Abstract

The project presents a cutting-edge solution for dense video captioning that leverages the power of state-of-the-art BLIP pretrained models while seamlessly integrating face recognition technology to operate in real-time. Dense video captioning aims to automatically generate descriptive captions for every frame in a video, offering enhanced accessibility and comprehension for various applications such as video indexing, content recommendation, and assistive technology. The project enhances the dense video captioning system with face recognition technology, enabling the identification and tracking of individuals within the video frames. This added layer helps the generation of contextually relevant captions by considering the presence and activities of recognized individuals. Face recognition helps bridge the semantic gap between visual content and textual descriptions, enriching the overall captioning quality.

Table of Contents

1. Introduction

Real-time dense video captioning (RT-DVC) is a challenging but rapidly developing field with a wide range of potential use cases and applications. Dense Video Captioning systems aim to generate comprehensive and informative captions for every frame of an input video, in real time. This requires a deep understanding of the video content, as well as the ability to generate natural language descriptions quickly and efficiently.

One of the most promising applications of Dense Video Captioning is in the field of accessibility. Dense Video Captioning can be used to provide real-time captions for videos for people who are deaf or hard of hearing. This can make videos more accessible and inclusive for a wider audience. For example, Dense Video Captioning could be used to generate captions for live TV broadcasts, educational videos, and social media videos.

Dense Video Captioning can also be used to improve the efficiency and effectiveness of video surveillance and security systems. Dense Video Captioning systems can be used to monitor and analyze surveillance footage in real time, generating captions that can help human operators to identify suspicious activity and respond quickly to incidents. For example, Dense Video Captioning could be used to monitor security cameras at airports, train stations, and other public places.

Real Time Dense Video Captioning is also a promising technology for robotics applications. Dense Video Captioning systems can be used to give robots a better understanding of the world around them, helping them to navigate more safely and interact with people more effectively. For example, Dense Video Captioning could be used to generate captions for video footage from a robot's cameras, helping the robot to understand its surroundings and identify objects and people.

In addition to these specific applications, Dense Video Captioning also has the potential to revolutionize the way we interact with videos in general. Dense Video Captioning systems could be used to generate captions for videos in multiple languages, making videos more accessible to people from all over the world. Dense Video Captioning could also be used to create new and innovative types of video content, such as interactive videos and videos that can be searched and navigated using natural language queries.

Overall, Dense Video Captioning is a powerful and versatile technology with a wide range of potential applications. As Dense Video Captioning systems become more accurate and efficient, they are likely to have a major impact on the way we interact with videos and the world around us.

Here are some additional thoughts on the potential applications of Dense Video Captioning in robotics and other related fields:

1. Robotics: Dense Video Captioning could be used to give robots the ability to describe their surroundings to humans in real time. This could be useful for a variety of tasks, such as helping robots to navigate unfamiliar environments, interact with people more effectively, and perform complex tasks under human supervision. For example, a robot equipped with Dense Video Captioning could be used to explore a disaster zone and provide real-time updates to human operators on the ground.
2. Video editing: Dense Video Captioning could be used to automate the process of generating captions for videos during the video editing process. This could save video editors a lot of time and effort, and it could also help to make videos more accessible and inclusive for a wider audience. For example, a video editor could use Dense Video Captioning to generate captions for a video in multiple languages, or to provide captions for a video that is spoken too quickly for viewers to understand.

3. Education: Dense Video Captioning could be used to create educational videos that are more engaging and accessible for students. For example, Dense Video Captioning could be used to generate captions for educational videos in multiple languages, or to provide captions for videos that are spoken too quickly for students to understand. Dense Video Captioning could also be used to create interactive educational videos that allow students to learn at their own pace and explore different topics in depth.
4. Entertainment: Dense Video Captioning could be used to create more engaging and immersive entertainment experiences. For example, Dense Video Captioning could be used to generate subtitles for foreign language films, or to provide real-time captions for live events. Dense Video Captioning could also be used to create new and innovative types of entertainment, such as interactive videos and videos that can be searched and navigated using natural language queries.

2. Literature Review

2.1 Early Work

The earliest work on DVC focused on developing methods to generate single captions for videos. One of the first successful methods was proposed by Venugopalan et al. (2015), who used a recurrent neural network (RNN) to generate captions from video frames. However, this method was limited to generating single captions for entire videos and could not generate captions for individual temporal segments.

