# Towards Answering Health-related Questions from Medical Videos: Datasets and Approaches

**Deepak Gupta**[†], **Kush Attal**[‡*], **Dina Demner-Fushman**[†]
[†]LHNCBC, National Library of Medicine, National Institutes of Health, MD, USA
[‡]NYU Grossman School of Medicine, New York University, NY, USA
[†]{firstname.lastname}@nih.gov
[‡]Kush.Attal@nyulangone.org

## Abstract

The increase in the availability of online videos has transformed the way we access information and knowledge. A growing number of individuals now prefer instructional videos as they offer a series of step-by-step procedures to accomplish particular tasks. Instructional videos from the medical domain may provide the best possible visual answers to first aid, medical emergency, and medical education questions. This paper focuses on answering health-related questions asked by health consumers by providing visual answers from medical videos. The scarcity of large-scale datasets in the medical domain is a key challenge that hinders the development of applications that can help the public with their health-related questions. To address this issue, we first proposed a pipelined approach to create two large-scale datasets: HealthVidQA-CRF and HealthVidQA-Prompt. Leveraging the datasets, we developed monomodal and multimodal approaches that can effectively provide visual answers from medical videos to natural language questions. We conducted a comprehensive analysis of the results and outlined the findings, focusing on the impact of the created datasets on model training and the significance of visual features in enhancing the performance of the monomodal and multi-modal approaches for medical visual answer localization task.

**Keywords:** Multimodal Learning, Video Localization, Medical Video Question Answering

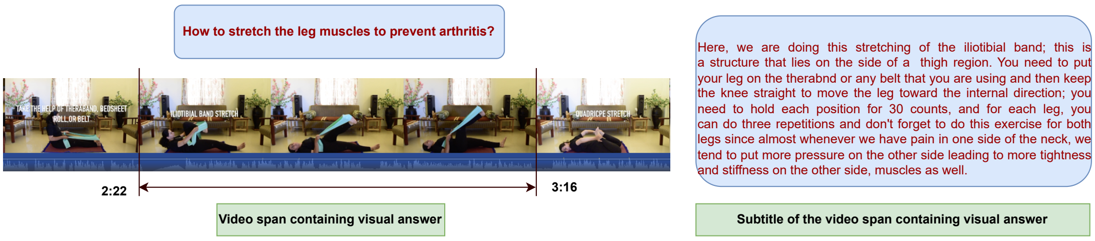
## 1. Introduction

An effective multimodal system that can enhance the ability to interact with the visual world, which encompasses images and videos, using a natural language query, has always been a coveted goal in artificial intelligence (AI) applications. These multimodal AI systems have the potential to revolutionize the fields of education, healthcare, and entertainment by enabling individuals to communicate with machines in a natural language that emulates human conversation. The emergence of large language-vision models and the availability of language-vision datasets has greatly improved the performance of many language-vision tasks, such as visual captioning (You et al., 2016; Pan et al., 2020; Anderson et al., 2018), visual question answering (Lei et al., 2018; Khan et al., 2021; Lei et al., 2020), and natural language video localization (Hendricks et al., 2017; Chen et al., 2019). Natural language video localization (NLVL) is one such language-vision understanding task whose goal is to semantically identify a temporal segment within an untrimmed video that is semantically aligned to a language query. Due to its applications in various downstream tasks, such as video retrieval (Francis et al., 2017), relation detection (Rodriguez-Opazo et al., 2021), and visual question answering (Lei et al., 2018), there has been a growing research interest in this direction. However, much of the advancement in NLVL is confined to open-domain, partially due to the availability of large-scale datasets. A specialized domain, such as the medical and healthcare domain, where there is a multitude of applications of NLVL task, remains unexplored.

Consider a health-related question, "*how to stretch the leg muscles to prevent arthritis?*" (*cf.* Fig. 1); the textual answer to this question may not be appropriate to act upon for a consumer with limited medical understanding. In this case, a short visual answer will be helpful for the consumer to follow as it offers visual assistance in the form of a step-by-step demonstration. In order to provide visual answers to the consumer's question, a multimodal system should be capable of identifying relevant videos and locating the appropriate segments from the videos, which can be considered as the audio-visual answer. Providing audio-visual answers from videos can cater to a wider audience, including those with reading difficulties or language barriers. Motivated by this, in this work, we focus on the task of medical visual answer localization, with the goal of locating visual answers to medical/healthcare-related questions. The small size of the existing dataset (Gupta et al., 2023) hinders the development of sophisticated neural-based approaches, which leads to sub-optimal performance on medical visual answer localization task. To address these issues, we first pro-

[*]Work done during the Postback at NLM, NIH.

*LREC-COLING 2024*, pages 16399–16411
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0


How to stretch the leg muscles to prevent arthritis?
TAKE THE HELP OF THERABAND, BEDSHEET ROLL,OR BELT
ILIOTIBIAL BAND STRETCH
QUADRICPE STRETCH
2:22
3:16
Video span containing visual answer
Subtitle of the video span containing visual answer

Here, we are doing this stretching of the iliotibial band; this is a structure that lies on the side of a thigh region. You need to put your leg on the therabnd or any belt that you are using and then keep the knee straight to move the leg toward the internal direction; you need to hold each position for 30 counts, and for each leg, you can do three repetitions and don't forget to do this exercise for both legs since almost whenever we have pain in one side of the neck, we tend to put more pressure on the other side leading to more tightness and stiffness on the other side, muscles as well.

Figure 1: An example of a health-related question and its video answer from the video.

pose a pipelined approach to automatically create large-scale datasets for the medical visual answer localization task. Second, we propose monomodal and multimodal approaches utilizing the created datasets. The proposed approaches achieve substantial performance improvement on multiple evaluation metrics for the medical visual answer localization task. We summarize the contributions of this work as follows:

1. We proposed a three-stage pipelined approach to automatically generate the datasets for the medical visual answer localization task. The human evaluation confirms that the approaches used in the pipeline are effective in generating high-quality datasets for the visual answer localization task.
2. We created two large-scale datasets[1], HealthVidQA-CRF and HealthVidQA-Prompt, for the task of medical visual answer localization. The former consists of $23,345$ question-answer-video triplets from $11,683$ medical videos, while the latter has $52,711$ triplets from $13,990$ medical videos.
3. We proposed an effective Cycle-Consistent Answer Localization (CCAL) approach, which outperforms the existing approaches on benchmark datasets. Later, we integrated the visual information from multiple visual encoders into the CCAL framework and performed the experiments that showed that the created HealthVidQA-CRF dataset can be used to achieve better performances with multimodal approaches.
4. We performed a detailed analysis of the results and highlighted the effects of the created datasets in model training, as well as the role of the visual features in improving the performance and benchmarking of the created datasets with monomodal and multimodal approaches.

