Multi-Moment Video Retrieval

Student~
Abu Shahid
B20CS003

Mentor~
Dr. Anand Mishra
Dept. of CSE

Video Moment Retrieval

YouTube

500+

Hours of video uploaded every minute

1 Billion

Hours of video watched by people everyday

Youtube Statistics 2019

1. Inspecting videos is time-consuming
2. It is hard to find the desired moments

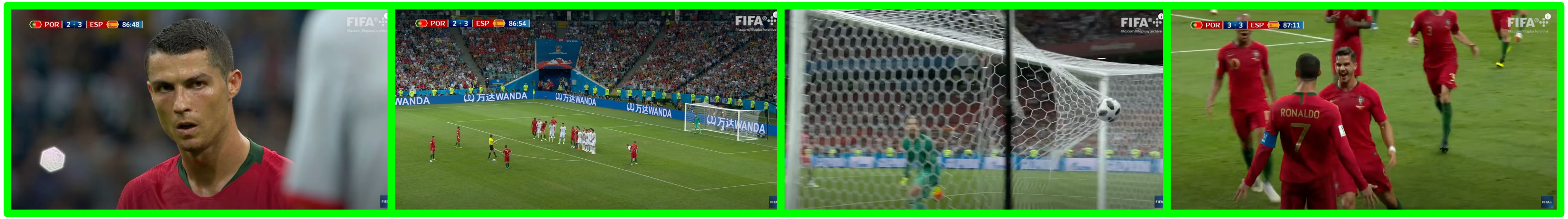
Portugal v Spain | 2018 FIFA World Cup

E.g. finding the goal that tied the match

Video Moment Retrieval

Show me the goal that brought the game to a tie

Portugal v Spain | 2018 FIFA World Cup



Desired Moment: 1:29:49 to **1:32:44**

Retrieving Multiple Moments from a Video

Problem Statement: Given a text query **T** and a video **V** (w/ transcript), localize the relevant segments $\mathbf{S=\{s_i\}}$ in V that are relevant to T.

Example:

Video: Best smartphones released in 2023
Text Query: "How is the battery life of iPhone 14 vs OnePlus 14?"

Retrieved Segments:
    S1: Segment of the video showing the features of iPhone 14
    S2: Segment of the video showing the features of OnePlus 14

Retrieving Multiple Moments from a Video

There is a prior existing work “QVHighlights”

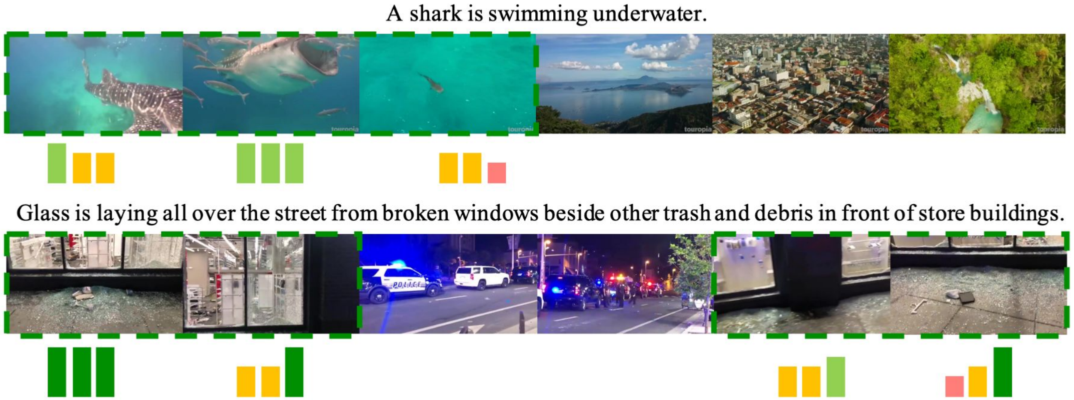
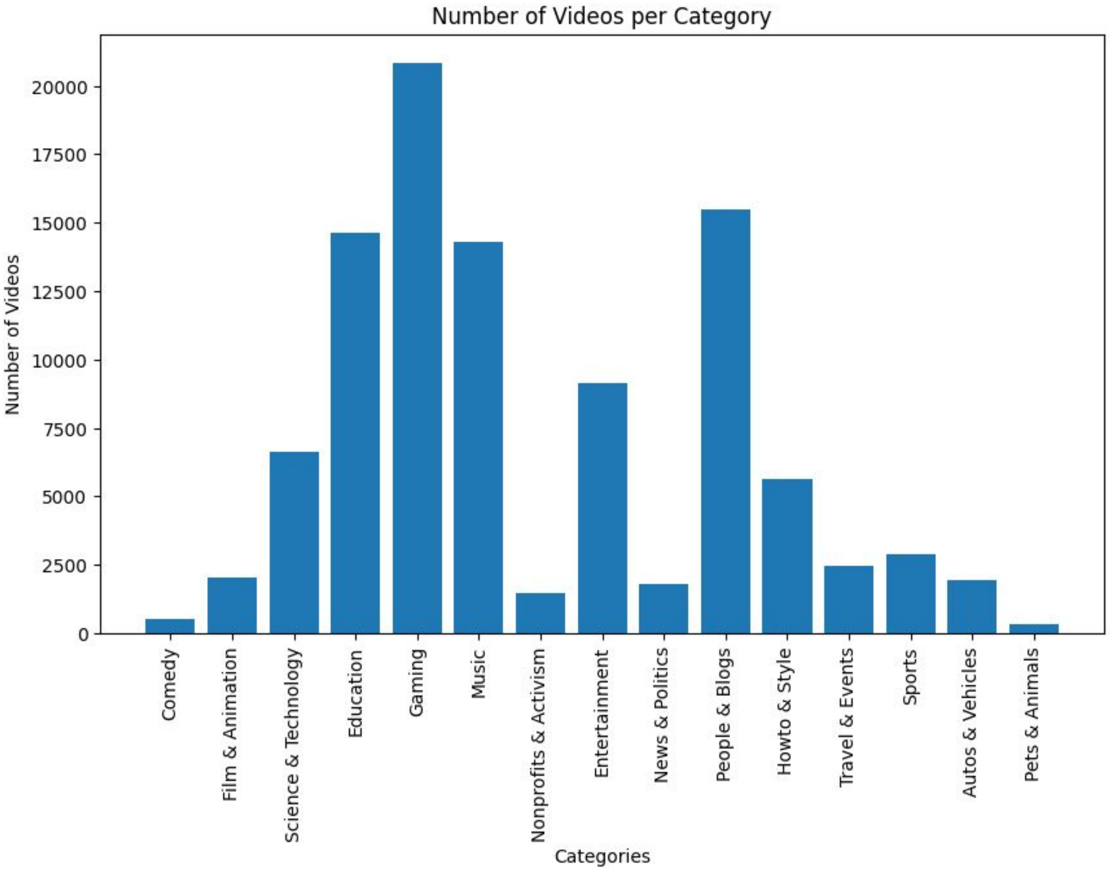


Figure 1: QVHIGHLIGHTS examples. We show localized moments in dashed green boxes. The highlightness (or saliency) scores from 3 different annotators are shown under the frames as colored bars, with height and color intensity proportional to the scores.

