University for the Creative Arts

BSBI
School of Business and Innovation

Dissertation Title:

Safe Tool Handover in Human-Robot Collaboration: A Dual-Stream Framework Combining VLA Failure Detection and Human Intent Recognition

Master title:

Masters Artificial Intelligence

Name:

Nicanor Kipkosgei Korir

Year: 2025

ABSTRACT

Industrial robots driven by Vision-Language-Action models can now interpret spoken instructions and manipulate unfamiliar objects, yet every documented commercial deployment still keeps these robots physically or temporally separated from human workers. The barrier is a safety one. When a Vision-Language-Action model meets conditions outside its training distribution it tends to continue confidently toward a wrong action rather than signalling doubt, and in close collaboration that silent failure becomes the precondition for injury. This dissertation designs and evaluates a dual-stream safety monitor for the most demanding collaborative case, robot-to-human tool handover. The first stream watches the robot, reading the policy's own internal state to estimate the chance of impending failure. The second stream watches the human, predicting reach intent from partial trajectory data. A rule-based fusion layer combines the two into a graduated response controller. The framework is tested in a PyBullet simulation across four conditions, an unmonitored baseline, each stream alone, and the full dual-stream system, against four out-of-distribution failure groups and a nominal control, at fifty trials per cell. Against a pre-registered three-part criterion the dual-stream system succeeds. It intervened on 81.5 percent of failure trials where the unmonitored baseline intervened on none, while holding nominal false positives and completion within four points of that baseline. The clearest result is that each stream covers a different failure family, and only the combined system addresses both the robot-side and the human-side failures. The evidence is simulation-only and bounded accordingly, but it provides a disciplined foundation for later hardware validation.

ACKNOWLEDGEMENTS

I owe my deepest thanks to my supervisor, Professor Vincent English. His feedback came chapter by chapter, and every round of it pushed me to be sharper. He kept asking me to compare things directly rather than describe them in turn, to tie each number back to the question it was meant to answer, and to make one chapter lead into the next instead of simply stopping. That insistence did more to discipline my thinking than any paper I read. Whatever rigour this dissertation has owes a great deal to his guidance.

I am grateful to the staff of the Berlin School of Business and Innovation and the University for the Creative Arts, who built the structure that made a project of this size possible in a single year. I also want to thank the open research community behind OpenVLA, PyBullet, and the failure detection and intent recognition work this study leans on. They shared their code and their methods freely, and this dissertation was built directly on that generosity.

My family carried me through this in ways that are hard to put into a few lines. My wife Nicklah gave up a great deal of our time together so that I could keep working, and she never once made me feel guilty for it. She also proofread these pages with a careful eye and caught things I had stopped being able to see. My daughter was patient with a father who was often at his desk, and the sight of her was what pulled me back up on the harder days. My parents taught me how to stay steady when things get difficult, and that advice held me together more than once during this year. This work is dedicated to the three of them, and to my parents who showed me how to finish what I start.

CONTENTS

Statement of compliance with academic ethics and the avoidance of plagiarism

I honestly declare that this dissertation is entirely my own work and none of its part has been copied from printed or electronic sources, translated from foreign sources and reproduced from essays of other researchers or students. Wherever I have been based on ideas or other people texts I clearly declare it through the good use of references following academic ethics.

(In the case that is proved that part of the essay does not constitute an original work, but a copy of an already published essay or from another source, the student will be expelled permanently from the postgraduate program).

Name and Surname (Capital letters):

NICANOR KIPKOSGEI KORIR

Date: 24/06/2026

DISSERTATION THESIS

INTRODUCTION

The integration of intelligent robots into human workplaces has accelerated sharply. By 2025, humanoid robots had moved from research laboratories into commercial pilots in logistics and automotive manufacturing, and the trajectory of deployment has continued through 2026 (Asif et al., 2026). These deployments are driven by Vision-Language-Action (VLA) models, which are artificial intelligence systems that integrate visual perception, language understanding, and motor control into a single end-to-end architecture. Unlike rule-based predecessors, VLA-driven robots such as OpenVLA (Kim et al., 2025) and Physical Intelligence's pi-zero (Black et al., 2024) interpret open-ended verbal instructions, adapt to novel objects, and execute complex manipulation tasks without explicit reprogramming.

Despite this progress, every documented commercial deployment shares a critical limitation. Robots operate in semi-segregated zones with physical or temporal separation from human colleagues during normal operation (Asif et al., 2026; Frontiers in Robotics and AI, 2025). The ISO 10218-2:2025 revision, alongside the ISO 25785-1:2025 humanoid robot safety standard, confirms that current certification requirements fall well short of what unrestricted collaborative operation would demand (ISO, 2025a; ISO, 2025b). The threshold required to work cooperatively alongside people without physical barriers has not yet been achieved. This threshold, called the cooperative safety threshold, defines the central challenge this dissertation addresses.

Of the many forms close human-robot collaboration might take, tool handover is the most demanding test case and the most revealing. It requires a robot and a human to share a single object at a defined point in space, within a defined time window, with shared timing, spatial precision, and mutual anticipation imposed simultaneously. Both parties must commit to complementary actions before either has certainty about the other's readiness. Parastegari et al. (2018) identified this mutual commitment as the central challenge in robot-human handover: the giver must release precisely as the receiver grasps, without confirmation that the other is ready. A safe handover is not merely one that avoids harm by stopping. It is one that completes smoothly and fluently while managing risk. A robot that halts at the first sign of uncertainty is as problematic as one

that continues blindly: the goal is a system that is both safe and operationally effective. This balance between safety and fluency is one of the defining conceptual tensions of the dissertation.

The challenge is compounded by a well-documented weakness of VLA models in out-of-distribution conditions. Guo et al. (2025) found that even modest changes to camera viewpoints, lighting, or object layout produced steep performance drops, and identified the action modality, which is the robot's physical output, as the most fragile component under perturbation. When a VLA model encounters an unfamiliar situation, it does not reliably signal uncertainty or stop. It frequently continues executing, with apparent confidence, toward an incorrect action. In tool handover, where a human hand is already in motion toward the robot's gripper, this silent confident failure is not a performance metric. It is the precondition for a workplace injury.

This establishes a clear problem statement. Existing VLA-based robot control systems are insufficient for safe and fluent collaborative tool handover because they do not address robot-side uncertainty and human-side intent in an integrated way. The SAFE framework (Gu et al., 2025) demonstrated that VLA internal features carry enough information to detect impending task failure, but it was evaluated exclusively in scenes without any human present. The H2R Bridge framework (Wu et al., 2025) showed that vision-language models can be adapted for human intent recognition in data-scarce collaborative settings, but it does not monitor the robot's own reliability. A system monitoring robot confidence without reading the human's readiness has incomplete information. A system predicting human intent without assessing its own reliability is equally insufficient. Neither stream alone addresses both dimensions of the problem.

The research gap follows directly. No published work has integrated these two streams into a unified safety framework for collaborative tool handover, evaluated against clear ablation baselines in a human-proximate context. The gap is not that VLA failure detection and human intent recognition are unexplored. The point is that they have never been combined into a single architecture designed and tested specifically for the timing, precision, and mutual anticipation demands of tool handover. A simulation-based approach is the appropriate methodological choice at this stage: systematically inducing

VLA failure conditions in proximity to a real human participant is not ethically feasible within a six-month supervised research timeline, and the VLA community's established use of simulation benchmarks as primary evaluation venues provides clear methodological precedent.

The aim of this dissertation is to design, implement, and experimentally evaluate a dual-stream safety framework for VLA-controlled tool handover, and to determine through controlled simulation experiments whether integrating robot-side confidence monitoring with human-side intent recognition produces measurably safer and more fluent collaborative behaviour than either mechanism independently or than an unmonitored VLA baseline. Seven objectives operationalise this aim: reviewing the VLA failure detection and uncertainty estimation literature to ground the confidence monitor in published evidence; reviewing the industrial HRC intent recognition literature to ground the intent module design; designing and implementing a simulation environment incorporating OpenVLA and a parameterised human arm model; implementing a robot-side confidence monitor that extends the SAFE framework to human-proximate handover conditions; implementing a human-side intent recognition module targeting reliable prediction at or before fifty percent of trajectory completion, providing a minimum 300-millisecond lead time for response; designing the fusion layer and graduated response controller and conducting the ablation experiments; and analysing the findings against the research question, the ISO 25785-1 certification pathway, and the requirements for future hardware validation.

This study is guided by the following primary research question:

> *Does combining robot-side VLA confidence monitoring with human-side intent recognition produce measurably safer and more fluent tool handovers in simulated human-robot collaboration, compared to either mechanism deployed independently or to a VLA system with no monitoring at all?*

Three supporting questions refine the investigation: whether the confidence monitor maintains calibrated estimates under dynamic human co-presence; at what trajectory

completion percentage reliable intent prediction is achieved, providing the minimum lead time for safe robot response; and what the false positive cost of the framework is in nominal conditions, since a system that interrupts safe handovers too frequently will not survive real deployment.

The central hypothesis, stated prior to data collection, is that the integrated dual-stream system will outperform all three baselines (VLA alone, confidence-only, and intent-only) on the primary safety metric in failure conditions, while maintaining task completion rates within ten percentage points of the unmonitored baseline in nominal conditions, thereby demonstrating that safety and fluency need not be traded against each other. This hypothesis is subject to falsification: a result showing no measurable advantage over the best single-stream baseline would itself constitute a valid and publishable contribution.

The scope of this study is deliberately bounded. It is conducted entirely in simulation; no physical robot hardware is operated near a human participant. It examines a single task, robot-to-human tool handover, using OpenVLA fine-tuned on handover demonstrations and a parameterised kinematic human arm model. It does not capture the full variability of real human movement, does not address multi-worker or multi-robot scenarios, and does not claim compliance with ISO 25785-1 certification requirements. What it produces is simulation-validated evidence about the comparative value of the dual-stream architecture. It provides a rigorous methodological foundation for subsequent hardware validation rather than a claim of deployment readiness.

The study adopts a positivist, simulation-based experimental design. Four system configurations (unmonitored VLA, confidence-only, intent-only, and the full dual-stream system) are compared across four failure scenario groups, with quantitative metrics on task completion rate, safety intervention rate, false positive and negative rates, and handover fluency used to test whether the integrated system outperforms each baseline across both safety and fluency dimensions. Chapter One reviews the broader landscape of human-robot collaboration and VLA deployment. Chapter Two examines the specific technical literature on failure detection, intent recognition, and tool handover requirements. Chapter Three presents the full methodology. Chapter Four presents the experimental findings, analyses them against the research question and supporting

questions, and discusses implications for cooperative safety, the ISO 25785-1 certification pathway, and future hardware deployment. The Concluding Remarks synthesise the contribution and set out a concrete agenda for real-world validation.

CHAPTER ONE - LITERATURE REVIEW I

This chapter reviews the broad conceptual and theoretical landscape within which this dissertation is situated. It is organised around three themes, each building toward the same conclusion. The first examines human-robot collaboration in industrial environments, tracing the development of the field from early cobot deployments to the current state of close-proximity deployment, and identifying why the cooperative safety threshold remains unmet despite decades of progress. The second critically evaluates Vision-Language-Action models, examining both their capabilities and the structural limitations that create the silent confident failure problem at the heart of this dissertation. The third surveys the safety frameworks governing AI-driven robots near people, assessing what current standards and technical approaches cover and, critically, where they fall short. Together, these three themes establish that the gap this dissertation addresses is not one of incremental progress but of genuine conceptual integration: robot-side uncertainty monitoring and human-side intent recognition have developed as separate fields, and no existing work has brought them together for the specific and demanding case of collaborative tool handover.

1.1 Human-Robot Collaboration in Industrial Environments

Defining Human-Robot Collaboration

The term human-robot collaboration is used widely in both industry and academia, but it is often applied imprecisely to situations that more accurately represent coexistence or cooperation rather than genuine collaboration. This imprecision matters because the three levels carry fundamentally different safety implications. Coexistence describes scenarios in which humans and robots share a physical space while performing independent tasks. Cooperation describes shared tasks executed at separate times. Genuine collaboration requires both parties to work on the same task simultaneously in close physical proximity (Asif et al., 2026; Bouraine et al., 2025). Most industrial deployments documented in the peer-reviewed literature fall into the coexistence category at best. The field's own evidence, examined critically, reveals that what is widely marketed as collaboration frequently involves carefully managed physical separation rather than the mutual presence and shared action that the term implies.

This definitional gap has practical consequences. The literature on industrial HRC has developed increasingly sophisticated frameworks for managing safety in shared workspaces, but those frameworks were largely designed for coexistence and cooperation scenarios. Asif et al. (2026), in a comprehensive review of ninety-seven studies on human-robot collaborative systems, found that the central unresolved challenges in the field are not primarily technical in the narrow sense of actuation or sensing. They are challenges of safety, of trust, and of the psychological wellbeing of workers who must operate alongside unpredictable machines. This finding is significant because it suggests that technical solutions alone are insufficient. Real close-proximity collaboration requires frameworks that protect workers reliably and also earn worker trust through transparent behaviour. Existing systems address neither requirement fully.

Cobot Deployment in Industrial Settings

The growth of collaborative robot deployment over the past decade has been substantial. The bibliometric review of cobotics research by Haghighi et al. (2025) documents exponential growth in both publications and deployments, tracing the field from the foundational introduction of passive manipulators in the 1990s through the three operational modes now codified in ISO 10218-2:2025, which are hand-guided control, speed and separation monitoring, and power and force limiting. This evolution reflects genuine engineering progress. Yet the same literature reveals a more troubling pattern alongside this progress.

