

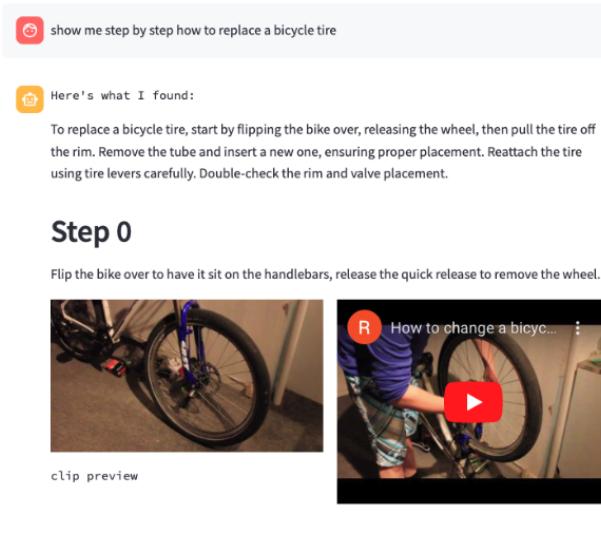
Video Enriched Retrieval Augmented Generation Using Aligned Video Captions

Kevin Dela Rosa

Snap Inc.

Santa Monica, California, USA

kevd1337@gmail.com



Step 0

Flip the bike over to have it sit on the handlebars, release the quick release to remove the wheel.



[clip preview](#)



Step 1

Pull the tire up and over the edge of the rim to remove it, release the tube to take it out.



[clip preview](#)

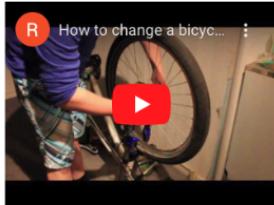


Figure 1: Sample output of chat assistant app leveraging retrieved video content stored in the form of aligned video captions

ABSTRACT

In this work, we propose the use of "aligned visual captions" as a mechanism for integrating information contained within videos into retrieval augmented generation (RAG) based chat assistant systems. These captions are able to describe the visual and audio content of videos in a large corpus while having the advantage of being in a textual format that is both easy to reason about & incorporate into large language model (LLM) prompts, but also

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

MRR 2024, July 18, 2024, Washington, DC

© 2024 Copyright held by the owner/author(s).

Step 2

Insert the new tube carefully, ensuring the air nozzle is correctly in place, and fit the tire back over the rim.



[clip preview](#)



Step 3

Verify the rim, place the tire back on by leveraging technique with tire levers, making sure the valve comes out last.



[clip preview](#)



Sources:

1. [How to change a bicycle tire / flat tire](#)
2. [Cycling Tips How To Change A Tire Like A Pro](#)

CCS CONCEPTS

- Information systems → Information retrieval;
- Computing methodologies → Visual content-based indexing and retrieval.

KEYWORDS

Retrieval Augmented Generation, Cross-modal Retrieval, Multi-modal Retrieval, Large Language Model Applications, Chatbots

ACM Reference Format:

Kevin Dela Rosa. 2024. Video Enriched Retrieval Augmented Generation Using Aligned Video Captions. In *Proceedings of the 2024 SIGIR Workshop on Multimodal Representation and Retrieval (MRR 2024)*. ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION

Video content in the forms of YouTube shorts, TikToks, Instagram Reels or the like are quickly becoming many people's main form of content ingestion online. At the same time, following the initial deluge of work on large language models there has been a recent surge in work to understand videos, with big corporate systems like OpenAI GPT-4 Vision and Google Gemini incorporating basic image and video chatting capabilities into their chat AI applications, as well as academic systems like [6] [5], [10], [4].

Surprisingly while there are many works that touch on video understanding at various levels, there have been relatively few works that have used videos in a retrieval augmented generation (RAG) [2] context; some notable related works include EgoInstructor [9] where the authors introduce a retrieval augmented multimodal captioning model that retrieves relevant exocentric videos as references to generate the captions for egocentric videos, also in [7] they use retrieved text to generate answers for questions about an input video. Additionally in [11] the authors improve multimodal query (image + text) to image retrieval using a large scale (query image, instruction, target image) triplet dataset. Those respective applications are great, but in this work we focus on bringing the context of a large video corpus itself into responses of a retrieval augmented generation chat bot setting.

