工艺波动与品控难题：AI 质检、参数优化在车间的落地

作者：**SVSAC 硅谷战略顾问委员会**

核心判断

车间 AI 项目的价值，应体现在减少真实缺陷和工艺损失上；模型演示再漂亮，也不能代替生产现场的改进。对企业而言，值得优先推进的是：优先选择损失可计量、数据可追溯的质量问题，把误检漏检、工况变化和人工复核纳入同一验收。本篇结合公开资料与分析，讨论企业可以如何识别机会、组织能力并安排下一步投入。

关键词：制造质量；视觉检测；过程漂移；异常检测；参数优化；质量验证；数据治理

SVSAC - sv-sac.org

车间 AI 项目的价值，应体现在减少真实缺陷和工艺损失上；模型演示再漂亮，也不能代替生产现场的改进。

我们的建议是：优先选择损失可计量、数据可追溯的质量问题，把误检漏检、工况变化和人工复核纳入同一验收。企业可以先据此梳理自己的产品、客户、能力与投入边界，把讨论推进到可以分工、可以验证的具体任务。

对管理层来说，这个议题值得放进经营议程。围绕“先做什么、凭什么做、做到哪一步再扩大投入”形成共识，往往比同时铺开多个方向更有推进价值。下文保留相关资料、比较和推演，供企业判断自身条件与行动顺序。

1. 核心问题、结果变量与方法

1.1 从“发现缺陷”转向可验证的质量结果

制造现场所谓质量波动，至少有三个容易混为一谈的对象。第一是产品的真实状态：尺寸、强度、表面、连接完整性、功能寿命等特性是否落在规定范围内。第二是测量与判定：检验系统能否在规定条件下区分合格、不合格以及需要复核的边界件。第三是生产过程：输入材料、设备状态、环境、换型、操作与控制参数如何共同生成产品状态。真实产品状态是目标，检验结果只是对目标的测量，过程数据则是可能的先行信息。把三者视为同一变量，会造成循环论证：模型训练标签来自既有检验，模型又以复现既有标签的能力宣称改善了真实质量。

因此，本篇不以模型分数作为最终结果变量。质量结果至少应落到独立定义的产品特性或缺陷事件上，并同时记录其测量方法、判定规则和观察窗口。若目标是降低外观缺陷，主要结果可以是按预先定义的缺陷类别与产品数计算的缺陷率；如果目标是降低漏放行，则必须追踪模型未拦截、后续独立检验确认的不合格件；若关心过程稳定性，还应观察关键特性随时间、批次、设备或材料变化的分布，而非只比较均值。报废、返工、复检工时和停线属于重要的后果变量，但它们不能替代产品符合性本身。

一个可解释的因果链可以写作：原材料、设备与环境状态影响工序输出；工序输出形成可测质量特性；检验系统采样并产生判断；阈值和处置规则决定放行、拦截、返工或停线；后续独立确认揭示漏检和误报；这些结果再影响最终交付质量与成本。AI 可能介入链条的多个位置，但在任一位置改变后，后续环节也会响应。例如，检测阈值调严，理论上可能提高异常召回，却会增加误报和复核负担；一旦复核能力不足，积压、跳过复核或误判都可能抵消检测收益。这个机制是待现场数据检验的假设，不能仅由模型 ROC 曲线推出。

SVSAC | sv-sac.org

本篇的核心核心问题是：在缺陷稀少、产品变型和过程状态随时间变化的制造环境中，怎样区分视觉识别性能、过程预测价值、参数干预效果与闭环质量改善，并为每种主张配置相称的验证证据？它分解为三项：一，离线或公开数据集表现如何映射为现场的漏检、误报和复核工作量；二，预测性信号在何种时间顺序和工艺机制下能支持参数调整，而非只与产品质量相关；三，从建议到实际控制的每次变化，如何保留安全约束、追溯和回退能力。

研究范围限定于一般制造现场的质量检验与工艺调整，不假定特定行业法规、设备协议或工厂数据。本篇采用定向范围梳理，而非系统综述或元分析。纳入材料包括：原始同行评议论文及其公开实验说明、论文所引数据集的发布页面、制造参数优化的实体试验研究，以及政府网站发布的法律原文。分析时依次记录研究是图像基准、实体设备试验还是现场运行资料；预测单位是像素、图像、产品、批次还是时间窗；指标是区分能力、召回、缺陷率、过程特性还是经济结果；是否有对照和独立确认。由于缺少统一的企业原始资料、共同缺陷定义和成本结构，本篇不汇总所谓“行业平均准确率”或投资回报率，也不作跨研究的数值合并。

1.2 四类技术任务与证据边界

“AI 质检”常覆盖四种任务，边界首先由输出及其下游动作界定。

第一类是视觉检测。输入通常是相机或其他成像设备捕获的图像，输出可能是图像级异常分数、缺陷类别、框或像素掩膜。其直接问题是：给定成像配置与被评估的产品总体，模型能否对目标缺陷排序、分类或定位？它不自动回答一个完整产品是否合格，也不回答缺陷成因。成像视角未覆盖的区域、被遮挡的缺陷、光照变化和相机污染，都构成模型外的盲区。

第二类是过程状态与质量特性预测。输入是设备传感器、工艺配方、材料批次、环境或在线测量信息，输出是某一时点或未来窗口的状态、质量特性预测或漂移报警。此类模型的价值取决于预测提前量与可操作窗口：提前几十秒识别的漂移若来不及停机或取样，统计上可预测却没有现场决策价值。随机拆分时间序列可能让相邻窗口同时落入训练和测试集，因而高估对未来批次、换型或设备状态的外推能力。

第三类是参数优化或推荐。算法依据历史数据、物理约束或新的试验，给出候选的温度、速度、压力、配比、进给量等设定。推荐的可信度不只取决于目标函数的预测误差，还取决于搜索范围是否覆盖真实可运行区域、输入变量是否可

SVSAC | sv-sac.org

控、目标是否把质量与产能或能耗冲突明确表达，以及试验样本是否足以辨别因果效应。参数–质量的相关性预测不等于“把参数设为某值就能改善质量”。

第四类是闭环控制。控制系统将测量、估计、决策与设备动作连续或周期性连接，动作可能影响安全、设备寿命、产品一致性与人员作业。离线模型推荐与闭环控制之间，至少还隔着可验证的执行接口、控制时延、异常处理、权限授权、状态估计、工程安全限值和回退机制。本篇将直接自动更改机台参数视作需要单独论证的高风险控制设计，不把“AI 建议”或数字孪生演示称为已经实现的自主控制。

这四类任务彼此可能组成系统，但不应共享一个模糊成功指标。视觉模型像素 AUROC 较高，既不表示每个缺陷类别都在既定阈值下达到所需召回，也不表明实际批次缺陷率下降；预测模型能预警漂移，也不表示工艺变更造成改善；优化器找到历史数据中目标函数的高值，也不说明该组合安全可生产；控制器稳定执行，则仍需单独验证其运行窗口内产品质量与异常恢复。后文据此分别讨论检测、预测、优化和控制的证据要求。

2. 由图像分数到整件质量：检测链的转换条件

2.1 缺陷率决定检测指标如何转成工作量

把分类指标用于生产决策时，必须恢复现场的基率。设目标缺陷在待检产品中的比例为 (p)，模型在选定阈值下的灵敏度（缺陷召回）为 (r)，特异度为 (s)。在每件产品都接受一次判定、且参照检验可靠的简化条件下，每检一件产品的漏检期望数是 (p(1-r))，误报期望数是 ((1-p)(1-s))。在低缺陷率下，即使特异度看上去很高，正常件的规模也可能使误报数量超过真缺陷数量。反过来，精确率会随缺陷基率变化；数据集上较高的精确率不能未经重算移植到缺陷更稀少的车间。

例如，仅作算术示例：若十万件中缺陷占千分之一、灵敏度为 95%、特异度为 99%，简化期望为约 5 个漏检和约 999 个误报。此处数值不是任何工厂的估计，也不是模型基准；它说明应当把分类性能换算成复核量、漏放行后果与实际基率，才能谈阈值。若人工复核每件误报需要几十秒，千件级误报便可能占用一段完整班次；若把阈值放宽以减轻队列，漏检风险则可能上升。成本比较至少需要每类误判的后果、复核时间分布、复核准确性与产线节拍，而非把漏检和误报视为等价错误。

实际产线往往还有“弃权”与复检区间：分数不确定时暂缓判定，由人工或更精密测量复核。增加第三个状态能将模型不确定性显式交给作业流程，却会引入

SVSAC | sv-sac.org

额外工位、等待和人员培训成本。弃权比例应纳入吞吐和复核负荷的结果变量，不能只报告模型在被自动判定样本上的准确率。并且，人工复核不是天然无误的金标准。若复核人员不知道模型的警报来源，能降低一类确认偏差；若质量标准本身存在分歧，则应通过重复测量、裁决与量具分析估计标签不确定性，而不是把单次历史标签当作无噪声真值。

质量的后续追踪还需要定义“同一件产品”。图像、工件序列号、批次、工位、设备配方与实验室检测记录必须能按时间关联。如果模型报警后产品被返工，而台账只保留返工后的最终结果，原始缺陷就可能从数据中消失；若只对疑似件做破坏性检测，则未报警产品缺少同等强度的参照，漏检率便不可识别。比较部署前后缺陷率时，应该保持抽样复核或独立终检策略足以观察未报警件，否则系统越会“挑出问题”，标签分布越可能随其自身行为改变。

2.2 从像素定位到作业判定

像素级定位任务将每个图像位置标记为缺陷或正常，适合检查模型是否大致覆盖标注区域，但其指标不直接对应产品拒收。一个掩膜可以在视觉上覆盖主要缺陷，却未覆盖功能关键区域；也可以在像素层面有大量轻微错位，而整件分类仍正确。反过来，一个很小但关键的缺陷像素比例极低，像素平均指标可能被大面积正常背景主导。产品层面的判定要经过聚合规则，例如“任何关键区域超限即拒收”、按缺陷尺寸分级，或按多个视角联合判定。聚合规则必须先由工程和质量责任人界定，再评估模型，而非由离线数据表现反向选择。

MVTec AD 可作为受控视觉研究的文献线索。本篇可读的 MVTec AD 2 相关工作转述其含 5,354 张图像、15 类对象与纹理；[4]这属于后续论文的二手说明，不是本篇对 Bergmann 等原论文或数据的独立核验。[1] 原始训练、测试及标注协议未核，不据此断言其具体拆分。对于任何图像基准，都必须另外判断采集与拆分是否代表目标产线；测试集缺陷混合比例不能直接作为工厂缺陷基率。

MVTec LOCO AD 进一步区分表面结构异常与逻辑异常，包含 3,644 张、五类对象图像。作者刻意纳入数量、位置或组合违反常规的情况，例如零件缺失或装配逻辑错误；论文也说明拍摄中使用规定位置、机械对准与为任务选择的照明和镜头[2]。这一设计很有价值，因为工件表面完整并不保证装配正确。但它同时提醒，benchmark 中的“接近真实应用”仍指任务设计与成像设置有工业启发，不意味着已在持续生产、换班、维护和材料变化条件下开展前瞻现场试验。

可读的 MVTec AD 2 相关工作还转述 Real-IAD 约有 15 万张图像、30 种对象；[4]本篇未取得 Wang 等原论文，不把二手规模数字当成对其多视角、样本级

SVSAC | sv-sac.org

评价或隔离协议的核实。[3] 对一个产品采用多幅图像时，应预先规定如何汇总分数，并核对采集是否来自目标部署、缺陷是否覆盖当地失效模式、跨时间与批次是否隔离。这些是我们提出的评价要求，不能由数据集名称或规模代替。

MVTec AD 2 提出八个新异常检测场景、超过 8,000 张图像，重点覆盖透明或重叠对象、不同照明和细小缺陷等难点；作者报告若干方法在其新场景上的像素区域指标明显低于传统基准表现[4]。这类扩展并不能直接替代本厂验证，却构成一项有用的反证：模型在常用数据上的高分未必可迁移到挑战性照明、缺陷尺寸或正常状态波动。研究上应报告跨场景性能变化与失败案例，而不是只展示一个类别平均值。

InsPLAD 列为运行资产图像研究的待核文献线索。[9] 原文未取得，因此不将图像数量、地点、资产类别、缺陷标签粒度或遮挡分布写成已核实结果。本篇仅提出任务区分：运行资产巡检与制造工件在线检验的观察单位、成像机会和后续动作可能不同。是否存在这种差异及其幅度，应由可读数据说明或实际采集记录确定，不能凭“真实环境”标签推断工厂部署收益。

由此可见，公开数据集的合理用途是筛选方法、暴露已定义条件下的边界，并使算法比较可复现；不能直接证明在线漏检、质量成本或客户投诉改善。要把基准证据推进到车间，应使用与目标产品、相机、光照、速度、班次和缺陷定义相匹配的时间外样本；对每一个缺陷和批次报告分层召回、误报/千件、弃权率和处理时延；并保留独立终检抽样，让未报警产品也有机会成为已知真值。

2.3 数据拆分与罕见事件造成的偏乐观

图像数据如果来自同一工件的连续帧、多个裁切或同一批次的近重复拍摄，随机按图像拆分会让高度相似的图像同时进入训练与测试。模型由此可能利用夹具、背景、纹理、位置或批次特征，而非真正识别缺陷机制。对于需要对未来生产作出判断的任务，测试划分应按工件、批次、生产日期、设备或生产线隔离；哪个层级是独立单位，取决于实际外推目标。模型拟部署到同一台设备的下一周，至少要测试时间后移；拟迁移到不同设备或工厂，则该设备/工厂应作为独立外部测试集。一个拆分方案无法同时证明所有迁移能力。

另一个被忽视的问题是评价单位不等同。像素 AUROC 把所有正常与异常像素的分数排序汇总；图像 AUROC 按图像汇总；产品误拒率则按每件产品在固定阈值和固定处置规则下计数；批次逃逸率还取决于抽样和后续检查。AUROC 不指定操作阈值，也不揭示某一规定漏检约束下需要承担多少误报。应补充阈值下

SVSAC | sv-sac.org

的混淆矩阵、精确率与召回、PR 曲线、分缺陷/分批次结果和置信区间。低基率时，精确率与误报负荷尤其应在代表性现场总体中估计。

验证也应分开三个数据阶段。开发集用于模型选择与阈值探索；锁定的时间外测试集只用于一次性评估预先选定的模型与阈值；部署后监测则观察分布漂移、复核队列、标签迟到和潜在故障。若每次看测试结果再调阈值，测试集已成为开发数据。若新缺陷类型在部署后持续增加，却只在既有类别上汇报平均分，则整体均值可能掩盖关键失效。合理报告需包括模型版本、图像预处理与硬件配置、阈值依据、验证总体及未覆盖范围。

3. 预测、优化和控制：从关联到可操作变化

3.1 预测的价值来自提前量而非分数

工艺波动不一定在最终检验图像中可见。温度、振动、电流、压力、进给或原料属性可能在质量特性恶化前先变化，因此时间序列模型有机会提供预警。但“先于缺陷出现”须由按时间顺序的评估证明。标签对齐错误会制造虚假的预测能力：例如，若用已发生缺陷后的设备报警预测该缺陷，模型只是在检测后果；若窗口跨越产品出站时点，未来数据会泄漏；若同一批次的相邻样本随机分割，模型可能学会批次特征。

可操作的预警需要把预测对象和决策窗口一起定义。若系统每分钟打分，而调机、测量和重新开线至少需要半小时，预测提前量小于恢复周期便可能没有处置空间。相反，提前很久的告警可能因设备状态和材料条件随后变化而失去准确性。可以按产品序列或生产区段记录预警至缺陷的时间差、未发生异常的报警暴露时间、检测延迟和复核结果，并用固定提前窗定义召回与误报。最终要检验的不只是预测是否正确，还包括告警是否触发了可执行动作，动作是否改变了预先指定的产品特性。

过程预测具有相关性但未必可干预性。若传感器读数与缺陷同受原材料批次影响，调节该传感器对应参数未必能消除缺陷；若某设备自动补偿改变了观测到的压力，预测变量还可能是控制反馈后的结果。要主张参数变化会改进质量，需要受控试验或有可信识别策略的观测研究，且必须描述同时发生的维护、换料、换型和操作班次。预测能用于排序复检或提前取样，不自动构成因果解释。

3.2 参数优化是受限试验问题

贝叶斯优化等序贯方法适用于每次实验代价较高、响应函数未知、变量连续且可以在约束域内试验的情境。优化器依据已观测到的参数–结果数据拟合代理

SVSAC | sv-sac.org

模型，再选择下一批候选点以平衡利用与探索。制造业中，实验代价并非单件材料价格：还包括设备准备、换型清洁、稳定时间、测量、操作者工时、设备磨损、废品处置以及生产节拍损失。若“每个点要做多件”以估计质量方差，实验成本又进一步增加。算法减少试验次数有潜在价值，但必须与基线设计比较，并核算每个有用数据点的总成本。

Maier 等的磨削研究是参数优化而非一般图像检测的具体例子：研究对象为制造硬质合金刀片的杯形砂轮磨削，考虑进给速度和切削速度，使用嵌入旋转砂轮的光纤/高温计测量接触区温度，并把表面粗糙度和磨削烧伤等质量要求及温度安全约束纳入受约束贝叶斯优化[5]。作者报告在有限试验中为未知工件与刀具组合寻找可行参数。其意义在于优化器并非仅追逐单一“质量分”，而把质量与安全约束显式纳入，且实验发生在实体设备上。外推则受限于特定机床、刀具、工件材料、测温装置与试验协议，不能据此断言其他工艺能以同样试验次数获得收益或实现全自动生产。

Guidetti 等的等离子喷涂研究仍为待核线索，[6]本篇未取得原文，不将其并行试验、设备状态处理及数值或实体结果当作已核实方法。本篇拟检验的问题是：若设备、粉末批次或热状态变化，固定参数是否仍对应同一响应？在昂贵测量与受限可行域下，序贯搜索是否比经验基线或传统试验设计降低总试验成本？这些条件命题需要本地对照，不能从未读案例得出答案。

但序贯优化的高效率依赖几个常被藏起来的前提。训练数据应覆盖足以拟合响应的区域；每次试验的运行条件要有记录；测量质量不能随试验批次变化；不可行点要能安全中止；目标函数权重要由质量与生产责任共同确定。若企业只有少数历史点且这些点集中于现行设定附近，模型在未试区域的不确定度可能很大。若优化器建议极端参数，安全边界之外的预测分数没有工程含义。实际流程应先规定安全可行域，由工艺工程师审查候选点，先在小规模试样或非关键时段试验；稳定性与质量确认后才扩大验证，而不是让算法直接覆盖 PLC 设定。