QVHIGHLIGHTS: Detecting Moments and Highlights in Videos via Natural Language Queries (Lei at al., NeurIPS’22)

Categories in the 100K YouTube Dataset

- QVHighlights [prior work]
  - Three categories only
    - Daily Vlog,
    - Travel Vlog,
    - News
- 15 categories in 100K YouTube dataset



Demo: Video Moment Retrieval

Video:

Query:

Chef makes a pizza and cuts it up

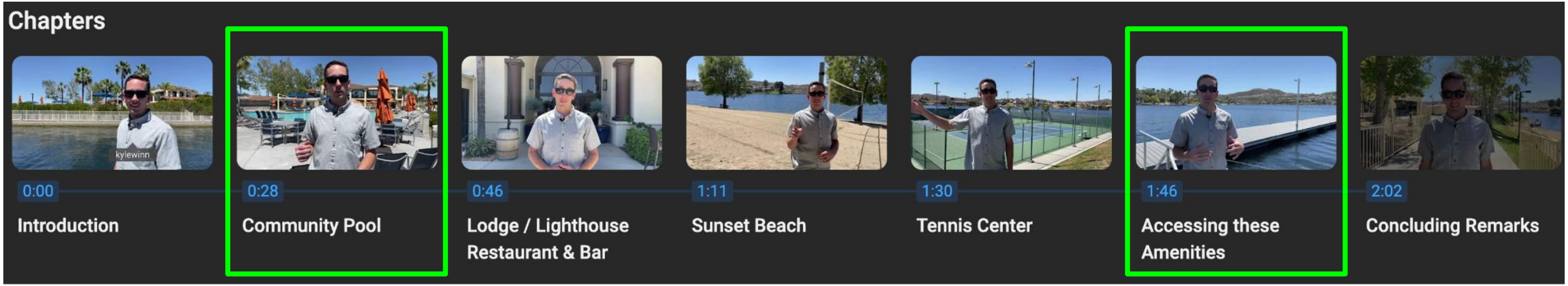
Predicted Moment**:** [start: 1:47, end: 2:02]

(Data Source: QVHighlights)

Retrieving Multiple Moments from a Video

Example

YouTube Video: 4 Great Canyon Lake Amenities In 1 Stop

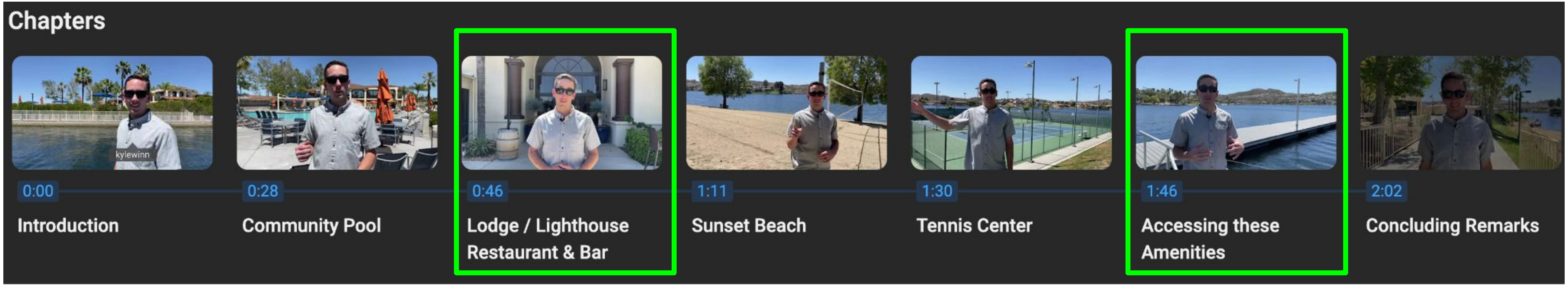


GPT User Query: "Can you access the Canyon Lake pool by boat?"

Retrieving Multiple Moments from a Video

Example

YouTube Video: 4 Great Canyon Lake Amenities In 1 Stop



GPT User Query: "Are there any restaurants near the Canyon Lake lodge with a lake view?"

Few concerns –

No MMVR dataset exists that suits our purpose. Need to make our own.