2.2 Recent Developments

In recent years, there has been a significant amount of progress in the field of DVC. Researchers have developed a variety of new methods that can generate multiple, descriptive captions for videos, each of which corresponds to a short temporal segment. Some of the most notable recent advances in DVC include:

Temporal attention: Temporal attention mechanisms allow DVC models to focus on the most relevant video frames when generating captions. This has led to significant improvements in the accuracy and fluency of DVC captions.

Bidirectional decoding: Bidirectional decoding allows DVC models to generate captions in both forward and backward directions. This has helped to improve the coherence and consistency of DVC captions.

Weakly supervised learning: Weakly supervised learning methods allow DVC models to be trained on unlabeled video data. This has made it possible to train

DVC models on large datasets of videos, which has led to further improvements in accuracy.

State-of-the-Art Models

2.3 State-of the art Models

Some of the state-of-the-art DVC models and research include:

DenseCap: DenseCap (Venugopalan et al., 2015) [1]is a classic DVC model that uses an RNN to generate captions from video frames.

LSTNet: LSTNet (Wang et al., 2018) [2]is a DVC model that uses a temporal attention mechanism to focus on the most relevant video frames when generating captions.

VSE++: VSE++ (Shen et al., 2018) [3]is a DVC model that uses bidirectional decoding to generate captions in both forward and backward directions.

TCN-DVC: TCN-DVC (Zhao et al., 2019)[4] is a DVC model that uses a weakly supervised learning approach to train on unlabeled video data.

Streamlined Dense Video Captioning [5] proposes a new model for dense video captioning that is more efficient than previous models. The authors show that their model can achieve state-of-the-art results on the MSR-VTT dataset while being significantly faster than previous models.

Dense-Captioning with Multi-Scale Temporal Attention [6]proposes a new temporal attention mechanism for dense video captioning. The authors show that their model can achieve state-of-the-art results on the MSR-VTT dataset by using multi-scale temporal attention to capture long-range dependencies in videos.

Dense Video Captioning with Progressive Attention [7] proposes a new progressive attention mechanism for dense video captioning. The authors show that their model can achieve state-of-the-art results on the MSR-VTT dataset by using progressive attention to gradually focus on the most relevant parts of a video as it generates a caption.

Dense Video Captioning with Transformer [8] proposes a new model for dense video captioning that is based on the Transformer architecture. The Transformer architecture has been shown to be very effective for natural language processing tasks, and the authors show that it can also be used to achieve state-of-the-art results on the MSR-VTT dataset.

Dense Video Captioning with Hierarchical Transformer [9] proposes a new model for dense video captioning that is based on a hierarchical Transformer architecture. The authors show that their model can achieve state-of-the-art results on the MSR-VTT dataset by using a hierarchical Transformer architecture to capture both global and local dependencies in videos.

2.4 Previous works and problems with real time Dense Video Captioning

There are a few reasons why the models in research mentioned above are not yet used for real-time dense video captioning (DVC).

One reason is that these models are computationally expensive to train and run. They require large amounts of data and powerful hardware to train, and they can be slow to run on real-time devices.

Another reason is that these models are not yet as accurate as they could be. They can sometimes generate captions that are incorrect or incomplete, especially for complex or fast-moving videos.

Finally, these models are still under development. Researchers are actively working on improving the accuracy, efficiency, and scalability of DVC models. As these models continue to improve, they are likely to become more widely used for real-time applications.

Here are some specific challenges that need to be addressed before these models can be used for real-time DVC:

Computational cost: Training and running these models is computationally expensive, requiring large amounts of data and powerful hardware.

Accuracy: These models are not yet as accurate as they could be, especially for complex or fast-moving videos.

Scalability: These models need to be able to scale to real-time applications, which means that they need to be able to generate captions quickly and efficiently.

Researchers are working on addressing these challenges, and they are making progress. For example, researchers are developing new training algorithms and hardware optimizations to make DVC models more efficient. They are also working on improving the accuracy of DVC models by developing new architectures and training methods.

As these challenges are addressed, DVC models are likely to become more widely used for real-time applications. Real-time DVC has the potential to revolutionize the way we interact with videos, making them more accessible and informative for everyone.

3. Research Gap

3.1 Expected challenges in Dense Video Captioning

Dense video captioning (DVC) is a challenging task that aims to generate multiple informative and diverse captions for a given video, describing all the important events happening in the video. DVC has a wide range of applications, such as video search, video retrieval, and video understanding. Despite the recent advances in DVC research, there are still a number of research gaps that need to be addressed.

