QCaption: Video Captioning and Q&A through Fusion of Large Multimodal Models

Jiale Wang[1], Gee Wah Ng[1], Lee Onn Mak[1], Randall Cher[2], Ng Ding Hei Ryan[2], Davis Wang[2]

[1]Q Team, Home Team Science and Technology Agency, Singapore
{wang_jiale, ng_gee_wah, mak_lee_onn}@htx.gov.sg
[2]Department of Computer Science and Engineering, Nanyang Technological University, Singapore
{rand0017, rng039, dwang023}@e.ntu.edu.sg

Abstract—This paper introduces QCaption, a novel video captioning and Q&A pipeline that enhances video analytics by fusing three models: key frame extraction, a Large Multimodal Model (LMM) for image-text analysis, and a Large Language Model (LLM) for text analysis. This approach enables integrated analysis of text, images, and video, achieving performance improvements over existing video captioning and Q&A models; all while remaining fully self-contained, adept for on-premises deployment. Experimental results using QCaption demonstrated up to 44.2% and 48.9% improvements in video captioning and Q&A tasks, respectively. Ablation studies were also performed to assess the role of LLM on the fusion on the results. Moreover, the paper proposes and evaluates additional video captioning approaches, benchmarking them against QCaption and existing methodologies. QCaption demonstrate the potential of adopting a model fusion approach in advancing video analytics.

Index Terms—Large Multimodal Model, Large Language Model, Video Analytics, Model Fusion

I. INTRODUCTION

A. Background

Large Large Language Models (LLM) such as GPT-3 [1], PaLM [2], and LLaMA 2 [3], have captured the world's attention for its remarkable proficiency in understanding and generating human-like text, transforming how we interact with and generate information.

Recent advancements [4] in integrating images and text within a unified latent space have catalysed the development of Large Multimodal Models (LMM) such as GPT-4(Vision) [5], Gemini [6], and LLaVA [7]. These models possess the capability to interpret images in addition to text, marking a significant leap beyond purely linguistic models, moving AI closer to emulating human intelligence that perceives the world through both visual and textual dimensions. These opened up a whole new dimension of use cases, from image captioning and search, to asking questions about an image.

Despite their multi-modality, most LMMs are primarily designed for interpreting images and text [5]–[7], lacking the capability to directly process videos. Video analytics, a crucial research domain with extensive applications, holds particular significance for applications like that in homeland security, where critical evidence often exists in video formats such as CCTV footage, body-cam recordings, and phone video recordings of incidents. Developing a versatile AI model and pipeline capable of captioning, analysing, and executing downstream tasks like report writing for videos — akin to current LMMs' functions with images — could significantly unlock new automation possibilities, significantly boosting productivity in domains such as homeland security, indexing multimedia databases for search [8] or building video recommendation engines.

B. Key contributions

In this paper, we present QCaption, a comprehensive video captioning and Q&A pipeline that surpasses existing LMM approaches for video analysis, yielding 44.2% and 48.9% improvements in video captioning and Q&A tasks, respectively. QCaption operates entirely on-premises without the need for external APIs. By conceptualising video captioning as a task of multi-image captioning and caption aggregation, QCaption fuses three models — a key frame extraction model, an LMM for combined image and text interpretation, and an LLM dedicated to text analysis — into a video analytics pipeline. This approach allows for the simultaneous analysis of text, images, and video content. Our system builds on the works of LLaVA [7] and Vicuna [9], enhancing their capabilities for video content. We demonstrate that QCaption outperforms existing video LMMs, including Video-LLaVA [10], Video-ChatGPT [11], and Video-LLaMA [12], for video captioning and Q&A.

Our key contributions include:

- The development of QCaption, a novel video captioning pipeline that enables the integrated analysis of text, images, and videos within a single pipeline, unlike previous LMMs that are limited to just image-text or image-video analysis only. Its modular design also allows the frame extraction model, LMM, and LLM to be swapped out as needed.
- Through our experiments, we showcase that fusing three models, each adept at different tasks, provides an effective method for video analysis, showing improvements of up to 44.2% and 48.9% in video captioning and Q&A tasks, respectively, over our baselines.

- We conduct ablation studies to evaluate the impact of each component within the fusion pipeline across various tasks.
- We also proposed and evaluated other video captioning methods, by adapting different key frame sampling approaches and by fusing LLaVA LMM and Vicuna LLM in various ways, and evaluated how they perform against baselines and QCaption.