3.3 自动闭环的证明标准更高

闭环控制要求感知、状态估计、决策、执行与反馈在可接受时延内连续工作。过程模型的预测误差、通信丢包、传感器漂移或控制执行器迟滞都可能导致动作不合适。传统控制器在其设计范围内有既定的稳定性和联锁逻辑；机器学习策略可能改变状态转移行为。因此，AI 模型的输出进入设备之前，应明确运行域、硬约束和控制权限。对不确定或分布外输入，安全策略应能拒绝建议、保持

SVSAC | sv-sac.org

已验证设定或交由人员处理；对超限动作，应由独立硬联锁裁决，而非依赖同一模型自我判断。

在可追溯的初期运行中，模型可只生成建议，由操作人员接受或拒绝；每次建议、理由、版本、人员决定、设备实际参数与产品后续检验应关联到同一工件或批次。影子模式可先记录模型建议而不改变设备，用以比较建议与现行决策；随后可做有权限约束的小规模干预验证。影子模式能检验分数与工作流兼容，却不能证明动作改变会产生预期的产品结果。进一步的受控上线需设置停用条件，并保留快速回退到已验证配方的能力。这些是工程上可审计的防护建议，不代表任何通用法规已经规定唯一实施路径。

4. 现场验证、成本与数据治理

4.1 以可归因结果设计试运行

试运行应从一个可观察的质量事件开始，而不是从“需要上 AI”的技术愿望开始。可先选定产品、工位、缺陷定义和最终质量指标；在不改变现行放行规则的影子阶段，记录模型输出、人工检验和独立终检之间的对应关系。此阶段有助于估计候选阈值下每班报警量、缺陷类别覆盖与成像失效，但不会给出模型干预对缺陷率的因果效应。随后若准备改变放行、抽检或参数决策，应把操作策略定义为干预，并在安全边界内比较新旧规则；如果产品或材料批次同时变化，就需要分层、同期对照或其他识别策略，避免把自然波动误当作 AI 作用。

验证集应反映目标运行总体而不是只由易检样本组成。预先确定抽样窗口、缺陷确认方法、标签裁决和失访规则；对模型报警与未报警件都按计划抽样复核。若所有报警件都复核而正常件只抽极少量，还需用抽样权重估算真实漏检和误报，并报告置信区间。类别分布、班次、设备、产品版本和环境应被保留为分层变量；一次短暂的试运行只能支持该时间与总体范围内的判断。模型、相机、镜头、灯源、推理硬件、预处理和阈值改变，都应视作可能影响结果的系统版本变化。

试运行的主要结果最好与关键质量特性及预先定义的缺陷率相连，同时记录误拒、漏放行、复检等待、节拍、返工、报废、停线与安全事件。仅比较准确率可能掩盖交接成本：例如模型减少了终检时间，却增加了异常件转运或工程师诊断；若没有观察总链条，这只是成本从一个工位迁移到另一个工位。质量改善的估计还应区分立即检测效果和后续过程改进效果：更早发现缺陷会改变拦截时点，但只有识别根因并采取有效措施，才可能减少后续生产中的缺陷发生。

SVSAC | sv-sac.org

4.2 不把硬件采购价当作总成本

生产视觉系统的成本至少包括光学与照明设计、固定与防护、传送带触发或机器人同步、数据存储、推理计算、与 MES/QMS 或 PLC 的接口、工程集成、样本标注、复核训练、维护校准和模型更新。参数优化还需计入试件与破坏性检测、设备停机或低速、配方切换、稳定过程、工程审核、重复试验和恢复现行参数的成本。云服务或外部调试可能涉及数据传输和权限管理，现场边缘计算则需要硬件冗余、补丁、监控和生命周期维护。不能因为算法本身开源或推理延迟很低，就推断部署总成本也低。

经济评价应将误判成本与现场容量显式写入。若漏检引发客户退货、下游停线或安全后果，其成本可能远高于一件误拒的复检成本；但若异常标记导致整批隔离，误报也可能触发高昂的交付延迟。不同产品、合同和缺陷严重度会改变权重，因此本篇不提供通用成本比或 ROI。更可靠的做法是以本企业可审计的历史成本构建区间情景，分别计算灵敏度、特异度、复核容量和缺陷基率变化下的后果，并把关键假设交由质量、生产和财务共同审阅。只有实际新增收益减去集成、运维、返工转移和停机成本后仍为正，才可称为经济收益；试验论文“减少试验次数”或供应商案例“提高准确率”都不能替代这一账本。

4.3 生产影像、工艺记录与隐私安全

机器视觉数据看似只拍产品，也可能包含工人面部、工牌、手部操作、声音、位置或可识别的生产班次信息。是否构成个人信息取决于内容与可识别性，不能仅凭摄像头用途称为“不含个人信息”。中国《个人信息保护法》将与已识别或可识别自然人相关的信息纳入个人信息范围，要求处理遵循明确合理目的、最小必要范围等原则；并要求个人信息处理者采取内部制度、分类管理、加密或去标识化、权限控制和应急预案等措施[7]。法律同时对处理敏感个人信息、利用个人信息进行自动化决策、委托处理、向其他个人信息处理者提供、公开以及向境外提供等特定活动规定了事前影响评估要求[7]。这些条文是否适用于某个具体系统，需要根据实际处理目的、字段和主体关系作个案判断；本篇不对具体企业作合法性结论。

《数据安全法》对数据作广义界定，涵盖以电子或其他方式记录的信息，并将收集、存储、使用、加工、传输、提供和公开列入数据处理活动；其第二十七条要求开展数据处理活动建立全流程数据安全管理制度、开展培训并采取必要技术措施[8]。制造数据即使不包含个人信息，也可能涉及工艺配方、客户产品、设备弱点或合同保密义务。将全部车间图像上传到外部训练服务、把不同客户数据

SVSAC | sv-sac.org

混合训练、长期保留原始影像或开放远程维护账户，都应先做用途、访问、保存期限、脱敏、分区与供应商权限评估。网络隔离并非万能控制，现场也要管理可移动介质、调试笔记本、备份与模型文件。

一套可执行的治理方案应在采集之前标明每类数据的业务目的和责任人；对相机视野进行遮挡或裁切以减少无关人员信息；在可能时使用产品标识符替代人员身份；设置基于岗位的访问权限和留痕；定义原始图像与标注的留存、删除、纠错和安全事件响应；对供应商远程访问实行审批、短时授权和会话审计；对模型训练集、权重、配置文件和推理日志一并纳入版本及备份。若将数据传出境、转交第三方或用于劳动绩效评估，风险和法律依据均可能发生变化，应由企业合规与法律专业人员结合具体事实审查。以上是风险治理的研究性建议，不等于对某种部署模式的法律意见。

5. 本篇证据与分析的边界

现有可读材料支持的证据范围并不相同：LOCO AD 及 AD 2 的相关原段涉及受控视觉任务，磨削研究仅以摘要支持其对象和约束说明，不能核全部成本对照；等离子喷涂及资产巡检原文未读，不计入已核案例。本篇据此区分图像方法比较与实体参数干预，但不声称所有相关论文均已验证，或已建立跨产线收益。现场质量体系仍须独立复核结果。

因此，后续章节的分析单位将由技术模块转向实施系统：如何构建按工件与批次关联的数据链，如何按业务基率和复核能力制定阈值，如何在按时间、设备与批次隔离的验证集上测量失效，如何以有约束的试验评估参数变化，以及如何在保持可追溯和安全回退的前提下扩大使用。此处提出的机制链与验证方案尚不是来自本次分析者的企业实测；它们是由原始研究的任务设定、方法限制和现行法条推导出的可检验命题。若真实部署数据不支持这些命题，结论应随证据修正。

6. 测量系统与真实标签：模型学习到的究竟是什么

6.1 把产品状态、检验操作与参考标准分开

真实质量缺陷不是直接存放在数据库里的标签，而是依据产品规格、测试程序和人员判定，通过测量过程被识别出来的状态。视觉标签尤为依赖操作定义：一道划痕多长、多深、位于何处才算不合格；颜色偏差是否影响功能；一处缺损处在临界等级时是否放行。若规范允许多个等级，二元“好/坏”标签会丢掉边界信

SVSAC | sv-sac.org

息；若各班组口头标准不同，训练集中的相同外观也可能被标注为不同类别。模型准确拟合这些标签，最多证明它在给定定义与标注流程下复制了某种判定，并不自动证明该判定对应真实使用风险。

NIST/SEMATECH 的测量过程表征章节把重复性、再现性、稳定性、偏倚和校准纳入测量系统分析，并强调测量不确定度应对应实际工作环境[10]。连续尺寸的量具研究可以通过重复测量分解设备、操作员与零件的变异；视觉检验的输出往往是属性类别，更需要关注同一检验者对同一零件能否重判、不同检验者之间能否一致、与独立参考标准是否一致，以及这些一致性在班次、疲劳、照明或缺陷严重度变化时是否稳定。只报告全体准确率会将“判得一致”与“判得正确”混在一起：检验者可能彼此一致但共同偏离工程标准，也可能平均准确率尚可，却在关键缺陷类别中系统漏判。

因此，建立训练或验证标签前，应定义参考标准的来源与裁决程序。可由多名经过训练的检验员独立查看匿名样件，再由预先指定的质量/工程专家根据量具、功能试验或拆解结果裁决分歧；所有标签要保留原始判定、裁决依据、产品状态和版本。若所谓“金标准”本身无法由更精密的测量确认，就应保留有争议/不确定标签，或把结果按参考方法的测量误差报告，而不是把多数票称为客观真值。判定标准修订后也需版本化，以免前后训练样本的“缺陷”定义发生不可见变化。

Simion 及 Marques 等文献保留为人员判定与属性测量研究的待核线索。[11,12] 本篇未取得原文，撤回具体企业、产品、培训安排及改善效果的断言。本篇的测量建议是，在人工判断参与标签或复核时，用适当的参考样件、重复判定和盲化安排评估一致性；若培训发生，须在可比样本与独立参考下检验效果。这一建议不是上述案例效果的转述，也不证明模型降低了工厂缺陷率。

6.2 人员差异、边界件与选择性抽检

人员意见不同不必都归结为“培训不足”。差异可能来自缺陷等级边界不明确、观察角度不同、样件表面状态改变、检验时间压力，或参考标准无法分辨细微差别。可使用盲态重复判定估计同一人重复性，使用多检验员交叉判定估计再现性，并按缺陷类别、严重程度及临界状态分层报告一致率、混淆矩阵和不确定范围。Kappa 等一致性指标可以校正偶然一致，但会受样本患病率/类别构成影响，不应脱离原始一致率和类别数量单独解读。特别是“几乎全是合格件”的样本，表面上很高的整体一致率可能掩盖少数缺陷样件上的意见分裂。

SVSAC | sv-sac.org

航空发动机叶片检验文献保留为待核研究线索。[13] 未取得原文，本篇不使用“50 人、26 张照片、两次判读、2,600 次观察”等数值，也不确认其随机顺序、技能分层或结果。可独立提出的核心问题是：同一图像的重复判断、不同检验员的判断及与参考判定的一致程度是否相同？这种测量设计须另行落实；照片判读不能自动代表实体检验，更不能据此证明 AI 替代人工更安全。

离线样品的来源和抽取规则同样决定模型所学分布。若训练标签主要来自客户退回件、报废件或人工已挑出的异常样本，模型见到的缺陷更显著、更容易定位；现场真正难以检出的边界缺陷反而可能极少。若抽检制度只在设备报警后加密抽样，确认数据会集中在已报警区段；“未报警产品无缺陷”的记录则只是没有检查，并非独立阴性。以这类数据计算召回率会有验证偏倚。应记录抽样概率、来源、选取条件与参照检测强度，对非随机抽取样本使用适当权重或单独界定估计对象；当无法对未报警样品执行更严格检测时，应明确说漏检率尚未被识别。

追溯单位必须匹配质量事件的生成与处置尺度。单张图片适合图像级检出，但同一件工件常有多视角、多帧、重复过站；同一批次又共享原料、机台与配方状态。若目标是整件拒收，评价单位应按序列号汇总全部视角和决策规则；若目标是防止整批逃逸，需要再按批次汇总并考虑抽检；若要定位过程原因，还需连到工序、设备、时间、班次、刀具寿命和物料批次。关联键缺失或返工后覆盖原始记录，会使报警与最终质量无法连接，也使模型影响质量的路径难以追踪。

7. 视觉质检链与拒识复核：从成像条件到批次决策

7.1 相机不是中性的输入端

工业视觉检测起于光与表面相互作用，而非模型选择。表面反射率、粗糙度、曲率、透明度、污染物、振动和曝光共同决定缺陷是否呈现在图像中。镜头分辨率与景深决定微小缺陷是否可见；光源的角度、颜色、偏振和频闪同步可能让同一划伤从明显变为不可见；工件定位、夹具磨损和输送速度会改变几何投影与模糊。若照明设计没有提供足够的信噪比，换更大模型不会恢复相机从未记录的纹理。若光学条件随维护、批次表面或温度变动，模型输出漂移可能首先是测量系统漂移，而不代表工艺质量改变。

因此视觉系统的验证不能只拿一份预先裁切、中心对齐的图像集。应把相机、镜头、光源、曝光、焦距、视野、触发与工件定位作为检验程序的一部分，保存配置版本并确认安装后复现。对每类缺陷绘出视野覆盖图，说明正反面、遮挡面、边缘、孔洞和装配后被遮挡区域的可见性；只有成像可见的缺陷才可作为

SVSAC | sv-sac.org

当前视觉模型的目标。使用多视角时，需要确保完整产品判定采用固定的图像聚合逻辑，而不是将各视角当成独立产品样本。MVTec LOCO AD 论文所述机械对齐和任务专用照明/镜头，恰好说明公开任务能够通过精心配置成像来解决某一识别问题；真实设备的漂移与换型还要另行评估[2]。

成像设计还有一种反直觉权衡：提高图像分辨率有利于看清微小缺陷，却增加传输、存储和推理负荷；降低曝光可冻结高速移动物体，但可能损失暗部细节；更强补光可以增加对比，也可能让反光材料过曝。工程选择应以缺陷和误判成本为条件，不存在脱离产品属性的“最佳相机规格”。对比采集方案时，需同时量化可见缺陷覆盖、样品重复性、每件采集时间、在线可用率及成像失败重拍率。图像处理前的去噪、归一化或裁切也需锁定版本，因其可能滤掉关键纹理或造成边界缺陷消失。

7.2 四种粒度下的错误含义

视觉任务输出通常从局部像素分数开始，但质量动作发生在工件或批次层级。像素定位关注“哪些位置不正常”；图像级分类关注“一张图是否异常”；工件级判定把多个图像、视角或局部模型结果汇总为接收、拒收、复查；批次级决策则综合抽样、工序稳定性与放行风险。每升一级，错误定义都改变。某像素预测为缺陷但产品整体合格，未必构成错拒；某个视角漏掉一个小缺陷但其他视角发现，工件级仍可正确拒收；若一批中只有一件关键缺陷，抽检策略未覆盖该件则批次级风险依然存在。论文报告的一个层级指标不能代替下游业务层级的失误概率。

低基率进一步改变告警队列的组成。设缺陷基率为 (p)、灵敏度为 (r)、特异度为 (s)，若所有产品一视同仁检测，预期误报数量为 ((1-p)(1-s)N)，真阳性为 (prN)。即使保持相同灵敏度和特异度，缺陷率下降也会降低阳性预测值、扩大需要复核的正常件数。Saito 与 Rehmsmeier 基于模拟和不平衡分类示例指出，在负例远多于正例时，ROC 曲线可能不能直观呈现阳性预测的可靠性；精确率-召回率曲线能显示阳性预测中真阳性所占比例，并随类别比例而变化[14]。该论文实验背景主要是生物信息与分类方法，不是工厂质量案例；这里引用它支持指标解释，不把其样例数值当成工业性能。生产验证应同时报告基率、每千件告警量、阈值下的漏检与误报、精确率/召回及每种缺陷严重度结果。

阈值不是“模型自动学会的质量标准”，而是把预测分数转成行动的业务规则。以下约定异常分数越高表示越可疑，分数达到拦截阈值即转入拦截或复核。对于相同样本和固定模型，收紧放行政策对应降低该拦截阈值：拦截集合扩大或

SVSAC | sv-sac.org

不变，并可能增加误报；提高拦截阈值则会缩小或保持拦截集合，也可能漏过更多缺陷。若模型以正常分数输出，阈值方向必须相应转换，不能用“更严”代替操作定义。真实关系还取决于分数校准、产品变体、缺陷类别和成像变化，所以不能用一个全局阈值隐去类别差异。决策者应根据预设的漏放行上限、复核容量、误拒代价和处置时间确定自动放行、自动拦截与拒识区间。对超过模型适用范围或输入质量不合格的样本，应先拒识并复拍/人工检查，而不是输出高置信度的“合格”。

7.3 复核负担与现场拒识规则

人工复核是一项受限产能。设产量 q 为件/小时，进入人工告警队列的比例为 a，平均每件复核耗时 t 为秒，则理论需求为 qat/3600 人时/生产小时；有 K 名同等可用复核员时，平均负荷比为 qat/(3600K)。这是忽略搬运、记录、切换和工程等待的简化核算。平均负荷小于 1 也不保证等待可接受，因为告警可集中到达、耗时有差异、人员会休息。应按实际峰值和处理时长验证队列，并预定超载时的停线、降速、加人或恢复已验证流程；不能通过静默放行消除积压。

拒识应当与错误分类分开统计。模型可能因新产品、相机失焦、图像模糊、遮挡或分布外纹理而无法作可靠判断；拒识使这种不确定性显式化，并将样本送往复拍或人工测量。若把拒识样本从性能分母剔除，表面准确率会被人为抬高，因此应同时报告拒识率、复查后的最终判定、复检耗时以及拒识产品最终去向。对所有拒识都再次拍摄也会产生新的图像并可能改变缺陷可见性，需要保存首次和复拍图像及对应结论，避免只留一张“最好看”的图像。

现场复核规则还要避免“报警导致标签更精确、未报警却没人看”的选择性确认。部署期间可按已知抽样概率从报警与非报警总体中抽取样件，由不知道模型输出的参考检验流程复判；样本抽样和复核计划应覆盖班次、产品、设备状态和缺陷严重程度。对成本极高或破坏性检测，可通过随机抽取子样本、分层设计及逆概率权重估计总体表现，报告区间和无法识别的子群；如果法规或客户要求规定全检，AI 只能作为辅助，不得用实验性模型取消原有检查。