1. This dataset maybe more **audio/transcript** heavy than **video** heavy

2. Motivation for retrieving multiple clips? Is it a narrow use case?

A Literature Survey on Video Moment Retrieval

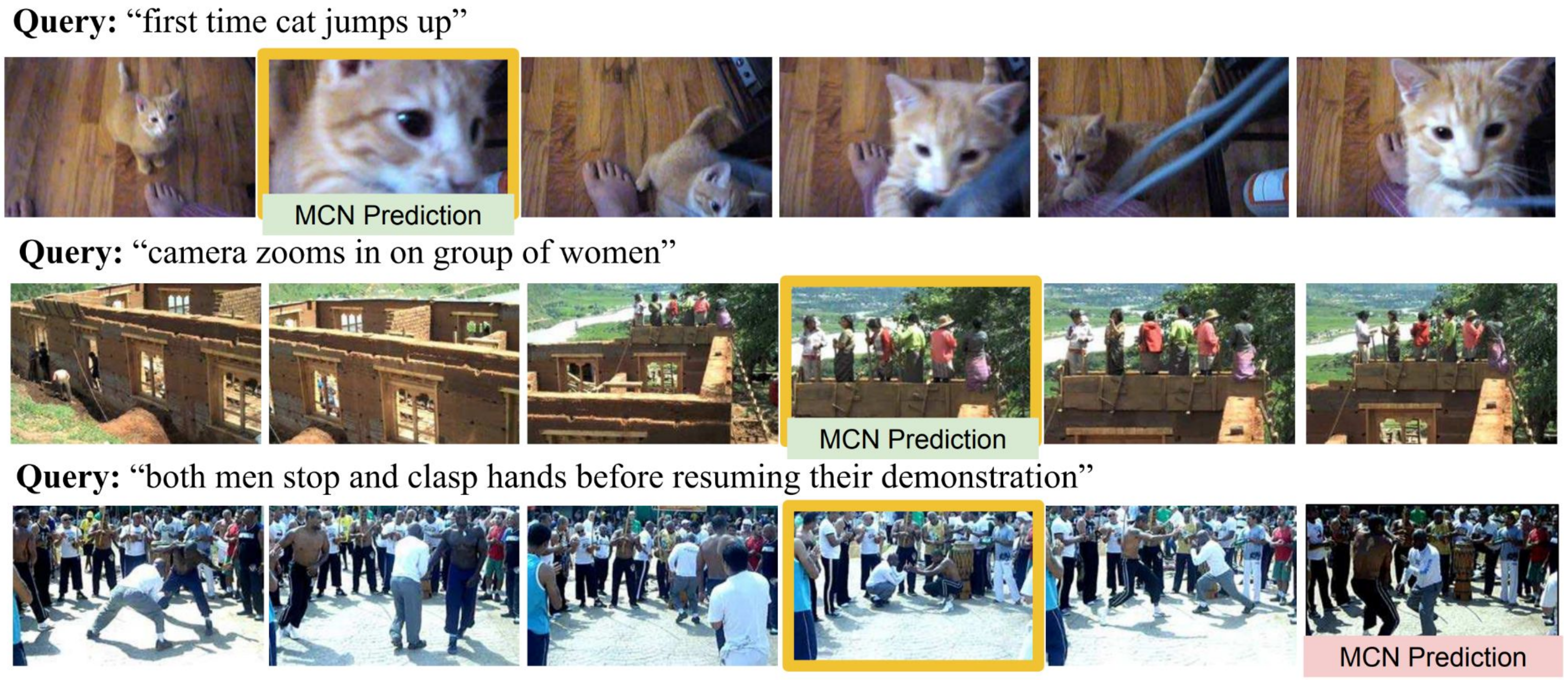
Localizing Moments in Video with Natural Language
[Hendricks et al., ICCV 2017]

Hendricks et al., ICCV'17

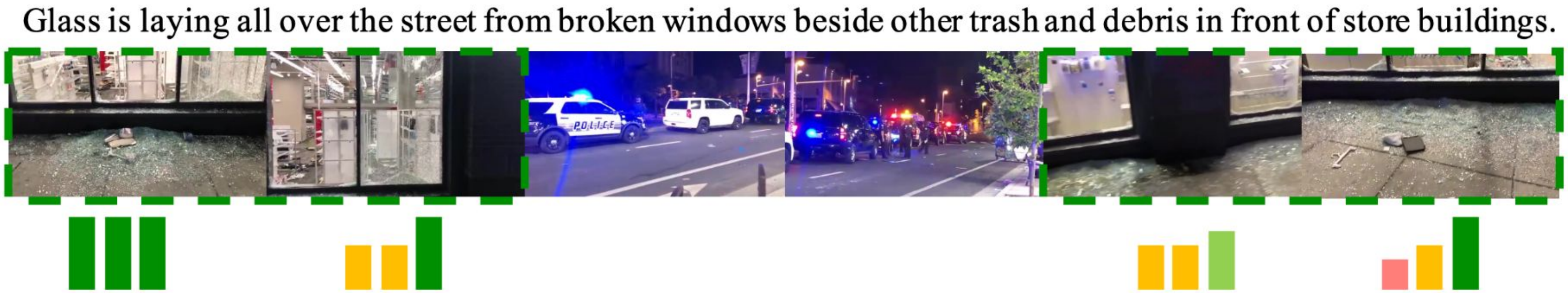
- Led the early efforts for the VMR task
- Aimed to localise moments using simple scene descriptions
- Introduced a suitable dataset **DiDeMo** (Distinct Describable Moments)
- **DiDeMo** contains 40K+ queries and videos

Localizing Moments in Video with Natural Language
[Hendricks et al., ICCV 2017]

Visualization of Moment Prediction



QVHighlights: Detecting Moments and Highlights in Videos via Natural Language Queries **[Lei et al., NeurIPS 2021]**



Lei et al. introduced **QVHighlights Dataset**:

- Large dataset (10K+ queries and videos)
- Supports multiple moments retrieval
- Introduces "saliency scores" for moments
- Longer videos (upto 2.5 minutes)
- More diverse categories (vlogs and news reports)

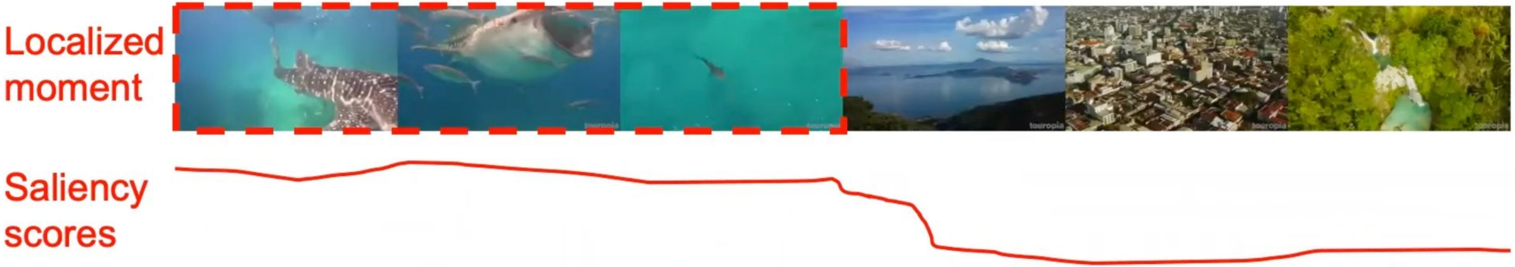
QVHighlights: Detecting Moments and Highlights in Videos via Natural Language Queries **[Lei et al., NeurIPS 2021]**

Follows the **Task Setup**

- ➔ Input

A shark is swimming underwater.

- ➔ Output



QVHighlights: Detecting Moments and Highlights in Videos via Natural Language Queries [Lei et al., NeurIPS 2021]

Visualization of Moment Prediction

Figure ☐ Prediction visualization. Predictions are shown in solid red boxes or lines, ground-truth are indicated by dashed green lines. *Top* row shows a correct prediction, *bottom* row shows a failure.

