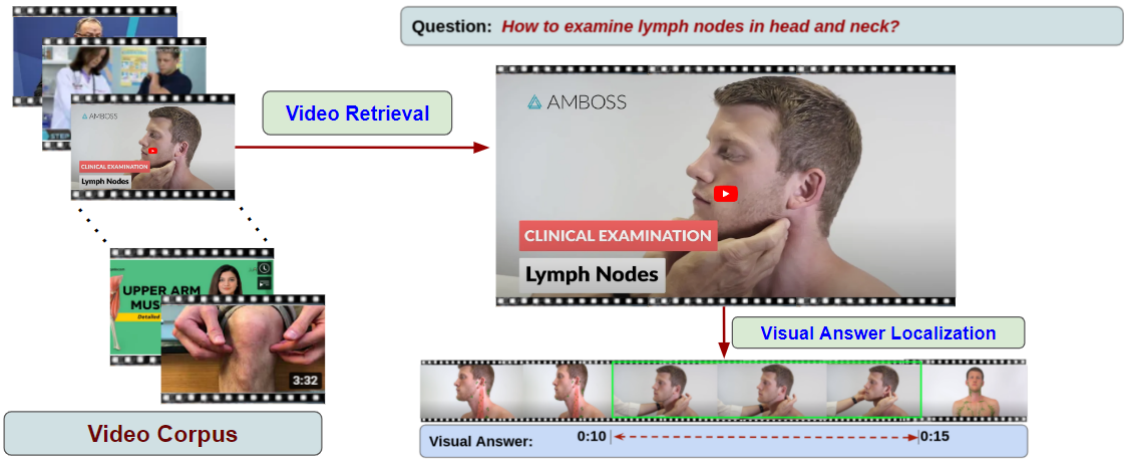

Figure 1: Visualization of the proposed video corpus visual answer localization (VCVAL) task. The VCVAL task consists of two sub-tasks: video retrieval and visual answer localization.

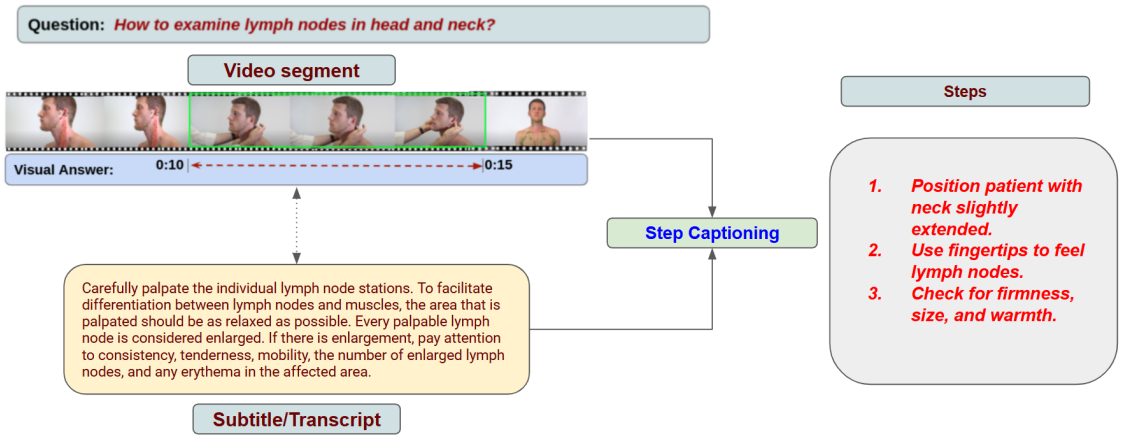

Figure 2: Visualization of the proposed query-focused instructional step captioning (QFISC) task.

to consider the video (visual) and subtitle (language) modality to generate (*cf.* Figure 2) the natural language caption. The applications of the QFISC task include the accessibility of video with the textual query, easy indexing of the videos with corresponding step caption, multimodal retrieval, etc.