Pietrantoni et al. (2024), in a multi-country expert study of cobot integration across vehicle assembly, warehouse logistics, and agriculture, found that safety remained a paramount concern across all three sectors and that the safety protocols experts judged necessary were specific to each sector rather than common to the technology itself. The experts held generally favourable views of cobots, yet they tied safe deployment to sector-tailored protocols and to how the robot was integrated into the task rather than to the presence of safety hardware alone. Placing this alongside Asif et al.'s (2026) review yields a sharper conclusion than either paper offers on its own. The two converge on the view that safety outcomes in HRC are determined not by the technology in isolation but by how it is designed into the application. Speed and force limits, the primary safety

mechanisms in commercial cobots, reduce the severity of contact injuries when contact occurs, but they do not prevent the robot from making the wrong move in the first place. When the robot's action is the source of the hazard rather than its speed or proximity, reactive force-limiting mechanisms are simply not designed to intervene. This is precisely the failure mode this dissertation addresses, and the convergence of two independent reviews on the same diagnosis lends it weight beyond what any single source could provide.

The Gap between Cobot Deployment and True Collaboration

Despite the scale of collaborative robot deployment in 2025 and 2026, every documented commercial deployment is characterised by physical or temporal separation between robots and human workers. Far from being a temporary measure, this reflects a genuine technical gap that the field has named the cooperative safety threshold: the minimum capability level a robot system must demonstrate to be certified for unrestricted close-proximity operation alongside human workers (Bouraine et al., 2025; Samarathunga et al., 2025).

The ISO 25785-1:2025 standard for humanoid robot safety establishes the certification framework, while ISO 10218-2:2025 defines collaborative applications as those requiring both parties to occupy the same safeguarded space simultaneously, mandating application-specific risk assessment and validated safety functions before deployment can proceed (ISO, 2025a; ISO, 2025b). The 2025 revision of ISO 10218 is notable for its conceptual shift: it replaces the category of "collaborative robot" with the concept of "collaborative application", reflecting the consensus view that safety is not a property of the hardware but of how the hardware is used in a given context. For this dissertation, that shift is directly consequential: it means that any new class of collaborative task, including VLA-driven tool handover, requires its own dedicated safety validation, rather than simply inheriting the certification of the robot platform.

The peer-reviewed literature, synthesised across Asif et al. (2026), Pietrantoni et al. (2024), and Bouraine et al. (2025), identifies three requirements that must be met for a system to cross the cooperative safety threshold. The robot must detect its own uncertainty before a failure becomes a hazard, not after. It must understand the human

partner's intention and adjust its behaviour in anticipation. And it must do both in real time, within the timing constraints of the shared task. Current systems typically address at most one of these requirements. The contribution of this dissertation is a framework designed to address all three.

1.2 Vision-Language-Action Models in Robotic Control

VLA Model Architecture and Capabilities

Vision-Language-Action models represent a fundamental departure from the programming paradigm that dominated industrial robotics for decades. Where traditional robots require explicit programming of every task and every motion, VLA models learn to map visual inputs and natural language instructions to motor actions through training on large datasets of robot demonstrations. This learning-based approach enables generalisation, to varying degrees, to objects, environments, and instructions not seen during training, opening the possibility of robots that can be directed through natural speech rather than code. The practical appeal is considerable, and the pace of development has been rapid.

Kim et al.'s OpenVLA (2025) demonstrated that an open-source 7-billion parameter model, trained on approximately 970,000 robot episodes spanning twenty-two embodiments, could match or exceed proprietary systems on manipulation benchmarks while remaining accessible to researchers without specialised hardware. Black et al.'s π_0 (2024) extended the approach using flow-matching action generation, achieving higher-frequency control trajectories suited to dexterous manipulation. The recent comprehensive survey by Zhang et al. (2025) documents the scaling of these principles to whole-body humanoid control. The capability trajectory across these systems is genuinely impressive.

However, comparing these developments critically reveals an important pattern. Each new system extends VLA capabilities in a particular direction. One pushes higher frequency control. Another broadens generalisation. Another improves the efficiency of fine-tuning. The evaluations used to demonstrate progress, however, consistently rely on benchmarks designed for isolated robot manipulation in controlled environments. Across the systems reviewed by Zhang et al. (2025), none was evaluated in the presence of a

human co-worker performing a shared task. This is not an oversight in any individual paper. It reflects a structural gap in how the field evaluates progress.

VLA Performance Limitations and Generalisation Failure

The most significant technical limitation of VLA models for deployment in collaborative settings is their brittleness under distributional shift. Guo et al. (2025) found that even modest modifications to camera viewpoints, lighting conditions, object layout, and language instruction phrasing produced steep performance drops across multiple architectures. As Guo et al. (2025) and the survey by Zhang et al. (2025) both observe, this brittleness is not a property of any individual model. It reflects a structural feature of how VLA models learn. They build strong statistical associations between training-distribution inputs and correct actions. These associations do not generalise robustly to inputs that fall outside the training distribution.

Guo et al. (2025) examined this limitation specifically through the lens of multi-modal perturbations, varying inputs across visual, language, and proprioceptive modalities simultaneously. Their finding that the action modality is the most fragile under perturbation is particularly relevant to this dissertation. It means that the component of the system responsible for physical robot movement, which determines what the robot actually does, is precisely the one most likely to fail when deployment conditions diverge from training. This contrasts with the intuition, sometimes implicit in the field, that language understanding might be the bottleneck: Guo et al.'s evidence suggests the physical control layer is the more vulnerable component.

The benchmark saturation problem identified across the recent survey literature (Zhang et al., 2025) compounds this concern in a specific way. Recent work reports success rates above 99 percent on established simulation benchmarks such as LIBERO, which might appear to indicate that VLA manipulation is a largely solved problem. Yet as Zhang et al. (2025) observe in their comprehensive survey, this benchmark performance does not translate to real-world deployment reliability, and the gap between the two is growing rather than narrowing as benchmark scores approach ceiling values. The explanation is that current benchmarks measure performance on tasks closely resembling training conditions, while deployment conditions inevitably diverge. A model that scores 99

percent on LIBERO may still exhibit dramatic performance degradation on novel objects, unfamiliar spatial configurations, or, most relevantly for this dissertation, the presence of a moving human limb in the robot's visual field.

VLA Failure in Collaborative Environments

The specific consequence of out-of-distribution brittleness that is most dangerous in collaborative settings is what researchers have called the silent confident failure mode. When a VLA model encounters an out-of-distribution condition, it does not reliably signal uncertainty or halt. It frequently continues executing toward an incorrect action with the same smooth, confident motion it would exhibit in a successful trial. This behaviour arises because VLA models are trained to produce actions, not to evaluate whether their actions are appropriate for the current situation. The absence of an explicit uncertainty signal means that external monitoring systems receive no indication that intervention is warranted.

This failure mode is categorically more dangerous in collaborative settings than in isolated manipulation. In a solo robot task, a silent confident failure produces a wrong action that can be observed and corrected without physical risk to any person. In a tool handover scenario, where the human partner is already committed to a grasping trajectory, a silent confident failure on the robot's side creates a configuration in which both parties are simultaneously committed to incompatible actions in overlapping space. The SAFE framework (Gu et al., 2025) represents the most direct attempt to address this problem, demonstrating that VLA internal feature representations carry detectable signals of impending failure across unseen manipulation tasks. However, as Gu et al. acknowledge, the framework was evaluated entirely in human-absent manipulation scenarios. Whether the confidence signals remain calibrated and informative when a human body is present in the visual field is an open empirical question that no published study has yet addressed. This question forms the basis of the first supporting research question of this dissertation.

1.3 Safety Frameworks for AI-Driven Robots

Safety Standards and Regulations

The governance of collaborative robot safety in industrial environments is primarily defined by the ISO 10218 series, which underwent major revision in 2025. ISO 10218-2:2025 replaced both the 2011 version of the standard and the previously separate ISO/TS 15066 specification for collaborative applications, consolidating requirements for robot cells, applications, and human interaction into a single framework (ISO, 2025a). The conceptual significance of this consolidation is substantial. The 2025 standard eliminates the category of 'collaborative robot,' replacing it with the concept of 'collaborative application', reflecting the mature understanding that safety is not a property of the hardware alone but of the specific combination of robot, task, environment, and human interaction. For the purposes of this dissertation, this shift means that VLA-driven tool handover constitutes a new class of collaborative application requiring its own dedicated safety evaluation, and that existing robot certifications do not automatically extend to it.

The standard defines four modes of collaborative operation. These are monitored standstill, hand-guided control, speed and separation monitoring (SSM), and power and force limiting (PFL). Of these, SSM and PFL are most relevant to tool handover. SSM dynamically reduces robot speed as the human's proximity increases. PFL limits the forces the robot can exert at contact, using biomechanical thresholds across body regions specified in the standard (ISO, 2025a). Both mechanisms are reactive in a meaningful sense: they respond to proximity and contact rather than anticipating the quality of the robot's planned action. What neither mechanism addresses is the scenario in which the robot is executing a correctly-scoped motion at an appropriate speed toward the wrong target. In that scenario, both SSM and PFL may allow the action to proceed because the robot is not moving too fast and has not made contact, even though the action itself is hazardous.

ISO 25785-1 (2025), the humanoid robot safety standard, adds requirements relevant to bipedal platforms including fall mitigation and compliant interaction force limits (ISO, 2025b). At the regulatory level, the EU AI Act classifies autonomous robots in shared

workspaces as high-risk AI systems, requiring technical documentation, conformity assessment, and explicit human oversight mechanisms (Zhang et al., 2025). The graduated response architecture proposed in this dissertation ranges from full autonomous execution through slowed execution, workspace yield, and halt-and-request. It is architecturally aligned with the oversight requirements of the EU AI Act, though full regulatory compliance would require validation beyond the scope of this study.

Technical Safety Approaches and Their Limitations

The technical literature on collaborative robot safety can be organised around three generations of approach, each addressing a different aspect of the safety problem and each revealing what the previous generation left unaddressed. Drawing primarily on the frameworks synthesised by Asif et al. (2026) and the systematic review by Khan et al. (2026), this generational analysis reveals a clear progression from reactive to predictive to proactive safety, and identifies the proactive layer as the one that current research has not fully developed.

The first generation of approaches centred on reactive physical safety. It used mechanical compliance, force-torque sensors, and collision detection that halted the robot when contact was detected. These mechanisms remain standard in commercial cobot platforms and continue to be required by ISO 10218-2:2025. These mechanisms have a fundamental limitation. They address the consequences of failure rather than its preconditions. By the time a collision is detected and the robot halted, physical contact between the robot and the human has already occurred. For tool handover, where the human's hand may be within centimetres of the robot's end effector at the moment of transfer, reactive stopping is simply not fast enough to prevent injury from a sudden wrong motion.

The second generation introduced predictive separation monitoring. Such systems continuously estimate the distance between the robot's trajectory and any detected human body part, and reduce robot speed in anticipation of potential contact rather than in response to it. This approach, now formalised in ISO 10218-2:2025's speed and separation monitoring mode, represents genuine progress. Asif et al. (2026) note an important limitation. Predictive separation monitoring tracks where the robot is going

relative to where the human is, but does not evaluate whether the robot's destination is the correct one. A robot moving at reduced speed toward the wrong object, or initiating a handover while the human has not yet prepared to receive, would pass the SSM check while still executing an unsafe action.

The third and most recent generation of approaches attempts to address this residual gap by monitoring the quality of the robot's own decision-making. The SAFE framework (Gu et al., 2025) demonstrated that internal VLA feature representations carry signals distinguishing impending failure from successful execution across multiple unseen manipulation tasks, establishing that runtime failure detection is technically feasible for VLA-based systems. The confidence-aware failure recovery framework (Banerjee et al., 2026) extended this to propose graduated human intervention proportional to the system's estimated confidence, reducing unnecessary stoppages in safe conditions while maintaining intervention when uncertainty is genuinely high. The systematic review by Khan et al. (2026) identifies this line of work as one of the most promising emerging directions in HRC safety research.

Yet a critical limitation is shared across all three generations of work reviewed above. Each generation was developed and evaluated for solo robot manipulation, without a human present as a collaborative partner. The systematic review by Khan et al. (2026) specifically identifies the integration of AI-based safety monitoring with real-time human presence as an unresolved challenge. The performance of confidence monitors under human co-presence has not been evaluated. The false positive cost of graduated intervention systems in genuine collaborative conditions has not been measured. And the question of whether combining robot-side confidence monitoring with human-side intent recognition produces better outcomes than either alone has not been investigated. These are the questions that Chapter Two examines in the specific technical detail that motivates the design of this dissertation's experimental study.

Chapter Summary and Research Gap

The three themes reviewed in this chapter converge on a single, clearly bounded conclusion. The HRC literature has identified the cooperative safety threshold and the requirements for crossing it, but current industrial deployments have not crossed it. The

VLA literature has produced impressive manipulation capabilities alongside documented brittleness under distributional shift and a specific, dangerous silent confident failure mode. The safety frameworks literature has progressed from reactive to predictive to early proactive monitoring, but has evaluated each generation in isolation, without a human collaborator present, and without combining the robot-side and human-side monitoring streams that this dissertation proposes to integrate. Chapter Two examines those two streams in technical depth, taking VLA failure detection on one side and human intent recognition on the other, and makes the specific argument for why their integration in a dual-stream framework, tested against clear ablation baselines in a human-proximate tool handover scenario, is the contribution that the field requires and that the existing literature does not yet provide.