## 2. Related Work

**Natural Language Video Localization:** requires modeling the cross-modality interactions between video and natural language to retrieve relevant segments from the video. Hendricks et al. (2017) proposed a Moment Context Network (MCN) that learns a shared embedding for video temporal context features and LSTM language features. The proposed video temporal context features integrate local and global video features and temporal endpoint features, which indicate when a moment occurs in a video. Later, they proposed the Moment Localization with Latent Context (MLLC) (Hendricks et al., 2018), which models video context as a latent variable. Liu et al. (2018b) develop a cross-modal retrieval technique to retrieve moments from the video responding to a given query. Other works also exploit the temporal relationship to tackle NLVL problems (Liu et al., 2018a; Zhang et al., 2019a; Liu et al., 2018c). Zhang et al. (2020a) propose span-based question answering to solve NLVL task. They propose the VSLNet approach based on a query-guided highlighting strategy to search for the target moment within a highlighted region. Later, they extend VSLNet to VSLNet-L (Zhang et al., 2021), which employs a multi-scale split-and-concatenation approach.

**Medical Visual Answer Localization:** Gupta et al. (2023), first introduced the task of medical visual answer localization (MVAL) and created the MedVidQA dataset having $3,010$ question-answer-video triplets. Later, Gupta and Demner-Fushman (2022) utilized this dataset and organized the shared task of retrieving the answer segments from the video against the health-related question. The majority of the participants utilized the pre-trained language models (Beltagy et al., 2020; Zaheer et al., 2020; Choromanski et al., 2020) to solve the MVAL task as a reading comprehension (Rajpurkar et al., 2016) problem. Li et al. (2022b) attempt to solve the MVAL problem by introducing the visual highlight prompts into the pre-trained language model (PLM) for enhancing the joint semantic representations of subtitles and video frames. Li et al. (2023) introduce the Cross-Modal Contrastive Global-Span (CCGS) method for the video corpus visual answer localization task.

[1] The created resources are publicly available on https://bionlp.nlm.nih.gov/

As an alternative, this study focuses on introducing new large-scale medical visual answer localization datasets and proposing approaches to effectively solve the MVAL problem. Specifically, we proposed an automatic approach to construct the video, question, and visual answer triplets by effectively handling the multiple subtasks (selecting instructional videos, detecting visual segments, and generating instructional questions) of the proposed approach. Additionally, we also introduce an intuitive mono-modal approach for localizing the answer from the videos by utilizing the cycle consistency loss, which is extended to a multimodal approach as well, where we explored a variety of visual features to represent the video frames.

# 3. Generating MVAL Dataset

This section describes our methodology for automatically creating a large medical instructional visual answer localization dataset.

## 3.1. Video-question-answer Triplets

This subsection deals with choosing the medical instructional videos and determining the visual segments in videos that could serve as the visual answer to the medical or health-related questions. Furthermore, we describe our methodology for generating medical instructional questions from the subtitles of the videos.

### 3.1.1. Selecting Medical Instructional Videos

In the first step of generating video-question-answer triplets, we aim to select the medical videos that can be used in subsequent steps of the dataset creation. We leverage the videos from '*Personal Care and Style*,' '*Health*,' and '*Sports and Fitness*' categories within the HowTo100M (Miech et al., 2019) dataset. To be included in a medical instructional visual answer localization dataset, **(1)** a video should describe a health-related topic, such as diseases, medical conditions, symptoms, drugs, treatments, medical exams, and procedures, etc., **(2)** a video should clearly demonstrate a step-by-step medical procedure providing enough details to reproduce the procedure and achieve the desirable results. We observed that the automatic video category labeled in the HowTo100M dataset was not always accurate. Moreover, some of the videos were also not instructional in nature, which hindered the development of the medical instructional visual answer localization dataset. To address this issue, we utilized the MedVidCL, a video classification dataset from Gupta et al. (2023), which has a total of $4,217$ training videos annotated for '*Medical Instructional*,' '*Medical Non-instructional*' and '*Non-Medical*.' We fine-tune the BigBird$_{\text{Base}}$ (Zaheer et al., 2020) model on the training set of the MedVidCL dataset by extracting the video subtitles, which yield the F1-score of $94.28$% on the test set of MedVidCL. The fine-tuned video classification model was used to label the subset ('*Personal Care and Style*,' '*Health*,' and '*Sports and Fitness*' categories videos) of the HowTo100M videos into medical instructional videos. This process yielded $15,664$ medical instructional videos that we used in the subsequent steps of the dataset creation.

### 3.1.2. Detecting Visual Answer Segments

Having the videos obtained in the previous stage, we consider a medical instructional video $V$ with raw subtitle/caption list $C_V = \{c_1, c_2, \ldots, c_m\}$ and corresponding time-stamps $T_V = \{t_1, t_2, \ldots, t_m\}$ of length $m$, where $c_i$ is the $i^{th}$ span of subtitle having the time stamps $t_i$. The issue with the raw subtitles is that they are not segmented and often overlap with the previous subtitles. To alleviate this issue, we concatenated the subtitle list $C_V$ and formed a sequence of words $W_V = \{w_1, w_2, \ldots, w_{|W_V|}\}$. In the next step, we hypothesize that the subtitle describing a visual answer in the video corresponds to a particular topic. Towards this, we aim to obtain the topic-aware segment from the $W_V$, and utilize the DeepSegment [2] model to segment the $W_V$ into $k$ topic-aware segments $S_V = \{s_1, s_2, \ldots, s_k\}$, next, we align the time-stamps $T_V$ to the topic-aware segments and obtain the aligned time-stamps $\hat{T}_V = \{t_1, t_2, \ldots, t_k\}$. With the topic-aware segments $S_V$ and corresponding aligned time-stamps $\hat{T}_V$ of the video $V$, Our goal is to identify topics that describe visual answers in the videos and subsequently divide the input sequence into contiguous segments representing distinct topics. We propose two approaches for detecting visual segments: **(1)** XLNet-CRF Model, and **(2)** XLNet-Prompt Model, which are described below:

**XLNet-CRF Model:** This approach utilizes the pre-trained `XLNet` (Yang et al., 2019) model to encode the segments and make the decision by using the conditional random field (Lafferty et al., 2001) based tagger to tag the boundaries of the visual segments.

