ViTA: An Efficient Video-to-Text Algorithm using VLM for RAG-based Video Analysis System

Md Adnan Arefeen[1,2], Biplob Debnath[1], Md Yusuf Sarwar Uddin[2], and Srimat Chakradhar[1]
NEC Laboratories America[1], University of Missouri-Kansas City[2]
{aarefeen, biplob, chak}@nec-labs.com, muddin@umkc.edu

Abstract

Retrieval-augmented generation (RAG) is used in natural language processing (NLP) to provide query-relevant information in enterprise documents to large language models (LLMs). Such enterprise context enables the LLMs to generate more informed and accurate responses. When enterprise data is primarily videos, AI models like vision language models (VLMs) are necessary to convert information in videos into text. While essential, this conversion is a bottleneck, especially for large corpus of videos. It delays the timely use of enterprise videos to generate useful responses.

We propose ViTA , a novel method that leverages two unique characteristics of VLMs to expedite the conversion process. As VLMs output more text tokens, they incur higher latency. In addition, large (heavyweight) VLMs can extract intricate details from images and videos, but they incur much higher latency per output token when compared to smaller (lightweight) VLMs that may miss details. To expedite conversion, ViTA first employs a lightweight VLM to quickly understand the gist or overview of an image or a video clip, and directs a heavyweight VLM (through prompt engineering) to extract additional details by using only a few (preset number of) output tokens. Our experimental results show that ViTA expedites the conversion time by as much as 43%, without compromising the accuracy of responses when compared to a baseline system that only uses a heavyweight VLM.

1. Introduction

Many enterprises routinely collect and archive videos for post-analysis. These archived videos enable enterprises to extract valuable business insights, investigate incidents, prevent disasters, and obtain answers to various inquiries. For example, surveillance systems play a vital role in upholding public safety, monitoring public areas, airports, transportation centers, and critical infrastructure through video cameras [26]). Moreover, there is a growing use of video analytics for patient monitoring in hospitals and healthcare facilities [28]. City planners have installed thousands of traffic cameras, with approximately one thousand of these cameras archiving 230 TB of video data every month [11]. Analyzing video feeds from traffic cameras for traffic monitoring, congestion management, and incident detection requires a deep understanding of the information within videos [23].

Large language models (LLMs), such as ChatGPT [4], have demonstrated their ability to engage in human-like text-based conversations. However, LLMs cannot answer questions related to data not seen during training. Thus, LLM cannot generate correct answer when processing enterprise data. As a remedy, the retrieval-augmented generation (RAG) method [13] is widely adopted where enterprise data is provided as context to the LLMs. However, conventional RAG systems operate under the assumption that enterprise information is presented in textual form. If enterprise data exists in non-textual formats like video or images, traditional RAG systems need an extra component to transform non-textual media into text.

Figure 1 illustrates a RAG-based system designed to address questions related to enterprise videos. Originally designed to handle textual data, RAG-based systems are adapted for videos through a conversion process that transforms videos into text. To initiate the process, the video RAG system converts the video file into text, dividing it into non-overlapping clips. Each clip undergoes analysis by a vision language model (VLM) [19, 27], recording the output in text format. The VLM can process entire clips or subsets of frames, extracting visual information in textual form. Collating text information from all clips generates a lengthy document for the video file, which is then segmented into chunks. Each chunk is embedded into a vector by an embedding model and subsequently stored in a vector database. Once the video-to-text conversion is complete, upon receiving a query, the video RAG system embeds the query, and conducts a semantic search using the embedding vectors of chunks to retrieve relevant chunks from the vector database. These retrieved chunks form the context, which is combined with the query to generate a prompt. Finally,

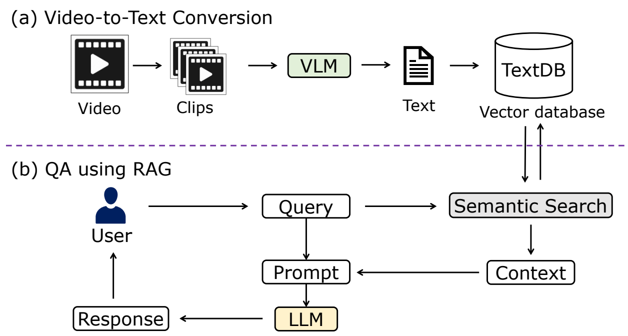

Figure 1. An overview of the RAG-based video analysis system. It operates in two phases: a) the video is initially converted into a text document, which is then chunked into smaller parts and stored in a vector database; b) queries are answered using a Large Language Model (LLM), leveraging query and contextual information retrieved from the vector database through semantic search.

the prompt is fed into a Large Language Model (LLM) to generate a response for the query.