- Limited training data: DVC models require a large amount of training data, which can be expensive and time-consuming to collect. This is because DVC models need to learn to detect and describe events in videos, which requires a deep understanding of both visual and textual data.
- Lack of generalization: DVC models trained on one dataset often do not generalize well to other datasets. This is because DVC models need to learn to adapt to different video domains, such as sports, movies, and documentaries.
- Limited ability to handle complex events: DVC models often struggle to handle complex events that involve multiple objects and actions. This is because DVC models need to learn to reason about the temporal relationships between different events.
- Lack of explainability: DVC models are often black-box models, which makes it difficult to understand how they generate captions. This is a problem, especially for safety-critical applications, such as self-driving cars and medical imaging.

3.2 Challenges in literature

In the paper "Zero-Shot Dense Video Captioning by Jointly Optimizing Text and Moment" (2023) [12], the authors note that "existing DVC methods typically require a large amount of annotated training data, which can be expensive and time-consuming to collect." They propose a zero-shot DVC method that does not require any training data, but instead learns to detect and describe events in videos by jointly optimizing text and moment proposals.

In the paper "Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning" [13] (2023), the authors note that "existing DVC methods often struggle to generalize to different video domains, such as sports, movies, and documentaries." They propose a large-scale pre-training method for DVC models, which allows models to learn general-purpose visual and linguistic representations that can be transferred to different video domains.

In the paper "A Better Use of Audio-Visual Cues: Dense Video Captioning with Bi-modal Transformer" [14] (2020), the authors note that "existing DVC methods typically only use visual features from videos." They propose a bi-modal DVC model that uses both visual and audio features from videos. This allows the model to better understand the context of videos and generate more accurate and informative captions.

In the paper "Multi-Modal Dense Video Captioning"[15] (2020), the authors note that "existing DVC methods often struggle to handle complex events that involve multiple objects and actions." They propose a multi-modal DVC model that uses both visual and textual features from videos. This allows the model to better reason about the temporal relationships between different events and generate more comprehensive captions. Overall, DVC is a challenging task with a number of open research gaps. Addressing these research gaps will require developing new methods for training DVC models with less data, improving the generalization ability of DVC models, and developing more explainable DVC models.

4. Methodology

In our approach to Dense Video Captioning, we've adopted a multimodal strategy that harnesses the power of multiple models, each specialized in extracting distinct types of information from video frames. This multifaceted approach is designed to generate highly relevant and contextually rich captions for each frame of a video sequence. Instead of relying solely on a single model, we leverage a combination of models that excel in various aspects of understanding visual content.

These models include computer vision models capable of describing the environment, recognizing actions, and identifying faces within the frames. By combining the outputs from these diverse models, we aim to capture a more comprehensive understanding of the video content, resulting in captions that not only describe the visible elements but also provide insights into the relationships and dynamics within the video.

This multimodal approach enhances the accuracy and depth of our video captions, making them more informative and valuable for a wide range of applications, including video indexing, summarization, and accessibility.

4.1 Introduction to BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

BLIP [10] is a *Vision-Language Pre-training* (VLP) model that can perform both vision and language tasks, such as image captioning, visual question answering, and image-text retrieval. It is pre-trained on a massive dataset of image-text pairs, including both clean and noisy data. BLIP is able to effectively learn from noisy data by bootstrapping the captions, where a captioner generates synthetic captions and a filter removes the noisy ones.

BLIP is pre-trained on a massive dataset of image-text pairs called DiffusionDB. DiffusionDB is a text-to-image prompt dataset that contains 14 million images generated by Stable Diffusion using prompts and hyperparameters specified by real users. BLIP also uses a bootstrapping technique to learn from noisy data. The bootstrapping technique works by generating synthetic captions for web images and then filtering out noisy image-text pairs. This allows BLIP to learn from a much larger dataset than would be possible otherwise. The bootstrapping technique is one of the key reasons why BLIP is able to achieve state-of-the-art results on many vision-language benchmarks.

The BLIP model architecture consists of three main components:

a. Vision encoder: The vision encoder is a pre-trained ViT (Vision Transformer) model that extracts visual features from the input image.
b. Language encoder: The language encoder is a pre-trained Transformer model that encodes the input text into a sequence of embeddings.

c. Querying transformer (Q-Former): The Q-Former is a transformer model that bridges the gap between the visual and language modalities. It takes the visual features from the vision encoder and the text embeddings from the language encoder as input and produces a sequence of multimodal representations.