**(1) Segment Encoding**: This module takes the segments $S_V = \{s_1, s_2, \ldots, s_k\}$ as input and processes them, and returns the encoded representation of the segment. Particularly, we obtain the hidden state representation $H_i$ of each segment $s_i$ of token length $|s_i|$ using the last layer of `XLNet`

[2] https://github.com/notAI-tech/deepsegment/tree/master

model. Thereafter, we choose (using operation Select (; last)) the hidden state representation of the last token as the segment representation. Formally:

$$\begin{aligned} H_1, \ldots, H_k &= \texttt{XLNet}(s_1, \ldots, s_k), \text{ where } H_i \in \mathcal{R}^{|s_i| \times d} \\ h_1, \ldots, h_k &= \texttt{Select}([H_1, \ldots, H_k], \text{last}) \end{aligned} \tag{1}$$

**(2) Segment Sequence Processing:** This step aims to process the encoded segments in the form of a sequence. However, the hidden state representations $\{h\}_{i=1}^{i=k}$ of the segments obtained in the previous step do not hold the inherent notion of the segment order. To tackle this, we first introduce the positional information of the segments in the form of positional embedding. Specifically, we augmented the positional embedding $p_i$ into the segment representation $h_i$ to obtain the position-aware segment representation $h_i^*$. Formally:

$$h_1^*, \ldots, h_k^* = (h_1 + p_1), \ldots, (h_k + p_k), \tag{2}$$

To process the segment representation $\{h^*\}_{i=1}^{i=k}$, we employed a Transformer-Encoder layer (Devlin et al., 2019), which utilizes the attention mechanism (Vaswani et al., 2017) to transform segment hidden states into rich and context-aware segment representations.

$$u_1, \ldots, u_k = \texttt{Transformer-Encoder}(h_1^*, \ldots, h_k^*) \tag{3}$$

Given the context-aware segment representations $U = \{u\}_{i=1}^{i=k} \in \mathcal{R}^{k \times d}$, we use a feed-forward network to project each segment representation $u_i$ into $c$-dimensional (exhibits the B-Seg, I-Seg and O tags) score $l$ as follows:

$$l = \mathbf{W}U + \mathbf{b}, \text{ where } \mathbf{W} \in \mathcal{R}^{d \times c}, \mathbf{b} \in \mathcal{R}^{c} \tag{4}$$

**(3) CRF-based Segment Tagging:** The output score $l$ obtained in the sequence processing step does not account for the dependencies across output labels. Segment labeling is one such task in which the label assigned to the preceding segment plays a crucial role in guiding the current segment to make accurate predictions. To achieve this, we utilized the CRF, which models the tagging decisions jointly. More formally, given the segments $S_V$ and prediction $y = \{y_1, y_2, \ldots, y_k\}$, the score $\mathcal{S}$ is computed as follows:

$$\mathcal{S}(S_V, y) = \sum_{i=2}^{k} M[y_{i-1}][y_i] + \sum_{i=1}^{k} l_i[y_i] \tag{5}$$

where $M$ is the matrix that contains the transition score between two subsequent labels. To train the network, the model maximizes the log probability of the correct segment sequence. In the testing phase, a sequence of predicted labels $y^*$ that maximize the score $\mathcal{S}$ is chosen as the final segment label sequence.

**XLNet-Prompt Model:** Inspired by the success of prompting (Liu et al., 2023) that aims to bridge the gap between pre-training and fine-tuning of the language model, we also explore prompt-based fine-tuning to tag the boundaries of the visual segments.

**(1) Prompt Tuning:** For the task of detecting visual segments, we develop prompts, which are a set of template $T(;)$ and label words $\mathcal{V}$. For each segment $s_i \in S_V$, we apply a prompt template and convert $s_i$ into prompt input $s_i^p$ for XLNet model. The prompt template usually has a [MASK] token, which needs to be filled by a label word $v \in \mathcal{V}$. We fed the prompt input $s_i^p$ into the XLNet model and computed the hidden state representation of $h_i^{\texttt{[MASK]}}$. Thereafter, we compute the probability that label $v$ can fill the [MASK] token. Formally,

$$\begin{aligned} s_1^p, \ldots, s_k^p &= T(s_1, \ldots, s_k), \\ H_1, \ldots, H_k &= \texttt{XLNet}(s_1^p, \ldots, s_k^p) \\ h_1^{\texttt{[MASK]}}, \ldots, h_k^{\texttt{[MASK]}} &= \texttt{Select}([H_1, \ldots, H_k], \texttt{[MASK]}) \\ p(\texttt{[MASK]} = v | s_i^p) &= \frac{exp(h_i^v . h_i^{\texttt{[MASK]}})}{\sum_{\hat{v} \in \mathcal{V}} exp(h_i^{\hat{v}} . h_i^{\texttt{[MASK]}})}, \forall v \in \mathcal{V} \end{aligned} \tag{6}$$

Finally, we map the segment labels 'B-Seg,' 'I-Seg,' and 'O' to the label words $\mathcal{V}$ to obtain the segment label.

**(2) Template and Label Words:** We performed a series of experiments with multiple templates for the visual segment detection task. With the supervised data, the pre-trained language model can be fine-tuned to maximize the log-likelihood of the correct segment labels. The label words used for the label to token mapping are as follows: 'B-Seg': '*first*,' 'I-Seg': '*next*,' 'O': '*other*.'

**Training and Evaluation of Models:** To train the models, we utilize the MedVidQA dataset, where video, question, and visual answer segment's beginning and end time-stamps are provided. We followed the strategy discussed above to transform the video subtitle into segments and marked each segment as start ('B-Seg'), intermediate ('I-Seg'), or other ('O') segments. The MedVidQA dataset has a total of $2,710$, $1,450$, and $1,550$ visual segments in training, validation, and test sets, respectively. We train XLNet-CRF and XLNet-Prompt models on the training set of the MedVidQA dataset, tune the hyper-parameters on the validation set, and evaluate the performance on the test set of the MedVidQA and MVAL task (Gupta and Demner-Fushman, 2022) ($1,530$ visual segments) datasets. While collating the subtitles and segmenting them based on DeepSegment, we observed that some text from the segment may fall in the previous and next segments. Therefore, while computing the true positive for the segment label, we relax it via a window of $w \in \{1, 2, 3\}$. Given a window of size $w$, we consider the predicted segment a correct segment if it is off by $w$ segments to either left or right in the segment sequence. Following this, we evaluated the XLNet-CRF models with multiple competitive approaches and reported the performance in Table 2. For the XLNet-Prompt approach, we experimented with multiple approaches and obtained the best F1-score of $0.5212$ and $0.5905$ (w=3) with prompt "This is the [MASK] step <SEG> " on the MedVidQA and MVAL test datasets, respectively. The detailed performance comparison of the XLNet-Prompt model on multiple templates is reported in Table 1.

