Coarse to Fine Frame Selection for Online Open-ended Video Question Answering

Sai Vidyaranya Nuthalapati*
vidyaranya.ns@gmail.com

Anirudh Tunga*
Blue Wave AI Labs
anirudh.tunga@bwailabs.com

Abstract

The central aim of Video Question Answering (VideoQA) is to provide answers to questions posed in natural language, relying on the content of the given videos. However, when applied to video streams like CCTV recordings and live broadcasts, the solver encounters more intricate challenges. In such scenarios, the segment of the video needed to answer a specific question is often a small component of the entire video. To address these complexities, a recent and innovative problem domain called Online Open-ended Video Question Answering (O^2VQA) has been introduced[18].

In this paper, we propose an architecture based on multimodal foundational transformers for the O^2VQA task. The architecture comprises three modules. The first module is responsible for the coarse selection of the target video segment relevant to answering the question. The second module refines this coarse segment by leveraging a Temporal Concept Spotting mechanism, enabling the capture of temporal saliency and resulting in the identification of frames most critical for addressing the question. Lastly, we employ an end-to-end Video-Language Pre-training model to provide the answer. To evaluate our proposed model, we conduct experiments on the publicly available ATBS dataset[18]. The results showcase the superiority of our approach over current state-of-the-art models.

1. Introduction

In recent years, there has been a remarkable surge in research focused on enhancing the understanding of multimodal models [5, 3, 13] focussed on vision and language. Among these areas of study, Video Question Answering (VideoQA) stands out as a prominent field due to its potential to enable interactive AI systems that communicate with the dynamic visual world using natural language.

Video Question Answering (VideoQA) involves a model that receives a sequence of frames and corresponding natural language questions as input and produces answers to those questions. The model needs to handle multimodal inputs and comprehend various relationships in the data, which include recognizing subject interactions, enumerating diverse objects, and discerning cause-and-effect relationships among actions depicted in the video.

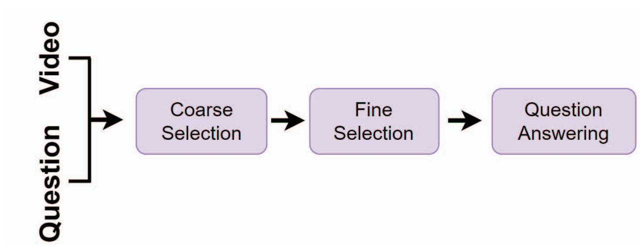

Figure 1. Overview of the proposed architecture for O^2VQA. We use three modules, namely Coarse frame selection, Fine frame selection, and Video Question Answering module

Video Question Answering (VideoQA) has gained significant popularity in recent studies [19, 40]; however, it remains a challenge due to its requirement for models to possess a comprehensive understanding of videos to provide accurate responses to questions. In comparison to Image Question Answering, commonly known as Visual Question Answering, VideoQA presents notably more complex hurdles. Primarily, the vast number of frames contained within videos often includes irrelevant information not pertinent to the specific question at hand. Moreover, the questions in VideoQA go beyond mere recognition of visual elements such as objects, actions, activities, and events; they demand the ability to infer intricate semantic, spatial, temporal, and causal relationships among these elements [37, 14]. This multifaceted nature of VideoQA tasks significantly augments the complexity involved, thereby making it a compelling and continuously evolving domain of research [45].

Existing research in VideoQA has primarily focused on short, fixed-length videos [35, 11, 37, 23]. However, in practical real-life scenarios, videos obtained from recordings, live streams, and CCTV footage are often much longer

*equal contribution

and vary in duration. This presents a significant challenge since the relevant part of the video needed to answer a question constitutes only a small portion compared to the entirety of the video.

To bridge this gap and make VideoQA systems more applicable to real-world situations, a novel task called Online Open-ended Video Question Answering (O^2VQA) was introduced by the authors of CEO-VQA paper [18]. In this task, the VideoQA model is presented with a video of unpredictable length and a related question. The objective is to enable the model to autonomously identify the relevant/target section of the video, collect enough information from it, and then terminate to answer the question based on that specific section. This new task aims to address the challenges posed by long and variable-length videos, facilitating better adaptation of VideoQA systems to practical applications.