While a RAG-based system for enterprise videos can address a wide range of queries, a major challenge lies in the time required for a VLM model to generate textual descriptions from video clips. For instance, processing a frame using the InternLM-XComposer2 [6] model, with a maximum token limit of 64, takes an average of 3.8 seconds on a server equipped with an NVIDIA GeForce RTX 3090 GPU. Thus, it would take more than a day for this VLM model to analyze a 24-hour surveillance video and generate text descriptions, assuming it processes one frame every three seconds. This poses a significant challenge for various applications, such as preventing law enforcement agencies from conducting timely analyses of criminal incidents captured in the video.

The quality of the output and the time required to generate textual descriptions vary among different VLMs (see Section 2.1 for detailed information). A lightweight VLM (i.e., with fewer parameters) runs faster, but its output contains limited information, which hinders our ability to answer many queries due to a lack of information. Conversely, a heavyweight VLM (i.e., with more parameters) produces output with richer information, but it takes longer to generate text. The time needed for text generation depends on the limit set for the number of output tokens. In this paper, our main goal is to extract richer information from videos to enable answering a wide variety of queries.

To this end, we present ViTA , a novel algorithm for rapidly generating textual descriptions of video clips using VLM models. Rather than relying solely on a heavy-weight vision language model (VLM), ViTA employs a hybrid approach that combines both lightweight and heavy-weight VLMs. Initially, the lightweight VLM generates a text description, which then serves as input for the heavy-weight VLM. Leveraging the lightweight VLM for the initial text generation allows ViTA to reduce the maximum output token limit of the heavy-weight VLM, subsequently minimizing its latency. ViTA effectively utilizes the efficiency of the lightweight VLM to guide the heavy-weight VLM in producing responses with fewer tokens, thereby achieving significant reductions in latency during the video-to-text conversion process. Our experiments show that ViTA reduces the latency of video-to-text conversion by as much as 43% through its hybrid approach, while preserving the accuracy of the query results.

2. ViTA

2.1. Motivation

Existing vision language models (VLMs) vary in capability of generating response. We consider the VLMs with less parameters as lightweight models and the VLMs with large parameters as heavy-weight VLMs. Due the less parameters, lightweight VLMs runs faster than heavy-weight VLMs. For efficient conversion of video to text, VLMs must handle video clips rapidly. A lightweight VLM can achieve faster inference but may only capture a subset of information. Conversely, a heavyweight VLM can capture more detailed information but at the cost of slower inference speed. Usually, heavyweight VLMs provide a maximum output token limit parameter to control the amount of information generated, thereby controlling the inference speed. The appropriate level of detail required for describing an image or a video clip depends on the specific scenario.

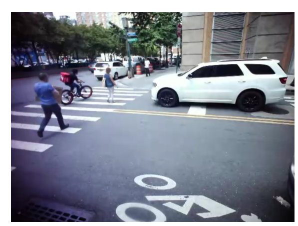

Figure 2. A sample image from the StreetAware [25] dataset.

For illustration, let us consider the image shown in Figure 2 related to a street in New York city from the StreetAware [25] dataset. At first, we execute the BLIP model [14] model, which is a light-weight Vision-Language Model (VLM) for caption generation given an image. It generates the following textual description:

BLIP: A man riding a bike down the street.

Following the utilization of the BLIP model [14], we proceed to generate a textual description employing the InternLM-Xcomposer2 model [6], and LLAVA-1.5-7b model [20]. The lightweight BLIP model has 247.4 million parameters, while the InternLM-Xcomposer2 and LLAVA models have 7 billion parameters. Utilizing the prompt "Compose a descriptive narrative", we employ various maximum output token limits, specifically 16, 32, 64, 128, and 256. This diversity allows us to observe the models response under different token lengths. Subsequently, we observe the generated response, with each response distinguished by a color corresponding to the token limit used for its generation as follows.

