



CAPSTONE PROJECT PRESENTATION

- CHANDHU HRISHITESH P
- ANIK VAHORA

Navigation in low light vision

Problem Statement

Most vision-language models (VLMs) are trained on diverse, well-lit, general-purpose image-text datasets. However, they often fail to perform accurately on **domain-specific videos**, especially those recorded under challenging conditions like **nighttime campus environments**. This project investigates the limitations of existing VLMs on such data by extracting images from **nighttime video recordings** of a campus and evaluating their performance. The goal is to analyze model errors and work toward building or fine-tuning a **more robust, domain-specific VLM** that can better understanding at night images.

Importance

Improving night time video analysis enhances safety in autonomous navigation, security monitoring, people with disabilities and urban planning.

Project Objectives

- Extract relevant frames from video recordings
- Evaluate multiple vision-language models for performance
- Conduct detailed error analyses to identify pitfalls.
- Make domain specific accurate VLM.

Video Recording and Route Details

Recording Details

Departmental Coverage in Basic Science Block:

Each evening and night

Department wise $(4 \times 2) = 8$ videos

Floor wise through stairs and lift $(7 \times 2 \times 2) = 28$ videos

Captured multiple videos in campus from main gate to brahmagiri in 2 different routes

(Main gate -> Main building -> Library -> Lhc a -> CID -> Amul -> Brahmagiri.)

(Main gate -> Girl's co-op -> Basketball court -> Chemical department -> Amul -> Brahmagiri.)



Image Extraction Process from Video

```
import cv2
import os

cam = cv2.VideoCapture("C:\\Users\\user_name\\Downloads\\g to 1 morning lift.mp4")

try:
    if not os.path.exists('data2'):
        os.makedirs('data2')
except OSError:
    print('Error: Creating directory of data')

currentframe = 0

while True:
    ret, frame = cam.read()
    if ret:
        name = './data2/frame' + str(currentframe) + '.jpg'
        print('Creating...' + name)
        cv2.imwrite(name, frame)
        currentframe += 1
    else:
        break

cam.release()
cv2.destroyAllWindows()
```

1 Extraction Steps

1. Load video using OpenCV in Python
2. Extract frames at specific intervals to reduce redundancy
3. Save frames for subsequent VLM input

2 Sample Extracted Frames

Frames depict urban scenes at different points in the route, capturing diverse lighting and object configurations for model evaluation.

WHAT IS VISION – LANGUAGE MODEL (VLM)

A Vision-Language Model (VLM) is a type of artificial intelligence that can **understand both images and text**, and more importantly, **link them together**. It learns to connect visual content (like photos or video frames) with language (like descriptions, captions, or questions), enabling it to perform tasks that require understanding across both domains.

How It Works

- **Image Encoding:**
The model first processes the image using a vision backbone (like a CNN or Vision Transformer) to convert it into a set of meaningful numerical features.
- **Text Encoding:**
Simultaneously, a language model (like BERT or a transformer-based encoder) processes the input text and turns it into another set of features.
- **Alignment/Training:**
The model is trained to align these two sets of features so that the image and its correct description are closer together than incorrect ones.
- **Prediction Tasks:**
Once trained, the VLM can generate captions, answer questions about images, retrieve matching text for a given image (or vice versa).

Evolution of Vision Tasks (2012–2022)

1 Image Classification (2012–)

Task: Assign a single label to an entire image.

Breakthrough: AlexNet (2012) on ImageNet.

Later models: VGG, ResNet, EfficientNet.

2 Object Detection (2014–2018)

Task: Detect and localize multiple objects with bounding boxes.

Key models: R-CNN, Fast/Faster R-CNN, YOLO, SSD.

3 Image Captioning (2018–2022)

Task: Generate natural language descriptions of images.

Early methods: CNN + RNN models (Show & Tell).

Advancements: Attention mechanisms (Show-Attend-Tell), Transformer-based models.

Advanced Vision-Language Models (2023–)

Goal: Joint understanding of images and text in a unified framework.