Prevailing methods for O^2VQA [18] adopt a two-step approach whereby they first predict the target video segment and subsequently employ a question-answering module on the chosen segment. Nonetheless, this approach encounters two significant challenges. Firstly, any inaccuracies in the target segment selection module reverberate into the question-answering module. For example, if the segment selection module selects an excessive number of frames beyond the actual ground truth, it introduces substantial noise into the subsequent question-answering process. Secondly, existing approaches for the question-answering module employ distinct text and video frames encoders [8, 4], subsequently fused using a transformer to predict the answer [32]. Unfortunately, this not only escalates the model's parameter count but has also demonstrated inferior performance compared to using a unified backbone capable of concurrently encoding all modalities [33, 1]. Thus, based on the work of Wang et al. [33], we adopt a pre-trained multi-modal foundational model featuring a unified backbone for effectively encoding both video and text within our question-answering module.

To address the O^2VQA challenge, we present the Coarse to Fine Frame Selection – Video Question Answering (CoFFS-VideoQA) model, comprising three essential modules:

(a) a coarse frame selection module that identifies the target segment in the video,

(b) a fine frame selection module to extract frames relevant to the given question, and

(c) an end-to-end Video-Language pre-trained question-answering (VideoQA) module.

In the coarse frame selection module, we leverage frame and question embeddings to gauge their similarity. Employing a Fibonacci sampling technique, we sample frames and utilize similarity thresholds to locate the boundaries of the target segment in the video. Subsequently, the Fine frame selection module uniformly samples frames from the identified segment and computes saliency scores for each frame, aiding in the identification of crucial frames essential for answering the question. This step effectively filters out noisy frames, enhancing the overall performance of the VideoQA module.

The VideoQA component employs an end-to-end model that learns both video and language representations from raw data. Following the pre-train and fine-tune approach used in recent works [33], we adapt a model pre-trained on video-text matching and masked language modeling to suit the O^2VQA task.

By combining these three modules, our CoFFS-VideoQA model presents a holistic approach to address the challenges of the O^2VQA task, as depicted in Figure 1. The model excels at selecting pertinent frames and delivering precise question-answering capabilities.

Our proposed method is tested on the publicly available ATBS [18] dataset and we show an improvement over the current state-of-the-art (SOTA) model.

2. Related Work

Traditional methods for VideoQA used frame-level and clip-level information using various techniques - graph neural networks [35], cross-modal attention [14, 23], attribute-based attention [41], hierarchical attention [44, 43], multi-step progressive attention memory [17], and multi-head attention [22]. With the success of Transformer architecture for vision and language tasks, pre-training on large-scale datasets with video-text pairs, and fine-tuning for downstream tasks, like VideoQA, has become the norm [37, 33, 46]. Pre-training has shown improved performance on downstream tasks. Pre-train and then fine-tune approach has been used in recent works on Visual Storytelling [42], and Text-to-Video retrieval [38]. In most of these works, the pre-training is done on video-text pairs and then finetuned on video-language tasks. In this work, following [33], we use a model pre-trained on video-language tasks, and then fine-tune it to VideoQA.

CLIP has demonstrated remarkable effectiveness in handling image-text tasks. Recently, researchers have also directed their attention towards employing CLIP for video-text retrieval tasks [10, 26] and video recognition [16, 24, 27, 28]. In the context of video recognition, some studies [24, 28] adopt CLIP as a reliable initialization for the vision encoder, while others [16, 27] utilize it to model video-label interactions.

Although these approaches partially capture the video-label interactions, a more recent work [34] highlights the potential of CLIP in modeling bidirectional cross-modal interactions. Drawing inspiration from the success of CLIP in modeling such interactions [34], we leverage CLIP in our Fine frame selection module to identify the salient frames.

