Tech4Biz

Excellence Guaranted!

CASE STUDY 03

INTELLIGENT VISUAL INSPECTION WITH REASONING

AI Vision and Edge Language Model Quality Assurance on the AMD Kria KR260

Machine vision finds and measures the defect. A quantized language model on the same device retrieves the applicable standard operating procedure, derives the disposition, and writes back to MES with the reason attached. The output is a decision an auditor can read, not just a bounding box.

The Business Problem

Where conventional machine vision stops

A deployed inspection system detects a defect and reports it. That is genuinely useful, and it is also where the automation ends. The operator at the station still has to decide whether the part is accepted, reworked, rejected, quarantined or escalated, and whether the line should stop.

That decision is not a vision problem. It requires the tolerance table for that part, the current revision of the standard operating procedure, the customer specific concession if one is in force, and an understanding of what the defect means for the assembly downstream.

What goes wrong

- **Delay at the station.** The part waits while the operator consults a document or a supervisor.
- **Inconsistency between shifts.** Two operators looking at the same scratch reach different dispositions, and neither is written down in a way that can be reviewed.
- **Over-rejection.** Under time pressure the safe answer is to reject, which is scrap the plant did not need to make.
- **Escapes.** Under throughput pressure the fast answer is to accept, which is a customer complaint later.
- **Unauditable decisions.** When a customer asks why a part shipped, the record shows a measurement and a disposition with nothing connecting them.
- **Knowledge concentration.** The people who can make the marginal calls reliably are few, and they retire.

The gap in one line

> THE REQUIREMENT
>
> Turn a measurement into a disposition, automatically, using the plant's own standard operating procedure, and record the reasoning alongside the result so that a customer auditor can follow it years later.

Why an edge language model is the right tool

The reasoning required is not complex arithmetic. It is reading a procedure, comparing a measured value against a stated tolerance, applying the exceptions, and stating the conclusion in a sentence. That is precisely what a small instruction tuned model does well, and it is precisely what a rules engine does badly once the procedures change frequently and are written in prose.

Why it has to run on the line

- Station cycle time is measured in seconds. A cloud round trip does not fit inside it.
- Production data, defect imagery and process capability are among the most sensitive things a plant holds, and most manufacturers will not let them leave the site.
- The line must keep running when the site link does not.

Solution Overview

A Kria KR260 sits at the inspection station. It triggers with the conveyor, detects and measures the defect on the programmable fabric, retrieves the governing procedure from local storage, derives the disposition, and writes the result and the reason back to MES over OPC-UA.

Layered architecture

LAYER 6 ACTION
MES, ERP and SCADA writeback
Disposition, reason, evidence reference and routing instruction.

LAYER 5 REASONING
4-bit quantized language model
Measurement compared against the retrieved tolerance, disposition derived with a stated reason.

LAYER 4 GROUNDING
SOP and tolerance retrieval
Version pinned procedures, tolerance tables and active concessions held locally.

LAYER 3 METROLOGY
Measurement in engineering units
Calibrated pixel to millimetre conversion, sub-pixel edge extraction.

LAYER 2 DETECTION
YOLO and segmentation on the DPU
Defect located, classified and masked at line rate.

LAYER 1 ACQUISITION
Encoder triggered capture
Industrial camera, controlled lighting, part in register.

Division of responsibility

Vision does	The language model does
Find the defect	Read the applicable procedure
Classify the defect type	Apply the tolerance and exceptions
Measure it in millimetres	Derive the disposition
Report a confidence	State the reason in one sentence
Produce an evidence image	Choose the routing instruction

> EXPLAINABILITY IS THE DELIVERABLE
> The commercial value here is not that the decision is automated. It is that the decision is automated and explained. A disposition that carries the rule it was derived from is an audit record. A disposition on its own is an assertion.

The confidence gate

Below a configured detection confidence, or where retrieval fails to return a governing procedure, the part is escalated to a human rather than dispositioned. This is a hard gate, not a tuning preference.

