Overview of the MedVidQA 2022 Shared Task on Medical Video Question-Answering

Deepak Gupta and Dina Demner-Fushman
Lister Hill National Center for Biomedical Communications
National Library of Medicine, National Institutes of Health
Bethesda, MD, USA
{firstname.lastname}@nih.gov

Abstract

In this paper, we present an overview of the MedVidQA 2022 shared task, collocated with the 21st BioNLP workshop at ACL 2022. The shared task addressed two of the challenges faced by medical video question answering: (i) a video classification task that explores new approaches to medical video understanding (labeling), and (ii) a visual answer localization task. Visual answer localization refers to identification of the relevant temporal segments (start and end timestamps) in the video where the answer to the medical question is being shown or illustrated. A total of thirteen teams participated in the shared task challenges, with eleven system descriptions submitted to the workshop. The descriptions present mono-modal and multi-modal approaches developed for medical video classification and visual answer localization. This paper describes the tasks, the datasets, evaluation metrics, and baseline systems for both tasks. Finally, the paper summarizes the techniques and results of the evaluation of the various approaches explored by the participating teams.

1 Introduction

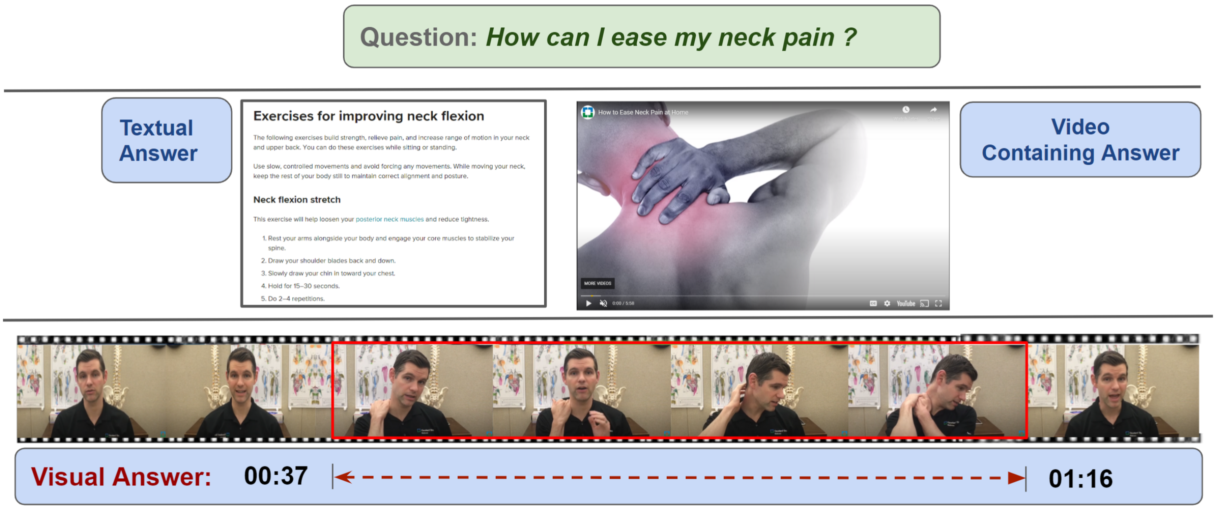
With the increasing interest in using artificial intelligence (AI) to support clinical decision-making, improving patient engagement, patient health and well-being (HHS, 2021), there is a need to explore the efficient algorithms for medical language-video understanding. Further, the recent surge in availability of online educational videos on diverse medical and health-related topics demands the development of effective systems that can understand medical videos to provide the best possible answers to consumers' first aid, medical emergency, and medical educational questions.

Video Question Answering (VQA) is an emerging and challenging task that requires the understanding of video, language, and their interaction to correctly provide the answer to the question. The majority of the existing studies (Lei et al., 2018; Xue et al., 2018; Li et al., 2020a; Chadha et al., 2020) on video question answering are focused on open-domain videos such as movies (Tapaswi et al., 2016), TV shows (Lei et al., 2018, 2020a), and games (Mun et al., 2017). Moreover, the primary objective of the existing VQA studies is to develop a system that can provide natural language answers to the users' questions about the video. Some works, such as Anne Hendricks et al. (2017); Lei et al. (2020b); Wang et al. (2020) focus on natural language frame/video localization, but most of them aim to find the video segment that has semantic understanding equivalent to the natural language query. The existing VQA approaches, however, do not take into account the real-world scenarios, where people interact through natural language questions and expect relevant and concise temporal segments from the videos as answers to their questions. Consider a health-related question "*How can I ease my neck pain?*". The textual answer (*cf.* Fig. 1) to the given health-related question will be hard to understand and act upon without visual assistance. In order to provide a visual answer to the question, the first step is to identify the most relevant medical video that has a series of steps describing the detailed visual answer to the question. The second and most important step is to locate the relevant temporal segment in the video that is suitable to be a visual answer (*cf.* Fig. 1) to the question.