#### 3.1.3. Generating Instructional Questions

Given the visual answer segments $S_V^a = \{s_1, s_2, \ldots, s_r\}$ and aligned timestamps $\hat{T}_V^a = \{t_1, t_2, \ldots, t_r\}$ of length $r$ of video $V$ detected using the approaches discussed in Section 3.1.2, the goal of this component is to generate instructional questions focusing on medical or health-related topics. Toward this, we built parameterized question generation models and optimized the parameters using the segment-question pairs available in the MedVidQA dataset. We explored monomodal and multimodal approaches to generate instructional questions by utilizing the respective modality from the video. Considering an answer segment $s_i \in S_V^a$ and respective timestamp $t_i = (t_i^s, t_i^e)$, where $s$ and $e$ denote the start and end timestamp, in the vision-based monomodal approach, we consider the frames $f_i = \{f_i^s, f_i^{s+1}, \ldots, f_i^e\}$ spanning between $t_i^s$ and $t_i^e$ in the video $V$ and train a Transformer-based encoder-decoder (Enc-Dec) model to generate the question. For the language-based monomodal approach, we collate all the sub-segments from $s_i$ and form the sequence $s_i = \{w_i^1, w_i^2, \ldots, w_i^{|s_i|}\}$ and fine-tune the pre-trained language models (PEGASUS, BART, and T5) on the MedVidQA dataset to generate the question. For the multimodal experiment, we fine-tune the UniVL (Luo et al., 2020) pre-trained language-vision model by using the frames $f_i$ and $s_i$ from the MedVidQA dataset. We have reported the performance from monomodal and multimodal approaches in terms of BLEU (Papineni et al., 2002), and BERTScore (Zhang et al., 2019b) in Table 3. It can be observed from Table 3 that the T5 model outperformed other approaches on the BLEU-4 metric, which has been a standard evaluation metric for question generation (Du et al., 2017; Gupta et al., 2020; Dong et al., 2019) task. Therefore, we consider the T5-based question generator to generate the instructional question.

### 3.2. HealthVidQA: large-scale Medical Instructional VideoQA dataset

We have utilized the procedure outlined earlier and created two large-scale **Health Vid**eo **Q**uestion **A**nswering (HealthVidQA) datasets: HealthVidQA-CRF and HealthVidQA-Prompt. In the HealthVidQA-CRF, we used the XLNet-CRF approach to identify the visual segments and used the T5-based question generation approach to generate instructional questions. The second dataset (HealthVidQA-Prompt) used the prompt-tuning approach to identify the visual segment. Similar to the former dataset, T5-based question generation was used to generate the instructional questions. We present the dataset analysis and human evaluation in the following subsections.

#### 3.2.1. Dataset Analysis

The HealthVidQA-CRF has $23,345$ video-question-answer triplets from $11,683$ medical videos. Each video has an average of $2$ visual answer segments of a duration of $73.39$ seconds. The average generated question length is $9.58$ words with a maximum length of $19$ words. To benchmark the dataset, we split the HealthVidQA-CRF dataset into the train $(18,754)$, validation $(2,355)$, and test $(2,236)$ sets. We have also performed the analysis of the HealthVidQA-Prompt dataset. The HealthVidQA-Prompt dataset has $52,771$ video-question-answer triplets from $13,990$ medical videos. We observed that each video has an average of $3.77$ visual answer segments of a duration of $33.88$ seconds. For the generated questions, the statistics match the HealthVidQA-CRF question. On average, the generated question length is $9.72$, with a minimum and maximum length of $5$ and $19$, respectively. We observe that the HealthVidQA-Prompt dataset has shorter segments, thus a shorter subtitle length, which leads to more visual segments in this dataset.

#### 3.2.2. Human Evaluation

We have followed an automated way to create the dataset, which could lead to noisy samples. In the dataset creation process, the possible causes of the errors could be in **(1)** selecting medical instructional videos, **(2)** detecting segment containing a visual answer, **(3)** generating valid instructional questions and **(4)** alignment of generated question and predicted segment. To assess the quality of the dataset, we devise a series of human evaluations to assess the aforementioned causes of the errors. In our human evaluation setups, we randomly choose $308$ samples from the generated datasets, and a total of three annotators judge the questions, segments, and videos to provide the as-

| Id | Template | MedVidQA | | | MVAL | | |
|---|---|---|---|---|---|---|---|
| | | F1-Score (w=1) | F1-Score (w=2) | F1-Score (w=3) | F1-Score (w=1) | F1-Score (w=2) | F1-Score (w=3) |
| 1 | [MASK] <SEG> | 0.4014 | 0.5016 | 0.5701 | 0.3747 | 0.4769 | 0.5380 |
| 2 | [MASK] [SEP] <SEG> | 0.3152 | 0.4535 | 0.5130 | 0.3463 | 0.4361 | 0.4918 |
| 3 | <SEG> [SEP] [MASK] | 0.3142 | 0.4226 | 0.5098 | 0.2985 | 0.4152 | 0.5401 |
| 4 | This is the [MASK] step where <SEG> | 0.2983 | 0.4275 | 0.4885 | 0.4225 | 0.4945 | 0.5583 |
| 5 | This is the [MASK] step where [SEP] <SEG> | 0.3283 | 0.4740 | 0.5279 | 0.3821 | 0.4864 | 0.5596 |
| 6 | This is the [MASK] step <SEG> | 0.3381 | 0.4642 | 0.5212 | 0.4064 | 0.5418 | 0.5905 |
| 7 | This is the [MASK] step [SEP] <SEG> | 0.3217 | 0.4896 | 0.5710 | 0.3595 | 0.5064 | 0.5807 |
| 8 | [MASK] I am going to <SEG> | 0.3042 | 0.4474 | 0.5124 | 0.3825 | 0.4854 | 0.5602 |
| 9 | [MASK] I am going to [SEP] <SEG> | 0.2861 | 0.4536 | 0.5343 | 0.2766 | 0.4188 | 0.5051 |