One potential reason for relatively few works in this space is that accessing videos in a large scale can be a daunting engineering endeavor, given video information's relative large size and multimodal nature. Another potential reason for a lack of related works can be attributed to the relative difficulty of collecting video data, and further compounded by the time consuming effort of evaluating the retrieval and generation stages' output manually.

In this work, we propose the use of "aligned visual caption" transcripts (see example in Figure 2) in the context of a chat assistant. In Section 2 we detail the process of preparing aligned video captions, then describe a video data set we curated for this work, and provide commentary on how these compare to using different signals from videos in conjunction with popular LLMs under the task of video summarization as a proxy for the model's capabilities in general video understanding. Then in Section 3 we describe an experiment that aims to automatically measure feasibility of using these transcripts in a retrieval augmented generation context. Then in Section 4 we describe a sample AI chat application architecture that leverages the aligned video caption representation of videos to illustrate the ease of integration. For sample demo application, LLM prompts, evaluation scripts and dataset pointers, see: <https://github.com/kdr/videoRAG-mrr2024>

2 ALIGNED VIDEO CAPTIONS

"Aligned Video Caption Transcripts" are temporally synced scene descriptions of a video in the form of machine generated visual captions and the associated subtitles or automatic speech recognition transcripts. In this study we curated a dataset based on public youtube videos sampled from Panda-70M [1], which contains individual clip segments and a general visual scene caption learned from a set of open source video captioning models. Specifically we sampled roughly 2,000 videos from each YouTube category present in Panda-70M [1], resulting in a dataset of 29,259 videos (1.5M video clips and corresponding visual captions) or roughly 215 days of footage. We then augment that dataset with subtitles gathered directly from YouTube's APIs and created the aligned transcripts as seen in Figure 2). General statistics shown in Table 1.

In order to verify that the information an LLM can generate from an aligned video caption transcript is roughly comparable to that of a multimodal LLM, as a sanity check we checked how semantically similar video summarizations generated by various LLMs were to those generated by GPT-4 Turbo using the aligned video caption transcript. We compared these generated summaries using BERTScore [12], which is an automatic summarization measure has been shown to correlate with human judgment on sentence-level and system-level evaluation. A total of 1.5K videos were summarized and evaluated, sampled uniformly from the original dataset.

In Table 2 we can see that the various configurations correlate significantly with the GPT-4 based ground truth. In particular we see that sending raw video frames and the automatic speech recognition (ASR) transcript to the GPT-4 scores a high BERTScore; so does the text only based settings using ASR, suggesting much of the information that the LLM is able to tap into resides in speech. Additionally we see that summarizations using Gemini 1.5 Pro with video based input and GPT 4 using video frames (i.e. first frame per scene) as input have similar scores as well, showing that these captions can produce similar quality output with having to send the entire set of frames to the LLM, greatly saving on context window and processing bandwidth at query time. For example, if the entire aligned video caption dataset was sampled at 1 frame per second (as is the case for popular LLMs like Gemini 1.5 pro) and assuming an image is resized to roughly fit the cost of 256 tokens for the LLM, you're looking at around 4.8 billion tokens including subtitles (roughly 69x bigger compared to using aligned visual captions). Note that Gemini 1.5 pro did not process audio signals in videos at the time this study was conducted.

3 VIDEO RETRIEVAL AUGMENTED GENERATION

In this section we determine if text embeddings over video derived data is feasible input for retrieval augmented generation, over the task of answering a provided general knowledge question using answers found in videos as support. In this experiment we use 1000 general knowledge questions generated via GPT 4 V as input to an embedding extractor. We also compare retrieval results using two multimodal embeddings, namely BLIP-2's [3] image feature extractor and CLIP [8] embeddings (ViT-L/14@336px). Then we retrieve the top K results as determined by a simple cosine similarity

