MLDA Deep Learning Week Hackathon

Problem #1: Tiktok Video Captioning Data Filtering Pipeline

Xponential Group:

- Ng Zheng Xun (CSC Y4)
- Lyu XingXiao (EEE Masters)
- Kevin Hu (EEE Masters)

Introduction to Problem Statement

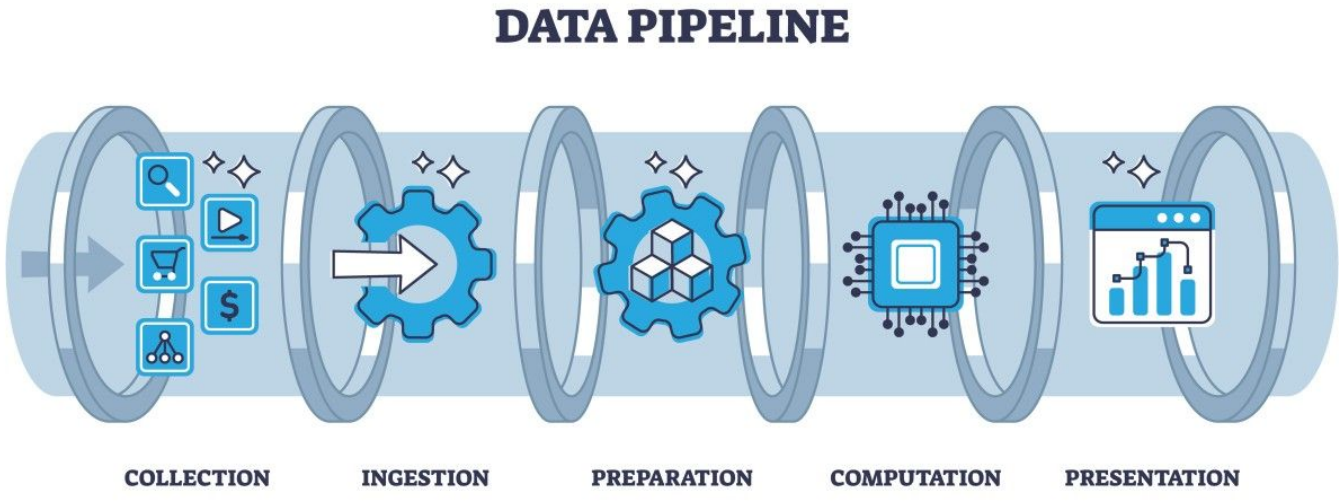
Given a Noisy Dataset of 1M Video-Caption pairs,

construct a pipeline to filter out the top 10% (10k pairs) in terms of quality

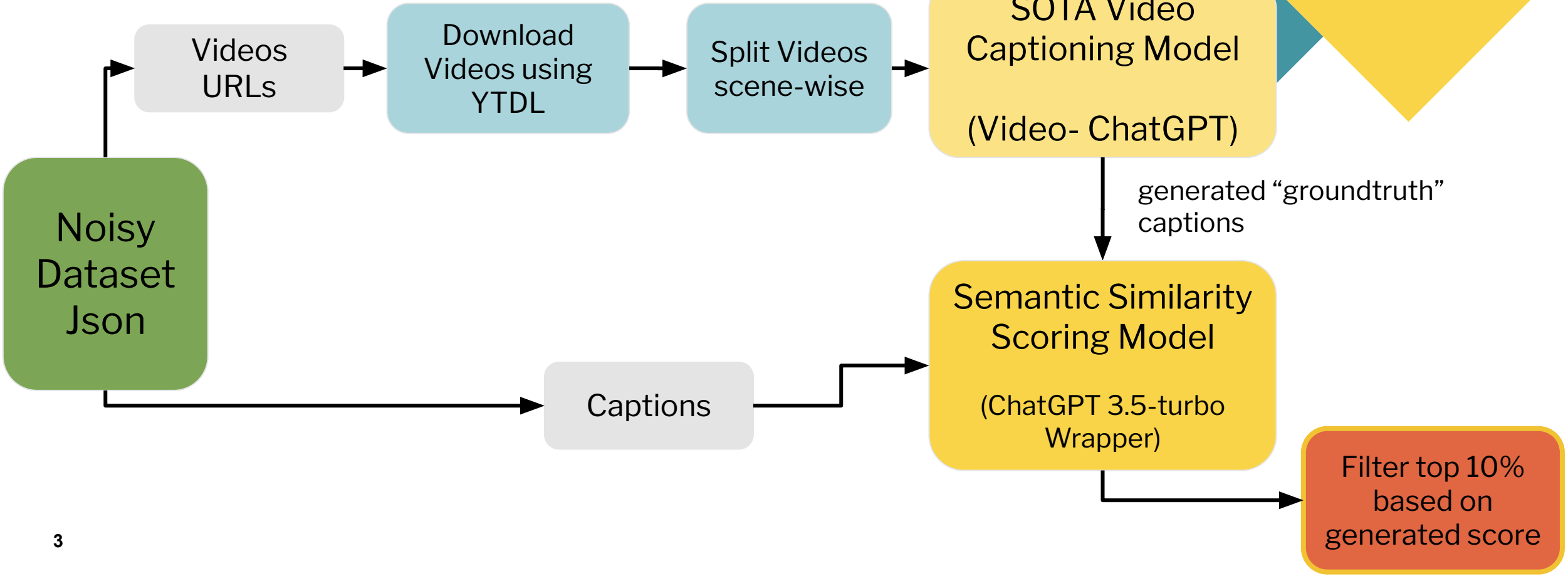
- crucial for improving model performance