Hardware, Optics and Acquisition

Station hardware

Element	Configuration
Platform	AMD Kria KR260
System on module	Kria K26, Zynq UltraScale+ MPSoC
Acceleration	DPU on the programmable fabric with Vitis AI
Memory	4 GB on-module DDR4
Camera	Industrial area or line scan, GigE Vision or MIPI
Trigger	Conveyor encoder, hardware triggered
Lighting	Controlled industrial illumination with strobe control
Control interface	PLC over OPC-UA or fieldbus
Storage	NVMe for models, procedures and evidence buffer

Why the KR260

The programmable fabric handles image pre-processing, the detection network and dequantization for the language model without contending for the application cores. That matters at line rate, where the vision stage and the reasoning stage are running against the same cycle budget.

Optics discipline

Measurement accuracy is decided by the optical setup, not by the model. A detection network cannot recover information the lens and the lighting did not capture, and no amount of reasoning downstream can compensate for a bad measurement upstream.

- **Telecentric optics** where dimensional accuracy matters, to remove perspective error across the depth of the part.
- **Controlled, repeatable lighting.** Ambient light contamination is the single most common cause of inspection systems that work in commissioning and drift in production.
- **Fixed working distance and rigid fixturing.** Any mechanical change invalidates the calibration.
- **Encoder triggering** so the part is in the same position in every frame, rather than time triggered capture that drifts with line speed.
- **Strobe synchronisation** to freeze motion without motion blur at production speed.

> **COMMISSION THE OPTICS FIRST**
>
> Every hour spent on optics and fixturing before the models are trained saves several later. The most common failure pattern in this class of project is a well trained model on an unstable optical setup.

Vision Pipeline and Metrology

Detection stage

01
Pre-processing on the fabric
Debayer, flat field correction, region of interest extraction.

02
Detection network on the DPU
YOLOv11 or RT-DETR, quantized to INT8 through the Vitis AI quantizer.

03
Segmentation for extent
A mask rather than a box, because the disposition depends on size and shape.

04
Sub-pixel edge refinement
Classical technique on the refined mask, because a network alone is not a metrology tool.

05
Conversion to engineering units
Calibrated pixel to millimetre transform applied, with uncertainty carried forward.

Calibration and uncertainty

- Calibration target imaged at commissioning and at defined intervals thereafter.
- Measurement uncertainty is stated, not implied. A depth of 0.15 mm with an uncertainty of 0.02 mm is a different decision near a 0.10 mm limit than one with an uncertainty of 0.08 mm.
- Gauge repeatability and reproducibility study run against the automated measurement, exactly as it would be for a manual gauge.

What the reasoning layer receives

```
{
  "part_serial"    : "SN-4471-22908",
  "station"        : "INSP-03",
  "captured_at"    : "2026-07-29T11:04:17Z",
  "detections": [
    { "class"       : "surface_crack",
      "confidence"  : 0.993,
      "location"    : "face_A, zone 2",
      "measurement" : {
        "type"        : "depth_mm",
        "value"       : 0.15,
        "uncertainty" : 0.02 },
      "extent_mm2"  : 4.8 } ],
  "evidence_ref"   : "img/2026-07-29/INSP-03/22908.png",
  "vision_model"   : "yolo11-surface-v7",
  "calibration_id" : "CAL-INSP03-2026-06"
}
```

NOTE WHAT IS ABSENT

The language model never sees the image. It receives a measured, calibrated, uncertainty bearing description. That keeps inference fast enough for line rate, keeps the reasoning testable against fixed inputs, and makes it impossible for the model to invent a defect that the vision stage did not measure.

The Reasoning Layer

Grounding corpus

The model answers from the plant's own controlled documents, held locally and version pinned to the part number and the current engineering change level.

- Manufacturing standard operating procedures
- Tolerance and acceptance tables
- Customer specific requirements
- Active concessions and deviations
- Rework routing definitions
- Defect classification standard

Decision contract

Condition	Required behaviour
Clear pass	Accept, record the rule applied.
Clear fail	Reject or route to rework per the procedure, with the reason.
Within measurement uncertainty of the limit	Escalate. Never resolve a borderline case in either direction.
Confidence below the gate	Escalate, with the detection shown to the operator.
No governing procedure retrieved	Escalate and flag the corpus gap.
Conflicting documents	Escalate. Do not choose between controlled documents.
Active concession applies	Apply it and name it explicitly in the reason.