Table 1: Performance comparison of the prompt-based segment detection approach on the test set of the MedVidQA and MVAL datasets.

| Models | MedVidQA | | | MVAL | | |
|---|---|---|---|---|---|---|
| | w=1 | w=2 | w=3 | w=1 | w=2 | w=3 |
| BERT-CRF (Devlin et al., 2019) | 0.3671 | 0.5191 | 0.6184 | 0.2923 | 0.4725 | 0.5712 |
| ALBERT-CRF (Lan et al., 2020) | 0.2981 | 0.4417 | 0.5854 | 0.2475 | 0.3997 | 0.4951 |
| ELECTRA-CRF (Clark et al., 2020) | 0.4110 | **0.5671** | 0.6225 | 0.3028 | 0.4167 | 0.5667 |
| RoBERTa-CRF (Liu et al., 2019) | 0.3860 | 0.5643 | **0.6462** | 0.3192 | 0.3974 | 0.5097 |
| XLNet-CRF (Yang et al., 2019) | **0.4183** | 0.5256 | 0.6112 | **0.3216** | **0.4904** | **0.5901** |

Table 2: Performance comparison (window-based F1-score) of the CRF-based segment sequence labeling on the test set of MedVidQA and MVAL datasets.

| Models | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | BERTScore |
|---|---|---|---|---|---|
| Enc-Dec (Vaswani et al., 2017) | 32.48 | 17.81 | 9.980 | 6.670 | 64.12 |
| UniVL (Luo et al., 2020) | 40.39 | 24.98 | 16.18 | 11.40 | 68.90 |
| PEGASUS (Zhang et al., 2020b) | 42.46 | 28.87 | 20.06 | 14.70 | 71.49 |
| BART (Lewis et al., 2019) | 43.06 | 28.53 | 19.22 | 13.71 | 69.59 |
| T5 (Raffel et al., 2020) | 43.97 | 30.42 | 20.99 | 15.43 | 70.23 |

Table 3: Performance comparison of the multiple language and vision models on medical instructional question generation on the test set of MedVidQA.

| HealthVidQA-CRF | | | HealthVidQA-Prompt | | |
|---|---|---|---|---|---|
| Medical-Instructional Videos | | | | | |
| Yes | No | | Yes | No | |
| 81.61 | 18.39 | | 81.61 | 18.39 | |
| Segment Containing Visual Answer | | | | | |
| Yes | No | Partial | Yes | No | Partial |
| 82.04 | 6.59 | 11.38 | 75.45 | 5.99 | 18.56 |
| Question Generation Assessment | | | | | |
| Correct | Incorrect | Partial Correct | Correct | Incorrect | Partial Correct |
| 77.38 | 13.10 | 9.52 | 62.59 | 35.97 | 1.44 |
| Segment Question Alignment | | | | | |
| Yes | No | Partial | Yes | No | Partial |
| 75.45 | 5.99 | 18.56 | 46.67 | 27.41 | 25.93 |

Table 4: Comparison of the human evaluation scores (on multiple criteria) for the datasets created using XLNet-CRF and XLNet-Prompt approaches.

sessment. Our human scores are the agreement scores amongst the annotators.

**1. Medical Instructional Videos**: Getting the medical instructional videos is the first step of the dataset creation pipeline; therefore, we performed the human evaluation and asked the annotators to mark the videos whether they were medical instructional or not.

**2. Segment Containing Visual Answer**: This evaluation assesses whether the predicted segment contains a visual answer to any medical instructional questions. We asked the annotators to mark '**Yes**' if the segment contains a complete illustration of a particular procedure, '**No**' if the segment does not contain any illustration, and '**Partial**' if the segment contains a partial illustration of a particular procedure.

**3. Question Generation Assessment**: With this evaluation, we evaluated the quality of the generated question. Towards this, we asked annotators to mark the question as '**Correct**' if the question is a well-formed and valid instructional question,'**Partial Correct**' if the question belongs to either well-formed, has minor errors or valid instructional question, and '**Incorrect**' if the question is either grammatically, semantically or pragmatically incorrect.

**4. Segment Question Alignment**: This evaluation assesses whether the visual segment and corresponding generated question are aligned with each other. We asked the annotators to mark the segment-question pair as '**Yes**,' or '**Partial**' if the segment contains a complete or partial illustration as a visual answer to the generated instructional question, '**No**' otherwise.

We performed the human evaluation on both the created datasets HealthVidQA-CRF and HealthVidQA-Prompt. The detailed human evaluation is depicted in Table 4. From the human evaluation, we observe that the HealthVidQA-CRF dataset is more accurate compared to the HealthVidQA-Prompt dataset; therefore, we chose HeathVidQA-CRF for benchmarking the medical visual answer localization task. We believe the HealthVidQA-Prompt dataset has some noise, but it can be helpful for training in low-resource settings (Fang and Cohn, 2016), bootstrapping (Papadopoulou et al., 2022), and building a scalable model (Cavinato et al., 2023).

# 4. Approaches to Visual Answer Localization

## 4.1. Cycle-Consistent Answer Localization (CCAL)

We proposed an approach to effectively locate the visual segments that contain the answer to the given medical instructional question. With the success of the reading comprehension-based approaches (Gupta and Demner-Fushman, 2022) for the medical visual answer localization task, we followed the reading comprehension-based approach where the task aims to locate the 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. Inspired by cycle-consistent training (Shah et al., 2019), given a medical instructional question $Q$, video $V$ and corresponding subtitle $S$, and visual answer timestamp $T = (T^s, T^e)$, we first employed a text-based reading comprehension model $f$ with parameters $\theta$, which takes question $Q$ and subtitles $S$ and predict the start and end time stamps of the answer $\hat{T} \leftarrow f(Q, S; \theta)$, where $\hat{T} = (\hat{T}^s, \hat{T}^e)$. Using the predicted answer $\hat{T}$ and subtitle $S$ of video $V$, we utilize a question generation model $g$ with parameters $\phi$, and generate question $\hat{Q} \leftarrow g(S, \hat{T}; \phi)$. Our hypothesis is that if the reading comprehension model $f$ predicts the answer $\hat{T}$ correctly for the question $Q$, then the generated question using $\hat{T}$ and subtitle $S$ will be semantically and syntactically similar to $Q$. Now, we will describe the specifics of the reading comprehension model $f$ and question generation model $g$.

