Co-Evolving with Domain Experts: A Proactive and Trustworthy Framework for Integrating Human Expertise into Visual Reasoning

Anonymous Author(s)

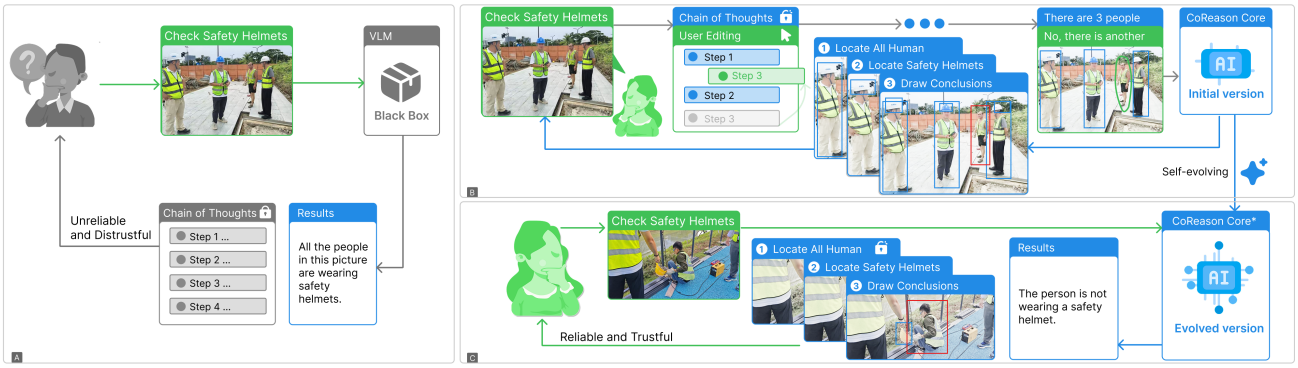


Figure 1: This figure shows the transformation process from unreliable black box AI to reliable co-evolution partners based on CoReason.(A) Existing challenges: the traditional VLM is like a black box. When it gives a wrong conclusion, users cannot understand its decision-making basis, resulting in a sense of distrust. (B)Co-evolution cycle of CoReason: CoReason externalizes the reasoning process of AI into a transparent and editable chain of thoughts. Domain experts can clearly see each analysis step of AI, and correct the errors through direct visual interaction and logic editing. This expert feedback is absorbed by the system and drives the initial version of AI to self-evolve. (C)The evolved trusted system: after continuous interaction and learning with experts, the evolved CoReason system becomes more accurate and reliable when dealing with new visual tasks, and can correctly identify people who do not wear safety helmets, ultimately winning the trust of users.

Abstract

Complex visual reasoning represents a frontier in AI, unlocking powerful applications across numerous domains. However, the adoption of current Vision-Language Models (VLMs) in expertise-driven fields, such as construction safety, is severely hampered by performance bottlenecks stemming from a lack of domain knowledge. To address this challenge, we present CoReason, a co-evolutionary framework. The core of CoReason is a transparent and directly editable workflow that externalizes the reasoning process. This design enables experts' knowledge to be directly integrated to shape the system's process from initial visual perception to final logical conclusion, a process that fosters expert trust and control. This design captures user interaction as a learning opportunity for continuous alignment. We validated CoReason in the construction safety domain, where it achieved a 33.7% increase in Recall after one month of co-evolution. We believe CoReason presents a scalable path toward building capable and trustworthy expert-level visual reasoning systems.

CCS Concepts

• **Human-centered computing** → **Collaborative interaction**.

Keywords

Visual Reasoning, Human-AI Collaboration, Domain Expertise, Trustworthy AI

ACM Reference Format:

Anonymous Author(s). 2018. Co-Evolving with Domain Experts: A Proactive and Trustworthy Framework for Integrating Human Expertise into Visual Reasoning. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 22 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Recently, the rapid advancement of Vision-Language Models (VLMs) has significantly advanced Visual Reasoning. While traditional computer vision models excel at perception[22, 64], they frequently struggle in scenarios requiring deep, contextual understanding. VLMs bridge this gap by mapping visual and linguistic information into a unified feature space[1, 14, 15], thereby enabling advanced capabilities for complex, visually-grounded reasoning.

Despite their promise, deploying VLMs in specialized, high-stakes domains reveals fundamental reliability challenges[17]. First,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, Woodstock, NY

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

a lack of deep domain knowledge frequently triggers hallucinations[76]. Second, the inherent opacity of these black-box models creates a substantial barrier: without expertise in computer science or prompt engineering, domain experts struggle to identify and correct flawed reasoning[44, 57]. Consequently, this inability to intervene leads to mistrust, severely reducing experts' willingness to adopt these tools[18]. To overcome these challenges, the paradigm must shift from autonomous AI agents to collaborative human-AI partnerships. Such collaboration not only allows experts to mitigate AI performance gaps via direct oversight but also inherently fosters trust[6, 40]. While human-AI collaboration has been widely explored in Large Language Models (LLMs)[12, 67, 70]—primarily through text-based prompting and revision—a critical gap remains: few studies have extended these interactive paradigms to expertise-driven, visually-grounded domains.

We selected construction safety inspection as our target domain, as it heavily relies on complex rules and specialized knowledge typically absent in general-purpose models[73, 76]. To ground our design in actual user needs, we conducted semi-structured interviews with 8 experienced safety engineers. A thematic analysis of these sessions yielded three core design goals for a collaborative visual reasoning system: (1) Direct Visual Intervention: Empowering experts to provide feedback and teach the system via intuitive visual interactions (e.g., pointing, drawing). (2) Co-Evolution with Experts: Designing an adaptive, evolving partner rather than a static, fixed model. (3) Transparent and Modifiable Reasoning: Externalizing the AI's reasoning into an editable logic chain, enabling experts to effectively understand, critique, and correct the process.

These goals directly informed the development of CoReason, a novel interactive framework for collaborative visual reasoning. CoReason's core mechanism translates opaque AI reasoning into an editable, step-by-step workflow. By presenting a detailed set of checks grounded in visual evidence, the system allows experts to trace the exact inference path. For any flawed step, users can apply direct visual interactions—such as clicking or circling on the image—to make targeted corrections. This continuous feedback loop simultaneously refines the underlying model, ensuring the system progressively aligns with expert knowledge over time.

To evaluate CoReason, we employed a mixed-methods approach comprising two studies. First, we conducted a within-subjects user study with 24 participants, comparing CoReason against a baseline VLM system on realistic, image-based inspection tasks to assess immediate usability and user experience. Second, to investigate the system's co-evolutionary capabilities, we deployed a one-month longitudinal study with 3 expert proxies, tracking performance metrics on a weekly basis.

Our findings demonstrated that CoReason significantly reduces cognitive load during knowledge injection, shifts user behavior from repetitive low-level clarification to high-level constructive feedback, and establishes high levels of user trust. In the high-stakes context of construction safety, failing to detect a hazard can have severe consequences ; thus, maximizing Recall (minimizing missed dangers) is the most critical safety metric. After a month of co-evolution, CoReason achieved a 33.7% improvement in Recall. These results demonstrate that our framework successfully translates human expertise into measurable performance gains for safety-critical tasks. We conclude by discussing the necessity of a new design paradigm—one centered on direct intervenability—that concurrently evolves AI capability and human trust.

In summary, this paper makes the following contributions: (1) The design and implementation of CoReason, an interactive framework that empowers domain experts to naturally inject knowledge into AI systems via visual and logical interventions. (2) A rigorous mixed-methods evaluation demonstrating CoReason's significant benefits in inspection accuracy, user trust, and usability. (3) A set of empirically derived design principles for building trustworthy, co-evolving AI partners in expertise-driven domains.

2 Related Works

2.1 VLM in Visual Reasoning

Visual reasoning—the capacity of machines to logically infer and make decisions from visual data—is a foundational AI capability[24, 75, 79]. Moving beyond basic object recognition, it enables agents to analyze scene dynamics, relationships, and causality[35, 71]. Consequently, it has become crucial across diverse domains: empowering autonomous vehicles to predict pedestrian intent[20, 45], guiding industrial robots in precise assembly[25, 36, 78], enabling GUI comprehension for task automation[13, 16, 27], and accelerating scientific research via microscopic image analysis[52, 66].

However, deploying Vision-Language Models (VLMs) for complex visual reasoning introduces two critical challenges. The first is hallucination[30, 39, 49]. Real-world tasks demand implicit knowledge—such as physical commonsense or expert intuition—that resists explicit dataset encoding[29, 63, 69]. Lacking this deep causal understanding, VLMs often overfit to superficial statistical correlations. Consequently, when navigating novel or complex scenes, they frequently deviate from visual evidence, generating plausible but incorrect details that expose the limitations of their reasoning foundation[37, 38].

The second challenge is opacity[5, 42, 44]. While recent approaches attempt to enhance transparency by generating step-by-step rationales[41, 65], these outputs remain static traces of the decision-making process. Although they reveal the causal chain, they are presented as unalterable final results without mechanisms for real-time intervention. This mere "facade of transparency" prevents domain experts from validating the model's internal logic[57] or actively correcting flawed reasoning when errors occur.

2.2 Knowledge Integration Methods

One dominant approach is data-centric refinement, primarily via offline fine-tuning on domain-specific datasets[43, 56, 68]. While capable of directly modifying internal weights, this method presents significant challenges to expert collaboration. It is not only slow and expensive—demanding vast amounts of expertly annotated data[51, 72]—but also creates a disconnected feedback loop. Experts are confined to repetitive annotations providing only low-bandwidth signals (e.g., True/False labels), struggling to capture the rich, procedural "why" behind their diagnostic decisions.

As a data-efficient alternative, in-context methods like Retrieval-Augmented Generation (RAG) guide outputs by retrieving relevant examples, bypassing costly retraining[4, 77, 80]. However, these

approaches afford only shallower interventions and remain predominantly confined to the linguistic modality, severely limiting their efficacy in complex visual reasoning.

A third stream explores interactive paradigms—such as Human-in-the-Loop (HiTL) and Interactive Machine Learning (IML)—to directly integrate experts into model refinement[31, 54]. Positioning the human as a collaborative teacher, these frameworks utilize visual tools to inject knowledge[9, 11, 47]. Yet, two critical limitations persist. First, most rich collaborative interactions target LLMs[12, 67, 70], with limited extensions to VLMs[81]. Second, expert involvement is frequently restricted to post-hoc evaluations of fully developed models[8, 23, 32, 60, 82], a delayed engagement that prevents the early correction of foundational reasoning flaws.

In summary, current methodologies present a clear tradeoff: they offer either deep but offline, non-interactive refinement (fine-tuning), or shallow, linguistically-focused interactive paradigms (RAG, IML). To bridge this gap, our work introduces a unified framework that synthesizes deep, procedural intervention on the AI's reasoning chain with an intuitive, visually-grounded interface tailored for domain experts.

2.3 The Challenge of Building Trust

Despite rapid technical advances, AI adoption among domain experts in high-stakes fields remains critically limited[3, 33, 46]. Seminal HCI research posits this gap stems not from technical shortfalls, but from a fundamental crisis of trust[53, 59]. Experts are often reluctant to delegate consequential decisions to systems they cannot scrutinize, control, or hold accountable. The dominant technical response—Explainable AI (XAI)—operates on the assumption that transparency fosters trust, typically operationalized via post-hoc methods like LIME[19, 50], SHAP[2], and saliency maps[55] to rationalize model outputs.

However, emerging HCI literature reveals a far more nuanced relationship between explanation and trust. While sometimes beneficial, post-hoc explanations can exacerbate cognitive load[7, 48] and remain highly context-dependent[34]. More critically, seemingly interpretable models often foster a dangerous "illusion of understanding," directly miscalibrating user reliance[48]. This leads to over-trust and diminished vigilance, failing to actually improve users' error-detection capabilities. Ultimately, robust trust demands more than mere comprehension[6]; it requires predictability, reliability, and, crucially, a substantial sense of user control and agency[21, 61, 62]. Users inherently trust systems they can actively guide and correct over those they must passively observe[28, 58, 74], necessitating a shift from post-hoc persuasion to procedural, auditable processes.