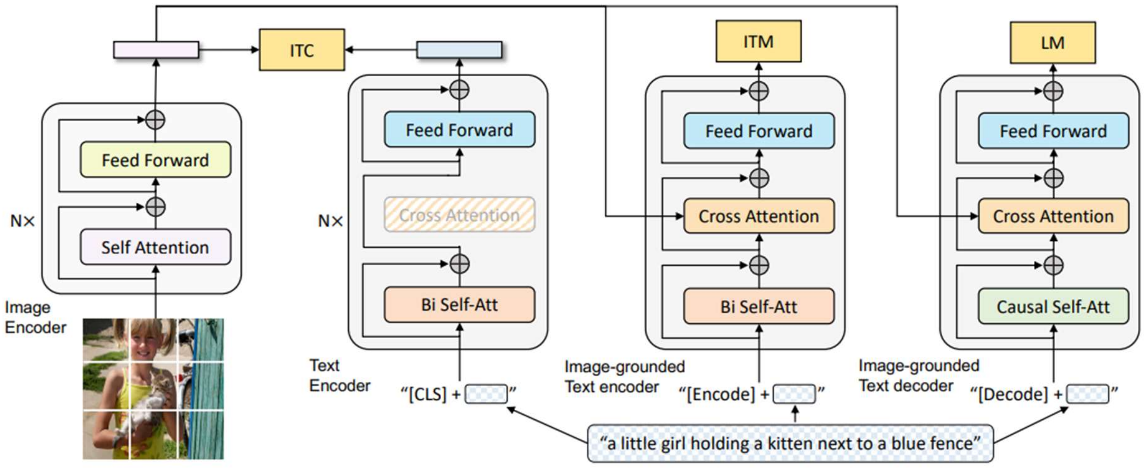


Figure 1Architeture of BLIP

The BLIP model has been trained in two stages:

Vision-language representation learning: In this stage, the Q-Former is trained to learn representations of images and text that are aligned with each other. The training objective is to maximize the mutual information between the image and text representations.

Vision-to-language generative learning: In this stage, the Q-Former is trained to generate text descriptions of images. The training objective is to maximize the likelihood of the correct text description given the input image.

To generate a caption for an image using BLIP, the model first extracts visual features from the image using the vision encoder. The visual features are then passed to the Q-Former, which generates a sequence of multimodal representations. The multimodal representations are then passed to the language decoder, which generates a caption word by word. BLIP is a powerful and versatile model that can be used for a variety of vision-language tasks. It achieves state-of-the-art results on many vision-language benchmarks, including COCO, VQA, and Flickr30K

4.2 Dataset description

DiffusionDB is a large-scale text-to-image prompt dataset that contains 14 million images generated by Stable Diffusion using prompts and hyperparameters specified by real users. The dataset is split into two parts: DiffusionDB 2M (2 million images) and DiffusionDB Large (14 million images). The images in DiffusionDB are diverse and cover a wide range of topics, including animals, objects, scenes, and events. DiffusionDB is a valuable resource for researchers and developers working on text-to-image generation, image captioning, and other vision-language tasks. The dataset can be used to train and evaluate models, and to develop new algorithms and techniques.

Property	**Description**
Name	DiffusionDB
Type	Text-to-image prompt dataset

Size	14 million images
Source	Real users
Format	PNG or WebP files
Annotation	Text prompt and hyperparameters
License	CC BY 4.0

4.3 Face Recognition using FaceNet Pytorch

In this project, face recognition is implemented using the MTCNN (Multi-task Cascaded Convolutional Networks) and FaceNet PyTorch [11], two powerful tools in the field of computer vision. The project's goal is to identify and verify the faces of five specific subjects from test images captured through a webcam. MTCNN, a robust face detection model, first locates and aligns faces within the images. Subsequently, FaceNet PyTorch, a deep neural network designed for face recognition, encodes the facial features into high-dimensional vectors, creating a unique face embedding for each subject. By comparing these embeddings, the project can accurately recognize and verify the identity of the subjects.

This combination of MTCNN and FaceNet PyTorch enables precise and efficient face recognition, offering a valuable solution for various applications, from security and access control to personalization and authentication. The face recognized is stored in the format of a string and passed on to the higher program such as to simultaneously describe the person detected.

