Pawan Perara

Behavioral Image Captioning:

Models and Tools Used

1. **BLIP (Bootstrapping Language-Image Pre-training)**: After evaluating multiple options, I selected the "Salesforce/blip-image-captioning-large" model from Hugging Face for generating initial image captions. This large variant of BLIP offers more detailed and accurate descriptions compared to the base model, with enhanced ability to capture nuanced visual elements.
2. **GPT-3.5 Turbo**: I utilized OpenAI's GPT-3.5 model to transform these detailed captions into comprehensive behavioral descriptions by inferring intentions, emotions, and social contexts.
3. **Streamlit**: For creating a simple, interactive user interface that allows image uploads and displays results.

Behavioral Inference Approach

I implemented a two-stage pipeline for behavioral analysis after comparing different methodologies:

1. **Stage 1 - Enhanced Caption Generation**: The BLIP-large model processes the input image and generates a detailed description of what's visible. I selected this larger model specifically for its improved ability to capture subtle visual cues that might indicate behavioral patterns.
2. **Stage 2 - Behavioral Analysis**: The enhanced caption is sent to GPT-3.5 with specific prompting to analyze and infer behavioral aspects such as:
    - Likely intentions of people in the scene
    - Emotional states suggested by posture or context
    - Social dynamics between subjects
    - Potential motivations for the observed activity

This approach leverages the strengths of both models - BLIP-large's superior visual understanding capabilities and GPT-3.5's contextual reasoning abilities.

Prompt Engineering

I carefully crafted the prompt for GPT-3.5 to guide it toward generating meaningful behavioral insights:

System prompt: "You are an expert at analyzing human behavior from descriptions. Provide a detailed behavioral analysis that goes beyond basic description. Infer intentions, emotions, and social context."

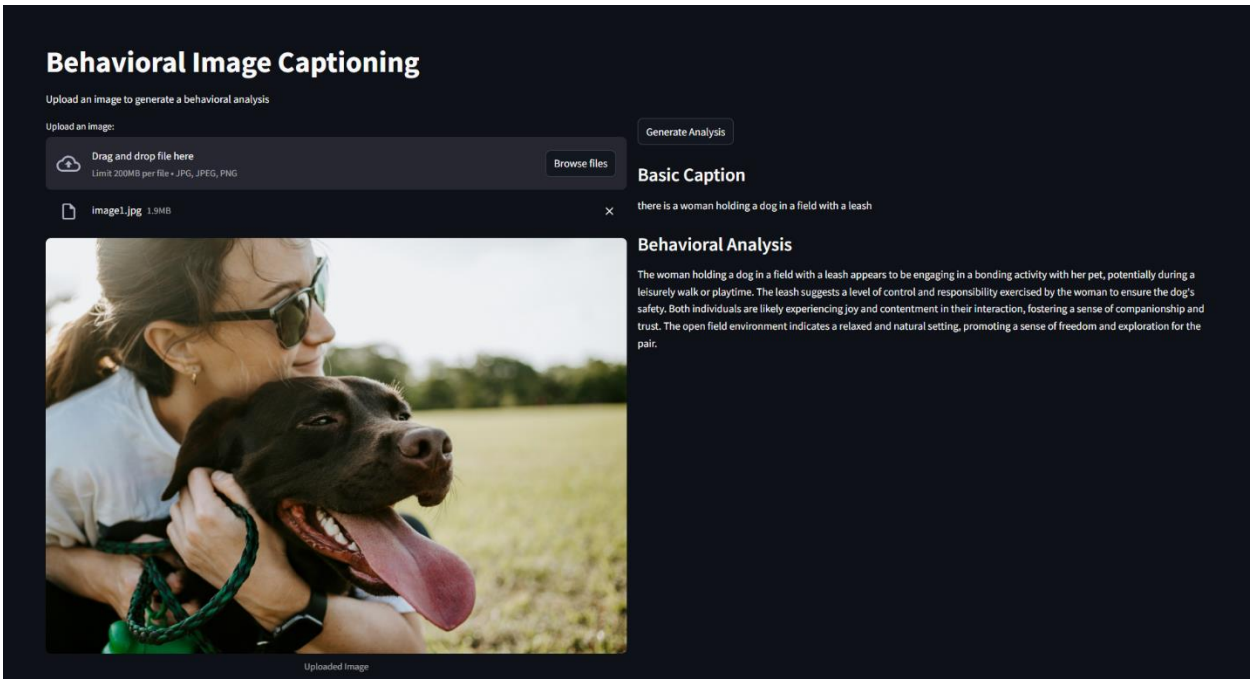
User prompt: "Based on this image description: '[BLIP caption]', provide a behavioral analysis of what might be happening. Only one paragraph with a maximum of 100 words. Stick strictly to what can be reasonably inferred without inventing details."