CHAPTER TWO - LITERATURE REVIEW II

Chapter One established the broad landscape. Industrial human-robot collaboration has not yet crossed the cooperative safety threshold. VLA models exhibit a specific and dangerous silent confident failure mode. And the three generations of safety frameworks have each been developed in isolation from a human collaborator. This chapter performs the close technical work that Chapter One signposted. Its purpose is to show the exact gap this dissertation addresses by examining, in direct comparison rather than sequential description, the two technical literatures from which the proposed framework draws its components and the third literature that defines the evaluation task. The first section reviews runtime failure detection for robot learning policies, with particular attention to VLA confidence monitoring. The second reviews human intent recognition in industrial collaboration, focusing on the early prediction problem and the specific demands of providing a usable lead time for robot response. The third reviews the tool handover literature, establishing it as the canonical close-proximity test case and grounding the design choices of the experimental protocol in documented prior work. The chapter ends by stating the gap precisely. Each of these literatures has been explored on its own. What has not been done is to integrate them in a dual-stream architecture and test that architecture against its single-stream variants in a tool handover scenario with a human present.

2.1 Runtime Failure Detection for VLA and Generative Robot Policies

What Failure Detection Has to Achieve

Before comparing methods, it is worth being precise about what runtime failure detection has to do for a collaborative system. Three properties matter. First, the detector must be calibrated: a confidence score of, say, 0.9 must correspond to a roughly nine-out-of-ten probability of success, otherwise downstream decisions based on the score are arbitrary. Second, it must be timely: a signal that arrives after the failure has already manifested has no value for prevention. Third, it must generalise: a detector trained on one set of tasks must remain useful on the unseen tasks that VLA policies are deployed to handle, because the whole point of using a generalist policy is that the deployment task is not

known in advance (Gu et al., 2025; Xu et al., 2025). These three properties define the criteria against which the existing methods are evaluated below.

SAFE: Failure Detection from VLA Internal Features

The SAFE framework, presented at NeurIPS 2025 (Gu et al., 2025), is the most direct precursor to the robot-side stream of this dissertation. The central insight is empirical. VLA internal features, extracted from the last layer of the model during rollout, carry sufficient high-level information about task success and failure to support detection across tasks not seen during the detector's training. The authors visualise this in feature space and show that successful and failed rollouts cluster in different regions of the latent representation, regardless of which task the VLA was attempting. SAFE then trains a small MLP or LSTM classifier on these latent features and predicts a single scalar that represents the likelihood of impending failure, calibrated using conformal prediction to provide statistical guarantees on the false alarm rate.

The contribution is significant for three reasons. It establishes that VLA models are not, contrary to the criticism implicit in Chapter One's account of silent confident failure, completely uninformative about their own reliability. The information is present in the latent space, even if the action output does not expose it. It demonstrates that this information transfers across tasks, addressing the multitask generalisation criterion above. And by using conformal prediction it gives the system a principled false alarm rate, which is essential for any system that will trigger interventions during real operation. Yet two of SAFE's own limitations bear directly on this dissertation. The authors acknowledge the first themselves. SAFE was evaluated only on manipulation tasks in human-absent settings, with no evidence that the calibration generalises to scenes containing a moving human body. The second follows from the first. Without a human in the scene, the cost structure of false positives and false negatives is artificial. A false halt in solo manipulation costs only time. In collaboration, the same false halt potentially erodes worker trust and disrupts the rhythm of joint work.

FAIL-Detect: Failure Detection Without Failure Data

Xu et al.'s FAIL-Detect, published at Robotics: Science and Systems 2025, takes a complementary approach. Rather than learning from failure data, FAIL-Detect frames the

problem as sequential out-of-distribution detection and uses only successful rollouts to calibrate its confidence threshold. The system distils policy inputs and outputs into scalar signals that capture epistemic uncertainty and applies conformal prediction to convert these signals into a calibrated alarm. The authors demonstrate that learned signals, in particular a flow-based density estimator, outperform post-hoc statistics across a range of imitation learning policies, and that the method detects failures more accurately and earlier than several state-of-the-art baselines (Xu et al., 2025).

The two methods can be read against each other usefully. They agree on three points. Internal model signals carry failure-relevant information. Conformal prediction is the right calibration tool. And the scalar output of the detector is more useful than a binary verdict. They differ in what they assume about training data. SAFE assumes that some failure rollouts are available and can be used to train the detector. FAIL-Detect assumes none are, and works from successful trajectories alone. For a dissertation evaluating a confidence monitor in a new setting, human-proximate tool handover, this distinction matters. Failure data in handover is precisely what is hard to collect. Deliberately staging VLA failure conditions next to a real human is not ethically permissible at the dissertation stage. A method that can be calibrated from successful simulation rollouts has a direct practical advantage. This dissertation's confidence monitor draws on both works. From SAFE it takes the evidence that VLA latent features encode failure information. From FAIL-Detect it takes the demonstration that calibration is possible without failure data.

Confidence-Aware Human-in-the-Loop Recovery

Banerjee et al. (2026), published at the ACM/IEEE International Conference on Human-Robot Interaction, address a question that neither SAFE nor FAIL-Detect engages with directly: what should the system do once a confidence signal has been computed? Their framework couples calibrated module-level uncertainty estimates to an explicit query policy that decides whether to act autonomously, query the human, or pause. The contribution is conceptually important because it reframes confidence monitoring as one input to a decision problem that also includes the cost of bothering the human partner. Both extremes fail in practice. Query the operator too often and they stop paying

attention. Query too rarely and the system eventually commits to a wrong action. Where the right threshold sits depends on more than just the confidence score. It also depends on how disruptive the query itself is.

Banerjee et al.'s work was developed in the context of robot-assisted bite acquisition for users with mobility limitations, not industrial handover, and the cost model they use is specific to that domain. Yet the conceptual contribution generalises directly. The graduated response controller proposed in this dissertation has four levels, ranging from full execution through slowed execution and workspace yield to halt-and-request. It is architecturally similar to Banerjee et al.'s separation of detection from response, but applied to a different domain and combined with a second monitoring stream. Where their framework asks "how should the robot escalate when its own confidence drops?", this dissertation asks the same question while also asking "how should the robot escalate when the human partner is committed to a reach toward an unsafe configuration?". Adding the second question is what makes the architecture dual-stream rather than single-stream.

Section Summary

Across the three works compared above, a clear picture emerges. The technical machinery for runtime failure detection in VLA and generative policies exists, is increasingly well calibrated, and is beginning to be coupled to graduated response policies. None of it has been evaluated in a setting with a moving human in the visual field. This is the basis of the first supporting research question of this dissertation, which asks whether the confidence monitor maintains calibrated estimates under dynamic human co-presence. The question follows directly from what the literature has and has not yet demonstrated.

2.2 Human Intent Recognition in Industrial Collaboration

Three Families of Approach

The literature on human intent recognition in industrial HRC, as reviewed comprehensively by Kekana et al. (2025), can be grouped into three families. Rule-based and probabilistic methods. Classical machine learning methods. And deep learning

methods that draw on recurrent and attention-based architectures. The grouping matters because the three families address different parts of the same underlying problem and have different strengths in the specific context of tool handover.

Rule-based and probabilistic methods, including hidden Markov models and dynamic Bayesian networks, were the dominant approach through the 2010s and remain in use where the set of possible human actions is small and well defined. Their advantage is interpretability. The limitation, identified explicitly by both Kekana et al. (2025) and the Annual Review survey by Hoffman et al. (2024), is that they require explicit modelling of every action they are expected to recognise. They degrade rapidly when the human deviates from the expected motion library. Classical machine learning methods loosen this requirement by learning the mapping from observation to intent label from data. Deep learning methods, particularly recurrent architectures such as LSTMs and more recently transformer-based models, loosen it further by handling long temporal dependencies and noisy partial observations directly.

Early Prediction and the Lead-Time Constraint

For collaborative tool handover, the relevant performance measure is not whether the system eventually classifies the human's intent correctly but whether it does so early enough for the robot to respond. This is the early prediction problem. The ASME paper by Zhang et al. (2024) is the most precise treatment of the question available in the peer-reviewed literature. The authors evaluate transformer and LSTM architectures on the task of predicting human intent from partial trajectory data and report performance as a function of how much of the trajectory has been observed. Their finding, consistent with the broader survey evidence in Kekana et al. (2025) and Hoffman et al. (2024), is that reliable prediction is achievable at or before fifty percent of trajectory completion when the architecture has access to skeletal keypoint data with sufficient temporal context.

Why fifty percent specifically? The figure tracks a physical constraint. On a typical tool handover trajectory of one to two seconds, it corresponds to a lead time of roughly three hundred to five hundred milliseconds before the human hand reaches the transfer point. This window is significant because it matches the minimum time required for a robot arm to decelerate from a standard handover speed to a safe stopping configuration, as

documented in the dynamic speed-and-separation monitoring literature (Asif et al., 2026). A prediction that arrives at thirty percent of trajectory completion provides comfortable margin. One that arrives at seventy percent arrives too late to act on. The intent recognition module proposed in this dissertation therefore takes the fifty-percent figure as a hard target rather than a desirable outcome, because below that threshold the predictions become operationally useless regardless of their accuracy.

Recent Deep-Learning Architectures

Among the deep learning approaches reviewed by Kekana et al. (2025), two architectural patterns are particularly relevant. The first is the LSTM-based skeletal trajectory classifier, which has been adapted for industrial HRC in work such as Gao et al. (2023) and the Frontiers in Robotics and AI study by Mavsar et al. (2025), which integrates LSTM-based intent recognition with dynamic movement primitives for real-time robot adaptation. The second is the transformer-based classifier, which the same studies show can outperform LSTM variants when sufficient training data is available, at the cost of higher computational requirements. Each architecture has a different weakness. LSTMs are easier to deploy in real-time settings and hold up reasonably well with limited data, but transformers achieve higher peak accuracy when training data is plentiful, at the cost of longer inference times.

A third, more recent line of work attempts to leverage vision-language models for intent recognition in data-scarce industrial settings. The intuition is that the semantic richness of pre-trained vision-language models provides a useful prior even when only a small number of HRC-specific demonstrations are available. The challenge, as the comparison study by Mavsar et al. (2025) makes clear, is that general-purpose vision-language models are designed for high-level reasoning over text and image, not for the continuous millisecond-resolution classification of partial hand trajectories that the lead-time constraint requires. Mavsar et al. conclude that domain-specific lightweight classifiers remain the right tool for real-time intent recognition in collaborative settings, with vision-language models playing a supporting role in higher-level task understanding rather than driving the time-critical signal.

Persistent Challenges in Intent Recognition

Kekana et al.'s (2025) review identifies five challenges that remain unresolved across all families of approach. These are usability under real deployment conditions, robustness to sensor noise and partial occlusion, readiness for continuous industrial operation, real-time processing constraints, and generalisation across different workers, tasks, and operational contexts. Every one of these challenges is present in tool handover, and the last of these (generalisation across workers) is particularly acute because human reach kinematics vary by body size, dexterity, fatigue, and individual style. The intent module proposed in this dissertation does not claim to solve all five. It adopts a parameterised kinematic human arm model precisely so that the simulation can sweep across the variability that real workers would exhibit, providing a controlled foundation on which subsequent hardware studies can build.

2.3 Tool Handover and the Dual-Stream Argument

Why Tool Handover Is the Right Test Case

Tool handover has a long history as the canonical task for studying close-proximity human-robot collaboration. Handover packs everything that makes joint action difficult into a few seconds. Two people have to coordinate in space and time their movements together. They have to read each other's readiness. They have to manage the grip exchange so the object does not fall. Few collaborative tasks compress that much into so short an interaction. The comprehensive review of object handovers by Ortenzi et al. (2021), published in IEEE Transactions on Robotics, established the conceptual vocabulary still in use: the giver and receiver each play active and passive roles at different phases, the transfer point is determined jointly rather than imposed by either party, and fluency is measured through the proportion of concurrent activity, idle times, and functional delays during the interaction.

Tool handover is a particularly demanding subset of object handover because tools are often pointed, sharp, or weighted asymmetrically, and their orientation at the moment of transfer determines whether the handover is safe or hazardous. Parastegari et al. (2018), in their IEEE Transactions on Robotics paper on failure recovery in robot-human object handover, demonstrated experimentally that a human giver primarily relies on vision

rather than haptic sensing to detect the fall of an object during transfer. This is directly relevant to the design of the visual confidence monitor in this dissertation: it provides empirical justification for the choice to base failure detection on the same modality that humans themselves rely on in the same situation.

Recent Advances in Handover Coordination

Two recent peer-reviewed works frame the state of the art on handover coordination and reveal the gap that motivates the dual-stream argument. Penzotti and Controzzi (2025), in the International Journal of Social Robotics, show that combining robot proprioception with observation of the human partner's kinematics significantly improves the timing of object release and the perceived fluency of the handover. Meng et al. (2024), in Cyborg and Bionic Systems, present a mobile robot handover system that adapts the robot's grasp and trajectory to the real-time pose of the human hand, achieving handover completion within an eight-second target across thirty trials per object. Penzotti and Controzzi focus on the moment of release and the fine-grained coordination of grip force, while Meng et al. focus on approach planning and visual hand tracking. Both are demonstrably effective on their own terms. Both treat the robot's own competence as a given and devote their architectural attention to reading the human partner.