3 Datasets

3.1 VCVAL

The VCVAL task comprises two subtasks: video retrieval and visual answer localization. For the video retrieval, we developed a video corpus considering the videos from '*Personal Care and Style*,' '*Health*,' and '*Sports and Fitness*' categories within the HowTo100M [13] dataset. We follow the strategy discussed in [14] to select the medical instructional videos from the HowTo100M dataset. We also added the video corpus of size 12,657 released for the TRECVID 2023 MedVidQA track [15]. This process yielded a total of 48,605 medical instructional videos, which we considered as a video corpus to retrieve the relevant videos against the query. To facilitate training and validation of the visual answer localization system, we provided MedVidQA collections [16] consisting of 3,010 human-annotated instructional questions and visual answers from 899 health-related videos.

We sampled a total of 150 distinct videos on different topics from the video corpus and created 52 medical instructional questions following the guidelines discussed in [16]. In particular, we follow the following guidelines:

- The question is looking for an instructional answer, and

Objective: **1)** to create step-by-step, well-formed steps of the visual instructional segment that can be considered the answer to the medical query, and **2)** provide the time stamps (start and end) where the step is being shown or the explanation is illustrated in the video.
Please note the following while creating the steps:

- A video can have multiple visual answer segments.
- For each visual answer segment, you need to create instructional steps.
- You can start writing a step with the action verb.

Well-formed Steps: We define a step to be a well-formed natural language step if it satisfies the following:

1. The step does not contain spelling errors.
2. A step should not be longer than seven (7) words.
3. Each step should clearly signify an action being demonstrated in the visual segment.
4. It should be a simple step focusing on a single action. Please avoid writing two actions in a single step.

Examples:

- Stabilize the arm with the board ✓
- Tie the elbow to the board ✓
- Stabilize the arm and tie the elbow with the board ✗

Time-stamps: A timestamp is a way to link to a specific step in the video. A time stamp needs to be entered in MM: SS format. E.g., 02:30 (2 min 30 sec), 10:56 (10 min 56 sec)

Figure 3: Annotation Guidelines for manual step captioning of QFISC task.

- The answer should be shown or the explanation must illustrated in the video for the created question, and
- A video snippet is necessary to answer the question, and
- The answer should not be given as text (e.g. definitional question) or spoken information without visual aid.

We consider the created questions as our test set for the VCVAL task. The created test questions are shown in Table 1.

3.2 QFISC

We utilized the open domain HIREST dataset [17] to train the system for the QFISC task. HIREST comprises 3.4K text-video pairs sourced from an instructional video dataset. Among these, 1.1K videos are annotated with moment spans pertinent to text queries. Each moment is further dissected into key instructional steps, complete with captions and timestamps, resulting in a total of 8.6K step captions.

To create a test collection, we used the manually annotated visual segments of the VCVAL task from TRECVID 2023 MedVidQA track [15]. We sampled 140 visual segments considering the distinct topics from the ground-truth collection of the TRECVID 2023 MedVidQA-VCVAL task and formulated a 90 visual segment-steps pair dataset. We provided the guidelines shown in Fig. 3 to the annotators to formulate the step captions. By following the guidelines, we created 90 visual segment-steps pair which was provided to the participants as test segments to generate the step captions.