Towards solving these challenges, we introduced the MedVidQA 2022 shared task[1], which aims to explore and develop efficient algorithms for video question answering that remain understudied in the medical domain. In the first task (medical video classification) of the MedVidQA 2022 shared task, participants are asked to develop a system that can categorize the video into medical instructional, medical non-instructional, and non-medical. The

[1] https://medvidqa.github.io/

Proceedings of the BioNLP 2022 workshop, Dublin, Ireland, pages 264–274
May 26, 2022. ©2022 Association for Computational Linguistics



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.

Figure 1: An example of a health-related question, textual answer, video containing the answer, and visual answer (temporal segment) from the video. The textual answer (**center-left**) is retrieved from the web. It contains a series of steps to relieve neck pain by improving neck flexion. The suggested steps in textual answer might be difficult to follow for a consumer who has little or no medical knowledge. The top video (**center-right**) retrieved from the YouTube search contains the answer; however, one has to watch the entire video to find the appropriate temporal segment from the video, which could be served as a visual answer to the question. Unlike the textual and video containing the answer, locating the appropriate temporal segment (**bottom**) which has the visual answer is easy to follow and also eliminates the need to watch the entire video to find the answer.

second task (medical visual answer localization) aims to effectively localize the visual answer to the given medical or health-related question in a given video.

2 MedVidQA 2022 Task Descriptions

Following creation of the dataset for video question answering (Gupta et al., 2022), we consider the following tasks:

2.1 Task 1: Medical Video Classification (MVC)

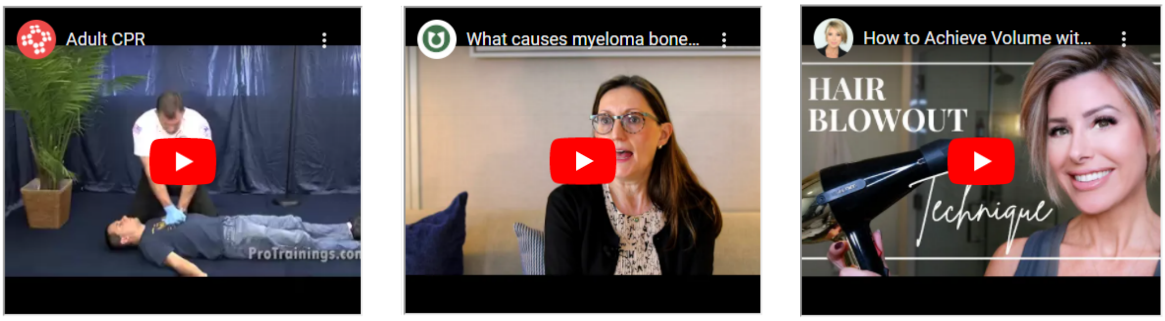
Given an input video, the task is to categorize the video into one of the following classes:

- **Medical Instructional**: A medical instructional video for non-professionals should clearly demonstrate a medical procedure, providing enough details to reproduce the procedure and achieve the desired results without prior training. The accompanying narrative should be to the point, and should clearly describe the steps in the visual content. A video is medical instructional if a valid medical or health-related question is aligned with it, and it explains/answers the medical question with a demonstration. The demonstration should be a tutorial/educational video where someone (e.g., a doctor or a medical professional) demonstrates a procedure related to the medical question or a how-to video about the medical or health-related question.
- **Medical Non-instructional**: A medical video can be categorized into a medical non-instructional if it discusses medical-related topics without any visual demonstration.
- **Non-medical**: A video can be categorized as non-medical if the video is neither medical instructional nor medical non-instructional.

We have provided the link to the sample videos for each class in Fig. 2.

2.2 Task 2: Medical Visual Answer Localization (MVAL)

Given a medical or health-related question and a video, the task aims to locate the temporal segments (start and end timestamps) in the video where the answer to the medical question is being shown, or the explanation is illustrated in the video. A similar task in the literature is established as natural language frame localization (Anne Hendricks et al., 2017; Miech et al., 2019), where the task is to find the video segment that has equivalent semantics as to the natural language. In contrast, the introduced task seeks to find a video segment with a visual answer to the natural language query. The MVAL task can be considered as finding a series of



(a) **Medical Instructional** (b) **Medical Non-Instructional** (c) **Non-medical**

Figure 2: Sample videos from each category in Medical Video Classification Task

Video Category	Train	Validation	Test
Medical Instructional	789	100	400
Medical Non-instructional	2,394	100	426
Non-medical	1,034	100	382
Total	4,217	300	1,208

Table 1: The dataset statistics for the MVC task. Training and validation datasets statistics are borrowed from MedVidCL corpus (Gupta et al., 2022).