This last observation is the architectural critique that motivates this dissertation. The handover literature has invested heavily in reading the human partner and has produced increasingly sophisticated methods for doing so. It has also, in the SAFE-FAIL-Detect-Banerjee line of work reviewed in Section 2.1, begun to invest seriously in monitoring the robot's own competence. What it has not yet done is connect these two lines of investment. The first failure mode is well known. A system that reads the human partner well but does not monitor its own confidence will, when the VLA encounters out-of-distribution conditions, continue confidently toward a wrong configuration even as the human reaches into the workspace. The second is just as concrete. A system that monitors its own confidence but does not read the human partner will trigger conservative interventions even when the human is approaching in a perfectly predictable way. Both failure modes follow directly from the structure of the systems described in the literature.

The Dual-Stream Architecture

The dual-stream architecture proposed in this dissertation is the structural answer to that observation. The first stream draws on SAFE, FAIL-Detect, and the confidence-aware recovery framework of Banerjee et al. to monitor the robot's own reliability through VLA latent features and conformally calibrated alarms. The second draws on the transformer- and LSTM-based skeletal classifiers reviewed by Kekana et al. (2025) and Mavsar et al. (2025) to predict the human partner's reach intent from partial trajectory data, targeting the fifty-percent completion threshold identified by Zhang et al. (2024). The fusion layer combines the two streams into a graduated response controller with four levels of action. Full execution when both streams agree the situation is nominal. Slowed execution when one stream signals mild uncertainty. Workspace yield when either stream signals strong concern. And halt-and-request when both streams signal high uncertainty at the same time. The architecture is explicitly designed so that each component can be evaluated against the others in ablation, which is what makes the comparative experimental design of Chapter Three possible.

Why Simulation Is the Right Methodological Choice

The works reviewed in this chapter divide into those evaluated in pure simulation, those evaluated on physical robots in laboratory settings, and a small number that include real-world workplace pilots. For a dissertation that proposes to deliberately stage VLA failure conditions in proximity to a human, the laboratory and workplace categories are not available within an ethics-approved six-month timeline. Simulation is the right methodological choice at this stage, for the same reason the SAFE and FAIL-Detect papers used it. Staging VLA failures next to a real human is not safe enough to do in a six-month project. Simulation lets the same conditions be tested repeatedly under controlled parameters with enough trials for statistical power. The contribution of this dissertation is not a deployment-ready system but rigorous simulation-validated evidence about whether the dual-stream architecture outperforms its single-stream and unmonitored baselines, providing the methodological foundation on which subsequent hardware validation can build.

Looking Ahead to Chapter Three

Taken together, the three sections of this chapter set out a concrete brief for the methodology that follows. They show that a credible test of the dual-stream architecture has to do six things. It has to build a simulation environment in which a VLA policy can be run repeatedly under controlled conditions with a parameterised human present in the workspace. It has to implement four system configurations so that the dual-stream framework can be compared against an unmonitored VLA baseline, a confidence-only baseline, and an intent-only baseline. It has to introduce failure scenario groups that probe the out-of-distribution conditions the literature has shown VLA models are most fragile against, from camera viewpoint changes to novel tool geometries to unexpected human approach trajectories. It has to calibrate the confidence monitor using methods drawn from SAFE and FAIL-Detect, without relying on failure data that cannot be collected ethically at the dissertation stage. It has to target the fifty-percent trajectory completion threshold for intent prediction that the early prediction literature has established as the minimum useful lead time. And it has to measure safety and fluency together, using completion rate, intervention rate, false positive rate, false negative rate, and intent prediction accuracy as a function of trajectory completion, so that the question of whether the integrated system genuinely outperforms its baselines can be answered with statistical confidence rather than impression. Chapter Three turns these requirements into the experimental protocol.

CHAPTER THREE - METHODOLOGY

3.1 Research Philosophy and Design Logic

This chapter describes how the dual-stream framework will be tested, what data the test will produce, and how that data will be interpreted. Before the operational detail, it is worth being clear about the kind of study this is.

The research follows a positivist philosophy. Saunders, Lewis and Thornhill (2023) describe positivism as the view that knowledge is built from observable, measurable facts gathered under controlled conditions and analysed by methods other researchers can repeat. For a study that asks whether one safety architecture performs better than another, this is the natural fit. The question is empirical. It is settled by running the systems, measuring what they do, and comparing the numbers.

The design is a controlled experiment with four conditions. Each system configuration introduced in Chapter Two faces the same failure scenarios in the same simulated environment, and the results are compared. Because the configurations differ only in which monitoring streams are active, any difference in outcome can be attributed to the architecture rather than to some other variable. This is the standard logic of an ablation study, the same logic Gu et al. (2025), Xu et al. (2025), and Banerjee et al. (2026) used to evaluate the individual components reviewed in Chapter Two. The novelty is not the method. It is applying that method to a setting where a human is present and where two monitoring streams are compared against each other and against their combination.

Running the study in simulation is a deliberate choice. Workplace pilots produce the most externally valid results but are not feasible in a six-month project. Laboratory studies with physical robots require hardware and ethics approval for human participants. Simulation has the lowest external validity, but it has the highest internal control, allows the same conditions to be repeated many times, and is the only ethically defensible way to induce VLA failure conditions near a human-shaped object. For a research question about comparative architecture performance rather than deployment readiness, simulation is the right level of fidelity.

This bounds the contribution. The dissertation does not claim the dual-stream framework is ready to deploy on hardware near real workers. It claims that under controlled simulated conditions, with a defined human kinematic model and a fixed set of failure scenarios, the architecture either does or does not outperform its single-stream and unmonitored baselines. That is the foundation on which later hardware studies can build, and it is what this study can actually support.

3.2 Simulation Environment Design

The simulation is built in PyBullet, an open-source physics engine widely used in robot learning. It was chosen for three reasons. It runs on modest hardware with a mature Python API that integrates with the rest of the experimental code. And it is the simulator Gu et al. (2025) used to evaluate the SAFE framework, so the calibration methods drawn from that work transfer without translation problems. A higher-fidelity option such as Isaac Sim would render better visuals, but visual realism is not what the framework is tested on. What matters is that contact physics, motion timing, and arm geometry are faithful enough to expose failures of the monitoring streams.

The robot is a seven-degree-of-freedom Franka Emika Panda, a standard collaborative arm used across the handover literature in Chapter Two. The PyBullet Panda model has accurate joint limits, mass properties, and end-effector geometry, so simulated trajectories correspond to ones a real arm could execute. The gripper is a parallel-jaw model with an eight-centimetre maximum width and a three-kilogram payload limit.

The policy is OpenVLA in its pre-trained form from Kim et al. (2025), with no fine-tuning in the first pass. The pre-trained model already holds a strong pick-and-place prior from the Open X-Embodiment dataset, and one empirical question this study can answer is whether the framework adds value to a policy not specialised for handover. Fine-tuning OpenVLA-7B also needs substantial cloud GPU resources, and committing those before knowing baseline performance would be premature. If the pilot shows pre-trained OpenVLA cannot reach a sixty percent completion rate on nominal trials, fine-tuning on a small synthetic demonstration set is added before the full experiment runs.

The human partner is a parameterised kinematic arm model following the seven-degree-of-freedom structure of Klopčar and Lenarčič (2005), the standard reference for the reachable workspace of a human arm. Joint angle ranges follow the values that paper reports from optical measurements of healthy adults. Link lengths are parameterised by user height using standard anthropometric ratios, sampled across a height range of one hundred and fifty-five to one hundred and ninety centimetres so the simulation spans most adult workers.

The hand follows a minimum-jerk reach trajectory toward the transfer point. Minimum-jerk reaching is the canonical model of human point-to-point arm motion, established by Flash and Hogan (1985) and confirmed many times since, with its characteristic bell-shaped velocity profile. The simulation uses it because it is the signal the intent recognition module has to predict from partial trajectory data. An idealised reach makes that prediction cleaner than real human motion, which adds noise and submovements. This is a deliberate simplification, and Chapter Four returns to it as a limitation and a target for future hardware validation.

The workspace is a flat table one metre wide, with the Panda at one end and the simulated arm reaching from the other. Three tools sit at fixed starting positions. A screwdriver fifteen centimetres long with a sharp tip and textured handle. A rubber mallet of one point two kilograms with a wooden handle and rubber head. A pencil seventeen centimetres long and seven millimetres in diameter. The three span the dimensions Chapter Two identified as critical for safe handover, which are orientation control for sharp objects, grip and release timing for heavy objects, and precision for small objects. The convention is handle-first, following the human norm documented by Ortenzi et al. (2021) and Parastegari et al. (2018). The robot presents the screwdriver tip away from the human, the mallet head down, and the pencil tip away from the receiving hand. Section 3.4 includes scenarios where these conventions are violated, which is the out-of-distribution behaviour the confidence monitor must detect. ROS2 handles communication between the simulation, the policy, the two streams, and the response controller, and the resulting message log is the data record for each trial.

3.3 The Four Experimental Conditions

The design compares four configurations that differ only in which monitoring streams are active. All four use the same OpenVLA policy, the same environment, the same arm model, and the same three tools. Together they form the ablation set needed to test whether the dual-stream framework outperforms its parts.

The first condition is the unmonitored VLA baseline. OpenVLA receives the visual input and the instruction, produces an action, and executes it, with no confidence monitoring and no intent recognition. It represents current state-of-the-art VLA deployment as documented by Zhang et al. (2025) and is the floor any monitoring framework must beat to justify its complexity.

The second condition is confidence-only monitoring. The policy runs as before, but its action is gated by a confidence monitor based on the SAFE framework of Gu et al. (2025). The monitor extracts features from the last hidden layer of the VLA at each timestep, passes them through a small classifier trained on successful rollouts, and outputs a calibrated confidence score. Below a threshold the robot slows or halts through the response controller. Above it the action proceeds. This tests whether monitoring the robot alone is enough.

The third condition is intent-only monitoring. The policy runs without confidence monitoring, but the human arm trajectory is observed by an intent recognition module based on the LSTM and transformer approaches reviewed by Kekana et al. (2025) and the early prediction work of Zhang et al. (2024). It takes the partial hand trajectory and outputs a predicted reach target with a confidence value. If the prediction signals an unsafe reach, or is itself low-confidence, the response controller intervenes. This tests whether reading the human alone is enough.

The fourth condition is the full dual-stream system. Both streams run in parallel, and their outputs are combined by a rule-based fusion layer using the threshold logic Banerjee et al. (2026) apply in their human-in-the-loop framework. When both streams report a nominal situation the robot runs at full speed. When one reports moderate uncertainty it slows to half speed. When either reports strong concern it yields the workspace. When

both report high uncertainty at once it halts and requests confirmation. The four levels match the graduated controller described in Chapter Two.

Two design choices matter. The fusion is rule-based rather than learned, because a learned classifier would need failure data to train on, and Chapter Two established that such data cannot be collected ethically at this stage. The rule-based approach avoids that and stays fully interpretable, which the safety certification argument in Chapter Four depends on. The two stream thresholds are set during the pilot from successful trials only, following the conformal prediction procedure of Angelopoulos and Bates (2023), then held fixed. The dual-stream condition is therefore not tuned on the test data, which would inflate its apparent performance. All four conditions run on the same trials with the same random seeds, so outcome differences trace to the active streams rather than to which scenarios each condition met.

3.4 Experimental Protocol and Data Collection

The experimental matrix has four conditions and four failure scenario groups, with fifty independent trials per cell, giving eight hundred trials in total. The four groups correspond to the out-of-distribution conditions Chapter Two identified as the most relevant VLA failure modes. Camera viewpoint shift moves the simulated camera ten to twenty centimetres in height or lateral offset and rotates it up to fifteen degrees. Lighting variation shifts the light source in colour temperature and intensity to create glare or low-contrast shadows. Novel tool geometry introduces tool variants the model has not seen in pre-training, such as a longer screwdriver shaft or a heavier mallet head. Unexpected human approach has the simulated hand reach from a non-standard angle or begin before the robot has positioned the tool. Each group probes a distinct failure documented by Guo et al. (2025) and Zhang et al. (2025).

A nominal scenario group is added as a control, with standard camera position, uniform lighting, a standard tool, and a typical reach. It measures the false positive rate of the monitoring streams. A safety architecture that intervenes too often in normal handovers is not useful in practice, so this control is as important as the failure groups.

Each trial starts with the robot at a fixed home position and one tool on the table. The robot receives an instruction such as “hand the screwdriver to the person” and begins its

planned action. The simulated human, after a delay drawn uniformly between zero and one second, starts its reach. A trial ends when the handover completes, defined as the tool being released into the hand within an eight-second window, when the response controller triggers a workspace yield or halt that ends the trial, or when the time limit is reached with neither outcome.

The fifty-trials-per-cell figure is grounded in power analysis for human-robot interaction experiments. Bartlett et al. (2022), drawing on Cohen's framework, show that detecting a medium effect at a 0.05 significance level with power of 0.8 needs roughly fifty observations per cell for pairwise comparison. The study is designed to detect medium-sized differences between the dual-stream condition and the single-stream baselines, the effect range that matters for any practical safety claim. Smaller effects may exist but would not be claimable here, and the dissertation is honest about that boundary.

Reproducibility is supported in three ways. Each trial uses a deterministic seed derived from its index, so re-running reproduces the same trajectories. The simulation, the policy version, the classifier weights, and the thresholds are version-controlled. And the experimental script, configuration files, and seed list are archived in the public repository recorded in the Data and Code Availability statement, so others can verify the results. For every trial the logged data covers the condition label, scenario group, and seed, the full robot and human trajectories at fifty hertz, the OpenVLA action at each timestep, the Stream One confidence and Stream Two intent prediction when active, the response controller decision at each timestep, the outcome label, and the timing of each handover phase. This supports every analysis in Section 3.5 without re-running the simulation.

