Text-based Video Question Answering

Thesis submitted in partial fulfillment
of the requirements for the degree of

*Master of Science in **Computer Science and Engineering** by Research*

by

Soumya Shamarao Jahagirdar
2021701015
soumya.jahagirdar@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
March 2024

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled **"Text-based Video Question Answering"** by Soumya Shamarao Jahagirdar, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Adviser: Prof. C V Jawahar

To my beloved ajja, ajji, amma, pappa, moushi, and brother

Acknowledgments

To the most transforming years of my research life at IIIT-Hyderabad. I remember being very nervous about the whole thought of pursuing studying at this university. The journey that I have been through for the past couple of years has not only taught me how to carry out research but has also taught me how to think, articulate, manage people and time, code, learning how to learn and unlearn things and accept that making mistakes and learning from yours and others mistakes is part of life. I only wish to work harder henceforth, applying all the learnings I was privileged to learn.

I would like to thank Prof. Jawahar for all the invaluable inputs and guidance throughout my research. I appreciate the honest feedback on where I should improve myself as a student in order to be a better researcher. I have learned a lot from him and have understood that research is always meant for the greater good. I have also learned that building a strong foundation is where the key strength lies and is where one should work even harder. I would also like to thank my co-advisor Prof. Dimosthenis. Prof. Dimos has always been the one to conduct our brainstorming sessions. The amount of intricate details he knows and observes even within a meeting of a couple of minutes is commendable and definitely something I would like to learn and implement in my future career. Prof. Dimos has this very positive energy around him that makes everything so much more interesting and keeps me motivated to work even harder. I appreciate and admire the way he communicates his ideas and offers constructive criticism with enough reasons and explanations that would make me rethink and delve even deeper into the ideas. I would like to thank Minesh Mathew, my mentor for his input and efforts that I looked for while formulating the task, and dataset and also designing different experiments. I am thankful for his never-ending help and his experience in how he carried out research that in turn have helped me a lot. I have learned a lot of minor yet very important details from him that will definitely help me in the future. I can't thank all of them enough for their exceptional guidance and enriching my research abilities.

I would like to thank my undergraduate advisors Prof. Shankar Gangisetty, Prof. Anand Mishra, and Prof. Uma Mudengudi. I owe a big thanks to Prof. Shankar for educating me on how to maintain discipline and conduct research during the early stages of my research career. He has not only been a great advisor but also an amazing mentor whom I have consistently turned to for guidance. I would like to express my gratitude to Prof. Anand Mishra for introducing me to the world of text-based scene understanding which has undoubtedly had a significant impact on my research up to this point. I would like to thank Prof. Uma Mudengudi under whose guidance, I conducted my first-ever research. She

provided me with a platform, time, a supportive peer group, and not to mention her own time and guidance, all of which exposed me to the world of computer vision and research!

I would like to thank all the seniors – Rudrabha, Siddhanth, Seshadri, Jayashree, Aditya Arun, Avijit, and Ajoy Sir – for their assistance whenever I needed help in understanding concepts or learning how things worked in general. I would like to thank Shubham, Madhav, Zeeshan, Varun, Shivanshu, and Rupak for being the best sport in the lab!! I would thank my dear friends, Siddharth for always being there and helping me understand where I go wrong, Darshan for never-ending motivation when I am down, and for conversations about papers we read on various video understanding topics, Darshana for always being the one to remind me to manifest the right things, George for being the best travel (I will miss the time at CVC) and project buddy, Vivek for our morning gym conversations. I would like to thank and appreciate my batch friends Aditya, Dhruv, Amruth, Prateek, Bhoomendra, Nayan, Madan, Shishir, and others who have been the best sport for the entire time at IIIT. I would like to thank Joy, Alik, Aparna, Aditya, Bipasha, Shanthika, Ravi, Pranav, Kiran, Anagha, Vansh, Akshat, and Manu for making the lab a fun place. I would also like to thank my CVC labmates for the great time in Barcelona.

I would also like to Rohitha mam, Aradhana mam, Cecilia mam, Tessy mam, Mahathi, Ram Sir, Varun Sir, Krishna Sir, Sony and Ann, and all the other lab staff members without whom my stay in the lab would have been difficult.

This is a special mention to all my IIIT-H best canine friends – Ceaser, Waffle, Batman, Paper, Lily, Trunks, Greyhound, Tulip, Auto, Goodboi, Rosy, Tofu, Leo, Alex, Gopi, Chess, Sheru, Chintu, Max, Spotty, Jeff, Cousines, browny, shiny, kcis trio, and Muscles, you all have been the best part of IIIT. I am so thankful to you all. You all have been my strength and home. I would miss you all the most!

Lastly, I would like to thank my family and friends outside IIIT. To my ajja – Your excitement in providing answers to my questions was filled with excitement and thoughtfulness. I loved talking to you about my projects. I wish you were here to read my thesis. I hope you are at peace. My ajji, my pillar – I thank you for being my best friend!!! To amma and pappa who have always believed in me and have always taught me to take the truthful and correct path. To moushi, my inspiration. To tamma for being the most sensible kid. I would like to thank my friends Amrutha, Srushti, Kartik, Rutvik, Tushar, Akshata, and others for always being there.

Abstract

Think of a situation where you put yourself in the shoes of a visually impaired person who wants to buy an item from a store or a person who is sitting in their house and watching the news on the television and wants to know about the content of the news being broadcast. Motivated by many more such situations where creating systems capable of understanding and reasoning over textual content in the videos, in this thesis, we tackle the novel problem of text-based video question answering.

Vision and Language are broadly regarded as cornerstones of intelligence. Though each of these has different aims – language has the purpose of communication, and transmission of information, and vision has the purpose of constructing mental representations of the scene around us to navigate and interact with objects. When we study both of these fields jointly, it can result in applications, tasks, and methods that, when combined go beyond the scope compared to when they are used individually. This inter-dependency is being studied as a newly emerging area of a study named "multi-modal understanding". Many tasks such as image captioning, visual question answering, video question answering, text-video retrieval, and more fall under the category of multi-modal understanding and reasoning tasks. To have a system that can reason over both text-based information and temporal-based information, we propose a new task. The first portion of this thesis focuses on the formulation of the text-based VideoQA task, by first analyzing the current datasets and works and thereby arriving at the need for text-based VideoQA. To this end, we propose the NewsVideoQA dataset where the question-answer pairs are framed on the text present in the news videos. As this is a new task proposed, we experiment with existing methods such as text-only models, single-image scene text-based models, and video question-answering models. As these baseline methods were not originally designed for the task of video question-answering using text in the videos, the need for a video question-answering model that can take the text in the videos into account to obtain answers became the need. To this end, we repurpose the existing VideoQA model to incorporate OCR tokens namely – OCR-aware SINGULARITY, a video question-answering framework that learns joint representations of videos and OCR tokens at the pretraining stage and also uses the OCR tokens at the finetuning stage.

In this second portion of the thesis, we look into the M4-ViteVQA dataset which aims to solve the same task of text-based video question-answering but the videos belong to multiple categories such as shopping, traveling, vlogging, gaming, and so on. We perform a data exploratory analysis where we analyze both NewsVideoQA and M4-ViteVQA on several aspects that look for limitations in these datasets. Through the data exploratory experiment, we show that most of the questions in both datasets

have questions that can be answered just by reading the text present in the videos. We also observe that most of the questions can be answered using a single to few frames in the videos. We perform an exhaustive analysis on a text-only model: BERT-QA which obtains comparable results to the multi-modal methods. We also perform cross-domain experiments to check if training followed by finetuning on two different categories of videos helps the target dataset. In the end, we also provide some insights into creating a dataset and how certain types of annotations can help the community come up with better datasets in the future.

We hope this work motivates future research on text-based video question-answering in multiple video categories. Furthermore, the pretraining strategies and combined representation learning from these videos and the multiple modalities that videos provide us will help create scalable systems and drive future research towards better datasets and creative solutions.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Being fortunate by having vision and working on the problems related to computer vision and multi-modal reasoning has motivated us to assist people who are visually impaired. These applications may range from overcoming their daily visual challenges to performing everyday tasks to being able to enjoy most of the facilities, which are not just for non-physically challenged people. Researchers have come up with many tools that empower people to snap a picture of an object and recognize what it is as well as where it can be purchased. Videos are multi-modal by nature, i.e., they are made up of visual, audio, and textual elements that make them information-rich.

To enrich this multi-modal approach to solving problems, researchers have introduced many tasks such as Image Captioning – where the task is to transcribe an image into natural language by describing what is in the image and Visual Question Answering (VQA) – where the task is to obtain an answer given an image and a question about the image, furthermore extended to videos i.e. Video Question Answering (VideoQA) – where the question about the entire video is asked and the task is to obtain the answer by having reasoning over multiple frames. To extend the usage of multi-modal content in the scene, two groups, [46] and [4], pointed out that VQA models fail catastrophically on questions requiring reading. Specifically, the VizWiz study [13] found that up to 21% of these questions involve reading and reasoning about the text captured in the images of a user's surroundings such as - 'What temperature is my overset to?', 'What denomination is this bill?'. Considering these questions, the models need to learn aspects such as: detecting the text, jointly reasoning about the detected text and the visual content, and deciding what is the appropriate answer to the question. Not to forget to mention that these were the questions [13] that were asked by visually impaired users for their assistance.

The reasoning processes required to tackle these challenges are not trivial to incorporate into a model. Taking into account the temporal dimension of an unfolding event requires reasoning over the evolution of certain actions, retrieving information from a specific time in the sequence, or a combination of the two. At the same time, recognizing the fact that the world around us is littered with textual information that often carries important semantics necessary to interpret the scene has spawned a new direction in VQA. Introducing the scene text modality in the process requires incorporating error-prone reading systems and connecting scene text semantics and literal transcriptions with the answer space.

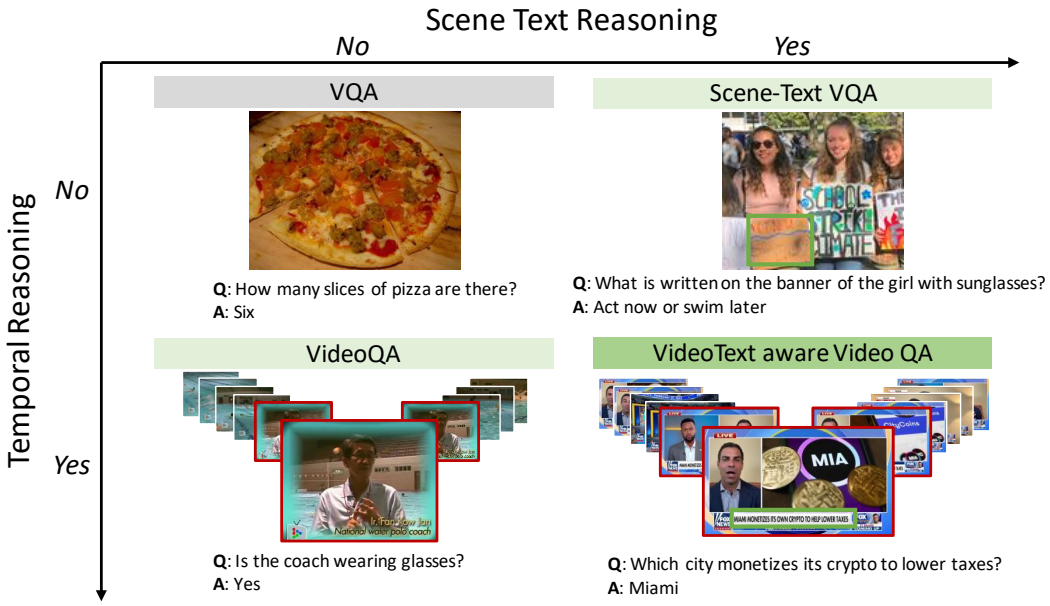


Figure 1.1 Comparison of tasks in VQA space. We address the task of text-based Video Question Answering, incorporating VideoText (VideoText is the textual content embedded in the videos) information (bottom right). We propose a new dataset of News Videos along with QA annotations grounded on video text and explore VQA models that jointly reason over temporal and text-based information.

In this work, we attempt for the first time to join these two lines of research and introduce the VideoText (VideoText is the textual content embedded in the videos) modality into Video Visual Question Answering. Various attempts to apply VQA to the video setting have been proposed [12, 26, 48, 58, 65]. Such VideoQA methods have put forward datasets and methods focusing on recognizing actions, emotions, activities, and reasoning over temporal, causal correspondences and knowledge graphs. However, they fall short in reasoning over the text appearing in the videos. Scene Text VQA [4, 46], on the other hand, focuses on methods that allow VQA systems to incorporate scene text in the reasoning process. On one hand, this entails extracting semantics from noisy textual input, and on the other hand, it requires dynamically expanding the answer space to incorporate new answer tokens afforded by the scene text [4,33,35,46,50]. Nevertheless, all scene text VQA methods are limited to processing a single image and cannot be readily extended to a multi-frame video input.

In the first part of this work, we attempt to combine multi-frame-based, VideoQA architectures with the scene text modality. To explore this novel research direction, we define a new task and associated dataset: NewsVideoQA. Motivated by the prominent function of scene text in news video snippets, and the complementary information it carries to the visual modality, we consider that Visual Question Answering over News Videos is an adequate task to advance in models that jointly reason over temporal and scene-text-based information. Since the proposed problem deals with answering questions based on reading text present in the videos, we start by building the dataset needed. We do this by first defining the task mathematically and explaining the major components of the ideal system for this setup. We also

present multiple baselines that are used as question-answering methods for related tasks in other areas, such as natural language processing, and also single frame-based VQA methods, and VideoQA models. We also repurpose the original VideoQA model to incorporate OCR tokens to process them along with the visual and question features.

In the second part of this work, we attempt to explore multiple other datasets, such as M4-ViteVQA, for the task of text-based video question-answering. This dataset contains videos from multiple categories, such as shopping, vlogging, traveling, and so on. We perform data exploratory analysis which is aimed to check for the distributions and different types of question-answer pairs in both NewsVideoQA and M4-ViteVQA datasets. The analysis aims to look into the type of question-answer that occurs the most in the datasets. This experiment allows us to gain a deeper understanding of the dataset by identifying the common structure in question-answer pairs and their distribution. It also helps to design and evaluate different methods. By analyzing question-answer types, we also aim to reveal potential bias in the dataset that can be improved in the later stages of creating bigger datasets. We later perform an exhaustive analysis of the text-only model BERT-QA model and its different variants. By doing so, we show that the current datasets lack in true multi-modal nature of the task setup. Towards the end, we present some possible directions that might lead to building better datasets that are truly multi-modal in nature.

1.0.1 Contributions

As mentioned in the previous sections, this thesis deals with video question-answering based on text present in videos. To this end, the following are our core contributions:

1. We introduce a new task of text-based Video Question Answering, in which models must have the ability to read and reason about the text in the videos (multi-frame input) to answer questions.

2. In the first work, We propose a new dataset: NewsVideoQA to explore the proposed task. This dataset comprises questions defined over the textual content in news videos and requires models to read and reason over it to obtain an answer.

3. We evaluate various baselines on the NewsVideoQA dataset. These baselines include simple heuristic methods, like text-only (machine comprehension) models, Scene-text VQA models, and VideoQA models. We also repurpose a video question-answering model to incorporate OCR tokens for the proposed task.

4. In the second work, we empirically show that current text-based VideoQA datasets have certain limitations, such as reasoning over only textual information and information from a single frame to obtain the answer to the question. We then provide insights on how to construct better datasets to make the task more realistic. We show that a simple repurposed text-only model like BERT-QA can achieve SOTA performance on both NewsVideoQA and M4-ViteVQA datasets. Along

with the NewsVideoQA dataset, we experiment with another text-based video question-answering dataset, M4-ViteVQA.

5. We provide possible guidelines that might lead to better formulation and data collection in the future that can boost the true nature of text-based video question-answering tasks.

1.1 Organization of the Thesis

The rest of the thesis is organized as follows:

1. In Chapter 2, we summarise existing works on VQA, Scene-text VQA, and Video Question Answering for both methods and datasets.

2. In Chapter 3, we shed light on the need for the newly proposed task of text-based Video Question Answering. We propose the NewsVideoQA dataset, its collection process, statistics, and data analysis, and repurpose different baselines to cater to NewsVideoQA.

3. In Chapter 4, we explore another dataset: M4-ViteVQA. We revisit NewsVideoQA and M4-ViteVQA and check for their limitations. We repurpose a text-only model that achieves almost the same performance and is better in most cases when compared to multi-modal methods, showcasing the limitations and bias in the current datasets. We also experiment on recent vision-langauge models like LLaVA and Gemini on these two datasets. We further detail some related works that explain the bias in the dataset and provide new guidelines that might help researchers come up with better datasets in the future.

Chapter 2

A Review: Utilizing of Text present in the Scene, Video Question Answering and Beyond

In this chapter, we first have a look at existing works on visual question answering, followed by works of text-based visual question answering and later works on video question answering. We discuss their shortcomings and provide the motivation for the need of the proposed task.