InternLM-Xcomposer2:
Token limit: 16, 32, 64, 128, 256

The image captures a bustling city scene, where life unfolds in its vibrant and dynamic form. A white SUV is seen making a left turn at an intersection, while a man on a red motorcycle is crossing the street, adding a splash of color to the urban landscape. The perspective of the photo suggests it was taken from the sidewalk, providing a ground-level view of the city's hustle and bustle. In the background, tall buildings stretch towards the sky, their towering presence a testament to the city's architectural prowess. Trees line the streets, offering a touch of nature amidst the concrete jungle. The colors in the image are predominantly gray and green, with hints of red and white adding contrast and depth to the scene.

LLAVA:
Max Token limit: 16, 32, 64, 128, 256

The image depicts a busy city street with a mix of vehicles and ped estrians. A white car is driving down the street, while a bicycle is parked on the sidewalk. There are several people walking along the sidewalk, some of them carrying handbags. In addition to the car and bicycle, there are multiple traffic lights visible in the scene, ensuring the smooth flow of traffic. The street is lined with trees, providing a touch of greenery to the urban environment. Overall, the scene captures the hustle and bustle of city life .

In Table 1, a comparison of the latency between a lightweight model BLIP [14] and heavyweight models InternLM-Xcomposer2 [6] and LLAVA [20] is illustrated. For BLIP, it only takes 0.22 seconds to generate the textual description. In contrast, for the heavyweight models InternLM-Xcomposer2 and LLAVA, it takes up to 8.94 seconds (40.6× slower) and 3.27 seconds (14.9× slower), respectively, to generate a textual description of the image shown in Figure 2.

Model	Max token limit	Tokens Generated	Latency (seconds)
BLIP [14]	—	—	0.22
InternLM-xcomposer2-7b [6]	1	1	0.64
	4	4	0.71
	8	8	0.91
	16	16	1.43
	32	32	2.17
	64	64	3.77
	128	128	7.01
	256	149	8.94
LLAVA-1.5-7b [20]	1	1	0.21
	4	4	0.24
	8	8	0.33
	16	16	0.53
	32	32	0.91
	64	64	1.68
	128	128	3.22
	256	129	3.27

Table 1. Latency of the lightweight BLIP model and heavyweight InternLM-Xcomposer2 model. The prompt is set to "Compose a descriptive narrative".

As depicted in Table 1, for both InternLM-Xcomposer2 and LLAVA, there is a clear correlation between generation of output tokens and latency. Higher token counts lead to longer processing times. Therefore, by strategically generating tokens only when essential and excluding unnecessary text, latency can be significantly reduced, facilitating rapid extraction of more comprehensive textual information.

2.2. Proposed Method

Instead of solely using a heavyweight vision language model (VLM) for generating textual description of a video clip, ViTA proposes a hybrid approach integrating both lightweight and heavyweight VLMs. Initially, the lightweight VLM generates a description, serving as input for the heavyweight VLM. This strategy reduces the heavyweight VLM's maximum output token limit, lowering its latency. By leveraging the lightweight VLM's efficiency to guide the heavyweight VLM, ViTA achieves substantial latency reductions in the video-to-text conversion process.

Figure 3 gives an overview of workflow of ViTA . The video undergoes segmentation into multiple clips using a scene detection algorithm, such as PySceneDetect [1]. Subsequently, each clip undergoes a two-step conversion process. Initially, ViTA employs a lightweight VLM (for example, BLIP [14]) to generate a textual description for the video clip. Following this, a heavyweight VLM (for example, InternLM-XComposer2 [6]) utilizes the textual description from the previous step along with a prompt as instructions to extract additional textual information for the clip. Finally, both textual descriptions are concatenated and added to the vector database.

To generate textual description for an image, after running the lightweight model, ViTA uses the following