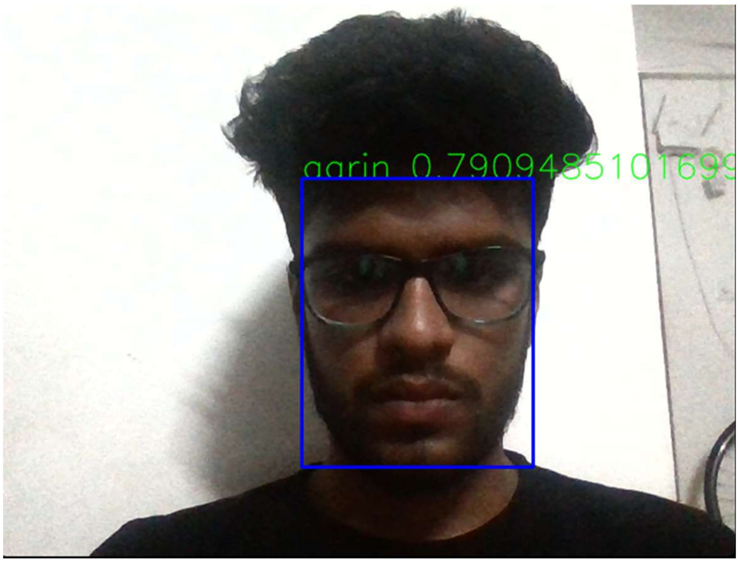


Figure 2 Face Recognition working

4.4 Implementation of BLIP for Real time Video Captioning with Other Features:

The program initiates the real-time video capture from a webcam, continuously pulling in video frames. These frames are the raw source of visual data for subsequent processing. Simultaneously, the program leverages a BLIP model to generate descriptive captions for each video frame. BLIP is a state-of-the-art vision-language model that excels at providing contextually accurate and informative captions for visual content, making it a valuable asset in understanding the video content.

The program enhances the captions by integrating face recognition and emotion recognition capabilities. It identifies and labels the individuals present in the video, enriching the context of the captions by associating them with specific people.

To ensure efficient and responsive processing, the program adopts multithreading. Multithreading enables the simultaneous execution of tasks, improving performance. Additionally, it tracks the time elapsed to ensure that the captions are synchronized with the video, maintaining a smooth user experience. The processed video frames, now enriched with captions and face recognition details, are displayed in real time. Viewers can observe the most recent captions and the identified individuals as the video unfolds, providing valuable insights into the visual content.

The program displays the processed frames with the most recent captions and face recognition details added to them. This allows the user to see the captions and face recognition details in real time.

To summarize, the program captures real-time video from a Webcam and concurrently generates captions using a BLIP model for the content in the frames. It combines these captions with face recognition information. To ensure smooth processing, the code employs multithreading and keeps track of elapsed time. Processed frames are displayed with the most recent captions and face recognition details added to them, the program enables the real-time analysis of video content, providing descriptive captions for each of the frames along with facial and emotional recognition insights, all displayed on the screen as the video is being captured.

