Video Captioning

Course Name: EE6180:Advanced Topics in Artificial Intelligence

Student Name: Anirud N, Hrushikesh Ajay Kant
Student Roll Number: CE21B014, EE22B108

---

1 Key Contributions

1.1 Anirud N

- Developed a multi-modal video analysis pipeline integrating diverse frame sampling methods (optical flow, image/caption embeddings) with a cascaded LLaVA architecture for progressive captioning, structured summarization, and narrative elaboration from video content.
- Engineered a configurable and resource-efficient system featuring dynamic model management for optimizing GPU memory, and a robust long-context handling strategy for LLaVA (via input chunking and iterative summarization) to enable detailed video-to-story generation from extensive visual data.

2 Code

The code for the project is given at https://github.com/Hrushikesh9807/immerso.

2.1 1. Hrushikesh Kant

- Developed the initial pipeline based on splitting the videos into clips of t seconds and then randomly sampling the frames for further processing. Incorporated off-the-shelf models like QWEN-2.5VL, LLAMA VLM for initial experimentation and understanding of the project.
- Developed the prompt and logic for image captioning via VLM. The VLM gives a structured output with Background, story, and actions. Wrote the recursive caption summarization pipeline, which takes in the captions and generates the final comprehensive and cohesive story based on the summary. Worked on Optical Flow methodology for sampling frames.

3 Literature Review and a few reproduced experiments

3.1 BLIP

The BLIP model is one of the fundamental models used for image captioning. It takes an image and returns the description of the image. It can have an external prompt or can work without an external text prompt as well. It has 3 parts - Image encoder, Image-text encoder, and Image-text decoder. Our earliest approach was to pass the frames through BLIP and get the summary of generations. The major drawback faced by us was in the description of the caption. Given the size of the model, BLIP just highlights the most important aspect of the image and ignores the subtle characters in the image. Given that we require the caption, which can be fed back to VLM to generate the image, this was not enough.

Ex.

Figure 1: a satanic crucifixion

As we can see, the generated caption is not even close to what is presented in the image. Moreover it is so verbose that it does not give any information about the background as well as another aspects of the image.

3.2 VideoLLava

Video-LLaVA is a large vision-language model (VLM) engineered to comprehend and interact with both static images and dynamic video content in a conversational manner. It extends the capabilities of models like LLaVA (Large Language and Vision Assistant) by incorporating sophisticated mechanisms for unified visual processing and instruction-tuned language generation. The core components include visual encoders, a shared projection layer, and a large language model (LLM). The Video-LLaVA architecture consists of the following key components:

- **LanguageBind Encoders (f_V)**: Used to extract features from raw visual signals (images or videos).
- **Large Language Model (f_L)**: A powerful LLM, such as Vicuna v1.5, serves as the core for reasoning and response generation.
- **Visual Projection Layers (f_P)**: Shared layers that map the extracted visual features into the LLM's input space.
- **Word Embedding Layer (f_T)**: Transforms tokenized textual queries into embeddings compatible with the LLM.

3.3 Video Blip

EILeV modified Video Blip with better training - The key idea is to structure the training data not just as simple `(video, text)` pairs, but as *meta-tasks* or *in-context learning (ICL)-formatted instances*. Each training instance is composed of:

- A set of n demonstration examples:

$$\{(\text{video}_1, \text{label}_1), (\text{video}_2, \text{label}_2), \ldots, (\text{video}_n, \text{label}_n)\}$$

- A query example:

$$(\text{video}_{\text{query}}, ?) \quad \text{or} \quad (\text{video}_{\text{query}}, \text{question}_{\text{query}}, ?)$$

- The expected output (target label or answer) for the query.

This structure encourages the model to learn from contextualized examples during training, mimicking the inference-time behavior of in-context learning.
We tried using VIDEO-BLIP for our use case: 2. Issues: Computationally very less efficient - The original implementation only allows first 10 seconds of the total clip to be uploaded and infered. It also does not provide any descriptive captions as such.

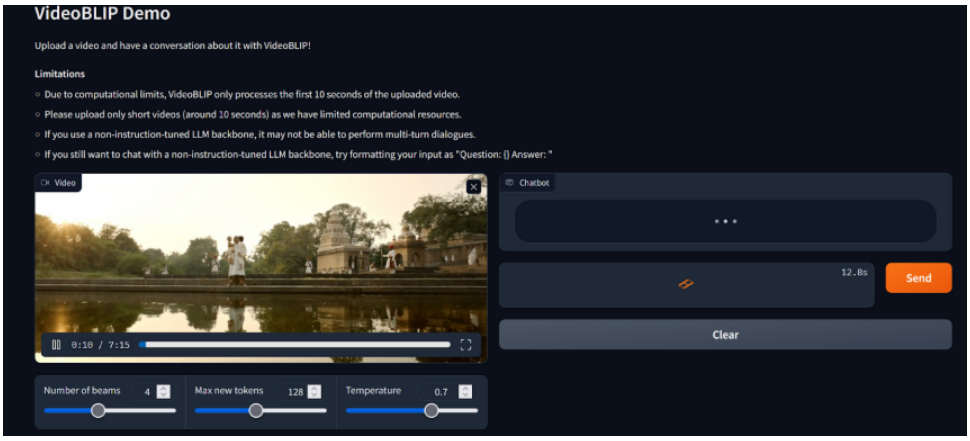


Figure 2: VideoBLip

3.4 VLOG-recent CVPR25 paper

Introduces a narration vocabulary (e.g., "Cut a potato" is a single token/entry). It uses a novel generative retrieval method. This means it retrieves entire, pre-defined narrations relevant to the video content and query, rather than generating them word by word. This results in a significant speedup (claimed 10x) for long video processing. It uses a lightweight language model (GPT-2). A special "retrieval token" is added to the end of the input sequence (video features + query). The language model processes the video and query, and the output embedding of this retrieval token is then used to perform a similarity search (retrieval) against the pre-defined narration vocabulary. This way, the language model's reasoning capabilities (understanding the query in context of the video) guide the retrieval process from the efficient narration vocabulary. The narration vocabulary embeddings themselves are pre-computed and fixed, saving computational cost during inference.