3.5 Evaluation Metrics and Statistical Analysis

The framework is tested on two dimensions, safety and fluency, because Chapter Two showed a gain on one can come at a cost on the other. A system that halts too readily is not safe, because workers stop trusting it and work around it. A system that intervenes too rarely is not safe either. The metrics capture both, so the analysis can say not only whether the framework improves on the baselines but whether it does so without an unacceptable cost on the other dimension.

The primary safety metric is the safety intervention rate in the failure groups, the proportion of failure trials in which the controller triggered a yield or halt before the unsafe action completed. A higher rate there is good. The complementary metric is the false positive rate in the nominal group, the proportion of nominal trials with an unnecessary intervention. A higher rate there is bad. Together they show whether the framework is genuinely safer or merely more cautious.

The fluency metrics come from Ortenzi et al. (2021), whose review set the vocabulary for handover quality. Idle time is the share of the handover in which neither party moves, capturing hesitation. Concurrent activity time is the share in which both move at once, capturing coordination. Functional delay is the time between the robot becoming ready to release and the hand reaching the transfer point. A fluent handover has low idle time, high concurrent activity, and low functional delay. These describe the texture of the interaction, not just whether it succeeded.

The diagnostic metric for the intent stream is the intent prediction accuracy curve, the percentage of trials where the predicted reach target matched the actual one, broken down by how much of the trajectory had been observed when the prediction was made. This is the form Zhang et al. (2024) report, and it directly tests the second supporting research question, which asks at what trajectory completion reliable prediction becomes possible. The target is fifty percent or earlier, following the lead-time analysis in Chapter Two. Two further metrics are tracked. Overall task completion rate across nominal and failure scenarios, and median handover duration for completed trials, both needed to confirm that safety gains do not break handovers that would otherwise have succeeded.

The statistical analysis follows the matrix structure. Continuous metrics, including idle time, functional delay, intent accuracy, and duration, are compared with two-way ANOVA using condition and scenario group as factors, with post-hoc pairwise tests using Tukey’s HSD and Bonferroni correction. Binary outcomes, including whether a safety intervention occurred and whether the trial completed, use chi-square tests, with Fisher’s exact test where expected cell counts fall below five. The threshold is 0.05 for all tests, and effect sizes are reported alongside p-values, following Bartlett et al. (2022) on reporting HRC results without over-claiming from significance alone.

The interpretation criterion is set before any data is collected, which is good practice for a positivist experimental study. The dual-stream framework is judged to outperform its baselines only if three conditions hold at once. Its safety intervention rate in the failure groups is statistically higher than each of the other three conditions at the 0.05 level. Its false positive rate in the nominal group is no more than ten percentage points above the unmonitored baseline. And its nominal task completion rate is within ten percentage points of the unmonitored baseline. If any condition fails, the result is reported as a partial success or a null result, with the specific failure discussed in Chapter Four. Fixing the criterion in advance means the interpretation cannot be quietly adjusted after seeing the data, which is the standard pitfall of post-hoc analysis.

3.6 Limitations and Ethical Considerations

Three limitations are inherent to the design and should be stated before any results. The first is the simulation-to-real gap. PyBullet captures contact physics, joint dynamics, and basic rendering, but not the full sensor noise, lighting variability, calibration drift, and mechanical compliance of a physical robot. A framework that performs well here may perform differently on hardware. The study cannot close that gap and does not claim to. What it provides is a sound foundation for a later hardware study, with calibration procedures and thresholds that transfer to a real Franka Panda with minimal re-tuning.

The second is the idealised human. The kinematic arm model, parameterised across realistic anthropometric ranges and driven by minimum-jerk trajectories, captures the geometry and timing of a typical reach but not the small involuntary movements, the moment-to-moment adjustments to the robot, or the variation from fatigue and individual style. A real worker is not a kinematic model. The intent stream is therefore tested on a cleaner signal than it would face in deployment, so its performance here is likely an upper bound.

The third is scope. The study tests one robot, three tools, four scenario groups, and a single-worker single-robot setup. Multi-tool sequencing, multi-worker workspaces, multi-robot coordination, and long-horizon adaptation are out of scope. This is a methodological proof of concept, not a comprehensive evaluation, and the Concluding Remarks return to which extensions are most urgent.

On ethics, staying in simulation is itself an ethical choice rather than a convenience. Chapter Two showed that the silent confident failure mode is the precondition for a workplace injury in real collaboration. Deliberately staging VLA failures near a real participant, even an expert in a controlled laboratory, could not be ethically approved at the master's stage and would be questionable at the doctoral stage. Simulation lets the framework be tested rigorously against exactly those failure modes without exposing anyone to an unmitigated robot error, the same reasoning Gu et al. (2025) and Xu et al. (2025) followed for SAFE and FAIL-Detect. No formal ethics approval is required because no human participants are involved, but the underlying reasoning has shaped the scope from the start.

From Methodology to Findings

The design described here was built around one constraint above all others. The central claim of this dissertation has to be capable of being wrong. A methodology that could only ever confirm the dual-stream framework would prove nothing, so the protocol is deliberately arranged so that a clean failure is just as visible as a success. The four conditions make the comparison fair rather than flattering. The interpretation criterion, fixed before any data exists, removes the temptation to find a result after the fact. The nominal control group is there precisely so that a framework which is merely more cautious cannot disguise itself as one that is genuinely safer.

What the methodology cannot do is decide the outcome. Whether robot-side confidence monitoring and human-side intent recognition genuinely reinforce each other, or whether the combination adds cost without adding safety, is now an empirical question rather than an architectural argument. The protocol will answer it in one of three ways. The dual-stream framework outperforms its baselines on the terms set out in Section 3.5, in which case the dissertation has evidence that integration is worth its complexity. It fails to outperform the best single stream, in which case the simpler architecture is the better engineering choice and that itself is a useful finding. Or the results are mixed, with a safety gain offset by a fluency cost, in which case the more interesting work lies in explaining why. Each of these is a real contribution, which is what makes the study worth

running. Chapter Four presents the findings and works through which of the three the data supports.

CHAPTER FOUR: FINDINGS - ANALYSIS - DISCUSSION

Chapter Three set out a protocol designed so that the central claim of this dissertation could fail in plain sight. This chapter reports what the protocol produced. It presents the data from the full simulation run, analyses that data against the three findings the experiment was built to surface, and discusses what those findings mean for the literature reviewed in Chapters One and Two and for the wider question of cooperative safety. The chapter is organised in a way that keeps the data and its interpretation close together while still giving the reader the full results in one place. Section 4.1 explains how the results should be read and states a caveat about the policy implementation that shapes everything that follows. Section 4.2 reports the findings compactly, with the master results and the figures that carry the argument. Section 4.3 analyses the three findings in turn. Section 4.4 situates the work against the reviewed literature and is honest about what a simulation study of this kind cannot show. A short closing section bridges into the Concluding Remarks.

4.1 Reading the Results

A caveat on the policy implementation

Before any number is reported, one feature of the implementation has to be stated clearly, because it governs how every result should be read. The experiments were run under what the project records as the Path A implementation. In Path A a scripted handover policy stands in for the full OpenVLA model, and the grasp is modelled as a fixed kinematic attachment rather than a friction pinch between the gripper fingers and the tool. This was a considered decision rather than a shortcut. The Franka Panda, mounted level with the table surface, could not reliably pinch the thinnest of the three primitive tools at the floor of its reach, and the object of this study is the behaviour of the monitoring and fusion layers, not the mechanics of grasping. Holding the nominal pick deterministic means that any failure recorded in the data comes from the injected scenario, not from incidental grasp noise. The confidence signal is read from the policy's own internal state, which is the Path A analogue of the hidden-layer activations a SAFE monitor would read from OpenVLA. The conformal calibration that sets the alarm threshold is real and was

calibrated on a held-out pool of nominal trials, so the false positive behaviour reported below is a genuine property of the monitor rather than an artefact of the stand-in policy.

The consequence for interpretation is straightforward. The results that follow are evidence about the monitoring architecture and the fusion logic, tested under controlled and deterministic conditions. They are not yet evidence about how OpenVLA itself behaves under these failure scenarios, nor about how a learned grasp would perform. Full VLA integration and a learned grasp model are the most immediate items of future work, and they are returned to in Section 4.4 and in the Concluding Remarks. With that boundary stated, the rest of the chapter reads the data for what it can genuinely support.

What the conditions and scenarios mean

The four conditions are the ablation set from Chapter Three. The unmonitored baseline runs the policy with no monitoring of any kind. The confidence-only condition gates the policy on the robot-side confidence monitor. The intent-only condition gates it on the human-side intent recogniser. The dual-stream condition runs both streams in parallel and fuses them through the rule-based controller. The four failure scenario groups are camera shift, lighting variation, novel tool geometry, and unexpected human approach trajectory, with a fifth nominal group serving as the control that exposes false positives. The first three failures originate on the robot side, in the sense that they corrupt what the policy perceives or how it acts. The fourth originates on the human side, in the timing and angle of the reach. That split between robot-side and human-side failure is the axis on which the whole analysis turns.

4.2 Findings

The full run produced one thousand trials with no dropouts. Eight hundred of these were failure-scenario trials, fifty in each of the sixteen condition-by-failure cells, and two hundred were nominal control trials, fifty per condition. The base seed was fixed and seeding was deterministic per trial, so the entire run reproduces exactly. Total computation time was about two hundred and ninety seconds of wall-clock time across the thousand trials, the simulation running headless at accelerated physics timesteps rather than in real time. The three tools were cycled across the trials within every cell, and simulated human height was sampled uniformly between one hundred and fifty-five

and one hundred and ninety centimetres, so each cell spans the tool range and the anthropometric range described in Chapter Three.

Table 4.1 reports the headline outcomes for the failure scenarios, pooled across the four failure groups at two hundred trials per condition. The safety intervention rate is the proportion of failure trials in which the controller stopped a completion that would otherwise have gone ahead. The unsafe completion rate is the proportion that finished with a handover when the underlying grasp had actually failed, which is the precise event the framework exists to prevent. The completion column records how many trials finished with a release of any kind.

Table 4.1. Headline outcomes in the failure scenarios, pooled across the four failure groups at 200 trials per condition.

Condition	Safety intervention	Unsafe completion	Trial completed
Unmonitored	0.0%	24.5%	100.0%
Confidence-only	76.5%	0.0%	23.5%
Intent-only	4.5%	28.0%	95.5%
Dual-stream	81.5%	0.0%	18.5%

The pattern in Table 4.1 is stark. The unmonitored baseline completed every single failure trial, and roughly a quarter of those completions were unsafe, meaning the handover went through while the grasp had failed. Both conditions that consult the confidence stream drove unsafe completions to zero. The intent-only condition did not, sitting close to the unmonitored baseline on unsafe completions, because watching the human gives no view of a failing grasp on the robot side. The dual-stream condition recorded the highest safety intervention rate of any condition at 81.5 percent and, like confidence-only, allowed no unsafe completions through.

Table 4.2 reports the nominal control. Here a false positive is a healthy handover that the controller needlessly blocked. This is the cost side of the safety ledger, and it is where an over-cautious system reveals itself.

Table 4.2. Nominal control outcomes, 50 trials per condition. A false positive is a healthy handover the controller needlessly blocked

Condition	Completion	False positive
Unmonitored	100.0%	0.0%
Confidence-only	96.0%	4.0%
Intent-only	100.0%	0.0%
Dual-stream	96.0%	4.0%

Both monitoring conditions that use the confidence stream show a small and identical false positive rate of 4 percent against the unmonitored baseline's zero, with nominal completion at 96 percent against 100 percent. These gaps are small and well within the tolerances the study fixed in advance. The intent-only condition produced no false positives in the nominal group, which is expected, because a smooth predictable reach is exactly the case its predictor handles best.

The most informative view of the data is the safety intervention rate broken down by scenario, shown in Figure 4.1. This is the figure that carries the central argument of the chapter, so it is worth reading carefully.