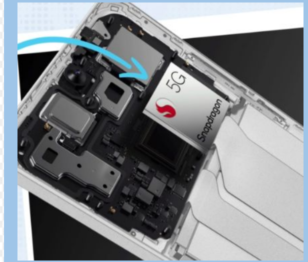
Current works are limited to

- Single moment retrieval
- Queries of simple scene descriptions
- Short videos with limited diversity of topics

Our aim is to retrieve multiple moments from a video where all the moments are jointly required to answer the query. Further we focus on *diverse categories* and *longer videos.*

Video: Pixel 6a vs One Plus 3CE Nord

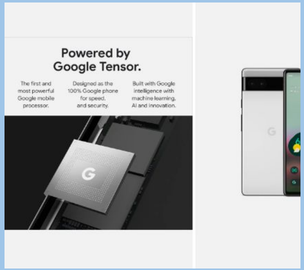
[One plus display]



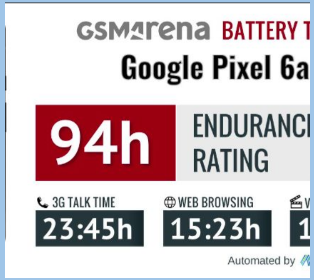
[1+ Processor]

[1+ battery]

[pixel display]



[pixel processor]



[pixel battery]

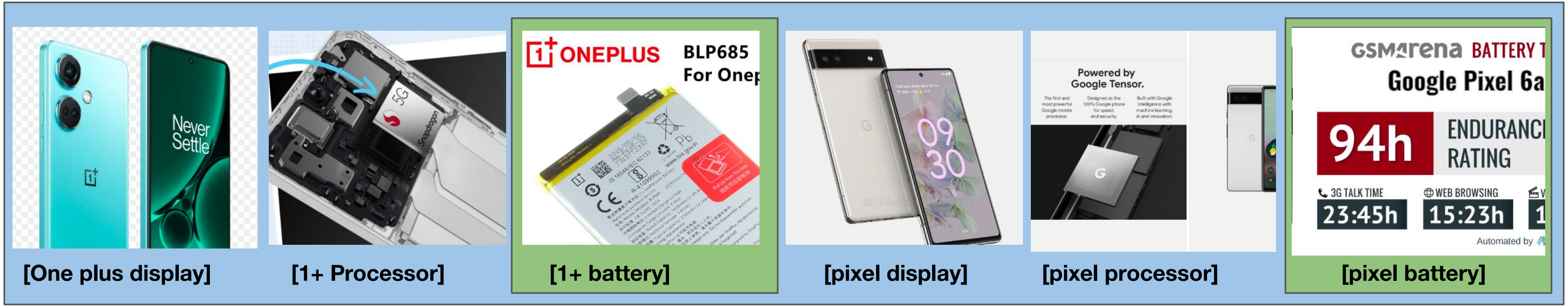
USER QUERY: How is the battery of OnePlus 3 compared to Pixel 6a?

Current works are limited to

- Single moment retrieval
- Queries of simple scene descriptions
- Short videos with limited diversity of topics

Our aim is to retrieve multiple moments from a video where all the moments are jointly required to answer the query. Further we focus on *diverse categories* and *longer videos.*

Video: Pixel 6a vs One Plus 3CE Nord



USER QUERY: How is the battery of OnePlus 3 compared to Pixel 6A?

Approach and Prompt Structure

We require

- **A large diverse dataset**
- **Covering various categories**

Approach:

- Give all the necessary context of a YouTube video to GPT
- Ask to generate user queries which require multiple chapters for answering

<context + rules>
<video title>
<description>
<categories>
Chapter annotations
    <chapter 1>
        <video captions>
        <subtitles>
    <chapter 2>
        <video captions>
        <subtitles>
    ...
<name of chapters for framing queries>

Prompt Template for LLM

Results so far…

Video: **TORTANG Talong 2 Ways, EGGY and CRISPY**

Intro

Preparation and Grilling

Eggy

Crispy

<Prompt> Generate queries targeting following two chapters

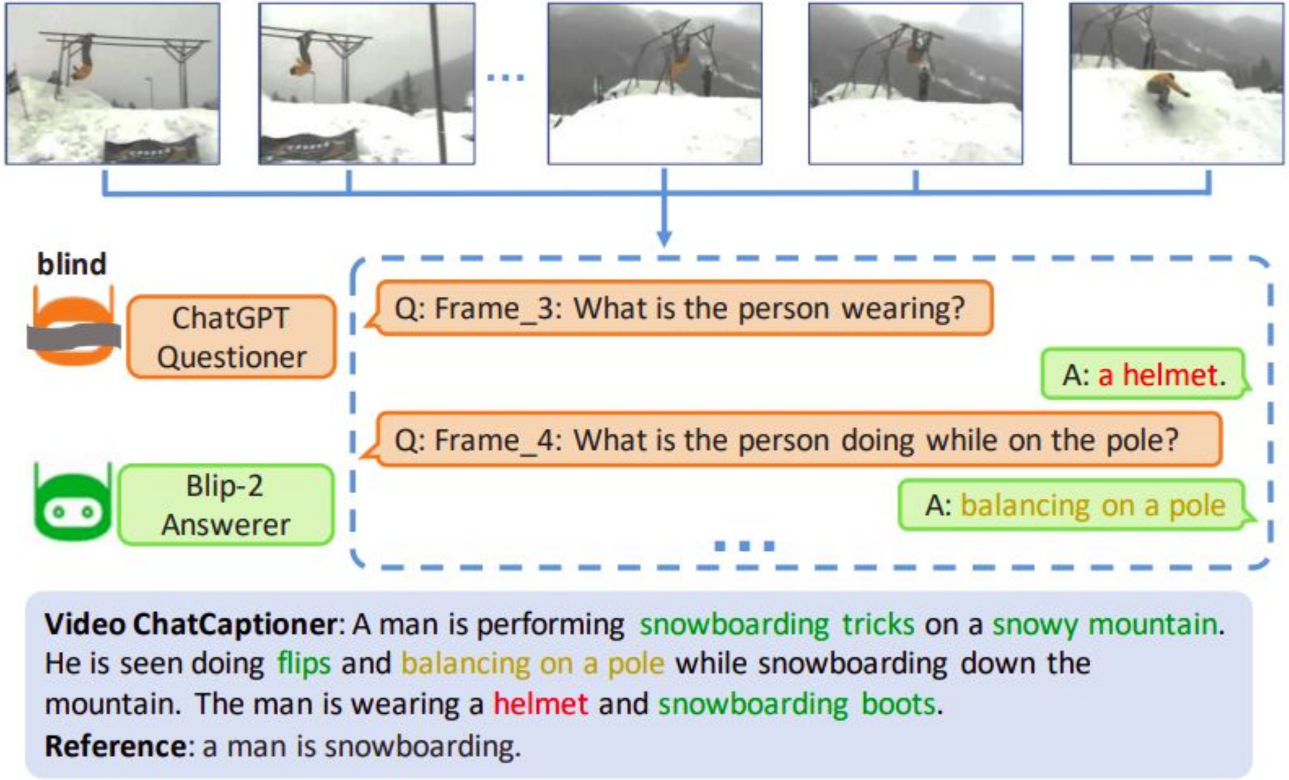
i. Preparation and Grilling
ii. Crispy

<Response> What steps are involved in making a crispy torta after grilling the eggplants?