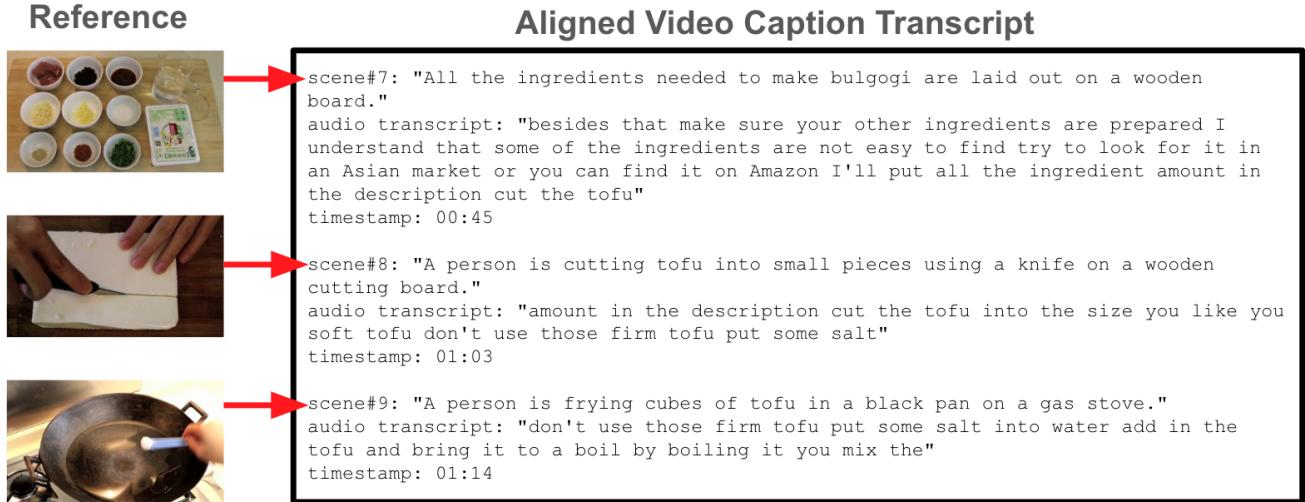


Figure 2: Sample aligned video caption transcript with corresponding example video frames from source scenes

| DATASET DIMENSION | TOTAL | MEDIAN |
|------------------------------|-------------|----------|
| Video Count | 29,259 | - |
| Scene Count | 1,476,462 | 31.00 |
| Video Duration (seconds) | 18,584,396 | 478.00 |
| Text Character Length | | |
| Title | 1,548,810 | 51.00 |
| Description | 30,565,705 | 780.00 |
| Title + Description | 32,114,515 | 833.00 |
| Visual Video Captions | 96,888,187 | 2,016.00 |
| Subtitles / ASR | 141,926,062 | 3,472.00 |
| Aligned Captions | 276,019,918 | 6,461.00 |

Table 1: Statistics for Aligned Video Caption Dataset

| LLM | PROMPT CONTEXT | BERT |
|---------------------------------------|-------------------------------|-------|
| Multimodal LLMs | | |
| GPT 4 V | Video Frames + ASR Transcript | 0.889 |
| Gemini 1.5 Pro | Original Video | 0.862 |
| GPT4 V | Video Frames | 0.860 |
| GPT 4 Turbo Varying Text Input | | |
| GPT 4 | ASR Transcript | 0.893 |
| GPT 4 | Visual Captions | 0.869 |
| GPT 4 | Title + Description | 0.858 |

Table 2: Generated video summary comparison against GPT 4 aligned visual captions based generation

against the entire 29K video dataset. Using the top K results we use GPT-4 as an automatic judge using the following metrics:

- **HIT@K**: in the top K retrieved results, does any retrieved document contain the information required to answer the posed question. We use this in lieu of recall given the difficulty of manually collecting ground truth over every video (also answers are free form sentences and can't simply be checked for existence via basic string comparisons)
- **QUALITY@1**: answer correctness / quality rating between 1-10, measuring quality of answers generated by GPT-3.5 turbo. In order to control for compounding factors due to the provided context in the LLM prompt, all answers were generated using the aligned video caption transcript of the retrieved result regardless of retrieval method

To generate the questions we first sampled 500 videos from the dataset, then provided the aligned video captions as context to GPT 4 and asked the LLM to generate general knowledge questions that the video could help answer but are not specifically tied to the source video, and from the resulting question set we uniformly sampled 1000 questions.

In Table 3 we can see that the text embeddings are able to find hits at a relatively low K using the aligned transcript and ASR. We also see that the relevance of results at very low K suffers for the cross-modal embedding configuration, but can ultimately catch up if you have a tolerance for higher K, i.e. LLM can handle processing more retrieved documents in its context window, which is encouraging for future extensions into multimodal querying.

