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

1. Problem Statement and Context

In semiconductor manufacturing, printed circuit board (PCB) inspection is a critical quality control step. Manual inspection is time-consuming and error-prone, while generic AI models often hallucinate defects that are not present. The objective of this task is to design a custom Vision–Language Model (VLM) that operates offline and assists inspectors by answering natural language questions about PCB defects.

The system must analyze a PCB image and return a structured response containing the defect type, precise bounding box coordinates, and a confidence score, while ensuring inference latency below 2 seconds.The assumed dataset consists of approximately 50,000 PCB images with defect bounding box annotations, but no human-labeled question–answer pairs. Therefore, the solution focuses on system design, training strategy, and reliability rather than full-scale implementation.

2. Model Selection

Based on the reference to **Transformers | Liquid Docs**, the selected model is a **compact transformer-based VLM** that:

- Performs well for **offline inference**
- Is **easy to fine-tune** on domain-specific data (PCB images)
- Has shown ability to **match or outperform larger models on some benchmarks**
- Trades off slightly higher inference time for better controllability and grounding

“The selected model follows the Liquid Foundation Model (LFM) transformer architecture proposed by Liquid AI, which emphasizes efficient offline inference, adaptive computation, and controllable behavior. Unlike conventional large-scale VLMs, Liquid’s transformer design prioritizes structured reasoning and grounding, making it suitable for safety-critical industrial inspection tasks.”

Comparison:

Criteria	Generic VLMs	Liquid Transformer
Offline inference	❌	☑
Fine-tuning ease	□□	☑
Hallucination risk	High	Low
Industrial control	Low	High

A lightweight transformer-based Vision–Language Model is selected for this task. Transformer architectures are well-suited for multimodal learning because they can jointly model visual and textual information through attention mechanisms. Compared to very large foundation models, a compact transformer-based VLM offers better control, lower computational cost, and improved reliability for industrial deployment.

The chosen model supports offline inference, parameter-efficient fine-tuning (LoRA or QLoRA), and quantization. These characteristics make it suitable for edge or on-premise environments commonly used in manufacturing facilities. Large generic VLMs are avoided because they prioritize fluent language generation over grounded, spatially accurate predictions, leading to hallucinations.

3. Architecture Design Strategy

The proposed architecture follows a modular vision–language design. A vision encoder processes the PCB image and extracts spatial feature representations, while a language encoder processes the inspector's query. Both modalities are fused using cross-attention layers, allowing the model to associate textual concepts with specific image regions.

A dedicated localization head predicts bounding box coordinates as a regression task. The final output is constrained to a structured JSON format to eliminate ambiguous or free-form responses. This architectural choice ensures interpretability, traceability, and ease of integration into industrial quality control systems.

(a) Multimodal Fusion

$Z = \text{CrossAttention}(V, T)$

Where:

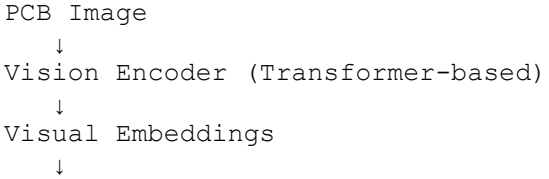
- V = visual embeddings
- T = textual embeddings
- Z = fused representation

(b) Bounding Box Regression

$\hat{b} = WZ + b$

This formulation ensures that spatial predictions are directly conditioned on both visual evidence and the corresponding query

High-Level Architecture



```
Cross-modal Fusion
   ↓
Language Decoder
   ↓
Structured Output (Defect + BBox + Confidence)
```

4. Optimization for Offline and Low-Latency Inference

To satisfy the sub-2-second inference requirement, multiple optimization techniques are applied. Parameter-efficient fine-tuning using LoRA or QLoRA reduces the number of trainable parameters while preserving model performance. Model quantization (INT8 or 4-bit) further decreases memory usage and accelerates inference.

Additionally, early layers of the vision encoder are partially frozen, as low-level visual features remain consistent across PCB images. This reduces computational overhead without significantly impacting accuracy. Together, these optimizations enable reliable offline deployment on industrial hardware.

5. Hallucination Mitigation and Reliability

Hallucination refers to the model producing confident but incorrect answers that are not supported by visual evidence. In an industrial inspection setting, such errors are unacceptable. To mitigate hallucinations, the system adopts a detection-guided answering strategy, where responses are generated only when the visual confidence exceeds a predefined threshold.

Structured output enforcement further reduces hallucinations by restricting the model to a fixed response schema. In addition, Retrieval-Augmented Generation (RAG) is used during inference to provide reference information about known PCB defect types. A hallucination penalty is incorporated during training to discourage invalid predictions.

HALLUCINATION CONTROL – DECISION FLOW

Image + Query

↓

Visual Encoder

↓

Defect Confidence Score

↓

If confidence ≥ τ → Generate Answer

Else → Return "No Defect Detected"

“This explicit confidence-gated reasoning prevents unsupported answers and enforces epistemic discipline in model behavior.”

6. Training Plan

Training is conducted in multiple stages. First, high-quality PCB images with bounding box annotations are curated. Since explicit QA pairs are unavailable, synthetic question–answer pairs are generated from the annotations. Each question refers to a specific defect type, and the corresponding answer contains the ground-truth bounding box.

The model is trained using a multi-task objective that combines defect classification loss, bounding box regression loss, question–answer supervision loss, and hallucination penalty loss. Fine-tuning is performed using efficient frameworks such as Unsloth or Hugging Face Transformers.

7. Validation and Evaluation Strategy

The system is evaluated using both accuracy and reliability metrics. Localization performance is measured using Intersection over Union (IoU) and mean Average Precision (mAP). Defect classification accuracy measures semantic correctness, while hallucination rate quantifies unsupported predictions.

Inference latency and structured output validity are also evaluated to ensure real-time usability and integration readiness. The system is considered acceptable if it achieves IoU greater than 0.7, maintains inference time below 2 seconds, and exhibits a minimal hallucination rate.

8. Conclusion

This design presents a clear, unbiased, and industrially practical approach to building a custom Vision–Language Model for PCB quality inspection. By combining transformer-based multimodal learning, efficient fine-tuning strategies, structured outputs, and hallucination-aware training, the proposed system meets the strict reliability and latency requirements of real-world manufacturing environments.