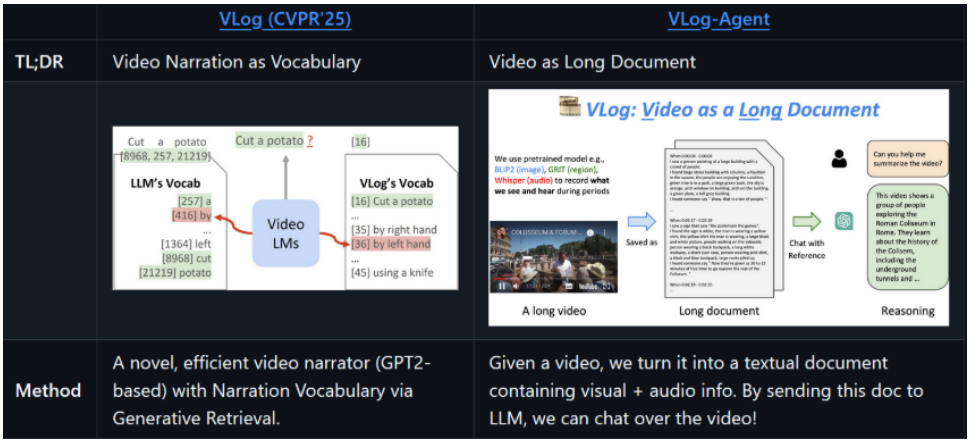


Figure 3: VLOG vs VLOG Agent

Experimenting this on our case: Vlog is a primitive one and gave not so descriptive answers. We realised that it was trained for smaller captions so it may not work for our case.

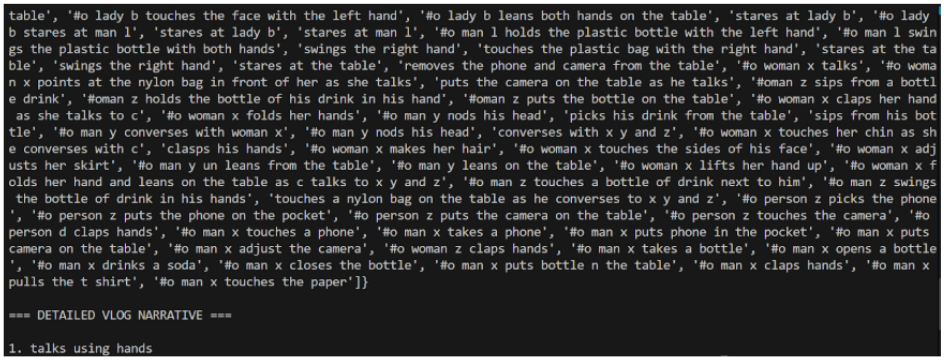


Figure 4: VLOG Performance

3.5 VLOG Agent

Given a video, we turn it into a textual document containing visual + audio info. By sending this doc to LLM, we can chat over the video. This also works with descriptive answers so this is more inclined with our usecase. 3. We have also set the pipeline to working for this. We need one help: **This pipeline works on openai access key for api calls, and we don't have any subscription for the same. If the company has the subscription for the same, this would be really helpful for us to infer this new model as well, and check how this works for our data.** 5.

We realised that the major bottlenecks in the present literature are:

- **Computational complexity:** It is computationally expensive to process an entire video clip. Therefore, we need to sample only those frames that are most relevant to the task.

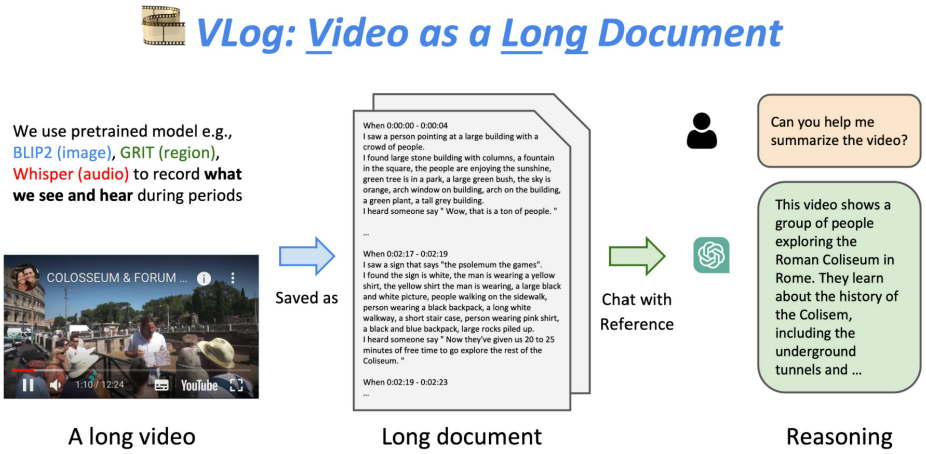


Figure 5: VLOG Agent

- **Domain adaptation:** There is a need to fine-tune models for *Desi movie* conditions. This includes challenges such as incorporating background cultural context, identifying actors in diverse visual conditions, and handling varied cinematographic styles. While our work in this project, we didnt work on fine tuning because we had a very small dataset- having only 30 clips. In future, this would also be a necessary work to look at, to ensure indic scene can be understood.

3.6 Video ReCAP

The video is a combination of the frames along with the audio data. The majority of the videos we used were sampled at 32 fps. The majority of the video captioning models use various techniques to sample the frames and then use the sampled frames for caption generation. The Video ReCAP paper does the same, but at a more fundamental level. The model contains three high-level components: a Video Encoder, a Video Language Alignment, and a Recursive Text Decoder. This paper mainly talks about captioning videos longer than an hour. It clips the full video into smaller videos of varying lengths and then generates the captions for those using the above model. Then, the recursive text decoder recursively generates the summary generated by earlier actions. The recursion occurs a total of three times. The intuition behind such methodology is the psychological understanding of humans of their surroundings. The tasks that we perform are made up of smaller microactions.