Structured output

```
{
  "part_serial"   : "SN-4471-22908",
  "disposition"   : "reject",
  "reason"        : "Depth 0.15 mm exceeds the
                     0.10 mm limit in SOP-QA-114
                     rev 7, section 4.2, by 45%.",
  "rule_ref"      : "SOP-QA-114#4.2",
  "sop_revision"  : "rev 7",
  "routing"       : "rework_line_3",
  "est_repair_min": 8,
  "escalated"     : false,
  "model_version" : "reasoner-2026-06-a"
}
```

> **CONSTRAINED GENERATION**
>
> The disposition field is generated under a grammar that permits only the defined enumeration. The model cannot emit a disposition that the quality system does not recognise, because the decoder will not let it. Free text is confined to the reason field, where it is exactly what is wanted.

Human in the loop

Escalated parts are presented to the operator with the detection overlay, the measurement, the retrieved procedure section and the model's stated uncertainty. The operator decides, and their decision is recorded against the same schema, which builds the dataset for tuning the gate.

Inference Stack and Memory Budget

Two models, one module

This station runs a vision network and a language model concurrently against the same memory and the same cycle budget. Sizing is therefore tighter than in a language-only deployment and is treated as a hard constraint from the start.

Component	Allocation
Detection network, INT8	~40 to 120 MB
Segmentation model	~60 to 150 MB
Quantized language model, 4-bit	~1.7 to 2.2 GB
Key value cache at working context	~120 to 200 MB
Procedure index resident portion	~100 to 200 MB
Frame buffers and evidence cache	Sized to line rate and retention policy
Runtime, OPC-UA stack and OS	~600 to 800 MB

STATUS OF FIGURES

These are design targets for the reference configuration. Final allocations depend on part complexity, line rate, resolution and retention policy, and are confirmed during the engineering phase.

Why four bit

At sixteen bit a three billion parameter model needs roughly six gigabytes for weights alone, which does not fit alongside the vision stack on a four gigabyte module. At four bit it comes down to approximately 1.7 gigabytes and the two workloads coexist.

Token generation is bound by memory bandwidth rather than arithmetic, because the full weight set streams out of DDR for every token. Halving the bytes per weight roughly doubles the achievable rate.

The DDR pipeline

TILING
Weights tiled to BRAM and URAM

DOUBLE BUFFERING
Transfer overlapped with compute over AXI

PREFETCH
Next layer fetched during the current one

ON-FABRIC DEQUANT
Four bit bytes cross the bus, not sixteen

ARBITRATION
Vision and language traffic scheduled against each other
Without this the two workloads contend and both degrade unpredictably.

Integration and Traceability

Integration surface

System	Interaction
MES	Disposition, reason, rule reference and evidence pointer written against the part serial.
ERP, SAP	Scrap, rework and yield postings through the existing interface.
SCADA	Station state, availability and alarm conditions.
PLC	Physical routing of the part to accept, rework or reject lanes.
OPC-UA	Primary structured interface to the plant floor.
MQTT	Event distribution to plant analytics where it exists.
Historian	Long term measurement and disposition trend storage.

Physical routing

The station writes a disposition. The PLC performs the physical routing. That separation is deliberate, so that a failure or a stall in the inspection node never leaves parts unrouted or the conveyor blocked. The PLC has a defined default behaviour for a station that does not respond in time.

The traceability record

Every inspected part produces a record that ties the physical part to every version of everything that contributed to the decision. This is what makes the system usable in a regulated or customer audited environment.

Recorded	Why it is needed
Part serial and timestamp	Ties the decision to the physical part.
Evidence image reference and hash	Proves the image was not altered after the fact.
Measurement and uncertainty	Shows what was actually observed.
Vision model version	Identifies which detector produced it.
Calibration identifier	Proves the measurement was traceable.
SOP document and revision	Shows which rule was in force.
Reasoning model version	Identifies the decision logic.
Disposition and stated reason	The decision and its justification together.
Operator override if any	Records where a human disagreed and why.

Worked Example, End to End

STEP 1 CAPTURE
Encoder triggered frame
Part SN-4471-22908 in register at station INSP-03.

STEP 2 DETECT
Surface crack, confidence 99.3 percent
Detected on face A, zone 2, segmented for extent.

STEP 3 MEASURE
Depth 0.15 mm, uncertainty 0.02 mm
Sub-pixel edge extraction, calibrated conversion applied.

STEP 4 RETRIEVE
SOP-QA-114 rev 7, section 4.2
Allowed tolerance 0.10 mm for this part and zone. No active concession.