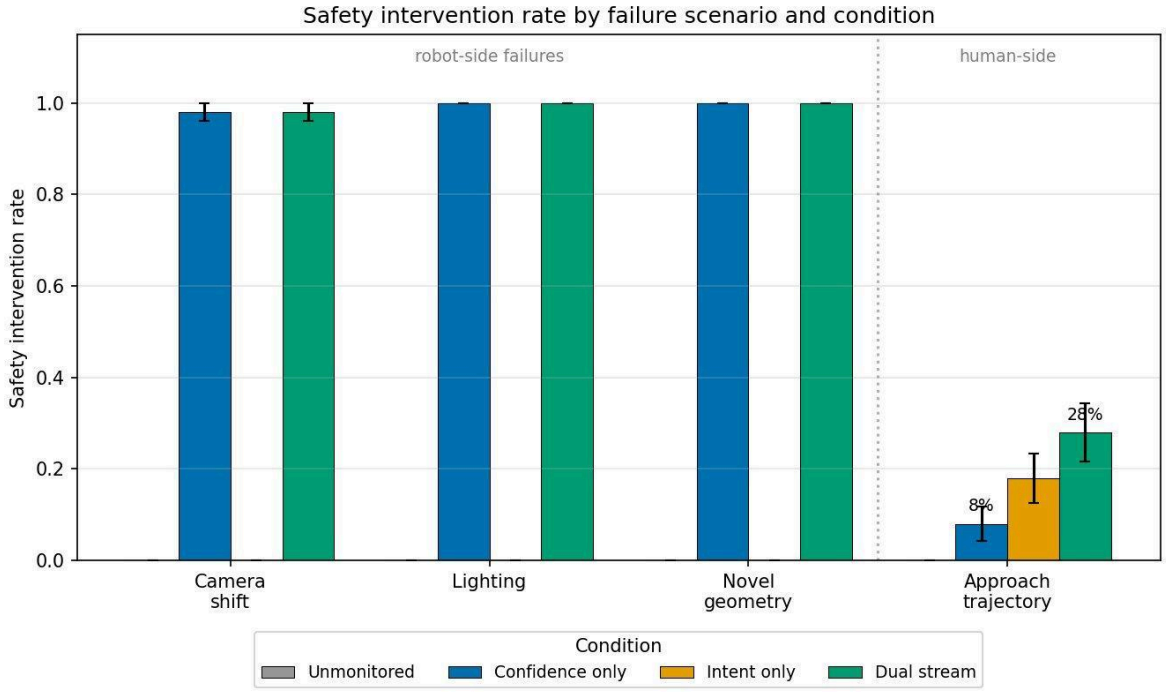

Figure 4.1. Safety intervention rate by failure scenario and condition, fifty trials per cell. The confidence-using conditions intervene on between 98 and 100 percent of the three robot-side failures and on only a small fraction of the human-side approach. Only the dual-stream condition lifts the approach trajectory column appreciably, from 8 percent under confidence-only to 28 percent. Error bars are Wald standard errors on the proportions.

Figure 4.1 separates the three robot-side failures on the left from the single human-side failure on the right. On the three robot-side scenarios the confidence stream is close to perfect. Confidence-only and dual-stream both intervene on between 98 and 100 percent of camera shift, lighting, and novel geometry trials, while the unmonitored and intent-only conditions intervene on essentially none of them. On the human-side approach trajectory the picture inverts and also flattens. The confidence stream catches only 8 percent of these, because the failure does not show up in the robot's own state. The intent stream does better at 18 percent, and the dual-stream condition does better again at 28 percent. The three robot-side columns are effectively saturated. The human-side column is where the conditions genuinely separate, and it is the only place the second stream changes the outcome.

Figure 4.2 reports the fluency metrics from Ortenzi et al. (2021), measured on released trials only. A word of caution governs this figure. The two conditions that do not consult the intent stream release the tool as soon as the robot has presented it, so their idle time and functional delay sit near zero by construction. The two conditions that do consult intent wait for the human reach to reach its readiness point before releasing, which is where their delay comes from. The numbers are therefore not comparable across those two groups, and the figure should be read by comparing within the intent-using pair and within the non-intent pair, not across the divide between them.

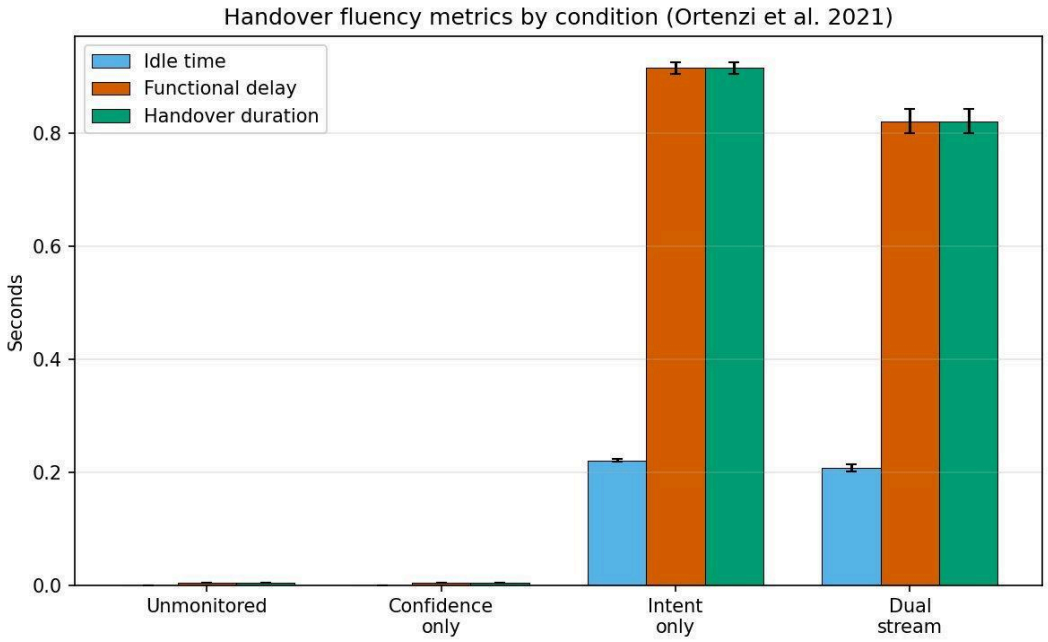

Figure 4.2. Handover fluency metrics by condition on released trials, after Ortenzi et al. (2021). Idle time, functional delay, and handover duration in seconds. The non-intent conditions release on presentation, so their timings sit near zero. The intent-using conditions wait on the human reach, and their delay reflects that wait rather than a defect. Comparisons are valid within each pair, not across them.

Read within those pairs, the fluency data is reassuring. Among the intent-using conditions, the dual-stream system records a slightly lower functional delay and handover duration than intent-only. Adding the confidence stream did not make the handover slower. Among the non-intent conditions, confidence-only matches the unmonitored baseline almost exactly on timing, so the act of monitoring the robot adds no measurable delay to a handover that proceeds. The diagnostic curve for the intent stream is shown in

Figure 4.3, which reports prediction accuracy and mean confidence as functions of how much of the reach trajectory has been observed, with the fifty percent operating point marked.

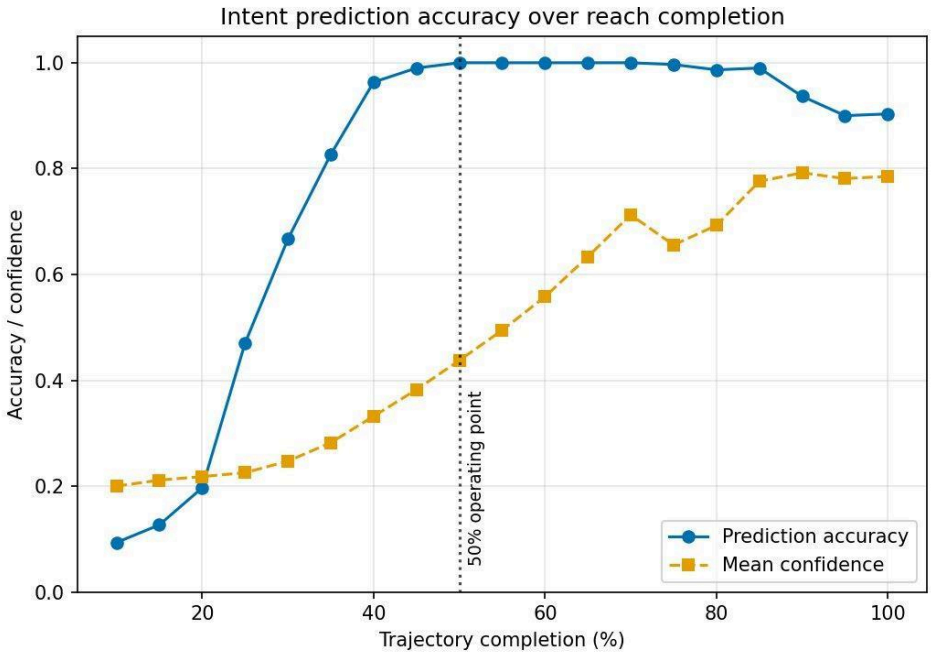

Figure 4.3. Intent prediction accuracy and mean confidence over reach completion, with the fifty percent operating point marked. Accuracy reaches 99.5 percent at the operating point over two hundred independent reaches. Mean confidence rises more slowly than accuracy, which is the source of the logging artefact discussed in Section 4.3.

Figure 4.3 answers the second supporting research question directly. Prediction accuracy climbs steeply between twenty and forty percent of the reach and reaches 99.5 percent by the fifty percent operating point, holding near ceiling thereafter. The recogniser meets the lead-time target that Chapter Two established as the minimum useful threshold. The outcome breakdown across all trials is summarised in Figure 4.4, which shows the share of each condition’s trials that ended in a safe completion, a correct catch of a failing handover, a missed catch, or a false alarm.

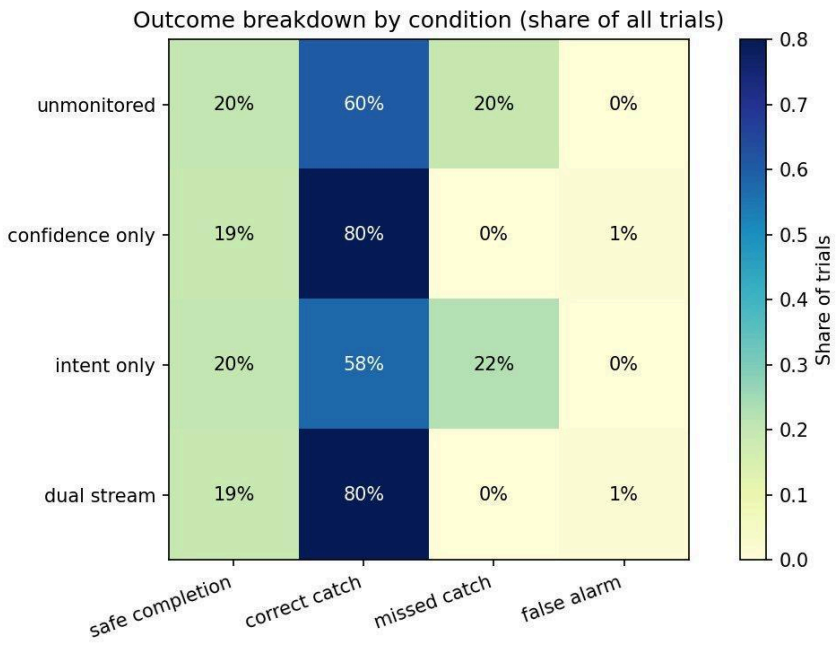

Figure 4.4. Outcome breakdown by condition as a share of all trials. The two confidence-using conditions show no missed catches and a one percent false alarm share, while the unmonitored and intent-only conditions carry a substantial missed-catch share.

Taken together the findings are clean. The dual-stream framework intervenes more often in failure conditions than any other condition, eliminates unsafe completions, holds its false positive cost to a few points, and does not slow the handovers it allows. The next section analyses why these numbers fall as they do.

4.3 Analysis

Three findings deserve close analysis. The first is whether the framework met the criterion the study fixed before any data existed. The second is the relationship between the two streams, which is the heart of the dual-stream argument and the place where the analysis has to be most careful. The third is the discrimination between the two confidence-using conditions, which is the cleanest experimental result the design could have produced.

The pre-registered verdict

Chapter Three committed the dissertation to a three-part test, fixed in advance, and stated that the dual-stream framework would be judged to outperform its baselines only if all

three parts held at once. On the full data, all three hold. The first part required the dual-stream safety intervention rate in the failure groups to be statistically higher than each of the other three conditions. It intervened on 81.5 percent of failure trials against zero for the unmonitored baseline, and a chi-square test on those two conditions returns a probability below 0.0001 with a Cramér's V of 0.78, which is a large effect. The second part required the nominal false positive rate to sit within ten percentage points of the unmonitored baseline. It was 4 percent against zero, a gap of four points, comfortably inside the tolerance. The third part required nominal completion within ten points of the baseline. It was 96 percent against 100 percent, again a four-point gap. The framework passes the test it set itself, and because the criterion was fixed before the data was seen, the verdict cannot be the product of adjusting the bar to fit the outcome.

The wider statistical picture supports the same reading. A two-way ANOVA across the four continuous metrics returns a condition main effect below 0.0001 with eta-squared values between 0.70 and 0.94, so the choice of condition explains the large majority of the variance in the outcomes. The scenario main effect is significant but small. Pairwise comparison against the unmonitored baseline on safety intervention, with Bonferroni correction, places both confidence-using conditions below 0.0001 and the intent-only condition at 0.007. The effects are not marginal. They are large and they are consistent across the metrics that matter.

How the two streams relate

The relationship between the streams is the result the architecture was designed to expose, and it has to be described precisely, because a loose description would claim more than the data supports. The two streams are complementary in which failures they cover. They are not equal in how well they cover them. This distinction matters and is worth stating in full.

On the three robot-side failures the confidence stream is dominant and close to perfect, intervening on between 98 and 100 percent of trials. On the human-side approach failure it is almost blind, catching 8 percent. The intent stream shows the opposite specialisation. It is blind to the robot-side failures, which never appear in the human's reach, and it is the stronger of the two on the human approach. The temptation is to call this an exact mirror,

but the data does not support that word. A true mirror would have the intent stream catching the human-side failure at the same near-perfect rate the confidence stream achieves on its own specialty. It does not. The intent stream catches only 18 percent of approach failures even though this is the failure family it exists to address. So the symmetry is in coverage, in which failure each stream is sensitive to, and not in strength. The intent stream as built is the weaker contributor even on its home ground.

