Sensmore Case Study

Aman Sidhu

Dataset Preparation and VQA Pair Generation

Data Collection

- Collected 12,141 frames from the provided video clips, sampled down to 745 for development purposes/resource constraints
- Using ffmpeg, sampled frames at 1 FPS from the video clips

VQA Generation Process

- Used GPT-4o to automate VQA data generation from detected objects and image context
- Generated three responses per frame using standardized operational prompts such as:
    - "Where should I go next?"
    - "Describe what you see in front of you."
    - "Go to the nearest pile and fill the bucket!"
- Final dataset: 1,500+ VQA pairs

Limitations

- Automated annotation pipeline lacks manual verification, creating a risk of hallucinated responses that don't match actual scene content
- The generic prompt set may not cover the full range of real-world operator queries, and would benefit from domain expert input on common operational commands
- Training data sourced exclusively from YouTube footage lacks diversity in:
    - Weather conditions (primarily clear, daylight conditions)
    - Lighting variations (dawn, dusk, shadows, artificial lighting)
    - Camera resolution quality
    - Geographic/site diversity (limited to specific job sites featured in videos)

Object Detection and Position Extraction

Vision Model Selection

- Used YOLO-World v2 (yolov8x-worldv2) for zero-shot, open-vocabulary object detection
- Enables detection without training on custom labeled dataset

Class Vocabulary

Sensmore Case Study

Aman Sidhu

- Selected detection classes: *['pile', 'heap', 'mound', 'dirt', 'gravel', 'sand', 'soil']*
- Experimented with more descriptive phrases ("pile of gravel", "dirt mound"), but achieved poorer detection rates
- Found "mound" performed better than "pile" in many scenarios, requiring class name experimentation

Position Encoding

- Detection outputs bounding box coordinates normalized to [0,1] range
- Provides semantic directional labels relative to the camera center:
    - *far-left, far-center, far-right*
    - *mid-left, mid-center, mid-right*
    - *near-left, near-center, near-right*
- Also outputs quantitative position (cx, cy) and relative size (area)

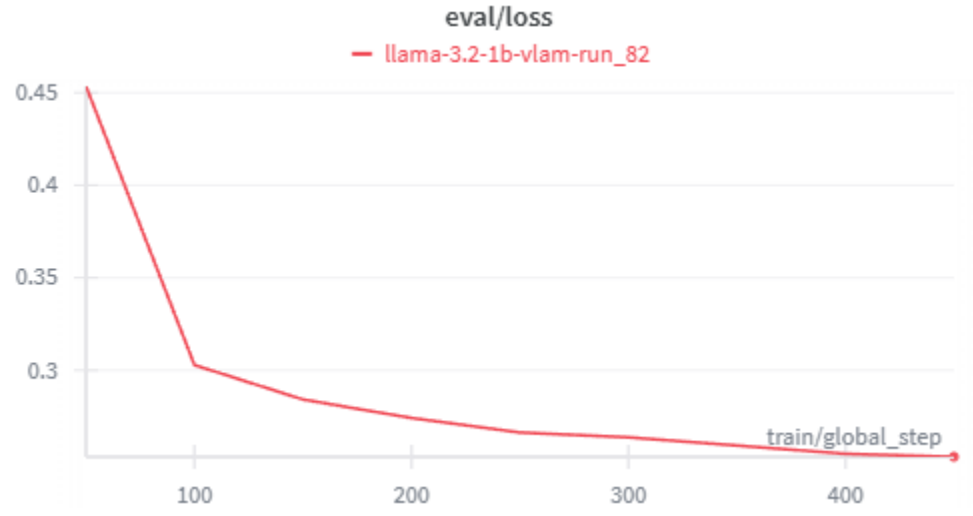
Limitations

- The open-vocabulary approach provides convenience but suffers from:
    - **False negatives**: Frequently misses multiple piles in the same frame
    - **Low confidence**: Many valid detections have confidence < 0.15
    - Suggests the need for a specialized model fine-tuned on construction site imagery
- Class name engineering required, detection performance sensitive to exact wording
- Vision-only approach limitations:
    - No depth perception or absolute distance estimation
    - Cannot determine pile volume or material composition
    - No real-time video stream processing (operates on individual frames)
    - Should be supplemented with LiDAR for accurate 3D position and volume data

Sensmore Case Study

Aman Sidhu

Integration and Demonstration

- Fined-tuned a unsloth/Llama-3.2-1B-Instruct model for this task
- A complete breakdown of the evaluation and training can be found here at this WandB report:
    - https://api.wandb.ai/links/aman-sidhu-mcgill-university/cj5yagy5
    - 