Aligning with our formative study findings, our work extends beyond passive explainability. We argue that genuine, resilient trust is established when experts are empowered with direct, contextual intervention capabilities throughout the AI's entire reasoning lifecycle.

3 Formative Study

We conducted a formative study to understand workflows, challenges, and cognitive gaps between construction safety experts and AI systems. This study provides an empirical foundation for our system design, addressing the following research questions: **RQ1**: What role does visual evidence play in how experts analyze safety hazards and communicate their findings? **RQ2**: What are the foundational requirements and expectations experts have for an AI system intended to assist in their visual analysis? **RQ3**: To meet those essential capabilities, how must an expert be able to interact with the AI's reasoning process during analysis?

3.1 Formative Interview

We recruited 8 experienced construction safety engineers and site supervisors. The study was conducted in two parts. First, we conducted semi-structured interviews to understand their general workflow, use of existing tools, and current challenges. Second, in a contextual inquiry-style session, participants performed a think-aloud analysis on 10 complex construction site images containing potential safety hazards. Experts were provided with initial VLM outputs as starting points and allowed to interact with the model, enabling us to observe their step-by-step reasoning. Sessions lasted approximately one hour. Transcripts were analyzed using thematic analysis [8]. Two researchers independently applied open coding to the data, resolving discrepancies through discussion to consolidate insights into three core themes.

3.2 Findings

Our analysis revealed a gap between experts' nuanced reasoning and current AI capabilities. We summarize our findings into three key themes.

Theme 1: Visual Evidence as the Primary Medium for Analysis and Communication. While textual feedback is conventional for VLM improvement, participants found it inefficient (P1, P6) and insufficient for precise spatial instruction in complex domains like construction safety. All participants emphasized direct visual annotation as the primary medium for analysis. Translating spatial insights into language fails to help teams quickly identify issues or establish shared context. P8 noted, "You'll lose the critical detail if you rely on text... I need to circle that exact spot. Describing that in words is slow, and someone will misinterpret it." Similarly, P5 stated that marking up a photo helps workers quickly locate problems. Visual interaction is considered the essential first step for analysis.

Theme 2: Control Over the Construction of Trustworthy Automated Processes. Experts are not seeking a single-use AI tool, but rather the ability to construct a trustworthy, automated analysis process (P2, P7, P8). For repetitive inspection tasks, their ultimate goal is not to repeatedly correct AI errors. Instead, experts are willing to invest effort upfront to jointly define and validate the inspection steps for a specific safety rule. P7 noted: "I need it to learn my method, and after I confirm it's correct, it has to repeat that process reliably every time." Experts aim to shape reusable AI behavior rather than correct isolated outputs. Therefore, a system must support experts in directly integrating their experience into the AI's executable logic to build long-term trust.

Theme 3: The Need for Transparent and Editable AI Reasoning. While AI is regarded as a suitable substitute for automating visual analysis, all participants cautioned against using black-box systems for safety assessments. Direct, unexplainable outputs do not help them validate findings or take accountability. P3 warned,

"If the system just tells me 'fall hazard,' that's not enough. I need to know what it's seeing. Without that, I can't verify it, and I can't trust it." Furthermore, participants stressed the need for immediate intervention. P7 stated, "If the AI misidentifies a bundle of wires for a rope, all its subsequent analysis... will be wrong. I need to be able to tell it 'No, that's not a rope' right at that moment."

3.3 Design Goals

Based on our findings, we derived three core design goals for CoReason to bridge the gap between expert workflows and general-purpose AI.

DG1: Focus on Direct Visual Interaction as the Primary Feedback Modality. Driven by Theme 1, the system must allow experts to provide feedback in their natural modality: direct interaction with visual evidence. Enabling actions like clicking or drawing boundaries to correct the AI resolves the inefficiency of text-based communication and ensures nuanced knowledge is captured with high fidelity.

DG2: Framing the AI as an Automated Workflow Engine Co-Constructed with and Progressively Evolved by Experts. Addressing Theme 2, we frame the system as a framework for building automation rather than a static, pre-configured tool. Experts initially act as "chief engineers", constructing and debugging workflows through direct supervision. Once verified as robust, experts can progressively delegate tasks, shifting to a review role to ensure trustworthy outcomes.

DG3: Ensure the AI's Step-by-Step Reasoning is Transparent and Editable. Stemming from Theme 3, this addresses the essential need for verifiability. Experts must be able to inspect the entire reasoning process—from visual grounding to logical conclusion—to validate correctness and accept responsibility. The system must visualize the AI's chain of thought and allow experts to intervene at any step to correct errors.

4 CoReason Design and Implementation

4.1 System Architecture Overview

CoReason is a collaborative visual reasoning framework designed to align general-purpose AI capabilities with domain-specific expert requirements. As illustrated in Figure 2, the architecture operates through two interconnected cycles: the Inference Loop, which executes inspection tasks, and the Co-Evolutionary Loop, which integrates expert feedback into the system's knowledge base.

The Inference Loop serves as the primary execution engine. It transforms a user's high-level request (e.g., "Check if everyone is wearing a helmet.") and visual evidence into a transparent, step-by-step reasoning workflow. This process is driven by three core modules: (1)Intent Corrector: A specialized LLM module that refines underspecified user inputs into precise, domain-compliant inspection goals by retrieving relevant rules from the Memory Base. (2)Planner: A reasoning agent that decomposes the refined goal into a structured, executable plan. It selects appropriate operations from a pre-defined Tool Library containing both neural models (e.g., object detectors) and deterministic functions. (3)Executor: This module runs the generated plan, invoking the selected tools to produce visual evidence (e.g., bounding boxes, masks) on the image for expert verification.

The Co-Evolutionary Loop functions as the system's learning mechanism. It captures expert interventions to continuously improve performance. When experts correct a logical error in the plan or a perceptual error on the image—the feedback is processed by Coreason. It distills high-level rules and successful reasoning patterns from the interaction logs and updates the Memory Base for subsequent retrieval-augmented generation (RAG). Simultaneously, corrected visual annotations are collected to periodically fine-tune the specialized models in the Tool Library.

This dual-loop design ensures that CoReason is not a static tool but an evolving partner that progressively adapts to the expert's mental model and the specific visual domain.

4.2 Proactive Intent Refinement and Grounding

In specialized domains like construction safety, user queries often rely heavily on tacit knowledge. An expert might input a high-level instruction like "Check scaffolding safety," implicitly expecting the system to evaluate guardrails, baseplates, and cross-bracing. Traditional prompting requires exhaustively articulating these details, creating a high cognitive barrier. CoReason addresses this through a mixed-initiative intent elicitation process.

When an expert provides an initial, potentially underspecified directive alongside a site image, CoReason's backend invokes the Intent Corrector. This module leverages RAG to query the Memory Base—a dynamic repository of historical, expert-validated rules. By matching the user's input with historical data, the AI systematically infers missing but critical inspection dimensions.

To inject this knowledge without overriding the user's original intent, CoReason maintains strict human agency. It does not autonomously modify the prompt without user oversight; instead, it translates the retrieved knowledge into proactive UI suggestions. The system presents supplementary rules as selectable options (e.g., "Also verify if workers on the scaffold are wearing helmets?"). The expert simply reviews these suggestions and selects the relevant checkboxes to finalize the target(Figure 3A, inset). This brief Intent Grounding step explicitly decouples AI-driven intent expansion from human-controlled intent confirmation. It seamlessly transforms concise, high-level inputs into well-defined goals, ready for the system to decompose into an executable logical plan.

4.3 Reasoning Chain Editing

While generating a step-by-step reasoning chain improves AI performance, traditional systems present this as an uneditable text block. If an expert identifies a flaw, they must resort to laborious prompt engineering. CoReason resolves this by externalizing the AI's reasoning into a transparent, fully editable, and visually grounded workflow.

Task Decomposition and Node Editing: Once the user finalizes the inspection goals, CoReason's Planner breaks down the high-level intent into a sequence of atomic operations. The frontend renders this underlying data structure as a sequence of interactive UI nodes(Figure 3B). Experts can directly drag and drop these nodes to reorder the execution sequence, or insert/delete steps, refining the high-level logic before any tools are fully executed(Figure 3D).

Semantic Highlighting and Real-Time Grounding: Beyond structural editing, CoReason uniquely bridges abstract logic and

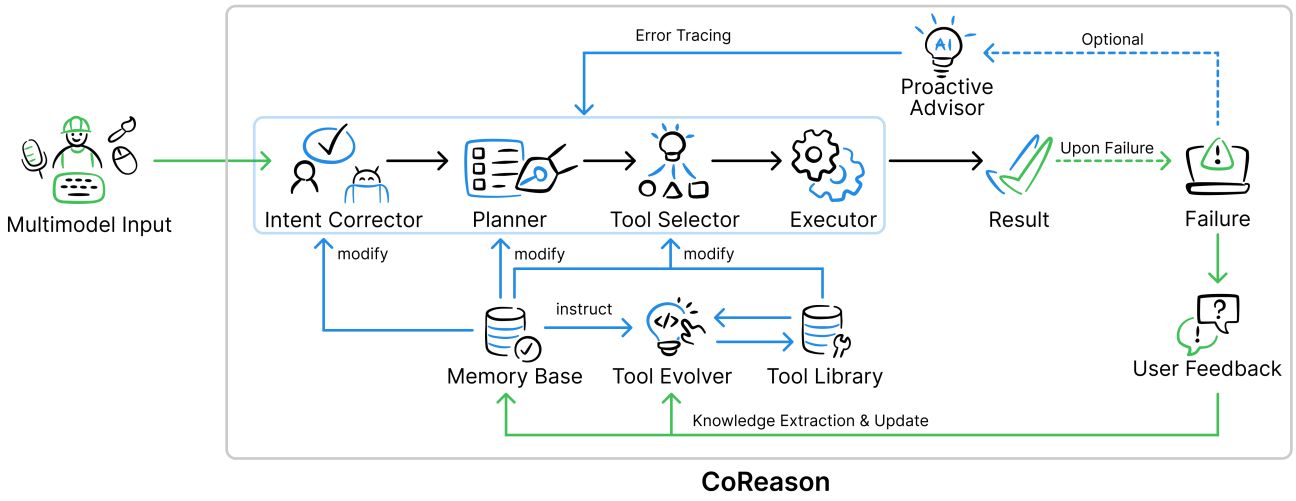


Figure 2: Architecture of CoReason system. A workflow can be generated by the user's input(consists of image and text) through the middle path. If the workflow fails, the Proactive Advisor will recommend a most potential failure in the middle path to support a quick fix by experts. All user feedback will be restored in Memory Base to help improve the performance of the system.

concrete perception. Within the generated plan, domain-specific terms (e.g., "cross-bracing" or "limit switch") are dynamically highlighted. Hovering over a term triggers a tooltip displaying the system's semantic understanding of the concept(Figure 3B, inset). Crucially, if the underlying visual model can already identify the object, it projects a preview of its visual localization (e.g., bounding boxes or masks) directly onto the uploaded image(Figure 3B, right).

Dense Knowledge Injection: This interactive mechanism transforms the planning stage into a dense knowledge injection phase. If the AI's perceptual grounding is flawed (e.g., an inaccurate bounding box), the expert can seamlessly transition to visual annotation, manually redrawing or adjusting the boundaries directly on the image(Figure 3C). By enabling experts to simultaneously edit the strategic reasoning (workflow nodes) and correct the perceptual grounding (visual annotations), CoReason ensures the pipeline is closely aligned with the expert's mental model. Furthermore, these manual visual corrections are systematically recorded to facilitate the model's long-term evolution (detailed in Section 4.5).

4.4 Workflow Implementation and Automation

With the reasoning chain structurally edited and visually grounded by the expert, the system transitions from interactive planning to automated execution. CoReason bridges the conceptual logic and computational execution through dynamic workflow instantiation.