II. RELATED WORKS

A. Pre-Large Large Language Models (LLM) era

Traditional Video Captioning and Q&A models employ Convolutional Neural Networks (CNN) to extract visual features from the video, then returning them as standalone classes or incorporating them into predefined sentence templates to generate a caption or answer a question [8], [13], [14]. Later methods also employ CNNs as the encoder and Recurrent Neural Networks (RNN) for the decoding phase (i.e., language generation) [15], [16]. However, the range of requests that can be accepted is highly limited because the model cannot comprehend complex natural language prompts, with use cases confined to just captioning and simple feature based Q&A.

B. LLM

LLMs, such as GPT-3 [1], have transformed NLP by enabling AI to comprehend and generate complex text. Utilising the transformer architecture, which employs self-attention mechanisms to prioritise words in a sentence to generate contextually relevant text [17]. Their effectiveness stems from extensive pre-training on large datasets and subsequent fine-tuning for specific applications, such as instruction-tuning for chatbots apt at emulating conversations between humans. Examples include LlaMA 2 [3], Vicuna [9], and Mistral [18]. These LLMs also form the backbone of LMMs, setting them apart from earlier Computer Vision and Video Captioning models.

C. Image Large Multimodal Models (LMMs)

LMMs build upon LLMs to integrate image processing capabilities with advanced language understanding. There are two main architecture paradigms [10]: using LLMs as schedulers, coordinating visual and text models for task-specific applications without end-to-end training, such as Hugging-GPT [19]; and LLMs decoders, aligning image and text data through a two-stage process involving initial auto-regressive pretraining and subsequent refinement with human instruction datasets, such as LLaVA [7], mPLUG-Owl [20], and Kosmos-2 [21]. This methodology enables image LMMs to effectively integrate and reason with multimodal data, producing content that aligns with human instructions.

D. Video LMMs

Building upon image LMMs, video LMMs like Video-LLaVA [10], Video-ChatGPT [11], Video-LLaMA [12] have been developed to handle complex video content, understanding prompts and generating coherent answers beyond simple labels, all of which allows a chatbot-like interaction to analyse videos. However, these models often work best on short clips of a few seconds to 1-2 minutes long and on videos with a single scene, facing challenges with multi-scene videos and capturing temporal relationships [10].

E. Fusion Techniques

Information fusion is a techinque for integrating data from multiple sensors or different modalities into a unified representation, often with the goal of enhancing the system's performance [22]. This technique is primarily divided into two categories: late fusion and early fusion [23]. Late fusion processes each modality independently and combines them just before the decision-making stage, whereas early fusion merges all modalities at the initial stages, often transforming them into a single latent space [23]. Although not always explicitly discussed, fusion techniques have been utilised in fields such as image/video captioning and question answering.

LLaVA [7] employs an early fusion method by integrating visual features Z_v — extracted from an input image X_v using the CLIP visual encoder [4] — with language embeddings through a projection matrix W. This maps the visual features Z_v to language tokens $H_v = W \cdot Z_v$. However, this approach is primarily limited to static images and requires significant adaptations and retraining to handle other information modalities, such as temporal information in videos. VideoLLaVA [10] adopts an early fusion approach as well, integrating images and videos into a common feature space. This enables the LLM to leverage a unified visual representation for learning. Similarly, early fusion strategies have been implemented in Video-ChatGPT [11] and Video-LLaMA [12] to analyse video content with LMMs.

Most video captioning and Q&A models have relied on a single-model, early fusion method, which often restricts their efficacy to shorter video clips and overlooks the nuances in longer videos. QCaption introduces a late fusion technique by deploying a pipeline of multiple pre-trained models, each specialising in different modalities. This approach enhances video analytics performance and enables the extraction of richer information from videos of varied lengths.

III. METHODOLOGY

A. QCaption Fusion Architecture

The QCaption fusion pipeline fuses three models (Fig. 1) Video-to-image keyframe extractor, 2) Image and text to text captioning and Q&A LMM, and 3) Text-to-text LLM.

In stage 1, given input video, a key frame extractor extracts frames from the video as a set of images, $\mathcal{S} = \{I_1, I_2, \ldots, I_{N-1}, I_N\}$.