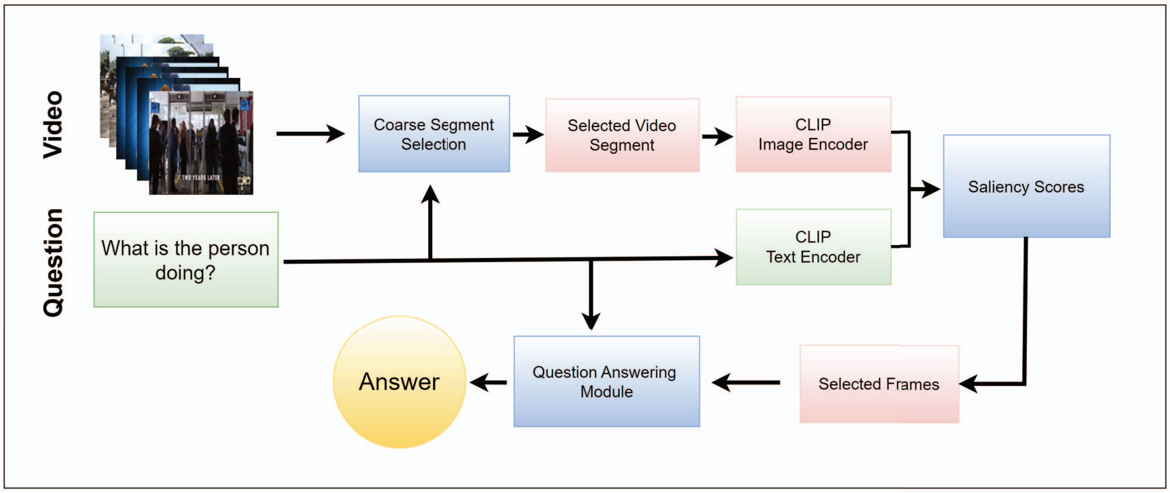

Figure 2. The proposed CoFFS-VideoQA module takes the question and video as input. Coarse frame selection predicts the target segment of the video, effectively removing the background component of the video. This is passed on to the Fine frame selection module which selects the top K frames which are required to answer the question. Finally, a Video question answering module takes the frames and the question as input and outputs an answer.

3. Proposed architecture

3.1. Problem Formulation

The task of O^{2}VQA consists of a video V and a corresponding textual question Q. The goal is to predict the correct answer for the given question. In open-ended Video Question Answering, responses are initially given in free-form natural language. However, a prevalent approach involves transforming the task into a classification problem by representing the answers using class labels [20, 33]. Following this, we model this as a classification task.

3.2. Overview

Precisely predicting the response in the O^{2}VQA problem necessitates the model's ability to identify the segment of the video relevant to the posed question. Additionally, a robust answering component is essential to predict the answer by considering both the identified video segment and the question. Keeping this in mind, our proposed CoFFS-VideoQA model for O^{2}VQA has three main components: Coarse frame selection, Fine frame selection, and VideoQA module. We expound further on each of the modules in the following sections

3.2.1 Coarse frame selection

To distinguish the relevant part of the video from the background video, we employ Coarse frame selection. To demonstrate the superior performance of the Fine frame selection and VideoQA modules, we build upon the Coarse frame selection module introduced by Kong et al. [18], which is also the current best-performing model for O^{2}VQA task.

In this module, we leverage the video and text encoders used in BridgeFormer [12]. The pre-training regime of Bridgeformer ensures that the encoders are capable of extracting fine-grained semantic associations across the text and video modalities. Specfically, we use a 12-layer vision transformer [9] to encode the video frames and DistilBERT [31, 32] to encode the text. The *cls* tokens from both encoders are utilized to compute the similarity between the video frames and the text. During video processing, when the similarity surpasses a specified threshold c_{max} at any point in the video, we consider that frame as the central frame of interest. Subsequently, we sample frames on both sides of the video using the Fibonacci sequence and calculate the corresponding similarities. Sampling ceases as soon as the similarity falls below another threshold c_{min}. The first and last frames selected through this process constitute the target segment, which serves as the input for the Fine frame selection module.