Capabilities:

- Visual Question Answering (VQA)
- Multimodal reasoning and dialogue
- Captioning, retrieval, and grounding

Notable Models:

- Flamingo (DeepMind), BLIP-2, GPT-4V, Gemini, LLaVA

Key Innovation:

- Few-shot and zero-shot generalization
- Unified multimodal input/output (e.g., “describe this image and answer questions”)

Exploring Vision-Language Models Implemented

Cosmos nematron

A model that understands both images and text. It's great at recognizing what's in a picture without needing extra training, and can match images with the right descriptions.

Perplexity

This model can look at a picture and describe it or answer questions about it. It learns from a mix of images and text, and then gets better with practice on specific tasks.

Hugging face smolVLM

A smaller, faster model that connects images and text well. It's built to work on devices with less computing power, while still doing a good job at understanding both visuals and words.

Few images with errors



SAMPLE ERROR - 1



This image was taken near basketball court in night conditions.

As we can see the text is hard to identify and the 3 vlms gave wrong outputs for this.

Output

Preview

JSON

The image captures a scene at night, dominated by a large blue sign with white text and arrows. The sign, which is the main focus of the image, is mounted on a pole and is illuminated by a nearby street lamp. It provides directions to various locations on the campus, including the swimming pool, sports complex, 1000V substation, chemical engineering department, and student activity center.

In the background, there are people engaged in different activities. Some are walking, while others are playing basketball on a court. The court is equipped with a hoop and a backboard, and it's surrounded by a fence. The area is well-lit, with several street lamps casting a warm glow on the scene.

The image also includes a road with a car driving on it, adding to the dynamic nature of the scene. The car's headlights are visible, cutting through the darkness. The overall atmosphere is lively and bustling, indicative of a typical night on a busy campus.

COSMOS

NEMATRON

Inaccurate text description of the sign board even in well lit conditions. It's assuming people's shadows in the background as cars and gave a detailed description of the car which is not present.

Directions (left/right) on the sign board are not captured by COSMOS.



The image shows a blue directional signboard at night, likely on a university or institutional campus. The sign lists several locations along with arrows indicating their directions:

- Swimming Pool (left arrow)
- Sports Complex (left arrow)
- 33 KV Substation (right arrow)
- Chemical Engineering Dept (right arrow)
- Student Activity Centre (right arrow)

The text on the sign is somewhat faded and difficult to read due to poor lighting and glare. In the background, there is a basketball court with people playing, and further back, a pathway with more people walking under streetlights. The surroundings include trees and open space, giving the scene an outdoor, campus-like atmosphere.


PERPLEXITY

This VLM gave a much better and accurate description of the image compared to cosmos nematron

SmoVLM-256M: The Smollest VLM ever 🌟

Play with [HuggingFaceTB/SmoVLM-Instruct-250M](#) in this demo. To get started, upload an image and text or try one of the examples. This demo doesn't use history for the chat, so every chat you start is a new conversation.

Chatbot



describe the image

A sign which reads 'chemical engineering' and 'student activity centre' points to the KV substation.

SMOLVLM

This VLM gave a single line output which is inconsistent sometimes and the output just says that it point to the KV substation which is completely wrong.

SAMPLE ERROR - 2



This image was taken near tennis court in night conditions.

As we can see the text is hard to identify and the 3 vlms gave different outputs for this.

Output

Preview

JSON

The image captures a serene night scene at a tennis court. The court, enclosed by a blue fence, is bathed in the soft glow of several lights that are strategically placed around it. The lights cast long shadows and create a play of light and dark on the court, adding a sense of depth to the scene.

In the foreground, a sign stands prominently. It's a circular sign with a black background and white text. The sign reads "Built by NITC 1972 Graduation Year 2022 6th Gifted to NITC", indicating the history and significance of the court.

The court is surrounded by trees, their leaves rustling gently in the night breeze. The trees form a natural boundary around the court, their dark silhouettes contrasting against the brightly lit court.

Overall, the image paints a picture of a well-maintained and cherished tennis court, a testament to the dedication and hard work of its builders and users.

COSMOS

NEMATRON

Inaccurate text description of the sign board even in well lit conditions. It says Graduation year 2022 while it's 2023 in the sign board and it reads NITK as NITC.