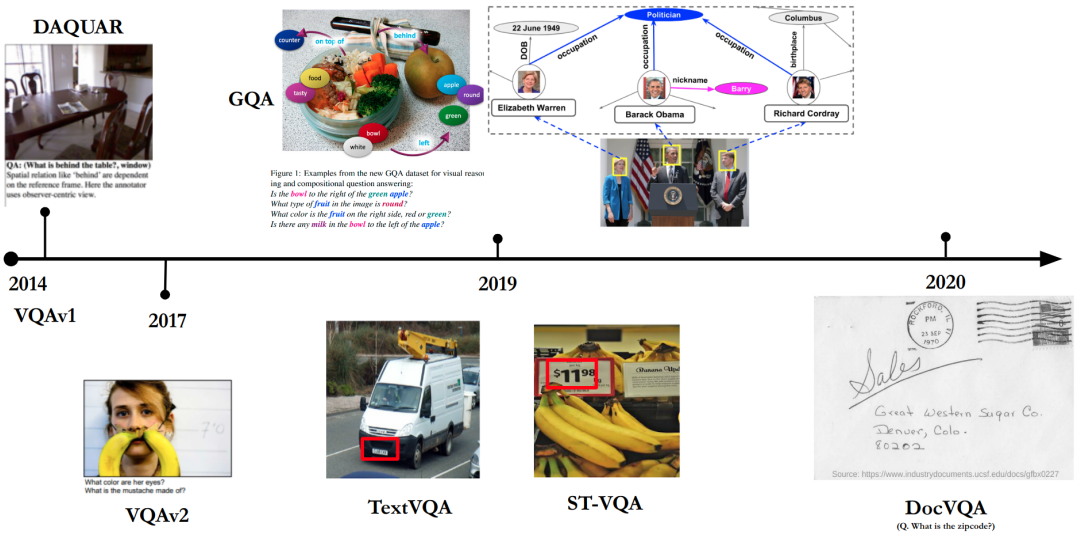


Figure 2.1 Timeline of VQA and VideoQA works. We present a timeline (does not contain all VQA works) of Visual Question Answering and major milestones over a couple of years.

2.1 Visual Question Answering

Visual Question Answering is a task of generating natural language answers when a question in natural language is asked related to an image. In recent years, many datasets, methods, and metrics have been proposed to enrich the research in this field of multi-modal learning. Researchers have looked into

different aspects of VQA, ranging from simple image question-answering, commonsense-based VQA, external-knowledge-based VQA, and scene-text-based VQA. Each of these different types brings in new contributions to the field and important applications which in turn help our society. Realizing the importance of scene-text understanding, Biten et al. [4] and Singh et al. [46] introduced two datasets on VQA on scene text. Specifically, text-based visual question-answering systems focus on scene text present in the images and reason over both textual and visual information in the scene to obtain answers. Another natural application of computer vision is to assist blind people, whether that may be to overcome their daily visual challenges or to break down their social accessibility barriers. In order for any research community to progress, it should begin with large-scale publicly-shared datasets that might help in the task of VQA for the visually impaired [13].

Methods. One of the first architectures in VQA was a method proposed in [2]. In this architecture, the image embedding is done using VGGNet [44] and question embeddings using LSTM. MLP was used to obtain the answer. Furthermore, a number of attention-based models [43] were proposed. These methods were based on generating spatial maps to highlight image regions that are relevant to answering the question. After the usage of transformers in the language domain, it was natural that multi-modal tasks like image captioning, VQA, and VideoQA also used transformer-based architectures. With the notable advancements in pretraining, vision and language community started building models that used vision-and-language pretraining (VLP) [47]. These VLP tasks typically adopt image-text matching (ITM) and masked language modeling (MLM) objectives on images and their corresponding captions during pretraining and then finetuned on downstream tasks such as VQA. Initial models such as LXMERT [47], UNITER [6], OSCAR [29] employed object detectors like Faster R-CNN [41] and YOLO [40] which were pretrained on the visual genome dataset [22].

2.2 Text-based Visual Question Answering

As we know traditionally VQA works in the literature focus on the visual content, and ironically fall short in answering the questions that require reading the text in the image. Knowing the importance of text in the images for answering visual questions, researchers have started focusing on text-based VQA [4, 33, 35, 46, 50]

2.2.1 Scene-text VQA

Two popular benchmarks for English scene text VQA are STVQA and TextVQA. These two datasets were introduced in parallel in 2019 and were quickly followed by more research on Multi-modal graph neural networks for joint reasoning on vision and scene text [11], Knowledge-Aware Visual Question Answering (KVQA) [42], and works on bilingual scene-text visual question answering [56].

ST-VQA is a dataset that contains a series of tasks of increasing difficulty for which reading the scene text in the context provided by visual information is necessary to reason and generate an appropriate

Table 2.1 Statistics of ST-VQA dataset. This table shows the number of images and questions from each dataset present in the ST-VQA [4] dataset.

Original Dataset	Images	Questions
Coco-text	7,520	10,854
Visual Genome	8,490	11,195
VizWiz	835	1,303
ICDAR	1,088	1,423
ImageNet	3,680	5,165
IIIT-STR	1,425	1,890
Total	23,038	31,791

answer. This dataset contains 31,791 questions over 23,038 images collected from different public datasets namely: ICDAR 2013 [19], ICDAR 2015 [18], ImageNet [8], VizWiz [13], IIIT Scene Text Retrieval [34], Visual Genome [22], COCO-Text [55]. Table. 2.1 shows the statistics of the number of images and questions obtained per dataset. There are 19,027 images - 26,308 questions for training and 2,993 images - 4,163 questions for testing. The dataset contains three tasks of increasing difficulty that stimulate different degrees of availability of contextual information.

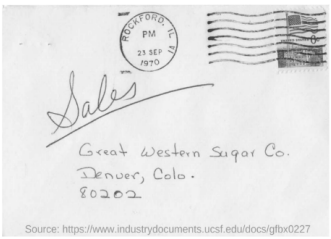
The TextVQA dataset has 45,336 questions over 28,000+ images sampled from specific categories of OpemImages dataset [20], that are expected to contain text. It uses the Open Images v3 dataset [20] as the source. It uses several categories in Open Images that fit the criterion which requires images containing text such as billboards, traffic signs, whiteboards, etc. Both these datasets consider the text present in real scenes to answer questions.

Methods. Along with the TextVQA dataset, Singh et. al. [46] introduced an approach: Look, Read, Reason $\&$ Answer (LoRRA). This model incorporates the regions (bounding boxes) in the image containing text as entities to attend to (in addition to object proposals). It is important to consider the fact that ST-VQA is a dataset that comprises images from other datasets specific to scene text, unlike TextVQA where the questions don't need to require scene text understanding. This means that the subset of questions in the ST-VQA dataset is larger when compared to the TextVQA dataset.

In the previous methods like LoRRA, the models were trained for the classification task to predict answers. The vocabulary used to train such methods restricts them from generating answers in an open-ended setting. These methods rely on the distribution of answers present in the training set which might not always be the ideal case, as the answers are scene text present in the image in most cases. Hu et. al. [16] proposes M4C that uses a multi-modal transformer-based model for STVQA and TextVQA. Unlike LoRRA, this model can generate answers of any length by combining tokens from a fixed vocabulary or the scene text tokens found in the image. In [63], Yang et al. introduce a text-aware pre-training (TAP) for TextVQA and Text-Caption tasks. TAP is designed to follow text-aware pre-training tasks to better fuse scene text (including both scene text words and their visual regions detected by OCR) with the text words (question tokens) and visual objects. The scene-text language includes two scene-text language pretraining tasks based on masked language modeling (MLM) and image-text (contrastive) matching (ITM) tasks. Specifically for MLM on the extended text input w with a probability of 15% random

Question: What does it say near the star on the tail of the plane?
Answer: Jet
Dataset: TextVQA

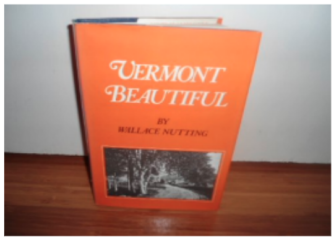
Question: What is the price of bananas per kg?
Answer: $11.98
Dataset: ST-VQA



Question: Mention the zip code written?
Answer: 80202
Dataset: DocVQA

Question: How many companies have more than 10K delivery workers?
Answer: 2
Dataset: InfographicVQA

Question: Does the foundation have any sunscreen?
Answer: Yes
Dataset: VizWiz



Question: What is the title of this book?
Answer: Vermont Beautiful
Dataset: OCR-VQA

Figure 2.2 Text-based VQA Dataset examples. Examples from different Text-based VQA datasets such as TextVQA [46], ST-VQA [4], DocVQA [33] and many others.

tokens are masked. For ITM, words from one image are polluted 50% of the time by replacing a text subsequence with a randomly selected one from another image.

2.2.2 Document VQA

Research in Document Analysis and Recognition has traditionally focused on information extraction tasks. Yet recently there have been works that look into document understanding through a Visual Question Answering perspective.

2.2.2.1 DocVQA

Mathew et al. [33] introduced Document Visual Question Answering (DocVQA) as a high-level task driving DAR algorithms to conditionally interpret document images. In Fig. 2.2, we show an example from the DocVQA dataset. DocVQA is a large-scale dataset of 12,767 document images of varied types and content, over which there are 50,000 question-answer pairs. The questions defined are categorized based on their reasoning requirements. The data is split randomly in an 80:10:10 ratio to train, validation, and test splits. The train split has 39,463 questions and 10,194 images, the validation split has 5,349 questions and 1,286 images and the test split has 5,188 questions and 1,287 images. A challenge [DocVQA challenge](https://www.docvqa.org/) on the same dataset was introduced to encourage more researchers to explore this new area. There are nearly 70+ submissions in the challenge, presenting multiple methods to solve this important task. The methods are ranked based on Accuracy score and ANLS (average Levenshtein distance). The majority of the questions in this dataset focus on information in tables, forms, and paragraphs among others.

2.2.2.2 Infographic VQA

InfographicsVQA [32] extended the task of DocVQA to infographic images. As infographics communicate information using a combination of textual, graphical, and visual elements, it becomes an important problem to build systems that can jointly reason over elements such as document layout, textual content, and, graphical elements. With more visually rich elements and answers that can be either extractive from a set of multiple text spans in the images, a multiple choice given in the question, or the result of a discrete operation resulting in a numerical non-extractive answer. Mathew et al. propose a dataset InfographicVQA that contains a diverse collection of infographics and question-answer annotations. It contains 30,035 questions over 5,484 images. The dataset contains four types of answer-source: image-span, question-span, multi-span, and non-extractive.

2.2.2.3 Others

DocCVQA [50] was a step towards better understanding document collections which was beyond just word spotting. The objective of DocCVQA was to extract information from a document image

collection by asking questions and expecting the methods to provide the answers. The dataset comprises 14,362 document images sourced from the Open Data portal. DocCVQA dataset has a setup of retrieval-answering tasks.

Landeghem et al. [23] propose a challenge that seeks to benchmark progress in understanding visually rich documents (VRDs). It presents a new dataset (DUDE) that contains novelties in terms of types of questions, answers, and document layouts based on multi-industry, multi-domain, and multi-page VRDs from various origins and dates. The variability of the dataset is such that, it motivates models that can answer natural yet highly diverse questions (e.g. regarding document elements, their properties, and compositions) for any VRD (e.g. drawn from potentially unseen distributions of layouts, domains, and types).

"Document Visual Question Answering" (DocVQA) challenge in ICDAR 2023 [38], presented a task on understanding business documents as a crucial step towards making an important financial decision. The competition was designed to find answers to the questions with minimal human supervision. Some problem-specific challenges included accurate understanding of the questions/queries, figuring out cross-document questions and answers, the automatic building of domain-specific ontology, accurate syntactic parsing, calculating aggregates for complex queries, and so on.

Tito et el. [51] proposes a new multi-modal hierarchical method Hi-VT5, that overcomes the limitations of current methods to process long multipage documents. Hi-VT5 is a multi-modal hierarchical encoder-decoder transformer built on top of T5. It is capable of naturally processing multiple pages by extending the input sequence length to up to 20,480 tokens without increasing the model's complexity. The encoder processes separately each page of the document, providing a summary of the most relevant information conveyed by the page conditioned on the question. This information is encoded in multiple special tokens [PAGE] tokens, which were inspired by the [CLS] token of the BERT model. Subsequently, the decoder generates the final answer by taking as input the concatenation of all these summary [PAGE] tokens for all pages. Furthermore, the model includes an additional head to predict the index of the page where the answer has been found.

2.3 Video Question Answering

One of the early attempts at VideoQA is a retrieval-based approach for factoid QA proposed by Yang et al. [62]. Their system relies on speech transcripts and external knowledge to answer the questions. One or more sentences from the transcript are returned as the output of the QA system, and the output is considered correct if the target answer is contained within the retrieved sentences. For QA evaluation, they used a private dataset containing only 40 QA pairs. Contrary to this work, our NewsVideoQA focuses particularly on the text appearing in the news videos and is defined over a much larger dataset.

More recent works in VideoQA [26, 48, 59, 65] require models to reason about the events taking place in videos, but disregard any textual information in the videos. Tapaswi et al. [48] introduced a dataset that aims to study story comprehension using video and subtitles. Zhou et al. [65] introduced

a large-scale VideoQA dataset that consists of videos of different activities. A method that gradually refines attention over the appearance and motion features is proposed in [59], along with an automatically generated dataset for VideoQA using subtitles. Yang et al. [61] and Maharaj et al. [31] focused on automatic generation of the VideoQA datasets. As the questions in [61] are automatically generated using captions, they are largely based on the visual appearance of objects and actions. Gupta et al. [12] explore knowledge-based question answering on news videos by proposing a new dataset. Questions in this dataset are primarily concerned with people seen in the videos, and the proposed models primarily rely on transcripts and an external knowledge base to find the answer. Questions in the above-mentioned works primarily require visual content and the transcripts of the videos to answer questions. Recently works such as [24,25,27,28] have introduced transformer-based models with different pretraining strategies and yield state-of-the-art performance on existing VideoQA datasets.

Table 2.2 summarises existing works on VideoQA. It can be seen that the majority of models focus on the visual content, transcripts, and external knowledge to answer the questions. The text seen in the videos is an important source of information critical to understanding the content of news videos and videos shot outdoors. However, existing works on VideoQA largely disregard text in the videos. This motivates the community to have a publicly available video question-answering dataset in which the questions require an understanding of the textual content in the videos to obtain the answers. Works such as [24, 25, 27, 28] have introduced transformer-based models with different pretraining strategies and yield state-of-the-art performance on existing VideoQA datasets.

Table 2.2 A comparative overview of VideoQA datasets. Datasets prior to our work, consider video, video + subtitles, video + knowledge base as input. Our work introduces a new line of research where the questions in the proposed dataset are framed based on textual content in the news videos. The column Synthetic Gen. indicates the dataset that is synthetically/automatically generated.

Dataset	Subtitles	Text in video	Type of videos	Synthetic Gen.	Free-form	#Video	#QA
VideoQA [67]	✗	✗	Cooking, movies	✓	✗	109K	390K
MSVD-QA [59]	✗	✗	YouTube	✓	✓	1.9K	50K
ActivityNet-QA [65]	✗	✗	YouTube	✗	✓	5.8K	58K
MSRVTT-QA [59]	✗	✗	YouTube	✓	✓	10K	243K
MoviesQA [48]	✓	✗	Movies	✗	✗	6.7K	6.4K
TVQA [26]	✓	✗	TV shows	✗	✗	21K	152K
HowtoVQA69M [61]	✓	✗	TV shows	✓	✗	69M	69M
QA News Videos [62]	✗	✓	Web videos	-	-	-	40
NewsKVQA [12]	✓	✗	News videos	✓	✗	5.8K	58K
NewsVideoQA (Ours)	✓	✓	News videos	✗	✓	3.0K	8.6K

Chapter 3

NewsVideoQA: Text-based Video Question Answering dataset on News Videos

In this chapter, we first define the proposed task which is given a news video clip and a question, we aim to obtain the answer by reading the text in the video. We later present details about how we plan to solve this task by proposing a new dataset, NewsVideoQA. We further explain the process of collecting data and statistics. We then explain different baselines and experiments.

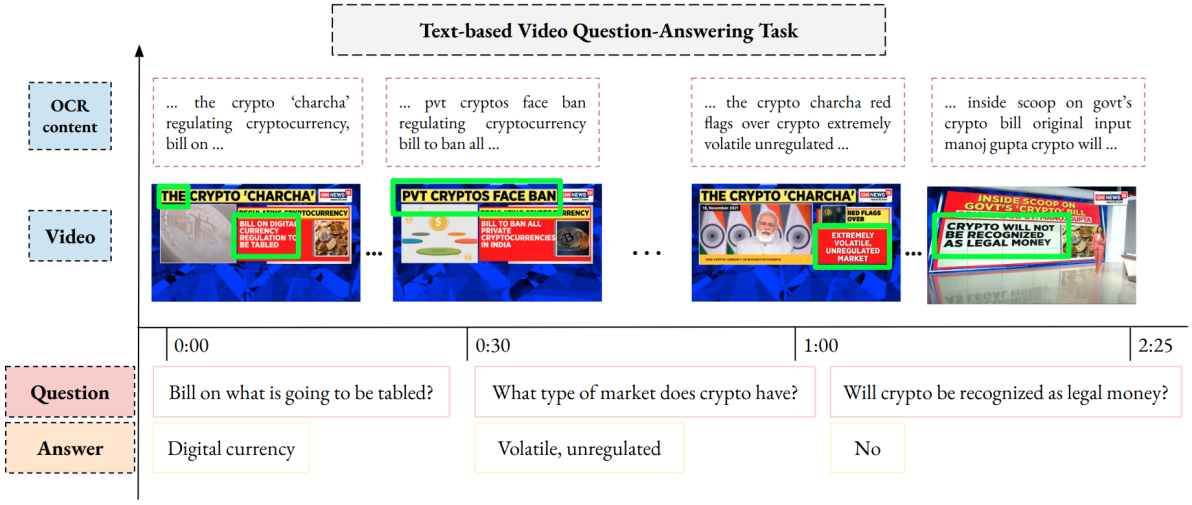


Figure 3.1 NewsVideoQA: Task and Dataset. Examples of the NewsVideoQA dataset showcasing the importance of text in the videos to answer questions. Examples from multiple topics are shown to indicate the variability in the type of questions.