In stage 2, each frame, I_i, for $i \in \{1, 2, \ldots, N\}$ is then passed through an LMM model, f_{LMM}, along with the input caption X_T, which is kept consistent for each frame, as denoted by equation 5. In other words, the input prompt posed to the video is repeated across all frames to caption them or generate an answer.
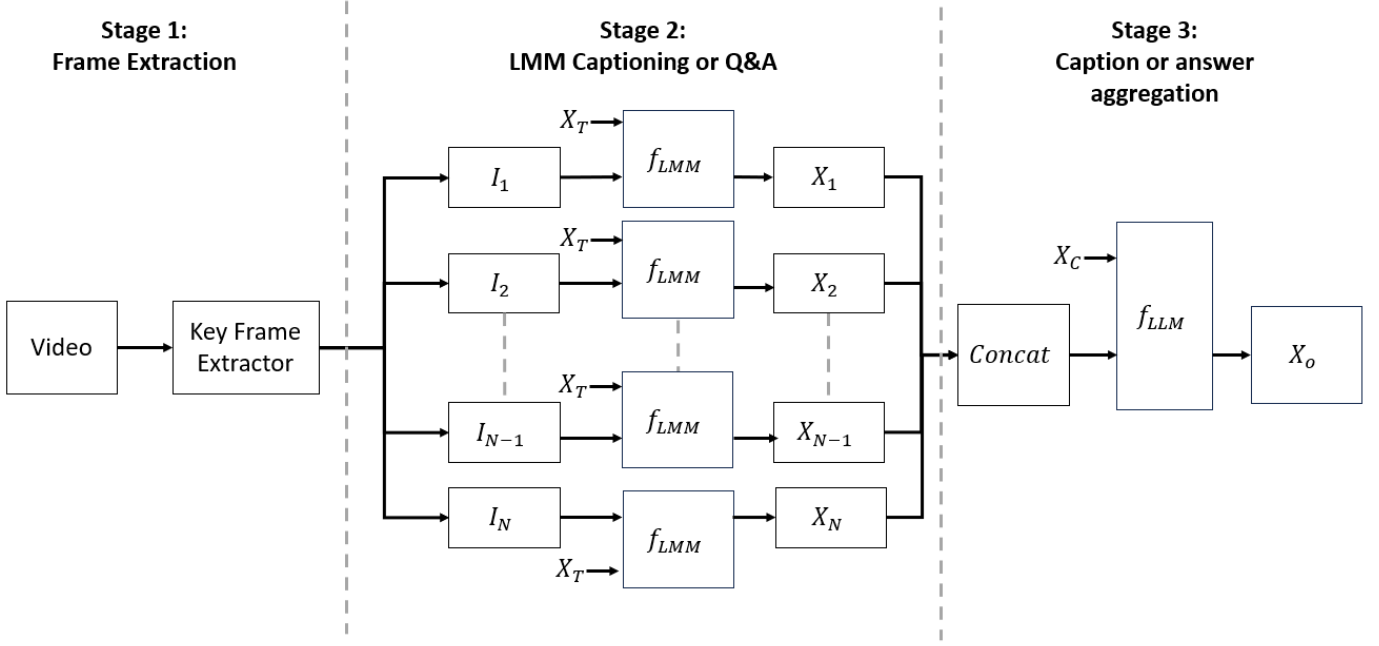
Fig. 1. QCaption multi-modal fusion pipeline.

$$X_i = f_{\text{LMM}}(I_i, X_T), \quad \text{for } i \in \{1, 2, \ldots, N\} \qquad (1)$$

Finally in stage 3, the captions are aggregated to generate a coherent and succinct answer that retains key details. Each individual captions are first combined through a simple string concatenation, per equation 2. Indexes are added in front of each caption prior to concatenation to denote their temporal position (i.e., earliest frame is denoted "0", next "1", and so on).

$$\mathcal{C} = X_1 \| X_2 \| \ldots \| X_N \qquad (2)$$

The concatenated block of captions, $\mathcal{C}$, is then passed into a LLM model, f_{LLM}, with a prompt X_C, specially designed to aggregate the captions together (section III-D). From experiments, the recommended prompt to use in this stage varies depending on the task and dataset. This produces the final caption or answer for the video, $X_o = f_{\text{LLM}}(\mathcal{C}, X_C)$.