8. 工艺时序、漂移与因果边界

8.1 让每条记录具有正确的时间含义

工艺模型的第一项质量检查不是换算法，而是检查时间戳。PLC、相机、MES、实验室检验和人工复核可能使用不同的时钟、采样周期和事件写入规则。传感器信号记录的是采集时刻，质量检测记录可能是出站时刻，实验室结果又可

SVSAC | sv-sac.org

能晚几个小时录入。若不统一时钟和单位、确认缓冲与批处理延迟，模型会把“之后才知道的结果”错配到更早的工件，产生看似提前的预测；也可能把真正的预警关联到下一件产品。每个特征应标清它在预测决策时是否已经可用，不能只记录数据库行的入库时间。

数据集要按目标外推范围切分。若目标是预测未来一周同一设备，训练和测试必须按连续时间截断，且测试在训练期之后；若目标是换设备，测试需留出整台设备；若目标是换供应商批次，应留出完整材料批次。随机把滑动窗口打散，会让相邻窗口共享绝大多数原始采样值；随机划分工件，则会让同一批次、同一刀具和相同配方纹理同时进入训练与测试。此时测试衡量的可能是相似窗口识别，而非未来状态预测。训练、调参、阈值选择和最终检验还要按先后顺序隔离，不能用测试期结果反复调整后仍称为独立验证。

标签时间的定义也要与业务问题一致。可按产品开始加工时刻、关键工序完成时刻或最终检测时刻定义预测起点，预先指定预测窗口和报警提前量；把全部连续告警折叠为一次事件，避免一个漂移事件被重复计成数百个命中；同时记录告警后是否停机、换料、人工调参或返工，以免把纠正后的未发生缺陷简单记为误报。若维修后报警消失，必须保留维修事件和处理时间，才能区分“模型错误”与“干预成功而截断结果”的两种解释。

8.2 把状态变化和干预放在同一条因果路径上

同一组参数与质量的观察关系，可能由多种共同变化产生。原料批次变化可能同时改变黏度和温度读数；刀具磨损既改变振动，也提高毛刺率；维护会重置传感器标定并更换零件；换班可能同时改变操作速度、抽检力度和配方。若模型观察到某传感器值与缺陷共变，不能据此判断该值是成因、代理指标还是后果。传感器本身也可能在工艺异常后才触发，使用它预测异常相当于把事后证据当作先行信号。

反馈控制会造成更隐蔽的选择偏倚。现行控制器可能在温度升高时自动降低速度，结果数据库只留下“高温且降速后仍合格”的记录；模型学到的可能是被保护后的状态。若按历史数据给出更激进的建议，却没有记录控制器是否动作和动作前后的值，模型所估计的关联就不代表取消或改变控制后的结果。类似地，操作员只会对报警批次加密检验；因此观察到某报警类别缺陷更多，可能部分是更多检查发现的结果。先记录检查策略与处理决策，再估计质量变化，才能避免把发现机会增加误认为真实缺陷增加。

SVSAC | sv-sac.org

漂移既可以指输入分布变了，也可以指输入与质量结果的关系变了，或标注规则发生了变化。灯源老化、相机重新对焦会改变视觉输入；供应商材料更换可能改变传感器与缺陷的关系；质量规范修订则会改变阳性标签比例。单一总体漂移分数无法告诉工程师该做什么。监测应按可解释分层变量报告图像质量、关键传感器分布、产品型号、批次、设备、标签延迟、缺陷类别和人审分歧；超出校准范围时，暂停自动处置并触发复测/工程核查。若报警触发调参，后续模型也必须将干预记录作为状态变量，而不是把经过调节的正常结果归为随机噪声。

预测性能与过程改进应各自估计。前者可报告事前固定窗口内的缺陷召回、每小时告警暴露、平均提前量和不确定性；后者需比较相同安全边界内应用某项处置策略后，最终产品结果是否改变。前者即使很高，如果没有可及时执行的动作，也不能直接降低缺陷；后者若同时伴随换料、维护、人工加严抽检，则需对共同变化作对照或清楚限定为观察性关联。若关键标签的延迟、样本漏配或未报警产品没有参考检验，应该报告不可估计，而不是以算法置信分数填补缺失的质量事实。

9. 公开研究与案例比较：不同证据各自能回答什么

9.1 图像基准与运行现场并非同一层级

LOCO AD 的已读成像设置和任务说明可支持受控视觉比较这一有限层级；[2]MVTec AD 与 Real-IAD 的规模在本次分析中仅见 AD 2 相关工作的二手转述。[4] 两篇原论文未取得，不把它们的具体训练、测试、标注或聚合协议计为已核。[1,3] InsPLAD 原文也未读，[9]不能用未经核实的场景差异排列数据集难度。我们主张按任务单位、采集条件及动作后果设计现场评价；基准图像结果不能单独估计产品逃逸率。

若数据来自运行资产而非固定工位产品，成像机会、视角及发现后动作可能不同；本篇将此作为待核的任务边界，而非 InsPLAD 已证实的特征。对于多图像产品，须定义缺失视图、重复拍摄和产品级汇总规则。仅有样本级分数并未计算复核人时、客户风险或返工成本，这些业务量须另外观测。

人员判定研究的三篇原文尚未取得，[11,12,13]因此不能把培训改善、叶片重复判读或特定企业流程相加为已核测量证据。本篇保留的机制问题是：标签由谁、以什么参照、在什么时点生成，复核是否消耗人时并存在误差？这些问题值得通过重复性、检验员间一致性及参考判定检验，但本篇不声称已经测得其改善幅度，也不外推通用人工错误比例。

SVSAC | sv-sac.org

9.2 实物参数试验检验的是可行优化，不是模型自动掌控

磨削研究的可读摘要支持其磨削对象及粗糙度、烧伤和温度约束说明，[5]但完整试验与成本对照未核；等离子喷涂研究原文未读。[6] 本篇由此提出条件性设计：工程变量可设定、质量响应可量测且可行域可表达时，才有条件比较序贯搜索与基线的试验成本。它不是已经证实的通用降本效果，更不允许凭历史数据直接自动改机；过程、测量、安全边界和初始数据都须先定义。

把视觉与过程优化放入同一质量链，至少可区分四种证据问题：第一，图像能否捕获并区分预先定义的缺陷；第二，模型是否跨时间、产品和设备保持判定；第三，过程信号能否在可行动的提前量内预测质量；第四，操作员或受约束控制策略采取行动后，独立的最终质量是否改变。第一项可以用受控基准筛选，第二项需要外部或时间隔离验证，第三项需要严格时间对齐和事件级评估，第四项则要求生产试验或有充分识别假设的观测设计。将这些层次合并成一个“AI 准确率”会抹去技术可能起效但组织链条未接通、或者基准分数提高但产品质量不变等重要结果。

公开项目的选择与披露范围须单独审查。本篇未建立成功、失败及中止项目的完整登记，不能估计发布偏倚方向或幅度，也未核验企业原始运行日志。上述人员判定案例原文不可读，不应称为已直接核验的企业级数据。供应商若只列成功指标而无基线、分母、产品组合、停机与验证费用，不能支持总体成效；算法基准、摘要层面的设备研究或数值算例均不能填补量产部署因果证据的缺口。

因此，跨研究的实质比较应落在机制：基准论文控制采集和类别以建立可比性；运行资产论文扩大环境变化以测试视觉稳健性；测量系统研究将人的定义与重复判断纳入标签质量；参数优化试验通过约束域和实体反馈探索有限变量；现场质量评价则需把模型输出、检验处置与最终产品特性关联。每种研究都能补足上一种的盲区，却不能代替其余环节。一个生产线项目可以据此形成阶段性的证据矩阵，并允许“不提升召回但显著降低复核时间”或“预测能力提升而质量结果不变”这样的混合结果被如实报告。

证据类型	主要观察单位	可以支持的判断	不能由此推出的判断
受控图像基准（MVTec AD、LOCO AD）	图像、像素、类别	方法在给定视角、标注与拆分下能否识别/定位	现场基率、漏放行、节拍与成本变化
运行资产图像（InsPLAD，原文	部件图像、资产类别	待核任务边界；本文不确认其场	制造工件在线全检的收益

SVSAC | sv-sac.org

待核）		景或稳健性结果	
拟议属性测量/复判研究（案例待核）	检验员、样件、重复判断	标签判定、检验员重复性及培训是否值得评估	AI 模型替代人工或降低工厂缺陷率
受约束参数研究（磨削限摘要，其余待核）	参数设定、试件、过程状态	可提出可行域内搜索的试验问题，尚不能确认具体试验效果	未经验证跨工艺迁移或直接闭环量产
现场策略比较	工件/批次、干预期、独立终检	具体决策规则是否改变最终质量与全链成本	超出受测产品、设备与运行窗口的普遍效果

该矩阵也说明“可解释性”不仅是展示热力图或参数重要度。对于视觉检验，解释应能回到可重复的缺陷定义和检验视野；对于过程预测，要说明信号何时出现以及是否有动作窗口；对于参数优化，应展示可行域、约束违反概率和实际试件验证；对于生产收益，应说明比较组、抽样、版本和总成本。模型内部可视化即使与专家直觉一致，也不能替代上述外部证据。若热图突出背景上的夹具标记而非缺陷本身，反而提示了数据拆分和部署迁移风险。

10 工艺变量与约束优化：从可预测到可改变

10.1 先问“能动什么”，再问“预测什么”

过程质量问题常被写成“用传感器预测缺陷”，但预测对象与可干预对象并不相同。设一段制造过程在时点 (t) 记录设备设定 (u_t)、材料与环境状态 (x_t)、传感器观测 (z_t)，并在工序后得到质量响应 (y_t)。其中，(u_t) 是操作者或控制器可以设定的候选变量，(x_t) 可能只能观测、分层或通过上游安排改变，(z_t) 是对过程状态的带噪测量，(y_t) 则是产品或过程结果。一个预测器学习 (P(y_t\mid u_t,x_t,z_t))，不自动意味着把 (u_t) 改成模型推荐值便会得到相应的质量变化。若历史记录中操作员只在设备异常时提高压力，压力与不良之间的关联就同时包含干预反应和异常状态，不能直接读作压力的因果效应。

因此变量清单需要将三类对象分开。第一类是可操纵的设定，例如进给速度、温度设定或喷涂距离；第二类是背景或状态，例如材料批次、工具磨损、湿度、夹持刚度、设备维护状态；第三类是过程与产品响应，例如粗糙度、涂层孔隙、尺寸偏差、温度峰值和最终检验结果。某些量会跨类变化：一个传感器读数

SVSAC | sv-sac.org

可以是预测输入，也可能是控制反馈；它若受到控制动作影响，便不能同时被当成未受干预的外生解释变量。变量定义应同时记录测量位置、采样频率、校准状态、时间戳、过滤和聚合规则。

优化问题也不应先简化为单一“质量分数”。对设定向量 (u)，工程目标可能包括最小化缺陷、保持关键特性在规格内、减少周期或材料消耗，同时不能突破安全、设备、法规和客户要求。可行集合 Ω 由这些硬边界构成，只有 (u\in\Omega) 的候选才进入效用比较。把严重缺陷折算成少量成本点与把每件产品的节拍改善计入效用，并非无害的数学便利：权重会决定系统推荐何种权衡。对涉及安全或客户关键特性的约束，先以不可违反的边界表示；若约束本身存在测量误差，则需针对不确定性设置工程裕量，而非把模型预测的均值当成真实响应。

磨削摘要说明粗糙度、烧伤和温度约束须同时考虑；[5]这足以提出多响应权衡问题，不足以核实完整优化收益。喷涂研究的试验选择与状态处理原文仍待核。[6] 本篇的工程推论是，材料、热处理、夹具、工具或测量方法改变时，应重新检验可行域，不能默认旧最优参数适用；这一条件应通过本地测量而非未读案例验证。

10.2 代理模型只在其试验域中“代言”

当每次设定都需消耗材料、机时或破坏性测量时，代理模型或贝叶斯优化可以用既有试验和新试验估计响应面，并选择下一组信息价值较高的设定。其作用是改善试验顺序和覆盖，而不是免除实体检验。实验域边缘尤其值得谨慎：如果候选点远离已有观测，模型不确定度应当进入决策；如果最佳点恰好贴着安全或规格边界，则所谓“最优”可能只是预测误差与约束估计误差共同放大的产物。

在实际运行之前，需把变量单位、取值域、不可组合条件、变化速率限制和恢复条件固定下来。候选设定先经过工程审查，再决定是否进入试件试验；试件结果由独立于建模的检验流程测量。模型可以给出推荐值与区间、约束违反可能性、适用材料/设备范围以及触发拒绝推荐的条件。只有平均响应而没有范围与适用边界的输出，不足以支持操作决定。缺少关键协变量或处于训练数据覆盖以外时，系统应明确标注“无建议”，而不是用最近邻外推产生看似精确的数值。

仿真与实物反馈的关系也需区分。仿真可用于排除明显不可行区域、比较控制策略或缩小昂贵试验范围，但仿真误差来自模型结构、材料参数、边界条件和未建模扰动。若计算预测某参数设置安全，而实测显示温升、振动或产品失效超出预期，仿真不能成为继续执行的理由。验证可按层推进：先核对单位与边界，

SVSAC | sv-sac.org

再检查候选点在已知工艺窗口内的物理合理性，随后以代表性材料和设备做受控试件，最后才考虑限域生产。每层均保留停止条件、回退设定和异常记录。

还要防止把相关性排序误读成操纵建议。特征重要性描述模型在其数据分布中的依赖，不等同于变量改变后的结果差异。若材料批次与设定共同变化，而批次未充分记录，模型可能把材料差异归给设定；若工艺人员根据预测告警主动调整参数，后续观测到的关联还受到处置策略影响。要支持参数干预，需要有物理试验、可信自然变异或明确因果识别设计，并在安全边界内评估。算法上最优与过程上可执行之间还隔着工程许可、切换时间、校准与返工成本。

11 从建议到执行：控制权限应逐级获得

11.1 四个运行阶段并不等于四次上线

对一个能够生成判定或参数建议的系统，较稳妥的运行边界可以分为影子计算、人工复核、受控建议执行和受限闭环四种状态。影子状态中，系统读取生产数据并保存输出，但不向操作员显示，也不改变检验或工艺动作；目的在于确认时钟、接口、输入缺失和结果追溯。人工复核状态可将模型提示呈现给授权人员，但决定仍由既有检验规程与工程权限作出，且需记录接受、拒绝及理由。受控建议状态只允许经核准的变量、范围、产品和工况，执行前由责任人确认；任何越界或输入异常自动撤销建议。受限闭环则由控制器在预先审批的窄范围内调整，必须具备独立联锁、回退和人工停止路径。

阶段差异不应仅以“模型准确率达标”衡量。影子阶段检查系统是否能按时生成与追溯输出；复核阶段还需测量人员队列、争议率、复核用时与错误升级路径；建议执行阶段需要观察参数变化能否被机台记录、是否存在未登记人工覆盖以及最终质量是否独立改善；闭环则额外需要故障安全、通信中断、漂移监测、极限保护和停机权限。各阶段都存在系统层失败可能：模型正确但接口传错工件ID，人员看见告警却无法及时复核，或控制逻辑把单位换算错误的变量写入设备。单体模型验证不会覆盖这些链路。

需要为每一建议设置来源与责任链：产生建议的模型、版本和输入数据；对应的工件、批次、工序、设备、时间；建议数值与允许范围；批准人、执行人、是否采用；回退/覆盖原因；随后由哪种独立检验确认结果。若建议进入工艺控制记录，应能与材料批次、设备程序、保养和最终质量追溯。这样的记录既支持审计，也能区分模型建议无效、模型未被采用、操作执行偏离和测量未捕获等不同解释。

SVSAC | sv-sac.org

11.2 拒识与回退不是模型的“附加功能”

模型适用域应由产品族、设备、成像/传感器配置、材料和时间范围共同定义。输入缺失、传感器漂移、未知产品、设备状态超界、运行压力导致超时、输入分布突然变化时，系统须回到经过批准的人工流程或既有控制逻辑。拒识条件要能被操作员理解并由系统日志证明触发，而不是在事后将困难样本从准确率分母中删除。对检测系统，超出置信范围的工件可转人工复判或隔离；对参数建议系统，超域即不输出可直接执行的值。

人工复核也需要产能设计。告警数量若在短时间超过复核人员处理能力，排队延迟会改变产品风险：质量隔离可能阻塞生产，紧急放行可能绕过检查。此时简单提高自动放行比例并非中性的“效率优化”。应基于实际值班安排和高峰到达率，预设队列阈值、溢出路线和停用规则，按缺陷风险区分优先级，但不能借风险评分自动降低强制检验要求。质量、生产和设备责任人需共同批准复核流程，并明确最终放行权。

每次模型更新都可能改变被拒识、送审或建议调整的产品集合。不能只比较新旧模型总体分数；更新前后应保留一段可比的旧版本输出或冻结样本集，在相同产品/缺陷/班次上分析差异，并在生产现场进行有限时段的影子验证。若新版本的平均表现改善但关键缺陷漏检增加，质量结果不能由总体平均掩盖。更新后出现非预期报警或越界动作时，应可回滚到上一已核准版本，并复核受影响产品范围。

最后，闭环权限由组织的安全工程、质量体系与设备控制架构共同决定，而不是由模型团队单方面授予。安全关键工序、法律/客户强制检验以及不可逆工艺，通常要求更严格的独立验证与审批。这里不提供未经现场工程评审的参数配方，也不意味着任何 AI 系统具有操作机台权限。研究评估应将“建议被采纳率”“机台实际变化”和“最终质量结果”分开记录，避免将算法输出本身当作已发生的生产改进。

12 生产试验：如何识别质量改善

12.1 先确定处理、结果和可识别的比较

现场评估不能从“AI 方案组缺陷更少”这句话开始。需要先定义处理是什么：是启用视觉告警、让人审查看模型建议、采用参数推荐，还是让控制器在限域内调整？同一个项目可能有多个环节，若其在试验期一并改变，结果只能归于整套策略，不能归因于某个模型。处理还应包括人员培训、班次覆盖、更新频率、拒

SVSAC | sv-sac.org

识与回退规则。否则“有 AI”组与“无 AI”组只在名称上不同，实际接触和执行不明。

主要结果应与质量目标相连，并由独立检验或不受模型告警直接影响的测量程序得到。可预先指定每千件最终确认缺陷、关键特性超规格比例、逃逸到下游/客户的风险指标等；同时记录漏检、误报、复判、返工、停机与节拍。检测系统输出本身不能作为唯一结果，因为同一模型既定义处理又产生结果标签，会形成循环评价。若最终质量只有抽检获得，抽检概率和规则必须在比较组之间可比；否则缺陷数变化可能来自检测强度，而非工艺改善。