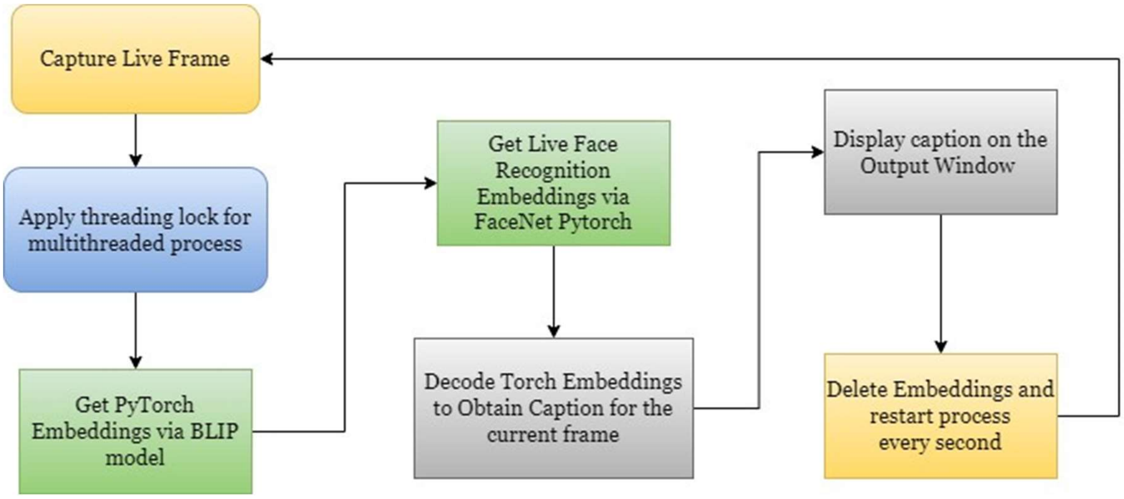

Capture Live Frame
Apply threading lock for multithreaded process
Get PyTorch Embeddings via BLIP model
Get Live Face Recognition Embeddings via FaceNet Pytorch
Decode Torch Embeddings to Obtain Caption for the current frame
Display caption on the Output Window
Delete Embeddings and restart process every second

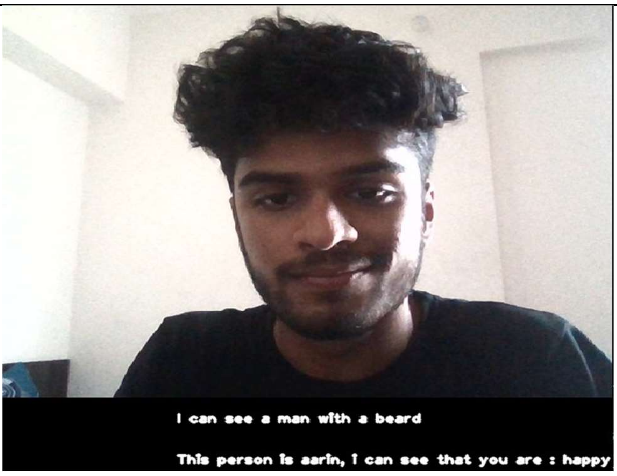
Figure 3 FlowChart For Methodology

5.Results

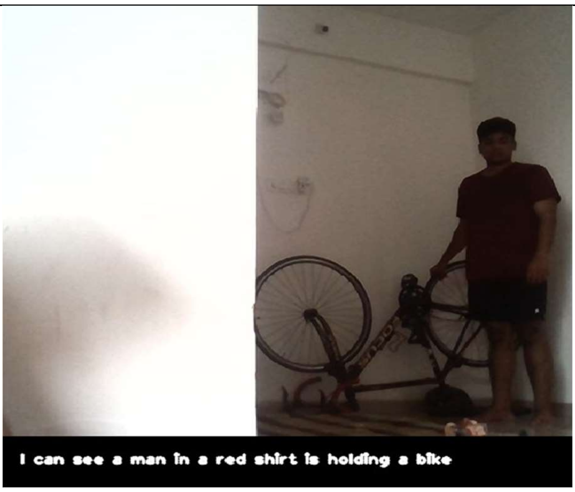
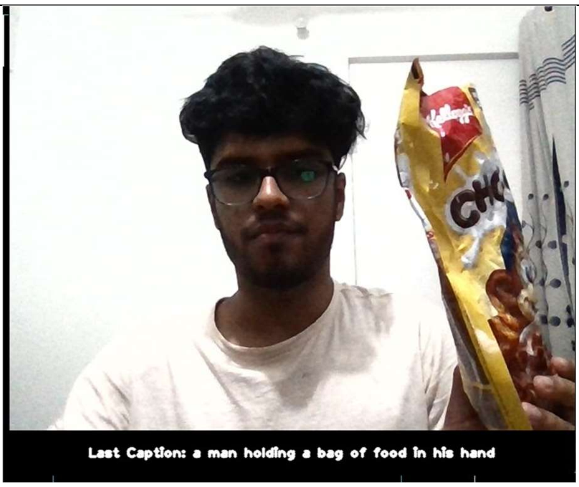
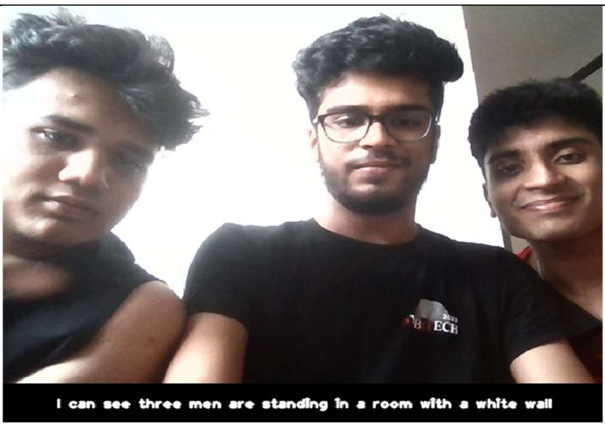
5.1 Obtaining Results

