



Video Enriched Retrieval Augmented Generation Using Aligned Video Captions

Kevin Dela Rosa, Snap Inc.

1st Workshop on Multimodal Representation and Retrieval @ SIGIR 2024
July 18, 2024





Introduction

Observations

- **Retrieval augmented generation (RAG)** is being widely adopted in LLM applications as a way to incorporate knowledge sources
- **Video is everywhere**, rising as primary content ingestion method
- **Surprisingly few** works integrate videos in RAG

Video can be intimidating / hard

- Large storage size
- Multimodality, lots of information
- Time consuming to collect database and evaluate methods

Goal

- Incorporate video content into RAG chat bots via **aligned video captions**

Reference

Aligned Video Caption Transcript



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 soft tofu don't use those firm tofu put some salt"

timestamp: 01:03

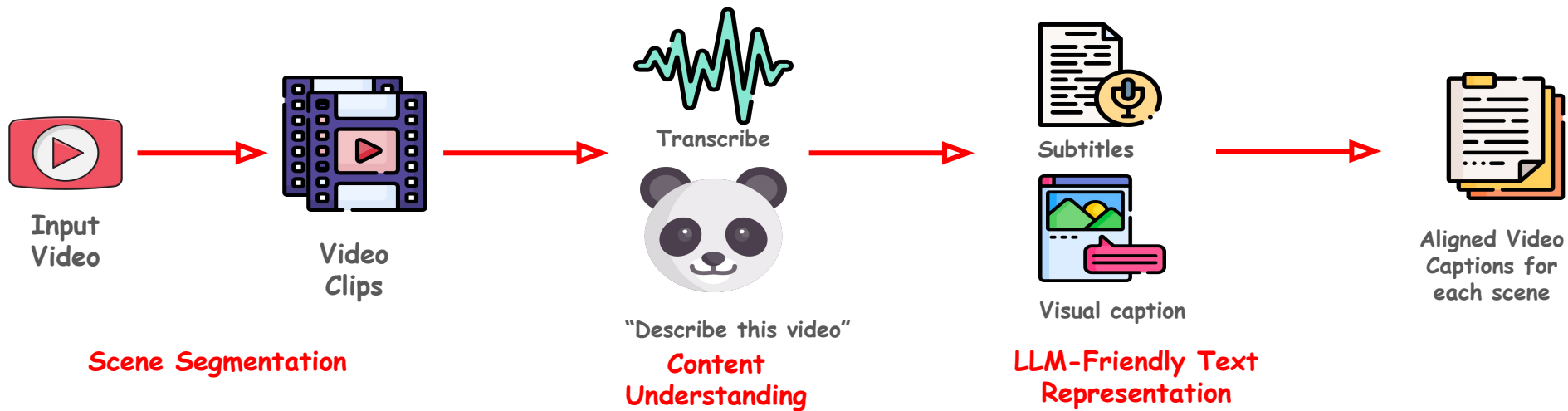


scene#9: "A person is frying cubes of tofu in a black pan on a gas stove."

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

Aligned Video Captions: Temporally synced scene descriptions combining visual captions and subtitles



Aligned Video Captions: General process of collecting aligned video captions. Content understanding steps can be adapted to specific use case (e.g. fine tune or prompt a video captioner for specific information)



Aligned Video Captions Dataset V1

Dataset

- Collected sample of clip segments and visual captions from [Panda-70M](#)
- YouTube videos from [HD-VILA-100M](#)

High Level Statistics

- 29,529 videos, 1.5M clips
- ~215 days of footage
- 1000 general knowledge questions



Video Understanding Feasibility

Verified information an LLM can generate using aligned video captions is roughly comparable to that of a multimodal LLM by comparing video summarizations.