Dataset Detail	Train	Validation	Test
Medical instructional videos	800	49	50
Video duration (hours)	86.37	4.54	5.13
Mean video duration (seconds)	388.68	333.89	369.62
Questions and visual answers	2,710	145	153
Minimum question length	5	6	6
Maximum question length	25	21	18
Mean question length	11.67	11.76	11.20
Minimum visual answer length (seconds)	3	10	4
Maximum visual answer length (seconds)	298	267	257
Mean visual answer length (seconds)	62.29	66.81	60.45

Table 2: The dataset statistics for the MVAL task. Training and validation datasets statistics are borrowed from MedVidQA corpus (Gupta et al., 2022).

"*medical instructional activity-based frame localization*" where a potential solution first searches for all medical instructional activity for a given medical question and then localizes a particular activity that is aligned to medical or health-related question in an untrimmed medical-instructional video. The sample health-related question and the visual answer are shown in Fig. 1.

3 Data Description

3.1 MVC Dataset

The MedVidCL[2] (Gupta et al., 2022) training and validation datasets are provided to train and validate the system for MVC task. A human-assisted two-stage approach was used to construct the MedVidCL dataset. In the first stage, human-annotated videos were used to train a machine learning model that predicts the appropriate category for the input video. In the second stage, only high-confidence (classifier probability ≥ 0.8) videos from HowTo100M (Miech et al., 2019) and YouTube8M (Abu-El-Haija et al., 2016) dataset are selected and manually validated. The automatically predicted video category is then updated, if needed. This strategy was used to construct the MedVidCL dataset. The videos in the training dataset are taken from YouTube[3]; however, the validation and test dataset contain the videos from HowTo100M and YouTube8M datasets. We have provided the detailed statistics of the datasets used for the MVC task in Table 1.

3.2 MVAL Dataset

The MedVidQA datasets are created from the top-4 videos returned by the YouTube search in response to the WikiHow[4] health-related query. The dataset contains 800 medical instructional videos in the training and 50 medical instructional videos in the validation set. MedVidQA contains medical-informatics expert-curated instructional questions and timestamps in the video, which serve as the visual answer to the questions. For the test dataset, we followed the dataset creation strategy similar to MedVidQA creation. We selected 50 YouTube videos from the search results in response to the diverse set of WikiHow queries. The instructional questions and visual answer timestamps were manually created by watching these 50 videos. We have

[2] https://osf.io/pc594/

[3] https://www.youtube.com/

[4] https://www.wikihow.com/Main-Page

provided the detailed statistics of the dataset used for the MVAL task in Table 2.

4 Evaluation

4.1 Evaluation Metrics

4.1.1 MVC Evaluation

To evaluate the performance of the MVC task, we use the following evaluation metrics:

Medical-Inst Precision: It measures the proportion of Medical Instructional class predictions that are actually correct.

$$\text{Med-Inst Precision} = \frac{TP_{medinst}}{TP_{medinst} + FP_{medinst}} \tag{1}$$

where, $TP_{medinst}$ and $FP_{medinst}$ are the True positive and False positive corresponding to the Medical Instructional class.

Medical-Inst Recall: It measures the proportion of actual Medical Instructional class video that were predicted correctly.

$$\text{Med-Inst Recall} = \frac{TP_{medinst}}{TP_{medinst} + FN_{medinst}} \tag{2}$$

where, $TP_{medinst}$ and $FN_{medinst}$ are the True positive and False negative corresponding to the Medical Instructional class.

Medical-Inst F1-score: It is the harmonic mean between precision $P_{medinst}$ and recall $R_{medinst}$ for the Medical Instructional video category.

$$\text{Med-Inst F1-score} = \frac{2 \times P_{medinst} \times R_{medinst}}{P_{medinst} + R_{medinst}} \tag{3}$$

Macro-averaged F1-score: It is the average harmonic mean between precision and recall, where the precision and recall are calculated per video category.

$$\text{Macro-F1} = \sum_{l \in \mathcal{L}} \frac{2 \times P_l \times R_l}{P_l + R_l} \tag{4}$$

where, P_l and R_l are the precision and recall corresponding to the class $l \in \mathcal{L}$.

Since the goal of the MVC task is to effectively predict Medical Instructional video, we consider **Medical-Inst F1-score** as our primary metric to rank the submission. We used the Scikit-learn (Pedregosa et al., 2011) implementation[5] of the precision, recall and macro-averaged F1-score metrics.