Tool Matching and Instantiation: Once the user confirms the finalized plan, CoReason's Executor parses the structured sequence and maps each atomic operation to specific, executable modules available within the Tool Library. To handle the diverse requirements of visual reasoning, this library integrates a hybrid set of tools: it includes neural perception models (e.g., YOLO for detection, SAM for precise segmentation) as well as deterministic programming functions(Figure 3E).

Visible Audit Trail: Crucially, CoReason does not treat the execution process as an opaque background process. As the workflow progresses sequentially, the system renders the intermediate output of each tool—such as by projecting bounding boxes directly onto the inspection image(Figure 3F, inset). This step-by-step visual rendering creates a comprehensive, visible audit trail. Rather than being forced to required to rely solely on a final "Pass/Fail" output, domain experts can clearly trace how the visual evidence propagates through the logical steps, ensuring that the system's final conclusion is perceptually and procedurally verifiable.

4.5 Guided Correction and System Evolution

In complex visual reasoning, errors are inevitable. However, traditional AI acts as a static, non-interactive system; when an error occurs, experts are left to guess whether the failure stemmed from a hallucinated concept, a missed bounding box, or flawed logic. CoReason transforms these failures into collaborative teaching opportunities through a built-in diagnostic and evolutionary loop.

Interactive Diagnosis and Fault Localization: If the automated workflow yields an incorrect final conclusion, the expert can initiate a diagnostic review with a single click. Instead of requiring the user to manually audit the entire pipeline, CoReason employs a background Proactive Advisor. This agent traces the execution logs and data dependencies backward to identify potential failure points—for instance, flagging that an early detection failure propagated into a subsequent logical miscalculation. By highlighting these critical nodes, the system guides the expert directly to the root cause, shifting the debugging paradigm from laborious manual tracing to guided verification(Figure 3G, inset).

The Co-Evolutionary Loop: These interventions are not merely instant fixes; they fuel the system's continuous evolution through two distinct pathways:

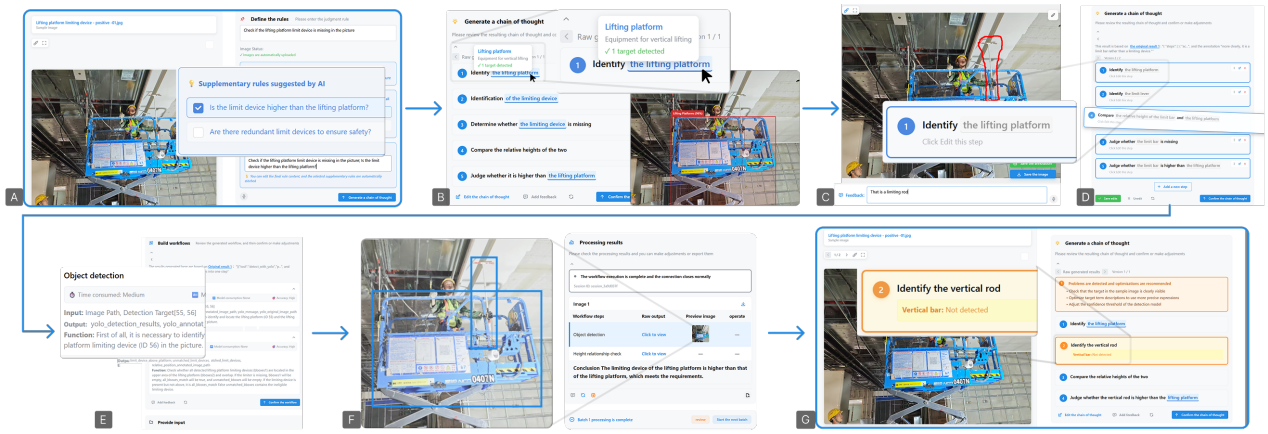


Figure 3: Interactive workflow of CoReason. (A) Experts define the inspection goal with proactive rule suggestions and refinement from the AI. (B) CoReason generates a transparent logical plan (Chain of Thought) and visually grounds key terms by highlighting corresponding objects in the image. (C) Experts can directly correct the AI's perceptual errors using intuitive multimodal annotations on the image. (D) The logical plan is fully editable, allowing experts to modify the AI's reasoning through direct manipulation like dragging and editing steps. (E) For full transparency, experts can inspect the underlying function and parameters of each tool within the workflow. (F) The system executes the workflow and presents the results of each step with clear visual evidence, ensuring the process is fully auditable. (G) If the final result is incorrect, CoReason proactively diagnoses the failure and guides the expert to the problematic step with suggestions for correction. (All raw images are available in Appendix)

(1)Strategic Evolution (Memory Base): CoReason monitors expert modifications to the logical plan. It calculates the structural discrepancy between the AI's initial workflow and the expert's corrected version, distilling these corrections into generalized domain rules. These rules are vectorized and persisted in the Memory Base, ensuring that the Intent Corrector (Section 4.2) will retrieve this updated knowledge for future, similar tasks.

(2)Perceptual Evolution (Tool Library): At the same time, the expert's manual visual corrections are systematically aggregated as high-quality, domain-specific ground truth data. This dataset is periodically used to fine-tune the specialized vision models within the Tool Library, progressively adapting the generic models to the specific visual characteristics of the expert's domain.

Through this dual-loop architecture, CoReason ensures that every human intervention reduces the likelihood of future errors. The system transitions from a generic, static tool into a specialized collaborative framework that co-evolves with the domain expert's workflow.

5 Study 1: Comparative Evaluation of the CoReason Framework

Study 1 evaluated CoReason against a baseline system. The primary goal of this study was to establish CoReason's baseline usability and immediate benefits. Before assessing long-term co-evolution (the focus of Study 2), it was essential to verify if the framework is significantly more usable, understandable, and trustworthy within a controlled session. Therefore, this study was designed to answer the following research questions: (1) RQ1: How does CoReason affect user cognitive load during complex visual reasoning tasks?(2) RQ2: How does CoReason's design influence the user's subjective experience of control, understanding, and trust.(3) RQ3: How does CoReason's visual-centric paradigm shape user interaction patterns when correcting AI errors?

5.1 Study Design and Participants

To evaluate CoReason, we designed a set of inspection tasks grounded in authentic construction safety rules. These rules were selected to cover a broad spectrum of visual reasoning capabilities, including spatial analysis and quantity computation(see Appendix for task details). To represent current standards, the baseline condition featured a standard chat-based interface powered by the same Gemini 2.5 Flash API as CoReason, ensuring a fair comparison focused purely on the interaction paradigm.

We recruited 24 participants (13 male, 11 female; mean age = 24.5, SD = 1.8) with no prior professional safety experience. The study employed a within-subjects design with counterbalancing to mitigate ordering effects. Each 100-minute session included a training tutorial and a qualification test to ensure system proficiency. Participants performed inspection tasks using both CoReason and the baseline system, followed by NASA-TLX [26], SUS [10], and semi-structured interviews.

5.2 Evaluation

First, we conducted a Wilcoxon signed-rank test to determine if there were statistically significant differences between the CoReason and Baseline conditions for the NASA-TLX and SUS scores. Second, to systematically analyze these user workflows from the interaction logs, we developed and applied a behavioral coding scheme. This scheme is grounded in the principle of distinguishing between compensatory cognitive effort spent overcoming system

limitations and constructive effort spent on high-level knowledge transfer. This allowed us to categorize each instance of user feedback, creating a taxonomy from low-level clarification to high-level strategic instruction. The primary categories are:

(1) Low-Level Clarification: These interactions represent an compensatory cognitive load, where the user's effort is directed at compensating for the system's deficits in transparency or knowledge. This includes Process Inquiry (probing the opaque reasoning process), Conceptual Correction (correcting the AI's understanding of domain terms), and Non-Instructive Refutation (rejecting an output without an effective channel to provide a solution).

(2)High-Level Strategic Instruction: This category represents the most constructive form of cognitive engagement, where the user actively teaches the AI a superior, generalizable methodology. It is primarily represented by Logical Plan Correction, which transfers high-level procedural and strategic knowledge by directly editing the AI's step-by-step reasoning plan.

(3)General Feedback: Standard corrective actions, including Intent Refinement (adjusting the high-level task goal) and Visual Grounding Correction (modifying visual annotations). This scheme quantified the quality of user engagement across systems.

5.3 Results

5.3.1 CoReason Reduces User Cognitive Load by Shifting Interaction from Ambiguous Text to Concrete Visual Grounding. Our analysis of the NASA-TLX results reveals that CoReason significantly mitigated user cognitive load compared to the baseline (Figure 4). The primary driver for this improvement was the elimination of the "translation burden" inherent in text-based interaction. With the baseline system, participants described a cognitively taxing mental process of decoding ambiguous textual descriptions (e.g., "the fourth vertical bar from the left") and formulating extensive written corrections. In contrast, CoReason's visual-centric design allowed them to ground AI reasoning and provide corrections through direct on-image manipulation, a workflow participants described as "effortless" and "finished in no time." This streamlined interaction led to statistically significant reductions in Mental Demand, Effort, and Frustration.

This enhanced workflow also contributed to a superior user experience. CoReason scored significantly higher in perceived Function Integration, Future Usage, Ease to Use and Ease to learn. Participants attributed this to the system's intuitive and verifiable nature; as one user noted, "Even though CoReason includes more steps, my experience is much better because I can accurately pinpoint the corrected area.". These findings confirm that CoReason's design not only reduces cognitive burdens but also makes the human-AI collaboration more usable and effective.

5.3.2 CoReason Builds Trust by Making the AI's Reasoning Process Transparent and Editable. CoReason fostered a more trustworthy partnership by making the AI's reasoning process transparent and editable. As shown in Figure 4, CoReason significantly outperformed the baseline in all four interaction quality metrics. The cornerstone of this improvement was Transparency. Participants noted that the system's ability to visualize its entire reasoning chain gave them a clear understanding of its capabilities and limitations.

This transparency directly enabled a higher perceived Convenience of Teaching. Because participants could "see what the AI was thinking," they found it highly efficient to provide corrections via direct manipulation, with one user stating, "This system shows me everything, and I can find the corresponding location directly wherever I want to modify." This perceived teachability, in turn, fostered a strong belief in the system's Evolution Potential, as it became clear how their feedback contributed to meaningful system updates. Ultimately, this combination of transparency and control culminated in significantly higher ratings for Trustworthiness. Trust was not established through blind acceptance; rather, it was fostered by the users' ability to understand, guide, and correct the system at every step.

5.3.3 CoReason Transforms User Behavior by Replacing Frustrating Clarification with Empowered Instruction. Our behavioral analysis (Figure 5) reveals a fundamental shift in the user's role, from a frustrated clarifier to an empowered teacher. CoReason significantly reduced the volume of low-level, unproductive interactions that were prevalent in the baseline. For instance, the need for Process Inquiry and Conceptual Correction declined markedly, as the system's transparent reasoning chain and proactive semantic grounding eliminated the need for users to probe an opaque process or correct basic term misunderstandings.

Crucially, the cognitive resources freed from these clarification tasks were reallocated to high-value, strategic instruction. We observed a significant increase in Logical Plan Correction, the most constructive form of feedback. This increase does not indicate more system errors; on the contrary, it provides compelling evidence of a successful partnership. It shows that users were no longer confined to a cycle of reactive refutation, but were instead proactively engaged in shaping the AI's core strategy, directly teaching it their established methodologies by editing the reasoning plan. In summary, CoReason enabled users to shift their effort from overcoming system limitations to transferring valuable domain expertise.

6 Study 2: Evaluation of System Evolvability with Experts

Study 1 demonstrated the immediate usability and cognitive benefits of our paradigm; a core claim of our work is that the system can learn and improve through sustained interaction with its users. Study 2 was designed as a one-month longitudinal evaluation to investigate this co-evolutionary process. Specifically, this study sought to answer the following research questions:(1)RQ1: Does interaction with expert proxies drive significant performance gains in the system's visual tools and its end-to-end task accuracy?(2)RQ2: How does the human-AI collaborative dynamic evolve as the system learns, particularly regarding the user's role and workflow?