ideally optimised to reduce

CO2 emissions



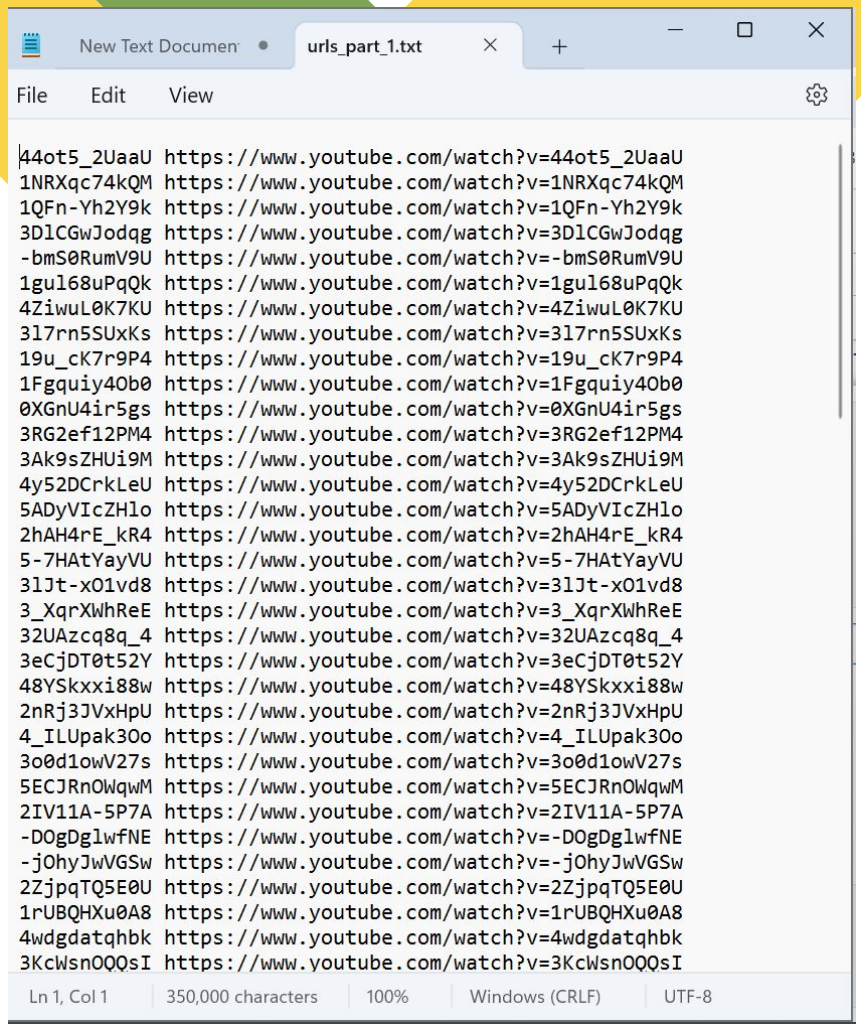
Overall Pipeline



Loading and Parsing Noisy Dataset Json

- Initial Idea: multiprocessing
  - ultimately not implemented due to time constraints, and the lack of GPU processing