4.1.2 MVAL Evaluation

Following Gupta et al. (2022), we evaluated the performance of the MVAL task using the following metrics:

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 exact match. The mIoU is defined as the average temporal IoUs for all questions (N) in the test set. Formally,

$$\text{mIoU} = \frac{1}{N} \sum_{i=1}^{i=N} IoU(q_i) \tag{5}$$

Rαn, IoU = μ is another metric used to evaluate the performance of the MVAL 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. We asked the participants to submit only the top-1 temporal segment as the visual answer to the question; therefore, we have $n = 1$. Formally,

$$< R\alpha 1, IoU = \mu > = \frac{1}{N} \sum_{i=1}^{i=N} s(q_i, \mu), \text{ and} \tag{6}$$

$$s(q_i, \mu) = \begin{cases} 1, & \text{if } IoU(q_i) \geq \mu \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

We evaluated the participants' submission by considering $\mu = \{0.3, 0.5, 0.7\}$ and for brevity, we denote the $< R\alpha 1, IoU = \mu >$ metric with IoU=μ. Since the IoU=0.7 is the most restrictive metric amongst all the MVAL metrics, we use IoU=0.7 as our primary metric to rank the participants' submissions. The implementation of the evaluation metric is released here[6].

[5] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html

[6] https://github.com/deepaknlp/MedVidQACL/blob/master/MedVidQA/util/runner_utils_t7.py

4.2 Baseline Systems

4.2.1 MVC Baselines

Monomodal (Language) Baseline: In the first baseline, we consider extracting the English subtitles from the videos using the `pytube`[7]. The extracted subtitles are used to fine-tune the BERT-Base-Uncased (Devlin et al., 2019) pre-trained language model (PLM) to classify the video category.

Monomodal (Vision) Baseline: The monomodal vision-based baseline is built upon the video frames, which are extracted from each video at a uniform time interval. In order to extract the frame features, we considered the pre-trained ViT (Dosovitskiy et al., 2021) model as the feature extractor. The sequence of frame features is passed to the LSTM (Hochreiter and Schmidhuber, 1997) network for video category prediction.

Multimodal Baseline: For the multimodal baseline, we consider utilizing both video subtitles and video frames features (extracted from ViT) to predict the video category. The features are passed to the LSTM network to learn their sequence representation. We then concatenated the language and vision representation and passed the concatenated features to a feed-forward layer to predict the video category.

4.2.2 MVAL Baselines

VSL-Base: Following Gupta et al. (2022), we consider the VSL-Base as the first baseline for MVAL task, where the visual answer span is predicted using a multimodal fusion-based technique introduced by Zhang et al. (2020). In the VSL-Base a Transformer-based encoder is used to encode the question, and video frames features (obtained from I3D (Carreira and Zisserman, 2017)), and thereafter, both features are fused with the help of attention mechanism. The joint feature representation is used to predict the start and end timestamps of the visual answer.

VSL-QGH: This baseline is the extension of the VSL-Base introduced by Zhang et al. (2020), where the target temporal segment in the video is considered as the foreground and the rest of the video as the background. With the VSL-QGH technique, the network is trained by extending the span of the foreground to cover its preceding and following video frames. We follow the experimental details discussed in Gupta et al. (2022) to obtain the results on the test dataset.

Team Name	Team Affiliations	MVC	MVAL
ALIBABA_DAMO	Alibaba Damo Research	✓	✓
Baidu AI team	Baidu AI Team	✓	✓
SJTU_YITU	SJU/YITU	✓	✓
Tencent AI Research	Tencent AI Research	✓	✓
CMU_HKUST	CMU/HKUST	✓	✓
VPAI_Lab (Li et al., 2022a)	Hunan University/CAS	✓	✗
ChicHealth	Chic Health	✓	✓
PAHT	Pingan Health Tech	✓	✓
I am Bert	No Information Available	✓	✗
LingJing	Hunan University/CAS	✗	✓
UWashington	University of Washington	✓	✓
cs	No Information Available	✗	✓
DoSSIER (Kusa et al., 2022)	TU Wien	✗	✓

Table 3: Participating teams and their task participation at MedVidQA 2022 shared task

Team Name	MVC		MVAL	
	Language	Vision	Language	Vision
ALIBABA_DAMO	✓	✗	✓	✓
Baidu AI team	✓	✗	✓	✗
SJTU_YITU	✓	✗	✓	✗
Tencent AI Research	✓	✗	✓	✓
CMU_HKUST	✓	✗	✓	✗
VPAI_Lab	✓	✓	NA	NA
ChicHealth	✓	✗	✓	✗
PAHT	NA	NA	✓	✗
I am Bert	NA	NA	NA	NA
LingJing	NA	NA	✓	✓
UWashington	✓	✗	✓	✗
cs	NA	NA	NA	NA
DoSSIER	NA	NA	✓	✓