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

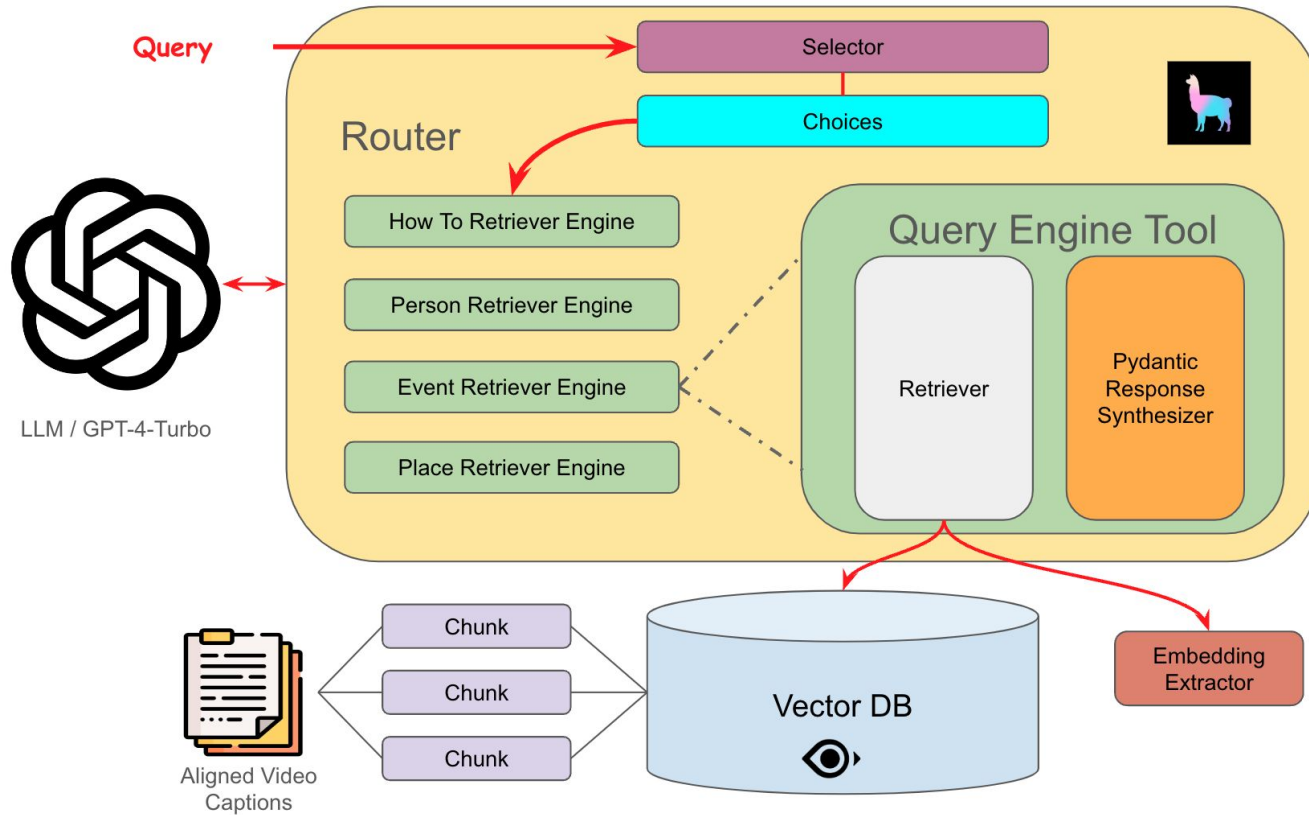


Video Retrieval Augmented Generation Feasibility

Performed relevant video retrieval using general knowledge questions as input

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



LLM application architecture integrating aligned video captions for video enriched RAG



show me the best steakhouses in new york city



Here's what I found:

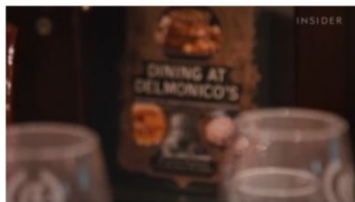
Two of the best steakhouses in New York City are Delmonico's, known for its historical significance and high-quality steaks, and Peter Luger Steak House, renowned for its dry-aged steaks and traditional atmosphere.

Delmonico's

One of the oldest and most iconic steakhouses in New York City, known for its history and top-quality steaks.

Known For:

- Historical significance
- High-quality steaks



clip preview



Peter Luger Steak House

A renowned steakhouse famous for its dry-aged steaks and traditional steakhouse ambiance.

Known For:

- Dry-aged steaks
- Traditional atmosphere



clip preview



Sources:

1. [The Best Steak In New York City | Best In Town](#)

Example: show me the best steakhouses in new york city

what are the top 5 places in england to visit for a harry potter fan?

Here's what I found:

For Harry Potter fans visiting England, these top 5 places offer a magical journey through iconic locations from the beloved wizarding world. From set tours and prop displays to real-life inspirations for Hogwarts, these destinations provide an enchanting experience for fans of the Harry Potter series.

Warner Bros. Studio Tour London - The Making of Harry Potter

A magical experience where Harry Potter fans can explore the sets, props, and costumes from the film series.

Known For:

- Set tours
- Prop displays
- Costume exhibits



clip preview

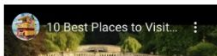


Platform 9¾ at King's Cross Station

A must-visit for Harry Potter enthusiasts, where they can take photos at the famous Platform 9¾ and explore the Harry Potter themed store.

Known For:

- Platform 9¾ photo ops
- Harry Potter merchandise



clip preview



Christ Church, Oxford

Harry Potter fans will recognize this iconic location as the inspiration for the Hogwarts Great Hall in the film series.

Known For:

- Hogwarts Great Hall inspiration



clip preview



Reptile House at London Zoo

A spot where Harry Potter aficionados can visit the location where Harry first discovers he can speak Parseltongue in the film.

Known For:

- Parseltongue discovery location



clip preview



Alnwick Castle

Known for being a filming location for various scenes at Hogwarts in the Harry Potter movies, offering a real taste of the wizarding world.

Known For:

- Hogwarts filming location



clip preview



Sources:

1. [10 Best Places to Visit in England - Travel Video](#)
2. [25 Best Places to Visit in Europe - Travel Europe](#)

Example: what are the top 5 places in england to visit for a harry potter fan



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

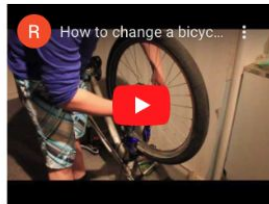


Step 1

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](#)

Example: show me step by step how to replace a bicycle tire



Closing Remarks

Summary

- Aligned video captions are effective for integrating video content into RAG systems
- Can have comparable performance to multimodal LLMs with reduced context window usage

Future Work

- Exploring domain-specific adaptations
- Enhancing audio signal processing
- Benchmarking diverse questions / tasks