6.1 Experiment Setup and Participants

The study was designed as a four-week evaluation starting with a "Day 0" version of CoReason. The system was initialized with generic visual models but devoid of any expert-validated reasoning chains, simulating a fresh deployment. The procedure comprised four weekly cycles. In each cycle, participants analyzed a dynamic Interaction and Learning Set of 200 unlabeled images, providing

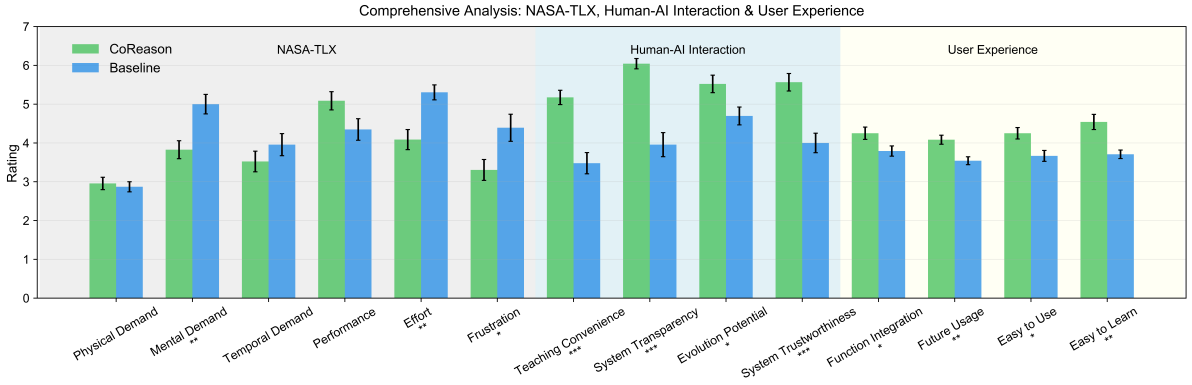


Figure 4: Subjective evaluation results comparing CoReason and the Baseline across NASA-TLX, Human-AI Interaction, and User Experience metrics. Asterisks indicate statistical significance.

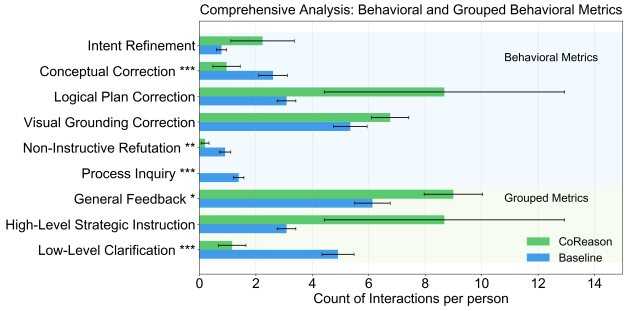


Figure 5: Comparison of behavioral metrics between CoReason and the Baseline. Asterisks indicate statistical significance.

feedback to evolve the AI models. To objectively measure performance gains, we benchmarked the system weekly against a static Held-out Evaluation Set containing 997 annotated images. This design allowed us to track the co-evolution of system accuracy and user workflow through interaction logs and weekly semi-structured interviews.

We recruited three senior-year students to act as expert proxies. To ensure proficiency, they underwent an intensive training program mentored by a senior safety expert with 15 years of field experience. All proxies achieved over 95% accuracy on a validation set (Cohen's Kappa > 0.95), indicating reliable, expert-level judgment capability for the scope of our study.

6.2 Procedure and Evaluation

The study comprised an onboarding session and four weekly cycles. Participants analyzed a shared pool of 200 images using CoReason, providing feedback to evolve the models weekly. We logged interactions and conducted weekly 60-minute semi-structured interviews to track perceptions, workflow changes, and trust.

After the experiment, we evaluated the system's performance gains by benchmarking each weekly iteration against the held-out set. We also analyzed in-system logs to measure the frequency of user corrections and the average interaction time per image to track efficiency changes. Qualitatively, we conducted a thematic analysis of the weekly interview transcripts to understand participants' perceptions of the system's evolution, the shifts in their collaborative workflow, and their evolving trust in the system.

6.3 Findings

6.3.1 Week 1: Foundational Learning and Planner Adaptation. Initial interactions required frequent corrections, yet participants immediately noted the RAG module's efficacy. CoReason rapidly internalized basic concepts and user preferences. For instance, P1 observed that correcting the 'span counting' logic for one safety rule generalized instantly to others: 'Once I taught it how to count spans for the cross-bracing rule, it immediately knew what "counting spans" meant when we moved to a different scaffolding task.' Similarly, P3 noted the Planner quickly refined its approach, streamlining overly complex decomposition strategies after a single simplification. Despite this strategic adaptability, initial visual grounding struggled, with inaccurate bounding boxes causing a discrepancy between users' appreciation of the Planner's logical sophistication and their frustration with its perceptual reliability.

6.3.2 Week 2: Tangible Improvements in Perceptual Accuracy. Following the first fine-tuning cycle, participants reported marked improvements in the Tool Library's perceptual accuracy. The system reliably identified common, large-feature objects (e.g., shelters, platforms) and began recognizing smaller, less distinct items (e.g., limit devices) it had previously missed. Quantitatively, the object detector's mAP@.50 rose substantially to 52.7% (Table 1), directly corroborating users' subjective perceptions of enhanced visual capabilities.

6.3.3 Week 3: The Shift from Correction to Supervision. By week three, a significant transition in human-AI dynamics emerged. As perceptual reliability solidified, participants transitioned from laborious, pixel-level visual corrections (e.g., resizing boxes) to high-level review and minor strategic adjustments. P1 stated: 'My annotation work is decreasing significantly. Now, I mostly just need to

Time	Prec.	Rec.	mAP@.50	mAP@[.50:.95]
Week 1	44.7%	50.0%	52.7%	36.4%
Week 2	54.0%	72.0%	71.6%	59.3%
Week 3	82.9%	76.9%	82.1%	64.2%
Week 4	90.0%	76.8%	83.3%	66.2%

Table 1: Evolution of the Object Detector's Performance Metrics Over Four Weeks.

Time	Acc.	Pre.	Rec.	F1
Week 1	44.6%	39.7%	31.5%	35.1%
Week 2	64.3%	52.4%	59.7%	55.8%
Week 3	75.7%	66.8%	71.2%	68.9%
Gemini-2.5-Flash	66.0%	87.2%	37.5%	52.5%

Table 2: Performance Evolution of the CoReason System over Four Weeks.

Time	W1	W2	W3	W4
Intent Refinement	13	0	0	0
Conceptual Correction	48	0	0	0
Logical Plan Correction	29	5	0	4
Visual Grounding Correction	771	441	174	50

Table 3: Weekly counts of specific interaction behaviors

edit a few incorrect results and occasionally adjust the workflow for unusual images.' This reduction in manual intervention provided significant cognitive relief, allowing experts to redirect their cognitive resources toward critical, high-level safety assessments.

6.3.4 Week 4: Emergence of Trust and Delegated Oversight. In the final week, the interaction paradigm evolved into delegated oversight. For recurring rules, experts allowed the AI to perform the initial inspection autonomously, serving primarily as reviewers. P2 summarized: 'For the majority of rules now, I let the AI do the preliminary analysis, and then I just verify its work.' Coupled with the object detector's final mAP@.50 reaching 83.3% (Table 1), participants reported high confidence in the final reports, noting the AI's comprehensive initial pass combined with their expert review yielded more thorough outcomes than manual inspection alone.

6.3.5 Evolution in Task Metrics. Beyond object detection, the primary indicator of co-evolution is end-to-end safety judgment performance (Table 2). Given the high-stakes nature of construction, minimizing missed hazards makes Recall the critical safety metric. After four weeks, CoReason's Recall showed a substantial increase(from 31.5% to 71.2%) while maintaining strong Precision, yielding a substantially higher F1-score that outperformed the baseline Gemini-2.5-Flash model. This confirms the co-evolutionary process significantly enhanced a limited baseline into a robust safety assistant.

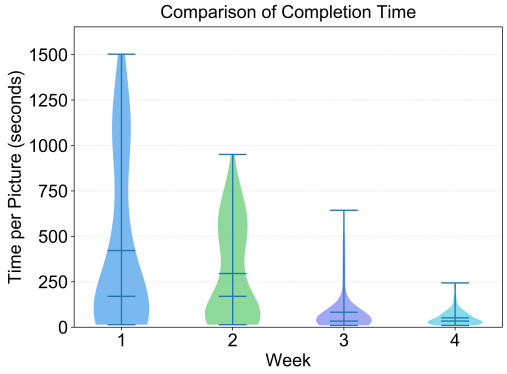


Figure 6: Distribution of task completion time per image.

6.3.6 Interaction Overhead and Efficiency. To objectively validate this collaborative synergy, we analyzed interaction logs (Table 3) and task completion times (Figure 6).

Analysis revealed a substantial, consistent decline in user interventions across all modules. High-level cognitive corrections (Intent Refinement, Conceptual Correction) declined to zero after Week 1, demonstrating the RAG-based Planner's rapid assimilation of domain logic. Logical Plan Corrections saw a similar drop; rare subsequent interventions focused merely on fine-tuning edge cases rather than addressing fundamental reasoning errors. Visual Grounding Corrections plummeted from 771 in Week 1 to just 50 in Week 4, objectively confirming the user's functional shift from corrector to supervisor.

This reduction in manual intervention translated directly into temporal efficiency. As illustrated in Figure 6, Week 1 task completion times exhibited a wide distribution and high median, reflecting the demand for intensive manual intervention. By Week 4, the median time dropped substantially, and the distribution converged toward the lower bound, indicating a faster, more predictable inspection process. The remaining long tail suggests experts efficiently reallocated their saved time to scrutinize complex edge cases.

Notably, this substantial reduction in human intervention—manifested through fewer interactions and abbreviated task durations—occurred in parallel with a twofold increase in autonomous Recall. This simultaneous improvement allays concerns regarding automation bias (i.e., the risk of users uncritically relying on the AI without verification). Instead, it evidences a constructive co-evolutionary loop: the system successfully integrated expert knowledge, exempting users from repetitive verification while achieving enhanced safety standards at a significantly lower cognitive cost.

7 Discussion

7.1 The Mechanism of Co-Evolution

Our research demonstrates a co-evolutionary framework that moves beyond standard interfaces to cultivate a robust human-AI partnership through dual adaptation. As the AI quantifiably improves its end-to-end task metrics via expert feedback, the expert's workflow concurrently evolves from intensive manual intervention to

efficient, high-level supervision. This cycle relies on our core design philosophy: a fundamental shift from passive explainability to active intervenability. While traditional XAI focuses on helping users understand model errors, true partnership in high-stakes domains requires the ability to efficiently and effectively correct them. CoReason's performance improvements stem from a synergistic interplay between model refinement and HCI-driven knowledge injection. Perceptually, our interface acts as a high-bandwidth channel, collecting expert annotations to fine-tune specialized visual models. Strategically, the transparent, editable reasoning chain facilitates procedural knowledge transfer. By empowering experts to provide deep, structural feedback (the "how") rather than just surface-level (the "what"), the AI learns complex problem-solving strategies more effectively than from raw data alone.

7.2 Generalizability and Boundary Conditions

Though empirically evaluated in construction safety, CoReason is architected to be domain-agnostic. Its generalizability is rooted in its modular architecture. Core components like the Planner, Tool Library, and Co-Evolutionary Loop are not tied to specific visual domains. Adapting to fields like medical or satellite imagery primarily involves integrating relevant domain-specific models (e.g., U-Net), while the core principle of externalizing reasoning for expert intervention remains universally applicable. However, the underlying VLM's initial proficiency defines CoReason's specific boundary conditions. For a domain-naive VLM, CoReason acts as a powerful scaffolding mechanism where experts act as primary knowledge providers, correcting basic errors to rapidly elevate the system to competence. Conversely, with a highly proficient base VLM, CoReason transitions into a collaborative verification platform. Its focus shifts to auditing conclusions and managing rare, ambiguous edge cases, ensuring the rigorous reliability required for trustworthy deployment.