3.1 Task definition

We define text-based Video Question Answering (VideoQA) as a multi-modal task that aims to automatically generate answers to questions asked about the textual content of a given video. In this task, we are provided with :

1. A video, denoted as V with set of frames $f_1, f_2, \ldots, f_n$.

2. Textual tokens extracted from the video using Optical Character Recognition (OCR) $t_1, t_2, \ldots, t_m$, denoted as T.

3. A question, denoted as Q with tokens $q_1, q_2, \ldots, q_k$.

The goal of this task is to generate an appropriate answer for the question Q based on the information contained in the V and the OCR tokens. Formally, for each question in the dataset, the text-based VideoQA task can be formulated as finding the answer A such that:

$$A_i = \arg\max_{a \in A} P(a|V = \{f_1, f_2, \ldots\}, T = \{t_1, t_2, \ldots\}, q_i)$$

The following are the major sub-models that are needed to solve the proposed task:

1. **Temporal component:** to reason and infer the answer based on temporal information from multiple frames and their visual information.

2. **Reading component:** This allows the methods to read the text in the image.

3. **Fusion component:** This is used to combine the feature from multiple modalities.

4. **Answering component:** This module is used to generate answers by the output of the fusion component.

3.2 NewsVideoQA Dataset

We attempt to solve this task by introducing a new dataset: **NewsVideoQA**. Motivated by the prominent function of scene text in news video snippets, and the complementary information it carries to the visual modality, we consider that Visual Question Answering over News Videos is an adequate task to advance in models that jointly reason over temporal and text-based information. Also, the availability of news videos in abundance and the type of information they contain, make them a good starting point. It is also important to note that as part of future research, news videos have similar news content in several languages which can further enhance the features of an assistive device that can be developed for such VQA systems.

We start this section by explaining the source of data collection, followed by the annotation step that contains two stages: (i) the annotation stage, and (ii) the verification stage. We also share the statistics and the analysis of the proposed NewsVideoQA dataset.

3.2.1 Data Collection

In order to collect videos from multiple news channels, we first looked into multiple categories of videos that were common throughout multiple news channels. This helped us obtain similar video category types along with variations in the displaying of text in the videos. While curating the videos, we made sure that the videos contained text in them. We downloaded the videos using the YouTube-DL library where all the videos were of varied length.

3.2.1.1 News Videos

We collect news videos from English news channels around the world. We obtain videos from the following YouTube channels: British Broadcasting Corporation (BBC), Australian Broadcasting Commission (ABC) Australia, India Today, Turkish Public Broadcaster (TRT) World, AL Jazeera, Cable News Network (CNN), NHK World Japan, Fox News, World is One News (WION), New Delhi Television Limited (NDTV), American Broadcasting Company (ABC) News, Cable News Network-News18 (CNN-News18), Connected TV (CTV) News, China Global Television Network (CGTN), and Intergovernmental Panel on Climate Change (IPCC). As mentioned above, while collecting the news videos, we manually ensure that the videos are text-rich because the proposed task relies on video question answering, which requires reading text. The collected videos are split into 10 seconds of non-overlapping clips. The proposed dataset contains 3,083 clips, with at least 20 videos from each channel. The average number of questions per video is 2.96. The maximum number of questions defined for a video in the dataset is 20. The minimum number of questions defined for a video is 1.

3.2.1.2 Annotation tool

We modify an existing in-house annotation tool to incorporate the annotations needed for text-based VideoQA. This tool was initially designed for the task of DocVQA [33]. We modify this to incorporate the news videos and obtain annotations for them. Using this tool, along with the question-answer pairs we also collected the timestamp at which the question was defined.

3.2.1.3 Question and Answers

The annotation process was organized into two stages. In stage 1, the annotators were instructed to define question-answer pairs based on textual information present in the news videos. Specifically, they were provided with the following instruction: 'Ensure that answering the questions generated requires reading of the text present in the news videos and should be related to the topic of that video'. Annotators were asked to frame factoid questions that can be answered by reading the text present in the news videos. They were also instructed to add a timestamp: the time (with up to 1-second precision) of the video when the question was framed.

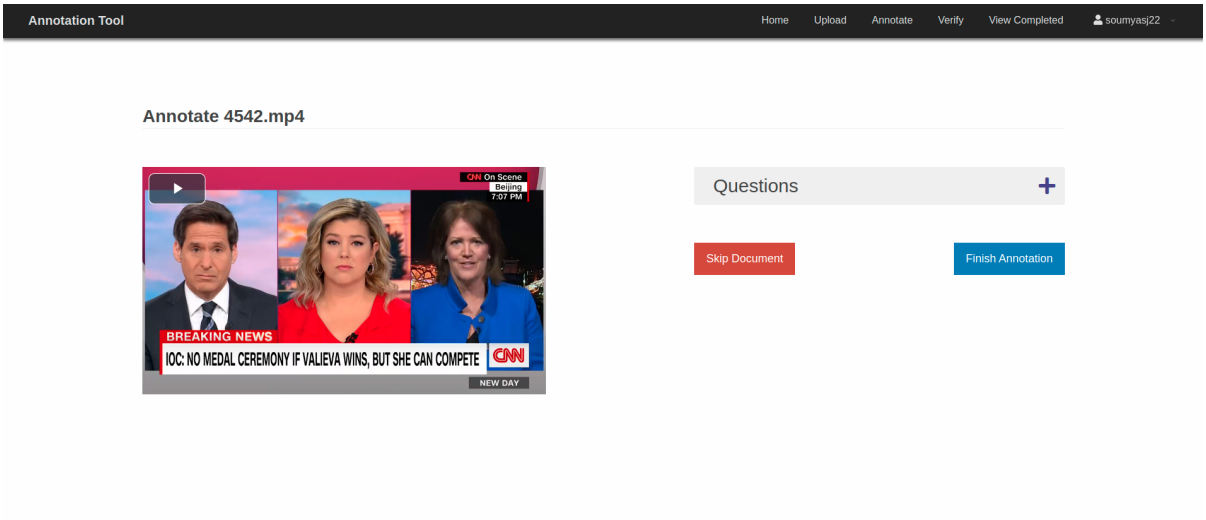


Figure 3.2 Image of Annotation tool. Annotation tool used for annotating QA pairs. The general setup for the annotation tool.

A second stage of verification was introduced to check the correctness of the data. Here, the annotators were asked to verify the data collected in the first stage. The annotators were shown the video question pair for a video clip and were asked to enter the answer and the timestamp and check the correctness of the question-answer pair based on its relevance to the textual content of the news video. They were asked to reject the questions with any grammatical mistakes in the questions or answers. During this stage, if the annotator finds a question-answer pair irrelevant to the topic or if the question was framed on the audio of the news videos, then such question-answer pairs were rejected from the dataset. A total of $1,200$ QA pairs were rejected after the verification step. An extra stage was also added where the authors reviewed randomly picked question-answer pairs and their correctness and relevance to the task proposed.

3.2.1.4 Statistics and Analysis

The NewsVideoQA dataset comprises $8,672$ questions framed on $3,083$ news videos. The data is split randomly in the 80-10-10 ratio to train, validate, and test split. The train split has $6,994$ questions over $2,407$ videos, the validation split has 714 questions over 330 video clips, and the test split has 964 questions over 346 video clips. Fig. 3.9a shows the distribution of question lengths for the questions in the NewsVideoQA dataset. The average question length is 7.04 words. Among the $8,672$ questions $7,008$ (80.81%) are unique. Higher diversity in questions is reflective of the fact that questions are based on textual content. Fig. 3.9d shows the top 15 most frequent questions and their frequencies. Fig. 3.10 shows a sunburst plot of the first three words of the questions. It can be observed from Fig. 3.10 that there is variability in the question types like questions starting with *"What"* that are likely related to the text in the videos, such as "What is the ...". We provide subtitles of the news videos using a publicly

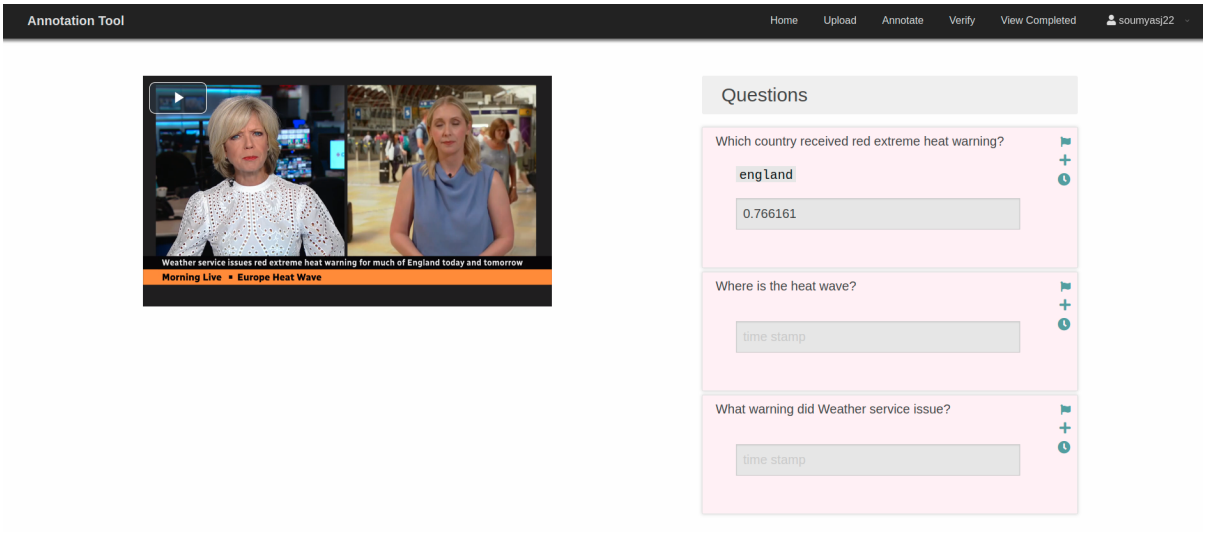


Figure 3.3 Example of annotation. Annotation tool with questions, answers, and respective timestamps.

available speech-to-text tool [49]. A total of $1,388$ (17.36%) questions can be answered with sub-titles of the videos. This low percentage is observed due to two reasons, (a) the smaller duration of the videos (10 seconds), resulting in incomplete sentences in the subtitles, and (b) all of the questions are based on the textual content of the news videos. In total, there are $4,150$ (47.85%) unique answers. The word cloud on the right in Fig. 3.4 shows the most common words in the answers. The answer space is broad and involves names of countries, events, games, people, etc. The distribution of answer lengths is shown in Fig. 3.9b. The average answer length is 2.02. The top 15 answers in the dataset are shown in Fig. 3.9e. We obtain OCR tokens using Google OCR. We uniformly sample the video at 2 frames per second and also retain the first frame of the video. Fig. 3.4 on the left shows the word cloud of OCR tokens. In Fig. 3.9f we show the top 15 OCR tokens present in the dataset. An average of 26.14 OCR tokens per frame is observed, and an average of 532.55 OCR tokens per video clip are observed in the dataset.

3.3 Baseline Methods

We evaluate three different methods as strong baselines for the newly introduced task of scene-text aware VQA on NewsVideoQA dataset. In this section, we briefly discuss the original methods and explain how these methods are adapted for the new task.

3.3.1 Heuristic methods and Upper Bounds

Inspired by heuristic baselines evaluated on scene text VQA [4, 46] and DocVQA [33] datasets, we evaluate the following heuristic baselines and upper bounds:

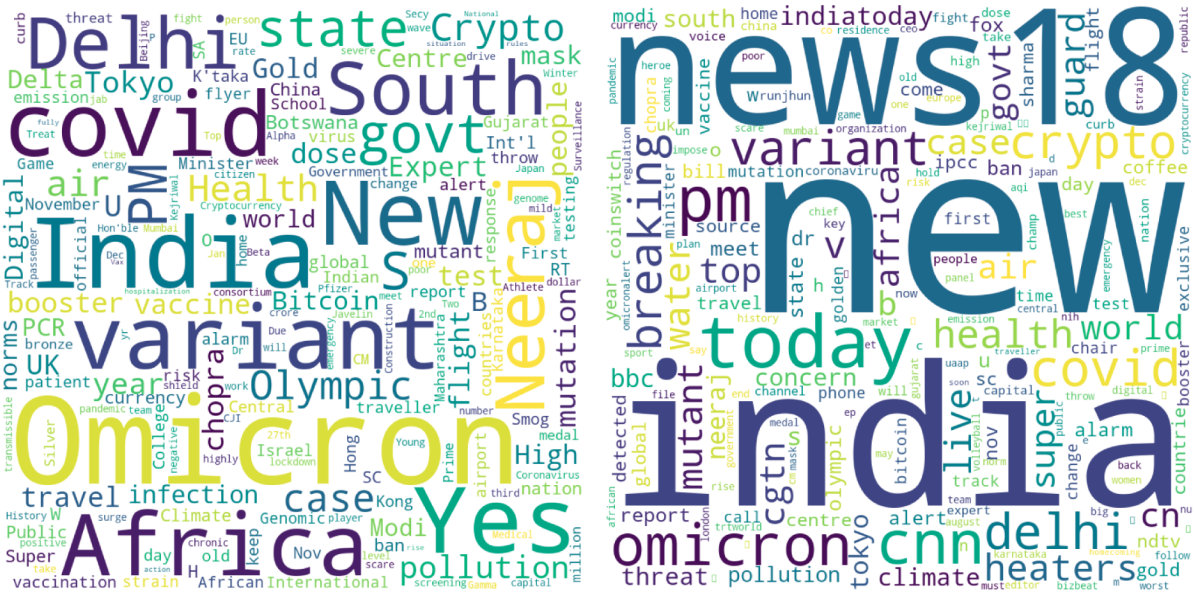
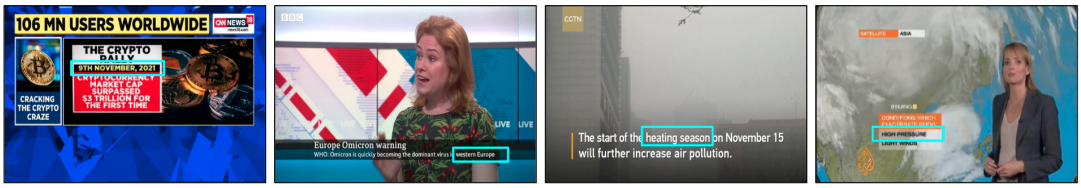


Figure 3.4 Word Clouds. Word clouds of words in answers (left) and word clouds of words in OCR tokens (right)



Question: When did crypto market surpass 3 trillion dollar for the first time? **Answer:** 9th november 2021	**Question:** Where is the omicron variant becoming dominant? **Answer:** western europe	**Question:** Start of what will increase air pollution? **Answer:** heating season	**Question:** What is the first condition that causes smog? **Answer:** High pressure

Figure 3.5 NewsVideoQA dataset samples. Examples of the NewsVideoQA dataset showcasing the importance of text in the videos to answer questions. Examples from multiple topics are shown to indicate the variability in the type of questions.

OCR tokens: ... toxic smog has returned to india's capital ... **Question:** Which country is facing toxic smog? **Answer:** india	**OCR tokens:** ... tokyo 2020 tokyo olympic events to be held without spectators ... **Question:** Which event is held without spectators? **Answer:** tokyo Olympic events	**OCR tokens:** ... bbc news un climate report 'wake-up call' ipcc ... **Question:** What are responsible for global warming? **Answer:** human activities	**OCR tokens:** ... pm hosts olympians pm shared ice cream with sindhu as promised ... **Question:** Who shared Ice Cream with the PM? **Answer:** sindhu

Figure 3.6 Extra examples. More examples from the NewsVideoQA dataset.

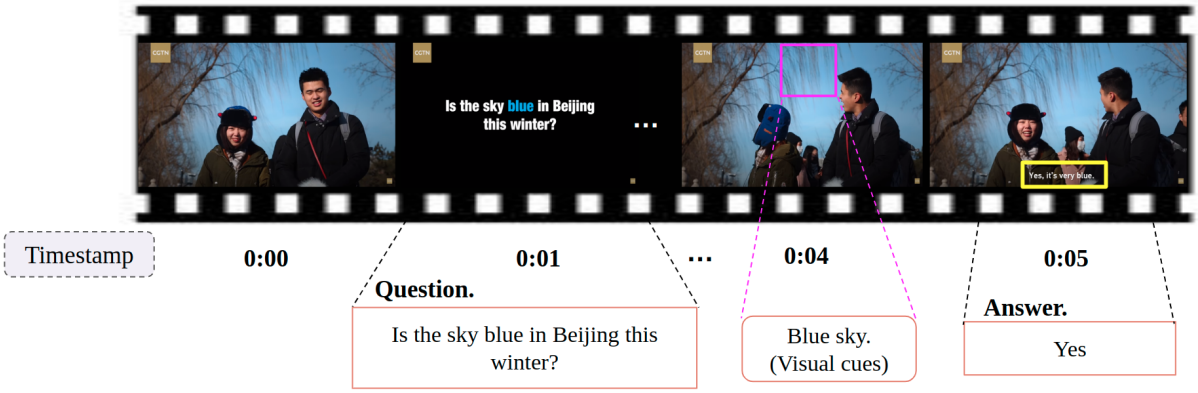


Figure 3.7 Specific Example. Examples of NewsVideoQA dataset showcasing the examples where a question is framed at a timestamp of 0:01 in a video. The answer to this question is in the timestamp of 0:05. Also, this question can be answered by visual cues in the video, but the text at 0:05 timestamp confirms the inference made by the text.