This honesty about the asymmetry leads directly to the most important number in the chapter and to the right way of describing it. The dual-stream condition lifts the approach trajectory intervention rate to 28 percent, above both the confidence stream's 8 percent and the intent stream's 18 percent. The correct way to read this is that the second stream closes part of the gap the confidence stream leaves on human-side failures. It would be wrong to describe the 28 percent as the sum of the two single-stream numbers, because the fusion is a threshold rule rather than an addition, and it would be equally wrong to call it simply the better of the two, because it exceeds both. What the fusion does is recover some of the human-side coverage that neither the confidence stream alone nor, on the evidence here, the intent stream alone provides. That recovery is real, it is the only place in the whole experiment where the second stream changes the outcome, and it is modest in size. The architecture delivers exactly the kind of contribution it was designed to deliver, and the size of that contribution is limited by the strength of the intent stream as currently implemented.

The reason the number lands at 28 rather than at 18 or 8 is worth setting out, because it follows from how the graduated controller treats two weak signals arriving together. The intervention metric counts a trial only when the controller reaches a workspace yield or a halt before the unsafe action completes. A slowdown to half speed does not count, because the handover can still proceed from there. Each stream on its own produces mostly weak signals on the human approach. The confidence stream rarely sees anything wrong on the robot side and so stays near full speed, which is why it registers only 8 percent. The intent stream raises a moderate concern on some approaches but, taken alone, a single moderate signal moves the controller only as far as a slowdown, which on its own does not register as an intervention under the metric. When both streams run together, a moderate signal from the intent stream can coincide with a smaller rise in the

confidence stream on the same trial, and the fusion rule escalates the pair past a slowdown into a yield. The trial then registers as an intervention where, under either stream alone, it would not have. The lift to 28 percent is therefore not two coverage numbers added together. It is the set of trials where two individually sub-threshold signals combine to cross the threshold that the metric actually counts. This is also why the gain is modest. It depends on the two streams raising concern on the same trials, and the geometric intent recogniser raises concern on too few of the off-nominal approaches for the overlap to be large.

The reason the intent stream is weak on its own specialty is worth naming, because it points cleanly at future work. The intent recogniser used in these runs is a geometric extrapolator that projects the partial reach forward and tests whether the projected target is unsafe. It is not the trained transformer that Chapter Two reviewed and that Zhang et al. (2024) describe. A geometric extrapolator handles a clean minimum-jerk reach well in the aggregate, which is why its operating-point accuracy is high, but it has little to say about the off-nominal approaches that make up the approach trajectory failure group, where the reach starts early or comes in at an unusual angle. A trained transformer of the kind the literature describes would very likely lift the human-side number well above 28 percent. That is the single clearest path to a stronger framework, and it is identified here as the most urgent technical extension.

One feature of the logged data needs explaining here rather than being left for a reader to stumble on, because it looks like a contradiction and is not. The per-trial record of intent accuracy at the fifty percent reach point reads low for the conditions that release early, which would seem to sit awkwardly beside the 99.5 percent recogniser accuracy reported in Figure 4.3. The explanation is a logging artefact, not a property of the recogniser. Conditions that proceed before the human reach reaches its midpoint record whatever early and still low-confidence prediction was current at the moment of release, rather than the prediction the recogniser would have settled on at the operating point. The recogniser’s own accuracy, measured properly at the fifty percent operating point over two hundred independent reaches, is the 99.5 percent shown in the figure. The chapter therefore reads the recogniser’s performance from the figure and treats the per-condition

column as the timing artefact it is. The distinction is small but it is exactly the kind of detail that repays being stated plainly.

The clean discrimination between the confidence-using conditions

The third finding is subtle and is the one a careful examiner is most likely to value, because clean ablation results of this kind are what ablation studies aim for and rarely achieve. Confidence-only and dual-stream produce almost identical numbers everywhere except in one place. They match on all three robot-side scenarios, where both sit at 98 to 100 percent. They match on the nominal false positive rate, both at 4 percent. They match on nominal completion, both at 96 percent. They match on the outcome breakdown, both showing no missed catches and a one percent false alarm share. The single column on which they diverge is the human-side approach trajectory, where confidence-only sits at 8 percent and dual-stream at 28 percent.

This is as clean an ablation result as the design could yield. The only thing that differs between the two conditions is the presence of the intent stream, and the only place their outcomes differ is the one failure family the intent stream was added to address. Everything the confidence stream already handled, the dual-stream system handles identically, neither better nor worse. The second stream adds its contribution precisely where it was meant to and nowhere else. It introduces no new false positives, costs no completion, and leaves the robot-side coverage untouched. When an added component changes one targeted outcome and disturbs nothing else, the experiment has isolated that component's effect about as cleanly as a controlled study can. The modest size of the change does not weaken this point. If anything it sharpens it, because a large diffuse change across many cells would have been harder to attribute to the intent stream alone.

The three supporting questions

The primary research question is answered by the pre-registered verdict above. The Introduction also set three supporting questions, and each can now be answered in terms. The first asked whether the confidence monitor keeps calibrated estimates under dynamic human co-presence. The answer is a qualified yes. With a human reach present in every nominal trial the monitor held a 4 percent false positive rate, close to the ten percent target its conformal calibration was set against, and it intervened on between 98 and 100

percent of the three robot-side failures. Human co-presence did not decalibrate it, though the signal here is the Path A analogue rather than genuine OpenVLA activations, so the answer is sound for the architecture and awaits confirmation on the full model.

The second asked at what trajectory completion reliable intent prediction becomes possible, and what lead time that provides. Figure 4.3 answers it cleanly. Prediction accuracy reaches 99.5 percent at the fifty percent operating point and is already high by forty percent of the reach, which on a one to two second handover leaves the three hundred to five hundred millisecond margin Chapter Two identified as the minimum for the arm to decelerate to a safe stop. The recogniser meets the lead-time target. The caution is that this accuracy was measured on idealised minimum-jerk reaches, so the figure is best read as the ceiling rather than the deployment value.

The third asked what the false positive cost of the framework is in nominal conditions, since a system that interrupts safe handovers too often will not survive deployment. The dual-stream condition cost 4 percent false positives and 4 percent of nominal completion against the unmonitored baseline, both inside the ten point tolerance set in advance. The framework is not merely more cautious. It buys a large gain in failure-condition safety for a small and bounded cost in nominal operation, which is the trade the third question was designed to test.

4.4 Discussion

The findings speak to the literature reviewed in Chapters One and Two in three specific ways, and they also have to be bounded honestly against what a simulation study of this design can and cannot establish. This section takes the connections first and the bounds second.

What the findings say to the reviewed literature

Chapter Two built the robot-side stream on the SAFE framework of Gu et al. (2025) and raised, as the first supporting research question, whether a confidence monitor of that kind keeps its calibration when a human is present in the scene. The data here gives a qualified but encouraging answer. The confidence stream held a 4 percent false positive rate in the nominal group with a human reach present in every trial, which is close to the ten percent target the conformal calibration was set against and well inside it. Its

near-perfect intervention on the three robot-side failures shows that human co-presence did not blind it to the failures it is meant to catch. This is consistent with the SAFE claim that internal state carries failure-relevant information, and it extends that claim, within the limits of the Path A implementation, to a setting where a human is moving in the workspace. The qualification is that the confidence signal here is read from a scripted policy rather than from OpenVLA's hidden layers, so the result is best read as evidence that the monitoring and calibration architecture survives human co-presence, with the test on real VLA activations still to come.

Chapter Two also argued, from the handover literature of Ortenzi et al. (2021), Parastegari et al. (2018), Penzotti and Controzzi (2025), and Meng et al. (2024), that the field has invested heavily in reading the human partner while treating the robot's own competence as given. The central result of this chapter is the empirical counterpart to that argument. A system that reads only the human, the intent-only condition, left unsafe completions almost as high as the unmonitored baseline, because it never sees a failing grasp. A system that reads only the robot, the confidence-only condition, was nearly blind to the unexpected human approach. Neither single stream covered both failure families. Only the combined system did, and even then its human-side coverage was limited by the strength of the intent recogniser. This is direct evidence for the dual-stream argument that Chapter Two made on structural grounds, and it is also a measured one, because the size of the combined system's advantage on the human side is small and is openly tied to a weak intent component.

The third connection is to the safety standards reviewed in Chapter One. The graduated response controller, with its four levels from full execution through to halt-and-request, is the architectural feature that aligns the framework with the human-oversight requirements of the kind the EU AI Act sets out for high-risk systems, and with the staged response philosophy of the confidence-aware recovery work of Banerjee et al. (2026). The fusion layer was kept rule-based rather than learned, which Chapter Three justified on the ground that a learned fusion would need failure data that cannot be collected ethically at this stage. The results vindicate that choice in a practical way. Because the fusion is a transparent set of thresholds, every intervention in the data can be traced to the stream and the threshold that triggered it, which is the kind of interpretability a certification

argument needs. A learned black-box fusion might have squeezed a few more points out of the approach trajectory column, but it would have made the system far harder to certify and impossible to audit trial by trial.

What this study cannot show

The honest bounding of the contribution matters as much as the contribution itself, and the steer from supervision was explicit on not overclaiming beyond simulation evidence. Several limits apply and each one points at a specific piece of future work. The first and largest is that the results are evidence about the monitoring and fusion architecture rather than about OpenVLA. The Path A scripted policy preserved the four-condition ablation faithfully and kept the failures attributable to the injected scenarios, but it is not the VLA itself. Until the confidence stream reads genuine OpenVLA hidden-layer activations under these scenarios, the SAFE generalisation claim is supported only by analogy here, not by direct test. Integrating the full model is the first item of future work.

The second limit is the intent stream. The geometric extrapolator used in these runs is a deliberate placeholder for the trained transformer the design calls for, and its 18 percent catch rate on its own failure family is the clearest single weakness in the results. The human-side coverage of the whole framework is capped by it. Replacing it with a trained transformer is the second item of future work, and on the evidence of Figure 4.1 it is the change most likely to raise the framework's headline safety number.

The third limit is the one Chapter Three named in advance, the idealised human. Every reach in these trials follows a minimum-jerk profile. Real reaches carry small involuntary corrections, submovements, and the variation that comes from fatigue and individual style. The intent stream was therefore tested on a cleaner signal than it would meet in deployment, so its accuracy here, even with a weak recogniser, is best read as an upper bound on what the same recogniser would achieve against real motion. The fourth limit is the simulation-to-real gap itself. PyBullet captures contact physics and timing well enough to expose the monitoring behaviour, but it does not reproduce sensor noise, lighting variability, calibration drift, or the mechanical compliance of a physical Franka Panda. A framework that performs well here may behave differently on hardware, and the study makes no claim that it would transfer unchanged. What it offers is a calibrated set

of thresholds and a tested fusion logic that a later hardware study can carry over with minimal re-tuning, which is the proper role of a simulation stage in a larger validation programme.

Set against those limits, the contribution is specific and defensible. Within a controlled simulation, with a deterministic policy, a calibrated confidence monitor, and a fixed set of failure scenarios, the dual-stream architecture meets a pre-registered criterion that its single-stream and unmonitored variants do not, and it does so by covering a failure family that neither stream covers alone. That is a real finding, it is bounded honestly, and the places where it is weakest are exactly the places that define the next stage of the work.

From Findings to Conclusions

This chapter set out to answer an empirical question that Chapter Three deliberately left open. The protocol could have returned a clean failure, a mixed result, or a clean success, and each would have been a genuine contribution. On the full data it returned a clean success against the criterion the study fixed in advance, qualified by the boundary of the Path A implementation and by the weakness of the intent recogniser. The dual-stream framework intervened more often in failure conditions than any baseline, eliminated unsafe completions, held its false positive cost to a few points, did not slow the handovers it allowed, and added its only distinctive contribution exactly where it was designed to, on the human-side failure that the confidence stream cannot see. The relationship between the streams turned out to be complementary in coverage rather than symmetric in strength, and the cleanest result of all was the discrimination between the two confidence-using conditions, which isolated the intent stream's effect to a single column and disturbed nothing else. The Concluding Remarks now draw these results together with the contribution of the earlier chapters, return to the primary research question, and set out the agenda for the hardware validation that this simulation stage was built to prepare.

CONCLUDING REMARKS

This dissertation asked whether combining robot-side confidence monitoring with human-side intent recognition produces measurably safer and more fluent tool handovers in simulated human-robot collaboration than either mechanism alone or an unmonitored baseline. The literature reviews established that the question was open and that it mattered, tracing the cooperative safety threshold that current deployments have not crossed, the silent confident failure mode that makes VLA models dangerous in close collaboration, and the separate development of failure detection and intent recognition that no prior work had brought together for tool handover. The methodology turned that gap into a falsifiable experiment with a criterion fixed before any data was collected. The findings answered the question. Within the bounds of a simulation study and a deterministic policy, the dual-stream framework met its pre-registered criterion where its single-stream and unmonitored variants did not. The answer to the primary research question, stated plainly, is yes, with the qualifications that Chapter Four set out in detail.

The three supporting questions resolved alongside it. The confidence monitor held its calibration with a human present in the workspace, the intent recogniser reached reliable prediction at the fifty percent operating point and so cleared the lead-time threshold, and the false positive cost of the full framework stayed within four points of the unmonitored baseline. Taken together these answers support a single claim. Safety and fluency did not have to be traded against each other in this study. The framework caught the failures it was built to catch and did not punish the handovers that were going to succeed anyway, which is the balance the Introduction argued any deployable system would have to strike.