QId	Question
Q1	How to assess for a dislocated hip through physical examination?
Q2	How to perform a simple exercise to relieve tension in the neck muscles?
Q3	How to get did of toenail fungus?
Q4	How to properly clean ear canal?
Q5	How to properly use a nebulizer?
Q6	How to reduce bleeding and inflammation in the gums?
Q7	How to apply heat or cold therapy to alleviate sciatica pain in the lower back?
Q8	How to handle a groin strain using exercise?
Q9	How to empty a urinary catheter bag safely and properly?
Q10	How to help a person in removing their prosthetic limb?
Q11	How to perform rescue breathing on an infant with a tracheal tube?
Q12	How to use an inhaler with a spacer?
Q13	How to prepare hydrogen peroxide solution for contact lenses?
Q14	How should contact lenses be correctly inserted?
Q15	How to use diaphragmatic breathing to manage asthma?
Q16	How can ice be used to treat calf tendonitis?
Q17	How do you use a water flosser to remove tonsil stones?
Q18	How to floss between an implant and the gum line?
Q19	How do you use a heat pack to help a stiff jaw?
Q20	How do you clean around wisdom teeth?
Q21	How to stop gums from bleeding?
Q22	How do you clear your nose with a nasal spray?
Q23	How to do leg squats to relieve a herniated disc?
Q24	How to stand up from a seated position after ACL reconstruction surgery?
Q25	How to do a self-exam on the lymph nodes in the neck?
Q26	How to do upright row-to-plank exercises for rotator cuff?
Q27	How to relieve pain from a pinched nerve in the neck?
Q28	How to brush your teeth with an electric toothbrush?
Q29	How to brush your teeth if you have braces?
Q30	How can I remove stains from my teeth?
Q31	How to clean your teeth with a handheld pressure washer?
Q32	How is the knee checked for arthritis?
Q33	How do you get rid of scar tissue in the foot?
Q34	How to stretch hip muscles to relieve hip pain?
Q35	How to massage around the ear to lessen jaw pain?
Q36	How do you stretch hamstrings for pain relief?
Q37	How to stretch to ease lower back pain from a pinched nerve?
Q38	How are squats done for relief from knee pain?
Q39	How can cracked feet be treated?
Q40	How is heat therapy used to treat stiff joints?
Q41	How is tape applied to the bicep for pain relief?
Q42	How do you self-massage the jaw for TMJ pain relief?
Q43	How is neck flexion measured?
Q44	How is CPR performed on a child?
Q45	How do you test for a torn Achilles tendon?
Q46	How do you check for a tight jaw?
Q47	How to do hamstring exercises to reduce knee swelling?
Q48	How to use a defibrillator for chest compressions?
Q49	How are metacarpal flexion and extension measured?
Q50	How do you ascend stairs with crutches?
Q51	How to do a glucagon injection?
Q52	How do you stop bleeding with a first aid kit?

Table 1: Test question for the Video Corpus Visual Answer Localization task.

4 Judgments

4.1 VCVAL

The participants were asked to retrieve the relevant videos (up to 1,000) for each question from the video corpus of having 48,605 videos. Additionally, the participants also had to provide the start and end timestamps from each retrieved video against a given question, which can be considered a visual answer to the question. In order to judge the relevant videos and corresponding visual answers in the videos, we performed the manual judgments by applying the pooling strategy with pool size=25 (first 10 with probability 1, next 5 with probability 0.3, next 5 with probability 0.2, and next 5 with probability 0.1) of all the submitted videos (12,184) and visual answers by the participants. We instructed a total of three assessors with the following guidelines to assess the video:

Objective:

1. To judge relevant videos with respect to medical/healthcare instructional questions. A video can be called relevant if it has a visual answer to the question.
2. For each relevant video, provide the time stamps (start and end) where the answer is being shown or the explanation is illustrated in the video.

Evaluating videos for relevance : The videos are judged as being *"Definitely Relevant"*, *"Possibly Relevant"*, or *"Not Relevant"* to the given question. The assessors were presented with videos from the submitted runs. They were instructed to determine if the video was definitely relevant, possibly relevant, or not relevant to the question. In general, a video is definitely relevant if it contains a visual segment that can be considered a complete visual answer to the question. A video can be considered possibly relevant if it contains a visual segment that can be considered a partial/incomplete visual answer to the question. If the visual segments from the videos do not provide any visual answers to the question, the video can be marked as not relevant. 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.

4.2 QFISC

We also performed a manual assessment of the participants' submitted step captions. To assess the generated steps, we performed manual judgments of all the steps to the participants' submitted runs. We instructed an assessor with the guidelines provided in Figure 4 to assess the system-generated steps.

