Video-Enriched Retrieval Augmented Generation Using Aligned Video Captions

Kevin Dela Rosa Aviary Inc. San Francisco, CA, USA kdr@aviaryhq.com

show me step by step how to replace a bicycle tire

Here's what I found:

To replace a bicycle tire, start by flipping the bike over, releasing the wheel, then pull the tire off the rim. Remove the tube and insert a new one, ensuring proper placement. Reattach the tire using tire levers carefully. Double-check the rim and valve placement.

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

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

Figure 1: Step-by-step bicycle repair instructions generated by the video-enriched AI assistant using aligned video captions

Abstract

In this work, we propose the use of "aligned video captions" as an intelligent mechanism for integrating video content into retrieval augmented generation (RAG) based AI assistant systems. These captions serve as an efficient representation layer between videos and large language models (LLMs), describing both visual and audio content while requiring significantly less context window space compared to traditional frame sampling approaches. We demonstrate how this representation enables more effective agent-based retrieval and generation capabilities, with captions that can be dynamically adapted through targeted prompting or finetuning of the underlying models. Our empirical evaluation across multiple LLM configurations shows that this approach achieves comparable performance to direct video processing while being more computationally efficient and easier to reason about in downstream tasks. Notably, the approach shows particular strength in procedural content like How-To videos, where aligned captions significantly outperform speech-only baselines.

WWW Companion '25, Sydney, NSW, Australia

^{© 2025} Copyright held by the owner/author(s). Publication rights licensed to ACM. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2,* 2025, Sydney, NSW, Australia, https://doi.org/10.1145/3701716.3716890.

CCS Concepts

• Information systems \rightarrow Information retrieval; • Computing methodologies \rightarrow Visual content-based indexing and retrieval.

Keywords

Multimodal Retrieval, Chatbots, Agentic Information Retrieval

ACM Reference Format:

Kevin Dela Rosa. 2025. Video-Enriched Retrieval Augmented Generation Using Aligned Video Captions. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3701716. 3716890

1 Introduction

Video content, particularly in the form of short-form videos across platforms like YouTube shorts, TikToks, Instagram Reels, and Snapchat Spotlights, has become the primary medium for online information consumption. Concurrent with this shift, AI agents powered by large language models have made significant strides in video understanding, with systems like OpenAI GPT-4 Vision and Google Gemini incorporating multimodal capabilities into their chat interfaces. However, the integration of video content into retrieval augmented generation (RAG) [2] architectures remains relatively unexplored, particularly in the context of intelligent agents that can effectively search and reason over large video corpora.

While several approaches have emerged for video understanding and retrieval, existing solutions often struggle with the computational demands and architectural complexity of processing video data at scale. Notable attempts include EgoInstructor [5], which uses retrieved exocentric videos to guide egocentric captioning, and approaches that leverage multimodal queries for enhanced retrieval [6]. However, these systems typically require extensive computational resources to process video frames and maintain complex cross-modal architectures.

In this work, we introduce an agent-centric approach using "aligned video caption" transcripts as an efficient intermediary between videos and language models. This representation enables AI agents to perform sophisticated retrieval and reasoning tasks over video content while significantly reducing the computational overhead typically associated with video processing. We demonstrate the effectiveness of this approach through:

- A scalable architecture for video-enriched RAG that enables efficient retrieval and reasoning over large video corpora
- An empirical evaluation showing comparable performance to direct video processing approaches while requiring substantially less computational resources
- A practical implementation demonstrating how this representation can be integrated into interactive AI assistants

Our code, LLM prompts, and dataset will be made available at https://github.com/kdr/avc

2 Aligned Video Captions

"Aligned Video Caption Transcripts" provide temporally synchronized scene descriptions combining machine-generated visual captions with associated speech recognition transcripts. We curated a dataset based on public YouTube videos sampled from Panda-70M [1], collecting approximately 2,000 videos per YouTube category for a total of 29,259 videos spanning 215 days of footage. The dataset comprises 1.5M video clips with corresponding visual captions, synchronized with subtitles from YouTube's APIs (see Table 1 for key statistics).