3.3. Fine Frame Selection

In this module, we employ uniform frame sampling from the target video segment to identify the most crucial frames (among the sampled frames) required to answer the question. The rationale behind uniform sampling is due to the observation that there is not much variation between a few consecutive frames, making it reasonable to treat them as representative samples. This process involves two steps: After predicting the start and end frames of the target segment, we compute a saliency score for each frame within

Method	What	Who	How	Where	When	Overall
ALPRO [21]	19.16	33.58	77.27	25.0	54.55	24.8
CEO-VQA [18]	19.20	41.68	**79.55**	25.0	54.55	27.1
CoFFS-VideoQA (Ours)	**21.26**	**43.54**	**79.55**	**28.57**	**60.00**	**29.2**

Table 1. Comparison of our proposed model for O^2VQA task with other models on ATBS [18] dataset. Best accuracies are in bold.

that segment. Based on these saliency scores, we select the top K frames with the highest scores. To achieve this, we draw inspiration from the approach presented in [34] and leverage a multi-modal foundational model to calculate the target segment-to-sentence level attention at a granular level.

Specifically, we establish word-frame attention for each word-frame pair by generating frame-level embeddings using CLIP [30]. Each frame within the target segment undergoes processing through the Vision Transformer [9], which constitutes the CLIP model, resulting in frame embeddings denoted as $\{\mathbf{v}_t \in \mathbb{R}^d \mid t = 1, 2, \cdots, T\}$, where $\mathbf{v}_t$ represents the embedding of the t-th frame, and T represents the total number of frames in the target segment.

In the context of the CLIP model, the question is encoded using the Transformer network [32]. Each word in the input question is represented by the output of the last self-attention block, forming the word representation $\{\mathbf{w}_p \in \mathbb{R}^d \mid p = 1, 2, \cdots, W\}$, where W is the total number of words in the question Q.

To determine the relevance of each frame to the question, we begin by computing the similarity between a word in the question and a frame. This similarity measure is then normalized using softmax across all frames, resulting in the word-frame similarity. This process yields the normalized similarity of a word with each frame.

Subsequently, for a given frame, we calculate the average of all the normalized word similarities corresponding to that frame, which forms the saliency score S_t for the particular frame.

$$S_{tp} = \frac{\exp\left(\mathbf{v}_t^\top \mathbf{w}_p / \tau\right)}{\sum_{t=1}^{T} \exp\left(\mathbf{v}_t^\top \mathbf{w}_p / \tau\right)}, \tag{1}$$

where S_{tp} is the similarity between the frame t and word p. τ is the temperature of the softmax function. v_t and w_p are the embeddings of t-th frame and p-th word respectively.

The saliency score for every frame is calculated by taking the mean of all the normalized word-frame similarities with respect to the current frame as follows:

$$\mathcal{S}t = \frac{1}{W} \sum_{p=1}^{W} S_{tp}, \tag{2}$$

where W is the number of words in the question and S_{tp} is normalized frame-word similarity.

Once the saliency scores are computed for all frames, we proceed to select the top K frames based on their saliency scores. These chosen frames are then utilized in the question-answering module, collectively forming a new video that serves as the input to the VideoQA module.

3.4. Video Question Answering Module

The third and final module in our architecture is Video Question Answering (VideoQA) module which predicts the answer given the most important frames. Unlike most existing works [36, 15] for VideoQA, which use separate encoders for text and video and later use complex fusion modules, we use a single backbone network for encoding both the video and the text.

With the recent surge in the use of multi-modal foundational models [7, 6] for many video-related tasks, we explore the usage of pre-trained models for VideoQA. We use the slightly modified and pre-trained ViT [9] model provided by All-in-one (AIO) [33]. Specifically, ViT has been modified to add support to text by using a learned word embedding to encode the question. On the other hand, to encode the input video, it is split into patches and passed through a fully-connected layer. Learnable position and modality type embeddings are added to each of the video and text embeddings.

Finally, the video and the text embeddings are concatenated and passed through multiple modified transformer layers, where each layer has a temporal token rolling module [33], multi-head self-attention layer [32] and fully-connected layer.

The token rolling module is mainly to model the temporal information in the ViT network. Token rolling has been shown to be effective in modeling the long-range dependencies between videos and text. We refer the reader to the original paper [33] for further details.

The *cls* token of the final ViT layer is passed through a two-layer fully connected (FC) layer to predict the answer label. The output of the final FC layer is passed through a softmax. The question-answering module is trained using cross-entropy loss.