### 4.1.1. Reading Comprehension Model

Our reading comprehension model deals with the subtitle $S$ of the video $V$, question $Q = \{q_q, q_2, \ldots, q_{|Q|}\}$ to predict the answer span in the subtitle. To effectively encode the longer subtitle $S$, we use the pre-trained Longformer model (Beltagy et al., 2020) to encode the subtitle $S = \{s_1, s_2, \ldots, s_n\}$ having $n$ subtitles. Specifically, we concatenate all the words from the subtitles and formulate the word sequence $W = \{w_1^1, \ldots w_1^{|s_1|}, \ldots, w_n^1 \ldots w_n^{|s_n|}\}$. Following the fine-tuning of the question-answering model, we packed the question $Q$ and subtitle word sequence $W$ to form a single-word sequence $C$. We fed the word sequence $C$ to the Longformer model and obtained the hidden state $h_i$ for each token $t_i \in C$. Thereafter, a linear layer is employed on the top of the hidden state to compute the span start logit and span end logit. We predict the start and end positions of the tokens; thereafter, we map these tokens to their corresponding subtitle and extract their time stamp. We call the predicted start and end positions of the time stamp of the answer as $\hat{T} = (\hat{T}^s, \hat{T}^e)$. We train the network by maximizing the log-likelihoods of the correct start and end positions of the answer. We denote the loss function of the network as $\mathcal{L}_f(T, \hat{T})$[3].

### 4.1.2. Question Generation Model

With the predicted start and end positions of the time stamp of the answer as $\hat{T} = (\hat{T}^s, \hat{T}^e)$, we map them to their corresponding word sequence $\hat{W}$ in $W$ and pass the word sequence $\hat{W}$ to BART (Lewis et al., 2019) to generate the instructional question unlike for the question generation task (*cf.* 3.1.3), where T5 was best, here we found that BART performed better in reinforcing the CCAL approach while generating instructional questions. We call the generated question $\hat{Q}$. The question model is trained by maximizing the log-likelihood of the correct question $Q$. We denote the loss function of the network as $\mathcal{L}_g(Q, \hat{Q})$. We train our proposed CCAL approach to minimize the following objectives:

$$\mathcal{L} = \mathcal{L}_f(T, \hat{T}) + \mathcal{L}_g(Q, \hat{Q}) \tag{7}$$

## 4.2. Multimodal Late Fusion with CCAL

We also benchmark a multimodal late fusion-based technique combined with the Cycle-Consistent Answer Localization approach as discussed in Section 4.1. Our aim is to obtain visual and language modality at each word of the word sequence $W$. Towards this, in the late fusion multimodal approach, we extracted the frame (one frame per second) features corresponding to the video $V$ utilizing a 3D ConvNet (I3D) model, which was pre-trained on the Kinetics dataset (Carreira and Zisserman, 2017). Thereafter, for each word $w \in W = \{w_1^1, \ldots w_1^{|s_1|}, \ldots, w_n^1 \ldots w_n^{|s_n|}\}$ in the subtitle, we obtained frame representation $F \in \mathcal{R}^{|W| \times d}$ by choosing the frame that lies in the corresponding subtitle time stamp. We consider $F$ as the input image of dimension $|W| \times d$ and pass this to the vision encoder to encode the frame representation and obtain the vision representation as follows: $\mathcal{V} \leftarrow p(F; \psi)$, where $\mathcal{V} \in \mathcal{R}^{d_v}$ and $p$ is the vision encoder of parameter $\psi$. Similar to the reading comprehension model discussed in Section 4.1.1, we obtained the hidden state representation $h_i \in \mathcal{R}^{d_l}$ and concatenated it with the vision representation $\mathcal{V}$ and obtained the multimodal representation $x_i \in \mathcal{R}^{d_l + d_v}$. Finally, we apply a feed-forward

[3] While computing the loss for the reading comprehension model, we calculate the cross-entropy loss between expected and predicted positions of the answer's starting and ending words within the word sequence denoted as $W$.

network with $relu$ activation to project the $x_i$ into language encoder dimension $d_l$ to use the pre-trained language model further as used in the reading comprehension approaches discussed in Section 4.1.1. The multimodal late fusion with the CCAL model is trained by following the objectives listed in Eq. 7. In our experiments, as vision encoder $p$, we utilized VIT-Base (Dosovitskiy et al., 2020), VITMAE(He et al., 2022), VideoMAE-Base (Tong et al., 2022), VAN-Base (He et al., 2022), and ConvNext-Base (Liu et al., 2022) vision-based models.

## 5. Results and Analysis

**Metrics:** We evaluated the performance of the system on the MedVidQA (Gupta et al., 2023) dataset. Additionally, we evaluated the best-performing approaches on the test set of the HealthVidQA-CRF dataset having $2,236$ samples. Following the previous works (Gupta and Demner-Fushman, 2022; Awad et al., 2023), we use "R@1 IoU = $\mu$" and "mIoU" for the evaluation of visual answer localization. For each test question, we measure the Intersection over Union (IoU) between the predicted and ground truth timestamps. "R@1, IoU@ = $\mu$" means the percentage of text queries with an IoU larger than $\mu$. The "mIoU" refers to the average IoU for all test questions.

**Experimental Setups:** We set the maximum token length of each segment to $128$ for the visual answer segment detection task. The models are trained with a batch size of $4$ with one layer of transformer encoder. The model parameters are updated using the AdamW (Loshchilov and Hutter, 2019) optimization algorithm with the learning rate of $4e-5$ and weight decay of $1e-4$.

We use the large version of pre-trained T5, BART, and PEGASUS models for question generation. The transformer-based encoder-decode model was trained with one layer of encoder and decoder, each with a hidden state size of $128$. We use the official repository[4] of UniVL with the default hyper-parameters to fine-tune for the question generation task. The pre-trained language models are trained with a batch size of $2$ with a source sequence length of $256$ and question generation target sequence length of $20$. We use the beam search to generate the question with beam size $5$. The model parameters are updated using the AdamW optimization algorithm with the learning rate of $4e-5$ and weight decay of $1e-4$.

We utilized the base version, pre-trained language, and vision models (T5, BART, Longformer, ViT, VideoMAE, ViTMAE, VAN, and ConvNeXt) from HuggingFace (Wolf et al., 2019) to perform the experiments. For RC and CCAL approaches, we set the maximum source sequence length to 1024, except for the Longformer model. The Longformer model was set to 4096. For visual features, we select one frame from each second of the video uniformly and extract RGB visual features with the 3D ConvNet that was pre-trained on the Kinetics dataset (Carreira and Zisserman, 2017). Each pre-trained Transformer model was trained with the AdamW optimizer, with a learning rate of $5e-5$ for ten epochs, early stopping with the patience of three epochs and a batch size of two.