Metric	Count
Videos	29,259
Total Scene Count	1,476,462
Visual Captions (chars)	96.9M
ASR Transcript (chars)	141.9M
Aligned Video Caption Transcript (chars)	276.0M

Table 1: Video Statistics and Transcript Characteristics

To validate our approach, we compared video summarization capabilities across different LLM configurations using BERTScore [7] to evaluate semantic similarity. Testing across 1.5K uniformly sampled videos revealed that aligned video caption transcripts enable performance comparable to direct video processing methods (Table 2).

LLM	Input	BERTScore
GPT-4 V	Video + ASR	0.889
Gemini 1.5 Pro	Full Video	0.862
GPT-4	ASR Only	0.893
GPT-4	Visual Captions	0.869

 Table 2: Summarization Performance vs GPT-4 with Aligned

 Video Captions

Notably, our approach achieves comparable performance while being more efficient - traditional 1fps sampling would require 4.8B tokens for our dataset, whereas our aligned video captions need 69x fewer tokens while maintaining high-quality results.

3 Video Retrieval Augmented Generation

To evaluate our agent-based retrieval system, we assess whether text embeddings over video-derived data enable effective retrieval augmented generation, particularly for answering general knowledge questions. Our framework combines GPT-4 as the reasoning engine with various embedding approaches for retrieval.

The experiment uses 1000 general knowledge questions generated via GPT-4V, comparing retrieval performance across different embedding strategies. We evaluate both traditional text embeddings and multimodal approaches, specifically:

- Text embeddings via text-embedding-3-small over various text representations
- Cross-modal embeddings using BLIP-2's [3] image feature extractor
- CLIP [4] embeddings (ViT-L/14@336px)

For each approach, we retrieve the top K results using cosine similarity against our 29K video dataset. The retrieval agent's performance is evaluated using two key metrics:

Reference	Aligned Video Caption Transcript
	<pre>scene#7: "All the ingredients needed to make bulgogi are laid out on a wooden board." audio transcript: "besides that make sure your other ingredients are prepared I understand that some of the ingredients are not easy to find try to look for it in an Asian market or you can find it on Amazon I'll put all the ingredient amount in the description cut the tofu" timestamp: 00:45 scene#8: "A person is cutting tofu into small pieces using a knife on a wooden cutting board." audio transcript: "amount in the description cut the tofu into the size you like you</pre>
	<pre>soft tofu don't use those firm tofu put some salt" timestamp: 01:03 scene#9: "A person is frying cubes of tofu in a black pan on a gas stove." audio transcript: "don't use those firm tofu put some salt into water add in the tofu and bring it to a boil by boiling it you mix the" timestamp: 01:14</pre>

Figure 2: Example of aligned video caption transcript showing temporal synchronization between visual scenes and speech transcripts in a cooking tutorial

- HIT@K: Measures the agent's ability to retrieve relevant information within the top K results. This metric is particularly important for agent systems that need to efficiently locate supporting evidence.
- QUALITY@1: A 1-10 rating assessing answer quality generated by GPT-3.5 turbo using the retrieved context. For consistency, all answers are generated using the aligned video caption transcript regardless of retrieval method.

The question dataset was built by sampling 500 videos and using GPT-4 with aligned video captions to generate generalizable questions that demonstrate content understanding.

Table 3 shows that ASR performs similarly to aligned transcripts in Overall retrieval (both achieving HIT@1 rates of 0.741 and comparable QUALITY@1 scores of 7.424 and 7.377 respectively). However, for How-To videos subset of data, where both visual and spoken content are critical, aligned transcripts show more substantial gains (QUALITY@1: 8.138 vs 7.385, HIT@1: 0.785 vs 0.708). This suggests that while speech dominates general video understanding, the integration of visual context becomes particularly valuable for procedural content and suggests further study is needed to understand impact in other domains. The cross-modal configurations catch up at higher K values (HIT@5: 0.908-0.923), suggesting promise for multimodal querying as LLM context windows expand.