3.5. Implementation details

The VideoQA module has been implemented using PyTorch [29]. We use Adam optimizer with weight decay [25] with an initial learning rate of 1e-4 and a weight decay of 0.01. The model is trained with a polynomial learning rate

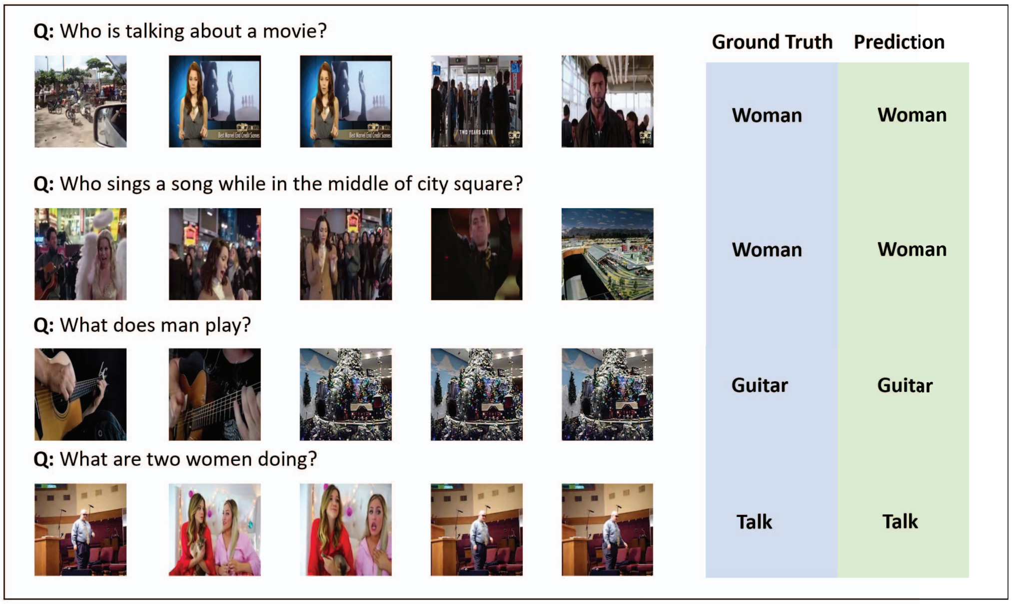

Figure 3. Examples from the ATBS [18] dataset where CoFFS-VideoQA predicts the answer correctly.

decay, with a warmup period of 2500 steps during which it is linearly increased from 0 to the initial learning rate of 1e-4. The decay power is set to 1.

For the 'Fine frame selection' (FFS) component, we use ViT-B/16 [34, 9] model pre-trained by CLIP [30]. Specifically, we use an input resolution of 224*224 for the frames. The vision transformer used for encoding the frames contains 12 layers, 12 heads and has a width of 768. Similarly, for encoding the text, we use a transformer with 8 heads, 12 layers, and with a width of 512. We note that this part of the architecture is not trained. Finally, we set the number of frames selected after FFS to be 7 i.e. K is set to 7.

4. Experiments

Within this section, an overview of the utilized dataset is presented, along with the corresponding results.

4.1. Dataset

For the evaluation of the O^2VQA task we use the Answer Target in Background Stream (ATBS) dataset [18]. In order to replicate real-world scenarios, Kong et al. adopt a *Background* + *Target* approach. This involves using a relatively long background video, simulating an online video stream, and a target short video clip containing essential information to answer the provided question. The target video clip is inserted into the background video, after choosing a random frame from the background video as the insertion point. This is to simulate the natural and random appearance of the target event within a dynamic video stream. The background videos are taken from the Distinct Describable Moments (DiDeMO) [2] dataset. The target video clips are extracted from the MSRVTT dataset [39].

Every video clip from the MSRVTT dataset [39] is paired with a unique background video that is randomly selected from the DiDeMo dataset [2]. Prior to further processing, both the frames of the target video clip and the background video undergo resizing to a uniform size of 224 × 224 pixels.

In total, there are 10k videos in the dataset and we follow previous work [18] to generate the train/val/test splits for a fair comparison.