3.7 QWEN-2.5 - VL

a. As this model takes the complete video clips as input, it requires a large amount of computing and has higher latency.

b. The recursion in the model was not up to date due to the varying lengths.

4 Proposed Methodology

The proposed methodology for video captioning and understanding is a sequential, multi-stage process. It begins with selecting representative frames from the input video, then generates textual descriptions for these frames, and finally processes these descriptions to create higher-level summaries and narratives. This entire pipeline is executed independently for each configured frame sampling strategy. The flowchart for the methodology can be seen in Fig 6.

4.1 Stage 1: Intelligent Frame Sampling

The goal of this stage is to extract a concise set of frames that best represent the video's content, reducing redundancy. The video is first divided into non-overlapping temporal clips of a predefined duration (e.g., CLIP_DURATION_SECONDS). Within each clip, frames are initially considered at a specific frame_sampling_rate. From these considered frames, one representative frame is selected per clip based on the chosen method. This process is repeated for each sampling strategy. We have implemented and compared the performance of the following 3 sampling strategies:

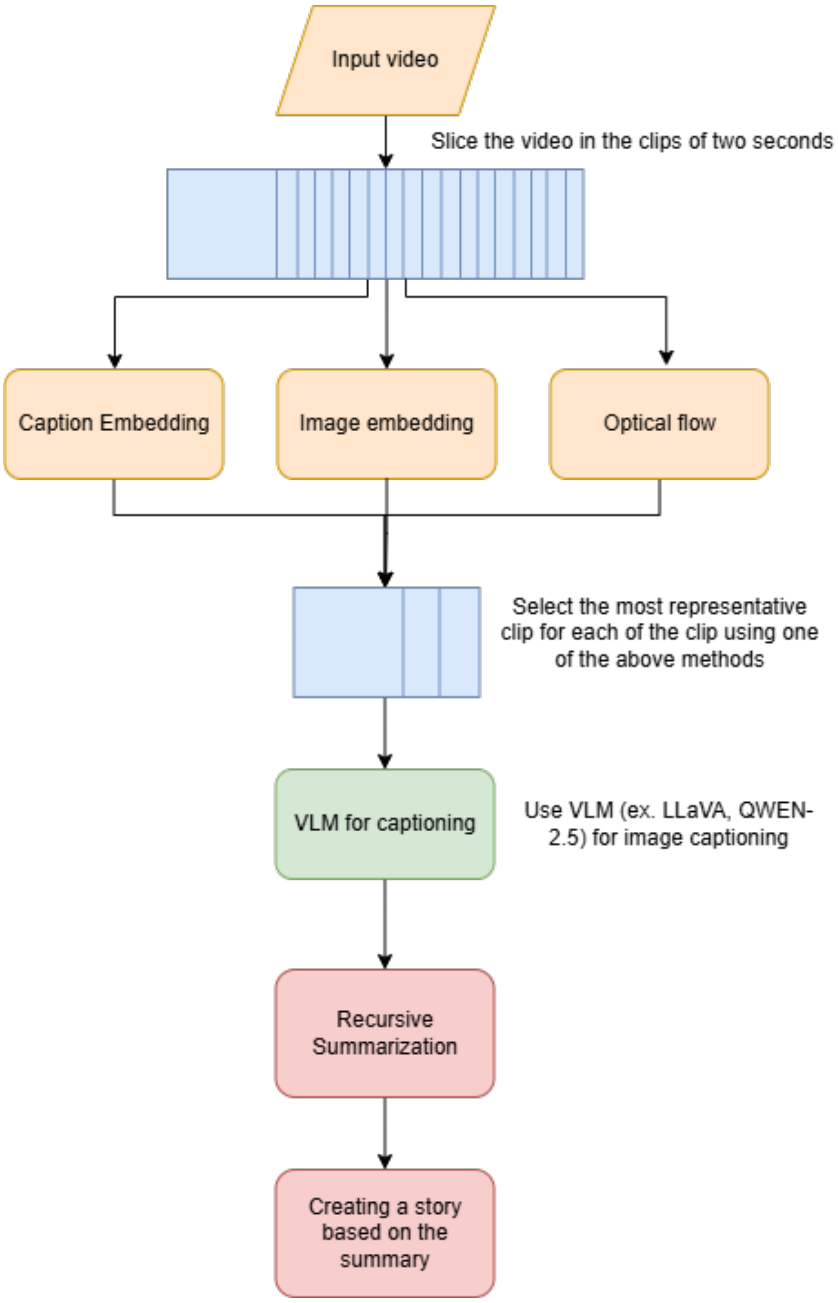


Figure 6: Pipeline flowchart

1. **Optical Flow-based Sampling**:
    - **Models Used**: None explicitly, uses OpenCV's `cv2.calcOpticalFlowFarneback`.
    - **Process**: For each pair of consecutively considered frames within a clip, dense optical flow is computed. The average magnitude of the flow vectors is calculated as a measure of motion.
    - **Selection Criterion**: The frame (from the considered set) that is the *first* frame of the pair exhibiting the highest average motion magnitude within the clip is selected. If no motion is detected or only one frame is considered, the first considered frame is chosen.
    - **Rationale**: This method prioritizes frames that capture significant action or change.
2. **Image Embedding Similarity-based Sampling**:
    - **Models Used**: A Vision Transformer (e.g., 'google/vit-base-patch16-224-in21k") for image feature extraction.
    - **Process**: Each considered frame within a clip is converted to an RGB PIL Image and fed into the pre-trained Vision Transformer to obtain a fixed-size image embedding.
    - **Selection Criterion**: The embeddings of all considered frames in the clip are aggregated, and their mean embedding (centroid) is computed. The frame whose embedding has the highest cosine similarity to this mean embedding is selected.
    - **Rationale**: This method aims to select the frame most visually representative of the average visual content within the clip.
3. **Caption Embedding Similarity-based Sampling**:
    - **Models Used**:
        - An image captioning model (e.g., 'microsoft/git-base-coco") to generate a textual description for each frame.
        - A sentence embedding model (e.g., 'all-MiniLM-L6-v2") to convert these captions into dense vector embeddings.
    - **Process**: For each considered frame within a clip:
        (a) A caption is generated.
        (b) The caption is converted into a sentence embedding.
    - **Selection Criterion**: Similar to image embedding, the mean embedding of all generated captions for the clip's considered frames is computed. The frame whose caption embedding is closest (highest cosine similarity) to this mean caption embedding is selected.
    - **Rationale**: This method selects frames that are most semantically representative of the described content within the clip.