### 5.1. Effect of HealthVidQA-CRF Dataset

We assess the effect of the created HealthVidQA-CRF dataset on the models trained and evaluated (*cf.* Table 8) on the MedVidQA test dataset. We chose the RC and CCAL models, which are the best-performing models on the MedVidQA test dataset. We begin by adding $10\%$ of the created HealthVidQA-CRF dataset into the training set of MedVidQA and trained the RC models. With the $10\%$ addition of the HealthVidQA-CRF dataset, we observe the absolute improvements of the $4.57$, $5.23$, $9.15$, and $6.79$ in terms of IoU=0.3, IoU=0.5, IoU=0.7 and mIoU metrics, respectively. The significant improvements signify that the created dataset is capable of providing additional informative samples, which is required to train an efficient visual answer localization system.

### 5.2. Effect of Visual Features

In another analysis, we aim to assess the effect of the visual features while adding a portion of the HealthVidQA-CRF dataset to train the visual answer localization system. Table 5, shows that multimodal approaches could not outperform the best-performing monomodal CCAL approach. We observed from Table 5 that CCAL+VAN obtained the lowest scores in terms of multiple evaluation metrics. We wanted to analyze the effect of the HealthVidQA-CRF dataset on this model. Toward this, we train the CCAL+VAN model with the 10% of the HealthVidQA-CRF dataset. The trained model achieved an absolute improvement of $4.58$, $10.46$, and $4.08$ in terms of IoU=0.5, IoU=0.7, and mIoU evaluation metric, respectively. These significant improvements signify that multimodal approaches need additional datasets to perform better on the task of visual answer localization.

### 5.3. Performance on HealthVidQA-CRF

We extend our experiments by evaluating the best-performing monomodal and multimodal on the test set of the HealthVidQA-CRF dataset.

[4] https://github.com/microsoft/UniVL

| Models | IoU = 0.3 | IoU = 0.5 | IoU = 0.7 | mIoU |
|---|---|---|---|---|
| VSLBase (Gupta et al., 2023) | 25.16 | 8.38 | 4.51 | 19.3 |
| VSLQGH (Gupta et al., 2023) | 25.81 | 14.2 | 6.45 | 20.12 |
| CCGS (Li et al., 2022a) | 67.1 | 50.32 | 27.74 | 47.11 |
| RC (Beltagy et al., 2020) | 61.44 | 47.06 | 29.41 | 45.02 |
| CCAL (T5-QG) | 67.32 | 49.67 | 35.29 | 50.58 |
| CCAL | **71.90** | **54.9** | **35.29** | **52.92** |
| CCAL+ViT | 69.28 | 50.33 | 31.37 | 50.24 |
| CCAL+VideoMAE | 66.66 | 49.02 | 30.07 | 48 |
| CCAL+ViTMAE | 66.66 | 52.29 | 33.33 | 52.2 |
| CCAL+VAN | 67.97 | 47.71 | 28.1 | 48.17 |
| CCAL+ConvNeXt | 67.97 | 52.94 | 32.03 | 50.12 |

Table 5: Performance compression of the multiple monomodal and multimodal approaches on MedVidQA test dataset. RC referees to the reading comprehension model. CCAL (T5-QG) denotes the CCAL approach with T5 as the question generator.

|  | Models | IoU = 0.3 | IoU = 0.5 | IoU = 0.7 | mIoU |
|---|---|---|---|---|---|
| 1 | RC | 51.11 | 32.51 | 17.30 | 36.39 |
| 2 | + 10% data | 63.01 | 41.32 | 23.39 | 44.76 |
| 2 | + 20% data | 66.41 | 44.68 | 25.72 | 46.96 |
| 3 | + 50% data | 68.52 | 46.29 | 26.74 | 48.01 |
| 4 | + 100% data | **70.04** | **49.82** | **30.01** | **50.35** |
| 5 | CCAL | 53.67 | 32.82 | 17.35 | 37.13 |
| 6 | + 10% data | 62.03 | 39.53 | 22.41 | 43.48 |
| 7 | + 20% data | 66.73 | 45.84 | 26.61 | 47.22 |
| 8 | + 50% data | **68.16** | **46.33** | **27.78** | **48.12** |
| 9 | CCAL+ConvNeXt | 52.45 | 30.84 | 16.21 | 35.19 |
| 10 | + 10% data | 56.57 | 36.18 | 20.04 | 40.48 |
| 11 | + 50% data | 69.14 | 47.99 | 29.65 | 49.1 |
| 12 | + 100% data | **72.05** | **50.4** | **30.68** | **51.07** |

Table 6: Effect of the portion of the HealthVidQA-CRF data on the performance of the HealthVidQA-CRF test dataset.

|  | Models | IoU = 0.3 | IoU = 0.5 | IoU = 0.7 | mIoU |
|---|---|---|---|---|---|
| 1 | CCAL+VAN | **67.97** | 47.71 | 28.1 | 48.17 |
| 2 | + 10% HeathVidQA-CRF | 66.67 | **52.29** | **38.56** | **52.25** |
| 3 | CCAL+ConvNeXt | 67.97 | 52.94 | 32.03 | 50.12 |
| 4 | + 10% HeathVidQA-CRF | 62.09 | 47.06 | 30.72 | 47.54 |
| 5 | + 20% HeathVidQA-CRF | **69.28** | 54.25 | 30.72 | 50.60 |
| 6 | + 50% HeathVidQA-CRF | 66.01 | 49.67 | 29.41 | 48.56 |
| 7 | + 100% HeathVidQA-CRF | 67.32 | **54.90** | **37.91** | **51.77** |

Table 7: Effect of the portion of the HealthVidQA-CRF on the performance of the multimodal approaches on the MedVidQA test dataset.

|  | Models | IoU = 0.3 | IoU = 0.5 | IoU = 0.7 | mIoU |
|---|---|---|---|---|---|
| 1 | RC | 61.44 | 47.06 | 29.41 | 45.02 |
| 2 | + 10% data | 66.01 | 52.29 | 38.56 | 51.81 |
| 3 | + 20% data | 67.32 | **54.25** | **35.95** | **51.84** |
| 4 | + 50% data | **68.63** | 52.94 | 33.99 | 51.6 |
| 5 | CCAL | **71.90** | **54.9** | 35.29 | **52.92** |
| 6 | + 10% data | 64.05 | 47.71 | 34 | 48.03 |
| 7 | + 20% data | 69.28 | 52.94 | 35.95 | 51.72 |
| 8 | + 50% data | 66.01 | 52.94 | **36.60** | 51.1 |

Table 8: Effect of the portion of the HealthVidQA-CRF data on the performance of the MedVidQA test dataset.