The image shows a well-lit outdoor tennis court at night, surrounded by a high blue fence. The court is illuminated by several bright floodlights, creating a clear and inviting atmosphere for night play. In the background, trees and foliage frame the scene, giving it a serene campus feel.

In the foreground, there is a prominent round sign mounted on a sturdy pillar. The sign bears the emblem of NITK (National Institute of Technology Karnataka) and reads:

“BUILT BY KREC 1973 BATCH IN GOLDEN GRADUATION YEAR 2023 & GIFTED TO N.I.T.K.”

This indicates that the tennis court was constructed as a commemorative gift by the 1973 batch of KREC (Karnataka Regional Engineering College, now NITK) on their golden jubilee graduation year in 2023.

A player dressed in sports attire can be seen on the court, suggesting the facility is in active use. The overall ambiance is peaceful, with the combination of artificial lighting and natural surroundings highlighting the court as a valued recreational spot on the campus.

PERPLEXITY

This VLM gave a much better and very accurate and gives mostly accurate results.



elaborate the image

A rectangular sign on a signboard next to a tennis court.



SMOLVLM

This VLM gave a single line output which doesn't even explain the scenario even when asked to elaborate.

Output Preview JSON

The image captures a scene from a dimly lit hallway. The hallway is characterized by a white wall on the left and a black wall on the right, creating a stark contrast. A sign, written in Hindi, hangs on the black wall, reading "अग्नि देवी मंदिर". The sign is positioned above a door, which is slightly ajar, revealing a hint of the room beyond. The floor of the hallway is adorned with a blue carpet, adding a touch of color to the otherwise monochrome setting. The overall atmosphere of the image is quiet and serene, with the open door inviting curiosity about the room it leads to.

The image captures a scene from a hospital, specifically the 5th floor. The viewer's perspective is from the entrance of a hallway, which is characterized by its white walls and a gray floor. The hallway is lined with doors on both sides, suggesting the presence of patient rooms or medical facilities.

On the right side of the hallway, there's a bulletin board, a common feature in hospitals for displaying important notices or information. The ceiling of the hallway is equipped with fluorescent lights, providing illumination to the space.

The most striking feature of the image is the red sign with white text that reads "5th Floor". This sign is prominently displayed on the ceiling, serving as a clear indicator of the floor level. The sign's bold color and placement ensure it's easily visible to anyone navigating the hallway.

Overall, the image provides a glimpse into the interior of a hospital, highlighting the functional design elements and signage that aid in navigation and communication within the facility.

Other common errors

There are many such errors which are completely different from the input image where it assumes the college is IIT Roorkee due to the training data set it was trained with or sometimes it assumes that it's a hospital. Many such common errors are included in excel sheet for easier comparison between all 3 VLMS and to finally compare with the VLM implemented by US.

EXCEL SHEET LINK :

<https://docs.google.com/spreadsheets/d/18mANyTd8zSLV9sUiwqCxnkARNc7k0S8ZYSnUr8xJp6o/edit?gid=214575619#gid=214575619>

https://docs.google.com/spreadsheets/d/1AGuA0WvwFgm uMmD9EffZcR6aj8z10yoPg3lb_NZ6vMI/edit?gid=1975964765#gid=1975964765

Error Analysis of the Models

1

Observed Issues

Misclassification of dark or occluded objects; difficulty in interpreting ambiguous shadows as distinct elements.

2

Causes

Limited contextual awareness under low-lit conditions and bias from training predominantly on daytime imagery.

3

Implications

Such errors reduce confidence in automated scene interpretation for safety-critical applications.

Next Steps and Project Timeline

1 Upcoming Actions

1. Implementation and fine tuning VLMs using night time image datasets, based on basic VLMs..
2. Evaluate additional VLM architectures
3. Train the model with the current data set of all the campus night images and videos.
4. Explore data augmentation for low-light conditions

2 Project Timeline

Current work include data collection completion, initial VLM evaluation, error analysis, and planned enhancements expected over the next 3 months.

THANK

YOU