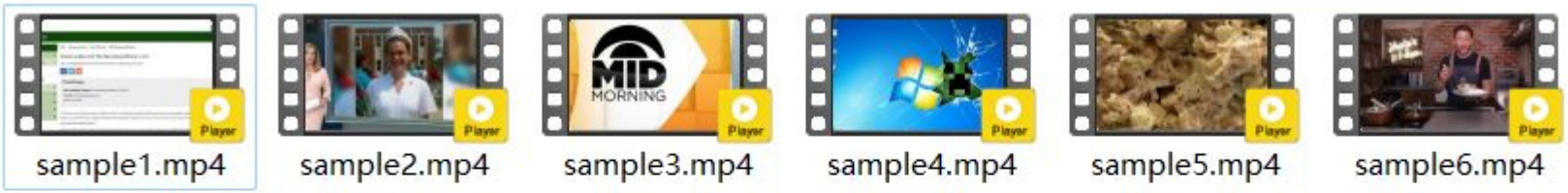
- Extracting youtube links and video ID from json file, and saving it as 3 separate text files, for multiprocessing



Automating the Download of Videos from URLs

- We used the **Python youtube-dl (YTDL) library** to do an automated download of the given URLs in mlda_data.json.

- We download videos in a lower resolution of 360p to optimise computational load