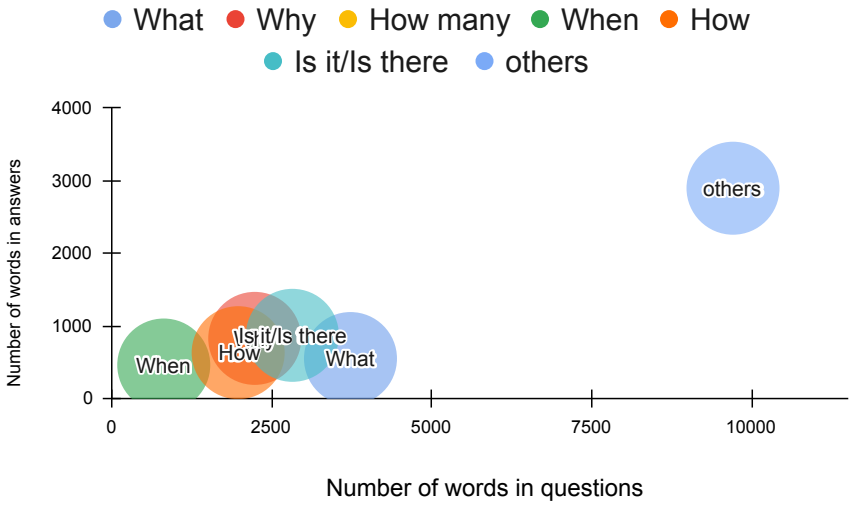
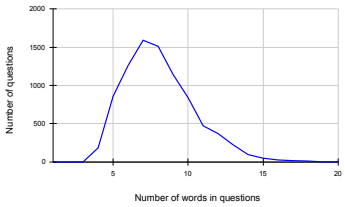
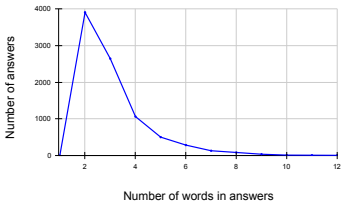


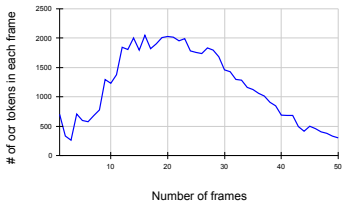
Figure 3.8 Question Distribution. A bubble chart with the number of words in question on the horizontal axis and the number of words in answers on the vertical axis based on the question type. Note that there is a diverse range of types of questions in the dataset. The question type *"What"* has a maximum count with questions such as, *"What could be the reason ...?", "What is the value..?", "What is one of the..?"* and so on.



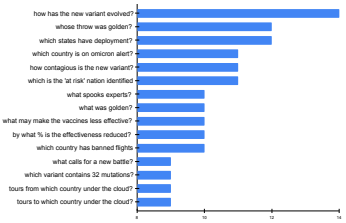
(a) **Questions with particular length.** The average length of questions in the dataset is 6.79 words.



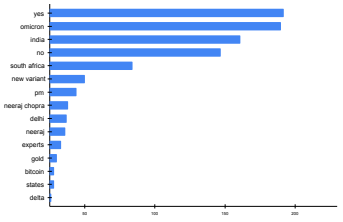
(b) **Answers with particular length.** The average number of words in the answers is 2.02 words.



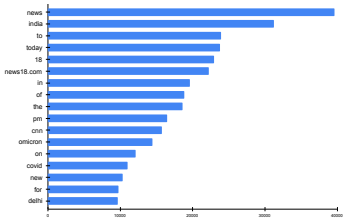
(c) **OCR tokens with particular length.** Average number of OCR tokens per frame is 26.14 tokens.



(d) Top 15 most occurring questions in the dataset.



(e) Top 15 most occurring answers in the dataset.



(f) Top 15 most occurring OCR tokens in the dataset.

Figure 3.9 Dataset statistics. Statistics for question, answer and OCR tokens in **NewsVideoQA dataset**.

(i) Majority answer: measures the performance when the most frequent answer in the train split is considered as the answer for all the questions in the test set. **(ii) Biggest OCR token:** measures the performance when the OCR token that occupies the largest area in the video is considered as the answer.

We compute upper-bound (UB) performance for the following cases:

1. **Vocabulary UB:** measures the maximum performance obtainable on the test set, if an answer is picked from a vocabulary of most common answers in the train split.

$$VocabularyUB = \max_{a \in V} P(a) \tag{3.1}$$

where $VocabularyUB$ is the upper bound for the Vocabulary case. a represents a possible answer. V is the vocabulary of the most common answers in the train split. $P(a)$ is the probability of answer a.

2. **OCRs Substring of single frame UB:** This measures the performance that can be obtained when we restrict our vocabulary to a list of OCR tokens of the frame on which the question was defined.

3. **OCR Substring of all frames UB:** measures the performance we can obtain if the answer in the test split is a substring in the concatenated list of OCR tokens from uniformly sampled frames of the video.

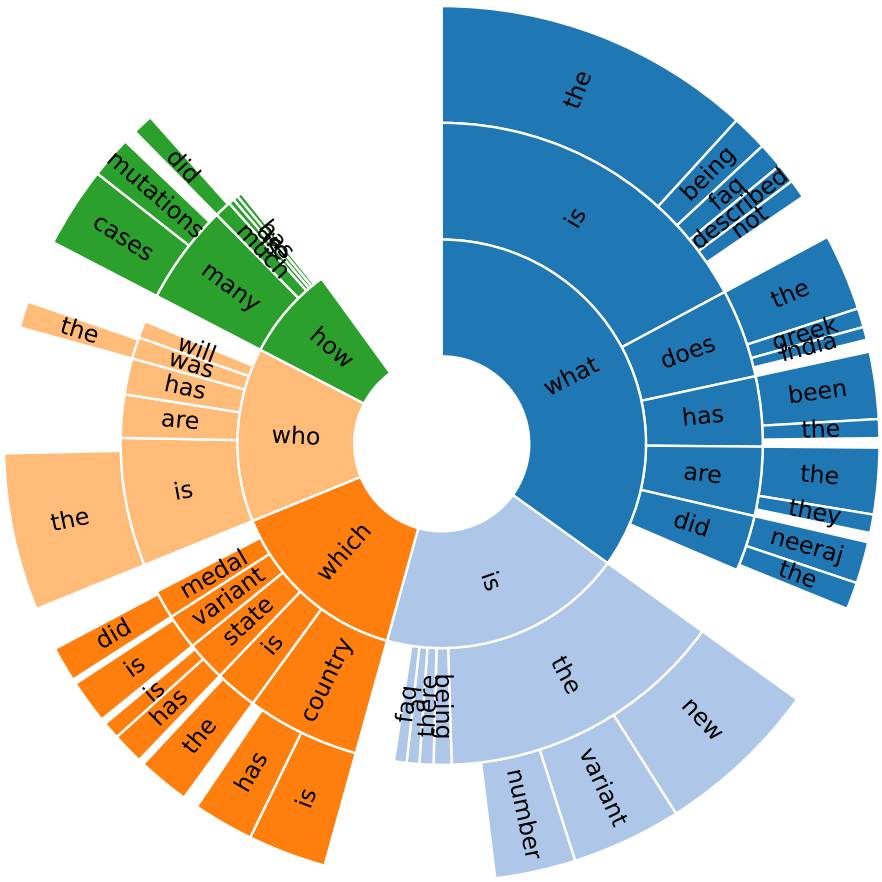


Figure 3.10 Sun-Burst for Questions in NewsVideoQA. Distribution of questions by their starting 3-grams. Note that there is a diverse range of types of questions in the dataset. The question type *"What"* has a maximum count with questions such as *"What is the ...?"*, *"What does the ...?"* and so on.

3.3.2 Reading comprehension model

As observed in the task definition, by design, almost all of the questions in NewsVideoQA are grounded on the text in the videos. For this reason, we evaluate a QA baseline that only considers the text in the videos to answer the questions. Specifically, we evaluate the BERT [9] QA model that is originally developed for extractive text-only QA. Extractive QA is the task of extracting a short snippet from the document/context on which the question is asked. The answer snippet is called a 'span' and the span is defined in terms of its start and end tokens. BERT is a transformer encoder-based method of pretraining language representations from unlabelled text. These pretrained models can be used later for downstream tasks with the addition of output suitable for the task at hand. BERT uses WordPiece embeddings [57] with a $30,000$ token vocabulary. This first token of every sequence is always a special classification token ($[CLS]$). The final hidden state corresponding to this token is used as the aggregate sequence representation for classification tasks at the time of pretraining and even in some cases for finetuning. Sentence-pairs (might differ from task to task such as ¡context, question¿ for question answering). A special token ($[SEP]$) is added to separate the two sentences. BERT is pretrained on two unsupervised pretraining tasks: (i) Masked LM, and (ii) Next Sentence Prediction (NSP).

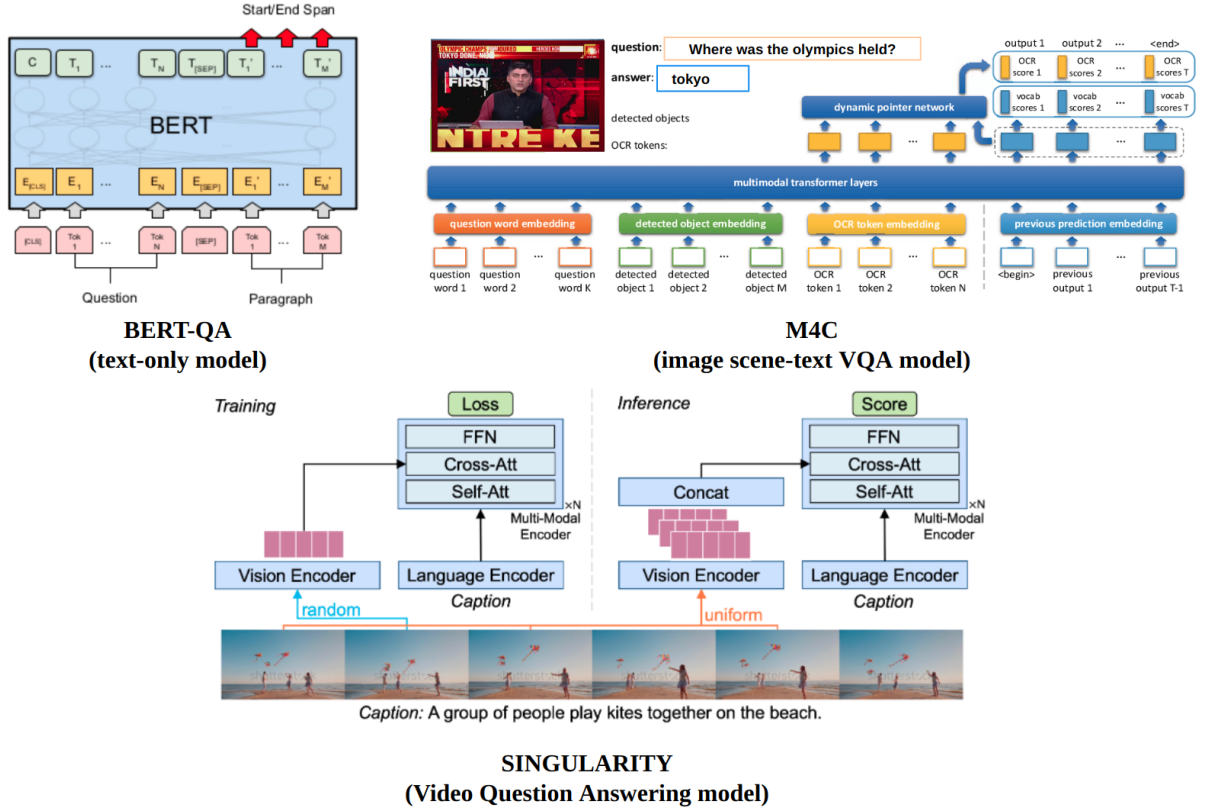


Figure 3.11 Baselines. BERT-QA - text-only model; M4C - single image scene-text based VQA model.

1. **Task 1: Masked LM** In this task some percentage of the input tokens are masked randomly, and then the task is to predict those masked tokens. For this task, the final hidden vectors correspond-

ing to the mask tokens are fed into an output softmax over the vocabulary, as in a standard LM. Around, 15% of all WordPiece tokens in each sequence at random are masked.

2. **Task 2: Next Sentence Prediction (NSP)** It is important to understand the relationship between two sentences, which is not directly captured by language modeling. In order to incorporate this, a binary task of next-sentence prediction is used. Specifically, when choosing the sentences A and B for each pretraining example, 50% of the time B is the actual next sentence that follows A and 50% of the time it is a random sentence from the corpus.

For finetuning, the Stanford Question Answering Dataset [39] which is a collection of 100K crowd-sourced question-answer pairs is used. Given a question and a passage from Wikipedia containing the answer, the task is to predict the answer text span in the passage. This was the original task on which the BERT model was finetuned and this finetuned version of BERT is called BERT-QA (BERT model with question-answering capability). The input question and passage is represented as a single packed sequence, with the question using the A embedding and the passage using the B embedding. A start vector $S \in R^H$ and an end vector $E \in R^H$. The probability of word i being the start of the answer span is computed as a dot product between T_i and S followed by a softmax over all of the words in the paragraph.

$$P_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_i}} \tag{3.2}$$

The score of a candidate span from position i to position j is defined as $S \cdot T_i + E \cdot T_j$, and the maximum scoring span where $j \geq i$ is used as a prediction. For the task of extractive QA, the additional layer, is an output layer that predicts these above-mentioned start and end tokens of the span of the answer. The training objective is the sum of the log-likelihoods of the correct start and end positions. For NewsVideoQA, we perform several experiments. In all of these experiments, we consider concatenation of OCR tokens as the context to the model. We convert the OCR tokens into a concatenated list of OCR tokens in default reading order for BERT-QA model.

3.3.3 VQA Model

To evaluate the performance of current VQA models on the NewsVideoQA dataset, we use the M4C [16] model which takes into account the text present in the frames of the news videos. Given a question and an image as inputs, this method extracts feature representations from three modalities – the question, the visual objects in the image, and the text present in the image. These three modalities are presented respectively as a list of question word features, a list of visual object features from an off-the-shelf object detector, and a list of OCR token features based on an external OCR system. For each modality, specific features are extracted and projected into a common d-dimensional semantic space.

1. **Embedding of question words.** Given a question as a sequence of K words, each word is embedded into a corresponding sequence of d-dimensional feature vectors ${x_k}^{ques}$ (where $k = 1, \ldots, K$) using a pretrained BERT [9] model.

2. **Embedding of detected objects.** Given an image, a set of M visual objects are obtained through a pretrained detector, Faster R-CNN [41]. Appearance features ${x_m}^{fr}$ using the detector's output from the m-th object (where $m = 1, \ldots, M$). A 4-dimensional locature feature ${x_m}^{b}$ from the m-th object's relative bounding box coordinates $[\frac{x_m in}{W_i m}, \frac{y_m in}{H_i m}, \frac{x_m ax}{W_i m}, \frac{y_m ax}{H_i m}]$, where $W_i m$ and $H_i m$ are image width and height respectively. Furthermore, location feature are projected into the d-dimentional space with two learned linear transforms (where d is the same as in the question word embedding), and are summed up as the final object embedding ${x_m}^{obj}$ as

$$ {x_m}^{obj} = LN(W_1 {x_m}^{fr}) + LN(W_2 {x_m}^{b}) \tag{3.3} $$

   where W_1 and W_2 are learned projection matrices. $LN(.)$ is layer normalization [3]. Also, the last layer of Faster R-CNN detector is finetuned during training.

3. **Embedding of OCR tokens with rich representations.** It is important to encode not only characters of text in image but also its appearance and spatial location in the image. Assuming that there are N OCR tokens in an image, for each OCR token the following features are extracted:

   (a) 300-dimensional FastText [5] vector ${x_n}^{ft}$: a word embedding with sub-word information

   (b) appearance feature ${x_n}^{fr}$ from the same Faster R-CNN detector in the object detection

   (c) 604-dimensional Pyramidal Histogram of Characters (PHOC) vector ${x_n}^{p}$ that captures what characters are present in the token. This is considered to be more robust to OCR errors and can be seen as a coarse character model

   (d) 4-dimensional location feature ${x_n}^{b}$ based on OCR token's relative bounding box coordinates.