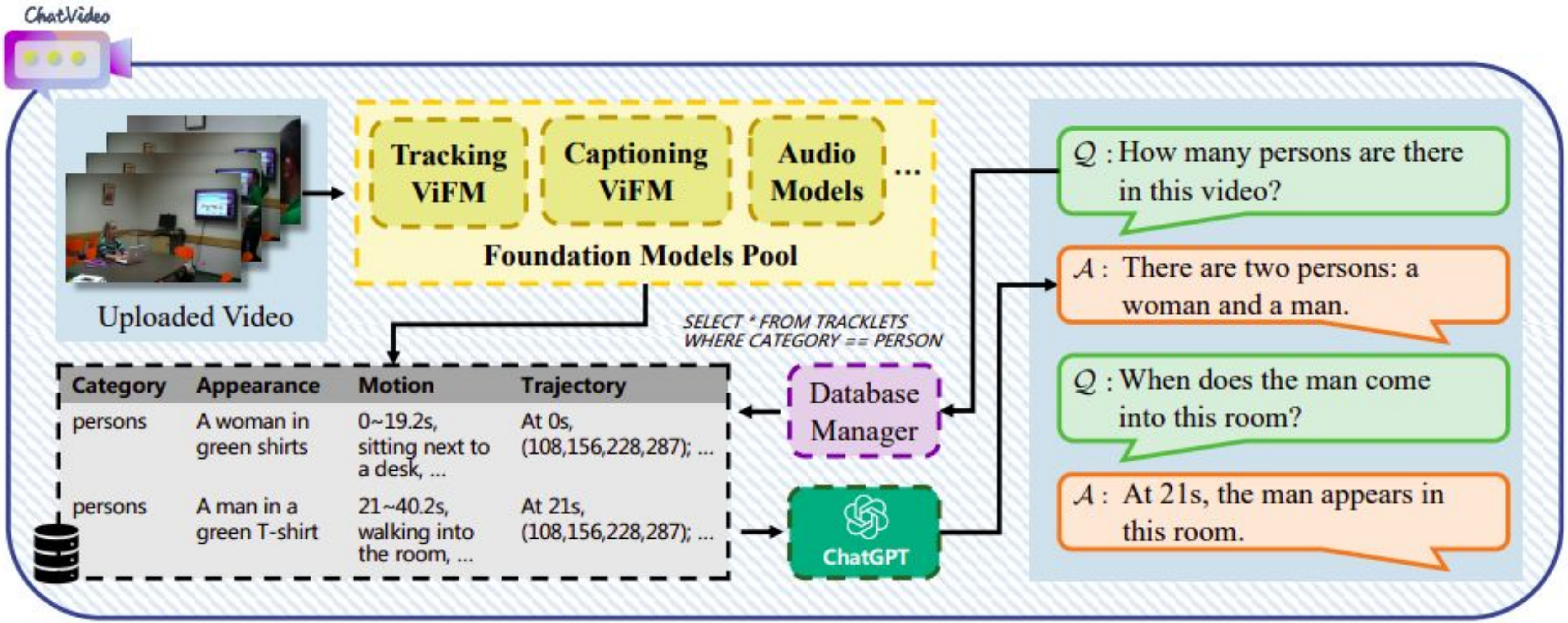
“Video-LLM” literature review

1. Chat-Video
2. VideoChat
3. Video ChatCaptioner
4. Video-ChatGPT

Video ChatCaptioner

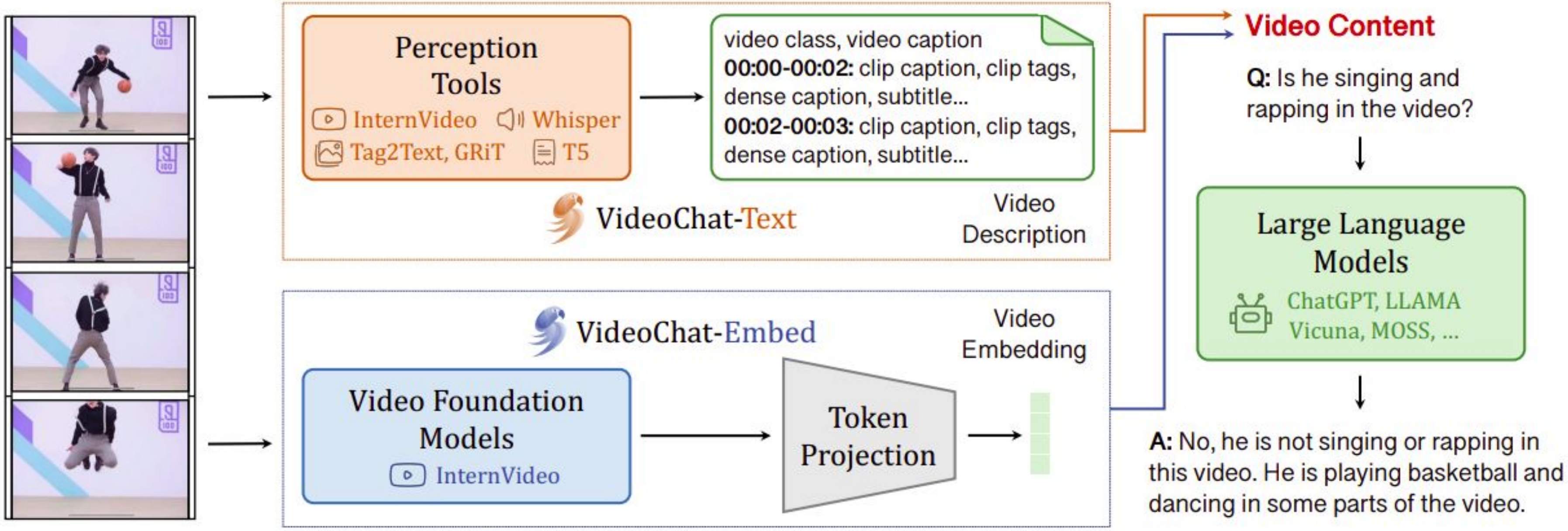


Chat-Video



https://arxiv.org/pdf/2304.14407.pdf

Video-Chat



The Input from perception tools and VFMs serves as context to LLMs which can now engage in question answering. https://arxiv.org/pdf/2305.06355.pdf

