**Project Report: Human vs Animal Detection using CLIP**

**Approach**

* We use CLIP (Contrastive Language–Image Pretraining) from HuggingFace to classify image content by comparing patches against natural language prompts (e.g., *"a photo of a lion"*, *"a photo of a human"*, etc.).
* Each input image or video frame is divided into a 3×3 grid of patches to localize objects.
* Each patch is independently classified using CLIP, and patches with confidence scores above 0.60 are accepted as valid detections.
* If patch-level predictions are ambiguous or below a confidence threshold (0.70):
  + We fallback to full-image classification to decide the dominant class.
  + This helps in cases where small patches do not contain enough visual context.

**Prompts Used**

We use prompt engineering to guide CLIP with 20+ class prompts:

* **Human-related**:  
  *"a photo of a human"*
* **Animal-related**:  
  *cow*, *goat*, *lion*, *monkey*, *zebra*, *dog*, *cat*, *elephant*, etc.

**Detection Output**

* **Each detected object is drawn with a bounding box and a label overlay.**
* **Results are saved to the detect/<run\_id>/ folder:**
  + **Annotated image: \*\_vis.jpg**
  + **Cropped detections: <uuid>\_<patch\_id>\_<label>.jpg**
  + **Detection metadata: JSON summary**
* **Terminal prints real-time alerts:**

For example -

Alert: Human detected - Man at (x1,y1,x2,y2) [Confidence: 0.83]

**Evaluation**

* The evaluate\_dataset() function runs a full dataset evaluation.
* It reports:
  + **Image-level accuracy** (majority class match per image)
  + **Patch-level accuracy** (correctly labeled patches)
  + **Confusion matrix** for class-wise breakdown
* Reports saved to:

classify/<run\_id>/evaluation\_report.json

**Challenges & Solutions:**

| **Challenge** | **Solution** |
| --- | --- |
| CLIP is not a detector (no bounding boxes) | We simulate detection via **patch-based classification** |
| Small patches may lack context | If no confident result, we **fallback to full-image classification** |
| Similar-looking animals cause confusion | Expanded prompts (e.g., “a photo of a goat” vs “a photo of a sheep”) help |
| Human diversity (man, woman, child) causes drift | Included **multiple human prompts** for robustness |

**Results:**

The evaluation on the test dataset showed**:**

* Image-level accuracy of approximately 85 percent
* Patch-level accuracy of approximately 87 percent
* Some misclassifications were observed between visually similar animals (e.g., lion and dog)

**Observations:**

* Human detection sometimes underperforms in small patches due to lack of contextual features like full body or facial structure
* Boosting human confidence when it exceeds a certain threshold improves overall recognition
* Animals with distinct color patterns like zebra and tiger perform very well.