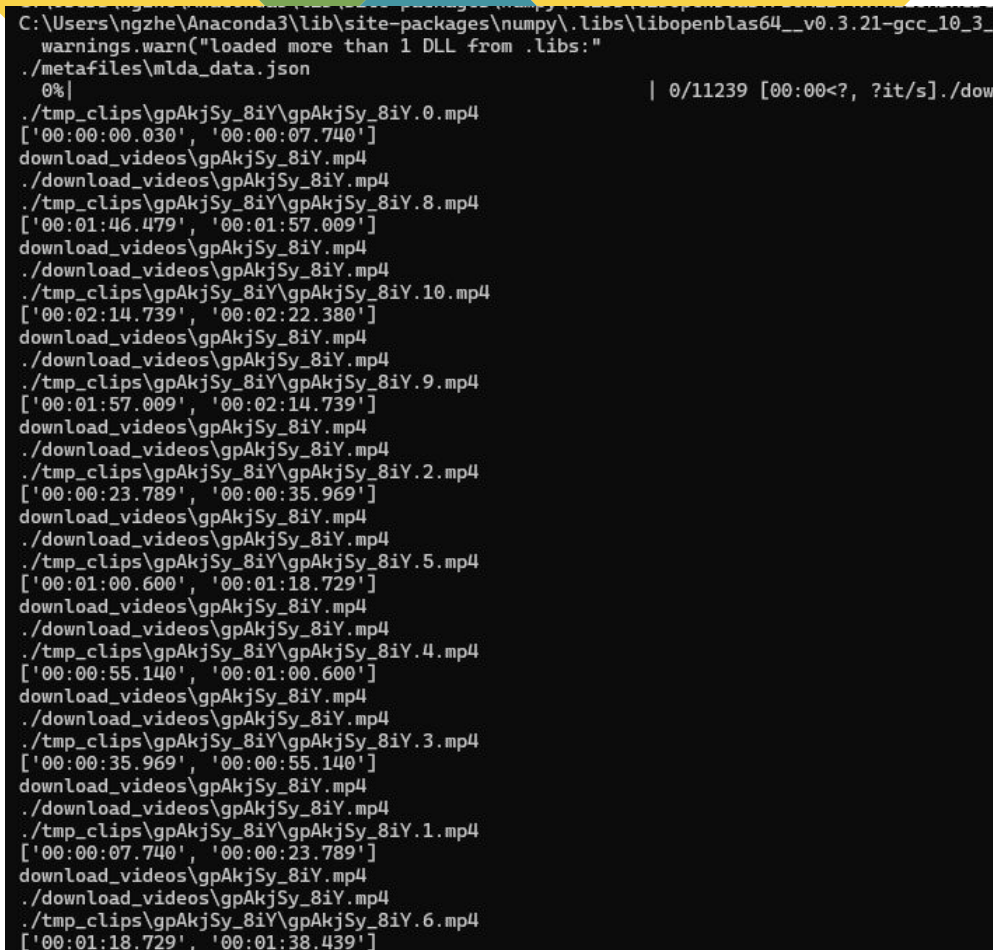


Splitting Downloaded Videos using Python script

- Clip videos from the downloaded video according to the timestamp shown in *mlda_data.json* through the given programme by Tiktok.

- much less intensive due to the script provided.

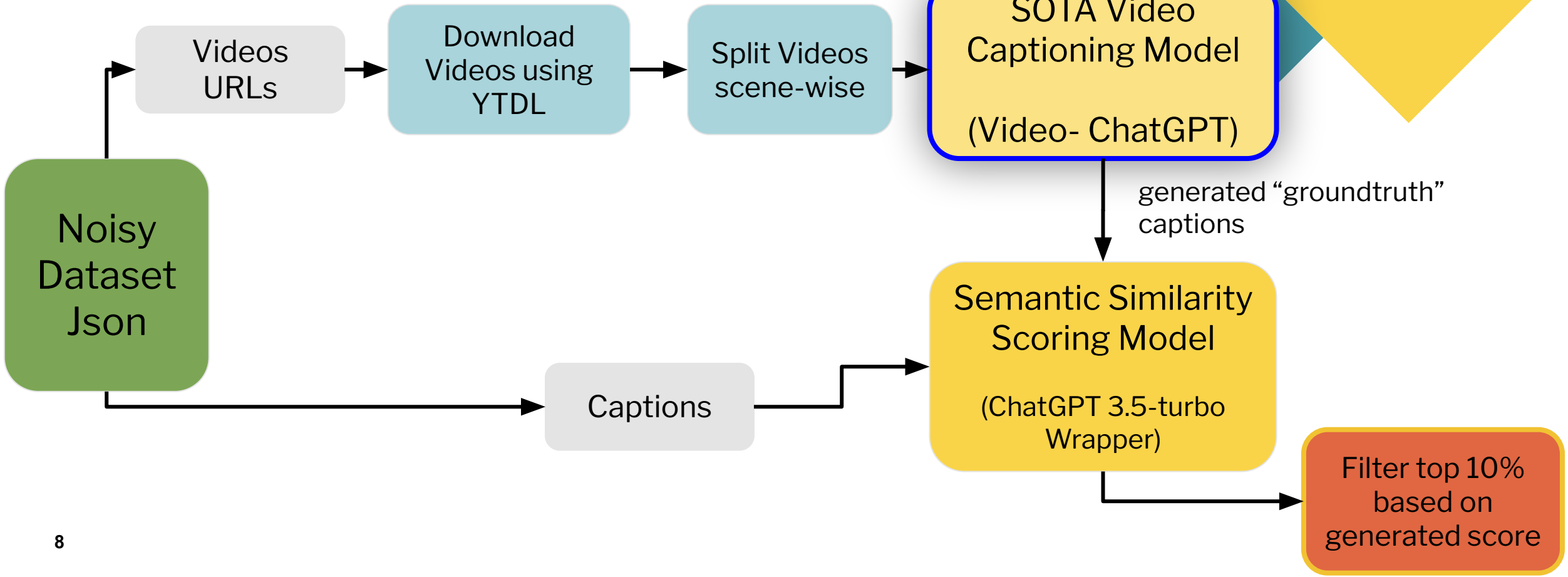
- After some debugging and tweaking some variables, `**python cut_videos_mlda.py --metafile mlda_data.json --resultfile cut_part0.jsonl**`