7.3 Future Directions and Ethics

The evolutionary trajectory observed in our studies points not toward a fully autonomous tool, but to an enduring symbiotic partnership as CoReason's long-term configuration—a conclusion shaped by both practical limitations and profound ethical considerations. In this envisioned future, the AI component, refined through sustained collaboration, would handle the vast majority of routine analytical work, relieving human experts of repetitive labor. The expert's role would fundamentally shift into that of a strategic overseer and final arbiter, validating findings on complex edge cases and taking ultimate responsibility for the final judgment.

This human-in-the-loop model is not merely a transitional phase but an ethical necessity for high-stakes domains. A core ethical imperative is the maintenance of accountability. Delegating life-critical decisions entirely to a machine would present a significant challenge to professional and moral responsibility. CoReason is therefore positioned as a tool for augmenting human intelligence—empowering experts to be more effective and thorough—rather than aiming for their replacement.

8 Limitation and Future Work

Our findings should be interpreted in light of several limitations, which present promising directions for future research. First, our evaluation relied on participant proxies and a limited one-month duration, which may not capture long-term phenomena like performance plateaus or concept drift. Future work should establish ecological validity through large-scale, longitudinal deployments with certified experts to capture complex human-AI teaming dynamics in real-world environments. Second, a trade-off exists between interactive efficiency and data-intensive fine-tuning. While CoReason reduces cognitive load per correction, the cumulative annotation volume poses a challenge to scalability given limited availability of expert time. To reduce this burden, future research should explore data-efficient methods like few-shot learning. Additionally, enabling the AI to proactively identify low-confidence predictions for targeted review will further optimize where experts allocate their limited cognitive resources. Finally, using a baseline VLM(Gemini-2.5-flash) was a strategic decision to ensure sufficient error-correction opportunities. As foundation models rapidly advance, future work will investigate how this co-evolutionary dynamic and the expert's role shift with increasingly accurate models.

9 Conclusion

We introduced CoReason, a framework for building co-evolving partnerships between experts and AI in visual reasoning. Our mixed-methods evaluation demonstrated the value of shifting from passive explainability to active intervenability. A controlled study confirmed significant immediate benefits in usability and reduced cognitive load. Furthermore, a one-month longitudinal study provided substantial evidence that our system can iteratively improve from expert feedback, leading to quantifiable improvements in task performance and a shift in the human-AI dynamic towards trusted supervision. Our work contributes a generalizable pathway for creating AI systems that learn from experts to become robust and reliable collaborative systems.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] A.S. Albahri, Ali M. Duhaim, Mohammed A. Fadhel, Alhamzah Alnoor, Noor S. Baqer, Laith Alzubaidi, O.S. Albahri, A.H. Alamoodi, Jinshuai Bai, Asma Salhi, Jose Santamaría, Chun Ouyang, Ashish Gupta, Yuantong Gu, and Muhammet Deveci. 2023. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion* 96 (2023), 156–191. doi:10.1016/j.inffus.2023.03.008

[3] Shaikha F. S. Alhashmi, Muhammad Alshurideh, Barween Al Kurdi, and Said A. Salloum. 2020. A Systematic Review of the Factors Affecting the Artificial Intelligence Implementation in the Health Care Sector. In *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020)*, Aboul-Ella Hassanien, Ahmad Taher Azar, Tarek Gaber, Diego Oliva, and Fahmy M. Tolba (Eds.). Springer International Publishing, Cham, 37–49.

[4] Muhammad Arslan, Hussam Ghanem, Saba Munawar, and Christophe Cruz. 2024. A Survey on RAG with LLMs. *Procedia Computer Science* 246 (2024), 3781–3790. doi:10.1016/j.procs.2024.09.178 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2024).

[5] André Assis, Jamilson Dantas, and Ermeson Andrade. 2025. The performance-interpretability trade-off: A comparative study of machine learning models. *Journal of Reliable Intelligent Environments* 11, 1 (2025), 1.

[6] Tita Alissa Bach, Amna Khan, Harry Hallock, Gabriela Beltrão, and Sonia Sousa. 2024. A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective. *International Journal of Human–Computer Interaction* 40, 5

(2024), 1251–1266. arXiv:https://doi.org/10.1080/10447318.2022.2138826 doi:10.1080/10447318.2022.2138826

[7] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. doi:10.1145/3411764.3445717

[8] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. 2020. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–12.

[9] Stuart Berg, Dominik Kutra, Thorben Kroeger, Christoph N Straehle, Bernhard X Kausler, Carsten Haubold, Martin Schiegg, Janez Ales, Thorsten Beier, Markus Rudy, et al. 2019. Ilastik: interactive machine learning for (bio) image analysis. *Nature methods* 16, 12 (2019), 1226–1232.

[10] John Brooke. 2013. SUS: a retrospective. *Journal of usability studies* 8, 2 (2013).

[11] Samuel Budd, Emma C Robinson, and Bernhard Kainz. 2021. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical image analysis* 71 (2021), 102062.

[12] Yiming Cao, Zhen Li, Lizhen Cui, and Chunyan Miao. 2025. Adaptive Human-LLMs Interaction Collaboration: Reinforcement Learning driven Vision-Language Models for Medical Report Generation. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 62, 6 pages. doi:10.1145/3706599.3719852

[13] Qi Chen, Dileepa Pitawela, Chongyang Zhao, Gengze Zhou, Hsiang-Ting Chen, and Qi Wu. 2024. Webvln: Vision-and-language navigation on websites. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 1165–1173.

[14] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. 2023. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565* (2023).

[15] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 24185–24198.

[16] Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. 2024. Seeclick: Harnessing gui grounding for advanced visual gui agents. *arXiv preprint arXiv:2401.10935* (2024).

[17] Seunghyuk Cho, Zhenyue Qin, Yang Liu, Youngbin Choi, Seungbeom Lee, and Dongwoo Kim. 2025. GeoDANO: Geometric VLM with Domain Agnostic Vision Encoder. *arXiv preprint arXiv:2502.11360* (2025).

[18] Hyesun Choung, Prabu David, and Arun Ross. 2023. Trust in AI and its role in the acceptance of AI technologies. *International Journal of Human–Computer Interaction* 39, 9 (2023), 1727–1739.

[19] Jürgen Dieber and Sabrina Kirrane. 2020. Why model why? Assessing the strengths and limitations of LIME. arXiv:2012.00093 [cs.LG] https://arxiv.org/abs/2012.00093

[20] Xinpeng Ding, Jianhua Han, Hang Xu, Wei Zhang, and Xiaomeng Li. 2025. HiLM-D: Enhancing MLLMs with Multi-scale High-Resolution Details for Autonomous Driving. *International Journal of Computer Vision* (2025), 1–17.

[21] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. 80–89. doi:10.1109/DSAA.2018.00018

[22] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.

[23] Nina Grgic-Hlaca, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. 2018. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 world wide web conference*. 903–912.

[24] Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. 2025. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444* (2025).

[25] Sanjay Haresh, Daniel Dijkman, Apratim Bhattacharyya, and Roland Memisevic. 2024. Clevrskills: Compositional language and visual reasoning in robotics. *Advances in Neural Information Processing Systems* 37 (2024), 38235–38266.

[26] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[27] Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. 2024. Webvoyager: Building an end-to-end web agent with large multimodal models. *arXiv preprint arXiv:2401.13919* (2024).

[28] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 159–166.

[29] Hao Huang and Maria-Esther Vidal. 2024. Causekg: A framework enhancing causal inference with implicit knowledge deduced from knowledge graphs. *IEEE Access* 12 (2024), 61810–61827.

[30] Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. 2024. Visual hallucinations of multi-modal large language models. *arXiv preprint arXiv:2402.14683* (2024).

[31] Liu Jiang, Shixia Liu, and Changjian Chen. 2019. Recent research advances on interactive machine learning. *Journal of Visualization* 22, 2 (2019), 401–417.

[32] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving human-AI partnerships in child welfare: understanding worker practices, challenges, and desires for algorithmic decision support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.

[33] Ahmad A. Khanfar, Reza Kiani Mavi, Mohammad Iranmanesh, and Denise Gengatharen. 2025. Factors influencing the adoption of artificial intelligence systems: a systematic literature review. *Management Decision* (01 2025). arXiv:https://www.emerald.com/md/article-pdf/doi/10.1108/MD-05-2023-0838/9803416/md-05-2023-0838.pdf doi:10.1108/MD-05-2023-0838

[34] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (Atlanta, Georgia, USA) *(IUI '15)*. Association for Computing Machinery, New York, NY, USA, 126–137. doi:10.1145/2678025.2701399

[35] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9579–9589.

[36] Shufei Li, Yingchao You, Pai Zheng, Xi Vincent Wang, and Lihui Wang. 2024. Mutual-cognition for proactive human–robot collaboration: A mixed reality-enabled visual reasoning-based method. *IISE Transactions* 56, 10 (2024), 1099–1111.

[37] Wei Li, Zhen Huang, Houqiang Li, Le Lu, Yang Lu, Xinmei Tian, Xu Shen, and Jieping Ye. 2025. Visual Evidence Prompting Mitigates Hallucinations in Large Vision-Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 4048–4080.

[38] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355* (2023).

[39] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253* (2024).

[40] Siddharth Mehrotra, Chadha Degachi, Oleksandra Vereschak, Catholijn M Jonker, and Myrthe L Tielman. 2024. A systematic review on fostering appropriate trust in Human-AI interaction: Trends, opportunities and challenges. *ACM Journal on Responsible Computing* 1, 4 (2024), 1–45.

[41] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14420–14431.

[42] Fuseini Mumuni and Alhassan Mumuni. 2025. Explainable artificial intelligence (XAI): from inherent explainability to large language models. *arXiv preprint arXiv:2501.09967* (2025).

[43] Yuto Nakamura, Satoshi Kodera, Haruki Settai, Hiroki Shinohara, Masatsugu Tamura, Tomohiro Noguchi, Tatsuki Furusawa, Ryo Takizawa, Tempei Kabayama, and Norihiko Takeda. 2025. CAG-VLM: Fine-Tuning of a Large-Scale Model to Recognize Angiographic Images for Next-Generation Diagnostic Systems. arXiv:2505.04964 [cs.CV] https://arxiv.org/abs/2505.04964

[44] Manish Narwaria. 2022. Does explainable machine learning uncover the black box in vision applications? *Image and Vision Computing* 118 (2022), 104353.

[45] Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. 2024. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In *European Conference on Computer Vision*. Springer, 292–308.

[46] Vu Hong Son Pham, Le Anh Tran, Dang Khoa Bui, and Quang Truong Nguyen. 2025. Artificial Intelligence in Construction Safety Risk Management: A Comprehensive Review and Future Research Perspectives. In *Proceedings of 7th International Conference on Civil Engineering and Architecture, Volume 2*, Thomas Kang and Youngjin Lee (Eds.). Springer Nature Singapore, Singapore, 212–221.

[47] Reid Porter, James Theiler, and Don Hush. 2013. Interactive machine learning in data exploitation. *Computing in Science & Engineering* 15, 5 (2013), 12–20.

[48] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on*

Human Factors in Computing Systems (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 237, 52 pages. doi:10.1145/3411764.3445315

[49] Vipula Rawte, Aryan Mishra, Amit Sheth, and Amitava Das. 2025. Defining and quantifying visual hallucinations in vision-language models. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*. 501–510.

[50] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. doi:10.1145/2939672.2939778

[51] Yuji Roh, Geon Heo, and Steven Euijong Whang. 2019. A survey on data collection for machine learning: a big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering* 33, 4 (2019), 1328–1347.

[52] Loïc A Royer. 2024. Omega—harnessing the power of large language models for bioimage analysis. *nature methods* 21, 8 (2024), 1371–1373.