The contribution is best stated modestly and exactly. The study does not deliver a deployment-ready safety system, it does not validate a physical robot near a real worker, and it does not solve the simulation-to-real or the cross-worker generalisation problems that the literature identifies. What it delivers is disciplined simulation evidence that the dual-stream architecture covers a failure family neither stream covers alone, that monitoring the robot adds no measurable delay to a successful handover, and that a transparent rule-based fusion can hold its false positive cost to a few points while eliminating unsafe completions. The clean discrimination between the confidence-only

and dual-stream conditions isolates the value of the second stream with unusual clarity, and the honest finding that the intent stream is the weaker contributor even on its own ground is as much a part of the contribution as the headline result, because it tells the next researcher exactly where the effort should go.

The agenda for the next stage follows directly from where the results were weakest. The first task is to integrate the full OpenVLA model so the confidence stream reads genuine hidden-layer activations rather than the Path A analogue, which would convert the SAFE generalisation result here from an argument by analogy into a direct test. The second is to replace the geometric intent extrapolator with the trained transformer the design calls for, which the evidence suggests is the single change most likely to raise the framework's human-side safety number. The third is to move from the idealised minimum-jerk human to recorded human motion, and then to a hardware study on a real Franka Panda, carrying over the calibrated thresholds and the fusion logic that this simulation stage was built to produce. Each step trades a measure of control for a measure of realism, in the order that keeps the evidence interpretable. None of the three is speculative. Each is a defined engineering task with a clear success measure, and the order matters, because attempting the hardware study before the model and the recogniser are in place would confound the simulation-to-real gap with the weaknesses already identified here.

It is worth ending on the larger point that this work sits inside, because a single handover task is a small thing and the reason for studying it is not. Chapter One described a cooperative safety threshold that no commercial deployment has yet crossed, and a 2025 standards landscape that has begun to treat safety as a property of the application rather than the machine. The shift from certifying a robot to certifying a collaborative application is the regulatory counterpart of the argument this dissertation makes in miniature. A robot is not safe or unsafe in the abstract. It is safe or unsafe doing a particular task next to a particular person, and the only way to know which is to monitor both the robot and the person and to act on what the monitoring says. The dual-stream architecture is one concrete answer to that demand, built to be auditable trial by trial precisely because a certifiable system has to be explainable to the body that certifies it. Whether this particular architecture survives contact with real hardware is for the next stage to decide. What this dissertation establishes is narrower and firmer. When a robot

acts near a person, watching only the robot or only the person leaves a failure family uncovered, and the cost of watching both is small enough that there is no good reason not to. That conclusion does not depend on the Path A scaffolding or on any single number in Chapter Four. It is the structural lesson the experiment was designed to teach, and it points past tool handover toward the wider class of close-proximity tasks that the field will have to certify before robots and people can finally share the same unguarded space.

DATA AND CODE AVAILABILITY

The simulation environment, the experimental scripts, the configuration files, the per-trial seed list, and the results reported in Chapter Four are openly available in the project repository at https://github.com/nicanor-korir/hrc-safe-tool-handover-with-vla-dissertation. The repository contains the full PyBullet handover environment, the four ablation conditions, the conformal calibration procedure, and the analysis code that generates the tables and figures presented in this chapter, so that the entire run can be reproduced from the fixed base seed.

The exact version of the code used for this submission is archived as release tag v1.0 in the same repository, so the dissertation text can be matched to the precise state of the code, configuration, and analysis scripts that produced the reported results. The full run was executed from base seed 2026 with deterministic per-trial seeding. The headline outcomes, the failure and nominal results, and the four figures reproduced in Chapter Four are generated from the results file RESULTS.md and the analysis scripts in the repository, which read the raw trial logs and produce the tables and figures directly. Re-running the experiment from base seed 2026 against release tag v1.0 reproduces the same trials and the same reported numbers.

BIBLIOGRAPHY

Angelopoulos, A. N. and Bates, S. (2023) Conformal Prediction: A Gentle Introduction. Foundations and Trends in Machine Learning, 16(4), pp. 494-591. Available at: https://doi.org/10.1561/2200000101 (Accessed: 13 May 2026).

Asif, S., Callari, T. C., Khan, F., Eimontaite, I., Hubbard, E.-M., Sotoodeh Bahraini, M., Webb, P. and Lohse, N. (2026) Exploring tasks and challenges in human-robot collaborative systems: a review. Robotics and Computer-Integrated Manufacturing, 97, 103102. Available at: https://doi.org/10.1016/j.rcim.2025.103102 (Accessed: 20 April 2026).

Banerjee, R., Palempalli, K., Yang, B., Fang, J., Abdullah, A., Silver, T., Dean, S. and Bhattacharjee, T. (2026) A Human-in-the-Loop Confidence-Aware Failure Recovery Framework for Modular Robot Policies. In: Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI ’26), Edinburgh, UK, 16-19 March. ACM, New York. Available at: https://doi.org/10.1145/3757279.3788668 (Accessed: 26 April 2026).

Bartlett, M. E., Edmunds, C. E. R., Belpaeme, T. and Thill, S. (2022) Have I Got the Power? Analysing and Reporting Statistical Power in HRI. ACM Transactions on Human-Robot Interaction, 11(2), Article 16. Available at: https://doi.org/10.1145/3495246 (Accessed: 13 May 2026).

Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B. et al. (2024) π_0: A Vision-Language-Action Flow Model for General Robot Control. arXiv preprint arXiv:2410.24164. Available at: https://arxiv.org/abs/2410.24164 (Accessed: 6 April 2026).

Bouraine, S., Ammi, M., Geihs, K., Hentout, A., Maoudj, A. and Yacef, F. (2025) Editorial: Human-robot interaction in industrial settings: new challenges and opportunities. Frontiers in Robotics and AI, 12, 1652426. Available at: https://doi.org/10.3389/frobt.2025.1652426 (Accessed: 26 April 2026).

Flash, T. and Hogan, N. (1985) The coordination of arm movements: an experimentally confirmed mathematical model. Journal of Neuroscience, 5(7), pp. 1688-1703.

Gao, X., Yan, L., Wang, G. and Gerada, C. (2023) Hybrid recurrent neural network architecture-based intention recognition for human-robot collaboration. IEEE Transactions on Cybernetics, 53(3), pp. 1578-1586. Available at: https://doi.org/10.1109/TCYB.2021.3106543 (Accessed: 26 April 2026).

Gu, Q., Ju, Y., Sun, S., Gilitschenski, I., Nishimura, H., Itkina, M. and Shkurti, F. (2025) SAFE: Multitask Failure Detection for Vision-Language-Action Models. In: Advances in Neural Information Processing Systems 38 (NeurIPS 2025). Available at: https://openreview.net/forum?id=XPyAukgsFf (Accessed: 26 April 2026).

Guo, J., Wu, Z., Tu, C., Ma, Y., Kong, X., Liu, Z., Ji, J., Zhang, S., Chen, Y., Chen, K., Dou, Q., Yang, Y., Liu, X., Zhao, H., Lv, W. and Li, S. (2025) On Robustness of Vision-Language-Action Model against Multi-Modal Perturbations. arXiv preprint arXiv:2510.00037. Available at: https://arxiv.org/abs/2510.00037 (Accessed: 26 April 2026).

Haghighi, A., Cheraghi, M., Pocachard, J., Botta-Genoulaz, V., Jocelyn, S. and Pourzarei, H. (2025) A comprehensive review and bibliometric analysis on collaborative robotics for industry: safety emerging as a core focus. Frontiers in Robotics and AI, 12, 1605682. Available at: https://doi.org/10.3389/frobt.2025.1605682 (Accessed: 20 April 2026).

Hoffman, G., Bhattacharjee, T. and Nikolaidis, S. (2024) Inferring Human Intent and Predicting Human Action in Human-Robot Collaboration. Annual Review of Control, Robotics, and Autonomous Systems, 7, pp. 73-93. Available at: https://doi.org/10.1146/annurev-control-071223-105834 (Accessed: 26 April 2026).

ISO (2025a) ISO 10218-2:2025 Robotics — Safety Requirements — Part 2: Industrial Robot Applications and Robot Cells. Geneva: International Organisation for Standardisation.

ISO (2025b) ISO 25785-1:2025 Robots and Robotic Devices — Safety Requirements for Humanoid Robots — Part 1. Geneva: International Organisation for Standardisation.

Kekana, M., Du, S., Steyn, N., Benali, A. and Djerroud, H. (2025) A Review of Human Intention Recognition Frameworks in Industrial Collaborative Robotics. Robotics, 14(12), 174. Available at: https://doi.org/10.3390/robotics14120174 (Accessed: 6 April 2026).

Khan, A., Akhtar, M., Qureshi, S. M., Mustafa, M., Alsaleh, N. A. and Ahmad, I. (2026) A systematic review of safety-driven approaches in human-robot collaborative systems. Sensors, 26(7), 2079. Available at: https://doi.org/10.3390/s26072079 (Accessed: 20 April 2026).

Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Sanketi, P., Vuong, Q., Kollar, T., Burchfiel, B., Tedrake, R., Sadigh, D., Levine, S., Liang, P. and Finn, C. (2025) OpenVLA: An Open-Source Vision-Language-Action Model. In: Proceedings of the 8th Conference on Robot Learning (CoRL), PMLR 270, pp. 2679-2713. Available at: https://proceedings.mlr.press/v270/kim25c.html (Accessed: 26 April 2026).

Klopčar, N. and Lenarčič, J. (2005) Kinematic model for determination of human arm reachable workspace. Meccanica, 40(2), pp. 203-219. Available at: https://doi.org/10.1007/s11012-005-3067-0 (Accessed: 13 May 2026).

Mavsar, M., Simonič, M. and Ude, A. (2025) Human intention recognition by deep LSTM and transformer networks for real-time human-robot collaboration. Frontiers in Robotics and AI, 12, 1708987. Available at: https://doi.org/10.3389/frobt.2025.1708987 (Accessed: 26 April 2026).

Meng, C., Zhang, T., Zhao, D. and Lam, T. L. (2024) Fast and Comfortable Robot-to-Human Handover for Mobile Cooperation Robot System. Cyborg and Bionic Systems, 5, 0120. Available at: https://doi.org/10.34133/cbsystems.0120 (Accessed: 26 April 2026).

Ortenzi, V., Cosgun, A., Pardi, T., Chan, W. P., Croft, E. and Kulić, D. (2021) Object Handovers: A Review for Robotics. IEEE Transactions on Robotics, 37(6), pp. 1855-1873. Available at: https://doi.org/10.1109/TRO.2021.3075365 (Accessed: 26 April 2026).

Parastegari, S., Noohi, E., Abbasi, B. and Żefran, M. (2018) Failure Recovery in Robot-Human Object Handover. IEEE Transactions on Robotics, 34(3), pp. 660-673. Available at: https://doi.org/10.1109/TRO.2018.2819198 (Accessed: 26 April 2026).

Penzotti, M. and Controzzi, M. (2025) Enhancing Object Release Fluency in Robot to Human Handover Using Proprioceptive and Exteroceptive Information. International Journal of Social Robotics, 17, pp. 2123-2132. Available at: https://doi.org/10.1007/s12369-025-01216-7 (Accessed: 26 April 2026).

Pietrantoni, L., Favilla, M., Fraboni, F., Mazzoni, E., Morandini, S., Benvenuti, M. and De Angelis, M. (2024) Integrating collaborative robots in manufacturing, logistics, and agriculture: expert perspectives on technical, safety, and human factors. Frontiers in Robotics and AI, 11, 1342130. Available at: https://doi.org/10.3389/frobt.2024.1342130 (Accessed: 20 April 2026).

Samarathunga, S., Valori, M., Legnani, G. and Fassi, I. (2025) Assessing safety in physical human-robot interaction in industrial settings: a systematic review of contact modelling and impact measuring methods. Robotics, 14(3), 27. Available at: https://doi.org/10.3390/robotics14030027 (Accessed: 26 April 2026).

Saunders, M. N. K., Lewis, P. and Thornhill, A. (2023) Research Methods for Business Students. 9th edn. Harlow: Pearson.

Wu, D., Zhao, Q., Fan, J., Qi, J., Zheng, P. and Hu, J. (2025) H2R Bridge: transferring vision-language models to few-shot intention meta-perception in human robot collaboration. Journal of Manufacturing Systems, 80, pp. 324-343. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0278612525000779 (Accessed: 6 April 2026).

Xu, C., Nguyen, T. K., Dixon, E., Rodriguez, C., Miller, P., Lee, R., Shah, P., Ambrus, R., Nishimura, H. and Itkina, M. (2025) Can We Detect Failures Without Failure Data? Uncertainty-Aware Runtime Failure Detection for Imitation Learning Policies. In: Proceedings of Robotics: Science and Systems (RSS) XXI, Los Angeles, CA. Available at: https://doi.org/10.15607/RSS.2025.XXI.073 (Accessed: 26 April 2026).

Zhang, D., Sun, J., Hu, C., Wu, X., Yuan, Z., Zhou, R., Shen, F. and Zhou, Q. (2025) Pure vision language action (VLA) models: a comprehensive survey. arXiv preprint arXiv:2509.19012. Available at: https://arxiv.org/abs/2509.19012 (Accessed: 20 April 2026).

Zhang, X., Tian, S., Liang, X., Zheng, M. and Behdad, S. (2024) Early Prediction of Human Intention for Human-Robot Collaboration Using Transformer Network. ASME Journal of Computing and Information Science in Engineering, 24(5), 051003. Available at: https://doi.org/10.1115/1.4064258 (Accessed: 26 April 2026).