The selected frames from each clip, for a given sampling method, are saved as JPEG images to a dedicated directory. Models used specifically for a sampling method are unloaded after its completion to free resources.

4.2 Stage 2: Image Captioning with LLaVA

For each set of frames sampled by a particular method in Stage 1, this stage generates detailed textual descriptions.

- **Model Used**: LLaVA (e.g., 'llava-1.5-7b-hf"), loaded in float16 precision.
- **Process**:
    1. The LLaVA processor and model are loaded.
    2. Saved frames (from Stage 1 for the current sampling method) are loaded in batches.
    3. Each image is presented to LLaVA with a specific prompt. The prompt structure for image captioning is:

```
USER: <image>
Describe this image in detail.
ASSISTANT:
```

Listing 1: LLaVA Stage 2 Captioning Prompt

4. LLaVA generates a caption for each image. The generation uses `max_new_tokens` (e.g., 150) and `do_sample=False` for deterministic, descriptive captions.
5. The generated text following "ASSISTANT:" is extracted as the caption.

- **Output**: A JSON file containing a list of objects, where each object maps an image filename (ID) to its generated caption.

The LLaVA model used for this stage is then unloaded.

4.3 Stage 3: Hierarchical Text Processing with LLaVA

This stage takes the raw captions from Stage 2 and performs text-to-text generation tasks using LLaVA to produce summaries and an elaborated story. The LLaVA model is loaded again for these text-only tasks.

1. **Initial Text Aggregation and Chunking (Conditional)**:
   - All valid captions generated in Stage 2 (for the current sampling method) are concatenated into a single string, with each caption numbered.
   - The total number of tokens in this concatenated text is checked against `LLAVA_STAGE3_MAX_INPUT_TOKENS_PER_CHUNK_FOR_SUMMARIZATION`.
   - **If too long**: The text is split into smaller chunks. The script uses a character-count-based splitting as an approximation for token-based chunking. Each chunk is then summarized individually by LLaVA. The prompt for chunk summarization is:

```
USER: Concisely summarize the key events, people, and settings from the
    following video frame captions. Focus on the narrative flow if possible:

{text_chunk}

ASSISTANT:
```

Listing 2: LLaVA Stage 3 Chunk Summarization Prompt

     The individual chunk summaries are then concatenated to form the input for the next step.
   - **If short enough**: The full concatenated text of captions is used directly.
2. **Final Structured Summary Generation**:
   - **Input**: The (potentially summarized and chunked) aggregated text from the previous step.
   - **Process**: LLaVA is prompted to generate a structured summary. Input text is truncated if it exceeds `LLAVA_EFFECTIVE_CONTEXT_WINDOW` minus `LLAVA_STAGE3_MAX_NEW_TOKENS_FINAL_SUMMARY`. The core instruction in the prompt is:

```
USER: You are a helpful video analysis assistant. The following text contains
    descriptions or summaries derived from video frames, sampled using the '{
    method_name}' method.
Based *only* on this provided text, provide an overall summary of the entire
    video sequence.
The overall summary must include:
1. The overall background or setting.
2. The main characters or subjects observed.
3. A continuous narrative or story of the actions and events in chronological
    order.

Please structure your answer *exactly* as follows, using these XML-like tags:
<Background>Details about the background and setting.</Background>
<Characters>Details about characters or main subjects.</Characters>
<Story>A continuous narrative of events based on the provided text.</Story>

Provided Text:
```

```
---
{summarized_input_for_final_step}
---

ASSISTANT:
```

Listing 3: LLaVA Stage 3 Final Structured Summary Prompt (Key Instructions)

- **Output**:
  - A JSON file containing the raw text output from LLaVA.
  - A JSON file containing the parsed sections (Background, Characters, Story) extracted using regular expressions from the XML-like tags.

3. **Story Elaboration (Conditional)**:

- **Input**: The "Story" segment extracted from the structured summary above. This step is skipped if no valid story segment is found.
- **Process**: LLaVA is prompted to elaborate on the extracted story segment, making it more descriptive and engaging while maintaining factual consistency. Input text is truncated if it exceeds `LLAVA_EFFECTIVE_CONTEXT_WINDOW` minus `LLAVA_STAGE3_MAX_NEW_TOKENS_STORY_ELABORATION`. The prompt for story elaboration is:

```
USER: You are a skilled storyteller. You are given a narrative segment that was
     extracted from a video analysis. Your task is to elaborate on this
    narrative, making it more descriptive, engaging, and flow like a continuous
     story. Maintain factual consistency with the provided narrative. Do not
    invent new core events not implied by it. Focus on detailing the actions
    and painting a clearer picture.

Provided Narrative Segment:
---
{story_segment_from_s3_summary}
---

Elaborated Continuous Story:
ASSISTANT:
```

Listing 4: LLaVA Stage 3 Story Elaboration Prompt

- **Output**: A JSON file containing the elaborated story text.

For all text generation tasks in Stage 3, a temperature (e.g., 0.7) is used to encourage more creative and varied outputs, and `do_sample=True` is set accordingly. The LLaVA model is unloaded after Stage 3 for the current sampling method. **NOTE:** Initially, we used llama in stages 2 and 3 using groq, and we used a model with higher parameters, but due to the GPU constraints, we had to resort to the models specified here. We do believe that using better models would give even better results.

4.4 Iteration and Final Output

The entire sequence from stage 1 (frame sampling for a specific method) to stage 3 (story elaboration) is repeated for each sampling method enabled in the configuration. This results in a set of captions, structured summaries, and elaborated stories for each frame sampling perspective.