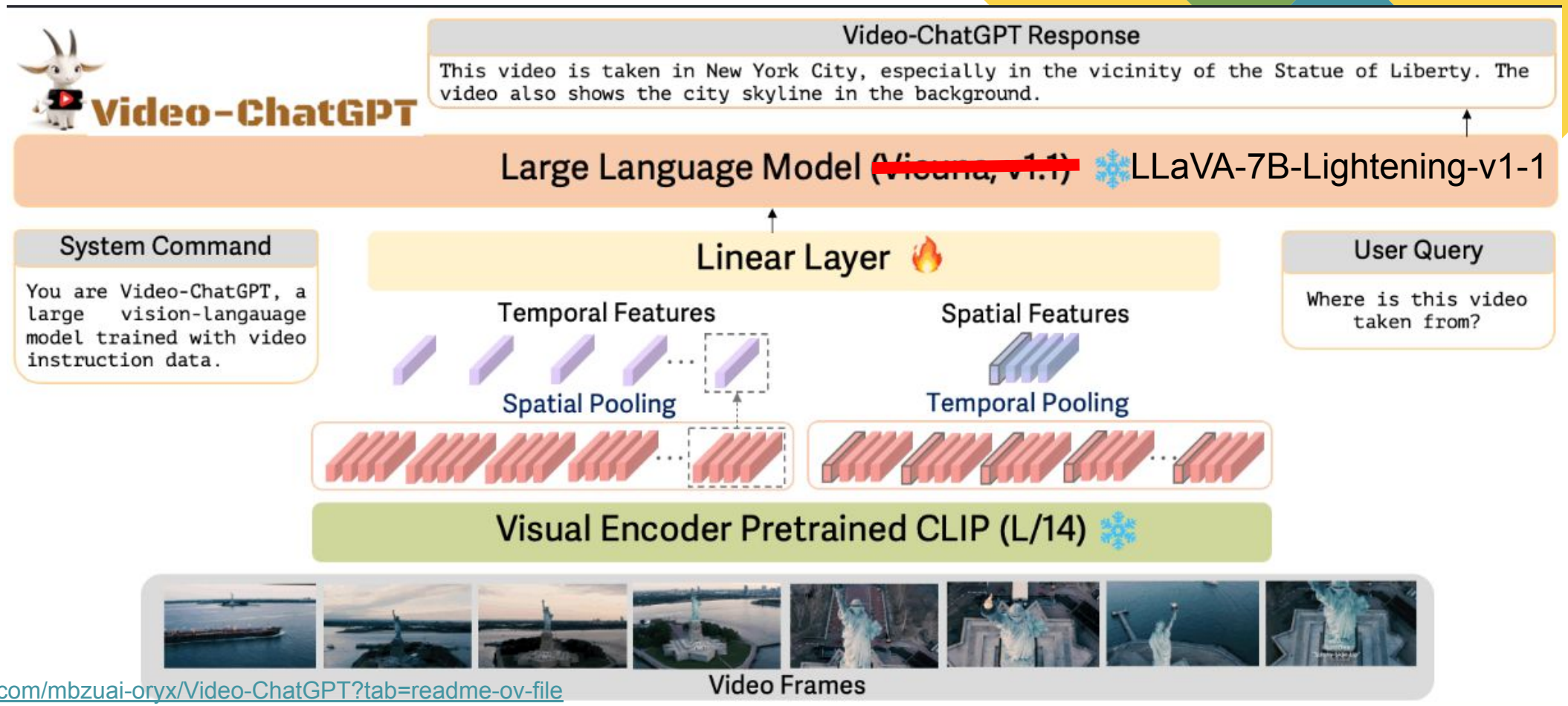
Splitting Downloaded Videos using Python script