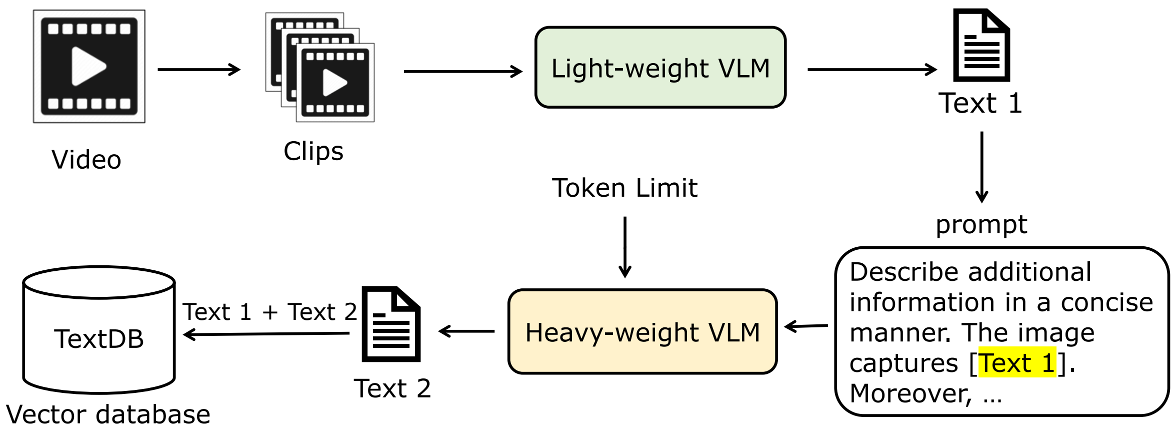

Figure 3. Overview of the workflow of our video-to-text conversion algorithm. Initially, a lightweight VLM is employed to generate a text description of a video clip. Subsequently, a heavyweight VLM is employed, utilizing the initial description as a prompt to generate additional information while limiting the output token count.

prompt to as an instruction to the heavyweight model:

```
Describe additional information
in a concise manner.  This image
captures  [lightweight model output].
Moreover, ...
```

As an illustration, for the image in Figure 2, when ViTA uses the textual description generated by the lightweight model BLIP [14], "a man riding a bike down the street", as a prompt to instruct the heavyweight model InternLM-Xcomposer2 [6] to generate additional information, with the maximum output token limit set to 32, it produces the following textual description:

```
The image captures a man riding a
bike down the street. Moreover, The
image also features a white SUV, a
person on a motorcycle, and another
individual walking.  The street has
markings for crosswalks and lanes,
with
```

In this manner, the initial textual description generated by the lightweight BLIP model aids the heavyweight InternLM-XComposer2 model in extracting additional textual information. Since an initial textual description is already generated by BLIP, ViTA directs the InternLM-XComposer2 to produce further details. ViTA establishes a maximum token limit to manage the latency of the heavyweight VLM model. Consequently, latency is reduced owing to this constrained output token generation in ViTA .

3. Evaluation

3.1. Datasets

For the evaluation of ViTA , we use two real-world video datasets i.e. StreetAware [25] and Tokyo MODI [12]. We consider a 46 minute clip for StreetAware dataset and a 2 hour video clip from Tokyo MODI dataset describing traffic scene with pedestrians. These datasets do not come with any query. We ask 5-7 questions on each videos, and generate the response using a large language model.

Dataset	Duration
StreetAware [25]	00:46:45
Tokyo MODI [12]	02:00:23

Table 2. Evluation dataset

3.2. Implementation Details

We utilize the `LangChain` framework [8] to design the our RAG based video processing system. Initially, we apply `PySceneDetect` [1] to each video to identify the clips. Subsequently, we sample one frame from each clip and employ the vision language model to generate a textual description. As a baseline, we utilize the `InternLM-Xcomposer2` [6] model with the following instruction prompt.

Baseline VLM Text Generation Prompt

```
Describe the clip in detail in a
concise manner.
```

In ViTA , for the lightweight VLM model, we utilize `BLIP` [14] to generate the initial textual description of a clip. Afterward, we incorporate the following prompt to generate additional information using `InternLM-Xcomposer2`:

> ViTA VLM Text Generation Prompt
> `Describe additional information in a concise manner. This image captures a scene from a street intersection showing [BLIP output].Moreover, ...`

After completing the video-to-text conversion for all clips, we consolidate all textual descriptions into a single text document. We then utilize `CharacterTextSplitter` [8] to segment the text into chunks, with a chunk size of 150 and a chunk overlap set to 0. Subsequently, we compute the embedding for each chunk using `OpenAIEmbeddings`[24] and store them in a vector database using `FAISS` [10]. During retrieval, we configure the system to retrieve the top 4 chunks related to the query, forming the context. Finally, we input the query and context into `gpt-3.5-turbo-1106` [24], with a temperature of 0. The maximum output token limit for the `InternLM-XComposer2` is set to 50 for the baseline, and 25 for ViTA . We conducted experiments on a server equipped with an AMD Ryzen 5950X 16-Core processor and an NVIDIA GeForce RTX 3090 GPU.