4 Video-Enriched AI Assistant

Our implementation demonstrates how aligned video captions enable sophisticated video retrieval and interaction. Figure 3 illustrates the architecture of an AI assistant that combines retrieval capabilities with contextual understanding to provide video-enriched responses. The system leverages an agentic LLM as its core reasoning engine, which coordinates a set of specialized tools to process user queries. When a query is received, the system uses a routing tool to analyze the intent and select the appropriate retrieval strategy (e.g., How-To, Person, Event, or Place retriever). The Query Engine Tool then vectorizes this query and performs a semantic search over the aligned captions, considering both content similarity and temporal relationships within videos. Retrieved information is structured using a Pydantic Response Synthesizer, which ensures consistent formatting and maintains semantic links to video timestamps.

This architecture enables contextually appropriate responses tutorial queries receive timestamped step-by-step instructions (as shown in Figure 1), while descriptive queries generate summaries with relevant video reference points. The system adapts its retrieval and presentation strategies based on query type, demonstrating how aligned video captions can support sophisticated human-AI interaction around video content.

5 Closing Remarks

Our study demonstrates how aligned video captions can serve as an efficient foundation for building intelligent retrieval agents over video content. By providing a structured representation that bridges visual content and language models, this approach enables sophisticated retrieval and reasoning capabilities while maintaining computational efficiency. The empirical results show that aligned video captions not only reduce computational overhead but also support flexible agent architectures that can adapt to different query types and user interaction patterns.

Looking forward, key areas for exploration include domainspecialized captioning agents, multi-agent architectures for collaborative understanding, and enhanced temporal reasoning in video-based dialogue systems. Additionally, exploring how agents can leverage both speech patterns and non-verbal audio cues could WWW Companion '25, April 28-May 2, 2025, Sydney, NSW, Australia

EMBEDDING	DATABASE	Overall				How To			
		QUALITY@1	HIT@1	HIT@5	HIT@10	QUALITY@1	HIT@1	HIT@5	
Multimodal Vision + Text Embeddings									
BLIP-2	Video Frames	5.199	0.482	0.801	0.895	5.569	0.477	0.846	
BLIP-2	Video Thumbnail	5.598	0.519	0.833	0.902	6.754	0.631	0.846	
CLIP ViT-L/14@336px	Video Frames	5.785	0.542	0.858	0.925	6.523	0.631	0.908	
CLIP ViT-L/14@336px	Video Thumbnail	5.889	0.553	0.859	0.926	6.431	0.600	0.923	
Text Embeddings									
embedding-3-small	ASR	7.424	0.741	0.936	0.969	7.385	0.708	0.954	
embedding-3-small	Visual Captions	6.605	0.650	0.878	0.932	7.723	0.754	0.908	
embedding-3-small	Title	6.503	0.629	0.905	0.950	6.354	0.585	0.908	
embedding-3-small	Title + Description	6.828	0.675	0.914	0.950	7.431	0.708	0.923	
embedding-3-small	Aligned Transcript	7.377	0.741	0.934	0.971	8.138	0.785	0.954	

Table 3: Comparison of retrieval performance and answer quality across different embedding approaches using answers generated with aligned video caption of the top retrieved document. Results shown for both the overall AVC dataset and How-To video subset demonstrate the increased impact of visual understanding in specific domains.



Figure 3: System architecture for video-enriched retrieval augmented generation using aligned video captions

enhance their ability to understand and communicate video context, particularly in interactive scenarios where emotional or tonal understanding is crucial. Video-Enriched Retrieval Augmented Generation Using Aligned Video Captions

References

- [1] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiangwei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. 2024. Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
- [2] 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. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Proceedings of the 34th International Conference on Neural Information Processing Systems (Vancouver, BC, Canada) (NIPS '20). Curran Associates Inc., Red Hook, NY, USA, Article 793, 16 pages.
- [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] 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

- [5] Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yue Zhang, Rui Feng, and Weidi Xie. 2024. Retrieval-Augmented Egocentric Video Captioning. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024), 13525–13536. https://api.semanticscholar.org/CorpusID:266693245
- [6] Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. 2024. MagicLens: Self-Supervised Image Retrieval with Open-Ended Instructions. In Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235), Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (Eds.). PMLR, 59403–59420. https://proceedings. mlr.press/v235/zhang24an.html
- [7] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net. https://openreview.net/forum?id=SkeHuCVFDr