某些产品要求和安全检查永远不应为了试验随机取消。随机化只能发生在真实存在不确定性、且两个方案都满足全部既定要求的选择之间。例如，可在相似设备单元之间比较“沿用现行检查”与“现行检查之外增加经验证的辅助告警”，但不得把客户强制的终检从任一组移除。参数实验应限制在工程批准窗口内，并按设备、材料批次或时间区组平衡，以免方案恰好分配到状态更好的机台。若风险评估不允许随机分配，则研究应使用匹配同期对照、断点、分阶段采用等设计，并明确需要的假设和残余混杂。

12.2 生产单元、干扰与实际采用

随机或准随机的分配单位取决于可能的干扰范围。若同一班组操作多条线，培训和告警习惯可能跨线扩散；若配方调整影响整批材料，单件分派可能污染同批对照；若复核人员同时看到处理组与对照组告警，判断标准会改变。此时应以线、班次、设备或完整批次作为分配单位，而不能把同一设备连续产出的每件产品当作独立样本。观察件数很大也无法弥补有效集群数量太少带来的不确定性。

生产环境还存在时间趋势和切换成本。不同星期的订单组合、刀具寿命、操作员、维护窗口、温湿度和原料批次可能与处理同时变化。可采用同期并行对照来减少共同时间冲击；若只能分阶段导入，需保留尚未导入或延后导入的单元，并检查处理前趋势及变更窗口。阶段转换时的培训、设备清洁、校准和首件确认会产生过渡期，需提前定义排除或单独报告，而不能看完结果后再挑选“稳定阶段”。

分析时应区分分配效果与采用效果。主分析可按原分配计算，即使部分操作员未采用建议、系统发生拒识或设备故障，仍保留在策略组内。这估计的是给出这一运行策略的总体影响。另可描述实际采用后的结果，但采用与否往往由工件难度、人员判断和设备状态共同决定，直接比较“采用的建议”和“未采用的建

SVSAC | sv-sac.org

议”会产生选择偏倚。任何按方案执行的次级估计都需要清楚说明识别条件，不能以其较大的效果取代分配分析。

在分析前固定结果、时间窗、样本剔除、分层和停止规则，并保存方案版本。生产团队不能为追求显著性而事后只报告某个缺陷类别、某条线或某个班次。应同时报告估计差异及不确定区间、分配与实际采用比例、数据缺失、设备异常、复核工作量和安全事件。若处理组因为告警更多而加大终检，观察到缺陷率增加可能是发现率增加；此时需要将检测强度与产品真实质量区分，并说明可由哪些独立测量识别。

12.3 影子、受控执行与可反驳判据

影子模式适合核查接口、运行时延和输出分布，却不能估计真实质量处理效果，因为人没有根据输出行动。人审阶段能够观察复核流程与决策影响，但若复核员知道处理组身份，可能更认真检查处理组产品。尽可能使用盲态终检，或者由独立检验团队按同一抽样比例判定。受控执行阶段可以使用并行生产单元，在安全许可范围内分配策略，记录现场偏离和污染。对过程优化，优先用工程批准的试件与小批量验证推荐方向，再进入受控生产；同时监测未被优化目标覆盖的关键质量特性。

可反驳的结论需要预先说明什么结果会否定预期机制。例如，若视觉辅助旨在减少漏检，则预期是独立终检中漏检事件下降，且复核工时没有升至不可接受水平；若告警仅减少人工检查量但最终缺陷不变，需要判断原检查是否过度、还是模型没有改变风险；若参数推荐降低一个目标响应却令其他关键特性恶化，复合目标未达成；若效果只在某一设备或材料批次出现，则应收窄适用范围。即便估计值方向有利，置信区间仍跨越实质性无效区间，也只能称证据不充分。

负向结果同样有信息：在既有过程已高度稳定时，新预测可能无可行动的提前量；操作员可能通过经验已在告警前处理问题；或者缺陷产生于模型未观测的上游条件。只有把数据接入、触发、复核、干预、最终质量各环节分别记录，才能区分“算法失败”和“干预链未执行”。如果一个方案在真实条件下无法稳定接入或超出复核容量，这本身就是部署策略的结果，不能通过排除这些运行失败后宣称模型有效。

SVSAC | sv-sac.org

13 全成本：误检、漏检与工艺优化的经济边界

13.1 成本模型从事件数开始，而不是从准确率开始

设每期产品量 N、缺陷基率 p、灵敏度 r 及特异度 s，则真阳性、漏检、误报和真阴性期望依次为 Npr、Np(1−r)、N(1−p)(1−s)和 N(1−p)s。若真阳性和误报每件均复核，平均耗时分别为 t 真、t 误秒，小时人工单价 w，则告警复核劳动成本为[Npr·t 真+N(1−p)(1−s)·t 误]w/3600。只有两类耗时相同，才能用共同 t_r 简化；下一算例明确采用该假设。拒识、复拍和未告警抽检另计，既有真缺陷处置只计方案相对基线的变化。漏检后果则按加工、隔离、返工、客户与安全事件分别界定，避免用一个固定单价掩盖严重度。

下面仅作可复算的算术示例，参数均为假设，不代表工厂调查或性能承诺。设某条线每月检测 (N=20,000) 件，历史缺陷基率 (p=0.5\%)，经现场独立验证后假设模型灵敏度 (r=90\%)、特异度 (s=98\%)。那么月缺陷件数期望 100 件，真阳性 90 件、漏检 10 件；良品 19,900 件，误报 398 件、真阴性 19,502 件。若每个告警复核 30 秒，且所有真阳性与误报均由人查看，告警复核时间为 488 件×30 秒=4.07 小时。此数值不是“AI 节省了 4 小时”：仍要扣除现行人工检查时间，也须计入无法及时处置的排队风险、质检员培训、软件维护和后续终检。

项目	假设值	可复算方式	解释边界
月检量	20,000 件	给定	示例产量，非企业数据
缺陷率	0.5%	给定	现场基率须由独立抽样估计
灵敏度/特异度	90%/98%	给定	仅示例现场阈值性能，不是数据集指标
真阳性/漏检	90/10 件	(N p r), (N p(1-r))	需要确认样本/产品判定单位
误报/真阴性	398/19,502 件	(N(1-p)(1-s)), (N(1-p)s)	误报量由低基率放大
告警复核时间	4.07 小时/月	((90+398)×30/3600)	未扣除既有人工检查，也不含队列峰值

SVSAC | sv-sac.org

13.2 不同账本不可相加两次

完整成本至少有三个账本。现金支出包括相机/传感器、照明、计算硬件、软件许可、网络改造、集成、标注服务、验证测试、维护和备件；应按使用寿命和实际产能分摊资本费用，而非把采购额全部记为首月成本。工时账本包括样本复判、争议裁决、数据清洗、设备工程、质量验证、操作员训练、异常处理和模型更新；这些是劳动时间，若已通过工时单价折算成现金，就不要再把同一工时作为独立“机会成本”重复加总。产能机会成本仅在人员或机台被占用导致可证明的产出损失时单列，需说明其与工资成本不同的计算依据。

风险账本可包括漏检后的返工、报废、保修或客户损失等期望成本，但需要用历史事件频率、严重度和暴露量估计范围。若某项客户损失已经按预期成本计入，则不能再把同一事件的“风险降低价值”按另一比例重复计入收益。风险可能难以货币化，尤其涉及安全或声誉；此时应将其作为单独的约束指标或情景，而不是为了得到正 ROI 而强行填入数字。一次严重失效概率低但后果极大的场景，简单期望成本也可能不能反映监管或客户不可接受的边界。

工艺优化的收益与视觉质检不同。可计入的结果可能是单位合格品的材料消耗、返工、节拍或设备磨损改变；但一个参数建议导致的产量提升若同时增加末端检验和废品，不应仅按理论产能估算。实验开发费用、工艺工程师时间、试验材料、停机切换、破坏性测试、测量系统改造和量产后的持续监控均应纳入。若新方案只减少试验次数而没有改变最终单位成本，价值可能体现在开发周期或探索覆盖面，但应明确商业决策实际重视哪个结果。

13.3 从技术指标到决策阈值的敏感性分析

单点估算掩盖了基率与现场分布的不确定性。上例若缺陷基率从 0.5%降到 0.1%，其余参数不变，20,000 件中预计 20 件缺陷、18 件真阳性、2 件漏检，但误报期望 399.6 件（约 400 件）；此时每个真阳性对应的误报数由约 4.4 个升至约 22.2 个。这个变化意味着模型即使保持相同的灵敏度和特异度，复核队列与阳性预测值仍会显著改变。反过来，若缺陷严重性较高，企业可能选择更低的异常分数拦截阈值（即收紧放行政策），以较多复核换取更低漏检，但这需要实际人员容量和终检规则支撑。

因此应把灵敏度、特异度、缺陷率、复核时长、单位人工成本、误报处置比例、漏检后果和设备可用率分别作为不确定输入，计算悲观/基准/乐观情景。尤其应展示达到盈亏平衡所需的条件：例如要使方案现金回收，需每月至少避免多少已核实的返工成本或减少多少可计价检验工时；要满足服务水平，告警队列的

SVSAC | sv-sac.org

高分位等待时间需低于多少。条件阈值要从实际成本合同和工艺要求推导，不能凭通用模板给一个行业常数。没有可靠的成本或基率数据时，正确结论是暂不能估计商业回报，而不是将模型分数转换成节省金额。

项目结果应报告增量净成本与质量风险的并列变化，而非用单一 ROI 覆盖取舍。可呈现每千件漏检数、误报复核小时、停机与返工、最终质量、数据/模型维护时间、资本摊销和适用的风险区间。若一个视觉项目增加了复核成本却降低了高危逃逸，决策取决于质量规范和风险偏好；若参数优化减少试验次数但没有实质改善最终合格率，也可能值得用于缩短研发周期，却不能宣称车间品控已经改善。成本核算的目标是让假设可被业务数据替换，并使不利结果仍能留在评价表中。

13.4 让示例转换成现场核算表

要把上面的算术替换成现场数据，第一步不是收集模型输出，而是对齐实体与时间分母。产量需定义为完成哪个工序的唯一产品数，返工件是否重复计数；缺陷率采用同一抽样规则中的独立确认件，还是使用全量终检，必须说明。模型告警应按唯一产品聚合，多张图像、多个框或重复告警不能各自成为一个“误报事件”。复核时间可用抽样计时获得，包括打开记录、调阅多视角、判断、登记及隔离操作，不宜只统计屏幕上的决策秒数。若不同缺陷要由不同资历人员复核，还需分层计算队列和成本。

建议将成本表按现行基线、影子模式、人审模式、受控执行模式分别列示，显示哪些费用是新增、哪些既有流程仍须保留。每种模式记录运行小时、产品数、告警数、拒识数、抽检数、复核人时、设备异常和最终缺陷。影子模式期间算法可能生成输出，但因没有影响处置，不能把任何产品质量差异归于算法；它的成本是接口与运行验证成本。人审阶段新增的是展示、复判、培训和流程协调费用；若仍保持全量终检，不能把模型覆盖的人工检查全部算成节省。受控执行阶段需再记录采用建议、实际执行、退回与覆盖，更新验证期的工时不得隐匿在常规维护中。

可以用简单差分表示：

增量总成本 ΔC=C 方案−C 基线。两侧 C 均按同一观察窗与产量口径，由互不重复的非劳动资源成本、劳动成本、可证明且未被前两项计入的产能机会成本及期望不良后果组成。资本投入的现金流与分摊费用分表展示，比较时采用同一口径；已计入服务合同价的劳动、返工工资或客户赔付，不在其他项再加一次。

SVSAC | sv-sac.org

各期必须使用同一产量或通过每千件标准化，且对设备停机、产品组合和原料批次变化作解释。若方案改变了产量或缺陷暴露，直接比较“月成本”会把规模效应误作效率效应；应同时报告每件合格产品成本、每千件缺陷事件成本，以及固定投入摊销。设备投资摊销年限属于财务假设，需以企业会计和设备利用计划为准。模型服务费按月、按调用或按产线计价的商业合同也要使用真实条款而不是宣传页中的理想价格。

敏感性分析还可回答“什么数据最值得补”。如果盈亏平衡主要受缺陷率控制，继续优化模型 AUC 可能不如建立代表性终检抽样；如果复核时间决定人力成本，测量复核任务分布比增加训练图像更有决策价值；如果漏检后果支配结果，则需先获得下游失效和返工记录。将输入按对净成本的影响排序，并标注数据来源、可信区间、负责人及更新时间，可以把评估从一次性商业论证转为可更新账本。只有当关键输入足够可靠时，才应把情景区间压缩为投资回报点估计。

即便财务净值为正，也不能自动满足质量治理要求。低风险零件的误报成本可能超过漏检成本，安全关键部件则可能存在不可折价的检验约束；两者需依据产品风险等级和合同要求区分。商业决策在质量约束内比较可行方案，而不是由模型在任意成本函数下自行决定风险接受水平。对尚未验证的现场性能，成本表应保留“待测”项，不能用基准性能直接代填。这样，模型、流程和财务假设都能在后续数据到来时被逐项证伪或更新，而不会把不确定性隐藏在一个看似精确的收益数字里。

对照设计还应将估计对象说清。若按产线随机分配，主估计量是被分配到策略的整条线在规定时间窗内的平均结果差异，不是“每次模型建议的成功率”；若采用分期导入，估计对象则可能是导入单元在实施期的平均策略效应，仍依赖时间趋势和无同期差异冲击等条件。分组数有限时，标准误不能仅以产品件数计算；可按分配单位聚类或基于实际随机化方案做推断。安全事件可设置独立停止规则，但停止并不会自动证明处理有害或有效，需记录触发标准和事件审查过程。对于不同缺陷类别、设备状态或班次的异质性分析，优先作为预先规划的次要结果；若仅在结果出来后挑选有利子群，结论须标为探索性。预先说明估计对象，才能防止把大量产品记录造成的表面精确误当成多个独立生产环境的证据。

SVSAC | sv-sac.org

14 按严重度定义验收：分类标签不是业务动作

14.1 质量判断要回答“怎样处置”，不只回答“像不像缺陷”

现场验收不应把所有异常图像合并为一个二分类准确率。质量特性可能是可返修外观、尺寸偏差、装配漏件、潜在疲劳裂纹或会引发功能失效的过程异常；同一个检出分数在不同处置制度下具有不同后果。模型输出最终影响的动作包括：直接放行、隔离待复核、返工、报废、加严抽样、停线调查或启动客户通知。验收对象应是从判定到动作的规则，而不是孤立的分数。这里的“缺陷严重度”作为分析维度，不替代企业现行风险分类、客户图纸、强制要求或安全评估。

应把三种错误分开计数。错拒是把可接受产品送入隔离、返工或报废路径，造成额外检查、节拍损失和可能的二次损伤；逃逸是把不合格产品放行，使风险流向下游、客户或使用现场；错误返工是对可修复产品采用不适当工艺，反而引入尺寸、表面或材料损伤。它们的损害通常不对称，也不共享相同评价单位：错拒按件数、复判时长或返工成本统计；逃逸需按缺陷严重度和流出位置追踪；错误返工要识别返修前后状态，而不能将“返工过”本身等同于已修复。

一项可以检验的验收方案会先为产品族/缺陷类型定义严重度与允许动作，再把模型阈值映射成动作。举例说，对于容易修复、不会改变后续功能的涂层外观瑕疵，系统可以将低置信样件送人工复判并对错拒率和复核积压设限；对于装配逻辑错误，如紧固件缺失或线束路径错误，是否允许模型参与自动放行应依据独立功能和最终装配检验，不能只靠局部图像评分；对内部裂纹或热处理异常，若相机没有物理可见通道，则视觉模型不应进入该缺陷的检测覆盖声明。上述情境是分析构造，说明验收标准如何依风险、可见性和可逆性改变，不是企业案例或通用行业等级。

14.2 不可逆后果改变阈值的容错结构

阈值选择不仅要比较误报与漏报个数，还要看动作是否可逆。错拒后经过复核可以恢复放行，说明损失可能被复核流程部分缓解；误放行后若产品已封装、混批或交付，追溯成本和风险暴露会扩大；若再加工会损伤零件，自动返工的错误动作也可能难以挽回。于是，模型可以在相对可逆的节点用于筛选或提示，却不必获得最终处置权。对于不可逆、高严重度结果，可设置强制复核、既有终检不变或模型仅作并行提示的约束。

可检验的问题是：在同一缺陷类别、同一基率和同一独立终检下，分级阈值是否降低高严重度逃逸，同时把低严重度错拒控制在质检能力内？分析应报告每个严重度级别的产品数、缺陷数、阈值、弃权和复核决定；如果模型没有足够阳

SVSAC | sv-sac.org

性样本，应说明风险区间很宽，不能因总体准确率高便宣布关键风险已通过。阈值方向也需明示：分数定义为异常度时，较高分是否意味着更严格拦截；阈值提高或降低何者会增加拒识，应依具体分数映射定义，不套用不带方向的“调高阈值改善质量”。

质量验收和模型训练标签需要保持独立。若同一批标注员按照同一模型建议复查，并把意见直接写回标签，系统容易把其自身判定循环强化。验收样本宜按产品、班次、设备与缺陷风险分层抽取，未报警件也需有独立参照检验机会。对罕见严重缺陷，可有意提高其验证抽样率，但估算整线基率时要按抽样概率加权；不加权地把富集样本当成现场分布，会高估阳性率并改变阈值经济含义。这样可同时保留“关键缺陷的检出能力”和“整个工位的实际负担”两个不同的问题。

15 学习范式：训练数据条件决定“未知缺陷”含义

15.1 监督分类、正常样本异常检测与合成异常

有足够、可靠且覆盖目标产品/缺陷的标注样本时，监督分类器可学习已定义缺陷类别并直接对错判成本做训练或阈值调整。但其输出语义依赖标签本体：标注集没出现过的缺陷不必然映射成“未知”，模型可能把它分到外观相近的已知类或正常类。产品改型、照明变化、材料批次和缺陷定义变化也可能使历史标签不再对应当前验收规则。监督方法适合回答已知类别的再现问题，不自动覆盖新缺陷发现。

PatchCore 列为异常检测方法的待核文献，[15]原文未取得，因此不将其特征记忆库实现或基准 AUROC 作为已核事实。本篇保留一般分析：若算法以正常样本作为参照，样本对产品位置、正常变体和成像条件的覆盖便值得检验；新批次的正常纹理可能触发误报，缺陷也可能未被分开。偏离参照的分数并非天然等于缺陷严重度或不合格概率，应由独立标注与校准核对。

