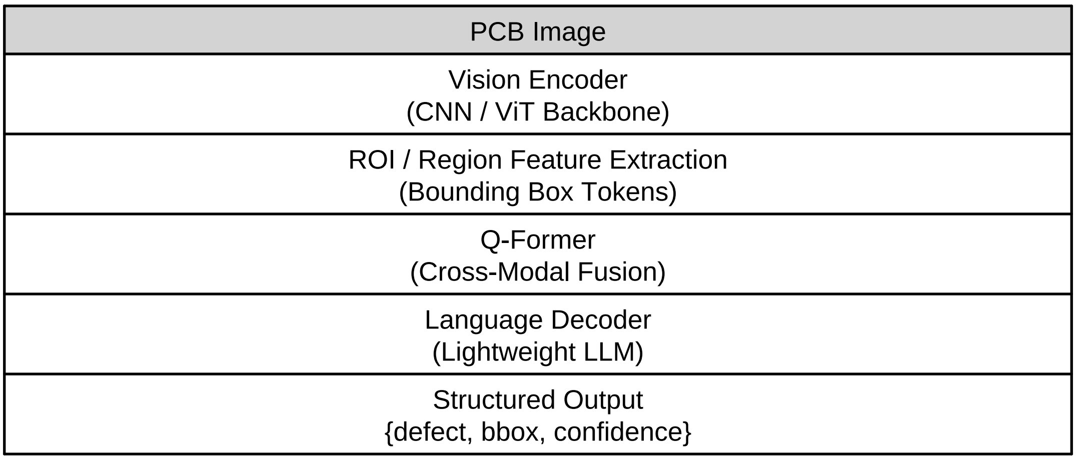
Assignment 3: Custom Vision–Language Model (VLM) Design for Industrial Quality Inspection

Candidate: Bashwaraj Sonkawade

System Architecture Overview



PCB Image

Vision Encoder
(CNN / ViT Backbone)

ROI / Region Feature Extraction
(Bounding Box Tokens)

Q-Former
(Cross-Modal Fusion)

Language Decoder
(Lightweight LLM)

Structured Output
{defect, bbox, confidence}

The architecture follows a modular BLIP-2-style design where the vision encoder extracts PCB-specific visual features. Region-level embeddings are created using bounding box information, which are fused with language tokens via a Q-Former module. A lightweight language decoder produces structured, grounded responses suitable for industrial inspection.

Key Design Principles

- Region-grounded visual representations
- Structured language generation (no free-form hallucination)
- Lightweight and offline deployable
- Sub-2 second inference latency