For all experiments below, we use $N = 8$, following the frame sampling procedure in Video-LLaVA training stage [10].

B. Key frame extraction

We employed three approaches to frame sampling to assess how it will affect performance depending on dataset used.

One approach is using the Katna [24] model for key frame extraction. Katna employs a selective approach in frame selection. Using its frame extractor module, it compares all the video frames that are sufficiently different from previous ones using absolute differences in LUV colorspace, as described in the Katna documentation. Moreover, Katna implements K-Means Clustering of frames using image histograms, enabling a structured organization of frames for enhanced analysis. The selection of the best frame from these clusters is determined based on the variance of Laplacian, facilitating image blur detection to ensure optimal frame representation.

As an alternative to Katna, we also applied a regular sampling method. Proposed by Gemini [6], frames were sampled at regular intervals under this approach. Each video was divided into equal segments and a frame was extracted from the midpoint of each segment for analysis.

A third approach involved random sampling, where frames are selected randomly from throughout the video, providing an randomised, varied set of frames for analysis.

C. Frame Captioning

Frame captioning involves generating text descriptions for individual frames within a video. This process turns visual information into textual descriptions, making it easier for computers to understand and interpret video content. We used LlaVA [7] to generate these captions. These models are trained on datasets comprising images and their corresponding descriptive captions. This training enables them to understand the correlation between visual elements and textual descriptions, where LLaVA specialises in generating captions for static frames.

D. Caption Aggregation

Caption aggregation is achieved through the use of the Vicuna LLM [9], which synthesises text descriptions from individual video frames into a coherent summary. This process enhances the understanding of video content by filtering out redundancies and selecting crucial information if different

frames yielded different information. This ensures that the final summary is both concise and informative, enhancing understanding for users. Prompt engineering was employed through a trial-and-error approach to determine an optimal prompt for caption aggregation.

E. Advanced Sampling - Multi clips method

As an alternative to the above frame captioning methods, we also propose the mutli-clips captioning method. The multi-clips method involves an approach to video captioning where we sample four 5-second clips at regular intervals from the video. This approach aims to capture more temporal information, for more efficient content management and comprehensive analysis. Once the video is segmented, each subsection is independently processed through VideoLLaVA [10] to generate the relevant captions. Following this, the captions generated from each subsection are then passed into the LLM for caption aggregation.

F. Ablation Studies and Other methods

1) Ablation Studies: We also performed ablation studies to evaluate the impact of the LLM stage on the fusion pipeline. In some experiments, the LLM was removed and the individual frame captions were joint to gather through naive string concatenation.

2) First N-frames method: We also implemented a method which returned the first N-frames for analysis. Similar to the previously mentioned techniques, we applied both an LLM and non-LLM approach to the captioning method.

Finally, various combinations of the above were tested. The goal is to evaluate a range of LMM-based video captioning strategies and whether any of them are more suitable for some datasets.

IV. RESULTS

A. Benchmark Datasets

To benchmark QCaption and other video captioning and Q&A approaches against existing work like Video-LLaVA [10], we employed three commonly reported datasets: YouCook2 [25] and MSR-VTT [26] for video captioning, and ActivityNet-QA [27] for video Q&A.

YouCook2 comprises untrimmed, third-person viewpoint videos from 89 global cooking recipes, each annotated with procedural steps in imperative English sentences. Following the Gemini paper's approach [6], we used the YouCook2 validation set comprising ≈450 videos.

MSR-VTT is a comprehensive large-scale video description dataset with a range of web clips extracted from a commercial video search engine, accompanied by an average of 20 associated captions per video, annotated by AMT workers. Topics depicted in MSR-VTT include news reports, animals, transportation, cooking, etc. We used the MSR-VTT validation set comprising ≈500 videos.

ActivityNet-QA is a fully annotated, large-scale VideoQA dataset with web videos and on average 10 QA pairs for each video. The videos are derived from the ActivityNet [28]

dataset, encompassing videos on a diverse array of complex human activities relevant to daily life. Likewise, following authors of Gemini [6], we employed the ActivityNet-QA test set comprising 800 videos.