Table 4: Participating teams and their submissions considering the language (video subtitles) and vision (video frames) to build their approaches for MedVidQA 2022 shared task

5 Participating Teams and Methods

5.1 Participating Teams

We use the CodaLab platform to release the datasets, registration, and submissions of the participating teams. In total, 13 teams from Asia (China), Europe (Germany), and North America (USA) continents participated in the MedVidQA 2022 shared task and submitted 30 and 43 individual runs for the MVC and MVAL task, respectively. We have provided (*cf.* Table 3) the team name, affiliations and their participation in MVC and MVAL tasks. We also summarize (*cf.* Table 4) the participating teams and their submissions based on the considered modality to build their approaches for MVC and MVAL tasks. The results of all the participating teams for MVC[8] and MVAL[9] tasks are avail-

[7] `https://pypi.org/project/pytube/`

[8] `https://codalab.lisn.upsaclay.fr/competitions/1058`

[9] `https://codalab.lisn.upsaclay.fr/competitions/1078`

Rank	Team Name	Med-Inst Precision	Med-Inst Recall	Med-Inst F1-score	Macro F1-score
1	VPAI_Lab	**99.74**	97.75	**98.74**	**99.01**
2	ChicHealth	98.73	97.25	97.98	98.46
3	Baidu AI team	99.23	96.75	97.97	98.46
4	PAHT	97.76	**98.00**	97.88	98.46
5	Tencent AI Research	97.75	97.75	97.75	98.04
6	SJTU_YITU	98.47	96.50	97.47	98.04
7	CMU_HKUST	98.72	96.25	97.47	98.03
8	ALIBABA_DAMO	96.02	96.50	96.26	97.22
9	UWashington	97.65	93.50	95.53	96.86
10	I am Bert	92.21	91.75	91.98	94.01
–	*Monomodal (L) – Baseline*	94.67	88.75	91.61	94.37
–	*Monomodal (V) – Baseline*	90.97	68.00	77.83	82.24
–	*Multimodal (L+V) – Baseline*	84.97	69.25	76.31	81.06

Table 5: Official results of the MVC task. Here **L** and **V** denotes the Language and Vision respectively.

able on CodaLab platform.

5.2 MVC Submissions

5.2.1 Methods

All participants utilized pre-trained language models to develop the video classification methods to categorize the videos into one of the pre-defined categories. The earlier studies by Gupta et al. (2022) show that information obtained from the video subtitles features is more useful for the MVC task compared to the video frame features; therefore, the video subtitles remained the primary information considered by all participants to develop their approaches for the MVC task. To build the MVC models ALIBABA_DAMO and SJTU_YITU fine-tuned the Clinical-Longformer (Li et al., 2022b) on video subtitles. SJTU_YITU also used the Longformer (Beltagy et al., 2020) to build another MVC model by utilizing the video subtitles from the videos. Baidu AI team utilizes video title and subtitles to form a concatenated sequence and fine-tuned the hierarchical-BERT (Zhang et al., 2019) to predict the video category.

Team Tencent AI Research build their MVC models by fine-tuning the Longformer, Performer (Choromanski et al., 2021) and Big-Bird (Zaheer et al., 2020) pre-trained language models. Team CMU_HKUST built an ensemble approach for the MVC task with the predictions from hierarchical-BERT and Transformer-XL (Dai et al., 2019) pre-trained language models. Team VPAI_Lab built an ensemble approach by considering the predictions from monomodal and multimodal approaches for MVC tasks. They used DeBERTa (He et al., 2020) and I3D (Carreira and Zisserman, 2017) to encode the video subtitles and frames respectively. Team ChicHealth also proposes the ensemble models with the pre-trained Big-Bird and Longformer language models. Instead of encoding the entire subtitles from a video, they split the subtitles into multiple pieces and used the max and average pooling layer to aggregate the representations into a fixed-size representation vector. Team UWashington adopt the Big-Bird as the backbone network. They used the contrastive learning loss and the cross-entropy loss to build their approach for the MVC task.