Table 3. Video-to-text conversion time.

Dataset	Method	Conversion Time
StreetAware	Baseline	00:36:15
	ViTA	00:20:24
Tokyo MODI	Baseline	01:18:02
	ViTA	00:46:22

3.3. Video-to-Text Conversion Time Reduction

Table 3 shows the comparison of video-to-text conversion time of ViTA and the baseline approach. We observe a significant latency reduction using ViTA for both datasets due to efficient utilization of heavy-weight VLM model with the help of auxiliary information from lightweight VLM models. For StreetAware 46 minute video, the video-to-text conversion takes 36 minute 15 seconds for the baseline approach, while ViTA reduces the conversion time to 20 minutes 24 seconds making the process faster by 43.7%. For Tokyo MODI video, ViTA reduces the conversion time by 40.58% by taking 46 minutes 22 seconds to transcript the 2 hr long video.

3.4. Query Results

3.4.1 StreetAware Video

We start querying over StreetAware video by asking the following queries to observe the context relevance and final response by LLM:

```
Q1.  Is there any bus?
Q2.  Is there any person with a
backpack?
```

The retrieved context by RAG is shown in Table 4. ViTA extracts the exact 3 clips belongs to the baseline and thus successfully responds the query `Q1`. As for `Q2`, although the retrieved context by baseline and ViTA seems different, by human evaluation on the video, we identify backpacks in those clips as shown in Figure 4.

(a) 00:22:26.786 (b) 00:45:45:000

(c) 00:45:51.429 (d) 00:02:02.143

Figure 4. Video clips related to Q2.

In Table 5, we present additional queries. For `Q3`, `Q4`, and `Q5`, the answers from ViTA match those generated by the baseline method. However, for query `Q6`, "Is there a person with a child", the baseline method fails to generate an answer, while ViTA successfully produces the correct answer. Overall, ViTA yields comparable (or even better) results compared to the baseline method.

3.4.2 Tokyo MODI video

For Tokyo MODI dataset, we ask the following queries:

```
Q1.Is there a large Sony billboard?
Q2.Is there any car seen making a
left turn?
```

Table 4. Query-response from the StreetAware video dataset. Context is retrieved using textual data implemented with RAG.

Q1	Is there any bus?
Baseline	**Context:** 00:34:30.000 The image features a blue and black bus driving on the street. The bus has a handicap accessible sign on its side, indicating that it is equipped to accommodate passengers with disabilities. . 00:34:24.856 The image captures a busy city street scene with a blue bus and a white van. The bus, which is on the left side of the image, displays the number 54 and has "UN MIRRLE" written on it. It appears . 00:34:33.213 The image features a blue and yellow bus driving on the street. The bus has an advertisement on its side that reads "Podcasts: yes. Surge pricing: never." There are also people visible inside the bus, including one person sitting near the front . 00:34:36.429 The image features a blue and yellow New York City bus driving down the street. The bus is adorned with an American flag on its side, indicating its pride in the country. Inside the bus, there are several passengers visible through the windows, including . **Response:** Yes, there is a blue and yellow bus with an advertisement for podcasts on its side. There is also a blue and black bus with a handicap accessible sign.
ViTA	**Context:** 00:34:24.856 This image captures a scene from a street intersection showing a bus is driving down the street in the city. Moreover, A white van is also present in the scene, and there are several people visible. The bus has a sign that reads " . 00:17:56.786 This image captures a scene from a street intersection showing a bus is seen in this image from the video. Moreover, The bus is driving on the street, and there are people visible in the image. The bus has a blue color with yellow . 00:34:30.000 This image captures a scene from a street intersection showing a bus is driving down the street in the city. Moreover, The bus is a blue and black vehicle with the number 7707 on it. It appears to be driving in the right . 00:34:33.213 This image captures a scene from a street intersection showing a bus is parked on the side of the road. Moreover, the bus is a blue MTA New York City Transit bus with the route B54 via Myrtle. The image also . **Response:** Yes, there is a bus driving down the street in the city. The bus is a blue and black vehicle with the number 7707 on it. Additionally, there is a blue MTA New York City Transit bus parked on the side of the road with the route B54 via Myrtle.
Q2	Is there any person with a backpack?
Baseline	**Context:** 00:09:54.643 The image captures a city street scene with a person crossing the road at a crosswalk. The individual is carrying a backpack and appears to be in motion, walking across the street. . 00:45:45.000 The image captures a bustling city street scene. A man with a backpack is crossing the street, walking in front of another person who is also carrying a backpack. The street is lined with parked cars and tall buildings, creating an urban atmosphere. There . 00:30:35.856 The image captures a bustling city street scene. A man with a backpack is crossing the street, while another person rides a bicycle in the background. The street is lined with cars parked on both sides and has a crosswalk painted with white stripes. . 00:05:02.142 The image captures a bustling city street scene. A blue car is driving on the right side of the road, while three people are crossing the street in front of it. The pedestrians are carrying backpacks and wearing jeans, suggesting they might be students . **Response:** Yes, there are people with backpacks in the city street scenes.
ViTA	**Context:** 00:02:02.143 This image captures a scene from a street intersection showing a person crossing the street with a bicycle. Moreover, the image also features a person crossing the street with a backpack. . 00:45:51.429 This image captures a scene from a street intersection showing a person walking across a street with a bike. Moreover, the image also features a person crossing the street on foot, and another individual walking with a backpack. . 00:22:26.786 This image captures a scene from a street intersection showing a person riding a bike down the street. Moreover, the image also features a person pushing a stroller and another individual carrying a backpack. . 00:02:05.356 This image captures a scene from a street intersection showing a person riding a skateboard down a street. Moreover, the image also features a person crossing the street with a backpack, and another individual walking on the sidewalk. . **Response:** Yes, there is a person with a backpack in the images.

```
Q3.Describe the buildings near the
street.
```