4.2. Results

We evaluate the performance of our proposed model on the ATBS dataset to demonstrate the improved performance of our model compared to the current SOTA models. We compare the top-1% accuracy of our model with other models. We see that the proposed model outperforms the current SOTA models on overall accuracy and also on almost all the question types establishing the superiority of our model.

4.3. Qualitative analysis

In this section, we present examples from the dataset demonstrating instances where the model performs accurately and instances where it makes mistakes.

Correct predictions: Figure 3 displays multiple videos alongside their corresponding questions and answers, both predicted by the model and the ground truth. Notably, the model exhibits a good performance in predicting answers by effectively filtering out irrelevant frames.

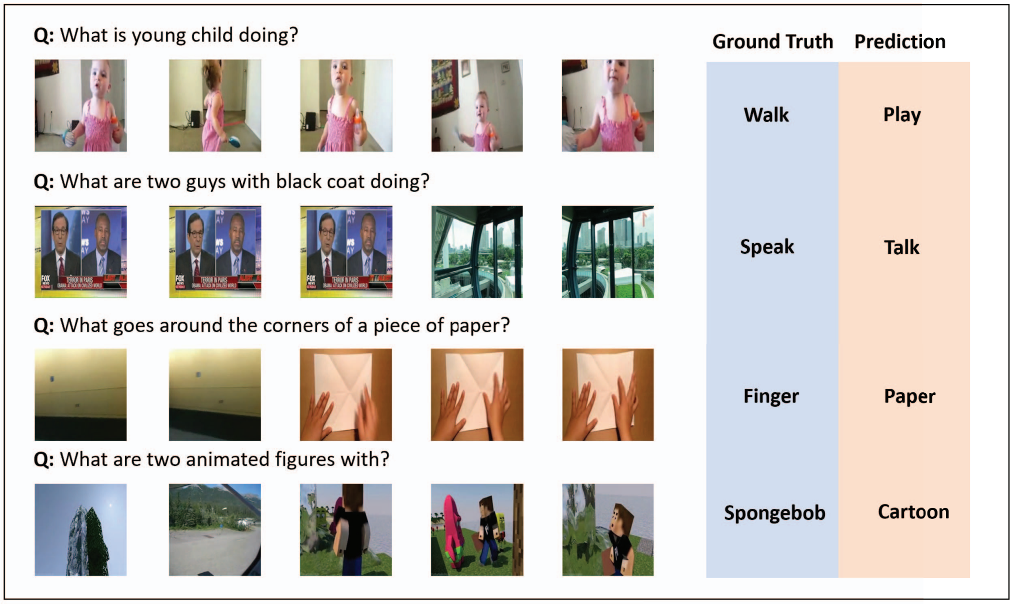

Figure 4. Examples from the ATBS [18] dataset where CoFFS-VideoQA predicts the answer incorrectly.

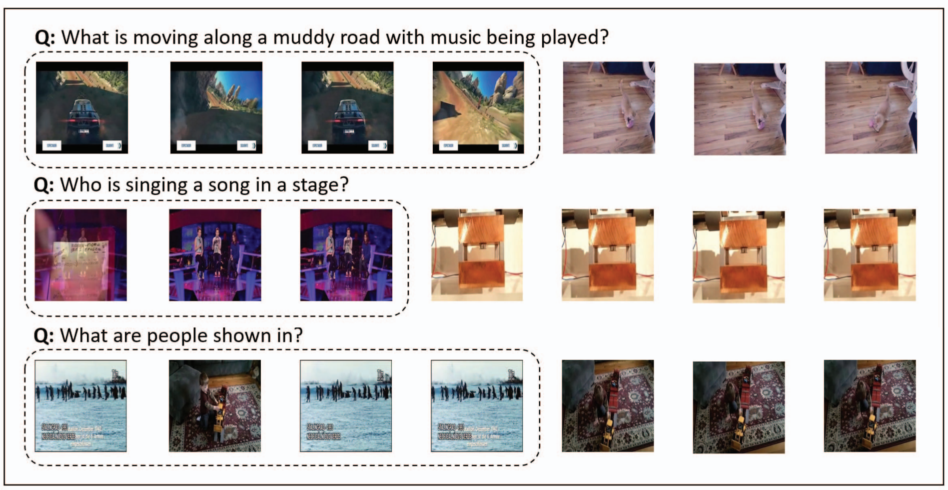

Figure 5. Inputs and outputs of the Fine frame selection (FFS) module: Each row corresponds to a video. A few sampled frames (which act as inputs to FFS) for each video are shown in each row. The dotted lines show the frames selected by the FFS module.