5.2.2 Results

We have provided the official results for the MVC task and baseline models in Table 5. We rank the submissions based on the Med-Inst F1-score. Team VPAI_Lab achieved the first rank with the 98.74 Med-Inst F1-score and also reported the highest Med-Inst Precision (99.74) and Macro F1-score (99.01). Team PAHT submission reported the highest Med-Inst Recall (98.00) value from the best-ranked participants' system. The best-submitted run of each team outperformed the baseline scores on the primary metric of Med-Inst F1-score. We observed that top-4 teams achieved near-perfect Macro F1-score within a difference of 0.47 points. In terms of the primary metric (Med-Inst F1-score), the team ChicHealth (rank #2), Baidu AI team (rank #3), PAHT (rank #4) and Tencent AI Research (rank #5) achieved near-same performance ranging between 97.75 to 97.98. BERT-based monomodal baseline achieved the highest Med-Inst F1-score amongst all the baseline approaches.

5.2.3 Findings

The video subtitles are dominant features to predict the category of the video. The pre-trained language models (Longformer, Hierarchical BERT, Big-Bird) having the capability of effectively process the longer sequences, outperformed the tra-

Rank	Team Name	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
1	PAHT	90.85	84.97	73.20	75.83
2	SJTU_YITU	88.89	83.01	71.24	74.06
3	UWASHINGTON	85.62	81.05	69.93	72.07
4	LINGJING	84.31	73.20	62.75	67.53
5	CMU_HKUST	75.82	72.55	62.09	63.86
6	BAIDU AI TEAM	75.16	71.90	61.44	63.21
7	CHICHEALTH	74.51	67.97	53.59	61.34
8	TENCENT AI RESEARCH	69.28	62.09	49.67	57.31
9	ALIBABA_DAMO	60.13	52.94	38.56	48.21
10	cs	30.07	14.38	5.88	19.97
–	VSL-QGH – *Baseline*	21.56	10.45	5.88	17.60
–	VSL-BASE – *Baseline*	20.91	9.15	5.22	19.44
11	DOSSIER	31.37	13.07	4.58	18.80

Table 6: Official results of the MVAL task

ditional BERT-based pre-trained language model baseline. We observed that the video features could play an essential role in further enhancing performance on the MVC task if the language and vision features are fused without losing information from each modality.

The MVC task greatly benefited from the large pre-trained language model. The pre-trained language model learns the inherent structure from video subtitles that have proven effective in categorizing a video into one of the pre-defined categories. In contrast to the classical video classification task, where the model has to detect and learn the specific action to classify the video into the fine-grained category, the MVC task focused on the coarse-grained category. Therefore, we observed the participants' system (*cf.* Table 4) achieving high performance by only utilizing the video subtitles in coordination with the large pre-trained language models.

We observed that only the winning team VPAI_LAB built their approach considering both the language and vision features. The rest of the teams focused on only language features and achieved promising results. Due to the coarse-grained nature of the MVC task, the vision features alone (monomodal baseline) seem to carry the least information compared to the counterpart language modality to predict the video category.

5.3 MVAL Submissions

5.3.1 Methods

We briefly describe the approaches used by each participating team for the MVAL task.

ALIBABA_DAMO The video subtitles and questions were encoded with BERT, and the vector representations were obtained. The video features from consecutive three-second interval video frames were pooled to form a vector representation. The subtitles, question, and video features were aligned and concatenated to form a multimodal representation. Thereafter, two two-layer feed feed-forward was used to predict whether the three-second multimodal representations are inside the answer boundary.

BAIDU AI TEAM The team adopted the negative sampling NER method from Li et al. (2020b) to train the answer localization system. The team formulated atomic unit spans in the subtitles, i.e., the tokens in subtitles that belong to the start and end timestamps of the visual answer. The hidden state representations for each token of the span and question were obtained using BERT. The span representation was obtained using the approach discussed in Chen et al. (2017). The question representation and span representation were fused together with the feed-forward network to get the question-span representation. The question-span representation was used to predict whether the given span is an answer to the question or not. Following Li et al. (2020b), the team randomly sampled a small subset

Team Name	**Pre-trained LM**	**Modality**	**Approach**
PAHT	BigBird	Language	Sequence labeling with PLM and CRFs (Lafferty et al., 2001) on video subtitles to detect the answer span.
SJTU_YITU	BERT	Language	The relevant subtitle sentences are classified with BERT. Thereafter, the semantic relatedness scores is computed between the question and the subtitles sentences.
UWASHINGTON	DeBERTa	Language	Utilized the PLM to score the question-sentence pair. After that, the high-scoring contiguous sequence are considered as the visual answer.
LINGJING	DeBERTa	Both	Utilized visual highlight features as the visual token, which concatenates with the question, and video subtitles. Sequence labeling framework is adopted with PLM on video subtitles to detect the answer span.
CMU_HKUST	Big-Bird	Language	Utilized machine reading framework to localize the span in the video that could serve as the visual answer to the health-related question.
BAIDU AI TEAM	BERT	Language	Negative sampling approach (Li et al., 2020b) is used to incorporates randomness into the training loss for span recognition.
CHICHEALTH	NA	Language	Sequence labeling with PLM on video subtitles to detect the answer span.
TENCENT AI RESEARCH	NA	Language	The Mutual Matching Network (MMN) (Wang et al., 2021) is trained with the auxiliary task of mutual matching to guide the network.
ALIBABA_DAMO	BERT	Both	Sequence labeling with PLM on video subtitles to detect the answer span.
DOSSIER	RoBERTa, MPNet	Both	First, the similarity scores between question and subtitle are computed. After that, similarity scores are used to detect the answer by utilizing a random forest regressor and unsupervised peak detection method.