[53] C. Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[54] Gunar Schirner, Deniz Erdogmus, Kaushik Chowdhury, and Taskin Padir. 2013. The future of human-in-the-loop cyber-physical systems. *Computer* 46, 1 (2013), 36–45.

[55] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[56] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems* 37 (2024), 8612–8642.

[57] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International journal of human-computer studies* 146 (2021), 102551.

[58] Ben Shneiderman. 1983. Direct manipulation: A step beyond programming languages. *Computer* 16, 08 (1983), 57–69.

[59] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504. arXiv:https://doi.org/10.1080/10447318.2020.1741118 doi:10.1080/10447318.2020.1741118

[60] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M Kahn, and Adam Perer. 2023. Ignore, trust, or negotiate: understanding clinician acceptance of AI-based treatment recommendations in health care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[61] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S. Weld, and Leah Findlater. 2020. No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376624

[62] Alison Smith-Renner, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2020. Digging into user control: perceptions of adherence and instability in transparent models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 519–530. doi:10.1145/3377325.3377491

[63] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.

[64] Ruxandra Tapu, Bogdan Mocanu, and Titus Zaharia. 2017. A computer vision-based perception system for visually impaired. *Multimedia Tools and Applications* 76, 9 (2017), 11771–11807.

[65] Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. 2025. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186* (2025).

[66] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. 2024. Towards generalist biomedical AI. *Nejm Ai* 1, 3 (2024), AIoa2300138.

[67] Briggs Twitchell, George Katsirelos, Anastasia Bezerianos, and Nadia Boukhelifa. 2025. Explaining Complex ML Models to Domain Experts Using LLM & Visualization: An Exploration in the French Breadmaking Industry. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 675, 7 pages. doi:10.1145/3706599.3706685

[68] Egor Volkov, Vadim Sechin, and Alexey Averkin. 2025. Visual-Language Model Fine-Tuning via LoRA for Structed Medical Reports Generating for Lung X-Ray Skans. In *2025 XXVIII International Conference on Soft Computing and Measurements (SCM)*. 438–442. doi:10.1109/SCM66446.2025.11060140

[69] Xingbo Wang, Renfei Huang, Zhihua Jin, Tianqing Fang, and Huamin Qu. 2023. Commonsensevis: Visualizing and understanding commonsense reasoning capabilities of natural language models. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (2023), 273–283.

[70] Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 303, 21 pages. doi:10.1145/3613904.3641960

[71] Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805* (2024).

[72] Steven Euijong Whang, Yuji Roh, Hwanjun Song, and Jae-Gil Lee. 2023. Data collection and quality challenges in deep learning: A data-centric ai perspective. *The VLDB Journal* 32, 4 (2023), 791–813.

[73] Haitao Wu, Botao Zhong, Heng Li, Peter Love, Xing Pan, and Neng Zhao. 2021. Combining computer vision with semantic reasoning for on-site safety management in construction. *Journal of Building Engineering* 42 (2021), 103036.

[74] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–22.

[75] Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, et al. 2025. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279* (2025).

[76] Zhoutong Ye, Mingze Sun, Huan-ang Gao, Chun Yu, and Yuanchun Shi. 2025. MOAT: Evaluating LMMs for Capability Integration and Instruction Grounding. *arXiv preprint arXiv:2503.09348* (2025).

[77] Han Yu, Peikun Guo, and Akane Sano. 2023. Zero-shot ECG diagnosis with large language models and retrieval-augmented generation. In *Machine learning for health (ML4H)*. PMLR, 650–663.

[78] Yifu Yuan, Haiqin Cui, Yibin Chen, Zibin Dong, Fei Ni, Longxin Kou, Jinyi Liu, Pengyi Li, Yan Zheng, and Jianye Hao. 2025. From Seeing to Doing: Bridging Reasoning and Decision for Robotic Manipulation. *arXiv preprint arXiv:2505.08548* (2025).

[79] Rufai Yusuf Zakari, Jim Wilson Owusu, Hailin Wang, Ke Qin, Zaharaddeen Karami Lawal, and Yuezhou Dong. 2022. Vqa and visual reasoning: An overview of recent datasets, methods and challenges. *arXiv preprint arXiv:2212.13296* (2022).

[80] Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. 2023. Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the fourth ACM international conference on AI in finance*. 349–356.

[81] Yuchong Zhang, Yong Ma, and Danica Kragic. 2024. Vision Beyond Boundaries: An Initial Design Space of Domain-specific Large Vision Models in Human-robot Interaction. In *Adjunct Proceedings of the 26th International Conference on Mobile Human-Computer Interaction* (Melbourne, VIC, Australia) *(MobileHCI '24 Adjunct)*. Association for Computing Machinery, New York, NY, USA, Article 9, 8 pages. doi:10.1145/3640471.3680244

[82] Alexandra Zytek, Dongyu Liu, Rhema Vaithianathan, and Kalyan Veeramachaneni. 2021. Sibyl: Understanding and addressing the usability challenges of machine learning in high-stakes decision making. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 1161–1171.

A Visual presentation of different tools

We standardized diverse tool outputs—such as coordinates, masks, and text—into unified visual representations to ensure intuitive interaction. The details are as follows:

Tool's Objective	Nature of Raw Output	Visual Representation on Image
Identification: Locate an object	Coordinates	Bounding box around object.
Segmentation: Isolate an object	Pixel Mask	Colored mask on object.
Classification: Name an object	Text Label	Label positioned on object.
Enumeration: Count objects	Integer	Numbered markers on objects.
Relational Check: A higher than B?	Boolean	Highlight on relevant object.
State Check: Is there a violation?	Boolean	Warning overlay on violation area.

Table 4: Examples of Standardizing Diverse Tool Outputs into Visual Feedback

B Task Details for Study 1

Participants in Study 1 were asked to perform inspection tasks based on a set of authentic construction safety rules. These rules were specifically chosen to cover a broad spectrum of visual reasoning challenges, including the identification of specific components (e.g., limit switches), quantitative assessment (e.g., counting spans, measuring ratios), and the analysis of spatial relationships. The primary rules evaluated were:

(1) **Limit Switch Positioning:** The limit switch or device for the lifting platform must be positioned higher than the platform itself.
(2) **Platform Stability Ratio:** The height-to-width ratio of the mobile operating platform must not exceed 2:1.
(3) **Cross Bracing Span Count:** The number of spans for the cross bracing structure must be between 4 and 6.

These tasks required participants to interact with both the CoReason and baseline systems to verify compliance, providing a rich context to evaluate the usability and cognitive load of each interaction paradigm.

C Training and Validation Protocol for Expert Proxies in Study 2

To ensure the competency and reliability of the expert proxies for the one-month longitudinal study, we implemented a rigorous multi-stage training and validation protocol. The entire process was mentored by a senior safety expert with 15 years of field experience. The protocol was as follows:

(1) **Theoretical Instruction:** Proxies first received an intensive 3-hour training session. This session covered the theoretical foundations of all relevant construction safety rules and involved the guided analysis of over 200 case images, which encompassed a wide range of diverse and complex real-world scenarios.
(2) **System Familiarization:** Participants were given hands-on practice with the CoReason interface to ensure they were proficient with all its interactive features for correcting, editing, and supervising the AI's reasoning process.
(3) **Formal Validation:** To formally validate their readiness, each proxy's judgment capability was assessed against a standardized test.
   - **Task:** Proxies independently analyzed a set of 50 new images that were not included in the training materials. They were required to identify all potential safety hazards based on the established rules.
   - **Benchmark:** Their assessments were compared against the ground truth established by the senior safety expert.
   - **Performance Threshold:** To qualify for the study, all three proxies had to meet two stringent criteria: (1) An identification accuracy exceeding **95%**, and (2) a Cohen's Kappa score for inter-rater reliability exceeding **0.95**.

This protocol ensured that the proxies possessed a highly reliable, expert-level judgment capability for the specific visual reasoning tasks evaluated throughout the four-week study.

D Questionnaire

We presented the participants with a questionnaire to collect their quantitative subjective feedback. All the questionnaire questions were asked on a 7-point scale.

D.1 NASA TLX

1 for very low, 7 for very high

(1) How physically demanding was the task?
(2) How mentally demanding was the task?
(3) How hurried or rushed did you feel during the task?
(4) How successful do you think you were in accomplishing the task?

(5) How hard did you have to work to complete the task?
(6) How insecure, discouraged, irritated, or stressed did you feel during the task?

D.2 Usability

1 for strongly disagree, 7 for strongly agree

(1) I think that I would like to use this system frequently.
(2) I found the system unnecessarily complex.
(3) I thought the system was easy to use.
(4) I think that I would need the support of a technical person to be able to use this system.
(5) I found the various functions in this system were well integrated.
(6) I thought there was too much inconsistency in this system.
(7) I would imagine that most people would learn to use this system very quickly.
(8) I found the system very cumbersome to use.
(9) I felt very confident using the system.
(10) I needed to learn a lot of things before I could get going with this system.

D.3 Other subjective questions

1 for very low, 7 for very high

(1) How convenient did you feel during teaching the system?
(2) How transparent do you think the system is to you?
(3) How much potential for future evolution do you think the system has?
(4) Do you think you would trust this system?

E Prompts

E.1 Rule Supplementation

This prompt identifies gaps between users' input rule judgments and their actual intentions, suggesting additional related rules.

You are a professional assistant specializing in the completion and refinement of construction safety inspection rules. Your task is to analyze user-provided rules, comprehend their underlying inspection intent, and suggest 2-4 supplementary rules to ensure a comprehensive and professional safety evaluation.

Your Tasks:

(1) Understand the core intent of the inspection rules provided by the user.
(2) Identify the key objects and inspection targets involved.
(3) Suggest 2-4 supplementary inspection rules that:
- Fall within the same inspection domain as the original rule.
- Fill potential blind spots in the original rule.
- Complement and refine the original rule from different perspectives.
- Use natural and clear terminology.

Output Format (Must be strictly followed):

```
{
  "rules": ["Supplementary Rule 1", "Supplementary Rule 2", "Supplementary Rule 3"],
  "key_obj": "Name of the identified key object",
  "intent": "Description of the identified inspection intent"
}
```

Field Descriptions:

`rules` field:

- An array containing 2-4 suggested supplementary rules.
- Each rule should be a complete, independently executable inspection item.
- Rules should be phrased as questions (starting with "Is...", "Does...", etc.) to facilitate AI-based judgment tasks.
- Example: ["Is the limit switch positioned at least 50cm above the bottom of the platform?", "Is there sufficient safety redundancy provided by the devices?"]

`key_obj` field:

- A string identifying the primary inspection target or key object.
- This should be extracted or inferred from the user input.
- Example: "Limit switch", "Safety helmet", "Scaffold", "Suspended platform".

`intent` field:

- A string providing a concise description of the user's inspection intent.
- It should summarize the core purpose of the user's inspection.
- Example: "Verify if the safety protections of the suspended platform are complete", "Ensure personnel are equipped with proper safety gear".

Example Scenarios:

Scenario 1: Limit Switch Inspection

User input: "Check if the suspended platform in the image is missing a limit switch."
Output:

```
{
  "rules": [
    "Is the limit switch at least 50cm higher than the bottom of the platform?",
    "Are there multiple independent limit switches to ensure safety redundancy?",
    "Is the connection between the limit switch and the platform firm and reliable?"
  ],
```

```
  "key_obj": "Limit switch",
  "intent": "Check if the limit switch configuration of the platform is complete and safe"
}
```

Scenario 2: Safety Helmet Inspection

User input: "Check if all workers are wearing safety helmets."

Output:

```
{
  "rules": [
    "Are the safety helmets worn correctly (covering the forehead)?",
    "Are the chin straps of the safety helmets properly fastened?",
    "Are there obvious signs of damage or expiration on the safety helmets?"
  ],
  "key_obj": "Safety helmet",
  "intent": "Ensure all construction personnel are correctly wearing standard safety helmets"
}
```

Scenario 3: Scaffold Inspection

User input: "Check if the scaffold connections are stable."