DRÆM 原文未取得，其网络构成、合成异常训练及泛化结果均保留待核，不作为已核方法比较。[16] 若项目拟采用合成异常，则应检验扰动与真实失效机制是否对应：仅改变颜色的扰动未必覆盖缺失、错装或尺寸关系错误。合成掩码可用于拟议试验，但不能免除真实样本验证，更不能在没有原文与现场证据时宣称已获得可迁移性能。

15.2 弱标签、拒识与校准不是同一问题

在现场，常见标签可能只到工件或批次级别，例如“此件已返工”而不知道哪张图、哪个像素或哪个工序产生问题。把工件级标签复制到每张图会把不确定定

SVSAC | sv-sac.org

位伪装成精确标注；多实例学习或弱监督可以利用不完整标签，但其性能仍受标签噪声、视角归属和产品追溯质量限制。另一种来源是自动光学检验的物理量阈值，它能够给出可计算的测量结果，却可能产生大量误报并依赖静态规格限值。后续人工复判也可能有未量化分歧，不能把弱标签说成“完全无需人工”。

拒识和校准解决的是决策不确定性，而不是异常发现本身。分类概率需要验证其在目标产品与环境中的校准；异常分数通常不是概率，需用本地验证集构造能解释的阈值或分位参考。对超出适用域、相机失焦、类别不明或高严重度边界样件，可让系统不作自动判定并交由复核。但拒识会形成工作量，必须报告拒识覆盖率及其在缺陷/正常类别、班次和产品上的分布。只报告“已接受样本”的准确率会隐藏模型把困难件都送走的情况。

因此选择范式应由数据和处置问题共同决定：已知缺陷稳定、标注和产品追溯可靠，监督分类可能清楚对应类别；缺陷罕见但合格样本丰富，一类检测可作为新颖性筛查；工件级标签不精确，可探索弱监督并保留不确定性；新产品冷启动且异常类型未知，可使用正常样本模型，但需验证正常变体覆盖与罕见缺陷风险。无论选哪类，最终验收仍要独立抽样确认错拒、逃逸和复核量。算法范式不能替代缺陷定义，也不应替代既有强制质量门。

15.3 推理时间不等于完整工位节拍

EfficientAD 仍为待核方法线索。[17] 本篇未取得原文，撤回学生–教师及自编码器结构、约 2 毫秒和 600 幅/秒的事实断言，具体数值保留于待核记录，不作为选型或产能依据。现场评估必须量测包括曝光、传输、多视角汇总、告警、复核和分拣的端到端时延分布，并报告错过执行窗口的后果；即使将来核到模型推理时间，也不能把它等同完整节拍。

方法比较也应采用可复现的统一测试：相同产品分组、相同现场阈值选择规则、同一显卡/边缘设备、同一输入尺寸、同一图像聚合、同一拒识与告警输出。一次测试可分别记录模型推理、成像读入、预处理、产品级判定和执行指令确认时间。如果算法节省 2 毫秒但工位光照重拍平均耗时数秒，瓶颈不在网络；若通过降分辨率获得更快推理却看不到细微裂纹，则速度提升牺牲的是缺陷可见性。单纯模型排行榜不能替代端到端量测。

SVSAC | sv-sac.org

16 与 SPC 和测量系统并用：何时规则比复杂模型更好

16.1 稳定性、能力与异常分数回答不同问题

统计过程控制（SPC）从按时间顺序取得的过程测量出发，判断过程变异是否持续在已表征的范围内，以及是否出现需要调查的变化。过程稳定性不是“产品已合格”的同义词；过程能力则比较稳定过程的分布与规格界限。NIST/SEMATECH 工程统计手册明确把稳定性评估放在能力评估之前：如果均值、方差或分布随时间改变，用当前总体估计能力可能混合了不同状态[18]。即便过程稳定，也可能稳定地产生规格外产品；反之，一个短期能力数字看似良好，也可能掩盖近期漂移或不同批次的混合分布。

控制图所依据的测量系统必须能区分过程变化和量具噪声。若相机清洁、曝光、量具校准、工装更换或操作员判读会产生与目标缺陷同量级的变化，首先需要理解测量变异；模型在噪声标签上训练，可能只把测量程序的改变学成“质量变化”。同样，工艺控制图如果样本次序不正确或采样间隔受报警决定，序列结构可能失真。SPC 不能修复错误追溯或模糊标签，但它的时间序列视角能把“过程何时变了”与“某件产品被模型打了高分”区分开。

SPC 与 AI 并非简单替代关系。传统图表适用于定义明确、连续可测且采样与分布条件可核验的质量特性；其信号通常指向一个超出基线的变化，需要工程师调查原因。图像异常模型适用于空间形态复杂、缺陷变化难以手工列尽的视觉信息，但异常热图不等于根因，也不天然提供过程是否稳定的统计解释。一个低维、直接对应规格的尺寸值若可以由量具稳定测得，单值/移动极差图、设定限或预先定义的规则更易审计；用神经网络替代它，可能增加维护和验证负担而没有增加可行动信息。

16.2 组合模型的价值来自补足信息，不是叠加“智能”

自监督多变量 SPC 研究 SSMSPC 把时序信号表示学习放进传统的控制图视图：模型对时间序列的变换和位置任务学习表示，再在表示上使用 Hotelling (T^2) 统计量和控制界限；作者以两个真实 CNC 铣削数据集评价，并描述在一台运行中的铣削机部署监测[19]。这类研究说明 AI 表示可以与 SPC 统计告警结合，并使异常时段与工序元数据共同展示。它不是“AI 已取代 SPC”的证据：输出仍采用控制图式界限，异常提示仍需操作员检查工序原因，研究对象限于特定铣削过程和数据集。

Pittino 等文献原文未取得，设备场景、方法对比及“接近 1 的召回”等结果均不计为已核证据。[20] 本篇讨论的独立问题是：监测告警能否对应到过程阶段、

SVSAC | sv-sac.org

可调查原因和独立质量结果？即使某方法在已记录异常上表现良好，也不能自行证明告警之后的动作改善产品质量；须分别验证检测、处置和结果之间的关系。

“用简单规则还是 AI”可以转成可检验的比较：在相同时间窗、样本单位和预定误报预算下，既有控制图/规格限能否发现目标变化；AI 增加了哪些早期信号、哪些新型异常或更少的无效复核；新增计算、建模与更新成本是否抵消检测收益。结果变量不应只对比 AUROC。还要报告事件提前量、报警确认率、每班人工调查次数、遗漏事件严重度和测量系统变化。若简单图表同等及时、错误更少且维护人员能解释，复杂模型就没有获胜；若高维波形或图像提供了经独立验证的补充信号，模型才可能成为现有 SPC 的辅助层。

对于强自相关、换型或按产品族分布的过程，直接套用独立同分布假设构造阈值可能使告警频率失真。时间序列模型也不是万能解：若状态切换由已知换模、保养或材料批次触发，按工序分段、记录状态或使用已知工程边界，比让黑箱自动学习混合基线更可诊断。产品规格限回答客户接受边界，统计控制限回答相对历史变异是否异常，两者不应相互替换。高分异常分数也不能自动推导“过程失控”，除非分数的基线与变化规则经过相应验证。

17 多工位、多视角与过程信号：关联错误会伪造缺陷关系

17.1 时间戳相近不等于来自同一件产品

多相机或多工序链条常跨越传送带、缓存区、分支、返修和重新上线。一个部件可能在工位 A 被多角度拍摄，在工位 B 组装后再次成像，之后通过量测设备或抽检获得终检结果。若只用时间戳近邻关联影像与结果，节拍波动、缓存、停线或人工取件就可能把上一件的模型分数配给下一件产品。关联看似高比例成功仍可能在少数产品上错位，而恰恰是稀有缺陷样本对质量结论最重要。

每条记录最好有持久产品/批次标识、工位标识、检查事件号、相机配置和时间；没有唯一序列号时，需要评估时间戳、载具 ID、设备循环计数和产品类型组合的碰撞概率，并对低置信匹配显式标记，而不是默认为正确。多视角图像须定义产品级聚合：所有视角都过关才放行、任一视角拒识即复核，还是按缺陷部位决定视角权重？聚合逻辑会改变整件误拒与漏放行率，因此模型评估和生产规则必须一致。若某视角失焦后被系统静默丢弃，整件判定就不再有声明的覆盖范围。

可构造一个无实测数据的错位演示：传送带正常时每件产品触发两张图，某次短停后 PLC 计数器比相机缓存多一件；系统仍以时间排序拼接，造成一个缺陷

SVSAC | sv-sac.org

图像与前后两件中的错误终检标签对齐。模型可能在错误映射中学到“某班次/配方会导致缺陷”，实际关系只是时间同步偏差。调试时可用已知标记件和人工注入的停线、跳件、返工重进场景，核验关联是否自动失效；这些是测试协议建议，不是已发生的工厂事故。

17.2 工件、焊点、批次与过程窗是不同的关联层级

SMT AOI 数据文章保留为待核数据线索，[21]不使用未经核实的 132 天、板级流转或具体标签错误描述。为说明观察单位，本篇构造如下分析：焊点测量应在焊点层核真值；整板是否送复核应在板层核队列；返修拥堵则观察等待、返修与重检事件。这是拟议记录设计，不是对该数据集的实测描述。同板多个焊点不能作为相互独立的 PCB 样本计算业务负担。

从传感器时间序列匹配影像时，需要定义物理因果窗口：哪一段温度、振动或压力信号作用于当前零件，哪个变化属于上游加工、哪个属于后续工序；传感器采样和视觉触发时钟是否同步，时钟漂移与时延如何估计。仅把产品序列号与所有历史信号连接到数据库，还不足以证明这些信号发生于产品被加工时。若数据对齐后发现预测能力，仍要考虑混批和串行设备上的相邻件关联；检验分组时应把同批或相邻窗口留在同一侧，以防模型记住背景状态。

对拟议 AOI 流程，应分别保存自动测量、阈值触发、人工复判、返修和终检，而不把最后一列标签当成产品原始状态。SMT 文献原文未读，[21]本篇不声称其已记录这些环节或已证实特定漂移。训练若使用复判结果，应先评估参考判断可靠性及进入复核的选择机制，并保存版本、时间、复核依据与产品改动，以便区分状态变化和标签变化。

多站点关联链路的验收可围绕断裂点展开：产品 ID 重复或缺失时停止自动合并；设备时间戳偏差超过已校正范围时保留未关联状态；返工件重新上线时建立新检查事件同时连接原件；一个批次需要隔离时，系统须显示其起止范围及追溯依据；多相机中某一视角无图或延迟时，执行预设拒识/补拍动作。每种动作都产生可查询事件日志。目标不只是“数据管道不断”，而是避免关联错误成为貌似稳定的模型信号、标签或质量趋势。

分析构造的情境	质量特性与判定单位	错拒/逃逸的不对称	可接受的 AI 角色及可检验结果
反光外观件的微小划痕	缺陷长度/位置与整件外观等级	轻微划痕错拒可能增加复判和抛光；深划痕逃逸	将边界件送复判；比较盲态等级一致率、误拒

SVSAC | sv-sac.org

		可能影响外观或后续涂层	率与复核时长
连接器装配漏装/错位	连接器或子组件是否按装配关系就位	错拒增加拆检；漏放行可能导致电气功能不良或整机返工	图像提示并保留功能测试；按组件和独立电测追踪逃逸
不能由当前相机看见的内裂纹	材料/部件内部状态	视觉“正常”被误解为无缺陷会形成覆盖声明错误	模型不承担此缺陷检出；由适用的无损检验/工艺监控承担验收

该表中“可接受角色”取决于现行质量计划与风险授权，不是通用建议值。若外观等级在客户图纸中定义，验收时应按其边界抽取样件并用盲态评审；如果不同检验员对边界件的判定不一致，先报告参照标准的不确定性。对于功能类错误，可用独立电气/装配测试确认图像判定是否对应实际后果；如果该项独立测试本身也存在敏感度限制，不能把它当成绝对真值。项目可将缺陷严重度、动作可逆性、发现时点、下游传播范围作为分层变量，分别报告置信区间；不宜把几个稀有严重事件与大量轻微外观件平均后给一个准确率。

多视角和批次聚合也要写进阈值验收。比如系统对每张图都得分，不能在校准阶段用“有任一图像异常则拒整件”的规则，在上线后又改成对多个视角取平均却不重新验证。若换线或操作员补拍改变了视角数量，样本级分数分布也会改变。故需要冻结产品级聚合、弃权和重拍规则，并在缺陷分布、视角缺失及实际复核容量下重新估计错拒和逃逸，不以单图像 ROC 直接推断整件决策。

在生产关联验证中，至少保留一批人工确认的校验件贯穿相机到最终处置：已知 ID 件应在各工位按预期次序出现；在缓存、停线、返工重入和扫码失败情形下系统应识别错位或拒绝自动汇总；同一件多角度影像应归并到一个工件决策，同时保留每幅原始图和触发时间。若某条测试无法还原工件历史，则不把该记录用于离线模型训练或效应估计，而应先修复追溯或报告关联覆盖率。这样测的是追溯系统在故障情况下是否能保持质量归因，而非仅仅能否正常运行时顺利拼接。

SVSAC | sv-sac.org

18 现场验证协议：让结果能被复核、推翻和延迟观察

18.1 预注册的不是模型，而是比较问题

现场验证应在第一件产品进入比较前形成冻结方案，方案版本可在技术迭代中更新，但每次变更需保留日期和理由。预注册的核心是估计对象：在哪些产品、工位、设备和时间范围内，比较哪两种处置策略，关心何种最终质量结果。若处理是“启用视觉告警后仍保留现行终检”，结果对应的是整个辅助策略；若处理是“使用 AI 建议改变参数”，则包括建议展示、工程批准、实际执行和回退规则，不应只把算法给出的数值称为处理。

协议至少列出产品族、图纸/软件/配方版本、机台与相机配置、运行时段、纳入/排除条件、分配单位、对照策略、主要和次要结果、测量者、盲态安排、分析单位、样本连接键、停机/安全停止规则、缺失记录处理、版本更新与偏离处置。排除规则应可由生产日志事前判定，例如新产品试制或设备维修窗口；不得在看到不利结果后再排除某条线或某些班次。协议附有各方签字和质量授权：质量负责人核准验收、工程核准参数范围、安全责任方核准联锁和停止条件，模型团队不独自决定产品放行。

协议项	需冻结的定义	变更时的处理
目标总体	产品族、版本、工位、设备与调查时间窗	新产品/配方单列，不混入原推断范围
处理策略	输出、阈值、复核、拒识、执行、回退与人员培训	任何会改变动作的更新视为新策略或单独记录
对照条件	现行合规检验、统计规则或另一已批准策略	不因实验撤除强制检验或安全措施
主结果	独立最终质量、严重缺陷逃逸或其他预先指定结果	检测定义改变时并列报告旧/新口径
关联与样本	工件/批次/工序键、抽样概率、重复图像聚合	关联失败件标未匹配，不能默认为正常
停止规则	安全事件、设备边界、队列/节拍服务水平	触发后暂停干预，按规则审查而非直接删数据

SVSAC | sv-sac.org

18.2 抽样和盲态终检构造可信参照

若现行全检可独立确定产品状态，可保留所有组一致的终检；若只有抽检，则需记录抽样概率、取样时点及是否受模型告警影响。建议建立分层抽样框架：在正常告警、未告警、拒识、高严重度类别、产品批次和设备状态中保留可追溯概率，保证模型没有提示的产品也有被参照检查的机会。严重缺陷样本可以为能力评估而富集抽取，但分析现场缺陷率时按纳入概率加权，并将两个结果分别报告。

终检人员应尽可能不知道该产品来自策略组还是对照组，也不应看到模型热图、预测类别或复判意见。若因设备或工序限制无法盲态，协议需说明哪些信息可见，并使用独立复核者判断重要结局。对边界缺陷设置参考标准：例如尺寸类用校准量具和规定位置，装配类用独立功能/关系检查，外观类由两名经培训检验员盲评并依据预定裁决规则处理分歧。持续记录复测差异、无法判定比例和标准件漂移，避免把参照测量误差隐藏在模型准确率之下。

一次独立终检也可能在破坏性试验、客户使用或下游加工后才能观察结果。对延迟结果应指定追踪期限和产品去向，分别报告“出站终检”“下游发现”“客户退回/投诉”等结果。没到追踪期或失访的产品是尚未观察，而不是没有缺陷；如果策略组与对照组的追踪完整度不同，比较就会偏。可先报告随访覆盖率和缺失原因，再进行预定的敏感性分析；不得将未观察的延迟事件自动记为零。

18.3 分配单位和重复测量

方案按产品随机分配前，应评估不同产品之间的干扰。如果同一操作员根据系统建议改变整个班次的检验习惯，按单件分组会污染对照；若调整参数作用于一批材料或设备热状态，则分配单位可需提升至完整批次、工艺窗口、机器或班次。选择集群单位会减少有效独立单元数，估计不确定性应按实际分配层级计算，而不是把图像、焊点和工件数量当成独立分配次数。若在相同产品上反复采图，应在样本层级归并并建模重复观测，不让一件多视角产品贡献多个独立产品权重。

如风险不允许随机分配，可以采用并行符合要求的线、分阶段导入、匹配同期批次或时间切片，但事前说明假设：对照能代表同一时期无干预结果，产品结构与设备条件可比，没有只在导入线发生的同期工程改造，干预前结果趋势可观察。设计越接近事后对比，未测混杂越难排除；应把结论表述为关联并说明敏感性，不包装成随机试验。若采用阶梯式实施，各组在未启用期继续执行合规现行检验，组间切换时间和培训窗口也要进入模型或单独报告。

SVSAC | sv-sac.org

18.4 停止和报告的预先约定

安全、客户和工艺停止规则应先于效果分析。越界温度、联锁失效、危险缺陷逃逸或复核队列超过已批准处置能力时，可暂停 AI 动作并恢复原流程；停止事件需记录触发证据、受影响产品范围、处置和复启条件。试验结束后报告每组分配和实际采用数、抽样覆盖、拒识、错拒、逃逸、返工、复核时间、停机、缺失、模型版本和估计区间。主要结果不显著并不等同于没有差异；如果置信区间仍覆盖有实际意义的改善与伤害，结论是精度不足，应继续观测或限制结论范围。

协议还应预先说明多重结果和中间查看。若每日监控决定是否继续收样，提前定义停止界限和查看频率；否则反复查看并在碰巧有利时结束，会夸大效果。关键缺陷类别、主要缺陷位置或设备状态的子组分析尽量事先指定；其他观察标为探索性，等待后续验证。原始决策记录、无法匹配事件和策略偏离需保存，防止结果被整理成只有模型成功样本的回顾性叙事。