We take the FIRST FOUR videos from the above folder as our toy dataset

Overall Pipeline



Generate "groundtruth" Caption with SOTA Video Captioning Model



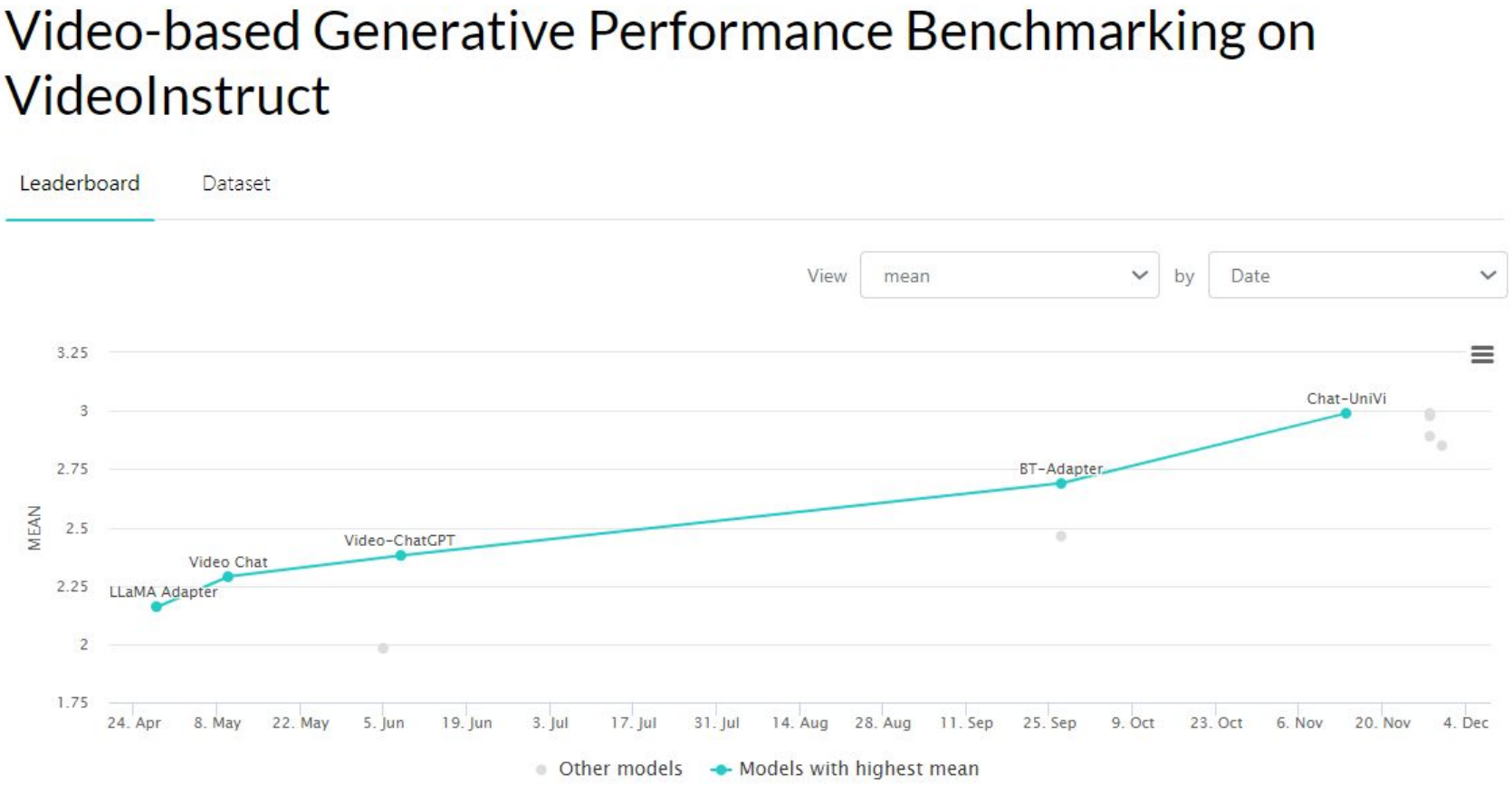
https://github.com/mbzuai-oryx/Video-ChatGPT?tab=readme-ov-file

Generate "groundtruth" Caption with SOTA Video Captioning Model

- Usage of State-of-the-Art (SOTA) Video Captioning Model **VideoChatGPT to generate a groundtruth caption**

- Model was chosen as it gives **SOTA and industry-leading results** on several video captioning benchmarks, such as the VideoInstruct benchmark

Generate "groundtruth" Caption with SOTA Video Captioning Model



https://paperswithcode.com/sota/video-based-generative-performance?p=video-chatgpt-towards-detailed-video

Generate "groundtruth" Caption with SOTA Video Captioning Model

- Model can also do Question-Answering, but **we focus on the Video Captioning instead**
-
    - Interestingly, the original authors a**pproached the Video Captioning task as a variation of Question-Answering,** whereby the question is simply the provided noisy caption.