This prompt engineering encourages GPT-3.5 to:

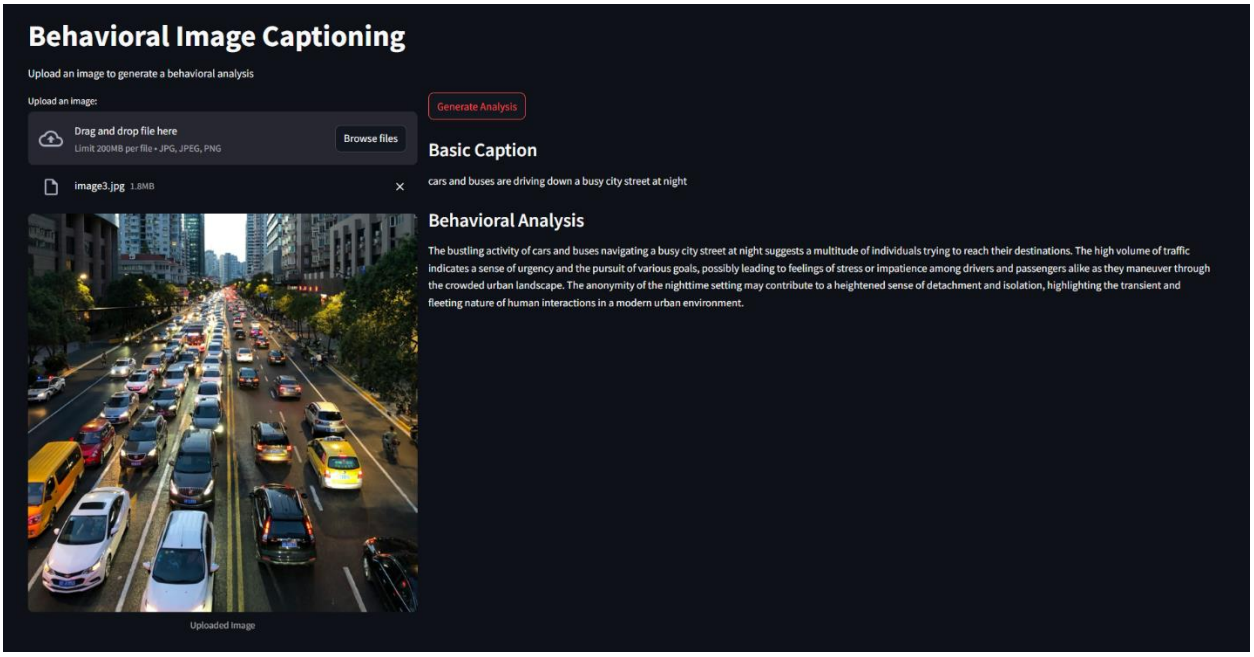
- Focus on behavioral aspects rather than just repeating the visual description.
- Make reasonable inferences about human intentions and emotions.
- Consider social and contextual factors that might explain the observed scene.
- Generate insights that go beyond what's explicitly visible in the image.
- Strictly limit response length to prevent over-generation of potentially false details.
- Include explicit instructions to avoid inventing information not supported by the image description.
- Enforce a single paragraph format for conciseness and clarity.