5 Evalaution

5.1 Metrics for VCVAL Task

The VCVAL task consists of two sub-tasks: video retrieval (VR) and visual answer localization (VAL). We evaluated the performance of the video retrieval system in terms of Mean Average Precision (MAP), Recall@k, Precision@k, and nDCG metrics with $k = \{5, 10\}$. We follow the `trec_eval`[1] evaluation library to report the performance of participating systems.

For the VAL task, if the predicted (retrieved by the system) video is from the list of relevant videos (marked by the assessor; we called it ground-truth video), then we compute the overlap between the retrieved and relevant video by the following metrics:

[1] https://github.com/usnistgov/trec_eval

Objective: To assess the system-generated steps on human evaluation criteria such as completeness, accuracy, and coherency. The assessor must provide the score on the Likert scale (1-5, 1 signifies least and 5 denotes most) for the system-generated steps using the human evaluation criteria listed below:

- **Completeness**: Whether the generated steps contain all the necessary steps to perform the certain task to achieve the desired output.
- **Accuracy**: Whether the generated steps are accurate to what is being demonstrated in the visual segment.
- **Coherency**: Whether the steps logically connect to each other. The steps should follow the correct ordering for the performed task to achieve the desired output.

Figure 4: Human assessment guidelines for evaluating the system generated steps in QFISC task.

1. **Mean Intersection over Union (mIoU):** For a given question q_i, IoU is computed as the ratio of intersection area over union area between predicted and ground-truth temporal visual answer segments. It ranges from 0 to 1. A larger IoU means the predicted and ground-truth temporal visual answer segments match better, and IoU = 1.0 denotes an exact match. The mIoU is defined as the average temporal IoUs for all questions (N) in the test set. Formally,

$$\text{mIoU} = \frac{1}{N}\sum_{i=1}^{i=N} \texttt{IoU}(q_i) \tag{1}$$

2. **IoU = μ** is another metric used to evaluate the performance of the VAL system. It denotes the percentage of questions for which, out of the top-n retrieved temporal segments, at least one predicted temporal segment having IoU with ground truth is larger than μ. Formally,

$$< R@n, IoU = \mu > = \frac{1}{N}\sum_{i=1}^{i=N} s(q_i, \mu), \text{ and} \tag{2}$$

$$s(q_i, \mu) = \begin{cases} 1, & \text{if } IoU(q_i) \geq \mu \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

We evaluated the participants' submission by considering $\mu = \{0.3, 0.5, 0.7\}$ and for brevity, we denote the $< R@n, IoU = \mu >$ metric with IoU=μ $n = \{1, 3, 5, 10\}$

Algorithm 2 Calculate Alignment Score

Input: Predicted step *pred_step*, Ground truth step *gt_step*, Overlap weight $\alpha = 0.5$, ROUGE weight $\beta = 0.5$
Output: Alignment score S

```
overlap ← calculate_time_overlap(pred_step, gt_step)   ▷ Compute time overlap between predicted and ground truth steps
rouge_score ← rouge.compute(
        predictions = [pred_step["caption"]],
        references = [gt_step["caption"]])              ▷ Calculate ROUGE-L score for captions
S ← α × overlap + β × rouge_score["rougeL"]             ▷ Weighted combination of overlap and ROUGE-L
return S
```

5.2 Metrics for QFISC Task

We evaluated the performance of the step captioning task on two fronts: **(1)** how close the system-generated step caption is to the ground truth step captions, and **(2)** how well the predicted step segment aligns with the ground truth step segment. We measure the closeness in two ways:

1. With the help of predicted timestamps and sentence level similarity (ROUGE-L) of the step, we align (*c.f.* Algo 1, 2) a generated step to one of the ground truth steps. Towards this, we define the following:

Algorithm 1 Align Predicted Steps with Ground Truth Steps

Input: Predicted steps P, Ground truth steps G, Overlap threshold θ
Output: True Positives (TP), False Positives (FP), False Negatives (FN)