For query `Q1` "Is there a large Sony billboard?", the answer from ViTA matches the baseline method. However, for query `Q2` "Is there any car seen making a left turn?", ViTA fails to provide an answer due to a lack of generated information. To generate details about the left turn, more information extraction is necessary. Interestingly, for query `Q3` "Describe the buildings near the street", the baseline approach generates an incorrect answer, while ViTA produces a correct one. Overall, ViTA yields comparable results compared to the baseline method.

Table 5. Additional query-response from StreetAware video dataset.

Q3	Is there a person running at the intersection?
Baseline	**Response:** Yes, there is a person running at the intersection.
ViTA	**Response:** Yes, there is a person running at the intersection at 00:09:57.856 and at 00:35:47.143.
Q4	Is there any person with a skateboard?
Baseline	**Response:** Yes, there are people with skateboards in the city street scenes captured in the video clip.
ViTA	**Response:** Yes, there are people with skateboards in the video clip.
Q5	Is there any collision?
Baseline	**Response:** Based on the provided context, there is no mention of any collision occurring.
ViTA	**Response:** Based on the provided context, there is no mention of any collision occurring at the street intersection in the captured scenes.
Q6	Is there any person with a child?
Baseline	**Response:** I don't know.
ViTA	**Response:** Yes, at 00:24:28.929, there is a person crossing the street with a child.

4. Related Work

Traditional video understanding models have focused on foundational task-specific model design [2, 3, 5, 7, 15, 18, 31]. These models stand out by incorporating different neural network designs and employing various training procedures. They excel in analyzing short video clips. However, the focus of this paper is on analyzing long videos.

Recent years have witnessed notable advancements in multimodal learning, particularly in bridging the gap between video and language understanding, as evidenced by recent work [16, 17, 22]. Researchers have explored the combination of a fixed Language Model (LLM) with a vision model and a small set of adaptable modules, a technique known as “visual instruction tuning” [21, 32]. This technique aims to refine the adaptable modules to enhance the ability of LLMs to generate textual descriptions for video content. While these models are deployed in systems for answering questions about videos [9, 29, 30], the majority of them focus on short video clips or images. In this paper, we aim to enhance the runtime efficiency of the video-to-text extraction process for long videos within a retrieval-augmented generation (RAG) [13] based system.