   Each feature is projected into d-dimensional space and is summed up (after layer normalization) as the final OCR token embedding ${x_n}^{ocr}$:

$$ {x_n}^{ocr} = LN((W_3 {x_n}^{ft}) + W_4 {x_n}^{fr} + W_5 {x_n}^{p})) + LN(W_6 {x_n}^{b}) \tag{3.4} $$

A stack of transformer [54] layers is applied over these features in the common embedding space. Through the multi-head self-attention mechanism in transformers, each entity is allowed to freely attend to all other entities, regardless of whether they are from the same modality or not. This enables both inter-entity and intra-entity attention. In the end, answers are predicted through iterative decoding in an auto-regressive manner for a total of T steps, where each decoded word may be either an OCR token in

the image or a word from the fixed vocabulary of frequent answer words. Let ${z_1}^{ocr}, \ldots, {z_N}^{ocr}$ be the d-dimensional transformer outputs of the N OCR tokens in the image and V the words in the vocabulary that frequently appear in the training set answers. The fixed answer vocabulary score ${y_{t,i}}^{voc}$ for the i-th word (where $i = 1, \ldots, V$) is predicted as a simple linear layer as

$$ {y_{t,i}}^{voc} = ({w_i}^{voc})^T {z_n}^{dec} + {b_i}^{voc} \tag{3.5} $$

To select a token from the N OCR tokens in the image, a dynamic pointer network is augmented that is used to predict a copying score ${y_{t,n}}^{ocr}$ where $n = 1, \ldots, N$ for each token via bilinear interaction between the decoding output ${z_t}^{dec}$ and each OCR token's output representation ${z_n}^{ocr}$ as

$$ {y_{t,i}}^{ocr} = (W^{ocr} {z_n}^{ocr} + b^{ocr})^T (W^{dec} {z_t}^{dec} + b^{dec}) \tag{3.6} $$

To modify this setting to the NewsVideoQA dataset, we pair each question with the frame corresponding to the timestamp of the question defined and consider it as input to M4C. The objective is to generate the answer from either vocabulary or OCR tokens for that image.

3.3.4 VideoQA Model

In addition to the text-only QA model and the text-based VQA models, we evaluate the performance of NewsVideoQA on a recently proposed transformer-based Retrieval and VideoQA method called SINGULARITY [24]. This method studies the importance of temporal relations to answer questions. SINGULARITY is a vision-language model pretrained on many video and image captioning datasets [15, 21, 59, 60, 64, 65]. It consists of three main components: a vision encoder F_v, a language encoder F_l, and a multi-modal encoder H. The vision encoder is an image-level visual backbone. The language encoder is an arbitrary model such as BERT [9]. For the multi-modal encoder, a transformer encoder is used in which layer contains a self-attention, a cross-attention, and a feed-forward network (FFN). The cross-attention layer is used to gather information from encoded visual representation using the text as key.

Let a video V contain T frames as $V = [f_1, f_2, \ldots, f_T]$, and paired text as S. During training, a single frame is randomly sampled f_t from V as model input, where $t \in 1, \ldots, T$. Its encoded representation can be written as $F_v(f_t) \in \mathbb{R}^{L_v x D}$. For the text, the encoded representation is $F_l(S) \in \mathbb{R}^{L_l x D}$. L_v and L_l are encoded sequence lengths, D is hidden size. The prediction p is made as:

$$ p = H(F_l(S), F_v(f_t)) \tag{3.7} $$

Q, K, and V which denote the query, key, and value matrices of self- and cross-attention. During inference, the frames are sampled uniformly T_{test} as ${f_{Ti}}_{i=1}^{T_{test}}$. Each frame is encoded separately, and

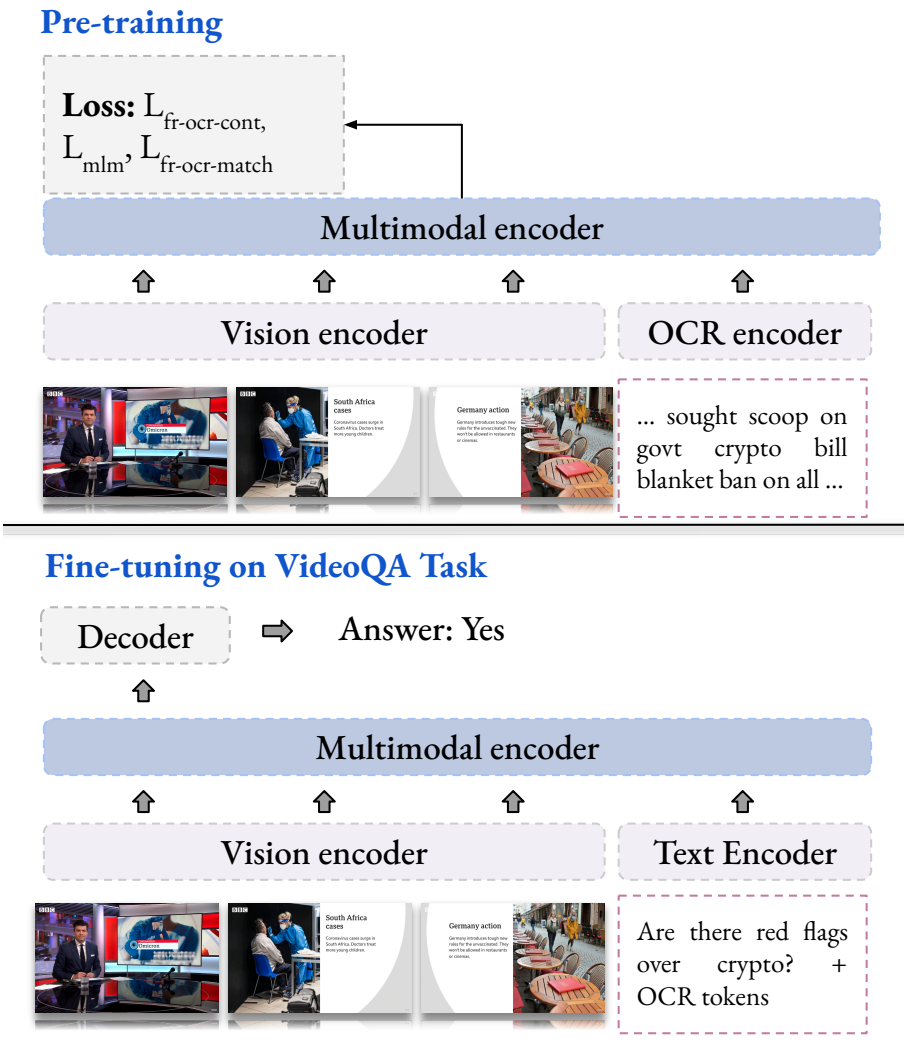


Figure 3.12 OCR-aware SINGULARITY. We extend SINGULARITY [24] for the task of text-based video question answering by incorporating OCR information by pretraining and finetuning on proposed NewsVideoQA dataset.

their encoded representations are concatenated as inputs to the multi-modal encoder to get a video-level prediction score:

$$p = H(F_l(S), [F_v(F_{T1}); \ldots; F_v(f_{TTtest})]) \tag{3.8}$$

where $[;]$ denotes concatenation, and $[F_v(F_{T1}); \ldots; F_v(f_{TTtest})] \in \mathbb{R}^{(T_{test} x L_v) x D}$. Early fusion design allows the model to make an informed prediction given the full context. For pretraining, each video/image is paired with its corresponding caption. The multi-modal encoder applies cross-attention to collect information from visual representations using the text as the key. Three pretraining objectives are defined: (i) Vision-Text Contrastive: a contrastive loss that aligns vision and text representations, (ii) Masked Language Modeling (MLM): predicts the masked visual and text contexts, and (iii) Vision-Text Matching: predicts the matching score of a vision-text pair with multi-modal encoder. For QA task, a multi-modal decoder is initialized from pretrained multi-modal encoder, which takes the outputs of multi-modal encoder as input. This generates an answer text with "[CLS]" as start token.

OCR-aware SINGULARITY.

We extend the original SINGULARITY model [24], Fig. 3.12 and propose a new **OCR-aware VideoQA** version that can read the text in the videos and thereby answer questions based on the text in the videos. To this end, we include the OCR tokens in the videos as additional input during pretraining and finetuning stages. At the time of pretraining, unlike the original model that uses image/video + caption pairs, we use image/video + OCR tokens pairs. Similar to the original model, the following three pretraining objectives are employed.

(i) Vision-OCR Contrastive loss: Adopted from the original SINGULARITY where the loss was Vision-Text contrastive, we adopt the same for OCR tokens in the frames. We obtain the OCR tokens of the frames at which questions were defined and consider this to be the ground truth video-text pair. This loss aims to align paired vision and language embeddings. Given the encoded vision embedding $F_v(f_{i,t})$, we use a projection head (with pooling) ϕ_v to project the embedding sequence into a vector representation $\phi_v(F_v(f_{i,t})) \in \mathbb{R}^D$. Here $F_{i,t}$ is the t-th frame in the i-th video in the training set, and t is randomly sampled from all available frames in this video. Notation of the j-th sentence is $\phi_l(F_l(S_j)) \in \mathbb{R}^D$.

The following equation is used to compute the similarity score $s_{i,j}$ of the video and text pair that is defined as their dot product:

$$s_{i,j} = \phi_v(F_v(f_i))^T \phi_l(F_l(S_j)) \tag{3.9}$$

Contrastive loss is applied to align paired vision-OCR embeddings:

$$p_i^v = \frac{\exp s_{i,j}/T}{\sum_j \exp s_{i,j}/T}, p_i^l = \frac{\exp s_{i,j}/T}{\sum_j \exp s_{j,i}/T}, L_{vtc} = -\sum_{i=1}^{n} (\log p_i^v + log p_i^l) \tag{3.10}$$

where T is a learned temperature parameter, and it is initialized as 0.07 following CLIP [37]. n is the total number of examples in the training set.

(ii) Masked Language Modeling: is a vision-conditioned masked language modeling loss. It aims to predict the masked text tokens from their (masked) textual context as well as the visual context. This loss is added at the last layer of the multi-modal encoder which follows the exact formulation present in BERT [9]. The only difference is that additional vision inputs are added and a higher mask ratio of 50% is used.

(iii) Vision-OCR Matching: similar to Vision-OCR Contrastive loss, this allows the models to improve the alignment between paired vision and OCR inputs by using the output of [CLS] token from the multi-modal encoder for binary classification. In essence, it says whether or not the input frame and OCR tokens pair match. Similar to the original model, we add a multi-modal decoder that has the same architecture as that of the multi-modal encoder. This decoder uses multi-modal encoder outputs as its cross-attention inputs. It decodes the answer with [CLS] as the start token.

3.4 Experiments

In this section, we explain evaluation metrics, and experimental settings and report the results. In all the experiments, we use the validation split of the dataset to save the best-performing checkpoints.

3.4.1 Evaluation Metrics

We use two evaluation metrics—Accuracy (Acc.) and Average Normalized Levenshtein Similarity (ANLS) [4]. Accuracy is the percentage of questions for which the predicted answer matches exactly with the target answer. The accuracy metric awards a zero score even when the prediction is a little different from the target answer.

ANLS is a Levenshtein Similarity-based metric that acts softly on minor answer mismatches that might stem from an error in recognizing text on the images (i.e., OCR errors). This metric has immunity to slightly wrong answers which might occur due to OCR errors. Since all the answers in our dataset are derived from text seen in the videos, we found ANLS to be a suitable metric for NewsVideoQA. More formally, ANLS can be defined as:

$$ANLS = \frac{1}{N}\sum_{i=0}^{N}(\max_{j} s(a_{ij}, o_{q_i})) \tag{3.11}$$

$$s(a_{ij}, o_{q_i}) = \begin{cases} 1 - NL(a_{ij},o_{q_i}) & \text{if NL}(a_{ij}, o_{q_i)} < T \\ 0 & \text{if NL}(a_{ij}, o_{q_i)} \geq T \end{cases} \tag{3.12}$$

where N is the total number of questions in the dataset, M is the total number ground-truth answers per question, a_{ij} are the ground truth answers where $i = 0, \ldots, N$, and $j = 0, \ldots, M$, and o_{q_i} is the

Question: What is the threat?

Ground Truth: omicron
BERT-QA: omicron
M4C: omicron threat
OCR-aware SINGULARITY: omicron

Question: What % are vehicles and transport contributions to pollution?

Ground Truth: 36%
BERT-QA: very poor
M4C: 40%
OCR-aware SINGULARITY: 17%

Figure 3.13 Qualitative results. for different baselines on the proposed task. Results for baselines are shown in green for the correct predictions and in red for the incorrect predictions.

network's answer for the ith question q_i. $NL(a_{ij}, o_{q_i})$ is the normalized Levenshtein distance between the strings a_{ij} and o_{qi}.

3.4.2 Experimental setup

We run a commercial OCR engine, Google OCR to obtain OCR tokens for the evenly sampled frames. We sample frames at 2 frames per second.

3.4.2.1 BERT-QA

In the case of NewsVideoQA, we use the OCR tokens of the sampled video frames as context for BERT-QA. We use the default OCR token ordering from the OCR system: top-left to bottom-right. To convert the NewsVideoQA dataset in SQuAD format, we find the first substring of the answer in the context, which is an approximation of the answer span as followed in [33]. We finetune the BERT QA checkpoint that is already pretrained and finetuned for QA on SQuAD dataset [39]. Specifically, we use the 'bert-large-uncased-whole-word-masking-finetuned-squad' checkpoint [10]. We train the BERT QA model starting from this checkpoint on the NewsVideoQA dataset for ten epochs with a batch size of 32 and a learning rate of $2e-05$.

3.4.2.2 M4C

For M4C, we use the official implementation along with default hyperparameters [45]. During training, we use a maximum of $24,000$ iterations. The model is trained using Adam optimizer, with a learning rate of $1e-4$ and a staircase learning rate schedule, where the learning rate is multiplied by 0.1

at $14,000$ and at $19,000$ iterations. The best model is saved based on the validation accuracy. The fixed vocabulary used for answer generation is $3,751$ words from answers in the train split of NewsVideoQA. Since M4C is a model for VQA on images, we train it using video frame + question pairs in the train split of NewsVideoQA. Similar to how BERT-QA was trained on NewsVideoQA, for each question the corresponding matching frame is found using the time-stamp information for each question that was collected during annotation. We call this as, M4C-Oracle.

3.4.2.3 SINGULARITY

We use the pretrained model of SINGULARITY and finetune it on NewsVideoQA. We finetune it for 20 epochs and all the hyperparameters and training settings are kept the same as in the official implementation. SINGULARITY uses a single frame while training, and 12 randomly sampled frames while testing.

3.4.2.4 OCR-aware SINGULARITY

The model is implemented in PyTorch [36]. Similar to how SINGULARITY was originally implemented, we initialize vision encoder using BEiT_BASE model which was previously trained on ImageNet-21K [8]. Using the NewsVideoQA dataset, we further pretrain the previously trained SINGULARITY (checkpoint: singularity_temporal_17m.pth). Original SINGULARITY uses a video/image-caption combination for pretraining. Instead, we combine the video with OCR tokens of the frames corresponding to the timestamp of the questions defined. Similar to original setting, in pretraining, a single frame is sampled from the whole video. We pretrain the model for 10 epochs. For the task of finetuning, we concatenate the question tokens with the OCR tokens of the frame on which the question is defined. An additional multi-modal decoder is initialized from pretrained multi-modal encoder (pretrained on NewsVideoQA dataset). It uses the multi-modal encoder outputs as cross-attention inputs and decodes the answer with start token as [CLS]. In the first half period, the learning rate is $1e-4$ with a warm up factor, followed by cosine decay to $1e-6$. We finetune the model for 20 epochs with a batch size of four. When training the model, a single frame is used, and when testing the model, 12 frames are used. At the time of inference, similar to [24], we use early fusion strategy, which takes all frames as model inputs (concatenation of all the frames considered at the test time) for directly making a more informative video-level prediction.

3.4.3 Results

In this section, we present results for multiple experiments performed on different baselines on the NewsVideoQA dataset.

In Table. 3.1, we show the results of heuristics and upper-bound baselines. 3.0% of the questions can be answered by predicting "yes" which is the most common answer in the train split. The Vocab Upper

Table 3.1 Heuristics and Upper bound baseline results. It can be seen that answers are substrings for more than 50 % of the serialized OCR tokens of a single frame corresponding to the timestamp of the question.

Heuristic Baselines	**Acc. (%)**
Majority answer	3.00
Biggest OCR token	1.03
Vocab Upper Bound	**76.58**
Substring single frame UB	53.05
Substring all frames UB	74.43

Table 3.2 Comparison of BERT-QA and M4C on single frame. Performance of the NewsVideoQA test set when trained and tested on single-frame information. It can be seen that both BERT-QA and M4C have decent performances but BERT-QA when given the correct information performs better.

Model	**#Frames for training**	**#Frames for testing**	**Type of frame at testing**	**Acc. (%)**	**ANLS**
BERT-QA [9]	1	1	random frame from the video	28.70	34.21
BERT-QA [9]	1	1	frame on which question was defined	46.55	56.81
M4C [16]	1	1	frame on which question was defined	28.49	32.17

Table 3.3 Performance of baselines on multi-frame setup. Quantitative results of BERT-QA, M4C, SINGULARITY, and OCR-aware SINGULARITY when trained on a single frame and tested in multi-frame setup. It can be seen that

Model	**#Frames for testing**	**Type of frame at testing**	**Acc. (%)**	**ANLS**
BERT-QA [9]	2	randomly sampled frames	15.03	17.65
BERT-QA [9]	2	frames on which question was defined	56.36	67.11
M4C [16]	2	frames on which question was defined	27.87	31.54
BERT-QA [9]	12	randomly sampled frames	53.86	65.27
M4C [16]	12	randomly sampled frames	30.68	34.90
SINGULARITY [24]	12	randomly sampled frames	4.82	5.78
OCR-aware SINGULARITY	12	randomly sampled frames (visual feat) + OCR tokens from a single frame	33.57	37.52
OCR-aware SINGULARITY	12	randomly sampled frames	32.47	35.56

Bound of 76.58% shows that many answers in the train split repeat in the test split as well. In Table. 3.2, we show the comparative results for BERT-QA and M4c when trained and tested on a single frame. It can be seen that when the OCR tokens of the frame on which the question was defined were given as the context to the model, BERT-QA performed better when compared to a random single frame. In Tab. 3.3, we show the performance comparison of all the baselines when trained in multi-frame setup. It can be seen that BERT-QA performs better when correct context is given assuming that we know the timestamp at which the question was defined. It can be seen that M4C (Text-based single image VQA model) performs well when they are tested on one frame and two frame settings (frames based on the timestamp of the question). The way these two frames are selected is: conside an example of question being defined at 4th second. Also, while we obtain the OCR tokens using GoogleOCR, we have sample 2 frames per second. Now we pick both the frames corresponding to 4th second and take respective OCR tokens. It can be seen that the performance of M4C reduces significantly when it is tested on 12 frames. For BERT-QA, the first row shows the performance when OCR tokens of a single random frame. This is followed by testing BERT-QA on OCR tokens of the frame on which the question was defined. From the table, it can be seen that non-VideoQA models perform poorly when the correct information required to answer the questions is not given as input to these models. SINGULARITY (without finetuning on NewsVideoQA) has poor performance compared to other baselines as the majority of the questions framed are based on textual content in the videos. It can be seen that after adding OCR tokens at the stage of pretraining and finetuning, the performance of the VideoQA model increases. We perform two experiments – a) when 12 frames are randomly sampled but the OCR tokens from the frame at which the question was defined were given and, b) when 12 random frames and OCR tokens of the random frames were given as an input to the model. It can be observed that the model learns to obtain the answer even when the OCR information from the correct frame was not given as input to the model. In Fig. 3.13, we show qualitative results from our experiments. The left example shows the predictions of baselines. As the frame contains less textual information all the baselines predict the correct answer. Whereas in the center and right example, the number of OCR instances increases thereby increasing the difficulty to obtain the correct answer.