Generate "groundtruth" Caption with SOTA Video Captioning Model

- There were **significant difficulties in setting up a suitable environment, and to get the model running.**

- **However, we were able to overcome these difficulties, and adapting the model to run on a CPU environment**

Generate "groundtruth" Caption with SOTA Video Captioning Model

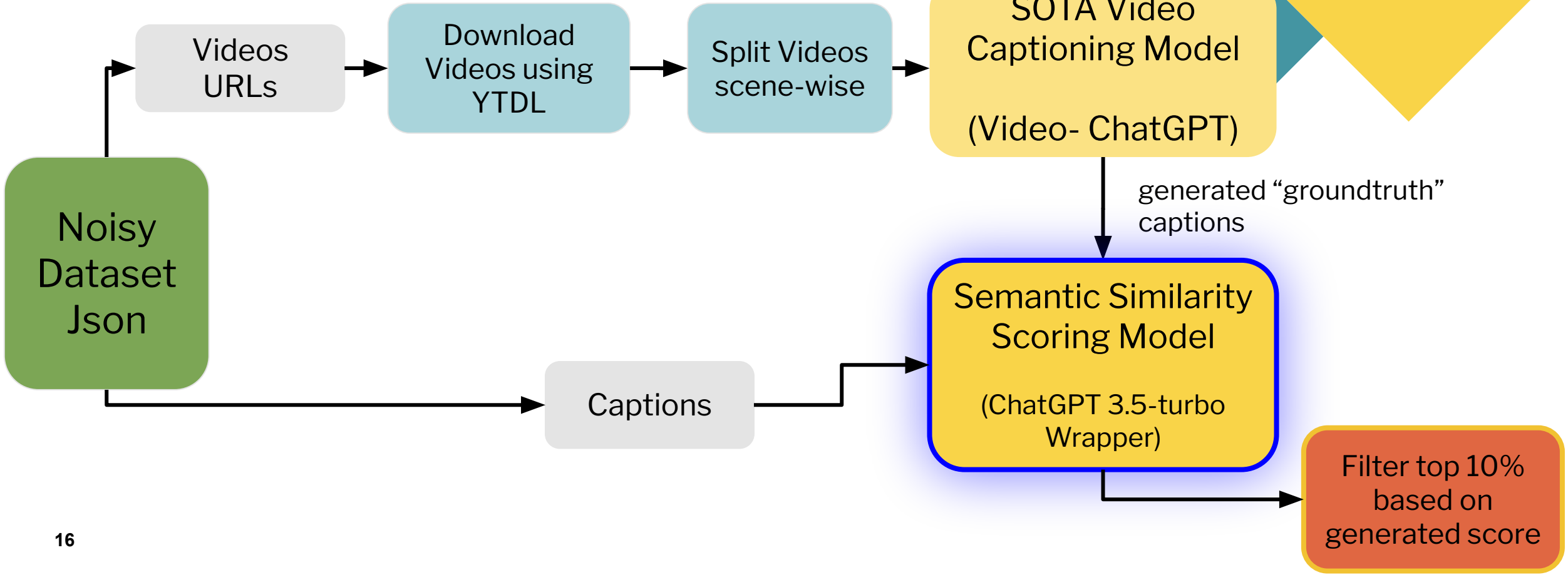
- Other models were considered, such as mPLUG-2 by Alibaba, but were ultimately not implemented due to time constraints. Ideally, we hoped to be able to do a comparison between the 2 models

- CLIP score was considered but as an **image-to-score metric, it was insufficient** and **unable to capture crucial context** or verbs (movement-based)

Generate "groundtruth" Caption with SOTA Video Captioning Model


```
 {
    "clip": "gpAkjSy_8iY/gpAkjSy_8iY.0_0.mp4",
    "caption": "a man in a suit and tie standing next to a news desk",
    "pred": "The man in the suit and tie is standing next to a news desk, and he is talking
to the camera. He is wearing a suit and tie, which suggests that he is likely a news
reporter or a guest on a news show. He is talking about the news and providing information
to the audience."
  },
  {
    "clip": "gpAkjSy_8iY/gpAkjSy_8iY.8_1.mp4"
    "caption": "a street that has power lines
    "pred": "The video shows a street with po
about the situation. He is wearing a suit and
  },
  {
    "clip": "gpAkjSy_8iY/gpAkjSy_8iY.8_2.mp4"
    "caption": "a woman standing in the middl
    "pred": "In the video, a woman is standin
by water. She is holding a microphone and spe
also speaking to the camera, and the news rep
reporter is talking about the flood and the w
reporter is also talking about the flood and
  },
```