Incorrect predictions: Figure 4 presents a collection of videos where the model encounters challenges in accurately predicting the answers. These challenges fall into three distinct categories:

a) Ambiguous answers: Within this category, instances arise where predicting the subject's activity in the video becomes intricate. The first case in the figure portrays this uncertainty, as it remains difficult to deduce whether the child is merely walking or playing.

b) Synonyms as answers: Notably, the model tends to generate responses that are synonymous with the ground truth. The second example within the figure exemplifies this occurrence, where the model predicts the answer as 'talk,' whereas the correct response is 'speak.'

c) Requires external knowledge: Certain questions necessitate knowledge beyond the model's training data. In the fourth example showcased in the image, the model's inability to predict 'Spongebob's' appearance arises from lacking any prior exposure to the character.

By categorizing these challenges, we gain insights into the limitations of the model's predictive capabilities in various contexts, prompting further investigation and potential improvements.

Fine frame selection (FFS): Figure 5 displays both the frames comprising the target segment, as predicted by the coarse segmentation module and the frames selected by the FFS module, which are deemed the most crucial ones for answering the question. Notably, the module adeptly identifies and picks frames of utmost relevance to the question, underscoring the significance of incorporating this module in ensuring robust and accurate question-answering capabilities.

5. Conclusion and Future Work

This work addresses the problem of using VideoQA for real-world use cases. Previous works concerning this problem extracted the target segment followed by a question-answering module which has a few drawbacks. Firstly, errors occurring in the target segment selection module have a cascading effect on the question-answering module. For instance, the extra number of frames picked by the segment selection module introduces substantial noise into the question-answering process. Secondly, regarding the VideoQA module, current approaches solely rely on separate text and video frame encoders to encode each modality independently. Later, a transformer is employed to fuse these encodings and predict the answer. While on one hand, this approach increases the number of model parameters, on the other hand, it performs less effectively compared to utilizing a unified backbone capable of simultaneously encoding all modalities[33]. To tackle these issues, we adopt a two-pronged approach. Firstly, we implement a Fine frame selection module to filter frames, allowing only relevant ones to be forwarded to the VideoQA module. Secondly, we leverage a unified backbone architecture to efficiently address the question and provide answers. We validated our model on the publicly available ATBS [18] dataset to show the efficacy of the model. In the future, we aim to integrate all three modules into a unified architecture and train them end-to-end.

References

[1] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221, 2021. 2

[2] L. Anne Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 5

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1

[4] G. Bertasius, H. Wang, and L. Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 2

[5] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5659–5667, 2017. 1

[6] S. Chen, X. He, L. Guo, X. Zhu, W. Wang, J. Tang, and J. Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*, 2023. 4

[7] S. Chen, H. Li, Q. Wang, Z. Zhao, M. Sun, X. Zhu, and J. Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *arXiv preprint arXiv:2305.18500*, 2023. 4

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 4, 5

[10] B. Fang, C. Liu, Y. Zhou, M. Yang, Y. Song, F. Li, W. Wang, X. Ji, W. Ouyang, et al. Uatvr: Uncertainty-adaptive text-video retrieval. *arXiv preprint arXiv:2301.06309*, 2023. 2

[11] J. Gao, R. Ge, K. Chen, and R. Nevatia. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6576–6585, 2018. 1

[12] Y. Ge, Y. Ge, X. Liu, D. Li, Y. Shan, X. Qie, and P. Luo. Bridging video-text retrieval with multiple choice questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16167–16176, 2022. 3

[13] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1

[14] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017. 1, 2

[15] J. Jiang, Z. Chen, H. Lin, X. Zhao, and Y. Gao. Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11101–11108, 2020. 4