3.5 Summary

To conclude, in this chapter, we first discuss existing datasets for video question answering and then look at their shortcomings. To fix the shortcomings, we propose a novel task of text-based video question answering and the NewsVideoQA dataset. For NewsVideoQA, we define and formulate the protocols followed in selecting and annotating the videos and question-answers Finally, we compute various statistics to highlight multiple aspects of the proposed NewsVideoQA dataset. Based on the experiments on mutliple baselines, it can be seen that, a text-only model outperforms both single image based scene-text VQA model and OCR-aware Video Question Answering model. In fact, there are two ways to look at this inference. The first, text-only model outperforming other methods which take

visual information as input, says that most of the questions require only textual information to answer the questions, and the second, VideoQA methods are unable to utilize the textual information along with textual information which leads to inferior performance. Also, the NewsVideoQA dataset contains videos from only multiple news channels but not other video categories. In the next chapter, we look at different baselines and their repurposed versions that exploit OCR information in the videos to answer questions.

Chapter 4

Beyond News Video: Exploring other video categories in text-based Video Question Answering

4.1 Introduction

In the preceding chapter, we looked into the different aspects of the text-based VideoQA dataset named as NewsVideoQA.

Figure 4.1 M4-ViteVQA Dataset Examples. Examples from M4-ViteVQA dataset. 'A' indicates the answers returned by TextVQA models and wrong answers are colored in red.

This dataset comprised a collection of question-answer pairs derived from news videos. Notably, news videos typically contain substantial textual information, which is not the case in many real-life scenarios. To understand the scenes comprehensively and address the challenges posed by multi-modal content in everyday scenes, we need to explore video categories that are more representative of common experiences. These categories encompass various aspects of daily life, such as shopping, vlogging, traveling, and more. While existing datasets like TextVQA [46], and ST-VQA [4] have made notable progress in the field of scene-text Visual Question Answering (VQA), they primarily concentrate on single, well-photographed images. However, for addressing the task of VQA within video categories that are more diverse and dynamic, a recent contribution, "M4-ViteVQA" [66], emerges in the research along the same timeline as that of NewsVideoQA.

The M4-ViteVQA dataset is specifically designed to serve as a comprehensive benchmark for Multi-category, Multi-resolution, and Multi-modal VQA. It spans across nine distinct video categories: shopping, traveling, driving, vlog, sport, advertisement, movie, game, and talking. This diverse set of categories ensures that the dataset is not limited to a narrow set of scenarios, but rather reflects the richness of everyday experiences that individuals encounter in their lives. This dataset contains videos from different resolutions such as 720p, 1080p, and 1176x664. Videos with rich text were selected manually from YouTube. The dataset also contains extra annotations such as "easy", and "hard" and labels to indicate the kind of information required to answer the question.

Table 4.1 Statistics of M4-ViteVQA dataset. The number of videos, frames, and questions in each category.

Category	# Videos	# Frames	# Questions
Shopping	847	155,275	3,892
Traveling	1,154	219,880	4,291
Driving	1,316	148,040	3,272
Vlog	947	168,715	2,897
Sport	665	133,979	2,072
Advertisement	623	113,108	1,264
Movie	719	103,429	1,449
Game	709	155,645	3,672
Talking	640	119,321	2,314
Total	7,620	1,317,392,	25,123

The benchmark defines 2 tasks with 3 settings. Task1 is the "regular QA task" and Task2 is the "domain adaption task". In Task1, QA pairs from all nine categories are present. Task1 is further divided into two data splits to check the robustness of the models. The first one is called *Task1Split1* which is divided according to the 7,620 cropped videos, the second one is called *Task1Split2* which is divided by the 1,150 raw videos. Task1Split2 is more challenging than Task1Split1 since the content of videos of the same category may be quite different (for example: various shopping venues and sports). In Task2, the model is supposed to be trained on seven categories while tested on the remaining two categories. Task2 requires the model to deal with unlearned content and completely different category-

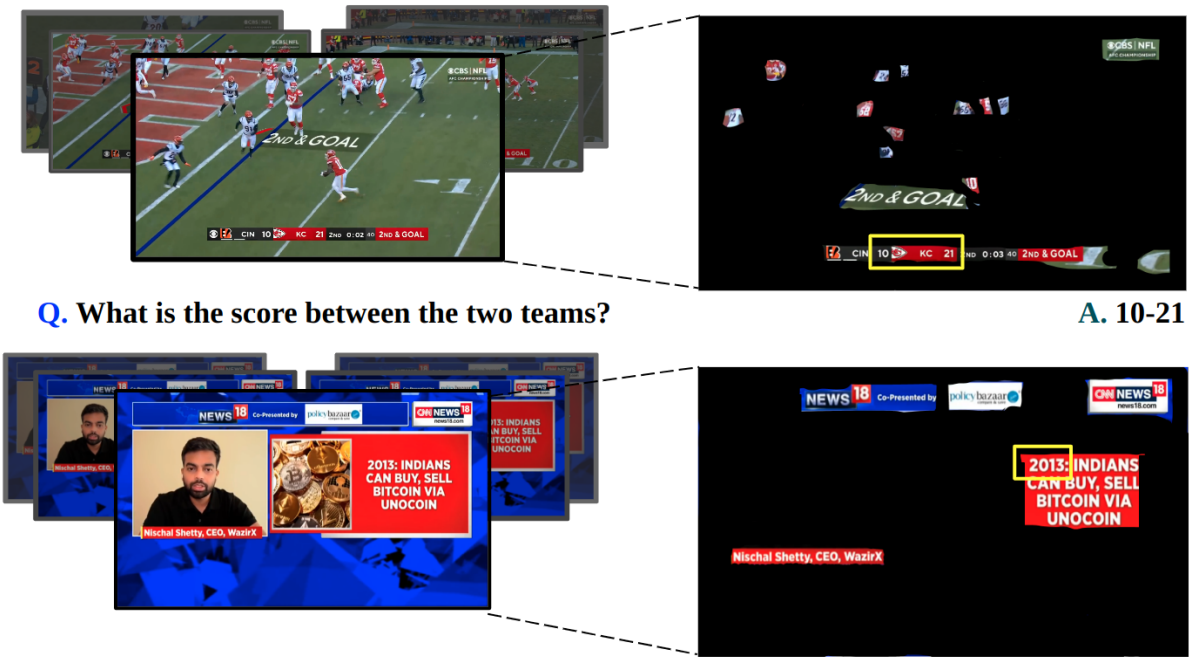


Figure 4.2 Limitations of Current Text-based VideoQA datasets – NewsVideoQA and M4-ViteVQA [17, 66]. Example illustrating two major concerns of existing text-based VideoQA datasets [17, 66]. Both examples showcase that only **textual information** from a **single frame** is sufficient to obtain answers to the questions.

specific questions, which is very challenging. Zhao et al. [66] also propose a baseline method called T5-ViteVQA. It consists of five transformers to conduct both textual and visual understanding as well as temporal reasoning over three modalities: texts from the video, a given question, and a video.

Recently, Hegde et al. [14] shed light that the question-answer pairs in the TextVQA datasets [4, 46] are more focused on the text in the image but given less importance to the visual features. As a result, these datasets tend to place more emphasis on the textual information present within the images while not giving adequate consideration to the underlying visual features. This text-centric bias can lead to the training of models that exhibit a skewed understanding of the overall visual context, which, in turn, results in the generation of biased answers.

To address this issue comprehensively, it is necessary that similar research is conducted within the task of text-based video question-answering datasets [17, 52, 66]. This would serve the dual purpose of shedding light on the limitations of existing datasets and emphasizing the need for datasets that have a better balance between textual and visual elements. This can be accomplished through a combination of human data exploratory analysis, which looks into the intrinsic bias and challenges within these datasets, and empirical testing using text-only methods that can support the observations in the ex-

ploratory analysis. By doing so, one can identify and propose numerous solutions to rectify any biases and shortcomings in the current datasets.

In this chapter, we look into different datasets of text-based video question-answering, where the task necessitates an understanding of both textual and visual components within video content. In the first stage, we first experiment by closely examining and understanding the above recently introduced datasets M4-ViteVQA [66], and NewsVideoQA [17], each offering unique and diverse categories for analysis. Firstly, we focus on conducting an in-depth exploratory analysis of these datasets. Through this analysis, we aim to resolve the extent of visual comprehension and multi-frame interpretation required to effectively answer the questions presented with both datasets. By exploring the nature of the questions and the visual context they contain, we gain insights into the specific challenges posed by these datasets. This analysis sets the stage to understand the inherent complexities and nuances associated with text-based VideoQA.

In the second stage, employ BERT-QA [9], a text-only model. This model relies solely on textual information to generate answers to questions, thus allowing us to measure the extent to which textual information alone can contribute to answering video-related questions. Notably, our experiments with BERT-QA reveal its effectiveness, as it attains results that are on par with the original multi-modal methods that consider both visual and textual cues. This finding emphasizes the significance of the textual information in the context of video-based question-answering.

Additionally, we explore the domain adaptation aspect by training models on one dataset and testing them on the other. Specifically, we train on M4-ViteVQA and evaluate on NewsVideoQA, and vice versa. This approach provides valuable insights into the challenges associated with cross-domain understanding, shedding light on the adaptability and generalization capabilities of models in the context of different categories.

Towards the end, we also provide additional research directions to build better datasets for video understanding for the videos containing text and provide future directions.

4.2 Benchmarking and Experiments

In this section, we present details of the exploratory analysis and the experiments we conduct. BERT-QA is a transformer-based encoder-only model pre-trained on a large corpus and further finetuned on SQuAD dataset [39] for question answering (Extractive QA). Extractive QA is the task of extracting a short snippet from the document/context on which the question is asked. The answer 'span' is determined by its start and end tokens. It is selected for its effective extractive QA performance, implementation ease, and finetuning, despite limitations like no answer generation or handling yes/no questions. Its ability in extracting answers from textual content makes it a suitable choice for tasks where answers are primarily found in the text of the video. To convert both M4-ViteVQA and NewsVideoQA datasets in SQuAD format, we find the first substring of the answer in the context, which is an approximation of the answer span as followed in [33].

Table 4.2 Data Exploratory Analsyis. Analysis of 100 random QA pairs from M4-ViteVQA and NewsVideoQA datasets.

Category	M4-ViteVQA (%)	NewsVideoQA (%)
Single Frame	92.0	95.0
Multi Frame	8.0	5.0
Visual Info	33.0	6.0
Textual Info	95.0	100.0
Frame crowded with text	18.0	64.0
Extractive-based	81.0	98.0
Reasoning-based	5.0	2.0
Knowledge-based	1.0	0.0

Table 4.3 Performance of BERT-QA on M4-ViteVQA Task1Split1. Performance comparison of BERT-QA model on M4-ViteVQA [66] dataset when the answer to questions is present in the concatenated list of OCR tokens from evenly sampled frames. This is for task1 split1.

Answer present in context	Finetuning	Acc.	ANLS	No. of QA pairs
No	×	9.03	17.05	1971
No	✓	21.96	32.18	1971
Yes	×	19.20	25.25	911
Yes	✓	47.42	55.14	911

4.2.1 Exploratory Analysis

For exploratory analysis, we randomly sample 100 QA pairs from both M4-ViteVQA and NewsVideoQA. For each QA pair, we check the following aspects: i) if the question can be answered by a single frame or needs multi-frame information, ii) if the question needs visual information and/or textual information to obtain the answer, iii) if the frame which is essential to obtain the answer, is crowded with text (approximately more than 15 OCR tokens). From Table. 4.2, it can be seen that for both datasets, information from a single frame is sufficient to obtain answers, which is counter-intuitive to the video question-answering task. From Table. 4.2, it can also be seen that most of the questions in both datasets need textual information to obtain answers. As M4-ViteVQA contains videos from multiple categories, it contains more questions of visual type compared to NewsVideoQA that contains only news videos. Since both datasets are designed for questions that require reading text to answer questions, this has resulted in minimal questions that require multi-modal information. We also check for the answer type: i) extractive, ii) reasoning based, and iii) knowledge-based, and combinations of each type. From Table. 4.2, it can be seen that most of the questions are extractive in nature and have fewer reasoning-based and knowledge-based questions. However, having more reasoning/knowledge-based questions is crucial, thereby creating the need for better methods beyond the scope of text-only models.

Table 4.4 Performance BERT-QA on M4-ViteVQA Task1Split2 Performance comparison of BERT-QA model on M4-ViteVQA [66] dataset when the answer to questions is present in the concatenated list of OCR tokens from evenly sampled frames. This is for task1 split2.

Answer present in context	Finetuning	Acc.	ANLS	No. of QA pairs
No	×	8.17	15.81	1321
No	✓	17.10	26.05	1321
Yes	×	20.30	27.19	532
Yes	✓	42.29	48.90	532

Table 4.5 Performance of BERT-QA on M4-ViteVQA Task2. Performance comparison of BERT-QA model on M4-ViteVQA [66] dataset when the answer to questions is present in the concatenated list of OCR tokens from evenly sampled frames. This is for task2.

Answer present in context	Finetuning	Acc.	ANLS	No. of QA pairs
No	×	10.89	18.41	762
No	✓	16.01	24.08	762
Yes	×	25.78	30.48	318
Yes	✓	38.05	43.82	318

4.2.2 BERT-QA experiments

M4-ViteVQA [66]: The M4-ViteVQA dataset consists of two tasks. The first task is divided into two splits and both splits contain evenly distributed question-answer pairs from all video categories in train-val-test sets. In the second task, the training set comprises videos from seven categories, while the question-answer pairs and videos in validation in test splits are exclusively sourced from the remaining two categories. Zhao et el. [66] also propose a multi-modal video question-answering method: T5-ViteVQA, that combines information from multiple modalities including OCR features, question features, and video features.

In our experiments on the BERT-QA model, we first sample frames at 1fps and order the OCR tokens of the frames to the default reading order based on the position of the top-left corner of the OCR token. We further concatenate the ordered OCR tokens which becomes the context of the BERT-QA model. After the training phase, we conduct two types of testing to evaluate the performance of the BERT-QA model. For the first type, we evaluate the model on the entire validation set without checking if the answer is present in the context. This experiment allows us to assess the model's overall ability to obtain answers. In the second type of testing, we specifically focus on questions that have answers in the context.

NewsVideoQA [17]: This dataset proposes questions on news videos. The dataset has timestamps for each question indicating the frame at which the question was defined. This work also proposes a re-purposed baseline: OCR-aware SINGULARITY, which was originally inspired by SINGULARITY [**?**]. OCR-aware SINGULARITY is a multi-modal transformer-based video question-answering model that

Table 4.6 Performance comparison of different baselines with BERT-QA on M4-ViteVQA. Performance comparison (Acc.) of M4C, T5-ViteVQA, and BERT-QA on the validation set of Task 1 Split 1.

Set	M4C [16]	T5-ViteVQA [66]	BERT-QA [9]
Easy	19.30	25.09	**25.49**
Hard	9.02	14.26	**16.34**
Text	17.26	23.08	**31.01**
Vision	18.36	**24.21**	18.82

combines information from OCR tokens, questions, and visual information from a randomly sampled frame.

In this work, we conduct two types of training on this dataset. In the first approach, we train the BERT-QA model using the OCR tokens of the single frame on which the question was defined (BERT-QA-SF: BERT-QA Single Frame). In the second approach, we concatenate the OCR tokens from frames sampled at 1fps which forms the context of the BERT-QA model. (BERT-QA-MF: Multi-frame). By conducting training in both single-frame (BERT-QA-SF) and multi-frame (BERT-QA-MF) setups, we aim to explore the impact of variations in the length of context on the performance of the BERT-QA model. These two training approaches provide insights into the model's ability to obtain answers based on either a specific frame or a broader contextual understanding derived from multiple frames.