Initialize $TP \leftarrow 0$, $FP \leftarrow 0$, $FN \leftarrow 0$
Set $g\_index \leftarrow 0$ ▷ Start from the first ground truth step
for each p_step **in** P **do**
    $best\_score \leftarrow -1$, $best\_g\_index \leftarrow -1$ ▷ Track highest alignment score and index
    for each g_step **in** $G[g\_index :]$ **do**
        $score \leftarrow$ calculate_alignment_score(p_step, g_step)
        if $score > best\_score$ **and** $score \geq \theta$ **then**
            $best\_score \leftarrow score$
            $best\_g\_index \leftarrow g\_index$
        end if
    end for
    if $best\_g\_index \neq -1$ **then**
        $TP \leftarrow TP + 1$
        $g\_index \leftarrow best\_g\_index + 1$ ▷ Advance to next ground truth step
    else
        $FP \leftarrow FP + 1$ ▷ No match found for this predicted step
    end if
end for
$FN \leftarrow |G| - TP$ ▷ Count remaining ground truth steps as False Negatives
return TP, FP, FN

- TP (True Positives): Represents the count of predicted steps that are present in the ground truth steps.
- FP (False Positives) : Represents the count of predicted steps that are not present in the ground truth steps.
- FN (False Negatives): Represents the count of ground truth steps that are not present in the predicted steps.

Following this, we compute the precision, recall, and f-score.

2. We use the *n-gram* matching metrics: BLEU [18], ROUGE [19], METEOR [20] and SPICE [21]. Additionally, we use sentence-level embedding-based metrics: BERTScore [22] as they capture the semantic similarity between the generated and ground truth captions. All the n-gram metrics are computed between predicted and ground-truth step captions.

To compute the alignment between the predicted step segments and the ground truth step segments, we use the intersection over union (IoU) metric. For a given step, IoU is computed as the ratio of the common segment to the union between the predicted and ground-truth segments. It ranges from 0 to 1. For the shorter step, where the step segment lasts, say, only 1-2 seconds, if the system-generated step segment does not match with the ground-truth segment, the system may end up with IoU=0. To deal with such a situation, we will use relaxed IoU, where we extend the segments by λ before computing the IoU. We compute the mean of the IoU for all the segments in the test set.

6 Participating Teams

We use the NIST-provided Evalbase platform[2] to release the datasets, registration, and submissions of the participating teams. In total, 7 teams participated in the MedVidQA track and submitted 17 individual runs for the tasks. We have provided (*cf.* Table 2) the team name, affiliations, and their participation in VCVAL and QFISC tasks.

[2] https://ir.nist.gov/evalbase

Team Name	Team Affiliations	VCVAL	QFISC
TJUMI	Tianjin University	✓	✗
PolySmart	City University of Hong Kong	✓	✓
NCSU	North Carolina State University	✓	✗
NCstate	North Carolina State University	✓	✗
UNCW	University of North Carolina Wilmington	✓	✗
UNCC	University of North Carolina at Charlotte	✓	✗
DoshishaUzlDfki	Doshisha University and University of Lubeck	✗	✓

Table 2: MedVidQA: Participating teams and their task participation at TREC 2024.

Team	RunID	MAP	R@5	R@10	P@5	P@10	nDCG
TJUMI	run-2	0.2164	0.1743	0.2219	0.3731	0.2923	0.3033
	run-1	0.2161	0.1767	0.2233	0.3808	0.2962	0.3017
NCstate	Seahawk_run-1	0.4119	0.2652	0.429	0.4808	0.4654	0.5738
PolySmart	3	0.1276	0.105	0.1668	0.2846	0.2481	0.239
	5	0.1008	0.063	0.1105	0.1692	0.1712	0.2237
	2	0.1105	0.115	0.186	0.2962	0.2673	0.1994
	4	0.0828	0.0824	0.1466	0.2231	0.2192	0.1618
	1	0.0884	0.1067	0.1603	0.2846	0.2231	0.1694
UNCC	mainrun1	0.0242	0.0242	0.0242	0.05	0.025	0.0453
NCSU	run1	0.0204	0.0204	0.0204	0.0577	0.0288	0.0447
UNCW	run1	0.0007	0.0007	0.0007	0.0038	0.0019	0.0013
Baseline	BM25	0.1743	0.1703	0.2588	0.3346	0.3	0.2812

Table 3: Performance of the participating teams on video retrieval subtask of the VCVAL task.