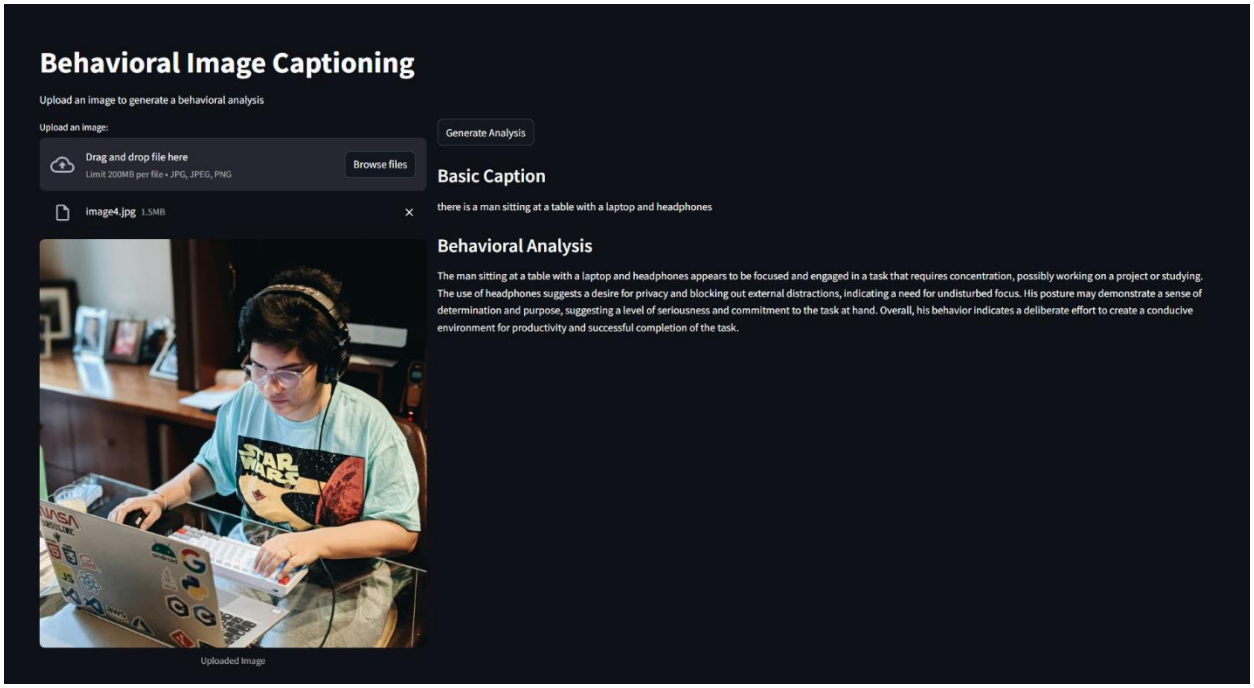
GitHub Link For project files: https://github.com/PinsaraPerera/AI-assessment.git