18.5 样本量和观察期从可识别的差异推导

试验样本量不应使用某个“AI 质检常用件数”。先从相同产品/工位的基线估计关键结果的发生率、班次/批次内相关性、测量误差、缺失率和可投入的独立分配单元，再定义质量团队关心的最小实际差异及错误方向的容许边界。对罕见逃逸，短期零事件不证明风险为零；如果方案期间没有观察到严重逃逸，要报告在实际分母下仍与数据相容的上界，而不是给出“零漏检”。对复核时长、图像处理延迟等连续结果，应按工件/批次重复结构估计方差，不能把同件产品的多张图当作独立信息来虚增精度。

观察期还应覆盖可能改变性能的周期：换班、原料补充、设备维护、换型、季节性环境或工具寿命阶段。若企业无法在一个完整变化周期内取得足够独立批次，协议需限制推论范围，并明确后续延长观察的触发条件。样本规划同时考虑参照检验资源：检验员能否以相同方式复查两组，严重缺陷富集后是否还能按概率还原整体基率，返工/下游结果是否能在预定观察窗内回收。数据量大不等于证据充分，关键是有多少独立的时间/设备/批次状态被覆盖，以及主要结果是否被可靠测量。

18.6 试运行操作次序

现场启动可以分为四个冻结点。第一，质量与工程共同签署合规基线、缺陷定义和参考测量方法。第二，记录设备、相机、阈值、系统接口与工件键的版本，在不改变现行处置的影子运行中验证关联和延迟。第三，开始预定分配或同

SVSAC | sv-sac.org

期比较，保持两组相同的强制检验、抽样和终检，盲态参照记录后再解盲分析。第四，按预定规则核查策略采用、偏离、缺失、事件严重度和延迟结果，先做主分析，再做明示的次要和探索性分析。每次冻结点需保存签署、输入版本和时间；若方案受安全暂停，则记录原分配和已发生处置，以便解释未完成样本。

18.7 试验结束后的审计与复现

验证数据冻结时，应保留未解盲的样本清单、抽样随机种子或规则、标签版本和原始输出文件；分析人员在完成主结果代码与口径后才取得分组标识。若无法对操作者盲态，可在结果计算前让独立数据管理员隐藏策略组标签。报告至少附一张样本流程图：符合条件数、因明确事前条件未纳入数、分配/实际接受数、无图/错配数、拒识数、失访数、参照检验数和最终分析数。每个差额都要有可复核原因。

复现时不仅需要模型权重，还需输入预处理、设备/相机版本、缺陷本体、阈值、聚合规则、随机分配表、参考测量程序和分析脚本版本。若业务上不能共享原始数据，可保存受控访问索引和执行环境，以便审计者在授权边界内重复结果。对于持续运行的质量系统，试验结束的模型版本并不自动成为永久标准；超出受测产品、工艺、设备或照明条件时，应重新做适用性确认。若长期结果与短期终检方向相反，需保留两者并重新评估机制，不用短期指标覆盖延迟失效。

协议可以预定继续、限制、暂停三类决策，但不得凭一个百分比或行业通用门槛自动触发。继续要求关键安全/客户约束均满足、结果精度足以支持目标结论且运行资源可承受；限制意味着只在已验证的产品/缺陷/工况内使用并保留人工关口；暂停则恢复已批准基线，查明测量、关联、模型或执行故障后重新评估。决策理由要绑定估计结果、区间、缺失和风险记录，这样未来审计才能还原为何当时采取该状态。

19 版本更新与反馈学习：质量数据进入模型前先经过治理

19.1 更新会改变“哪些产品被看到”

视觉质检系统上线后可能形成反馈回路：模型把某些图像判为异常，产品被送往人审或返工；只有被送审的样件获得详细标签；随后模型用这些标签更新，下一版本便主要从旧模型挑选的样本中学习。由此产生的训练集不是全线产品的随机截面，而是经过阈值、人员判断和工艺处置选择后的样本。若未告警件没有独立抽检，模型不知道其漏检；若错拒件被人工放行后标签记录覆盖，模型可能把争议样本当成可靠正常。

SVSAC | sv-sac.org

SMT 数据文章尚待核查，[21]本篇不以它证明具体标签生成流程。可提出的程序风险是：若测量触发报警、复核决定处置、返工又改变产品，最终 pass/fail 就可能混合初始状态、观察机会与处理后结果。为检验这种风险，更新数据应保存不可覆盖的事件历史，包括原始图像或信号、版本分数、阈值、决策、复核依据、产品改动、独立结果及时间顺序。

数据分层可区分主动抽样与常规生产。主动学习挑选罕见、边界或模型不确定样本，适合提升对局部难例的认识，却不能直接代表量产缺陷分布。另一组按既定概率从全量未告警与告警产品中抽检，用于估计错误和迁移；其抽样和复核成本需要单列。若新模型增加告警，后续“缺陷标签增加”可能只是发现机会扩大；若更新阈值减少人审，标签数量下降也不等于过程变好。上线前后分开标识模型版本、流程版本和抽样策略，才能比较真正同口径的结果。

19.2 更新门控需包含验证样本和现场状态

每次更新前，冻结一份不参与训练的验证样本，按产品版本、缺陷严重度、视角、设备和时间分层；样本中应保留难例与典型正常变体，并注明来源是主动挑选还是概率抽样。模型比较包括整件/批次动作、错误严重度、拒识、复核时长和端到端节拍，不能只看像素级 AUROC。验证集也会随产品和工艺变化而过时，故应规定复核周期和触发条件；新照明、新镜头、镜头维修、配方更改或新缺陷类别都可能要求重新界定适用范围。

将候选版本先进行离线重放和影子运行，所有输出与当前批准版并行保存而不改变处置。若候选版本的关键严重度缺陷漏检增加、置信分数方向变化、拒识骤增、某一视角被系统性忽略，或处理时延超出动作窗口，先拒绝推广并查明原因。通过影子门后，再按协议执行有限产品族/设备的受控发布。旧版和新版同时运行时，比较数据需对齐产品、批次、阈值与最终判定，避免新旧版本使用不同抽检规则造成表面差异。

变更授权需要质量系统责任人批准适用产品范围、回退条件和受影响库存追踪；设备工程师确认成像/传感配置和运行时延；数据负责人确认标签版本、训练样本来源和数据血缘。自动训练流水线可生成候选模型，却不自动授予量产权限。若无法重现训练数据、训练代码、模型权重、预处理配置或阈值，版本就无法被审计，发生逃逸后也不能可靠识别影响面。

19.3 反馈不能只收集模型喜欢的错误

复核采样需同时覆盖模型拒识和接收件、人工接受和否决、返修前后、相机重拍、终检与下游发现。否则模型收到的反馈偏向可见、易分类和被系统标出的

SVSAC | sv-sac.org

异常。运营上可以按风险抽查一部分自动放行件，并维持对高严重度缺陷的强制独立检查。抽样率不应在没有评估的情况下因新模型“更自信”而下降；模型置信度和真实错误概率之间的关系需按受测产品验证。

当标签争议发生，保存各方原始意见和裁决依据，不要只留最终多数票。必要时可设盲态仲裁员或测量专家复查边界样件，更新缺陷定义时维护新旧版本映射，但不把旧标签无痕覆盖。学习系统可以从新的缺陷类型中扩展类别或重建正常域；此操作应触发重新验证，不把新类别错误压回旧模型的正常分布。对短期不能确认为缺陷的未知样件，保留未知状态并隔离复查，比强行伪造训练标签更可取。

19.4 将变化类型和授权动作分开

视觉质量系统至少可能遇到四种变化。输入变化包括相机、光源、焦距、产品姿态和背景；产品/过程变化包括材料、表面处理、工艺配方或装配版本；标签变化包括新缺陷定义、客户标准改版、终检测量方法改变；决策变化则是阈值、拒识区间、视图聚合或放行规则调整。相同的“分布漂移告警”可以对应完全不同的原因和处置。只监控输入向量统计量，无法知道变化是否影响最终质量；只观察模型置信分布，也可能把产品改版和相机故障混在一起。

版本控制要把这些层次拆开：模型权重、特征预处理、相机/照明配置、缺陷标签本体、阈值与产品级聚合分别拥有版本号和生效时间，并指向适用产品/工位。设备调整即使没有更换权重，也可能使旧阈值失效；定义变化即使没有变化图像，也可能改变“何为缺陷”。适用域变更时，不能把之前的验证结果自动继承到新组合。可设置受控变更窗口，在代表性产品上并行记录旧/新输出及独立验收，保留一段稳定的旧规则用于回退。

19.5 反馈样本的授权和数据血缘

反馈数据并非只要存下就可以用于训练。复核人可能有个人信息或质量责任记录，图像可能包含员工、客户或保密工艺信息；允许用途、访问角色、保留期限和脱敏方式需按企业数据治理程序和适用规则评估。即便法律允许处理，合同、工艺保密和客户审核要求仍可能限制跨厂或第三方使用。模型训练集应能追溯到数据源、授权范围、标注定义、抽样方案、处理步骤和训练配置，限制把一条线或一个客户的产品数据无条件迁移到其他任务。

训练数据进入版本前可设置质量门：缺陷定义负责人确认标签语义；质量人员抽样复核正/负和争议样件；数据负责人报告各层抽样率和未标注部分；模型负责人说明训练/验证是否按批次、时间和设备隔离。对泄漏或关联错误，保存原始

SVSAC | sv-sac.org

样本但撤销其监督标签，记录修订原因和受影响的模型版本。若已部署版本基于错误标签作了放行决策，应按质量偏差程序追踪受影响范围，而不是仅通过再训练掩盖过去的处置。

更新目标也要限定：若目标是减少误报，不能只优化告警总量而忽略同一阈值下的高风险漏检；若目标是覆盖新缺陷，需要新增缺陷定义、真实样件和交叉批次验证；若目标是节拍，需确保最终质量与复核压力保持在准入界限内。每次版本迭代登记一个主要变化假设和相应失败判据，这样才能区分“新模型更好”与“流程、阈值或样本改变了”。

19.6 版本回退也需验证受影响产品

当错误标签或新版本问题被发现，回退模型只是未来决策的恢复，不会自动撤销已发生的处置。质量团队需要查询该版本在何时、哪些产品/批次/设备上生效，保存对应告警和放行决策，按风险界定需复检或隔离的范围。若模型仅作为影子工具，风险面可能限于数据和人员判断；若输出曾影响自动分拣或参数设定，产品状态和可逆性要求更深入的追溯。版本管理应支持回滚到可执行的模型、阈值、预处理及成像配置组合，而不是只恢复权重文件。

对多模型或多相机系统，回退也不能局部混合未经验证的组件。例如旧模型与新相机标定、新聚合规则组合，可能既非已批准旧版，也非完整验证新版。发生漂移或质量事件时，应有明确的“安全基线包”恢复，并进行首件确认、时间同步核查及独立质量抽检。事故后再训练的样本应保留事件标签和选择来源，避免训练出会在相同故障条件下重复失效的更新模型。

20 三种制造情境比较：缺陷可见性、决策单位和干预路径

以下三种情境为分析构造，并非企业调查。AutoVI 原文未取得，[22]其任务数量、结果差异及图像剔除规则不作为本篇已核事实。情境设计旨在区分表面外观、部件关系和过程参数问题，不能以某数据集名称替代现场采集、功能验证或质量收益评估。

20.1 反光小缺陷：高像素分辨率可能遇上低信噪比

第一类是反光金属或覆膜零件上的微划痕、压痕或细小污染。缺陷可能只占少量像素，随光源角度改变而时隐时现。验收前要定义照明/相机视野、表面处理和缺陷尺寸等级，且以整件的可放行质量标准确定真值。若所有相机影像都来自

SVSAC | sv-sac.org

固定位置、同一光照和清洁样件，PatchCore 或 DRÆM 类方法在公开图像基准上的能力不能直接外推到反光变化或油污条件[15][16]。

竞争解释包括：所谓模型“漏检”是光学系统未采到；或者所谓“识别提升”来自人工重新调整光源而非模型。测试可在相同产品上交叉记录光源状态、角度与曝光，使用盲态尺寸参照和重复拍摄，检验模型改进是否跨照明保持。若更换漫射罩、加偏振片或机械定位即可稳定成像，简单成像改造也可能比复杂异常网络成本更低；若缺陷纹理细节仍存在且规则方法难以表述，AI 才可能提供补充。

20.2 装配逻辑错误：局部零件正常不代表组合正确

第二类情境是零件各自外观正常、但组合顺序或关系错误，例如连接器错位、线束遗漏或管夹未闭合。这是分析示例，不是已核案例。局部表面分数未必对应关系错误，项目可比较产品级表示、对象检测或逻辑规则，并用独立功能检验评价。AutoVI 保留为待核文献，[22]不据未读任务结果证明这些方案有效；图像判定也不等于功能失效或客户退回的最终结果。

一个有力反证是：如果独立功能测试已经稳定且在节拍内检出逻辑错装，视觉模型可能没有增量价值；若每种车型的装配关系规则可直接从工程 BOM/工艺防错约束中验证，确定性规则也可能更安全。AI 的价值须在现行防错之外表现为降低特定漏检或复核劳动，并且不能通过撤去原有功能检查来换取“自动化”。对于新车型和未见装配结构，拒识与人工复核比按训练中熟悉车型强行给分更合理。

20.3 受约束参数优化：小范围搜索不等于接管工艺

第三类是表面粗糙度、温度、喷涂质量等随多个可调参数变化的过程问题。与纯视觉不同，算法建议会直接改变物理状态，必须面对安全边界、材料变异、传感器误差和工艺反应时延。第 10 章已说明实体试验与模型预测的差别[5][6]；这里的情境比较着重验收：缺陷分类任务可能以产品逃逸与误拒为结果，而参数优化要将受约束试件/小批量的响应、关键特性和变化过程纳入结果，不应仅看预测误差下降或搜索点数减少。

竞争解释包括设备保养、刀具新旧、材料批次或熟练工操作造成的改善。若实验组与对照组跨不同设备/班次，参数效应和设备状态就难分开。可在批准窗口内作小规模区组试验，在独立测量系统中确认每个设定及最终响应，并设置工程停止条件和回退。若规则经验能精准锁定安全可行域，优化算法只在边缘微调而其维护成本很高，算法可能不值得部署；若响应面随材料/设备变化，序贯设计也许更有效，但仍须定期重估可行域与不确定性。

SVSAC | sv-sac.org

20.4 比较落点是任务链，而非模型高低

三类情境的主要失效点不同：反光小缺陷可能在成像与像素分辨率处失败；装配逻辑错误可能在多对象关系和整件追溯处失败；工艺优化可能在混杂、约束和物理执行处失败。对应的比较单位分别是缺陷位置与整件、产品装配关系与功能结果、参数/工艺窗口与独立过程响应。统一使用一个“AI 准确率”会掩盖每种任务的核心错误。

现场选择不必偏爱复杂模型。先确定缺陷定义、动作和独立参照，再比较人工/规则/SPC 与候选 AI 在同一合规基线上是否增加可重复信号。若 AI 只增加告警而没有改善决策，选择更简单方案；若新信号减少高严重度逃逸或能在损害发生前触发可逆处置，且没有超出复核/控制边界，才有理由进入后续验证。不同场景中混合结果、失败和拒识都应留在报告中，不能以一类任务的高分宣称整套车间质量系统已被验证。

20.5 情境比较矩阵

情境	主要失效链	最强竞争解释	优先检验结果	可能更简单的方案	AI 继续验证的条件
反光小缺陷	光源/姿态变化 → 成像信噪比改变 → 像素分数波动 → 边界件误拒或漏检	成像质量改善而非模型学习造成性能变化；标签员对“细划痕”标准不一致	盲态严重度等级、重复成像一致性、整件误拒/逃逸与复核时长	光学改造、固定视野、工程规则或少量尺寸/位置测量	缺陷图像稳定可见且模型在换照明/批次后提供可复现增量
装配逻辑错误	单件局部正常 → 关系/顺序未验证 → 整机功能风险	现有功能测试、互锁或防错已覆盖；图像关联错件	组件关系判定与独立功能/装配终检；产品 ID 关联完整率	BOM/工艺规则、传感器互锁、治具防错	视觉能补足现有防错遗漏，并保留强制功能测试/高风险复核
受约束参数优化	观测混杂/测量延迟 → 建议值偏差 →	保养、批次、操作员或设备状态解释改善	受控窗口内实体测量、多个关键响应和独立质	工程 DOE、固定窗口、简单规则/控	序贯试验减少无效探索且实物响应可复现、约

SVSAC | sv-sac.org

	关键特性或安全界限受损		量结果	制图	束经工程核验

矩阵中的“优先结果”不是一个通用 KPI，而是对应每类失效机制的最直接观测。对于反光表面，可通过相同样件在规定光照条件下多次采集来估计视觉测量系统重复性，并把图像分数与检验员盲评和物理缺陷尺寸连接；对于装配关系，需将局部视觉输出与整件几何关系、功能测试和序列追溯同时核验；对于参数优化，则应在物理试验中同时采集目标响应、未优化的关键质量特性和安全约束。若同一方案在一个结果上改善、另一结果恶化，按事先声明的质量等级和约束处理，而非临时选择最有利指标。

这三种情境还呈现不同的可逆性。图像筛查在产品尚未放行前可通过复判纠正错拒；装配错漏若被包覆或交付，回溯代价增加；工艺参数一旦影响热历史、材料组织或表面形成，结果可能不可逆。因此控制权限可以随产品进程收紧：筛查阶段允许并行提示，功能/安全关口保留独立确认，参数改变只在批准窗口中小批验证。是否允许自动处置，要看发生错误时能否发现并恢复，而不只看测试集分数。

20.6 迁移决策建议的边界

这三种情境也不可被归并为“视觉 AI 是否有效”。反光表面模型的输入单位通常是多张图像及其整件决策；装配关系需要跨对象、工序和功能结果的组合；参数优化则面对传感器序列、可控变量和实际工件响应。它们的困难不只是模型结构不同，观测对象和处理结果本身也不同。一个视觉检测器可能可以换到相邻尺寸的同材质外观件，但无法因此声称能识别装配顺序；一项磨削参数优化策略即使能迁移到同型号机床，也不能自动转用于涂层或热处理。

跨产品转移需要先对比输入覆盖、目标定义和干预机制。若新产品维持相同成像几何和缺陷定义，可进行外部图像验证并抽检现场产品；若视角、表面、材料或装配规则改变，就应重新界定目标任务。对传感器模型，需检查采样率、测量点、设备动态及可调变量是否相同。迁移测试不仅比较总体分数，也要检查按产品/批次/缺陷类分层的召回、误报、弃权和阈值稳定性。新情境如果没有足够独立参照样本，结论只能是待验证的迁移假设。