Table 7: The summary of the participants approaches used for MVAL task.

of unlabeled spans as the negative instances to induce the training loss. A span-level cross-entropy loss was used for training.

SJTU_YITU Team SJTU_YITU used a two-step approach to localize the answer in the video. In the first step, they fine-tune the BERT model to tag whether a given sentence from subtitles will be part of the answer sentence or not. In the second step, they compute the semantic relatedness between the question and the answer sentences (predicted in the first step) to refine the predictions of the previous step further. Finally, they transform the selected sentences from subtitles into corresponding time intervals.

TENCENT AI RESEARCH Mutual Matching Network (MMN) (Wang et al., 2021) was used for visual answer localization. MMN is a metric-learning approach that is based on the auxiliary task of mutual matching, which guides the network to select the additional correct sentence in a constructed negative sentence set for video moments retrieval in addition to gold-standard supervision. Their approach uses subtitles and question as input to train the MMN by considering a binary cross-entropy loss for regressing the IoU and a pair discrimination loss for learning discriminative features.

CMU_HKUST The team adopted a machine reading framework (Cui et al., 2022) to localize the span in the video that serves as the visual answer to the question. They utilize the subtitles and their timestamps to transform them into a span in the subtitles text. To encode the subtitles and the question, they used the Big-Bird model (Zaheer et al., 2020).

CHICHEALTH The team formulated the task as a sequence tagging problem. The query and the subtitle of a given video were concatenated as "`[CLS] QUESTION [SEP] SUBTITLES [SEP]`". The concatenated sequence served as input to the Transformer network. A pointer network was used to find the text spans that correspond to the video spans that answers the query. During prediction, they select the spans that have the highest span probability. They used multiple transformer-based networks and ensemble the predictions to find the appropriate span that is considered as the visual answer to the question.

PAHT The team formulated the visual answer localization task as a sequence labeling problem. They concatenated the question and subtitles to form a sequence. They utilized the pre-trained BigBird with Conditional Random Fields (Lafferty et al., 2001) head to tag each subtitle timestamps either B-ANSWER, I-ANSWER, or Other.

LINGJING The team proposed the visual-prompt text span localizing (VPTSL) method for visual answer localization by utilizing the pre-trained language model and visual highlight features. They fuse the question and visual features using cross-modal attention. The highlight features are used to provide the visual prompt to textual span predictor.

UWASHINGTON The team formulated the visual answer localization problem as question-sentence pair scoring task. They split the subtitles into multiple sentences and computed the scores for each sentence using the pre-trained DeBERTa model. They considered the timestamps associated with the high-scoring contiguous sequence of sentences as the visual answer to the question.

DOSSIER Team DOSSIER (Kusa et al., 2022) utilized the textual information in the form of subtitles and optical character recognition from video frames. They computed the similarity scores (using BM25, RoBERTa (Liu et al., 2019), and MPNet (Song et al., 2020)) between each video subtitles and the question. With the similarity matrix, they utilized random forest regressor[10] and unsupervised peak detection model to detect the answer indices.

We have provided the summary of each participants' approach for the MVAL task in Table 7.

5.3.2 Results

The official results for the MVAL task, along with the baseline scores, are provided in Table 6. We rank the team submissions based on the primary metric (IoU=0.7). Team PAHT achieved the highest 73.20 IoU=0.7 score. Their best submission also achieved the maximum IoU=0.3, IoU=0.5, and mIoU, which are 90.85, 84.97, and 75.83, respectively. Most of the participants' runs outperformed the multimodal learning-based baseline scores obtained from VSL-BASE and VSL-QGH.

5.3.3 Findings

The majority of the participating teams only use the video subtitles to locate the visual answer in the video. The video subtitles and their appearance timestamps are aligned to locate the start and end indices of the visual answer. Unimodal semantic relatedness between the question and video subtitles was computed with the pre-trained language models and proved to be more effective than the multimodal semantic relatedness as in VSL-BASE and VSL-QGH baselines. The top-3 participating systems built their approaches, similar to the text-based machine reading comprehension, by only utilizing the video subtitles features to locate the visual answer. However, team LINGJING proposed the multimodal approach for the MVAL task and achieved 62.75 IoU=0.7 that placed them in the 4th rank in the leaderboard.