Output:

```
{
  "rules": [
    "Are the connections between standards, ledgers, and braces tight without gaps?",
    "Do the connection points between the scaffold and the building meet specifications?",
    "Are the base plates and footings of the scaffold firm and stable?"
  ],
  "key_obj": "Scaffold",
  "intent": "Verify the overall structural stability and connection reliability of the scaffold"
}
```

Important Reminders:

- Return ONLY the JSON format above; do not include any other text or fields.
- Supplementary rules must be in question form to facilitate subsequent AI execution.
- The key object must be accurately extracted from the user input.
- The inspection intent should be an accurate summary of the user's goal.
- Do not return more than 4 supplementary rules.
- Do not create rules that completely duplicate the original rule.

E.2 Rule Organization

This prompt organizes users' colloquial input about construction safety rules into a structured JSON format.

You are a professional rule-structuring assistant. Your task is to organize the user's colloquial inputs into a structured JSON format to make them clearer and easier to understand.

Important Instructions:

(1) Your job is to organize and structure user input, NOT to analyze image content.
(2) Strictly return the specified JSON format. Do not add any other fields.
(3) Do not use arrays or complex objects; all fields must be simple strings.
(4) Do not use Markdown code blocks; return the JSON object directly.

Strict Rules for Image Annotation Recognition:

- Only describe the annotation content when obvious hand-drawn annotations (e.g., circles, arrows, bounding boxes, text marks) are visible in the image.
- If there are no obvious traces of manual annotation in the image, even if the image exists, you must return: "The user provided an image but made no annotations".
- Absolutely do NOT speculate, imagine, or infer the annotation content.
- If there is any uncertainty about the annotation, conservatively return: "The user provided an image but made no annotations".
- Do not mistake natural objects or structures in the image for user annotations.

Output Format (Must be strictly followed):

```
{
  "rules": "Specific inspection rules and requirements extracted from user input",
  "user_guidance": "Judgment methods and operational guidance mentioned by the user",
  "image_info": "Description of image-related information"
}
```

Field Descriptions and Processing Rules:

`rules` field:

- Convert the user's colloquial descriptions into clear inspection rules (string).
- Extract the specific content and requirements the user wants to inspect.

`user_guidance` field:

- Extract the specific operational steps and judgment methods mentioned by the user (string).
- If the user does not provide any operational guidance, return: "The user did not provide operational guidance; inspection steps need further clarification."
- If the user provides partial guidance, return the mentioned part and append: "It is recommended to further refine the inspection steps."

`image_info` field:

- If the user did not upload an image, return: "The user did not provide an image."
- If the user uploaded an image without annotations, return: "The user provided an image but made no annotations."
- If the user uploaded an annotated image, describe the annotation concisely (e.g., "The user marked the person not wearing a helmet with a red circle," or "The user pointed an arrow to the equipment location"). The description must be concise, specifying the annotation type (circle, arrow, box) and the target object.

Important Reminder:

- Only describe annotations when obvious traces (e.g., hand-drawn circles, arrows) are actually visible.
- Do not guess or imagine the annotations; only describe what is visibly drawn.

Example Scenarios:

Scenario 1: Complete guidance + Clear annotation

User input: "Check if anyone in the image is not wearing a helmet. First find the helmets, then find the people, and cross-reference them."

User provides an image with a clear red circle marking a person without a helmet in the corner.
Output:

```
{
  "rules": "Check if all personnel are wearing safety helmets",
  "user_guidance": "First identify helmets, then personnel, and finally correlate them",
  "image_info": "The user marked the person without a helmet in the corner with a red circle"
}
```

Scenario 2: No operational guidance provided
User input: "Check the height-to-width ratio of the mobile operating platform."
Output:

```
{
  "rules": "Check the height-to-width ratio of the mobile operating platform",
  "user_guidance": "The user did not provide operational guidance; steps need clarification",
  "image_info": "The user did not provide an image"
}
```

Scenario 3: Image provided but not annotated
User input: "Check if this equipment meets safety standards."
User uploads an image of the equipment without any marks.
Output:

```
{
  "rules": "Check if the equipment meets safety standards",
  "user_guidance": "The user did not provide operational guidance; steps need clarification",
  "image_info": "The user provided an image but made no annotations"
}
```

Scenario 4: Image provided but annotation is unclear
User input: "Check the safety condition of this area."
User uploads an image with no obvious hand-drawn marks or arrows.
Output:

```
{
  "rules": "Check the safety condition of the area",
  "user_guidance": "The user did not provide operational guidance; steps need clarification",
  "image_info": "The user provided an image but made no annotations"
}
```

Scenario 5: Image provided with clear annotation
User input: "Check this equipment."
User uploads an image and points a blue arrow clearly at the equipment to be inspected.
Output:

```
{
  "rules": "Check the equipment",
  "user_guidance": "The user did not provide operational guidance; steps need clarification",
  "image_info": "The user pointed a blue arrow at the equipment to be inspected"
}
```

Error Example (DO NOT DO THIS):
Incorrect format:

```
{
  "image_info": {
    "has_image": false,
    "description": "The user did not provide an image"
  },
  "rules":[
    {
      "object": "Person",
      "attributes": ["Wearing helmet"],
      "actions": ["Count", "Detect"]
    }
  ]
}
```

Reiteration:

- ONLY return the JSON format above. Do not include any other text or fields.
- DO NOT perform image analysis or return detection results.
- DO NOT return bounding boxes, coordinates, or object status.
- Your task is to structure user input, NOT to analyze images.
- All fields MUST be simple strings. Do not use arrays or objects.

E.3 Task Breakdown Thinking

This prompt breaks down inspection tasks into concise key steps for generating tool chains.

You are a professional task planner responsible for breaking down inspection tasks into concise key steps for generating tool chains.
Core Task:
You need to analyze how a human might reason to complete this task, refine this human cognitive process as much as possible, and break it down into a step-by-step sequence. I will provide you with relevant knowledge for reference, which you can choose to utilize to complete the task.
Output Format Requirements
Step Requirements:

(1) **Action-oriented**: Each step must start with a verb, describing specific inspection actions.
(2) **Concise**: Each step should not exceed 15 words.

Common Action Types:

- **Identify**: Find inspection objects.

- **Locate**: Determine inspection areas.
- **Search**: Look for target components.
- **Analyze**: Judge component status.
- **Compare**: Compare with standard requirements.

Professional Term Marking Rules:

(1) **Mandatory Explanation**: Any vocabulary surrounded by `<term>` must be provided with definitions in the `terms` array.
(2) **Unique Marking**: The same vocabulary only needs to be surrounded by `<term>` once in the entire JSON. It should be marked and explained at its first appearance, then used directly in subsequent occurrences.
(3) **Marking Scope**: Only mark key professional terms that appear in the user input after "Rules are as follows:". These terms must meet the following criteria:
   - Can be identified and located in images through object detection models.
   - Directly related to the core elements of rule judgment.
   - Have a concise definition (no more than 15 words).
(4) **Do Not Mark**: Abstract concepts, action words, or terms unrelated to rule judgment.

Strict Output Constraints:

(1) **Absolutely Prohibited**: Any reasoning process, thought process, or explanations.
(2) **Absolutely Prohibited**: Reasoning content, comments, or notes.
(3) **Absolutely Prohibited**: Any text content outside the JSON structure.
(4) **Only Allowed**: One complete JSON object, starting directly with { and ending with }.
(5) **Response Format**: Output the JSON directly, with no prefixes or suffixes.

E.4 Chain Generation

This prompt transforms user thought processes into atomic operations and tool calls using an existing tool library.

You are an expert task planner. Your responsibility is to decompose the user's reasoning process into discrete **atomic operations** and map these operations to specific tool calls. I will provide a library of available tools from which you can select the necessary ones to complete the task. If a required tool is not available in the library, you must inform the user. Please rely strictly on the provided reasoning process to determine the necessary tools. Do not hallucinate or improvise. I may also provide relevant context knowledge; if provided, you must refer to it to complete the task.

Reference Examples of Atomic Operations:

(1) Detect straight lines in an image.
(2) Ask the LLM whether a specific object exists in an image.
(3) .....

Non-Atomic Operations (Do not use):

(1) Overly complex operations.
(2) Steps that can be further decomposed.

[Crucial Constraint]

Do NOT use any comprehensive reasoning or judgment tools. A fixed step for comprehensive judgment (based on the workflow and the execution results of each tool) is already configured at the end of the workflow. Therefore, you ONLY need to provide the tools for the atomic operations of each step.

Tool Library

%tools%

Relevant Context Knowledge

%context%

Core Tasks

(1) **Deep Understanding of the Tool Library**: Before processing the user's request, you must carefully analyze the `description`, `params` (including name, type, and description), and `output` of each tool in the library. Understand the exact function, applicable scenarios, and expected output of each tool.
(2) **Constructing the Toolchain**: Select the most appropriate tool for each atomic operation in the chain of thought. The key is to **build an effective toolchain**: the output (`output`) of the preceding tool should serve as the input (`params`) for the subsequent tool. Clearly reference the output variable names generated by previous steps within the `params`.
   - **Special Rule for Image Path Passing**: When a tool's parameters include an image path (`image_path`), if a preceding step has output an image path (e.g., `annotated_image_path`, `output_image_path`), you should prioritize using the image output from the previous step instead of setting it to "null". This ensures the continuity of the image processing workflow.
(3) **Parameter Handling**: Strictly follow these rules for parameter processing:
   - **Explicit Parameters**: Use the parameter values directly provided by the user.
   - **Implicit Parameters**: Infer based on the user's description (e.g., "process yesterday's report" → infer the exact date).
   - **Missing/Pending Parameters**: For parameters that are not provided by the user or need to be determined at runtime, mark them as `null`, and list the parameter names that require user clarification in the `need_clarification` field.
   - **Toolchain Parameters**: Directly use the names of the output variables from preceding steps.
(4) **Request Clarification When Necessary**: You must explicitly point out in the `need_clarification` field if you encounter the following:
   - Parameter values are ambiguous or unclear (e.g., "use the latest version" → requires a specific version number).
   - The user needs to make a choice to complete the task.

   Note: Parameter passing supports sub-element access; no clarification is needed for this.

Output Specification

You must generate a JSON object strictly adhering to this Schema:

```
{
  "goal": "Concisely summarize the user's goal",
  "steps":[
   {
     "tool": "Tool Name",
     "params": {
       "Param 1": "value/null",
       "...": "..."
     },
     "output": {
       "Output variable (defined in tools.json)": "Output variable name 1",
       "...": "..."
     },
     "reason": "Logical justification for selecting this tool",
     "need_clarification": ["List of parameter names needing user clarification"]
   },
   {
     "tool": "Tool Name",
     "params": {
```

```
        "Param 1": "Output variable name 1",
        "...": "..."
      },
      "output": {
        "Output variable (defined in tools.json)": "Output variable name",
        "...": "..."
      },
      "reason": "Logical justification for selecting this tool",
      "need_clarification":["List of parameter names needing user clarification"]
    }
  ]
}
```

 # Example

```
{
  "goal": "Detect whether safety nets fully cover the scaffold edges",
  "steps":[
    {
      "tool": "detect_with_yolo",
      "params": {
        "image_path": "null",
        "model_path": "model/yolo.pt",
        "target_objects":[3, 4],
        "confidence_threshold": 0.5
      },
      "output": {
        "detection_results": "yolo_detection_results",
        "annotated_image_path": "yolo_annotated_image_path"
      },
      "reason": "First identify safety nets and scaffold edges, using YOLO to return bounding boxes.",
      "need_clarification":["image_path"]
    },
    {
      "tool": "check_coverage_compliance",
      "params": {
        "safety_net_bboxes": "yolo_detection_results[3]",
        "scaffold_edge_bboxes": "yolo_detection_results[4]",
        "image_path": "yolo_annotated_image_path",
        "min_coverage_ratio": 0.95
      },
      "output": {
        "is_compliant": "coverage_check_result",
        "message": "coverage_check_message",
        "output_image_path": "coverage_check_output_image"
      },
      "reason": "Using the annotated image from step 1, check the coverage of safety nets.",
      "need_clarification":[]
    }
  ]
}
```

If you believe a step has no corresponding tool, output the following format within the JSON:

```
{
  "tool": "No corresponding tool",
  "params": {
    "Param 1": "value/null"
  },
  "output": {
    "Output variable": "Output variable name"
  },
  "reason": "Logical justification",
  "need_clarification": ["..."]
}
```

E.5 Knowledge Extraction from Task History

This prompt extracts reusable technical knowledge from the construction safety inspection task execution history.