7 Results and Discussion

7.1 VCVAL

VCVAL Task: The VCVAL task consists of video retrieval and visual answer localization subtasks. We presented the results of the video retrieval subtask in Table 3. We reported the results in terms of MAP, R@5, R@10, P@5, P@10, and nDCG. Since the relevancy of the videos is judged in terms of multi-level judgment, we consider nDCG as the primary metric for video retrieval subtask. Team NCstate achieved the best nDCG for the video retrieval subtask with a score of 0.5738. For the video retrieval subtask, we also developed a baseline using the BM25 lexical search approach where we used the subtitles of videos and indexed them using pyserini[3] with default hyperparameters. We retrieved the top 10 relevant videos for each question of the test set with the built index and reported the performance in Table 3. The baseline achieves a strong performance compared to many of the participant's approaches for video retrieval tasks.

The participating teams' visual answer localization subtask results are reported in Table 4. The table exhibits the detailed results with varying numbers of n and multiple evaluation metrics. We consider IoU=0.7 the primary metric for this subtask as it is the most strict metric, which signifies >= 70% overlap between the predicted and ground-truth visual answer segments. Team TJUMI achieved the best IoU=0.7 for the visual answer localization subtask with a score of 29.2 ($n = 1$).

7.2 QFISC

The results of the query-focused instructional step captioning by the participating teams are presented in Table 5, 6, 7. The table provides a detailed breakdown of the performance using various evaluation metrics including human evaluations on the appropriate metrics. We demonstrated the performance of the submitted runs on aligning ground truth and predicted steps in terms of precision, recall, and F-score as discussed in Section 5.2. We also show the effect of steps similarity threshold θ on the evaluation metrics as shown in Table 5. The lenient value of θ yields better results. Table 5 also provides further insight into the accuracy of the models predicted timestamps, showing how closely the predicted timestamps align with the actual timestamps of each generated

[3] https://github.com/castorini/pyserini

n	Team	RunID	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
1	TJUMI	run-2	34.62	21.15	15.38	26.01
		run-1	40.38	23.08	17.31	29.2
	NCstate	Seahawk_run-1	30.77	17.31	7.69	20.13
	PolySmart	3	21.15	13.46	5.77	14.13
		5	7.69	1.92	1.92	6.63
		2	17.31	11.54	1.92	11.52
		4	19.23	9.62	3.85	12.68
		1	13.46	7.69	1.92	9.17
	UNCC	mainrun1	0	0	0	0.04
	NCSU	run1	13.46	7.69	1.92	10.01
	UNCW	run1	0	0	0	0
3	TJUMI	run-2	55.77	38.46	30.77	43.15
		run-1	55.77	38.46	30.77	43.6
	NCstate	Seahawk_run-1	63.46	42.31	28.85	43.17
	PolySmart	3	32.69	23.08	11.54	24.15
		5	25	17.31	9.62	18.46
		2	32.69	23.08	11.54	23.73
		4	28.85	19.23	11.54	21.61
		1	36.54	26.92	9.62	25.93
	UNCC	mainrun1	0	0	0	0.04
	NCSU	run1	13.46	7.69	1.92	10.01
	UNCW	run1	0	0	0	0
5	TJUMI	run-2	59.62	44.23	36.54	47.32
		run-1	61.54	46.15	36.54	47.87
	NCstate	Seahawk_run-1	73.08	46.15	32.69	47.86
	PolySmart	3	36.54	26.92	17.31	29.19
		5	26.92	21.15	13.46	20.72
		2	36.54	26.92	15.38	29.63
		4	36.54	25	17.31	28.29
		1	44.23	32.69	15.38	31.22
	UNCC	mainrun1	0	0	0	0.04
	NCSU	run1	13.46	7.69	1.92	10.01
	UNCW	run1	0	0	0	0
10	TJUMI	run-2	63.46	48.08	42.31	50.46
		run-1	63.46	48.08	42.31	50.46
	NCstate	Seahawk_run-1	78.85	63.46	42.31	56.41
	PolySmart	3	51.92	38.46	23.08	38.57
		5	42.31	30.77	21.15	32.39
		2	48.08	42.31	23.08	39.77
		4	50	36.54	21.15	36.72
		1	51.92	38.46	25	38.54
	UNCC	mainrun1	0	0	0	0.04
	NCSU	run1	13.46	7.69	1.92	10.01
	UNCW	run1	0	0	0	0