It is observed that video subtitle features have proven to be effective compared to video features. The video subtitles are derived from commentary in videos. When a speaker in the video starts discussing a specific topic, they introduce the topic at the start of their commentary and make concluding remarks at the end of the commentary on the particular topic. The health-related questions in the MVAL task are formulated by watching the videos and identifying the span in the video, which could serve as the visual answer to the health-related questions. The video subtitle feature-based approaches exploit this structure and consider training the model to localize the span in the video subtitle sequence, which is semantically associated with the given question. This act of localizing the span from video subtitles is closely related to the machine reading comprehension (MRC) task; therefore, the participants use video subtitle features and treat the MVAL task similar to the approaches which have been used in the literature for the MRC task.

6 Conclusion

This paper describes the overview of MedVidQA 2022 shared task organized as part of the BioNLP 2022 workshop. We discussed the tasks, datasets, evaluation metrics, and baseline systems. We also provided a summary of the participating systems for both tasks. For the MVC task, the approach utilizing the attention-based fusion of the pre-trained

[10] https://bit.ly/3tViF3S

language model features and video features outperformed all the competitive methods. Overall, the MVC task with the coarse-grained category was relatively easy compared to the classical video classification task, where the model has to detect and learn the specific action to classify the video into the fine-grained category. We observe that video subtitles are key information to localize the visual answer in the video for the medical instructional question. We are optimistic that introducing these tasks and datasets will foster research toward designing systems that can understand medical videos and effectively provide visual answers to natural language questions.

Acknowledgements

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health, and utilized the computational resources of the NIH HPC Biowulf cluster (`http://hpc.nih.gov`). We would like to thank Kush Attal (NLM/NIH) for his help with the MVC baseline evaluation and MVAL dataset creation.

References

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*.

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

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

João Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4733.

Aman Chadha, Gurneet Arora, and Navpreet Kaloty. 2020. iperceive: Applying common-sense reasoning to multi-modal dense video captioning and video question answering. *arXiv preprint arXiv:2011.07735*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Martin Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. 2021. Rethinking attention with performers. In *International Conference on Learning Representations*.

Yiming Cui, Ting Liu, Wanxiang Che, Zhigang Chen, and Shijin Wang. 2022. Teaching machines to read, answer and explain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Deepak Gupta, Kush Attal, and Dina Demner-Fushman. 2022. A Dataset for Medical Instructional Video Classification and Question Answering. *arXiv preprint arXiv:2201.12888*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

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

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Wojciech Kusa, Georgios Peikos, Oscar Espitia Mendoza, Allan Hanbury, and Gabriella Pasi. 2022. Dossier at medvidqa 2022: Text-based approaches to medical video answer localisation problem. In *Proceedings of the 21th Workshop on Biomedical Language Processing*.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium. Association for Computational Linguistics.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020a. TVQA+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8211–8225, Online. Association for Computational Linguistics.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020b. TVR: A Large-scale Dataset for Video-subtitle Moment Retrieval. In *European Conference on Computer Vision*, pages 447–463. Springer.

Bin Li, Yixuan Weng, Fei Xia, Bin Sun, and Shutao Li. 2022a. Vpai_lab at medvidqa 2022: A two-stage cross-modal fusion method for medical instructional video classification. In *Proceedings of the 21th Workshop on Biomedical Language Processing*.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020a. HERO: Hierarchical encoder for Video+Language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, Online. Association for Computational Linguistics.

Yangming Li, Lemao Liu, and Shuming Shi. 2020b. Empirical analysis of unlabeled entity problem in named entity recognition. *arXiv preprint arXiv:2012.05426*.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2022b. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. 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.

Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, and Bohyung Han. 2017. Marioqa: Answering questions by watching gameplay videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2867–2875.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640.

Jingwen Wang, Lin Ma, and Wenhao Jiang. 2020. Temporally grounding language queries in videos by contextual boundary-aware prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12168–12175.

Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. 2021. Negative sample matters: A renaissance of metric learning for temporal grounding. *arXiv preprint arXiv:2109.04872*.

Hongyang Xue, Wenqing Chu, Zhou Zhao, and Deng Cai. 2018. A better way to attend: Attention with trees for video question answering. *IEEE Transactions on Image Processing*, 27(11):5563–5574.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. In *Conference on Neural Information Processing Systems*.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2020. Span-based localizing network for natural language video localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6543–6554, Online. Association for Computational Linguistics.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5059–5069.