对每种情境，继续验证应有可证伪且可操作的理由：输出提供有待检验的额外信号，能在适当时点与独立结果联系，并在工程批准的可行域内行动。动作后果须可发现、受控或隔离，错拒、漏检、返修和复核负担在授权范围内。可逆动

SVSAC | sv-sac.org

作应定义回退；热历史等不可逆试验则应限定试件、损失上限及停试条件，参数恢复不能撤销已改变的产品。条件不满足时，只能保留离线探索，不能声称车间质量已提升。

21 多工序质量传递：末端变好不等于本工序改进

21.1 质量结果如何穿过工序边界

当一件产品依次经过来料筛查、加工、装配、复检和放行时，末端合格率不是某一台设备或某个算法的单独属性。上游筛除会改变下游所接收产品的组成；返工会让同一件产品多次进入某工序；报废和隔离会从后续观察中移除困难件；复检强度变化则改变缺陷被发现的机会。因而，若模型上线后下游不良率下降，至少存在四种不同解释：本工序形成缺陷的概率真的降低；上游已把高风险件拦下；返工、报废或放行规则改变了被计入的产品集合；测量或检出方式变了。单看末端分子无法区分这些路径。

有用的分析对象是同一件产品在各工序的状态轨迹，而不是把不同工序的汇总百分比直接相减。对每件产品，可用不可重复的产品 ID 连接上游材料批次、工序时间、设备/刀具状态、参数设定、检查结果、返工/报废原因和最终放行结果。追溯链不必意味着工厂已有完美数据库；若 ID 在返工后丢失、标签在工序间重用或一个托盘混有多批零件，研究应把这些记录缺口列为识别限制，而不是用时间邻近关系默认为同一对象。

因果链可以简化为：来料与设备状态影响过程响应，过程响应影响中间缺陷；检查决定哪些件被放行、返工、隔离或报废；这些动作再决定下游输入及最终检验样本。检查策略本身因此同时是质量干预与样本选择机制。若只分析“到达末端的产品”，就对一个受上游检查影响的变量作了条件化；被筛除的件越多，留下的子群越容易看起来质量好，即使工艺本身完全未变。

21.2 一个可反证的分析性算例

以下数字是为说明分母变化而构造的算例，不是工厂观测。假设某工序每期接收 1,000 件，其中 100 件来自高风险来料批，真实缺陷率为 20%；其余 900 件真实缺陷率为 2%。在没有上游筛查时，预期缺陷数为 20+18=38，接收件口径的缺陷率为 3.8%。现在在前道增加 AI 辅助筛查，准确地将那 100 件高风险批次全部隔离，但本工序参数、设备和操作均未改变。下游本工序接收 900 件普通风险件，预期缺陷数仍为 18，缺陷率变为 2%。报告若只比较接收件缺陷率，会称下

SVSAC | sv-sac.org

降 1.8 个百分点；但它实际观察到的是来料构成改变，不能据此认定加工能力提高。

反过来，如果将所有原始来料仍纳入固定分母，按原始入组产品追踪到隔离、返工、报废、最终缺陷或放行，各路径就不会消失。要检验“工序改进”主张，可提出一个反证预测：在相同材料批次风险层、设备状态和预先规定的合格入料群内，处理后本工序新增缺陷率仍应下降；若只在未经构成调整的末端总体指标上下降，而层内差异消失，证据更符合筛选构成解释。这个检验也有边界：风险分层需要在干预前定义并可靠测量；若 AI 隔离规则创造了新的人工分类，不能事后把这些处理后标签当作基线协变量。

返工例须同时保留产品与检验事件。假设 100 件均初检，10 件不合格；这 10 件均返工并各复检一次，其中 8 件通过、2 件仍不合格。初检不合格率为 10/100；两轮共有 110 次检验、12 次不合格判定，检测事件不合格率为 12/110；产品口径为 98 件最终通过、2 件仍未通过。原初检失败数不能除以 108 或 110 就冒充产品不良率。三种指标分别回答首检、检验事件和最终状态问题，均不能证明返工件原本无缺陷。成本另列 10 次返工及复检、等待和损耗，数值仍是构造示例而非现场观测。

可以预先分别报告四个估计对象：按原始产品入组的最终逃逸风险；每个工序的首次通过缺陷率；返工/报废/隔离的累计比例；以及每件产品完整制造路径的总工时或资源消耗。一个 AI 策略可能降低逃逸但增加前道误拒，也可能降低首检不良却增加返工回流。只有把分母固定、各状态转移完整记录，才看得出它在质量链上的净作用。由于各路径的结果并不总能简单相加（例如报废件无法再观察终检），还需报告状态转移定义与删失原因，不能把缺失终检自动当作合格。

21.3 分离参数作用与批次、磨损和顺序

即使产品 ID 追溯完整，参数变更仍可能与材料批次、刀具磨损或生产顺序共同发生。例如，操作员在刀具使用后期降低进给速度，同时新批次材料恰好更均匀。最终粗糙度改善可能来自降速、材料差异、提前换刀，或这些因素的共同作用。若仅比较参数变更前后均值，时间趋势、设备维护和产品组合都是竞争解释。

分析上需要将“处理前已知状态”和“处理后的工序结果”分开。材料批次、设备编号、刀具累计加工量、开机至加工时间、工件型号、班组和环境条件可作为分层或设计区组变量；只有在参数设置之前测得的状态才能用于平衡处理。参数

SVSAC | sv-sac.org

改动后才出现的温度、振动或操作者处置，可能是干预传导路径，不应无条件作为协变量调整，否则会把真实作用路径扣除。若批次或刀具状态与处理完全重合，没有共同支持的对照工况，统计调整不能凭空补出缺失比较，应将因果结论降为无法识别。

工序顺序本身也可能改变结果。首件往往用于调机；连续生产带来热稳定或磨损累积；换型与停机后的首批可能更易偏离。将所有产品作为交换的独立样本，会把这些顺序依赖平均掉。更稳妥的记录至少保留参数生效时刻、实际执行时刻、工件顺序、停机/换刀/换批事件及结果测量时刻。若同一刀具下运行先后会互相影响，比较单位应覆盖能承受这种干扰的生产段，而不能把单件当作互不相关的处理单元。

对多个工序的改进归因，适合把局部“过程变化”与系统“流出变化”作为两条证据链。前者问在同一来料和设备状态下，本工序的独立测量响应是否改变；后者问整条路线的缺陷逃逸、返工、报废和周期是否改变。若本工序信号不变而最终结果改善，可能是上游筛选有效，这仍可能有价值，但价值应归属于筛选策略；若过程响应改善而终检无变化，可能是下游掩盖、样本量不足或终检灵敏度有限。两条链一致会增强机制解释，不一致则提醒研究者保留替代解释。

21.4 质量链上的价值与责任归属

工序优化的经济结果由整个流程决定。上游把缺陷提前发现，可能避免昂贵的后续加工，却也可能使隔离区占用、供应商索赔、补料和排产变得更频繁。前道模型的误拒成本与末端漏检的客户后果处于不同时间和责任位置，不宜仅以某工位的准确率排序。改进记账可以对每个状态转移标注发生工序、发现工序、责任规则和已发生资源消耗；跨部门收益再按明确约定归集，避免某环节通过把困难件送去下游复检来改善自己的指标。

这也改变“最优改进”的单位。对单站来说，提高该站通过率可能是目标；对整条产品路线，减少最终不合格、返工等待和不可逆损失才是更贴近客户结果的对象。两者可能冲突：若上游严筛，单站通过率变好，整体良率却因误拒增加而下降；若只对总路线评分，则可能看不见某工序对潜在重大缺陷的早期拦截作用。应把强制质量门槛和系统性结果同时列出，清楚说明哪个单位拥有决策权以及误拒/漏检如何裁决。

本章的分析框架给出可反驳的归因条件：若声称某工序参数导致改进，需要在有重叠的来料、设备状态和顺序条件下观察到局部结果变化；若声称 AI 筛查改善整条路线，应显示相同原始入组分母下的最终风险或资源结果变化，并完整

SVSAC | sv-sac.org

记录被筛除件和复检强度。任何一项均不充分时，结论应收窄至可观察到的环节，而不将“下游指标变好”泛化为工艺能力已经提高。

22 鲁棒改进：多个目标如何在工况变化下共存

22.1 单点最优不代表可持续的工艺选择

生产团队面对的往往不是一个标量目标。提高进给可能缩短节拍，却增加切削力、刀具磨损或尺寸波动；提高相机告警敏感度可能减少漏检，却增加误拒和人工复核队列；降低炉温也许降低能耗，却改变材料组织或使后道返工增加。质量、周期、能耗、设备寿命和复核容量的单位不同，也由不同责任人承担。将它们压成一个分数，必须有权重与代价换算；权重不是算法自己从数据中“发现”的自然答案，而是组织对风险、客户和资源优先级的选择。

多目标分析先识别不可协商的限制，再显示可选择的权衡。设决策设定为 u，工况情境为 x，若需最小化质量损失 Q(u,x)、周期 T(u,x)、能耗 E(u,x)及磨损 W(u,x)，同时满足安全约束 g(u,x)≤0、客户规格和复核容量 R(u,x)≤Rmax，则可行集合是各限制的交集。若安全或合同界限是硬条件，就不应先把违反它们的候选计入 Pareto 前沿，再靠综合分数“补偿”。硬约束定义谁可进入比较；软目标描述可行方案之间的权衡。

Pareto 有效只表示在现有目标、测量与候选集合下，没有另一方案在所有目标上都不差且至少一项更好。它不告诉管理者应选哪一个，更不保证候选集合包含实际工艺的最佳方案。由数据估计的前沿还带有模型不确定性、测量误差和工况变化；某点在均值预测下非支配，置信范围却可能被另一方案支配。决策者需要查看前沿的稳定性：小幅改变权重、测量误差或工况占比后，推荐是否剧烈跳变？若跳变，系统提供的应是需要人工取舍的范围，而非一个假精确的单一最优值。

22.2 局部可行与跨工况稳健不是一回事

某参数组合可能在新刀具、干燥材料和稳定供电时表现最佳，却在刀具磨损后越过温升或尺寸约束。另一组合在最佳工况下略慢，但对材料批次、环境湿度或设备状态波动不敏感。两者体现不同决策：优化平均工况下的平均表现，或要求在预定的合理工况集合内保持边界并控制尾部损失。所谓“鲁棒”必须说清楚针对何种变化、何种分位/最坏情形或概率标准，不能仅因输出是一个不确定性区间就自动称为稳健。

SVSAC | sv-sac.org

可以把工况划分为事先可识别的情境层（材料批次等级、刀具使用区间、设备状态或班次），先在各层估计目标与限制，再比较跨层表现。若生产需求有可信的情境权重，可报告加权平均，同时单独呈现关键层下的最低质量表现；若权重会随订单或供应情况变动，则用权重范围做敏感性分析。对于危害严重而不可逆的规格违反，采用“所有合理情境均满足”的鲁棒约束，可能比一个平均违反概率更合适；但这种保守度会缩小可行域，必须由质量/安全责任人解释其依据。

Jia 等的 CNC 铣削文献仍待核，本篇不确认其机型、目标构造或三项改善结果。[23] 本篇保留多目标评估要求：若参数可能同时影响效率、能耗和表面质量，应报告各响应的测量定义、权衡和可行域，而非只择一个综合分数最优点。具体耦合方向、幅度及跨工艺适用性都须由可读证据或本地试验确定；第 10 章的搜索顺序不能代替这种检验。

22.3 不确定性概率不等于安全许可

“约 95%概率满足约束”是模型与数据生成假设下的统计表述，不是设备认证、客户放行或法律授权。它依赖约束测量足够可靠、训练域和实际工况相符、变化机制可建模等条件。若安全边界定义遗漏、传感器量程漂移或工况超出试验支持范围，概率数值可能精确地描述错误对象。将这种概率直接显示为“可安全执行”会混淆算法内部置信度与工程批准。

Berkenkamp 等提出将性能目标和多个安全约束分开建模的 SafeOpt-MC，并在四旋翼参数调优实验中说明方法如何在已知初始安全集合和 GP 规律假设下探索[24]。该工作支持的是一种明确问题结构及假设条件，不是制造现场安全许可。Fiedler 等进一步分析 SafeOpt 类算法对有效不确定度界、函数范数假设和实际启发式实现的依赖，指出启发式替代可能使理论保证失效[25]。这恰好说明“带安全约束的优化器”这个标签并不等于任何特定系统已获得安全保证。

在车间，安全与质量约束还常有统计模型之外的来源：设备制造商限制、工艺评定、控制计划、法规、客户特殊特性、操作规程和质量授权。模型可用于在这些审批边界内排序候选试验、发现可能的冲突或估计超限风险；边界由有权限的工程与质量程序确定。若约束边界未知、难以测量或无法在当前试验中被安全检验，合理结果是排除该优化问题或保持人工工程设定，而不是让模型在无证据时自行扩展。

22.4 复核能力也是系统约束

多目标中的“复核容量”常被隐藏在模型之外。若检测阈值下调使更多边界产品进入人工队列，模型的召回提升可能转化为延迟、抽检不足或紧急放行。应把

SVSAC | sv-sac.org

单位时间内告警到达、每类告警的处理时间分布、值班能力和允许积压量转化为运行限制，并考虑峰值而非只看月均。一个在平均班次可处理的阈值，可能在换型或高缺陷短时聚集时失效。人为复核结果也不是免费且无误差的“最终真值”：判定疲劳、标准不一致和排队时间会影响其质量。

此类约束与产品质量门槛不同。客户规格是产品必须达到的条件；复核容量则决定组织能否执行规定动作。若两者冲突，合理响应可能是降低产速、增加资源、隔离产品或退回现行检查流程，而不是降低强制检验要求。把容量加入可行集合并不能合理化越权放行，只能揭示当前运行方案在操作上不可实施。

22.5 选择责任应与 Pareto 图一起呈现

假设存在三种经验证可行的方案：A 在表面质量最好但节拍慢、能耗高；B 节拍快但刀具更快磨损；C 各项处于中间但跨批次变化小。Pareto 图可显示三者相互不可支配，但选择仍需回答：质量等级是否允许 A 以外方案？磨损是否会造成计划外停机？平均节拍改善是否超过高峰复核或换刀负担？哪一类风险不可折算成成本？这些问题需要质量、生产、设备和客户责任人在部署前共同定义，而非在算法中隐藏偏好。

若评审只能给出一个经批准的运行点，研究报告仍应保留其相邻备选及其权衡，说明为何被排除。可采用约束优先法，例如先满足关键质量特性和安全界限，再在余下候选中减少能耗或周期；也可以对软目标给出明确权重。但若权重变化会令选择翻转，应报告敏感性并将决策标为价值判断。模型更新、材料变化或设备磨损若使原来的可行集合漂移，就需要重新确认边界，而不是把旧推荐当永久配方。

对 AI 质量系统的有限启示是：模型适合帮助描述多响应关系、聚焦可验证的工艺区域或发现折中候选；工程和质量职能仍负责定义不允许违反的边界，业务责任人负责说明权衡如何排序。系统的“最优”只在指定数据、工况、目标和限制下成立。若唯一可行方案已由成熟控制计划明确，或者复杂模型不会改变任何被授权决策，继续优化的技术收益可能为零；这也是多目标分析应允许得出的结果，而不是用算法复杂度证明部署必要。

23 讨论：从模型成绩到可归因的质量改善

23.1 不同证据回答的是不同层级的问题

本次分析区分识别能力、预测价值、参数干预效果与最终质量改善。可读的 LOCO AD 及 AD 2 相关原段支持受控视觉任务的有限比较；[2,4]MVTec AD 与

SVSAC | sv-sac.org

Real-IAD 的规模仅见 AD 2 二手转述，未核其原始协议。[1,3] PatchCore、DRÆM 与 EfficientAD 原文未取得，[15,16,17]不以其方法或性能作为已核对照。上述证据范围不能单独确定产线阈值、整件逃逸、复核负荷或客户风险。

运行环境数据值得关注，但 InsPLAD 和 SMT AOI 原文在本次分析中均未取得，[9,21]不能称为已核现场采集与标签流程证据。我们提出的研究要求是审查背景、视角、时间和复判如何影响记录；即使将来确认数据来自现场，仍需另验标签误差和策略对照，才能讨论部署净收益。现场性本身不等于真值可靠或因果识别。

磨削摘要支持特定过程及约束的说明，[5]但喷涂与 CNC 原文不可读，不计其具体设备和试验结果。[6,23] 安全优化的作者预印本摘要提供目标与安全约束可区分建模及保证条件需审查的线索；[24,25]正式版证明与完整实现未核。本篇因此仅提出受约束现场试验设计，不能预估其他产线收益，也不能把算法安全条件等同质量授权、客户放行或法规认证。

证据之间因此不宜按“更真实”简单排序，而应根据主张选择：模型是否区分图像异常，用独立基准或现场留出图像检验；告警是否能及时行动，用时间外生产记录和作业队列检验；调参是否改变产品质量，用有共同支持的受控实物对照检验；上线策略是否改善整线结果，用固定原始分母、完整返工/报废追溯和独立终检检验；闭环控制是否符合安全边界，则需要工程危害分析、独立联锁和授权程序。任何单一类型证据都不能代替另一层。

23.2 支持证据、反例与可接受的竞争解释

可读 AD 2 原段说明扩展视觉任务包含照明、透明物及微小缺陷等挑战，[4]磨削摘要提示需同时考虑多个过程约束。[5] 这为现场可检验假设提供有限依据，尚不证明任一拟用算法具有可迁移净收益。未读的异常检测及 CNC 文献仅作待核线索，[15,16,17,23]不以其具体方法或结果补强结论。部署能否降低废品，须由独立终检、成本和干预对照回答。

相反证据并非总是模型失效。若过程早已稳定，历史规则或操作员经验能够在产品进入关键工序前消除风险，额外预测可能没有增量提前量；低基率下，少量召回改进可能带来大量复核；未见缺陷可能被正常样本模型接受，正常但未见的产品变体又可能被大量拒绝。若改善只发生在告警产品或更严格检查的班次，检测强度变化、筛选和关注效应可能解释观察结果。若模型提升后实际参数未改变，或者推荐被频繁拒绝，最终质量没有变化与“预测能力没有用”并不矛盾，可能只是干预链没能执行。