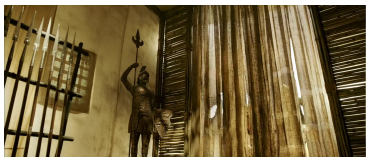
5 Results

We have experimented on a few videos from the given dataset. We have used videos 4, 10, and 22 for the metric calculation. The peculiarity in each of the videos is as follows -

1. 4.mp4 - This video contains a fight scene.
2. 10.mp4 - This video contains a long conversation.
3. 22.mp4 - This video depicts a traditional celebration.

Moreover we have generated captions for 3 other videos downloaded from YouTube to check how the model reacts to videos from various genres. You can see the outputs in the "/results" folder. We did not have the ground truths for the above videos, so we did not compute any metrics on them. They are

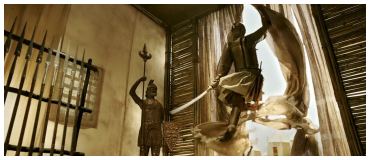
(a) Image 1



(b) Image 2

(c) Image 3

(d) Image 4



(e) Image 5

(f) Image 6

(g) Image 7

(h) Image 8

(i) Image 9

(j) Image 10

(k) Image 11

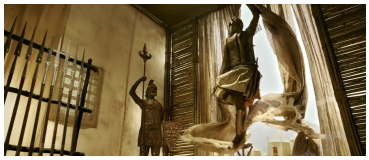
Figure 7: Set of all the images sampled using OPTIC FLOW for the video- 4 fight scene sequence in Bajirao Mastani movie

(a) Caption Embedding Frame 1

(b) Caption Embedding Frame 2

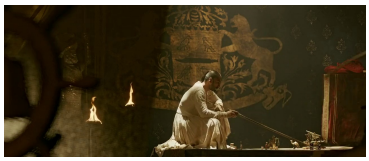
(c) Caption Embedding Frame 3

(d) Caption Embedding Frame 4



(e) Caption Embedding Frame 5

(f) Caption Embedding Frame 6



(g) Caption Embedding Frame 7

(h) Caption Embedding Frame 8

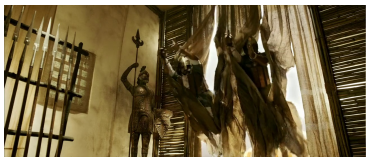
(i) Caption Embedding Frame 9

(j) Caption Embedding Frame 10

(k) Caption Embedding Frame 11

Figure 8: Set of all the images sampled using CAPTION EMBEDDING for the video — 4 fight scene sequence in *Bajirao Mastani* movie

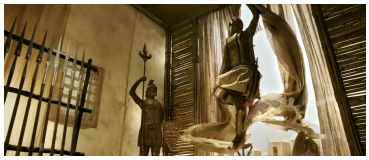
(a) Image Embedding Frame 1



(b) Image Embedding Frame 2

(c) Image Embedding Frame 3

(d) Image Embedding Frame 4



(e) Image Embedding Frame 5

(f) Image Embedding Frame 6

(g) Image Embedding Frame 7

(h) Image Embedding Frame 8

(i) Image Embedding Frame 9

(j) Image Embedding Frame 10

(k) Image Embedding Frame 11

Figure 9: Set of all the images sampled using IMAGE EMBEDDING for the video — 4 fight scene sequence in *Bajirao Mastani* movie

considered an experiment to see how well the model generalizes.
Below are a few frames and output captions for video 4. From Figures 7, 8 and 9, we could see that caption embedding has some sort of redundancy in the frames sampled- Frames 1, 2, and 4 are very similar. So sampling them 3 times is redundant when they have the same meaning. Image embedding, on the other hand, shows a much better sequence of the sampled frames, and this helps in easier understanding of how the video moves with time, and the proper representative images have been observed from the sampled images.
The generated caption is detailed and consists of information about even the minute aspects of the video. The captions generated by image embedding:


```
{
  Image EMbedding: "final_story_for_image_embedding": "In the midst of a bustling
    encampment, a warrior prepared for the impending battle, meticulously adjusting his
    armor and ensuring his sword was securely fastened to his belt. The air was alive
    with tension as he donned his helmet, its visor glinting in the fading light of day
    .\n\nThe scene shifted, and the clash of steel on steel echoed through the air as
    two armored warriors engaged in a fierce duel. The ground shook beneath their feet
    as they exchanged blows, their movements a blur of color and steel. The sound of
    clashing metal and the grunts of exertion filled the air, as the intensity of the
    battle subsided, and the warriors parted, their chests heaving with exhaustion.\n\
    nIn stark contrast, a tranquil scene unfolded, as a man sat relaxed, puffs of smoke
    curling lazily from the pipe clenched between his teeth, his eyes closed in
    contentment. The warmth of the sun on his face and the gentle rustle of leaves in
    the background created a sense of peaceful reprieve from the intensity of the battle
    .\n\nBut the serenity was short-lived, as the scene shifted to a vibrant, theatrical
     performance, where a troupe of costumed actors took to the stage, their elaborate
    attire and dramatic gestures drawing the audience into their world of make-believe.\
    n\nThe setting darkened, and the atmosphere grew somber, as a lone figure, broom in
    hand, swept the cold, grey floor of a dungeon-like chamber, the only sound the soft
    swish of the bristles against the ground.\n\nThe scene shifted once more, and a
    weary traveler, or perhaps a wanderer, sat on a simple bed, his eyes cast downward,
    lost in thought, the soft, golden light of the room enveloping him in its
    tranquility.\n\nNext, a majestic statue came into view, a figure of a warrior,
    shield in hand, standing proudly, possibly part of a larger work of art, its beauty
    and grandeur awe-inspiring.\n\nBut the tranquility was short-lived, as the scene
    shifted, and a warrior stood, his eyes fixed intently on some point in the distance,
     his hand grasping the hilt of his sword, his body tense with anticipation, as if
    preparing for a fight or honing his sword skills.\n\nThe scene shifted once more,
    and a figure, kneeling on the ground, held a sword, his eyes fixed on the blade, his
     body tense with concentration, possibly preparing for battle or practicing his
    sword skills.\n\nFinally, the story culminated in a dramatic, climactic scene, as a
    warrior, or samurai, stood poised, ready for action, in a dark, dimly lit room, the
    only sound the soft rustle of his clothing, the air thick with tension, as if part
    of a play or performance, the audience holding its collective breath in anticipation
     of the action to come.",
  "source_story_segment": "The story begins with a man preparing for a battle or a
    historical reenactment. The scene then shifts to a dramatic fight between two
    armored warriors. Next, a man is seen relaxing with a pipe, possibly after the
    battle. The story then takes a turn to a theatrical performance, where a group of
    costumed actors are seen on a stage-like setting. \n\nThe story shifts to a dungeon-
    like setting, where a man is seen sweeping the floor with a broom. The scene then
    cuts to a man sitting on a bed, possibly a traveler or someone on a journey. Next, a
     statue of a man holding a shield is seen, possibly part of a larger artwork. \n\
    nThe story then takes a dramatic turn, with a man preparing for a fight or practice
    his sword skills. The scene then shifts to a man kneeling on the ground, holding a
    sword, possibly preparing for a battle or practicing his sword skills. \n\nFinally,
    the story ends with a dramatic scene, where a man dressed as a warrior or samurai is
     seen standing in a dark room, ready for action. The scene appears to be part of a
    play or a performance."
}
```