5. Conclusion and Future Work

In this paper, we introduce ViTA , a novel algorithm designed to accelerate the conversion of videos to text within a retrieval augmented generation (RAG) framework for video processing system. By capitalizing on the characteristics of vision language models (VLMs) output generation, ViTA restricts the number of tokens to reduce latency during conversion. To preserve accuracy, ViTA utilizes a lightweight VLM to initially extract information from video clips. Subsequently, it instructs the heavyweight VLM to leverage the text generated by the lightweight model for extracting additional details. Supported by empirical evidence, we conclude that employing a cascading approach, where the lightweight model extracts initial information and guides the heavyweight VLM in further extraction, ViTA significantly reduces the latency of the video-to-text conversion process.

While ViTA employs a lightweight VLM to extract general scene information, there may be instances where both lightweight and heavyweight VLM outputs convey identical information. In such cases, ViTA can be further optimized to avoid invoking the heavyweight VLM for generating additional information redundantly. Therefore, intelligent prompt engineering plays a crucial role in minimizing these redundant invocations of the heavyweight VLM. We plan to address this issue in our future work.

References

[1] Pyscenedetect: Video scene cut detection and analysis tool. https://www.scenedetect.com/, 2023. 3, 4

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, 2021. 7

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021. 7

[4] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023. 1

[5] Biplob Debnath, Oliver Po, Farhan Asif Chowdhury, and Srimat Chakradhar. Cosine similarity based few-shot video classifier with attention-based aggregation. In *ICPR*, 2022. 7

[6] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024. 2, 3, 4

[7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *CVPR*, 2019. 7

Table 6. Query-response from Tokyo MODI video dataset. Context is retrieved using textual data implemented with RAG.

Q1	Is there a large Sony billboard?
Baseline	**Response:**Yes, there is a large Sony billboard displayed on the side of a building in the bustling city scene. Additionally, there is a large billboard advertising Sony in the background of the street scene. The image captures a bustling city street from an elevated perspective, with numerous people walking and vehicles driving in front of a Sony building.
ViTA	**Response:** Yes, there is a large Sony billboard in the street intersection scene.
Q2	Is there any car seen making a left turn?
Baseline	**Context:** 00:44:55.067 The image captures a bustling city street scene, viewed from an elevated perspective. The street is filled with numerous pedestrians and vehicles, including cars, buses, and motorcycles. A white bus can be seen driving on the right side of the road, . 00:28:25.067 The image captures a bustling city street scene from an aerial perspective. Numerous people are walking on the sidewalks, and some are crossing the street at crosswalks. Cars can be seen driving along the road, with one car making a left . 00:48:30.000 The image captures a busy city street scene with numerous people and vehicles. There are two buses visible, one on the left side of the road and another towards the center-right. A car is also present in the middle of the scene. . 00:49:00.000 The image captures a bustling city street scene from an aerial perspective. The street is filled with numerous people walking and crossing the road, while several cars are also present, including buses and vans. A white car can be seen driving on the right . **Response:** Yes, a car can be seen making a left turn at 00:28:25.067 in the video clip.
ViTA	**Context:** 00:12:00.067 This image captures a scene from a street intersection showing a police car is seen in this image. Moreover, The image captures a busy street intersection in an urban area. A police car is visible on the right side of the road, . 00:05:15.067 This image captures a scene from a street intersection showing a street scene with people walking and cars. Moreover, there is a bus crossing the street, and a police car is present. The image also features several traffic lights and billboards . 00:44:30.067 This image captures a scene from a street intersection showing a street scene with people walking and cars. Moreover, there are several vehicles, including cars and buses, moving through the intersection. Pedestrians can be seen crossing the street at . 00:37:30.000 This image captures a scene from a street intersection showing a street scene with people walking and cars. Moreover, The image also features a large Sony billboard on the right side, and there are several traffic lights visible at various points in . **Response:** There is no specific mention of a car making a left turn in the provided context. Therefore, it is not possible to confirm if a car was seen making a left turn.
Q3	Describe the buildings near the street.
Baseline	**Context:** 00:54:40.067 The image captures a bustling city street scene, viewed from an elevated perspective. The street is teeming with people crossing the road and walking on the sidewalks. Numerous vehicles, including cars and buses, are navigating through the area. Build . 00:43:45.067 The image captures a bustling city scene from an aerial perspective. The street is filled with vehicles, including cars and buses, navigating through the area. Pedestrians can be seen walking on the sidewalks, going about their day. Buildings line . 00:40:20.000 The image captures a bustling city street scene, viewed from an elevated perspective. The street is filled with numerous pedestrians and vehicles, including cars and buses. A large Sony building can be seen on the right side of the frame, adding to the urban . 00:38:20.067 The image captures a bustling city street scene, viewed from an elevated perspective. The street is teeming with people crossing the road and walking on the sidewalks. Numerous vehicles, including cars and buses, are navigating through the area. Build . **Response**: The buildings in the area are not specifically described in the provided context.
ViTA	**Context:** 01:03:05.067 This image captures a scene from a street intersection showing a street scene with people walking and cars. Moreover, there are several cars, a bus, and a van on the road. The street is lined with tall buildings, including a . 01:59:15.067 This image captures a scene from a street intersection showing a street scene with people walking on the sidewalk. Moreover, there are several vehicles, including cars and buses, on the road. The street is lined with tall buildings, some of which . 01:16:50.067 This image captures a scene from a street intersection showing a street scene with people walking on the sidewalk. Moreover, there are several vehicles, including cars and buses, on the road. The street is lined with tall buildings, some of which . 01:56:20.067 This image captures a scene from a street intersection showing a busy street with people walking and cars. Moreover, there are several buildings, including a Sony store and a Modi building. The street is bustling with activity, featuring numerous . **Response**: The buildings are tall and line the street, with some of them including a Sony store and a Modi building. The street scene shows a bustling area with people walking and cars.

