Final Project Proposal
Emerging Trends in Explainable AI (XAI)

Title: Explainable Vision-Language Models for Safety Monitoring in Manufacturing Videos

1. Project Objective

To build an explainable AI system that analyzes videos from manufacturing environments to detect the presence or absence of essential safety equipment (e.g., helmets, gloves, vests), using Large Vision-Language Models (VLMs). The system will not only predict compliance but also generate interpretable visual and textual explanations that help humans understand why the model made its decision, thus enabling trust, transparency, and accountability in automated safety monitoring.

2. Why This Topic is Important

Workplace safety is a global concern, especially in manufacturing environments where personal protective equipment (PPE) can prevent serious injuries. Manual monitoring is time-consuming, inconsistent, and error-prone. Automating this task with explainable AI can save lives and improve trust in AI systems. This project brings together socially impactful technology, real-world application, and AI transparency—making it important both to me personally and to the wider world.

3. What Has Been Done Before

Conventional deep learning models like YOLO and Faster R-CNN have been used for PPE detection. However, these models are typically black boxes and lack interpretability. Recent advances in Vision-Language Models (e.g., BLIP-2, LLaVA, MiniGPT-4) allow for multimodal understanding and natural language generation. While these models can describe images, their explainability in safety-critical tasks remains underexplored. This project addresses that gap.

4. General Approach

1. Extract frames from manufacturing safety videos using OpenCV.
2. Feed each frame into a Vision-Language Model like BLIP-2 or LLaVA.
3. Use structured prompts (e.g., 'Is the worker wearing a helmet?') to get predictions.
4. Extract textual justifications and generate attention/Grad-CAM visualizations for XAI.
5. Present results in a user-friendly interface (e.g., Streamlit or Jupyter Notebook).

5. Data Sources

- Roboflow PPE Dataset
- Kaggle Safety Helmet Wearing Dataset
- Custom or synthetic data extracted from publicly available videos

6. What is Unique About This Project

This project uniquely combines the interpretability of XAI with the multimodal capabilities of VLMs to solve a real-world problem. While many models can detect safety violations, this system explains its reasoning visually and textually, helping human reviewers trust and verify model outputs in high-risk environments.

7. Final Artifact

The final deliverable will be a web-based application or notebook interface where users can upload a video, receive frame-by-frame safety predictions, and view both textual justifications and visual explanations (e.g., heatmaps, attention maps).

8. Project Timeline

Week 1 (Apr 1): Select VLM, collect videos, set up VLM-based frame analysis with prompts
Week 2 (Apr 7): Add XAI features (textual explanations), evaluate interpretability
Week 3 (Apr 15): Build a user-facing artifact (web app or notebook),