STEP 5 REASON
Limit exceeded by 45 percent
Outside measurement uncertainty of the limit, so the case is clear rather than borderline.

STEP 6 ACT
MES updated, PLC routes to rework 3

Recorded output

Part	SN-4471-22908
Defect	Surface crack, face A zone 2
Confidence	99.3 percent
Measurement	0.15 mm depth, uncertainty 0.02 mm
Rule applied	SOP-QA-114 rev 7, section 4.2
Disposition	**Reject**
Reason	Tolerance exceeded by 45 percent
Routing	Rework line 3
Estimated repair	8 minutes

Illustrative. Thresholds, dispositions and routing definitions are taken from the customer's own controlled procedures.

THE BORDERLINE CASE, FOR CONTRAST

Had the measurement been 0.11 mm with an uncertainty of 0.02 mm, the limit would sit inside the uncertainty band. The system would escalate rather than reject, and the operator would see the measurement, the band and the rule. Getting this case right is what separates a usable inspection system from one the quality team quietly stops trusting.

Quality System, Drift and Revalidation

Fitting the plant's quality system

An automated disposition is a change to the control plan. It has to be introduced the way any other change to the control plan is introduced, with evidence rather than with a demonstration.

Requirement	How it is met
Measurement system analysis	Gauge study run against the automated measurement before it is trusted for disposition.
Control plan update	Station documented as a control point with defined reaction plan.
Document control	Procedures retrieved by the system are the controlled revisions, synchronised from the document management system.
Change control	Model and procedure updates follow the plant's change process, not a software release process.
Records retention	Traceability records retained per the customer and regulatory requirement.
Non-conformance handling	Escalated parts route into the existing non-conformance process unchanged.

Drift is the real operational risk

Inspection systems rarely fail suddenly. They degrade, because something physical changed and nobody connected it to the detection performance.

- **Optical drift.** Lens contamination, lamp ageing, fixture movement, ambient light ingress.
- **Process drift.** A new supplier, a new surface finish, a tooling change that alters how defects present.
- **Population drift.** The defect mix changes, so the model sees a distribution it was not calibrated against.
- **Document drift.** The SOP is revised and the local corpus is not resynchronised.

> **REVALIDATION GATE**
>
> Any change to lighting, optics, fixturing, part revision or governing procedure triggers a revalidation before the station is permitted to disposition automatically again. Until it passes, the station runs in advisory mode and every part is escalated. This should be enforced by the system, not by process discipline alone.

Continuous monitoring

Confidence distribution, escalation rate, disposition mix and operator override rate are tracked over time. A rising override rate is the earliest reliable signal that something physical has changed.

Deployment and Lifecycle

Release management

Vision model, reasoning model, procedure corpus and application are released as one signed bundle with a single version number. Every traceability record names that version, so any decision can be reproduced later with the exact components that made it.

BUILD
Quantize, index, evaluate, sign
Regression set of known parts with known correct dispositions must pass.

STAGE
Deployed in advisory mode
Runs alongside the current process without dispositioning.

VALIDATE
Agreement measured against operators
Disagreements reviewed individually, not in aggregate.

ACTIVATE
Automatic disposition enabled
Confidence gate initially set conservatively and relaxed with evidence.

ROLLBACK
Previous bundle retained
Reversion to advisory mode is always available.

Delivery phases

Phase	Scope	Exit criteria
Phase 0 Assessment	Defect taxonomy, SOP audit, optics feasibility, MES interface review, sample part collection.	Agreed scope and measurable defect list.
Phase 1 Optics and metrology	Station optics, lighting, fixturing and calibration established. Gauge study run.	Measurement capability demonstrated.
Phase 2 Detection	Detection and segmentation models trained on plant imagery and deployed to the DPU.	Detection performance at line rate.
Phase 3 Advisory mode	Reasoning layer live, dispositions generated but not applied, compared against operators.	Agreement rate and escalation behaviour accepted by quality.
Phase 4 Automatic disposition	Control plan updated, station operating as a control point.	Yield, escape and false reject targets held.
Phase 5 Expansion	Additional stations, part families and defect classes.	Repeatable deployment package.

Benefits and Measurement

Benefits realised

Explainable decisions
Every disposition carries the rule, the revision and the measurement it was derived from.

Automated quality calls
Routine dispositions stop waiting on an operator to consult a document.