- To make training more computationally efficient, I fine-tuned the LLM using LoRA adapters
    - Was able to train the LLM on a local Nvidia RTX 2070 GPU with 8 GB of VRAM

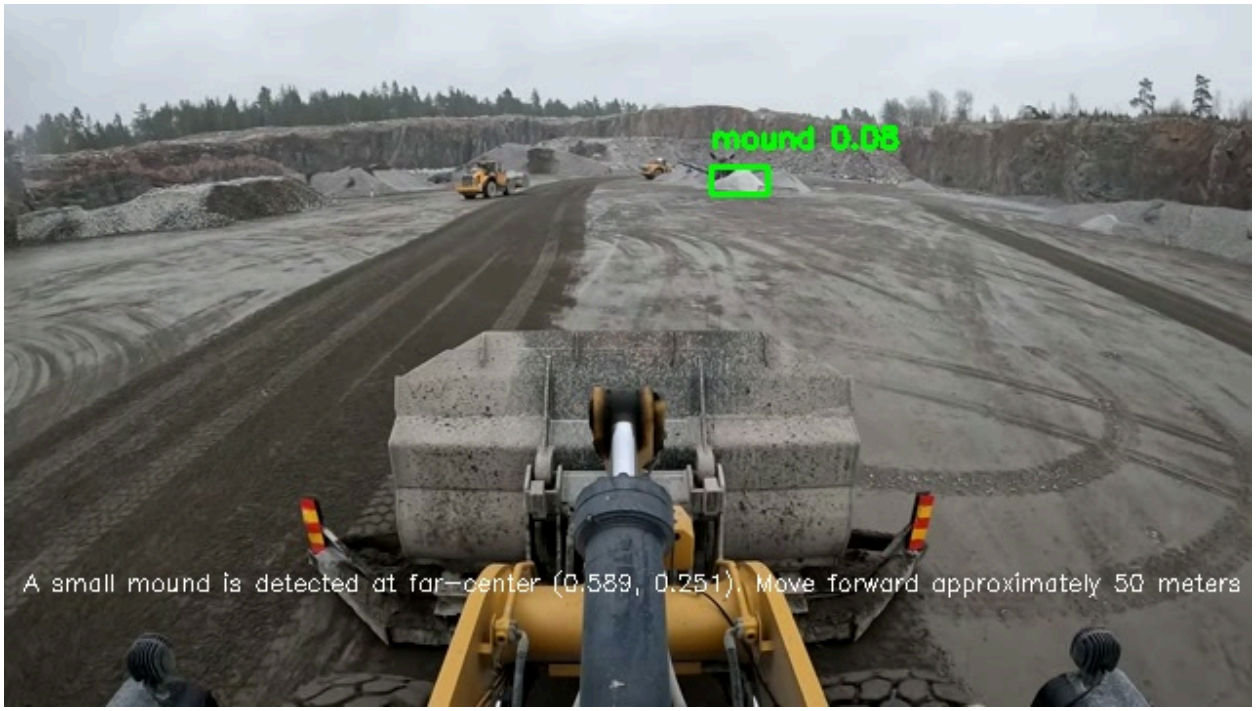
Sensmore Case Study

Aman Sidhu

Inference Pipeline

1. YOLO-World detects objects in the frame
2. Detections formatted as structured text context
3. LLaMA generates a suitable, natural language response

Example Output



[SCENE] Construction site view from wheel loader.

[DETECTIONS]

1. mound (conf=0.08) at far-center (0.589, 0.251) size=0.002

VLAM Response:

A small mound is detected at far-center (0.589, 0.251). Move forward approximately 50 meters to reach it.

Sensmore Case Study

Aman Sidhu

Potential Real-World Applications

- This proof-of-concept design does show a promising application for future embodied AI systems with a more rigorous training and evaluation pipeline
- Future applications should consider the types of actions the agent would be responsible for such as scooping and dumping piles, or being able to lay pile content down flat as to make a bedding as in https://www.youtube.com/watch?v=wc72kf9DWaY.

Further improvements and integration into existing robotics systems

---

- Depending on the construction site environment, if the agent is expected to work alongside human workers, there needs to be a level of collision avoidance and detection in these scenarios
- Fine-tune the vision model detector on labelled construction site images to improve accuracy and reduce false negatives
    - Also consider camera placement since claw apparatus takes up a lot of the view
- Increase modalities to include GPS and IMU for localization, and LiDAR for accurate distance measurements
- Optimize for edge deployment. Currently, 1B parameter models require running on accelerated hardware