B. Evaluation Metrics

1) CIDEr: For the task of video captioning using YouCook2 and MSR-VTT datasets, after captions were generated using the various pipelines in our experiments, we compared them against the ground truth captions (from the dataset) using the **CIDEr** metric [29]. This aligns with the metric used by authors of Gemini [6]. Originally developed for image captioning but later adopted for video captioning, CIDEr quantifies that for given video I_i, how well its generated caption c_i aligns with the consensus of a set of ground truth descriptions $S_i = \{s_{i1}, \ldots, s_{im}\}$. During evaluation, each caption is represented by a set of *n-grams* ω_k comprising 1-4 words. The Term Frequency Inverse Document Frequency (TF-IDF) for each *n-gram*, $g_k(s_{ij})$, is then computed using equation 3, which accounts for how often *n-grams* in the generated caption is present in ground truths, and that *n-grams* not in ground truth should not appear in generated caption. Note that $h_k(\cdot)$ the number of times a *n-gram* occurs in a caption.

$$g_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_{\omega_l \in \Omega} h_l(s_{ij})} \log \left(\frac{|I|}{\sum_{I_p \in I} \min(1, \sum_q h_k(s_{pq}))} \right), \tag{3}$$

The CIDEr score for *n-gram* of length n is then computed using equation 4, which compares the average cosine similarity between generated and ground truth captions.

$$\text{CIDEr}_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{g^n(c_i) \cdot g^n(s_{ij})}{\|g^n(c_i)\| \|g^n(s_{ij})\|} \tag{4}$$

Scores from individual *n-grams* are then summed up to yield the final CIDEr score, per equation 5.

$$\text{CIDEr}(c_i, S_i) = \sum_{n=1}^{N} w_n \text{CIDEr}_n(c_i, S_i), \tag{5}$$

2) Video-ChatGPT metric: To evaluate the quality of answers generated for Q&A tasks using the ActivityNet-QA dataset, we employed the Video-ChatGPT evaluation protocol [10], also used by Video-LLaVA [10] and Gemini [6]. The metric was designed to evaluate the text generation capabilities of video-based conversational models (e.g., video Q&A). This evaluation encompasses five critical dimensions: Correctness of Information, Detailed Orientation, Contextual Understanding, Temporal Understanding, and Consistency. For each input question for a video, every generated and ground truth pair are evaluated in the above dimensions using ChatGPT "gpt-3.5-turbo" version, called through the OpenAI API in the evaluation script.

TABLE I
PERFORMANCE COMPARISON ON VIDEO CAPTIONING AND Q&A BENCHMARKS, BETWEEN EXISITING WORK AND QCAPTION.

S/N	Method		Video Captioning		Video Q&A	
			YouCook2 (val)	MSR-VTT (val)	ActivityNet-QA (test) Accuracy	Score
1	Baselines	VideoLLaMA	-	-	12.4	1.10
2		VideoChatGPT	-	-	35.2	2.70
3		Video-LLaVA	19.9	28.6	42.5	3.53
4	QCaption	Katna + LLaVA + LLM	**28.7**	36.8	61.6	3.74
5		Regular Sampling + LLaVA + LLM	26.3	36.5	**63.3**	3.87
6		Random Sampling + LLaVA + LLM	26.0	35.1	58.2	3.78
7		Multiclips: Clip Sampling + Video-LLaVA + LLM	25.5	**41.7**	53.1	3.52

3) Qualitative: We also performed qualitative experiments using videos of a few minutes long, each featuring different scenes or point-of-views, to compare the prowess of each pipeline in capturing key details and the sequence of events depicted.

C. Baseline

To validate the proposed QCaption multi modal fusion pipeline, we compared it against the following notable LMM-based video captioning models:

- Video-LLaMA [12]: LMM that employs a common visual encoder to understand both images and videos.
- Video-ChatGPT [11]: LMMs that assign a unique encoder to each modality, for the ability to understand images or videos via multiple projection layers.
- Video-LLaVA [10]: LMM that aligns image and video representations before projection, which enables a unified visual representation of both image and video. Employs the Vicuna-v1.5 LLM, ViT-L/14 visual encoder, and LlaMA text tokeniser, just like LLaVA-v1.5 [7]. Video-LLaVA outperformed both Video-LLaMA and Video-ChatGPT.

The baselines all employ an early fusion approach; QCaption utilises late fusion.

QCaption is designed to be LMM and LLM agnostic, one can swap out any part of the pipeline for other models. For clarity of comparison, we built QCaption on top of LLaVA-v1.5 [7], a LMM for image captioning and Q&A that shares ample similarities to the architecture of Video-LLaVA. The LLM used is Vicuna-v1.5 [9].