Faster inspection
The decision happens in line rather than at a separate review station.

Standardised reasoning
The same defect receives the same answer across every shift and every operator.

Reduced over-rejection
Parts within tolerance stop being scrapped defensively under time pressure.

Fewer escapes
Parts outside tolerance stop being accepted under throughput pressure.

Audit ready record
Measurement, rule, revision and disposition stored together against the part serial.

Knowledge retained
Marginal case judgement stops living only in a handful of experienced operators.

KPI framework

Measure	What it tells you
False reject rate	Good parts scrapped. Usually the largest recoverable cost.
Escape rate	Bad parts shipped. Usually the largest risk.
Escalation rate	How much human attention the station still consumes.
Operator override rate	Whether the reasoning is trusted, and the earliest drift signal.
Station cycle time	Whether inspection is still the bottleneck.
Disposition consistency	Variation across shifts, which should approach zero.
Time to disposition	Delay removed from the line.
Audit response time	How long it takes to answer a customer query about one part.

BASELINE FIRST

Capture current false reject rate, escape rate and inspection cycle time before deployment. These are the numbers the business case rests on, and they are almost never recorded well enough beforehand to reconstruct afterwards.

Engineering Considerations and Risks

Known hard problems

Risk	Mitigation
Measurement accuracy	Optics and metrology established and gauge studied before any model work begins. The reasoning layer inherits whatever the vision stage gives it.
Procedure drift	Corpus synchronised from the document management system, with the revision recorded in every decision.
Borderline dispositions	Measurement uncertainty carried through and used to force escalation near a limit, rather than resolved silently.
Defect class imbalance	Rare defects are the expensive ones. Targeted collection and, where necessary, physical sample creation.
Optical and process drift	Revalidation gate enforced by the system on any physical or revision change.
Quality team acceptance	Advisory mode for long enough to build evidence, with disagreements reviewed case by case.
Line rate contention	Vision and language memory traffic explicitly arbitrated rather than left to contend.

Open questions for the customer

- Which defect classes are in scope, and can each of them actually be measured optically at line rate?
- Are the governing procedures available in a controlled electronic form that can be synchronised, or do they exist as scanned documents and tribal knowledge?
- What is the station cycle time, and how much of it is available to inspection?
- Which MES is in use, and what interface does it expose for disposition writeback?
- Is the plant subject to customer audit, IATF, GxP or another regime that constrains automated disposition?
- Who owns the decision to change the control plan, and what evidence will they require?
- What is the retention requirement for evidence images, and does the storage sizing reflect it?

RECOMMENDED FIRST STEP

A single station in advisory mode on one part family. It produces dispositions and reasons without applying them, which gives quality a measured agreement rate against their own operators before anything changes on the control plan, and gives the business case real false reject and escape numbers rather than estimates.

Contact

Website

https://tech4bizsolutions.com

Contact details

If you would like to know more about Tech4Biz Solutions
and our products please contact us via email contact@tech4biz.io

Address

Bangalore
NGR Arcade,
HSR Layout Sector 1,
Bangalore, Karnataka

Surat
306-307,
Millionaire Business Park,
LP Savani Road, Adajan,
Surat, Gujarat - 395009



Tech4Biz

Excellence Guaranted!

Tech4Biz is a provider of comprehensive engineering and IT solutions for businesses of all sizes. From cloud computing and data management to security, networking and embedded intelligence, we have the expertise and experience to help your business stay ahead of the curve.

Our mission is to provide businesses with the best possible technology solutions. We understand that a reliable and efficient technology foundation is crucial in an increasingly competitive marketplace. That is why we offer a wide range of services, from managed IT and cloud computing to edge and embedded AI engineering.

This communication contains general information only, and Tech4Biz Solutions Private Limited is not, by means of this communication, rendering professional advice or services. Architecture descriptions, performance characteristics, memory budgets and integration details in this document represent a reference design and design targets for the stated configuration. They are confirmed against the specific customer hardware, workload and duty cycle during the engineering phase. Third party product names are the property of their respective owners. Before making any decision or taking any action that may affect your finances or your business, you should consult a qualified professional adviser. Tech4Biz Solutions Private Limited shall not be responsible for any loss whatsoever sustained by any person who relies on this communication.

© 2026 Tech4Biz Solutions Private Limited