The caption embedding results:


```
{
  "raw_summary_for_caption_embedding": "Here is the structured summary:\n\n<Background>
    The setting appears to be a combination of medieval and modern elements, with rooms,
     chairs, tables, and windows. There are also hints of a performance or stage setting
    , with a dark room and a curtain. The atmosphere ranges from intense and action-
    packed to calm and focused, with a sense of drama and intensity building throughout
    the sequence.</Background>\n\n<Characters>The main character is a man who appears in
     various costumes and settings, often wearing a white robe or medieval armor. He is
    seen engaging in various activities, such as sword fighting, smoking a pipe, and
```

```
    holding a walking stick. He is often the main focus of the scene, suggesting that he
     is the protagonist of the story. There are also hints of other characters, such as
    a person observing a statue, a bird, and other people in the background of certain
    scenes.</Story>\n\nThe story begins with a intense sword fight between two people in
     medieval costumes. The scene then shifts to a calm and focused atmosphere, where
    the main character is seen sitting on the ground, wearing a white robe and holding a
     pipe. He then appears in another scene, holding a stick and sitting near a window,
    possibly posing for a picture. The story then takes a dramatic turn, with the main
    character seen holding a sword, standing in a room with a tapestry on the wall. He
    is then shown in a defensive stance, kneeling on the ground and holding a sword. The
     story builds up to a dramatic and intense moment, where the main character is seen
    standing in a dark stage, ready for action. The final scene shows another intense
    sword fight, similar to the first scene, suggesting that the story has come full
    circle.</"
}
```


The optic flow results :


```
    {
  "final_story_for_optical_flow": "Here's the elaborated continuous story:\n\nIn a flash
     of steel and valor, our tale begins with a knight standing tall and proud, his
    armor gleaming in the light, his eyes fixed intently on some unseen foe. He is ready
     for battle, his muscles coiled like a spring, his sword at the ready.\n\nBut in a
    sudden shift, the scene changes, and we find ourselves in a tranquil setting, where
    a man, serene and focused, goes about some quiet task. His white robe is a stark
    contrast to the armor of the knight, and his movements are deliberate and calm.\n\
    nThe pace quickens, and we are transported to a world of performance and art, where
    an actor, brandishing a stick as a prop, takes center stage. The stick becomes a
    sword, a magic wand, a conductor's baton, as the actor weaves a spell of imagination
     around us.\n\nNext, we find ourselves in a historical setting, surrounded by the
    trappings of a bygone era. A statue of a warrior stands proud, its stone eyes gazing
     out into the distance, its sword at the ready.\n\nBut drama intervenes, and we are
    shocked into attention by a woman in distress, clutching a cell phone as if it were
    a lifeline. Her eyes are wide with fear, her face pale with anxiety.\n\nThe scene
    shifts once more, and we find ourselves back with the man in the white robe, this
    time sitting in a chair, holding the stick that was once a sword, a magic wand, a
    conductor's baton. The contrast between his calm demeanor and the woman's distress
    is striking.\n\nThe pace quickens again, and we find ourselves in a world of action
    and adventure. A statue of a man holding a sword becomes the backdrop for a group of
     warriors, their swords flashing in the light, engaged in a lively sword fight. The
    clash of steel on steel echoes through the air, the warriors' cries and shouts
    filling the space.\n\nThe scene changes, and we find ourselves in a theater, where a
     man in medieval armor stands on a stage, the spotlight shining down on him. He is a
     knight, a warrior, a hero, and his presence is commanding.\n\nThe drama deepens,
    and we find ourselves in a darkened room, where two men are engaged in a fierce
    sword fight. The air is thick with tension, the only sound the clash of steel on
    steel.\n\nFinally, the story concludes with a scene of high drama and action, as two
     warriors, dressed in medieval finery, engage in a fierce sword fight in a dungeon-
    like setting. The sound of steel on steel echoes through the space, the warriors'
    cries and shouts filling the air, as the drama reaches its climax.",
  "source_story_segment": "The story that emerges from the captions is one of drama,
    action, and performance. It begins with a knight standing ready for battle, and then
     shifts to a man in a white robe, possibly engaged in a calm and focused activity. \
    n\nThe scene then changes to a performer or actor, holding a stick as a prop, and
    then to a statue of a warrior, standing in a historical setting. \n\nThe story takes
     a dramatic turn with a woman in distress, holding a cell phone, and then shifts to
    a man in a white robe, sitting in a chair, holding a stick. \n\nThe scene then
    changes to a statue of a man holding a sword, and then to a group of warriors
    engaged in a lively sword fight.\n\nThe story continues with a scene from a play,
    featuring a man in medieval armor, standing on a stage, and then shifts to two men
    engaged in a sword fight in a darkened room.\n\nFinally, the story concludes with
    two people dressed in medieval armor, engaging in a sword fight in a dungeon-like
    environment.\n\nThroughout the sequence, the story is marked by a sense of drama,
    action, and performance, with medieval elements and historical settings providing
    the backdrop for the action to unfold."
}
```