Performance benchmarks for Video-LLaMA and Video-ChatGPT were adapted from the Video-LLaVA paper; results for Video-LLaVA were from experiments closely following the experimental procedure stipulated in the paper [10]. For YouCook2 and MSR-VTT captioning datasets, the CIDEr [29] metric was used (section IV-B1); for ActivityNet-QA Q&A dataset, the Video-ChatGPT evaluation approach (section IV-B2) was adopted.

D. QCaption

Using the QCaption pipeline (section III-A), two frame extraction methods were tested: using Katna key frame sampling (section III-B) and Regular Sampling (taking frames at equally spaced intervals in the video). 8 frames were sampled for each video. For both approaches, frames were annotated using LLaVA-1.5 LMM and the captions were aggregated using the Vicuna-v1.5 LLM. Both methods yielded improvements over the baseline Video-LLaVA approach for some captioning and Q&A tasks.

Referring to Table I, Using Katna + LLaVA + LLM yielded 44.2%, 28.7%, 44.9% improvements for video captioning (YouCook2, MSR-VTT) and Q&A (ActivityNet-QA) tasks respectively. Using Regular Sampling + LLaVA + LLM yielded improvements of 32.2%, 27.6% and 48.9% respectively on captioning (YouCook2, MSR-VTT) and Q&A (ActivityNet-QA).

The results indicate that regular sampling of frames yields better results for Q&A tasks, but key frame extractions (using Katna) is more useful for captioning. In general, in tasks such as general scene captioning or answering general Q&As, treating video captioning as a multi-image captioning approach — by late fusion of frame extraction algorithms, LMM, and LLM — performs better compared to constructing a video-image-text model from scratch, which was the case of the baselines. Additionally, this strategy greatly reduces the resources needed for data curation and training of a new model.

E. Advanced Sampling - Multiclips

QCaption took this idea a step further by breaking a long video analysis problem into one of analysing multiple small videos. Similar to the image-based pipeline, this approach first samples short clips of 5s long at regular intervals, and then iteratively caption them using Video-LLaVA. The captions are then aggregated using a LLM. Per Table I, this approach yielded notable 45.8% and 24.9% improvements for both video captioning (MSR-VTT) and Q&A (ActivityNet-QA) tasks.

F. Ablation Studies and Other methods

We also conducted a comparison of QCaption with other possible video captioning methods and performed ablation studies by disabling or swapping out certain stages of QCaption, all of which is detailed in Section III-F. Referencing table II, the best video captioning approach varies depending on the task and type of answers expected in a given dataset.

| S/N | Method | | Video Captioning | | Video Q&A | |
			YouCook2 (val)	MSR-VTT (val)	ActivityNet-QA (test) Accuracy	Score
1	QCaption	Katna + LLaVA + LLM	**28.7**	36.8	61.6	3.74
2		Regular Sampling + LLaVA + LLM	26.3	36.5	**63.3**	**3.87**
3		Random Sampling + LLaVA + LLM	26.0	35.1	58.2	3.78
4		Multiclips: Clip Sampling + Video-LLaVA + LLM	25.5	41.7	53.1	3.52
5	Ablation Studies	Regular Sampling + LLaVA (no LLM)	21.3	**52.3**	55.9	3.47
6		Random Sampling + LLaVA (no LLM)	21.1	**52.3**	55.0	3.42
7	Other methods	First N frames + LLaVA (no LLM)	9.4	42.9	45.1	3.34
8		Sample 1 frame only + LLaVA	13.9	34.1	57.4	3.76

For the YouCook2 dataset, which predominantly features detailed captions outlining specific ingredients and recipe steps, we discovered that regular sampling of frames, caption generation via LMM, and subsequent integration through simple string concatenation (without LLM aggregation) still outperforms the baseline with a 7.04% improvement. With the use of the LLM, as proposed by QCaption, the results show a further improvement of 19.0%. Conversely, employing the first N frames or single frame sampling method all yields vastly inferior results, owing to the loss of information from lack of samples. The above all supports the effectiveness of the Fusion approach proposed by QCaption.