We performed these experiments in data incremental setup, where we first utilized the MedVidQA training set to train the model and validated its performance on the HealthVidQA-CRF test dataset. Thereafter, we added the HealthVidQA dataset in an incremental manner and analyzed its impact on the model's performance towards the HealthVidQA-CRF test dataset. We evaluated the performance of the RC model on the HealthVidQA test dataset and reported the results in Table 6. Thereafter, we trained the model with an additional 10%, 20%, 50%, and 100% of the HealthVidQA dataset along with the MedVidQA training set and obtained the results. We also evaluated the performance of the CCAL+ConvNext model on the HealthVidQA test dataset. The experimental results show that with the additional HealthVidQA dataset, the CCAL+ConvNeXt model outperformed the RC and CCAL approaches.

### 5.4. Error Analysis

We analyzed the failed predictions where the overlap between the predicted and ground truth segment was $< 0.2$ using the CCAL approach. We categorize the major errors: **(1)** model falsely predicted the start segment where the video has multiple instructions, **(2)** model could not predict the specific segment correctly for the fine-grained questions ('*How to perform rescue breathing on an infant with a trach tube?*') compare to the coarse-grained counter-parts ('*How to perform rescue breathing on an infant?*'), and **(3)** CCAL model which focuses on subtitles of the video could not predict the segment where the visual information is required, though the multimodal approaches were able to reduce this type of error.

## 6. Conclusion

In this work, we presented a pipeline to automatically create medical visual question-answering datasets focusing on health-related questions and their visual answers in the videos. With the proposed pipeline, we build two large-scale medical visual question-answering datasets, HealthVidQA-CRF and HealthVidQA-Prompt. We performed in-depth human evaluations on the created datasets, and the evaluation shows the former dataset is better aligned with the human annotations. We also proposed a monomodal and multimodal CCAL approach for medical video question-answering task that achieved state-of-the-art performances and set competitive baselines for future research. The detailed experiments and analysis show that the created datasets help in improving the performance of the MedVidQA system. We believe that the created datasets can be used to provide the solution by pre-training/fine-tuning language-vision models for medical visual answer localization task.

## Ethics Statement

**Dataset generation and models**: The datasets presented in this work leverage the existing publicly available HowTo100M dataset. The subsets generated through leveraging the HowTo100M dataset may inherit gender, age, geographical, and cultural biases from the public instructional videos available online. General ethical considerations that apply to visual recognition models may apply to our models.
**Limitations**: The health topics covered in our datasets are limited to those that are available in the HowTo100M dataset videos, which may leave some health topics under-explored.

## Acknowledgments

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). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We would like to thank Debora Whitman, Denise Hunt, Melanie Huston, Dorothy Trinh, Preeti Kochar, Cathy Smith, Luke Zic, and Arun Crispino for their annotations and contributions to this work.

## Bibliographical References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

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

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

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Lara Cavinato, Noemi Gozzi, Martina Sollini, Margarita Kirienko, Carmelo Carlo-Stella, Chiara Rusconi, Arturo Chiti, and Francesca Ieva. 2023. Explainable domain transfer of distant supervised cancer subtyping model via imaging-based rules extraction. *Artificial intelligence in medicine*, 138:102522.

Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. 2019. Localizing natural language in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8175–8182.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. 2020. Rethinking attention with performers. In *International Conference on Learning Representations*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, 32.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Meng Fang and Trevor Cohn. 2016. Learning when to trust distant supervision: An application to low-resource pos tagging using cross-lingual projection. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 178–186.

Danny Francis, Paul Pidou, Bernard Merialdo, and Benoit Huet. 2017. Natural language access to video databases. In *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, pages 78–81. IEEE.

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

Deepak Gupta, Hardik Chauhan, Ravi Tej Akella, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Reinforced multi-task approach for multi-hop question generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2760–2775.

Deepak Gupta and Dina Demner-Fushman. 2022. 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, Dublin, Ireland. Association for Computational Linguistics.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2018. Localizing moments in video with temporal language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1380–1390. Association for Computational Linguistics.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. 2017. Localizing moments in video with natural language. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5804–5813.

Humair Raj Khan, Deepak Gupta, and Asif Ekbal. 2021. Towards developing a multilingual and code-mixed visual question answering system by knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1753–1767.

John D Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020. 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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Bin Li, Yixuan Weng, Bin Sun, and Shutao Li. 2022a. Learning to locate visual answer in video corpus using question. *arXiv preprint arXiv:2210.05423*.

Bin Li, Yixuan Weng, Bin Sun, and Shutao Li. 2022b. Towards visual-prompt temporal answering grounding in medical instructional video. *arXiv preprint arXiv:2203.06667*.

Bin Li, Yixuan Weng, Bin Sun, and Shutao Li. 2023. Learning to locate visual answer in video corpus using question. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Bingbin Liu, Serena Yeung, Edward Chou, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. 2018a. Temporal modular networks for retrieving complex compositional activities in videos. In *The European Conference on Computer Vision (ECCV)*.

Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018b. Attentive moment retrieval in videos. In *The 41st*

*International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 15–24.

Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018c. Cross-modal moment localization in videos. In *Proceedings of the 26th ACM International Conference on Multimedia*, pages 843–851. Association for Computing Machinery.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

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*.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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

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.

Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10971–10980.

Anthi Papadopoulou, Pierre Lison, Lilja Øvrelid, and Ildikó Pilán. 2022. Bootstrapping text anonymization models with distant supervision. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4477–4487.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.

Cristian Rodriguez-Opazo, Edison Marrese-Taylor, Basura Fernando, Hongdong Li, and Stephen Gould. 2021. Dori: Discovering object relationships for moment localization of a natural language query in a video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1079–1088.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6649–6658.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659.

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. *Advances in neural information processing systems*, 33:17283–17297.

Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. 2021. Natural language video localization: A revisit in span-based question answering framework. *IEEE transactions on pattern analysis and machine intelligence*, 44(8):4252–4266.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2020a. 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.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020b. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Songyang Zhang, Jinsong Su, and Jiebo Luo. 2019a. Exploiting temporal relationships in video moment localization with natural language. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 1230–1238, New York, NY, USA. Association for Computing Machinery.

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