Prompt Engineering

1. Went through 10+ versions of prompt refinement
   a. Followed by 10+ human evaluation steps
2. Tried 0-shot and few-shot prompting techniques
3. Tried various flavour of context stripping
4. Tried multi-modal GPT4

Results so far…

Video: **TORTANG Talong 2 Ways, EGGY and CRISPY**

Intro

Preparation and Grilling

Eggy

Crispy

<Prompt> Generate queries targeting following two chapters

i. Preparation and Grilling
ii. Crispy

<Response> What steps are involved in making a crispy torta after grilling the eggplants?

Recognised common error categories

- “And” template is used very frequently
    - Two independent queries joined by word ‘and’
- Chapters picked may not have relation
- Use of “Intro” and “Outro” chapters
- Lot of videos are sequential (Step-based videos)
- Queries can be answered from non-selected chapters
- Lack of visual-based queries
- Most videos are digital or monologue (lack of diversity)
- Granularity of chapters
- AI made videos

Previous Issues with queries and their status/solution

- Use of ‘And’ template persists
  - Alleviated to a good extent
- Chapters picked may not have relation-
  - Instances reduced
- Use of “Intro” and “Outro” chapters
  - Instances reduced
- Lot of videos are sequential or completely random
  - No proper way of alleviating
- Queries can be answered from non-selected chapters
  - No proper way of alleviating post chapter selection freedom
- **Lack of visual-based queries**
  - **Can be solved by making prompts crisper**
- Videos lack of diversity; AI made videos
  - No proper way of alleviating

Baseline 0(w/ single MR)

Experiment:

- 1000 single-VMR samples: (video, query, gt_chapter)
- Used ground truth chapter segmentations of the YT video
- Used pre-trained **ImageBind** model for encoding Video, Transcript and Query
- Given query and video; for each chapter, we measure:
  Query-Video (s_{QV}) and Query-Transcript (s_{QT}) Similarity.
- Evaluation: Classification task - accuracy

Baseline 0(w/ single MR)

Experiment:

Method	Accuracy (%)
Using S_{QV} only	28.9
Using S_{QT} only	16.0
Using $(a*S_{QV} + b*S_{QT})$ [a = 0.5, b = 0.3]	32.0

Table: Classification accuracy for single-moment retrieval (using ground truth chapter segmentations) for 1000 samples (Pre-trained ImageBind)

Dataset v-1

Sample

ID: **TuQwxE1tQdw,**

Query : **"How do soldiers incorporate their surroundings in an urban training environment, and what impact does discipline have on these exercises?",**

Ground truth chapters: **['Urban Training', 'Shooting Exercise']",**

Chapter-duration info: **"{""Marines Train with Green Berets"": {""start_time"": 0, ""end_time"": 10, ""duration"": 10}, ""Room Clearing"": {""start_time"": 10, ""end_time"": 40, ""duration"": 30}, ""Shooting Exercise"": {""start_time"": 40, ""end_time"": 50, ""duration"": 10}, ""Urban Training"": {""start_time"": 50, ""end_time"": 59, ""duration"": 9}}"**

Dataset v-1

Chapterwise audio, video accessed using video ID

Sample

ID: **TuQwxE1tQdw,**

Query : **"How do soldiers incorporate their surroundings in an urban training environment, and what impact does discipline have on these exercises?",**

Ground truth chapters: **['Urban Training', 'Shooting Exercise']",**

Chapter-duration info: **"{""Marines Train with Green Berets"": {""start_time"": 0, ""end_time"": 10, ""duration"": 10}, ""Room Clearing"": {""start_time"": 10, ""end_time"": 40, ""duration"": 30}, ""Shooting Exercise"": {""start_time"": 40, ""end_time"": 50, ""duration"": 10}, ""Urban Training"": {""start_time"": 50, ""end_time"": 59, ""duration"": 9}}"**

Dataset v-1

Sample

ID: **TuQwxE1tQdw,**

Query : **"How do soldiers incorporate their surroundings in an urban training environment, and what impact does discipline have on these exercises?",**

Ground truth chapters: **['Urban Training', 'Shooting Exercise']",**

Chapter-duration info: **"{""Marines Train with Green Berets"": {""start_time"": 0, ""end_time"": 10, ""duration"": 10}, ""Room Clearing"": {""start_time"": 10, ""end_time"": 40, ""duration"": 30}, ""Shooting Exercise"": {""start_time"": 40, ""end_time"": 50, ""duration"": 10}, ""Urban Training"": {""start_time"": 50, ""end_time"": 59, ""duration"": 9}}"**

Dataset v-1

Chapterwise audio, video accessed using video ID

Sample

ID: **TuQwxE1tQdw,**

Query : **"How do soldiers incorporate their surroundings in an urban training environment, and what impact does discipline have on these exercises?",**

Ground truth chapters: **['Urban Training', 'Shooting Exercise']",**

Chapter-duration info: **"{""Marines Train with Green Berets"": {""start_time"": 0, ""end_time"": 10, ""duration"": 10}, ""Room Clearing"": {""start_time"": 10, ""end_time"": 40, ""duration"": 30}, ""Shooting Exercise"": {""start_time"": 40, ""end_time"": 50, ""duration"": 10}, ""Urban Training"": {""start_time"": 50, ""end_time"": 59, ""duration"": 9}}"**

Dataset v-1

Chapterwise audio, video accessed using video ID

Sample

ID: **TuQwxE1tQdw,**

Query : **"How do soldiers incorporate their surroundings in an urban training environment, and what impact does discipline have on these exercises?",**