For the ActivityNet-QA dataset, focusing on Q&A tasks, the inclusion of LLMs emerged as crucial for generating coherent and precise answers. Our ablation studies demonstrated that eliminating the LLM stage led to a performance decline of up to 10.3%, compared to QCaption's LLM-integrated methods. Interestingly, employing a single-frame sampling strategy yielded outcomes on par with QCaption's top-performing method, suggesting that straightforward questions might be effectively addressed using just one indicative frame from the video. Nonetheless, the combination of QCaption's random sampling with LMM and subsequent LLM processing consistently delivered superior results, underscoring the significance of this fusion approach in video Q&A.

For the MSR-VTT dataset, eliminating LLM-aggregation led to improvements, with up to a 43.3% increase in performance over the LLM-enabled QCaption method. This likely stems from MSR-VTT datasets employing much shorter, often one-word long descriptions. Using a LLM could potentially abstract away that one-word that was correctly captured during the video captioning process. Interestingly, employing the first N frames (without LLM) achieved results on par with QCaption's top-performing approach (Multiclips). This indicates that, for MSR-VTT videos, capturing the initial scene alone is often adequate for a comprehensive understanding of the video, likely due to the absence of significant scene transitions, and one-word answer. The case for MSR-VTT is rather unique, as general video captioning / Q&A tasks like YouCook2 and ActivityNet-QA require answers in complete sentences, which

could be effectively captured using the QCaption pipeline.

In all, QCaption's fusion pipeline consistently outperforms baselines like Video-LLaVA for both captioning and Q&A datasets. Notably, for both tasks where a comprehensive, complete answer is expected, the LLM is often an important final step to aggregate a coherent answer.

G. Qualitative Experiments

In Fig. 2, we curated two illustrative examples to meticulously assess the video captioning prowess of QCaption. We systematically juxtapose its performance against Video-LLaVA using a pair of carefully selected videos. These videos are deliberately chosen to encompass diverse camera angles, thereby elevating the intricacy of the video inputs. The primary objective is to gauge the nuanced capabilities of each model in the realm of video captioning.

Our overarching observation indicates that both QCaption and Video-LLaVA adequately interpret and articulate the content of the videos in accordance with the provided instructions. However, QCaption demonstrates a enhanced proficiency in capturing crucial information spanning multiple scenes within each video. Its prowess becomes evident in discerning changes in camera angles and perspectives, particularly highlighted in **purple** in the presented examples. This makes QCaption's fusion approach preferrable for understanding longer videos with multiple scene changes.

V. CONCLUSION AND FUTURE WORK

In this paper, we introduced QCaption, a video captioning and Q&A pipeline that stands out for its fusion of three distinct models: a key frame extraction model, an LMM for image-text analysis, and an LLM for text analysis. This integrated approach enables simultaneous analysis of text, images, and video, achieving superior performance both quantitative and qualitative experiments for video captioning and Q&A tasks, all while using pre-trained models. We also performed ablation studies to examine how certain components of the pipeline (or the lack thereof) can impact performance of different tasks, and proposed and bench-marked other possible video captioning pipelines.

[Input Instruction] Describe this video. Be succinct, limit to 1-2 sentences.

Video-LLaVA

The video shows a large statue of a man holding a torch and a book, with a body of water in the background.

Q Caption — Katna

The video features a statue of the Statue of Liberty in various locations and angles, including a low angle, high angle, and bird's eye view. The statue is also shown overlooking a river with a boat passing by and is lit up at night. The video also includes a view of the statue from the water and a statue located on an island in the middle of a river.

Q Caption — Regular Sampling

The video features a large boat traveling on the water near the Statue of Liberty, which is lit up at night. Other scenes include a statue of the Statue of Liberty flying a kite and a statue of a woman holding a torch on top of a building.

Q Caption — Multiclips: Clip Sampling

The video shows a variety of statues, including one of a woman holding a torch and a crown, and another of a woman holding a torch and a book. The statues are located in different settings, such as a park and on a hill overlooking a river.

[Input Instruction] Describe this video. Be succinct, limit to 1-2 sentences.

Video-LLaVA

The video shows a parking lot with a car parked in it. The car is parked in a spot where it is not allowed to park. The video also shows a security camera that is recording the parking lot.

Q Caption — Katna

The video shows a black car driving down a street at night, followed by a parking lot with a camera monitoring the area. The black car is also seen parked in front of a house with a camera watching the area. Other scenes include a garbage truck parked next to a white truck, a white van parked in a parking lot, and a person walking in the distance in a nighttime view of a parking lot.