Comparing the three caption generation methods, the `Image Embedding` approach demonstrably provides the most accurate and detailed description, aligning closely with its own source segment without introducing hallucinations. It better describes scenes of warrior preparation, combat, and theatrical performance with elaborations. In contrast, the `Optical Flow` method suffers from a significant hallucination, erroneously introducing a "woman in distress, clutching a cell phone," an object clearly out of place in a presumably historical movie setting like Bajirao. The `Caption Embedding` result, while

not directly repeating the cell phone error in its raw summary, appears to be influenced by a flawed source segment (similar to Optical Flow's) and introduces potentially unverified details like "modern elements" and a "walking stick." Thus, the `Image Embedding` stands out as the most reliable for generating contextually appropriate and detailed video captions based on the provided examples.

6 Comparison with Ground Truths

Things to note- this comparison based on ground truths may be misleading - because - the ground truths are very short and they are actually not very proper. an example:

```
In the clip from the movie BajiraoMastani visuals depict a statue of armored figures
, a woman on a horse, and a man in a mask, suggesting a tense atmosphere with
potential conflict brewing.
```

This is the ground truth for the video 4.mp4, which is a fight scene. As seen ,this ground truth is very small and not so descriptive when compared to the generated captions - so the similarly scores calculated in the coming sections might be misleading.

6.1 METEOR

METEOR score - The METEOR (Metric for Evaluation of Translation with Explicit Ordering) metric is used to evaluate the quality of machine-generated text by comparing it to human-written reference texts. METEOR aligns words in the generated output with the reference using a variety of matching strategies, including exact matches, stemmed word forms, synonyms, and even paraphrases. It then calculates a harmonic mean of precision and recall. This makes METEOR more sensitive to the actual meaning and structure of the text.

Video	Caption embedding	Image Embedding	Optic Flow
4.mp4 - fight scene	0.24	0.16	0.14
10.mp4 conversation	-	-	0.30
22.mp4 - celebration	0.13	0.14	0.16

Table 1: METEOR

6.2 ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics used to compare an automatically generated summary or translation against one or more reference texts. It mainly focuses on recall—how much of the reference content is captured by the candidate output. We have used the following 4 scores. ROUGE-N measures the number of matching n-grams between the model-generated text and a human-produced reference. We have used N=1 and 2. We have also used ROGUE-L, which is based on the longest common subsequence. The scores are as follows.

6.2.1 ROUGE-1

Video	Caption embedding	Image Embedding	Optic Flow
4.mp4 - fight scene	0.17	0.08	0.07
10.mp4 - conversation	-	-	0.22
22.mp4- celebration	0.07	0.08	0.11

Table 2: ROGUE-1

Video	Caption embedding	Image Embedding	Optic Flow
4.mp4 - fight scene	0.01	0.02	0.03
10.mp4 - conversation	-	-	0.06
22.mp4 - celebration	0.01	0.02	0.01

Table 3: ROGUE-2

6.2.2 ROUGE-2

6.2.3 ROUGE-L

Video	Caption embedding	Image Embedding	Optic Flow
4.mp4 - fight scene	0.13	0.06	0.06
10.mp4 - conversation	-	-	0.14
22.mp4 - celebration	0.05	0.08	0.08

Table 4: ROUGE-L

7 Experimentation

- We used videos other than the provided ones to look for generalization of the model. The videos are present in the folder 'videos_to_test_valuate'. Bwlow are the two captions generated from the videos.

7.1 Horror Genre

We evaluated our method on a horror based youtube video - (https://www.youtube.com/watch?v=BYysLg8Hzlo).

```
"parsed_summary": {
        "background": "The video series consists of various scenes featuring a diverse range of people, settings, and activities. The main focus is on a woman wearing a scarf, who makes funny faces throughout the video. The scenes include candid moments, relationships, and daily life events, with objects such as chairs, a clock, a handbag, a cell phone, a piece of paper, and a door.",
        "characters": "The main subjects of the video are a woman wearing a scarf who makes funny faces, and other people who appear in the background or foreground of the scenes.",
        "story": "The video begins with a woman wearing a scarf making a funny face in front of a mirror, and she appears in multiple instances throughout the video expressing different emotions or making funny faces. In one scene, she is leaning against a wall with her mouth open, and in another, she is in an elevator making a funny face. The overall theme of the video is a mixture of candid moments, relationships, and daily life events featuring the woman wearing a scarf and other people in the background or foreground."
    }
```

Listing 5: Horror

As you can see from the captions above- the algorithm performs, very poorly in identifying the context that this is a horror movie - claiming about "funny faces" and "mirror" while this is a scene in a lift/elevator, where reflection is visible. That is why this description is bad, as it hallucinates a lot and does not get the context of a horror movie where the women gets stuck in the elevator/lift.

7.2 Comedy Genre

This was taken from the Mr. Bean movie (https://www.youtube.com/watch?v=FG3ohfDASao), where Mr. Bean is doing a surgery, and this involves a lot of fun elements and satire.

```
"parsed_summary": {
        "background": "The video contains a series of scenes featuring people in various medical settings, showcasing different attire styles and interactions among individuals. The main subjects appear to be medical professionals and patients, with a focus on their professionalism and care for one another.",
        "characters": "The main characters in the video include doctors, nurses, patients, and other medical professionals. They are depicted in different attire, such as scrubs, lab coats, suits, and other casual clothing, reflecting the diverse nature of their roles and responsibilities.",
```

```
    "story": "The video begins with scenes of people in various medical settings,
   with some dressed in scrubs and others in suits. In one scene, a man in a suit is
   standing in a room and examining a patient, while another person is plugging in an
   electrical outlet. This sets the tone for the video, emphasizing the professionalism
    and dedication of the medical professionals.\n\nLater scenes depict a busy city
   street filled with traffic, showcasing the bustling and lively atmosphere of urban
   life. The video transitions back to the medical setting with scenes of doctors and
   medical professionals in scrubs, wearing masks, and smiling, indicating a caring and
    professional The video then displays a man in a red tie sitting on a couch,
   suggesting a more casual and relaxed atmosphere. This scene is followed by a series
   of scenes featuring people in bed, possibly sleeping or resting, and a group of
   doctors gathered around a patient in a hospital setting. The video concludes with a
   scene in an airport terminal, with people wearing ties and carrying handbags.\n\nThe
    overall narrative of the video revolves around the interactions and experiences of
   medical professionals and their patients in various settings, highlighting the
   diverse range of attire styles and the dedication of the medical professionals to
   their work."
   },
```