4 VIDEO ENRICHED CHAT BOT

In Figure 3 we illustrate the main components in a RAG based AI chat bot application that leverages the aligned video captions to return relevant answers and corresponding video clip source. The processing is as follows:

- (1) Based on the user query and a choice of tool descriptions, one retriever tool is selected; we created different tools that point to specific subsets of the video catalog

| EMBEDDING | DATABASE | HIT@1 | HIT@5 | HIT@10 | QUALITY@1 |
|--|---------------------|-------|-------|--------|-----------|
| Multimodal Embeddings: Cross-modal Text to Vision Match | | | | | |
| BLIP-2 | Video Frames | 0.482 | 0.801 | 0.895 | 5.199 |
| BLIP-2 | Video Thumbnail | 0.519 | 0.833 | 0.902 | 5.598 |
| CLIP ViT-L/14@336px | Video Frames | 0.542 | 0.858 | 0.925 | 5.785 |
| CLIP ViT-L/14@336px | Video Thumbnail | 0.553 | 0.859 | 0.926 | 5.889 |
| Text Embeddings | | | | | |
| text-embedding-3-small | ASR | 0.741 | 0.936 | 0.969 | 7.424 |
| text-embedding-3-small | Visual Captions | 0.65 | 0.878 | 0.932 | 6.605 |
| text-embedding-3-small | Title | 0.629 | 0.905 | 0.95 | 6.503 |
| text-embedding-3-small | Title + Description | 0.675 | 0.914 | 0.95 | 6.828 |
| text-embedding-3-small | Aligned Transcript | 0.741 | 0.934 | 0.971 | 7.377 |

Table 3: Video retrieval results and average quality of answer generated using aligned visual action of top retrieved document