Q Caption — Regular Sampling

The video shows a parking lot at night with a camera on the top right corner. There are several cars and trucks parked in the lot, including a black car parked in front of a house and a garbage truck parked in front of a building. A person is walking in the distance, and a blue car and a white van are also visible in the scene.

Q Caption — Multiclips: Clip Sampling

The video shows a man walking around a parking lot, a car driving away, a large truck turning into a parking lot, a black car driving down a street at night, a parking lot at night with a car parked in the middle, and the car being seen from different angles.

Fig. 2. Comparison of Video-LLaVA and QCaption's video captioning capabilities

Future work We observed that outcome of the caption aggregation stage is highly dependent on the input prompt for the LLM, for it determines what information is retained and abstracted away. We prompt engineered using a trial and error approach to qualitatively determine a prompt that yields decent results. There is room to dive deeper into how the aggregation prompt affects quantitative results, and whether the prompt to use varies between datasets. There is also room to test on more recent image LMM models like Gemini Pro [6] and GPT-4V [5] to further improve the performance while retaining the same pipeline. Finally, the modular approach to fusion also affords potential to incorporate other modalities, such as speech to text models into the pipeline, which further enrich the information that the LLM can work with, to produce more comprehensive captions and answers. QCaption can serve as a baseline for fusion-based approaches to video captioning that do not require costly architecture modification and retraining of models.

REFERENCES

[1] Tom B. Brown, Benjamin Mann, and Nick Ryde. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020-December, 5 2020.

[2] Aakanksha Chowdhery, Sharan Narang, and Jacob Devlin. Palm: Scaling language modeling with pathways. 4 2022.

[3] Hugo Touvron, Louis Martin, and Kevin Stone. Llama 2: Open foundation and fine-tuned chat models. 7 2023.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *Proceedings of Machine Learning Research*, 139:8748–8763, 2 2021.

[5] OpenAI. Gpt-4 technical report. 3 2023.

[6] Gemini Team. Gemini: A family of highly capable multimodal models. 12 2023.

[7] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. 4 2023.

[8] Moloud Abdar, Meenakshi Kollati, Swaraja Kuraparthi, Farhad Pourpanah, Daniel McDuff, Mohammad Ghavamzadeh, Shuicheng Yan, Abduallah Mohamed, Abbas Khosravi, Erik Cambria, and Fatih Porikli. A review of deep learning for video captioning. 4 2023.

[9] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. 4 2023.

[10] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munang Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. 11 2023.

[11] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models, 2023.

[12] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding, 2023.

[13] Adel Jalal Yousif and Mohammed H. Al-Jammas. Exploring deep learning approaches for video captioning: A comprehensive review. *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, 6:100372, 12 2023.

[14] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. *Proceedings of the IEEE International Conference on Computer Vision*, pages 2712–2719, 2013.

[15] Hanan Nasser Alkalouti and Mayada Ahmed Al Masre. Encoder-decoder model for automatic video captioning using yolo algorithm. *2021 IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021 - Proceedings*, 4 2021.

[16] Naveen Yadav and Dinesh Naik. Generating short video description using deep-lstm and attention mechanism. *2021 6th International Conference for Convergence in Technology, I2CT 2021*, 4 2021.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017-December:5999–6009, 6 2017.

[18] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. 10 2023.

[19] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, Yueting Zhuang, Zhejiang University, and Microsoft Research Asia. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. 3 2023.

[20] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality. 4 2023.

[21] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. 6 2023.

[22] Weishi Li, Yong Peng, Miao Zhang, Liang Ding, Han Hu, and Li Shen. Deep model fusion: A survey, 2023.

[23] Maciej Pawłowski, Anna Wróblewska, and Sylwia Sysko-Romańczuk. Effective techniques for multimodal data fusion: A comparative analysis. *Sensors (Basel, Switzerland)*, 23, 3 2023.

[24] Mayank Jain, Nitin Katyal, and Alok. Github - keplerlab/katna: Tool for automating common video key-frame extraction, video compression and image auto-crop/image-resize tasks, 2021.

[25] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 7590–7598, 3 2017.

[26] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language, 6 2016.

[27] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, pages 9127–9134, 6 2019.

[28] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.

[29] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June-2015:4566–4575, 11 2014.