Listing 6: Horror

As you can see, it doesn't understand the context at all, and considers it as a serious medical emergency with no hint of fun elements in this captions. For instance, it thinks that Mr. Bean is a professional doctor doing the surgery 'meticulosly' while we know that this is not the case, and the surgery was something fun to laugh at.
As we can see above, the frames are not enough to capture the overall emotion in the video. The first video is a scary scene from the horror movie, whereas the second scene is from the Mr Bean series. The model is correctly analyzing frames and summarizing frames, but is not able to get the storyline of the video in full. We humans have a baseline that defines what a horror or comedy is; for the LLM, everything is the same. We need to focus more in the long-form understanding and correlation between he frames. One of the approaches can be to create a super image by concatenating frames in nXn and use it for captioning along with the time data.

7.3 Requiem for a Dream

We also evaluated our performance on a particular clip of the movie from - Requiem for a Dream which is a psychological movie that is even tough for humans to understand. The clip we used is from - (https://youtu.be/5eTVGaa9KwA?feature=shared). As you can see, this is a very dynamic clip, which frames change every 1-2 seconds, which makes it very difficult for our algorithm to understand what's going on.
Caption embedding gave the results as:

```
   {
 "final_story_for_caption_embedding": "Here is the elaborated continuous story:\n\nThe
   morning begins with a quiet ritual, as a person carefully opens a pill bottle, the
   transparent plastic container revealing a multitude of tiny, colorful pills inside.
   The gentle rustle of the bottle and the soft clinking of the pills as they settle
   back into their container suggest a daily routine, perhaps a vital part of
   maintaining health and well-being.\n\nIn a warm and inviting atmosphere, an older
   person is shown savoring a steaming hot beverage, possibly a freshly brewed cup of
   coffee. The gentle steam rising from the rim adds to the cozy ambiance, hinting at a
    relaxing morning routine.\n\nThe scene then transitions to a serene outdoor setting
   , where a dog lies comfortably on the lush green grass, its fur glistening in the
   soft sunlight. As the camera lingers, it suddenly stands up, shaking off the slumber
   , and stretches its legs, as if awakening from a peaceful nap.\n\nThe focus then
   shifts to a more intimate moment, as a person's tongue is shown savoring a morsel of
    food, perhaps during a leisurely meal. The close-up shot emphasizes the simple
   pleasure of enjoying a meal, highlighting the senses of taste and smell.\n\nIn a
   different setting, an older woman is shown meticulously sorting pills at a table,
   her fingers moving deftly as she organizes the medication with precision and care.
   The shot emphasizes her attention to detail, underscoring the importance of
   maintaining health and well-being.\n\nThe narrative then takes a modern turn, as a
   person is shown holding a cell phone, their gaze fixed on the screen, perhaps
   checking updates or messages. This is followed by a close-up shot of a person's eye,
    the iris and pupil sharply defined, as if highlighting the importance of staying
   connected and awareness.\n\nIn stark contrast, a hand with grimy fingernails is
   shown, the dirt and grime a testament to hard work or outdoor activities. This
   gritty realism is quickly replaced by a serene and calming image \u2013 a beautiful
   black and white photograph of a waterfall, its gentle mist and soft focus evoking a
   sense of tranquility and peace.\n\nThe story comes full circle, as a person is shown
```

```
    holding a pill bottle once more, similar to the opening scene. The repetition
    emphasizes the daily routine, the importance of maintaining health and well-being
    that was woven throughout the narrative.\n\nThe final scenes bring a sense of
    closure, as a person is shown brushing their teeth, the gentle scrubbing motion and
    the soft foam underscoring the daily rituals that bookend the day. The closing shot,
    a close-up of a delicious-looking donut and doughnut holes, adds a touch of
    sweetness and indulgence, hinting at the rewards and pleasures that come with taking
    care of oneself."
}
```

As you can see, in this case, our algorithm performs very well, explaining every scene in detail. One problem is that the LLM still struggles to learn the connection between two frames that are far apart in time - ie, the connection between frame 1 and frame 10, for example. It considers this as two separate scenes, but in real they are the same scene, with a different scene in between these frames. This is another drawback observed - trying to identify the correlation between the frames that are distant apart in time is a challenge faced by this model.

8 Future Work

Video captioning consists of three important aspects, as mentioned above in the pipeline. We need to improve each of them to achieve greater accuracy.

1 Sampling - Identifying important frames - We need to increase the sampling rate for higher amounts of data being processed. The current sampling rate of 1 and 2 seconds assumes that any substantial action takes at least 1 second to manifest itself. Whereas there are certain specific niche use cases where this may not work. Ex. Firing a bullet. The impact might give the LLM an idea of what might have happened, but it won't record the bullet being fired.

2 Image captioning - Captioning the frame - There has been a lot of work done in image captioning. With the intent that we are going to create a movie from the text, the caption should include not-so-important details from the image as well. We have tried promoting the VLM for such information. Another important aspect in image captioning is character identification. In all the results until now, the LLM answers the characters in third person - "a person", "a fighter", etc. For combining longer videos, we need to map the faces of the character externally or ask LLM to assign a name that can be used further.

3 Summarization - Summarizing the captions of all the frames and generating a coherent caption based on the earlier ones - General summarization issues hamper he caption summarization. Longer context becomes an issue, specifically when the summarization captions do not have the earlier context. Maintaining the background information and other details becomes an issue as there are lots of scenes, and LLM tries to include and exclude some of them.