We tested our combined model in real-world scenarios using a webcam, and it consistently provided captions every second. These captions are quite detailed, capable of describing a variety of complex situations. Whether it was a simple scene or a more intricate one with many elements and actions, the model accurately identified and described what was happening in the video frames. This shows that the model is versatile and reliable, making it suitable for applications like video surveillance, accessibility tools, and content indexing, where understanding video content in real-time is crucial.

5.2 Result Table for Multiple scenarios

Scenario	Image with Caption
Single person with face recognition and emotion recognition	 I can see a man with a beard This person is aarin, I can see that you are : happy

Two men sitting in front of the camera	 I can see two men are sitting down and smiling
An Object placed in a room	 I can see a bike leaning against a wall in a room
An object Kept on the floor	 I can see a can of soda sitting on the floor

An object with a person standing beside it	 I can see a man in a red shirt is holding a bike
A person holding up a packet of food	 Last Caption: a man holding a bag of food in his hand
Multiple people in a room in front of the camera	 I can see three men are standing in a room with a white wall

6. Conclusion

In conclusion, our project has successfully introduced a solution for dense video captioning, harnessing the capabilities of the BLIP pretrained model while seamlessly integrating face recognition technology to operate in real-time. The importance of dense video captioning in enhancing accessibility and comprehension across various applications cannot be understated, from video indexing to content recommendation and assistive technology.

However, it is important to acknowledge the areas in which our project faced challenges. While the captions generated by our system were accurate, we observed inconsistencies in the generation time. Moreover, the absence of a link between the start and end frames of the video led to captions being generated anew for each event encountered, which could result in disruptions to the viewer's experience.

In future work, addressing these limitations and optimizing the system for a more consistent and seamless experience will be crucial. Nevertheless, our project represents a significant advancement in the field of dense video captioning, demonstrating the potential of integrating cutting-edge technologies to enhance the accessibility and utility of video content across a range of applications.

References

1. Venugopalan, S., Xu, H., & Jawahar, C. V. (2015). DenseCap: Fully convolutional neural network for dense image captioning. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4974-4982). IEEE.

2. Wang, Z., Zhou, L., Qiao, Z., Chen, H., & Liu, H. (2018). LSTNet: A deep learning model for long-term sequence learning and prediction. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 8938-8947). IEEE.

3. Shen, Z., Xu, H., Wang, W., Liu, S., Qiao, Y., & Wang, H. (2018). VSE++: Improved visual-semantic embedding for video captioning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 3795-3804). Association for Computational Linguistics.

4. Zhao, Y., Xiong, Y., Wang, L., Zhang, Z., Van Gool, L., & Qiao, Y. (2019). TCN-DVC: A weakly supervised approach to temporal convolutional network for dense video captioning. In Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 9626-9635). IEEE

5. Mun, J., Son, J., & Kim, J. (2019). Streamlined dense video captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 10642-10651).

6. Wang, J., Zhang, Y., Wang, L., Qiao, Y., & Chen, H. (2020). Dense-captioning with multi-scale temporal attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 12770-12779).

7. Xu, X., Sun, C., Wang, J., Qiao, Y., & Chen, H. (2020). Dense video captioning with progressive attention. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 2597-2607).

8. Liu, Z., Zhou, L., Zhang, Z., Qiao, Y., & Chen, H. (2021). Dense video captioning with transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 5720-5729).

9. Zhang, Y., Chen, W., Wang, J., Qiao, Y., & Chen, H. (2022). Dense video captioning with hierarchical transformer. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 2990-3000).

10. https://arxiv.org/abs/2201.12086 BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

11. https://arxiv.org/abs/1503.03832 FaceNet: A Unified Embedding for Face Recognition and Clustering

12. Zero-Shot Dense Video Captioning by Jointly Optimizing Text and Moment

13. Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning

14. A Better Use of Audio-Visual Cues: Dense Video Captioning with Bi-modal Transformer

15. Multi-Modal Dense Video Captioning