Ground truth chapters: **['Urban Training', 'Shooting Exercise']",**

Chapter-duration info: **"{""Marines Train with Green Berets"": {""start_time"": 0, ""end_time"": 10, ""duration"": 10}, ""Room Clearing"": {""start_time"": 10, ""end_time"": 40, ""duration"": 30}, ""Shooting Exercise"": {""start_time"": 40, ""end_time"": 50, ""duration"": 10}, ""Urban Training"": {""start_time"": 50, ""end_time"": 59, ""duration"": 9}}"**

x10.9k times

Dataset v-1

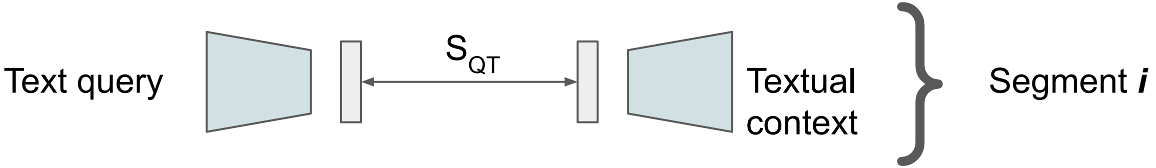
2500 videos * 5 queries = 12500 datapoints <video, query, timestamps>

- However, many data points were noisy.
- After cleaning, 10.9K data points
  - 8718 in train set
  - 2179 in test set
- Dataset generation pipeline is finalized
  - More data can now be generated as and when needed.

Baseline ideas

1. **Two stage model with text embeddings**
   (a) Temporal Segmentation into chapters -> Using existing model?
   (b) Retrieve multiple relevant segments from the video

Relevance Score for segment i, S^i



Rank and retrieve the chapters using this score

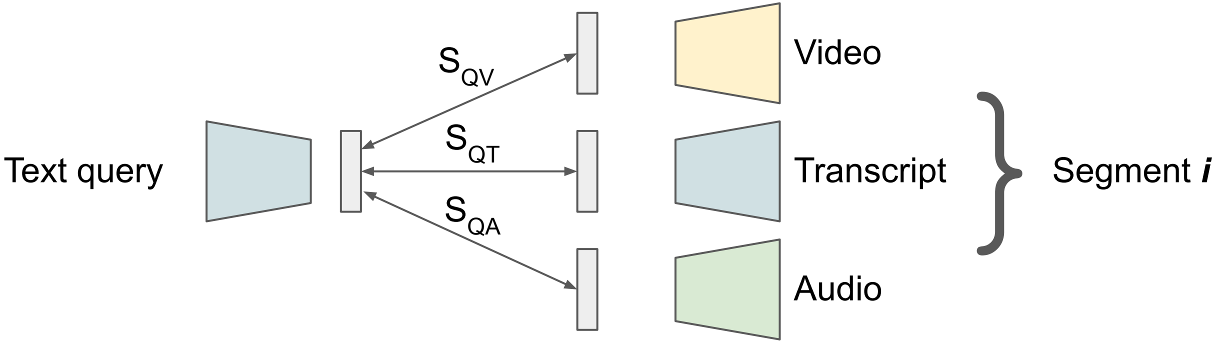
<u>First experiment</u>: Can we use ground truth chapters + pretrained models?

Baseline ideas

2. Two stage model text+audio+video embeddings

(a) Temporal Segmentation into chapters -> Using existing model?

(b) Retrieve multiple relevant segments from the video



Relevance Score for segment i

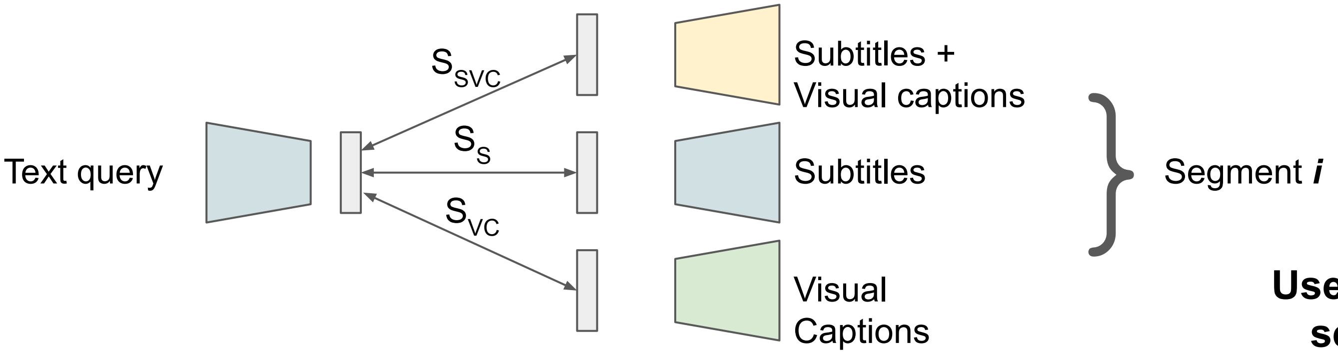
$S^i = a*S_{QV} + b*S_{QT} + c*S_{QA}$

and

If S^i > threshold “**1**” else “**0**”

<u>First experiment</u>: Can we use ground truth chapters + pretrained models?