4.2.3 Domain Adaptation Experiments

We conduct experiments to determine if the BERT-QA model can perform or generalize well with the out-of-domain context. This evaluation aims to determine if the model can provide accurate answers even in unfamiliar video categories and their corresponding contexts. To achieve this understanding, we perform several experiments. We check for the performance of the BERT-QA model trained on the **Source dataset** followed by testing on the **Target dataset**. We do this in two settings: i) without finetuning on the target dataset, and ii) with finetuning on the target dataset (Example: Train on NewsVideoQA and test on M4-ViteVQA in two settings i.e. with/without finetuning and vice-versa). By doing these, we try to examine the impact of domain shift and the importance of training the model on videos from diverse categories, where scene text serves as the textual content in one dataset that is M4-ViteVQA, as opposed to embedded text in NewsVideoQA. These experiments help us determine the model's ability to generalize and adapt to the specific categories of videos.

4.2.4 Evaluation Metrics and Experimental Setup

Frequently used in the majority of the works on scene-text based visual and video question answering, we use two evaluation metrics — Accuracy (Acc.) and Average Normalized Levenshtein Similarity (ANLS). Accuracy is the percentage of questions for which correct answers and predicted answers match exactly. Whereas ANLS is a similarity-based metric that acts softly on minor answer mismatches.

More details can be found in [4]. For all experiments, we train BERT-QA `bert-large-uncased-whole-word-masking-finetuned-squad` for 15 epochs with a batch-size of 16 on 4 GPUs with a learning rate of 2e-05.

4.3 Results

4.3.1 Quantitative Results

In this section, we present the results and analysis of different experiments. In Table. **??**, we show the results of the performance of T5-ViteVQA and BERT-QA on different tasks and splits on the validation set of the M4-ViteVQA dataset. It can be seen that a simple text-only model achieves comparable results and beats the scores of T5-ViteVQA for certain splits. The results indicate that we need more datasets that require information from multiple modalities and multiple frames which is a concerning limitation in the current datasets. It can be seen that the BERT-QA relies purely on the OCR output to infer and extract the answer. Therefore, if the OCR output is noisy or if the tokens are incorrectly ordered (errors in default reading order) the model might fail to find the right answer. However, since the ANLS metric acts softly on OCR errors, BERT-QA outperforms T5-ViteVQA on the ANLS metric.

In Table. 4.5, we show the performance of BERT-QA for the questions that contain answers in the context. We create this test set by checking if the answer is a substring of context. For each of the splits, nearly half of the original questions in the validation set have answers in the context. In Table. 4.6, we show the performance comparison—in terms of Accuracy—of two methods: i) M4C [16]: It uses a multi-modal transformer and an iterative answer prediction module. The model answers questions based on scene-text questions on a single image. ii) T5-ViteVQA: method proposed as a baseline in [66], with BERT-QA on the validation set of Task 1 Split 1. It can be seen that BERT-QA outperforms M4C and T5-ViteVQA on different sets. Here, the "sets" correspond to the type of questions which is provided with the dataset. These sets are: i) easy - answering requires information from a single frame, ii) hard - answering requires information from multiple frames, iii) text - answering requires only reading text, and iv) vision - answering requires both visual and textual information. Only for questions that require visual information, BERT-QA underperforms, yet still manages to obtain decent performance.

In Table. 4.7, we show results of the performance of different methods on the test set of NewsVideoQA [17] dataset. OCR-aware SINGULARITY is a model trained in a single-frame setup and is tested on a multi-frame setup (by combining visual and textual information from 12 frames - more details in [17]). This is followed by results of BERT-QA-SF i.e. trained on OCR context from a single frame and tested by picking a random frame. In the third row, we show the results of BERT-QA when tested with OCR tokens of the frame on which the question was defined (correct frame). In the fourth row, BERT-QA-MF: BERT-QA is trained and tested on a multi-frame setup. In Table. 4.10, we show the results of out-of-domain training performance on both [17, 66] datasets. It can be seen that testing a model initially trained on M4-ViteVQA (Source dataset) achieves decent performance on an out-of-domain

Table 4.7 Performance comparison of BERT-QA with OCR-aware SINGULARITY on NewsVideoQA dataset. We show the performance of OCR-aware SINGULARITY [17] and BERT-QA in different settings. BERT-QA-SF: single frame setup, BERT-QA-MF: multi-frame setup on NewsVideoQA. In the second column, we explain the type of testing. "12 random frames": considers visual and textual information from 12 random frames, "single random frame": OCR tokens of a random frame, "single correct frame": OCR tokens of the correct frame, "1 frame per second": OCR tokens of frames sampled at 1fps.

Baseline	Type of testing	Acc.	ANLS
OCR-aware SINGULARITY	12 random frames	32.47	35.56
BERT-QA-SF	single random frame	23.71	29.47
BERT-QA-SF	single correct frame	46.55	56.81
BERT-QA-MF	1 frame per second	**52.29**	**61.12**

Table 4.8 More experiments of BERT-QA on NewsVideoQA. In this table, we show the results of the performance of the BERT-QA model on the test set of the NewsVideoQA [17] dataset. For the random frame, we sample a frame randomly and consider its OCR tokens as context to the model.

Training data	Testing data	Ft	Acc.	ANLS
-	single random frame	×	16.78	22.47
single correct frame	single random frame	✓	23.71	29.47
-	single correct frame	×	33.29	43.43
single correct frame	single correct frame	✓	46.55	56.81
-	1fps-sampled-frame	×	31.31	40.60
single correct frame	1fps-sampled-frame	✓	51.25	62.67
1fps-sampled-frame	single random frame	✓	17.41	20.36
1fps-sampled-frame	single correct frame	✓	37.26	42.26
1fps-sampled-frame	1fps-sampled-frame	✓	52.29	61.12

Table 4.9 Domain Adaptation Experiments: Source dataset – M4-ViteVQA. Out-of-domain training performance for NewsVideoQA and M4-ViteVQA datasets. "Source dataset" corresponds to the dataset on which we train the model, and "Target dataset" corresponds to the dataset we test the model on, in this case is "NewsVideoQA" dataset.

Source dataset	Finetuning on target	Acc.	ANLS
NewsVideoQA	✓	52.29	61.12
M4-ViteVQA	×	40.39	51.86
M4-ViteVQA	✓	50.41	61.04

Table 4.10 Domain Adaptation Experiments: Source dataset – NewsVideoQA. Out-of-domain training performance for NewsVideoQA and M4-ViteVQA datasets. The "Source dataset" corresponds to the dataset on which we train the model, and the "Target dataset" corresponds to the dataset we test the model on, in this case, is "M4-ViteVQA" dataset.

Source dataset	Finetuning on target	Acc.	ANLS
M4-ViteVQA	✓	21.96	32.18
NewsVideoQA	×	7.86	12.68
NewsVideoQA	✓	22.17	31.95

NewsVideoQA (target dataset) and vice-versa. By further finetuning on the target dataset, the performance of the model increases. This indicates that the BERT-QA model can effectively generalize across domains through out-of-domain training.

4.3.2 Qualitative Results

In this section, we present qualitative analyses conducted on the two datasets which are NewsVideoQA [17] and M4-ViteVQA [66], to gain deeper insights. Fig. 4.4 showcases qualitative results obtained from the NewsVideoQA dataset. We compared the ground truth with the predictions made by the BERT-QA model before and after finetuning. The results demonstrate that finetuning helps and improves the model's ability to extract relevant answers related to the questions. Similarly for the M4-ViteVQA dataset, we show the qualitative results in Fig. 4.3. In Fig. 4.5, we present qualitative results from the NewsVideoQA dataset. We compare the predictions of the BERT-QA models using context from OCR tokens from randomly sampled frames and context from the frame on which the question was defined. The results indicate that text in the random frame is insufficient for the model to obtain accurate answers. However, when provided with OCR tokens from the frame where the question was defined, the model successfully obtains the correct answer. In Fig. 4.5, we show the results for the out-of-domain experiments.

Q. What is the result of her test?

Ground truth	: negative
BERT-QA without fine-tuning	: **tdy heh negative**
BERT-QA with fine-tuning	: **negative**

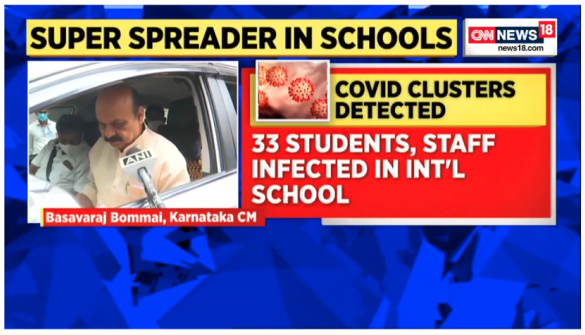
Q. Who is Bill Gross?

Ground truth	: pimco cofounder
BERT-QA without fine-tuning	: **worldwide exchange**
BERT-QA with fine-tuning	: **pimco cofounder**

Figure 4.3 Qualitative results from M4-ViteVQA dataset.

Q. How was delhi's air on nov 25?

Ground truth	: very poor
BERT-QA without fine-tuning	: **dirty**
BERT-QA with fine-tuning	: **very poor**



Q. In which school where 33 students and staff infected?

Ground truth	: int'l school
BERT-QA without fine-tuning	: **basavaraj bommai**
BERT-QA with fine-tuning	: **int'l school**

Figure 4.4 Qualitative results from NewsVideoQA. This figure shows the qualitative results of the NewsVideoQA [17] dataset. We show the ground truth, prediction without finetuning the BERT-QA model, and prediction after or with finetuning the BERT-QA model. It can be seen that finetuning helps the model to extract the answers relevant to the questions.

Figure 4.5 NewsVideoQA example with context from the correct frame and random frame. This figure shows the qualitative results of the NewsVideoQA [17] dataset. We show the ground truth, prediction of the BERT-QA model with context as the concatenated list of OCR tokens of the randomly sampled frame, and prediction of the BERT-QA model with context as the concatenated list of OCR tokens of the frame on which the question was defined. It can be seen that the text in the random frame is not sufficient for the model to obtain the answer. Whereas if we give OCR tokens of the frame where the question was defined, it obtains the correct answer.

4.4 Performance of Vision Language Models

Before the rise of general-purpose large language models and dedicated vision language models, previous works looked at specific task understanding such as multiple frame comprehension in VideoQA, scene-text understanding in ST-VQA and TextVQA, and knowledge-graph understanding in KVQA and so on. However, as the landscape of AI research evolved with the introduction of advanced models, it becomes increasingly important to evaluate their performance on established benchmarks. We can gain insights into their generalizability across diverse scenarios by subjecting models like LLaVA and Gemini to testing on these datasets. Moreover, such evaluations shed light on potential weaknesses or failure points, offering valuable feedback for refining these models. In this section, we delve into the performance metrics of two prominent large vision-language models: Gemini [1] and LLaVA [30].

LLaVA [30]. The model trained as a standard causal language model, taking language instructions (a user text prompt) as input, and returns a language response. The ability of the language model to handle images is allowed by a separate vision encoder model that converts images into language tokens, which are added to the user text prompt (acting as a soft prompt). LLaVA's language model is based on Vicuna [7] and the vision encoder is based on CLIP [37]. Vicuna is a pretrained large language model based on LLaMa-2 [53]. CLIP is an image encoder, pretrained to encode images and text in a similar embedding space using contrastive language-image pretraining. In our experiments, we test the performance of LLaVA on both NewsVideoQA and M4-ViteVQA datasets. We test this model in two

Table 4.11 Performance of Vision-Language Models on Text-based VideoQA datasets. We experiment with two Vision LLMs: Gemini-1.5 vision and LLaVA on both the M4-ViteVQA and NewsVideoQA datasets. Here # frames at testing shows the number of frames used at the time of testing to answer a question.

Dataset	Model	# frames at testing	Substring Match Accuracy (in %)
NewsVideoQA	LLaVA	1	22.7032
NewsVideoQA	LLaVA	5	28.4054
M4-ViteVQA	LLaVA	1	16.9964
M4-ViteVQA	LLaVA	5	28.4627
NewsVideoQA	Gemini	1	61.4649
M4-ViteVQA	Gemini	1	37.3032

setting, (i) When a single frame is given as an input to the model along with the question, (ii) When 5 evenly sampled frames are given as an input to the model along with the question. For NewsVideoQA dataset, as we know the frame at which the question was defined, we sample that frame as an input to the method (shown in first row of Tab. 4.11). For M4-ViteVQA single frame setting, we randomly sample a frame from the video (shown in third row of Tab. 4.11).

Gemini [1]. Gemini is a multimodal method, that can generalize and understand, operate across, and combine different types of information, including text, code, audio, image, and video. Gemini obtains state-of-the-art performance on multiple tasks. We obtain the free API key and use 'gemini-pro-vision' in order to obtain the answer to the questions for the given frame. Since the API call consumes time, we experiment with only single-frame QA for both datasets.

In the preceding sections, we employed two metrics to assess model performance: accuracy, which measures the exact match between predicted and actual answers, and ANLS, designed to account for subtle errors such as those stemming from OCR inaccuracies. However, these metrics fall short of evaluating the output of large vision-language models effectively. These models, trained on human chat data, generate open-ended, free-form responses, posing a challenge for zero-shot evaluation using metrics used previously. We attempted using prompts like "Please answer the questions briefly, i.e. 3-4 word answers." This led to shorter responses, though they were still at the sentence level, unlike the ground truth. To address this, we adopt a substring matching approach to evaluate the generated answers. Specifically, we examine whether the ground truth answer is a substring of the predicted answer, acknowledging that the predicted response may constitute an entire sentence rather than a single-word answer. Below is the equation to calculate substring match for single-frame setup:

$$\text{Substring Match} = \begin{cases} 1 & \text{if ground truth is a substring of predicted answer} \\ 0 & \text{otherwise} \end{cases}$$

Question: where were the omicron cases detected?
Ground Truth: botswana, hong kong & israel
LLaVA: the omicron cases were detected in hong kong and Israel.
Gemini: botswana, hong kong and israel

Question: which college reports 281 positive cases?
Ground Truth: k'taka college
LLaVA: ktaka college reports 281 positive cases.

Gemini: k'taka college

Figure 4.6 Qualitative results of Vision Language Models. Performance of Gemini and LLaVA on the NewsVideoQA dataset when a single frame on which the question was defined is given as input along with the question.

In the case of a multiple-frame setup, we combine all the generated answers into a single string and verify whether the ground truth answer is contained within this concatenated sequence. Below is the equation to calculate substring match for multi-frame setup:

$$\text{Substring Match (Multiple Frames)} = \begin{cases} 1 & \text{if ground truth is present} \\ & \text{in concatenated string of all generated answers} \\ 0 & \text{otherwise} \end{cases}$$

In Table. 4.11, we present the performance comparison of LLaVA and Gemini on both the NewsVideoQA and M4-ViteVQA datasets. The column ”Frames at Testing” indicates the number of frames utilized during testing to answer each question.

In the case of NewVideoQA, where the timestamp of the question is known, a single frame corresponding to that timestamp is selected. Additionally, we evenly sample five frames for multi-frame testing. For the M4-ViteVQA dataset, since the frame at which the question was defined is unavailable for single-frame testing, a frame is randomly sampled from the video. Similarly, for multi-frame testing, five frames are evenly sampled.

LLaVA's performance is evaluated in both single-frame and multi-frame setups on both datasets. It's notable that LLaVA achieves respectable performance in the single-frame setup without explicitly using OCR information but rather implicitly leveraging CLIP image encoding. Conversely, Gemini's

performance is evaluated only in the single-frame setup due to its time-intensive API usage. Despite this limitation, Gemini demonstrates commendable performance on both datasets with just a single frame.

In Fig. 4.6, we show the predictions of LLaVA and Gemini on the NewsVideoQA dataset when a single frame on which the question is defined was considered as input. It can be seen that LLaVA is able to generate partial answers in both examples, whereas Gemini generates correct answers. The potential reason for this could be that the LLaVA model being utilized has 7 billion parameters, which is significantly smaller in comparison to Gemini. In Fig. 4.9, we show the predictions of LLaVA and Gemini on the M4-ViteVQA dataset, where we sample a random frame as input. In the first example, LLaVA may have struggled due to the extensive text, leading to incorrect answers, while Gemini managed to provide accurate responses. In the second example, both LLaVA and Gemini mistakenly inserted the first name of the person. In Fig. 4.7 and Fig. 4.8, we show the predictions of LLaVA in multi-frame setup. We uniformly sample five frames from the video and pass each of these frames individually along with question as input. The predictions in blue suggest that for a few frames, the model can generate the expected answer if the correct information needed to obtain the answer is present. Otherwise, as seen in the red-colored predictions, it either generates it is not able to determine the answer or sometimes (in Fig. 4.8 [d] and [e]) it generates the wrong answer. Hallucination represents a significant challenge in evaluating these methods, constituting one of their major known drawbacks.

4.5 Future work

Based on the previous chapters, their conclusions, and insights, there are several potential future directions and areas of work that can be explored to advance the field of text-based video question answering. Non-extractive Answers: As noted, current text-based VideoQA datasets primarily focus on extractive answers, where the answer is a direct part of the video's text. Questions can be more open-ended, seeking interpretations, summaries, or explanations rather than direct text extraction. Future work can aim to develop datasets and models that prioritize generative or non-extractive answers but still would require reasoning over textual content present in the videos. Multi-frame Comprehension: To enhance the degree of visual understanding and multi-frame comprehension in VideoQA, future datasets could be designed to include questions that require reasoning across multiple frames or time segments of a video. Though was the original objective in mind but generating such questions became difficult in a trivial setup. Having better instructions for the annotators and actually designing scenarios where multi-frame understanding is needed might help. Using videos that contain both action and textual information might be a good place to start. Multi-modal Questions: Building on the idea of multi-modal learning, future datasets could incorporate questions that involve text, visual as well as audio information. These multi-modal questions might require models to combine cues from numerous modalities to provide accurate answers. This approach would better reflect the complexity of real-world information processing. Because current datasets have limitations, building methods that are truly multi-modal would prevent us from obtaining SOTA. Hence, have multi-modal questions that boost better multi-

modal VideoQA methods. Data augmentation: Currently, the process with which datasets are generated is manual, which becomes one of the hurdles in scaling up the dataset. Hence, building automated or semi-automated frameworks that can generate data would help both in scalable data generation and would be a step towards eliminating personal bias while framing questions. The emergence of models such as ChatGPT, LLaVA, Gemini, and various other multimodal Language and Vision Models (LLMs) signifies a shift towards exploring diverse video genres and complex content. Additionally, avenues like fine-grained video understanding and zero-shot learning present promising directions for further investigation. Moreover, leveraging these LLMs to generate richer and more abundant datasets warrants consideration, offering opportunities for enhanced model training and performance.