[8] Harrison Chase. LangChain. https://github.com/hwchase17/langchain, 2022. 2022-10-17. 4, 5

[9] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. Promptcap: Prompt-guided image captioning for vqa with gpt-3. In *CVPR*, 2023. 7

[10] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3), 2019. 5

[11] Ferdi Kossmann, Ziniu Wu, Eugenie Lai, Nesime Tatbul, Lei Cao, Tim Kraska, and Sam Madden. Extract-transform-load for video streams. *Proceedings of the VLDB Endowment*, 16 (9), 2023. 1

[12] Ferdinand Kossmann, Ziniu Wu, Eugenie Lai, Nesime Tatbul, Lei Cao, Tim Kraska, and Samuel Madden. Extract-transform-load for video streams. *arXiv preprint arXiv:2310.04830*, 2023. 4

[13] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*, 2020. 1, 7

[14] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2, 3, 4, 5

[15] Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatiotemporal representation learning. In *ICLR*, 2022. 7

[16] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 7

[17] Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. Lavender: Unifying video-

language understanding as masked language modeling. In *CVPR*, 2023. 7

[18] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, 2019. 7

[19] Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuan Yang, Jianfeng Wang, Lin Liang, Zicheng Liu, Yumao Lu, Ce Liu, and Lijuan Wang. MM-VID: Advancing Video Understanding with GPT-4V(ision). *arXiv preprint arXiv:2310.19773*, 2023. 1

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 3

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 7

[22] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 7

[23] Usha Mittal, Priyanka Chawla, and Rajeev Tiwari. Ensemblenet: A hybrid approach for vehicle detection and estimation of traffic density based on faster r-cnn and yolo models. *Neural Computing and Applications*, 35(6), 2023. 1

[24] OpenAI. `https://openai.com/product`, 2023. Accessed: July 11, 2023. 5

[25] Yurii Piadyk, Joao Rulff, Ethan Brewer, Maryam Hosseini, Kaan Ozbay, Murugan Sankaradas, Srimat Chakradhar, and Claudio Silva. Streetaware: A high-resolution synchronized multimodal urban scene dataset. *Sensors*, 23(7), 2023. 2, 4

[26] Yuto Shibata, Yutaka Kawashima, Mariko Isogawa, Go Irie, Akisato Kimura, and Yoshimitsu Aoki. Listening human behavior: 3d human pose estimation with acoustic signals. In *CVPR*, 2023. 1

[27] Showlab. VLog: Video As a Long Document. `https://github.com/showlab/VLog`, 2023. GitHub Repository. 1

[28] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. Video understanding with large language models: A survey, 2024. 1

[29] Yanan Wang, Michihiro Yasunaga, Hongyu Ren, Shinya Wada, and Jure Leskovec. Vqa-gnn: Reasoning with multimodal knowledge via graph neural networks for visual question answering. In *CVPR*, 2023. 7

[30] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. 7

[31] Yanyi Zhang, Xinyu Li, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe. Vidtr: Video transformer without convolutions. In *ICCV*, 2021. 7

[32] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023. 7