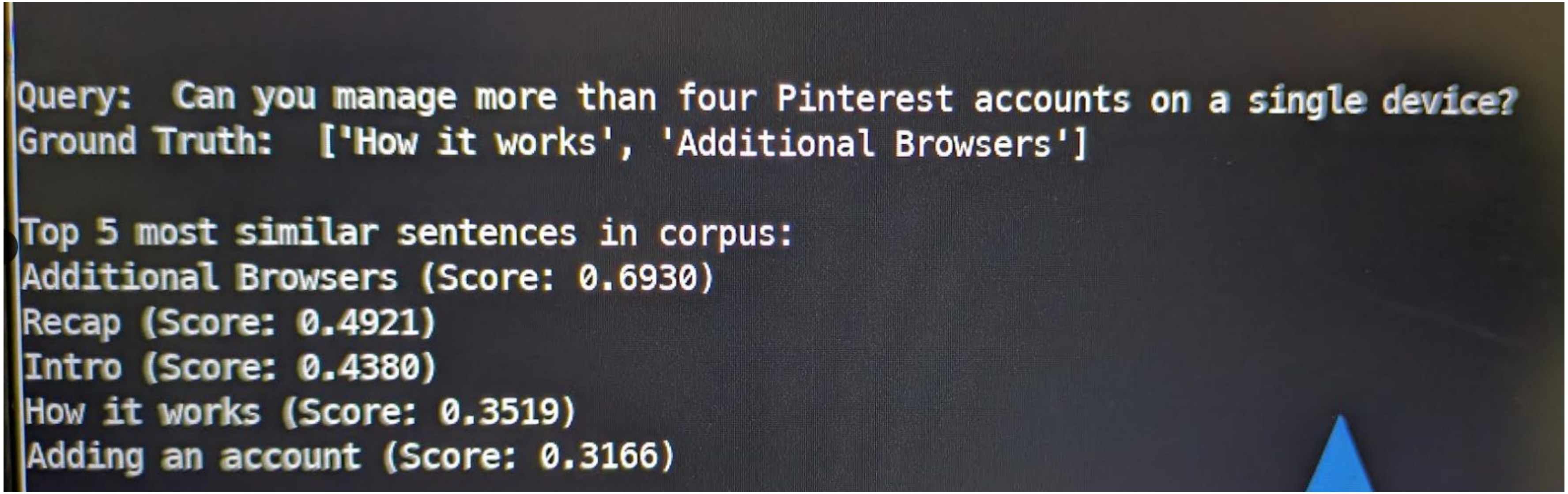
Baseline #1 setup



Use sentence-bert to calculate semantic textual similarity

- Do the above for each chapter
- Use these scores to rank and retrieve the relevant chapters
  - Top two ranks are taken are predictions

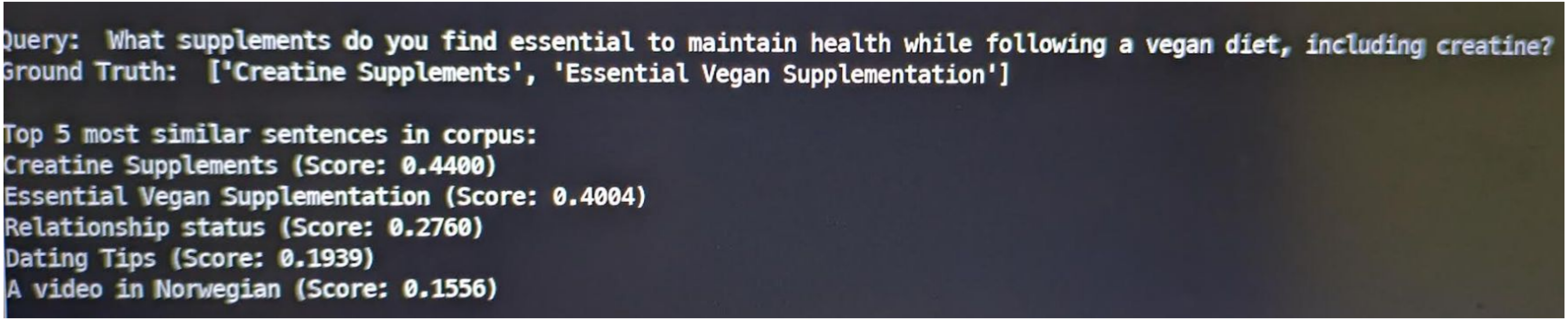
Sample results



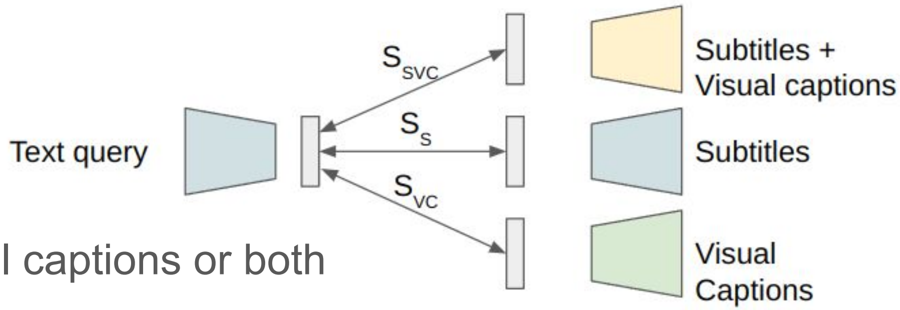
```
Query:  Can you manage more than four Pinterest accounts on a single device?
Ground Truth:  ['How it works', 'Additional Browsers']

Top 5 most similar sentences in corpus:
Additional Browsers (Score: 0.6930)
Recap (Score: 0.4921)
Intro (Score: 0.4380)
How it works (Score: 0.3519)
Adding an account (Score: 0.3166)
```

Sample results



Sentence Bert Experimentation



Input: Query + Chapterwise subtitles or visual captions or both
Task:

- Find semantic textual similarity bw Query & Chapters
- Rank and retrieve top 2
    - Assumption: We know 2 moments need to be retrieved

For a sample:-

If no ground truth segment in top-2 => 0.0
If one ground truth segment in top-2 => 0.5
If two ground truth segment in top-2 => 1.0

- Accuracy: For a sample, accuracy can be 0%, 50% or 100% as described above

Sentence Bert Experimentation

- Ground truth chapters: **[C, D]**
- Predicted/Retrieved Chapters: **[B, D]**

- True positive : [D]
- False positive: [B]
- False negative: [C]
- True negative: [A, E]

- IOU= TP/(TP + FN + FP)
- Precision= TP/(TP + FP)
- Recall = TP(TP + FN)

Sentence-bert baseline

S-BERT Method	Acc.	Avg IOU	Avg P	Avg R
Measured using dot-product similarity				
w/ Subtitles	48.8	0.414	0.497	0.495
w/ VisCaps	35.4	0.279	0.367	0.354
w/ Sub+VC	**50.0**	**0.426**	**0.510**	**0.513**

Table: Experimentation results from Sentence-Bert

Sentence-bert baseline

S-BERT Method	Acc.	Avg IOU	Avg P	Avg R
Measured using dot-product similarity				
w/ Subtitles	48.8	0.414	0.497	0.495
w/ VisCaps	35.4	0.279	0.367	0.354
w/ Sub+VC	**50.0**	**0.426**	**0.510**	**0.513**
Random selection				
_	29.2	0.216	0.30	0.27

Current/Upcoming work

- **ImageBind Baseline:**
  - Do the experimentation with audio, video and transcript in multi-moment retrieval setup
  - Find optimal values of ***a, b*** and ***c*** in $S^i = a*S_{QV} + b*S_{QT} + c*S_{QA}$

    … which gives the best result

- **LLM based baseline:**
  - Provide query and context to a LLM and ask the LLM to rank the chapters.
- Presently we are assuming availability of ground truth chapter boundaries. Plans to work on candidate-chapter-segmentation.
- Expand to more than 2 chapter retrieval.
- Plans to work on fine-grained retrieval.

Thank You

:)