4.6 Summary

In this chapter, we focused on the important task of understanding textual information within videos for question-answering. The study provides insights that current text-based VideoQA datasets majorly focus on extractive answers and insights that the degree of visual understanding and multi-frame comprehension in current datasets is limited for better VideoQA using text in videos. Additionally, the paper demonstrates the effectiveness of BERT-QA, a text-only model, in achieving comparable performance to original methods on both datasets and also looks into the domain transfer aspect, by comparing the performances by training on one type of dataset and testing on the other. In future developments, we hope to see datasets that prioritize non-extractive answers and incorporate multi-modal questions based on multiple frames to facilitate improved multi-modal learning.

[a] [b] [c] [d] [e]

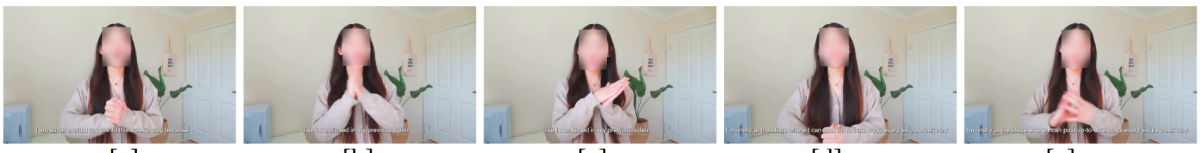
Question: what are the two precautions to be taken?
Ground Truth: wear masks, keep distance

Figure 4.7 Qualitative results of Vision Language Models. Performance of LLaVA on the NewsVideoQA dataset when we uniformly sample five frames. The following are the predictions of LLaVA for each of the considered frames: **[a] LLaVA.** In the image, there are two precautions to be taken. First, the man is wearing a face mask, which is essential for protecting himself and others from the spread of airborne viruses or bacteria. Second, the man is sitting in a chair with a pink wall, which suggests that he is in a public or shared space. In such environments, it is crucial to maintain a safe distance from others, practice good hygiene, and follow any guidelines or recommendations provided by local authorities or healthcare professionals to minimize the risk of infection or transmission. **[b] LLaVA.** In the image, there are two precautions to be taken. First, the man is wearing a mask, which is essential to protect himself and others from the spread of airborne viruses or bacteria. Second, the man is being tested for a disease, which is crucial to identify any potential health issues and receive appropriate treatment. These precautions highlight the importance of personal hygiene and healthcare in preventing the spread of diseases and ensuring public health. **[c] LLaVA.** In the image, a man is being tested for the COVID-19 virus. Two precautions to be taken are to maintain proper hygiene and to follow the guidelines provided by health authorities. These precautions include washing hands frequently, using hand sanitizer, wearing a mask, and practicing social distancing. Additionally, it is essential to stay informed about the latest developments and recommendations related to the virus and follow the guidelines provided by health authorities to minimize the risk of spreading the virus. **[d] LLaVA.** In the image, a man is being tested for COVID-19. The two precautions to be taken are to maintain social distancing and to wear a mask while waiting for the test results. These measures help reduce the risk of spreading the virus and protect the individual and others around them. **[e] LLaVA.** In the image, a man is being tested for COVID-19, and another man is wearing a hazmat suit. The two precautions to be taken are:1. The man being tested should follow proper hygiene practices, such as washing hands, using hand sanitizer, and wearing a face mask, to minimize the risk of spreading the virus. 2. The man wearing the hazmat suit should follow strict infection control measures, such as wearing a face mask, gloves, and other personal protective equipment (PPE), to ensure his safety and prevent the spread of the virus.

Question: who died second?
Ground Truth: Perfecto
LLaVA: Niko
Gemini: Perfecto died second

Question: what's the name of the man?
Ground Truth: Powell
LLaVA: Paul Powell
Gemini: Jerome Powell

Figure 4.8 Qualitative results of Vision Language Models. Performance of Gemini and LLaVA on the M4-ViteVQA dataset when a single frame on which the question was defined is given as input along with the question.



[a] [b] [c] [d] [e]

Question: How frequent is she going to post vlogs?
Ground Truth: every single week

Figure 4.9 Qualitative results of LLaVA on M4-ViteVQA dataset. Performance of LLaVA on the M4-ViteVQA dataset when we uniformly sample five frames. The following are the predictions of LLaVA for each of the considered frames: **[a] LLaVA.** Based on the image, the woman is going to post vlogs every week. She is holding a cell phone and a pink pen, which suggests that she is preparing to record a video or write a caption for her vlog. The fact that she is wearing a gray shirt and has long hair also indicates that she is likely to be consistent in her vlogging activities. **[b] LLaVA.** Based on the image, it is not possible to determine the exact frequency of the woman's vlogs. However, she is seen in the image with a vlog camera, which suggests that she might be planning to post vlogs regularly. The woman's actions, such as praying and making a video, indicate that she is engaged in creating content for her vlog. **[c] LLaVA.** Based on the image, it is not possible to determine the exact frequency of the woman's vlogs. However, she is seen holding her hands up, which might suggest that she is about to post a vlog. The woman's actions in the image indicate that she is likely to post vlogs occasionally or when she has something to share. **[d] LLaVA.** Based on the image, the woman is going to post vlogs every single day. **[e] LLaVA.** Based on the image, the woman is going to post vlogs every single day.

Chapter 5

Conclusion

In this thesis, we proposed and thoroughly investigated text-based video question-answering tasks and datasets. We looked at the literature, identified the gaps, proposed new tasks, and repurposed baselines to fill those gaps and provided limitations and future work. Here is a summary of the thesis.

In Chapter 2, we look at existing datasets and methods in the field of multi-modal reasoning, starting from Visual Question Answering, Scene-text-based VQA, and Video Question Answering. This literature review helps us understand the gaps in the existing works and gives us a strong motivation to propose a new task.

In Chapter 3, we propose a new task and an associated dataset. This task aims at obtaining answers to questions that require reading text in the videos. To this end, we propose a new dataset NewsVideoQA that enables research in this direction. We benchmark the dataset with multiple repurposed baselines.

In Chapter 4, we explore different video categories by understanding and experimenting with the M4-ViteVQA dataset. We perform data exploratory analysis and identify key limitations such as heavy bias in a single frame and text-only information to obtain answers to questions in both datasets. We also experiment with a text-only model achieving comparable performance as that of the original multi-modal methods on both datasets. We also show domain adaptation results on both datasets. Furthermore, based on the inferences obtained using extensive ablations, we present directions toward future work that might help us create better datasets. We also look into performance of current vision language models on NewsVideoQA and M4-ViteVQA datasets.

To conclude, in this thesis, we highlight the importance of text in videos and propose new tasks and datasets to over multiple shortcomings of existing works. We also present ways in which this direction can be improved to build better assistive systems. We hope this work will motivate future text-based video understanding research.

Related Publications

- **Watching the News: Towards VideoQA Models that can Read**, *Soumya Shamarao Jahagirdar*, Minesh Mathew, Dimosthenis Karatzas, C.V. Jawahar. *In IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2023*.
- **Understanding Video Scenes through Text: Insights from Text-based Video Question Answering**, *Soumya Shamarao Jahagirdar*, Minesh Mathew, Dimosthenis Karatzas, C.V. Jawahar. *In IEEE/CVF International Conference on Computer Vision, Workshops (ICCVW: Vision-and-Language Algorithmic Reasoning) 2023*

Bibliography

[1] R. Anil, S. Borgeaud, Y. Wu, J. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, D. Silver, S. Petrov, M. Johnson, I. Antonoglou, J. Schrittwieser, A. Glaese, J. Chen, E. Pitler, T. P. Lillicrap, A. Lazaridou, O. Firat, J. Molloy, M. Isard, P. R. Barham, T. Hennigan, B. Lee, F. Viola, M. Reynolds, Y. Xu, R. Doherty, E. Collins, C. Meyer, E. Rutherford, E. Moreira, K. Ayoub, M. Goel, G. Tucker, E. Piqueras, M. Krikun, I. Barr, N. Savinov, I. Danihelka, B. Roelofs, A. White, A. Andreassen, T. von Glehn, L. Yagati, M. Kazemi, L. Gonzalez, M. Khalman, J. Sygnowski, and et al. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805, 2023.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: visual question answering. *CoRR*, abs/1505.00468, 2015.

[3] L. J. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016.

[4] A. F. Biten, R. Tito, A. Mafla, L. G. i Bigorda, M. Rusiñol, C. V. Jawahar, E. Valveny, and D. Karatzas. Scene text visual question answering. In *ICCV*, pages 4290–4300. IEEE, 2019.

[5] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguistics*, 5:135–146, 2017.

[6] Y. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. UNITER: universal image-text representation learning. In *ECCV (30)*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer, 2020.

[7] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[8] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE Computer Society, 2009.

[9] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[10] H. Face. Huggingface models. `https://huggingface.co/models`. Accessed on 27 August 2022.

[11] D. Gao, K. Li, R. Wang, S. Shan, and X. Chen. Multi-modal graph neural network for joint reasoning on vision and scene text. In *CVPR*, pages 12743–12753. Computer Vision Foundation / IEEE, 2020.

[12] P. Gupta and M. Gupta. Newskvqa: Knowledge-aware news video question answering. In *PAKDD (3)*, volume 13282 of *Lecture Notes in Computer Science*, pages 3–15. Springer, 2022.

[13] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, pages 3608–3617. Computer Vision Foundation / IEEE Computer Society, 2018.

[14] S. Hegde, S. Jahagirdar, and S. Gangisetty. Making the V in text-vqa matter. In *CVPR Workshops*, pages 5580–5588. IEEE, 2023.

[15] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. C. Russell. Localizing moments in video with natural language. In *ICCV*, pages 5804–5813. IEEE Computer Society, 2017.

[16] R. Hu, A. Singh, T. Darrell, and M. Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *CVPR*, pages 9989–9999. Computer Vision Foundation / IEEE, 2020.

[17] S. Jahagirdar, M. Mathew, D. Karatzas, and C. V. Jawahar. Watching the news: Towards videoqa models that can read. In *WACV*, pages 4430–4439. IEEE, 2023.

[18] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. K. Ghosh, A. D. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, and E. Valveny. ICDAR 2015 competition on robust reading. In *ICDAR*, pages 1156–1160. IEEE Computer Society, 2015.

[19] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. Almazán, and L. de las Heras. ICDAR 2013 robust reading competition. In *ICDAR*, pages 1484–1493. IEEE Computer Society, 2013.

[20] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, A. Veit, S. Belongie, V. Gomes, A. Gupta, C. Sun, G. Chechik, D. Cai, Z. Feng, D. Narayanan, and K. Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017.

[21] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. Dense-captioning events in videos. In *ICCV*, pages 706–715. IEEE Computer Society, 2017.

[22] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017.

[23] J. V. Landeghem, R. Tito, L. Borchmann, M. Pietruszka, P. Józiak, R. Powalski, D. Jurkiewicz, M. Coustaty, B. Anckaert, E. Valveny, M. B. Blaschko, S. Moens, and T. Stanislawek. Document understanding dataset and evaluation (DUDE). *CoRR*, abs/2305.08455, 2023.

[24] J. Lei, T. L. Berg, and M. Bansal. Revealing single frame bias for video-and-language learning. In *ACL (1)*, pages 487–507. Association for Computational Linguistics, 2023.

[25] J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, and J. Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, pages 7331–7341. Computer Vision Foundation / IEEE, 2021.

[26] J. Lei, L. Yu, M. Bansal, and T. L. Berg. TVQA: localized, compositional video question answering. In *EMNLP*, pages 1369–1379. Association for Computational Linguistics, 2018.

[27] J. Li, D. Li, C. Xiong, and S. C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR, 2022.

[28] L. Li, Y. Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu. HERO: hierarchical encoder for video+language omni-representation pre-training. In *EMNLP (1)*, pages 2046–2065. Association for Computational Linguistics, 2020.

[29] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV (30)*, volume 12375 of *Lecture Notes in Computer Science*, pages 121–137. Springer, 2020.

[30] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[31] T. Maharaj, N. Ballas, A. Rohrbach, A. C. Courville, and C. J. Pal. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *CVPR*, pages 7359–7368. IEEE Computer Society, 2017.

[32] M. Mathew, V. Bagal, R. Tito, D. Karatzas, E. Valveny, and C. V. Jawahar. Infographicvqa. In *WACV*, pages 2582–2591. IEEE, 2022.

[33] M. Mathew, D. Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2199–2208, 2021.

[34] A. Mishra, K. Alahari, and C. V. Jawahar. Image retrieval using textual cues. In *ICCV*, pages 3040–3047. IEEE Computer Society, 2013.

[35] A. Mishra, S. Shekhar, A. K. Singh, and A. Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, 2019.

[36] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035, 2019.

[37] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[38] S. Raja, A. Mondal, and C. V. Jawahar. ICDAR 2023 competition on visual question answering on business document images. In *ICDAR (2)*, volume 14188 of *Lecture Notes in Computer Science*, pages 454–470. Springer, 2023.

[39] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*, pages 2383–2392. The Association for Computational Linguistics, 2016.

[40] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788. IEEE Computer Society, 2016.

[41] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[42] S. Shah, A. Mishra, N. Yadati, and P. P. Talukdar. KVQA: knowledge-aware visual question answering. In *AAAI*, pages 8876–8884. AAAI Press, 2019.

[43] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *CVPR*, pages 4613–4621. IEEE Computer Society, 2016.

[44] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[45] A. Singh, V. Goswami, V. Natarajan, Y. Jiang, X. Chen, M. Shah, M. Rohrbach, D. Batra, and D. Parikh. Mmf: A multimodal framework for vision and language research. `https://github.com/facebookresearch/mmf`, 2020.

[46] A. Singh, V. Natarjan, M. Shah, Y. Jiang, X. Chen, D. Parikh, and M. Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019.

[47] H. Tan and M. Bansal. LXMERT: learning cross-modality encoder representations from transformers. In *EMNLP/IJCNLP (1)*, pages 5099–5110. Association for Computational Linguistics, 2019.

[48] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, pages 4631–4640. IEEE Computer Society, 2016.

[49] S. Team. Silero models: pre-trained enterprise-grade stt / tts models and benchmarks. `https://github.com/snakers4/silero-models`, 2021.

[50] R. Tito, D. Karatzas, and E. Valveny. Document collection visual question answering. In *ICDAR (2)*, volume 12822 of *Lecture Notes in Computer Science*, pages 778–792. Springer, 2021.

[51] R. Tito, D. Karatzas, and E. Valveny. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognit.*, 144:109834, 2023.

[52] G. Tom, M. Mathew, S. Garcia-Bordils, D. Karatzas, and C. V. Jawahar. Reading between the lanes: Text videoqa on the road. In *ICDAR (6)*, volume 14192 of *Lecture Notes in Computer Science*, pages 137–154. Springer, 2023.

[53] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton-Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang,

R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023.

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[55] A. Veit, T. Matera, L. Neumann, J. Matas, and S. J. Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *CoRR*, abs/1601.07140, 2016.

[56] X. Wang, Y. Liu, C. Shen, C. C. Ng, C. Luo, L. Jin, C. S. Chan, A. van den Hengel, and L. Wang. On the general value of evidence, and bilingual scene-text visual question answering. In *CVPR*, pages 10123–10132. Computer Vision Foundation / IEEE, 2020.

[57] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.

[58] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang. Video question answering via gradually refined attention over appearance and motion. MM '17, page 1645–1653, New York, NY, USA, 2017. Association for Computing Machinery.

[59] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017.

[60] J. Xu, T. Mei, T. Yao, and Y. Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296. IEEE Computer Society, 2016.

[61] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, pages 1666–1677. IEEE, 2021.

[62] H. Yang, L. Chaisorn, Y. Zhao, S.-Y. Neo, and T.-S. Chua. Videoqa: Question answering on news video. pages 632–641, 01 2003.

[63] Z. Yang, Y. Lu, J. Wang, X. Yin, D. Florencio, L. Wang, C. Zhang, L. Zhang, and J. Luo. Tap: Text-aware pre-training for text-vqa and text-caption. In *CVPR*, 2021.

[64] Y. Yu, J. Kim, and G. Kim. A joint sequence fusion model for video question answering and retrieval. In *ECCV (7)*, volume 11211 of *Lecture Notes in Computer Science*, pages 487–503. Springer, 2018.

[65] Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, and D. Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, pages 9127–9134, 2019.

[66] M. Zhao, B. Li, J. Wang, W. Li, W. Zhou, L. Zhang, S. Xuyang, Z. Yu, X. Yu, G. Li, A. Dai, and S. Zhou. Towards video text visual question answering: Benchmark and baseline. In *NeurIPS*, 2022.

[67] L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann. Uncovering the temporal context for video question answering. *Int. J. Comput. Vis.*, 124(3), 2017.