Table 4: Performance of the participating teams on visual answer localization subtask of the VCVAL task.

θ	Team	RunID	Precision	Recall	F-score	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
0.4	PolySmart	LLaVA-NeXT-Video-32B-Qwen___GPT4o	31.54	35.06	32.17	34.87	30.9	17.06	24.13
	DoshishaUzlDfki	GPT_meta_prompt	35.64	50.36	39.08	46.22	38.85	20.13	32.04
		chatGPT_zeroshot_prompt	35.04	51.19	38.91	47.64	40.23	21.17	33.06
		mistral_fewshot_prompt	40.37	47.39	41.51	43.5	38.81	21.24	30.65
		mistral_meta_prompt	36.99	42.31	37.56	39.7	34.62	20.85	28.11
		CoSeg_meta_prompt	29.65	19.74	22.93	17.48	15.72	8.78	12.8
0.5	PolySmart	LLaVA-NeXT-Video-32B-Qwen___GPT4o	14.98	17.15	15.55	17.15	16.17	10.9	13.37
	DoshishaUzlDfki	GPT_meta_prompt	22.32	30.77	24.72	29.77	27.39	16.24	22.08
		chatGPT_zeroshot_prompt	23.49	32.59	26.1	31.59	29.96	18.24	23.69
		mistral_fewshot_prompt	25.56	31.25	27.08	30.36	27.88	18.37	22.15
		mistral_meta_prompt	23.54	28.73	24.88	28.51	25.76	18.27	21.05
		CoSeg_meta_prompt	16.8	11.28	13.16	10.57	10.35	7.71	8.73

Table 5: Performance of the participating teams on QFISC task focusing on closeness and alignment of the predicted steps with the ground-truth steps. The IoU is computed by considering the overlap between ground truth and predicted captions and the extension parameter $\lambda = 3$.

Team	RunID	BLEU-2	BLEU-3	METEOR	ROUGE-L	SPICE	BERTScore
PolySmart	LLaVA-NeXT-Video-32B-Qwen___GPT4o	17.04	11.29	25.42	29.54	23.66	86.29
DoshishaUzlDfki	GPT_meta_prompt	19.65	11.72	22.12	34.49	23.84	86.75
	chatGPT_zeroshot_prompt	19.88	11.47	22.18	34.51	23.83	86.69
	mistral_fewshot_prompt	17.42	10.07	19.49	33.69	23.87	86.74
	mistral_meta_prompt	15.52	9.07	18.01	32.17	22.62	86.56
	CoSeg_meta_prompt	4.87	2.05	11.38	23.15	17.19	85.55

Table 6: Performance of the participating teams on QFISC task focusing on closeness in terms of n-gram matching and semantic similarity. The results are shown considering the threshold $\theta = 0.4$.

step. This illustrates the model's effectiveness in estimating precise timing across the predicted steps.