SVSAC | sv-sac.org

竞争解释要沿质量链具体检验，而不是将未观察到的差异都归咎模型。结果变好可能来自光学照明改造、设备保养、刀具更新、来料批次更均匀、操作员培训、排除高风险产品或复检定义改变；结果变坏可能是模型召回提升后发现率提高，而非产品真实恶化。若处理与设备、时间、批次、人员完全重合，观察比较便无法把它们分开。多工序场景还可能因为前道筛选减少了下游缺陷分母；产品级追溯和未报警件独立抽检是识别这类机制的重要条件，而不是额外的数据便利。

因此，即便结果不利，也应判断支持何种机制：模型分数没有排序能力、测量系统不可重复、阈值带来的队列不可承受、工艺推荐改变不了物理响应，或现场没有授权/执行条件。这些失败意味着不同决策：前两类可能不适合继续开发；队列问题可能要调整流程资源；物理响应无效提示候选变量选错；授权边界不清则应留在影子或工程试件阶段。把所有问题汇成“AI 项目失败”会丢失可迁移知识，也会使成功主张难以被反驳。

23.3 可识别性与外部效度的界线

本篇没有企业原始日志、追溯数据、成本台账或本地试验，不能估计行业平均误报、投资回报或产线改善。LOCO AD、AD 2、SPC 和 SSMSPC 仅在已读相关原段范围内使用；磨削和安全优化分别限于摘要及作者预印本摘要。无法读取的案例、方法细节和数字不作为已核事实；MVTec AD 及 Real-IAD 二手规模说明明确标出转述来源。数值算例只作计算说明，本篇的机制分析与拟议设计仍须现场检验。

公开研究的外部效度取决于输入、标签、处理、结果及现场组织是否足够相似。基准中的图像常可重复拍摄和标注，而生产数据受到换型、遮挡、人员复判、缺陷稀有和追溯错误影响；实体工艺实验常在控制范围内改变设定，现场则可能同时发生维护、来料变更和赶工；机器人中的概率安全保证基于数学模型假设，制造设备还受客户、质量与法定流程约束。不同证据之间可以共享概念，但不能共享未经检验的结果数值。

可识别性需要多个环节同时成立：结果变量与单位预先明确；处理实际定义为阈值、建议策略或参数变更而非含糊的“有 AI”；分派或比较提供有解释力的对照；缺陷测量独立于模型输出；原始入组分母覆盖被隔离、返工和报废件；关键材料/设备状态有重叠或随机平衡；反馈、拒识和执行情况被记录；缺失与删失有明确定义。若这些条件不成立，复杂模型、更多数据或显著的离线差异不会自动修补识别缺口。

SVSAC | sv-sac.org

外推也应分层。相同相机、产品与缺陷定义下的时间外验证，主要检验时间稳定性；换批次可检验部分来料变化；跨设备验证涉及相机、控制和维护差异；跨产品线或行业迁移更可能改变缺陷机理、质量规范和动作成本。应把“本线当前版本已验证”“相邻批次待持续监测”“其他产品仍需外部验证”等层次分开声明。一次通过验证只能支持有限的范围与期限，不能让模型版本、设备或产品改变后继续沿用未经审查的结论。

23.4 相对单任务分数的增量

只看单个模型分数，会把测量、样本选择、操作动作和产品后果压缩成一个数字。本篇的增量不是另造一个统一 AI 成熟度指标，而是把证据与主张一一对应：图像分数不能替代整件判定；异常预测不能替代参数因果效应；某工序的通过率不能替代固定分母下的全路线结果；Pareto 前沿不能替代组织的风险选择；统计意义的安全概率不能替代工程授权。每一层都保留可以证伪的预期结果和竞争解释。

这也使“应该不采用 AI”成为实质性结论的可能选项。如果现有量规、规格限、控制图或功能测试可以可靠及时地发现问题；如果模型不能增加行动提前量；如果误报超过可复核产能；如果优化器提供的方案无法在经批准窗口内重复实现；或者质量结果不变而维护成本增加，那么项目停止或缩小范围是理性结果。AI 的必要性不由研究新颖性、模型复杂度、数据规模或供应商演示决定，而由在既有合规流程上提供的可重复决策增量决定。

23.5 结果不确定时，决策建议应按主张层级收窄

现场结果常不是清晰的“有效/无效”。例如模型对特定缺陷有更高召回，但置信区间很宽；复核工时增加，最终逃逸事件没有明显变化；或工艺响应改善，但关键质量特性和末端判定并未变化。此时，不能把未达到统计显著性直接解释为没有作用，也不能把有利点估计当成已经实现收益。应报告估计方向、区间宽度、样本和集群数量、测量可信度，以及该区间是否同时容纳有意义的改善与恶化。若精度不足以区分两者，结论是证据尚不能区分，而非项目成功或必然失败。

判断也应针对事先提出的机制，而不是事后挑选指标。若假设视觉告警让缺陷更早被拦截，主要证据是独立终检追踪的逃逸是否变化，并观察误拒和复核容量；若假设参数建议抑制波动，应先看到相同工况下的过程响应变化，再看产品结果是否沿预期路径变化；若只发现单位时间告警数下降，却没有相同分母、同

SVSAC | sv-sac.org

等抽检的产品结果，不能把工作量改善等同于质量改善。机制链断在哪一环，就只对该环节作结论，避免从一个中间指标跳到最终客户效果。

异质结果同样需要保留。整体平均不变可能与某材料、设备或缺陷层的改善相容，但只有这些层次事先定义、样本具有可比性且报告全部结果时，才能支持有限的条件性解释。反之，如果一个总体提升由容易产品主导、关键层恶化被平均掩盖，就不构成稳健改进。未来的现场研究应在项目开始前指定少量主要质量结果、次要过程结果和严重度分层，同时公开所有预设类别与不利事件；这既减少选择性报告，也让第三方能判断改进是普遍、局部还是由构成变化造成。任何数值门槛都须由具体质量风险、客户要求和检验精度决定，本篇不提供可通用套用的界值。

24 决策建议

本篇的核心问题是：在缺陷稀少、产品变型和过程状态变化的制造环境中，怎样判断 AI 质检、工艺预测与参数优化是否改善最终质量，而非只提高离线分数？回答是：必须沿测量–识别–处置–物理响应–独立质量确认这条链分别验证，并把结论限定在被观察到的层级。公开图像基准可以证明给定拆分下的算法识别能力；它不直接给出现场漏检率、复核负荷或最终合格率。过程模型只有在时间外预测早于可行动窗口、且告警后动作被执行时，才可能成为操作信号。参数优化需要在预先批准的可行域内进行实体试验，以独立测量的工件结果而非代理模型分数确认。全路线改善还必须固定原始产品分母，并追踪隔离、返工、报废、复检和放行，防止筛选改变样本后把下游变好误写成本工序改善。

这一结论是条件性的，不是 AI 天然优于传统方法的命题。若质量问题可被稳定量具、规格限、SPC、功能测试或简单防错及时处理，复杂模型可能没有可验证的额外价值；若异常结构难以用规则表示，现场数据有可靠追溯，且模型信号在既有流程之外改善了独立结果并维持可接受复核量，则 AI 可以作为辅助检测或试验排序工具。被识别的风险越严重、不可逆性越高，模型越需要保持提示或复核角色，并由质量和工程人员控制实际动作权限。多目标输出也只提供可行折中候选；客户规格、安全条件和组织风险偏好不会从 Pareto 图中自动生成。

对产业实践的有限启示是，先确定要改进的产品结果和决策单位，再选择模型；保留独立参照、可追溯分母和不利结果；将模型推荐与批准的工艺边界分开；以可否证的现场结果决定扩展、修改或停止。没有企业日志与本地对照实验时，当前材料不支持具体质量收益、成本节约、通用阈值或适用于所有车间的部

SVSAC | sv-sac.org

署顺序。本篇提供的是验证和归因框架，实际效果仍需在具体产品、过程、测量系统与授权条件下获得证据。

给企业的下一步建议

车间 AI 项目的价值，应体现在减少真实缺陷和工艺损失上；模型演示再漂亮，也不能代替生产现场的改进。因此，我们建议把下一步工作收敛为一项有明确交付物的专项任务。

具体起点是：优先选择损失可计量、数据可追溯的质量问题，把误检漏检、工况变化和人工复核纳入同一验收。在此基础上，管理层可以按证据充分程度、资源投入与业务影响确定优先级，先验证关键假设，再安排后续合作与投入。

要把上述判断用于实际经营，可以围绕具体产品或项目、目标市场、现有能力和决策难点，进一步形成质量损失诊断、场景优先级、质检与参数优化试点方案。这类工作需要把行业研究、企业诊断与合作条件核验衔接起来，让分析结果成为管理层可以使用的决策依据。

资料与分析范围

下列说明列明资料范围、分析依据与适用条件，便于读者辨别文中的事实、推演和建议。市场进入、交易安排及适用规则仍需结合具体项目核实。

制造质量不是单一的“缺陷识别准确率”问题。一个产品是否满足要求，取决于产品特性、测量系统、过程状态、抽检策略以及缺陷后续处置的共同作用。视觉模型可以给出图像级异常分数或缺陷区域；过程模型可以估计漂移或质量特性；优化算法可以从受约束的试验中建议参数；控制器则可能据此影响设备动作。它们位于不同决策环节，所需证据、失效模式与责任边界并不相同。将四者统称为“AI 质检”会把图像基准指标误读为整件产品的漏检率，把关联预测误作可干预的参数因果效应，或把离线建议误作已经验证的自动闭环。本篇以定向文献梳理与机制分析研究：视觉检测、过程预测和参数优化在何种条件下可能改善最终质量，而非仅提高离线分数。所依据的材料包括公开研究、机构统计资料与法律文本，判断仅就可读部分作出；只取得摘要或无法读取的，不计为全文核对。分析逐项区分研究对象、观察单位、指标及外推边界，不报告自有企业数据，不将基准性能视为车间实绩。结论仍是待现场检验的条件命题：测量追溯、阈值与成本、受控过程变更和独立终检必须连成可观察的工作链；单个高 AUROC 不足以证实这条链。

SVSAC | sv-sac.org

参考文献

[1] Paul Bergmann; Michael Fauser; David Sattlegger; Carsten Steger. MVTec AD—A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9584–9592 (2019). https://openaccess.thecvf.com/content_CVPR_2019/html/Bergmann_MVTec_AD_--_A_Comprehensive_Real-World_Dataset_for_Unsupervised_Anomaly_CVPR_2019_paper.html

[2] Paul Bergmann; Kilian Batzner; Michael Fauser; David Sattlegger; Carsten Steger. Beyond Dents and Scratches: Logical Constraints in Unsupervised Anomaly Detection and Localization. International Journal of Computer Vision, 130(4), 947–969 (2022). https://link.springer.com/article/10.1007/s11263-022-01578-9

[3] Chengjie Wang; Wenbing Zhu; Bin-Bin Gao; Zhenye Gan; Jianning Zhang; Zhihao Gu; Shuguang Qian; Mingang Chen; Lizhuang Ma. Real-IAD: A Real-World Multi-View Dataset for Benchmarking Versatile Industrial Anomaly Detection. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 22883–22892 (2024). https://openaccess.thecvf.com/content/CVPR2024/html/Wang_Real-IAD_A_Real-World_Multi-View_Dataset_for_Benchmarking_Versatile_Industrial_Anomaly_Detection_CVPR_2024_paper.html

[4] Lars Heckler-Kram; Jan-Hendrik Neudeck; Ulla Scheler; Rebecca König; Carsten Steger. The MVTec AD 2 Dataset: Advanced Scenarios for Unsupervised Anomaly Detection. *International Journal of Computer Vision*, **134**, article 175 (2026). https://link.springer.com/article/10.1007/s11263-026-02743-0

[5] Markus Maier; Alisa Rupenyan; Christian Bobst; Konrad Wegener. Self-Optimizing Grinding Machines using Gaussian Process Models and Constrained Bayesian Optimization. *The International Journal of Advanced Manufacturing Technology*, **108**, 539–552 (2020). https://link.springer.com/article/10.1007/s00170-020-05369-9

[6] Xavier Guidetti; Alisa Rupenyan; Lutz Fassl; Majid Nabavi; John Lygeros. Plasma Spray Process Parameters Configuration using Sample-efficient Batch Bayesian Optimization. 2021 IEEE 17th International Conference on Automation Science and Engineering (CASE), pp. 31–38 (2021). https://doi.org/10.1109/CASE49439.2021.9551499

[7] Standing Committee of the National People's Congress of the People's Republic of China. 中华人民共和国个人信息保护法. Official law text published by the Cyberspace Administration of China; adopted 20 August 2021 (2021). https://www.cac.gov.cn/2021-08/20/c_1631050028355286.htm

[8] Standing Committee of the National People's Congress of the People's Republic of China. 中华人民共和国数据安全法. Official law text published by the National Bureau of Statistics; adopted 10 June 2021 (2021). https://www.stats.gov.cn/gk/tjfg/xgfxfg/202503/t20250310_1958928.html

SVSAC | sv-sac.org

[9] André Luiz Buarque Vieira e Silva; Heitor de Castro Felix; Franscisco Paulo Magalhães Simões; Veronica Teichrieb; Michel Mozinho dos Santos; Hemir Santiago; Virginia Sgotti; Henrique Lott Neto. InsPLAD: A Dataset and Benchmark for Power Line Asset Inspection in UAV Images. International Journal of Remote Sensing, 44(23), 7294–7320 (2023). https://www.tandfonline.com/doi/abs/10.1080/01431161.2023.2283900

[10] C. M. Croarkin. NIST/SEMATECH Engineering Statistics Handbook, Chapter 2: Measurement Process Characterization. NIST/SEMATECH Engineering Statistics Handbook; NIST technical reference, updated 2017 (2003). https://www.nist.gov/publications/nistsematech-engineering-statistics-handbook-chapter-2-measurement-process

[11] Carmen Simion. Measurement system analysis by attribute, an effective tool to ensure the quality of the visual inspection process within an organization. *MATEC Web of Conferences*, **290**, 05004 (2019). https://doi.org/10.1051/matecconf/201929005004

[12] Carlos Marques; Nuno Lopes; Gilberto Santos; Isabel Delgado; Pedro Delgado. Improving operator evaluation skills for defect classification using training strategy supported by attribute agreement analysis. *Measurement*, **119**, 129–141 (2018). https://www.sciencedirect.com/science/article/abs/pii/S0263224118300472

[13] Jonas Aust; Dirk Pons. Assessment of Aircraft Engine Blade Inspection Performance Using Attribute Agreement Analysis. *Safety*, **8(2)**, article 23 (2022). https://www.mdpi.com/2313-576X/8/2/23

[14] Takaya Saito; Marc Rehmsmeier. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE, 10(3), e0118432 (2015). https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118432

[15] Karsten Roth; Latha Pemula; Joaquin Zepeda; Bernhard Schölkopf; Thomas Brox; Peter Gehler. Towards Total Recall in Industrial Anomaly Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 14298–14308 (2022). https://openaccess.thecvf.com/content/CVPR2022/html/Roth_Towards_Total_Recall_in_Industrial_Anomaly_Detection_CVPR_2022_paper.html

[16] Vitjan Zavrtanik; Matej Kristan; Danijel Skočaj. DRÆM – A Discriminatively Trained Reconstruction Embedding for Surface Anomaly Detection. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 8310–8319 (2021). https://openaccess.thecvf.com/content/ICCV2021/html/Zavrtanik_DRAEM_-_A_Discriminatively_Trained_Reconstruction_Embedding_for_Surface_Anomaly_Detection_ICCV_2021_paper.html

[17] Kilian Batzner; Lars Heckler; Rebecca König. EfficientAD: Accurate Visual Anomaly Detection at Millisecond-Level Latencies. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 127–137 (2024). https://openaccess.thecvf.com/content/WACV2024/html/Batzner_EfficientAD_Accurate_Visual_Anomaly_Detection_at_Millisecond-Level_Latencies_WACV_2024_paper.html

[18] NIST/SEMATECH. NIST/SEMATECH e-Handbook of Statistical Methods: Assessing Process Stability (3.4.5) and Process Capability (3.4.6). NIST/SEMATECH engineering statistical reference handbook (2012). https://www.itl.nist.gov/div898/handbook/ppc/section4/ppc45.htm

SVSAC | sv-sac.org

[19] Tobias Biegel; Patrick Helm; Nicolas Jourdan; Joachim Metternich. SSMSPC: Self-Supervised Multivariate Statistical In-Process Control in Discrete Manufacturing Processes. *Journal of Intelligent Manufacturing*, **35**, 2671–2698; version of record published 29 June 2023 (2024). https://link.springer.com/article/10.1007/s10845-023-02156-7

[20] Federico Pittino; Michael Puggl; Thomas Moldaschl; Christina Hirschl. Automatic Anomaly Detection on In-Production Manufacturing Machines Using Statistical Learning Methods. Sensors, 20(8), article 2344; published 20 April 2020 (2020). https://www.mdpi.com/1424-8220/20/8/2344

[21] Korbinian Pfab; Roman Eichler; Adarsh Mallandur; Marcel Rothering. Data of Automated Optical Inspection of Surface-Mounted Technology Electronic Production. *Data in Brief*, **53**, article 110110; published 30 January 2024 (2024). https://www.sciencedirect.com/science/article/pii/S2352340924000830

[22] Philippe Carvalho; Meriem Lafou; Alexandre Durupt; Antoine Leblanc; Yves Grandvalet. Detecting Visual Anomalies in an Industrial Environment: Unsupervised Methods Put to the Test on the AutoVI Dataset. *Computers in Industry*, **163**, article 104151 (2024). https://www.sciencedirect.com/science/article/pii/S0166361524000794

[23] Shun Jia; Shang Wang; Na Zhang; Wei Cai; Yang Liu; Jian Hao; Zhongwei Zhang; Yang Yang; Yang Sui. Multi-objective parameter optimization of CNC plane milling for sustainable manufacturing. Environmental Science and Pollution Research, online ahead of print, 30 December 2022; 2023 volume record (2022). https://pmc.ncbi.nlm.nih.gov/articles/PMC9803406/

[24] Felix Berkenkamp; Andreas Krause; Angela P. Schoellig. Bayesian Optimization with Safety Constraints: Safe and Automatic Parameter Tuning in Robotics. *Machine Learning*, volume **112**, pages 3713–3747 (published online 24 June 2021; volume issue 2023) (2021). https://arxiv.org/abs/1602.04450

[25] Christian Fiedler; Johanna Menn; Lukas Kreisköther; Sebastian Trimpe. On Safety in Safe Bayesian Optimization. Transactions on Machine Learning Research (TMLR), 2024; published 26 September 2024 (2024). https://arxiv.org/abs/2403.12948

SVSAC | sv-sac.org