Overall Pipeline



Generating a Semantic Similarity Score

- An NLP problem (comparison between 2 texts)
  - Several classical methods (n-grams, Character/Word Error Rate etc)

- Why Semantic Similarity? instead of Syntactic similarity?
  - There are multiple ways to express the same idea in natural language!
  - methods such as n-grams would not be useful / effective

Generating a Semantic Similarity Score

- Model Used: A **ChatGPT 3.5 turbo wrapper**

"You are an intelligent chatbot designed for evaluating the detail orientation of generative outputs for video-based question-answer pairs. "

"Your task is to compare the predicted answer with the correct answer and determine its level of detail, considering both completeness and specificity. Here's how you can accomplish the task:"

"-----"

"##INSTRUCTIONS: "

""- Check if the predicted answer covers all major points from the video. The response should not leave out any key aspects.\n"

"- Evaluate whether the predicted answer includes specific details rather than just generic points. It should provide comprehensive information that is tied to specific elements of the video.\n"

"- Consider synonyms or paraphrases as valid matches.\n"

"- Provide a single evaluation score that reflects the level of detail orientation of the prediction, considering both completeness and specificity."

GPT 3.5 TURBO

Generating a Semantic Similarity Score

- [{"score": 5}, {"q": "a man in a suit and tie standing next to a news desk", "a": "a man in a suit and tie standing next to a news desk", "pred": "The man in the suit and tie is standing next to a news desk, and he is talking to the camera. He is wearing a suit and tie, which suggests that he is likely a news reporter or a guest on a news show. He is talking about the news and providing information to the audience."}]

- [{"score": 3}, {"q": "a street that has power lines down on it", "a": "a street that has power lines down on it", "pred": "The video shows a street with power lines down on it, and a man is talking about the situation. He is wearing a suit and tie and is standing in front of a news desk."}]

- [{"score": 2}, {"q": "a woman standing in the middle of a flooded street", "a": "a woman standing in the middle of a flooded street", "pred": "In the video, a woman is standing in the middle of a flooded street, surrounded by water. She is holding a microphone and speaking to the camera. The man in the suit is also speaking to the camera, and the news reporter is sitting in front of a desk. The news reporter is talking about the flood and the woman standing in the flooded street. The news reporter is also talking about the flood and the woman standing in the flooded street."}]

Generating a Semantic Similarity Score

- After a few rounds of experimenting, we ran into

```
openai.error.RateLimitError: You exceeded your current quota, please check your plan and billing details. For more information on this error, read the docs: https://platform.openai.com/docs/guides/error-codes/api-errors
```

- Decided to stop here for the running and usage of the models, as the **above results from the model shows that the pipeline is functional, and operates as intended.**

GPT 3.5 TURBO

Sorting of Results

- After each group of videos is rated, the scores and corresponding serial numbers are reordered, and the top 10% with the highest scores are screened out.

Problems Solved

- Configured the required environments: pytorch, ffmpeg, pillow, etc.
- Without GPU, Video based performance benchmarking was modified for CPU processing and successfully produced results.
- Modified the format to meet the calling needs of different program modules.
- Submitted a functioning and complete data filtering pipeline integrating 2 SOTA models

Efforts to reduce CO2 emissions

- Optimize download resolution:
- On the premise of ensuring recognition accuracy, we try to reduce the resolution of the downloaded video to reduce the consumption of network bandwidth resources.

- Using Pre-trained Models:
- We opt to not train a SOTA model from scratch which would consume significant computational resources

Further Work/Optimisations

- UI / Web App to display progress/results so far
  - Current scores and serial numbers of all the clips may be dynamically updated / rearranged
  - User can view progress and terminate the data pipeline at any time
  - For easier usage, compilation of to .exe could be considered too
- Comparison between several Video Captioning and Semantic Comparison Models
  - Less of unfinished work, more of an optimisation
  - Always ideal to compare against industry benchmarks and other models (if possible computationally)

Thank You! :)

Produced by Xponential group