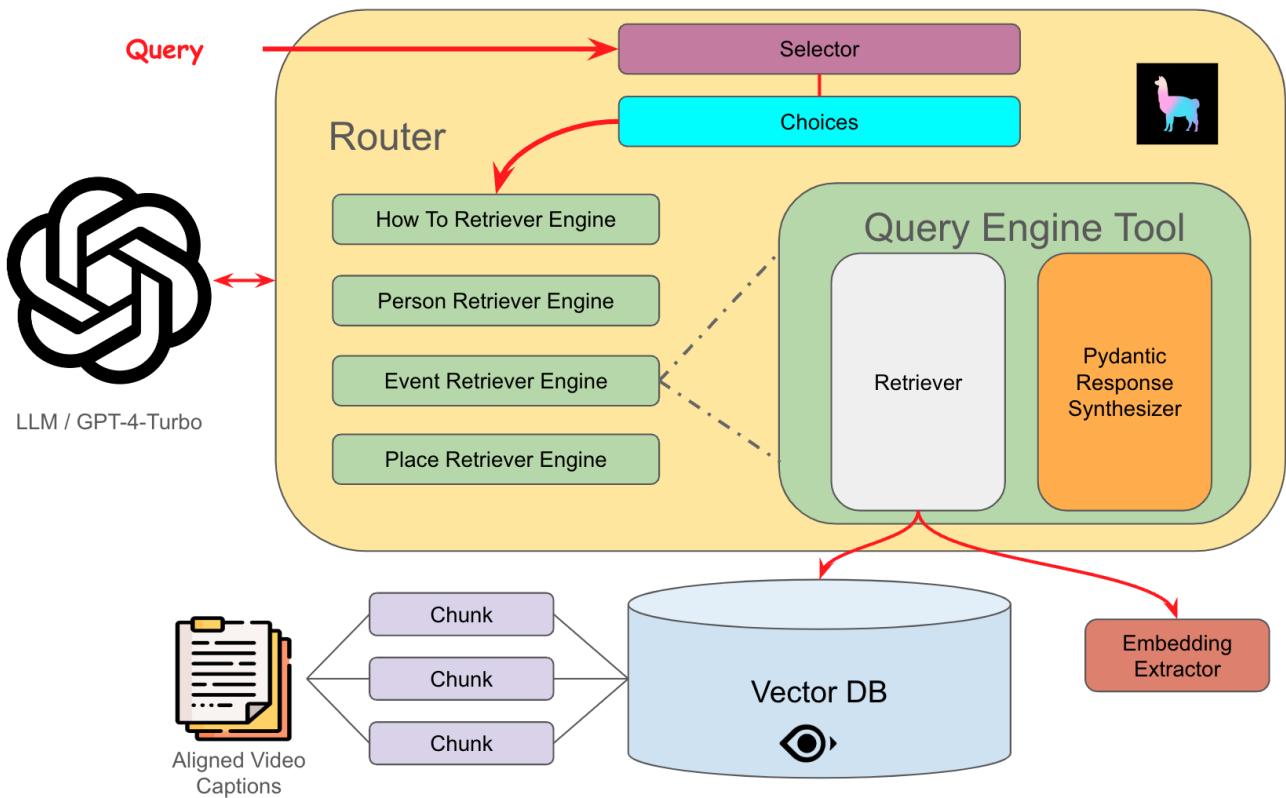


Figure 3: Example application architecture for integrating aligned video captions to enabled video enriched RAG

- (2) The selected query engine tool vectorizes the query and searches the vector database to retrieve (chunked) aligned video caption text blobs.
- (3) The query engine tool interprets the results and summarizes into a specific pydantic format customized for that answer type; for example a "how to" response should respond with a bulleted list of steps like in Figure 1, whereas a "place"

response would describe a location and why it is notable. Timestamps in retrieved docs help give the application pointers to specific parts of video to enhance user interaction

5 CLOSING REMARKS

In this study we show that aligned visual captions provide a compelling and adaptable representation of video information that can

easily plug into basic LLM application architectures. We curate a large scale dataset, demonstrate how to leverage the data representation to generate questions, and offer an automated procedure for measuring video RAG based question answering results.

This work gives us a glimpse into the potential of using aligned video captions representation, and is ripe for future exploration. For example, a key practical consideration in deploying this solution in the real world is the availability of video captioning models suitable for the intended use case. Another thing to consider is how one identifies meaningful video clip segments to be summarized by the captioning models in the first place. In future works it would be interesting to study how generic video captioning and clip segmentation methods fare on different video domains (e.g. general knowledge vs. surveillance, etc.) and contrast those with strategies that adapt the various components in the processing pipeline for a target domain. Moreover, the audio signal incorporated in this study focuses on the spoken word, and for other domains it may be interesting to incorporate other aspects of the audio as well (e.g. descriptions of music or shifts in loudness could give hints to the overall tone or emotion of the scene).

REFERENCES

- [1] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jia Ren, Ming-Hsuan Yang, and Sergey Tulyakov. 2024. Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers. *arXiv preprint arXiv:2402.19479* (2024).
- [2] Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *ArXiv abs/2005.11401* (2020). <https://api.semanticscholar.org/CorpusID:218869575>
- [3] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *ICML*.
- [4] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhui Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. VideoChat: Chat-Centric Video Understanding. *arXiv preprint arXiv:2305.06355* (2023).
- [5] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-LLAVa: Learning United Visual Representation by Alignment Before Projection. *arXiv preprint arXiv:2311.10122* (2023).
- [6] Muhammad MaaZ, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. *arXiv:2306.05424* (2023).
- [7] J. Pan, Z. Lin, Y. Ge, X. Zhu, R. Zhang, Y. Wang, Y. Qiao, and H. Li. 2023. Retrieving-to-Answer: Zero-Shot Video Question Answering with Frozen Large Language Models. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. IEEE Computer Society, Los Alamitos, CA, USA, 272–283. <https://doi.org/10.1109/ICCVW60793.2023.00035>
- [8] 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. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*. <https://api.semanticscholar.org/CorpusID:231591445>
- [9] Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yue Zhang, Rui Feng, and Weidi Xie. 2024. Retrieval-Augmented Egocentric Video Captioning. *ArXiv abs/2401.00789* (2024). <https://api.semanticscholar.org/CorpusID:266693245>
- [10] Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. *arXiv preprint arXiv:2306.02858* (2023). <https://arxiv.org/abs/2306.02858>
- [11] Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhui Chen, Yu Su, and Ming-Wei Chang. 2024. MagicLens: Self-Supervised Image Retrieval with Open-Ended Instructions. *arXiv preprint arXiv:2403.19651* (2024).
- [12] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=SkeHuCVFDr>