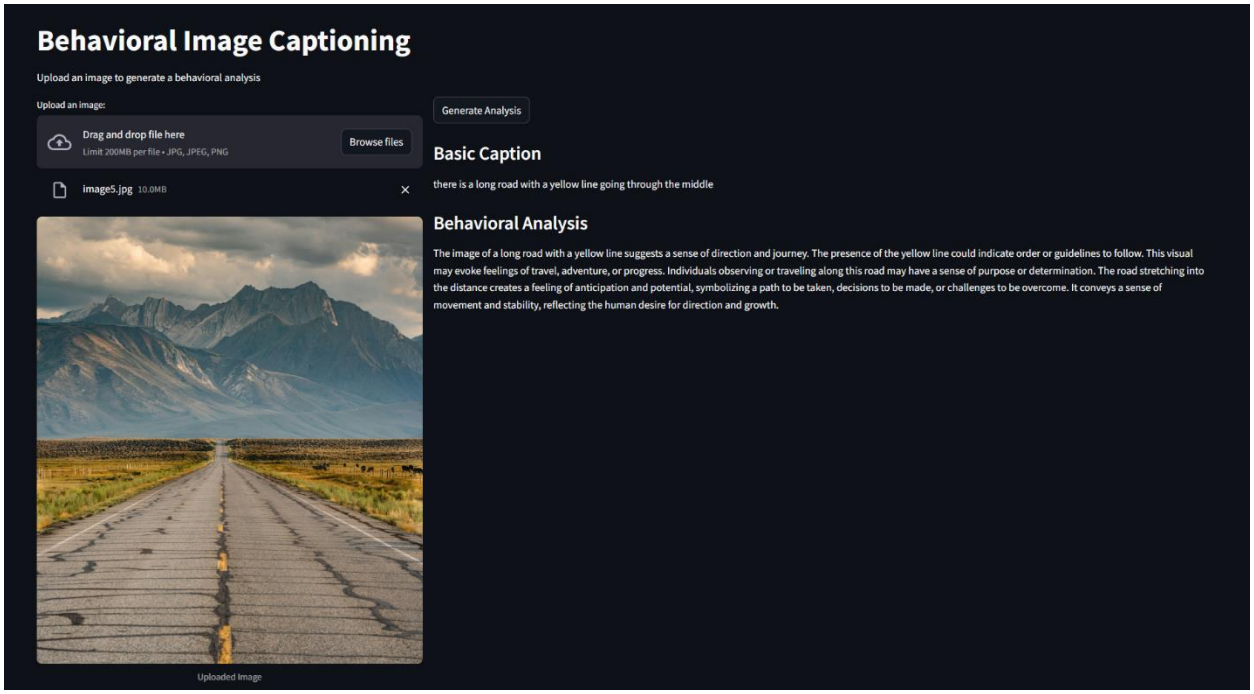
Example Outputs




Behavioral Image Captioning
Upload an image to generate a behavioral analysis
Upload an image:
Drag and drop file here
Limit 200MB per file • JPG, JPEG, PNG
Browse files
image2.jpg 2.0MB
Uploaded Image
Generate Analysis
Basic Caption
there are many children sitting at a table doing homework
Behavioral Analysis
The children are likely engrossed in their homework, exhibiting focused and diligent behavior as they navigate their tasks. Some may appear more engaged and enthusiastic, showcasing a strong sense of determination and eagerness to excel, while others might display signs of frustration or boredom, potentially seeking assistance or distraction. This setting suggests a structured and educational environment, providing an opportunity for social interaction, collaboration, and support among peers, fostering a sense of camaraderie and shared responsibility in achieving academic goals.


Behavioral Image Captioning
Upload an image to generate a behavioral analysis
Upload an image:
Drag and drop file here
Limit 200MB per file • JPG, JPEG, PNG
Browse files
image3.jpg 1.8MB
Uploaded Image
Generate Analysis
Basic Caption
cars and buses are driving down a busy city street at night
Behavioral Analysis
The bustling activity of cars and buses navigating a busy city street at night suggests a multitude of individuals trying to reach their destinations. The high volume of traffic indicates a sense of urgency and the pursuit of various goals, possibly leading to feelings of stress or impatience among drivers and passengers alike as they maneuver through the crowded urban landscape. The anonymity of the nighttime setting may contribute to a heightened sense of detachment and isolation, highlighting the transient and fleeting nature of human interactions in a modern urban environment.


Behavioral Image Captioning
Upload an image to generate a behavioral analysis
Upload an image:
Drag and drop file here
Limit 200MB per file • JPG, JPEG, PNG
Browse files
image4.jpg 1.5MB
Uploaded Image
Generate Analysis
Basic Caption
there is a man sitting at a table with a laptop and headphones
Behavioral Analysis
The man sitting at a table with a laptop and headphones appears to be focused and engaged in a task that requires concentration, possibly working on a project or studying. The use of headphones suggests a desire for privacy and blocking out external distractions, indicating a need for undisturbed focus. His posture may demonstrate a sense of determination and purpose, suggesting a level of seriousness and commitment to the task at hand. Overall, his behavior indicates a deliberate effort to create a conducive environment for productivity and successful completion of the task.


Behavioral Image Captioning
Upload an image to generate a behavioral analysis
Upload an image:
Drag and drop file here
Limit 200MB per file • JPG, JPEG, PNG
Browse files
image5.jpg 10.0MB
Uploaded Image
Generate Analysis
Basic Caption
there is a long road with a yellow line going through the middle
Behavioral Analysis
The image of a long road with a yellow line suggests a sense of direction and journey. The presence of the yellow line could indicate order or guidelines to follow. This visual may evoke feelings of travel, adventure, or progress. Individuals observing or traveling along this road may have a sense of purpose or determination. The road stretching into the distance creates a feeling of anticipation and potential, symbolizing a path to be taken, decisions to be made, or challenges to be overcome. It conveys a sense of movement and stability, reflecting the human desire for direction and growth.