We also computed the similarity between the predicted and ground-truth steps with multiple automatic metrics and showed the results in Table 6. Since QFISC is a generation task, it demands human evaluation due to the varying degree of model generation capacity, therefore we also computed the human scores as discussed in 4.2 and provided the detailed results in Table 7.

8 Conclusion

In the overview of the TREC 2024 MedVidQA track, we discussed the tasks, datasets, evaluation metrics, participating systems, and their performance. We evaluated the performance of the submitted runs using suitable automatic as well as manual evolution metrics. We hope that introducing the new tasks, developed topics/visual segments, and ground truth judgments along with the human evaluation of the generated system outputs, will be useful for fostering research toward designing video question-answering systems for healthcare needs.

Acknowledgments

This research was supported by the Division of Intramural Research (DIR) of the National Library of Medicine (NLM), National Institutes of Health. This work utilized the computational resources of the NIH HPC Biowulf cluster (https://hpc.nih.gov). We want to acknowledge George Awad Ian Soboroff for their feedback on planning the track, and the team at National Institute of Standards & Technology (NIST) for helping in the video assessments.

Team	RunID	Completeness	Accuracy	Coherenece
PolySmart	LLaVA-NeXT-Video-32B-Qwen___GPT4o	4.46	4.46	4.78
DoshishaUzlDfki	GPT_meta_prompt	3.8	3.76	4.48
	chatGPT_zeroshot_prompt	3.62	3.7	4.56
	mistral_fewshot_prompt	3.16	3.14	4.31
	mistral_meta_prompt	2.86	3.03	4.16
	CoSeg_meta_prompt	1.74	1.92	3.33
Ground Truth	NA	4.74	4.76	4.93

Table 7: Performance of the participating teams on QFISC task focusing on human evaluation of the predicted steps.

References

[1] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

[2] Shweta Yadav, Deepak Gupta, Asma Ben Abacha, and Dina Demner-Fushman. Reinforcement learning for abstractive question summarization with question-aware semantic rewards. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 249–255, 2021.

[3] Shweta Yadav, Deepak Gupta, Asma Ben Abacha, and Dina Demner-Fushman. Question-aware transformer models for consumer health question summarization. *Journal of Biomedical Informatics*, 128:104040, 2022.

[4] Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. Trec-covid: constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, volume 54, pages 1–12. ACM New York, NY, USA, 2021.

[5] Shweta Yadav and Cornelia Caragea. Towards summarizing healthcare questions in low-resource setting. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2892–2905, 2022.

[6] Shweta Yadav, tefan Cobeli, and Cornelia Caragea. Towards understanding consumer healthcare questions on the web with semantically enhanced contrastive learning. In *Proceedings of the ACM Web Conference 2023*, pages 1773–1783, 2023.

[7] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[8] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.

[9] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.

[10] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017.

[11] HHS. Artificial intelligence (ai) strategy. U.S. Department of Health and Human Services, 2021.

[12] Deepak Gupta and Dina Demner-Fushman. Overview of the medvidqa 2022 shared task on medical video question-answering. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 264–274, 2022.

[13] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019.

[14] Deepak Gupta, Kush Attal, and Dina Demner-Fushman. Towards answering health-related questions from medical videos: Datasets and approaches. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16399–16411, 2024.

[15] George Awad, Keith Curtis, Asad A Butt, Jonathan Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado, Eliot Godard, Lukas Diduch, Yvette Graham, et al. Trecvid 2023–a series of evaluation tracks in video understanding. In *Proceedings of TRECVID*, volume 2023, 2023.

[16] Deepak Gupta, Kush Attal, and Dina Demner-Fushman. A dataset for medical instructional video classification and question answering. *Scientific Data*, 10(1):158, 2023.

[17] Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oguz, Yashar Mehdad, and Mohit Bansal. Hierarchical video-moment retrieval and step-captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23056–23065, 2023.

[18] 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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[19] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[20] Satanjeev Banerjee and Alon Lavie. Meteor: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, volume 29, pages 65–72, 2005.

[21] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016.

[22] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2019.