[16] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, pages 105–124. Springer, 2022. 2

[17] J. Kim, M. Ma, K. Kim, S. Kim, and C. D. Yoo. Progressive attention memory network for movie story question answering. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition*, pages 8337–8346, 2019. 2

[18] W. Kong, S. Ye, C. Yao, and J. Ren. Confidence-based event-centric online video question answering on a newly constructed atbs dataset. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1, 2, 3, 4, 5, 6, 7

[19] T. M. Le, V. Le, S. Venkatesh, and T. Tran. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9972–9981, 2020. 1

[20] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7331–7341, 2021. 3

[21] D. Li, J. Li, H. Li, J. C. Niebles, and S. C. Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4953–4963, 2022. 4

[22] X. Li, L. Gao, X. Wang, W. Liu, X. Xu, H. T. Shen, and J. Song. Learnable aggregating net with diversity learning for video question answering. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1166–1174, 2019. 2

[23] X. Li, J. Song, L. Gao, X. Liu, W. Huang, X. He, and C. Gan. Beyond rnns: Positional self-attention with co-attention for video question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8658–8665, 2019. 1, 2

[24] Z. Lin, S. Geng, R. Zhang, P. Gao, G. de Melo, X. Wang, J. Dai, Y. Qiao, and H. Li. Frozen clip models are efficient video learners. In *European Conference on Computer Vision*, pages 388–404. Springer, 2022. 2

[25] I. Loshchilov and F. Hutter. Fixing weight decay regularization in adam. 2018. 4

[26] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 2

[27] B. Ni, H. Peng, M. Chen, S. Zhang, G. Meng, J. Fu, S. Xiang, and H. Ling. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, pages 1–18. Springer, 2022. 2

[28] J. Pan, Z. Lin, X. Zhu, J. Shao, and H. Li. St-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35:26462–26477, 2022. 2

[29] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 4

[30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4, 5

[31] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 3

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3, 4

[33] J. Wang, Y. Ge, R. Yan, Y. Ge, K. Q. Lin, S. Tsutsui, X. Lin, G. Cai, J. Wu, Y. Shan, et al. All in one: Exploring unified video-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6598–6608, 2023. 2, 3, 4, 7

[34] W. Wu, X. Wang, H. Luo, J. Wang, Y. Yang, and W. Ouyang. Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 4, 5

[35] J. Xiao, P. Zhou, T.-S. Chua, and S. Yan. Video graph transformer for video question answering. In *European Conference on Computer Vision*, pages 39–58. Springer, 2022. 1, 2

[36] J. Xiao, P. Zhou, A. Yao, Y. Li, R. Hong, S. Yan, and T.-S. Chua. Contrastive video question answering via video graph transformer. *arXiv preprint arXiv:2302.13668*, 2023. 4

[37] 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 *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 1, 2

[38] J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 2

[39] J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 5

[40] Z. Yang, N. Garcia, C. Chu, M. Otani, Y. Nakashima, and H. Takemura. Bert representations for video question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1556–1565, 2020. 1

[41] Y. Ye, Z. Zhao, Y. Li, L. Chen, J. Xiao, and Y. Zhuang. Video question answering via attribute-augmented attention network learning. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 829–832, 2017. 2

[42] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021. 2

[43] Z. Zhao, Q. Yang, D. Cai, X. He, Y. Zhuang, Z. Zhao, Q. Yang, D. Cai, X. He, and Y. Zhuang. Video question answering via hierarchical spatio-temporal attention networks. In *IJCAI*, volume 2, page 8, 2017. 2

[44] Z. Zhao, Z. Zhang, X. Jiang, and D. Cai. Multi-turn video question answering via hierarchical attention context reinforced networks. *IEEE Transactions on Image Processing*, 28(8):3860–3872, 2019. 2

[45] Y. Zhong, J. Xiao, W. Ji, Y. Li, W. Deng, and T.-S. Chua. Video question answering: Datasets, algorithms and challenges. *arXiv preprint arXiv:2203.01225*, 2022. 1

[46] L. Zhu and Y. Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755, 2020. 2