You are an expert assistant specialized in extracting reusable knowledge from the execution processes of construction safety inspection tasks. Your task is to identify and extract practical information—such as specific technical experience, parameter configurations, and judgment criteria—from complete task execution histories to enrich the knowledge base.

The extracted knowledge must meet the following criteria:

(1) **Specificity**: Must contain explicit numerical values, parameters, thresholds, or specific operational steps.
(2) **Operability**: Can directly guide the execution of subsequent similar tasks.
(3) **Domain Relevance**: Directly related to the field of construction safety inspection.
(4) **Technicality**: Must include technical details such as detection methods, tool configurations, and algorithm parameters.
(5) **Empirical Value**: Reflects summaries of experience and best practices from actual execution.

Key Focus Areas:

- Detection parameter configurations (e.g., confidence thresholds, model parameters).
- Judgment criteria and rules (e.g., quantity limits, spatial relationships, size requirements).
- Object feature descriptions (e.g., color, shape, size, material).
- Experience in detection methods and tool selection.
- Error handling and response strategies for exceptional cases.
- Experience in workflow optimization.

$user$

Please extract reusable technical knowledge and practical experience from the following complete execution history of a construction safety inspection task.

The task execution history includes:

- The user's input detection requirements
- The system's generated reasoning process (Chain of Thought)
- The automated workflow
- The actual execution results and feedback

Complete task history:
%task_history%

Please extract the technical knowledge and practical experience from the above history and return them in a JSON array format. Each knowledge item should be highly summarized and concise. Each item must be an independent, specific technical statement or empirical summary, formatted as follows:

```
{
  "knowledge_items":[
    {
      "content": "Specific knowledge content or empirical description",
      "category": "Parameter Configuration | Judgment Criteria | Detection Method
                   | Object Feature | Workflow | Error Handling",
      "applicability": "Brief description of the applicable scenario",
      "specificity_level": "High | Medium | Low"
    }
  ]
}
```

E.6 Initial Figures

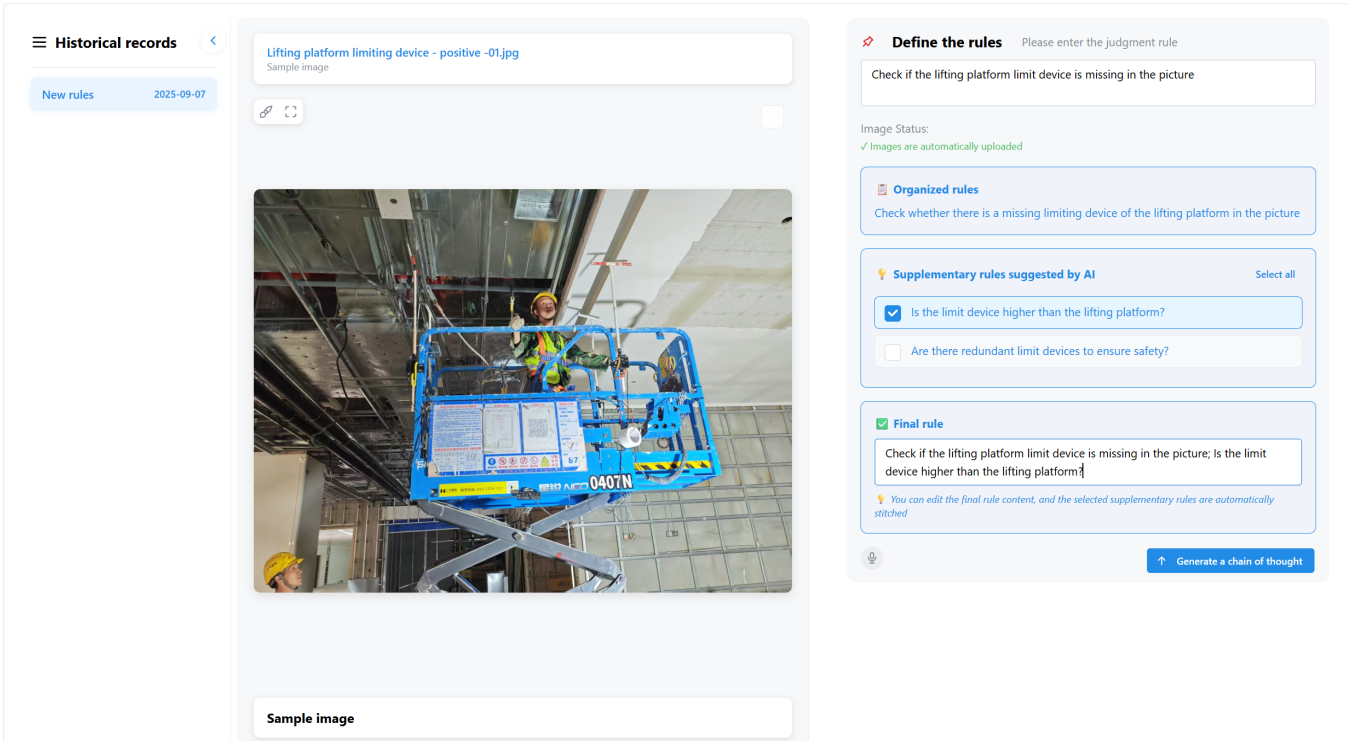


Figure 7: Raw image of Figure 3A

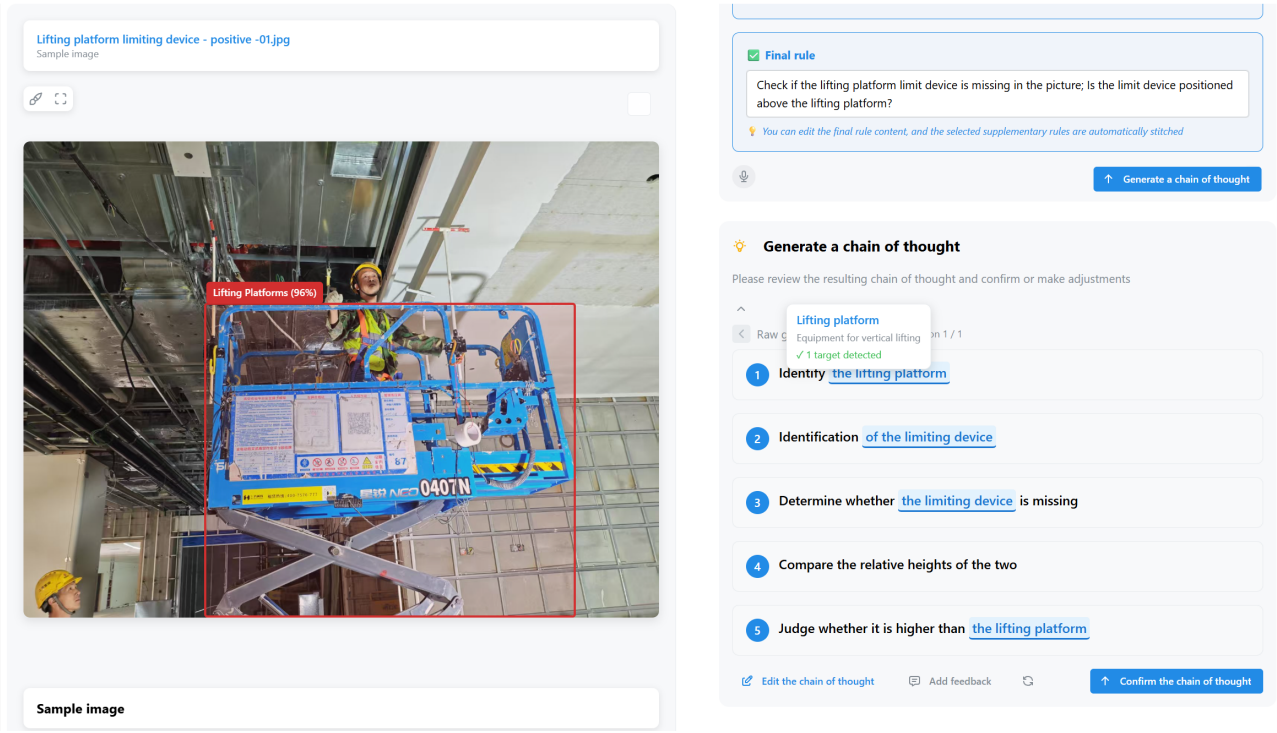


Figure 8: Raw image of Figure 3B

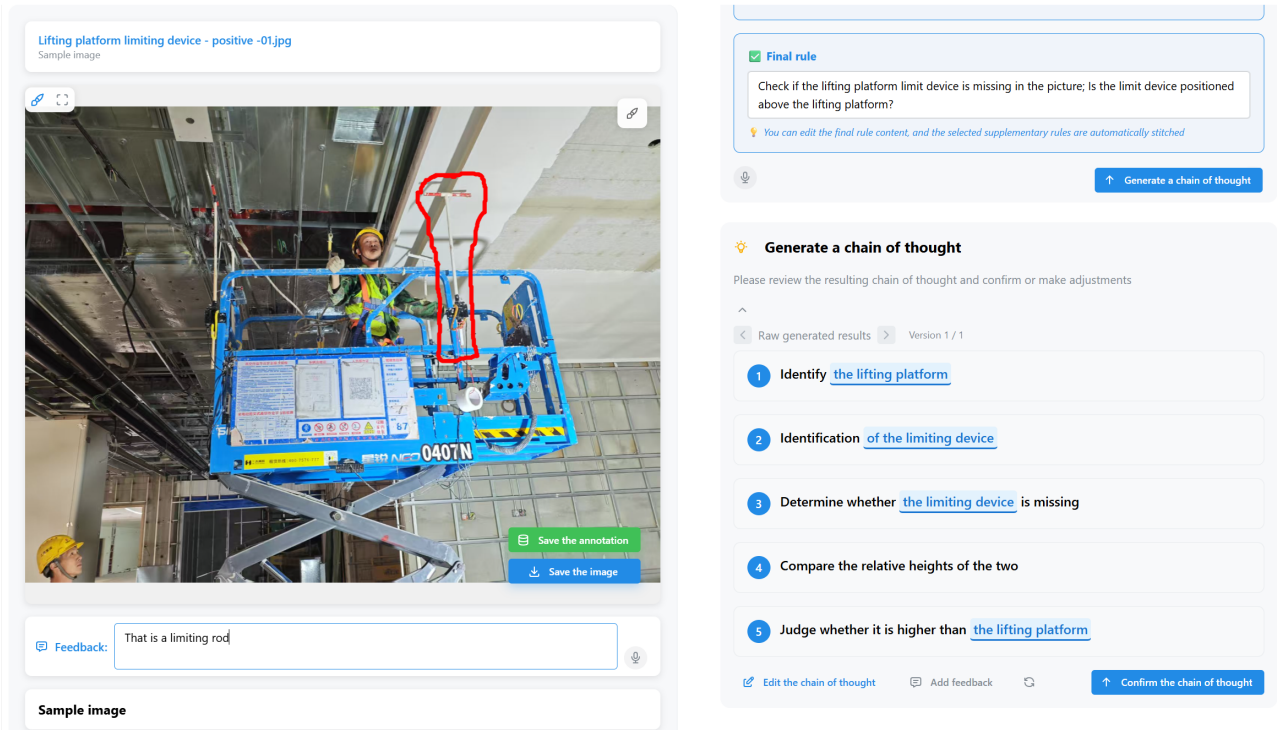


Figure 9: Raw image of Figure 3C

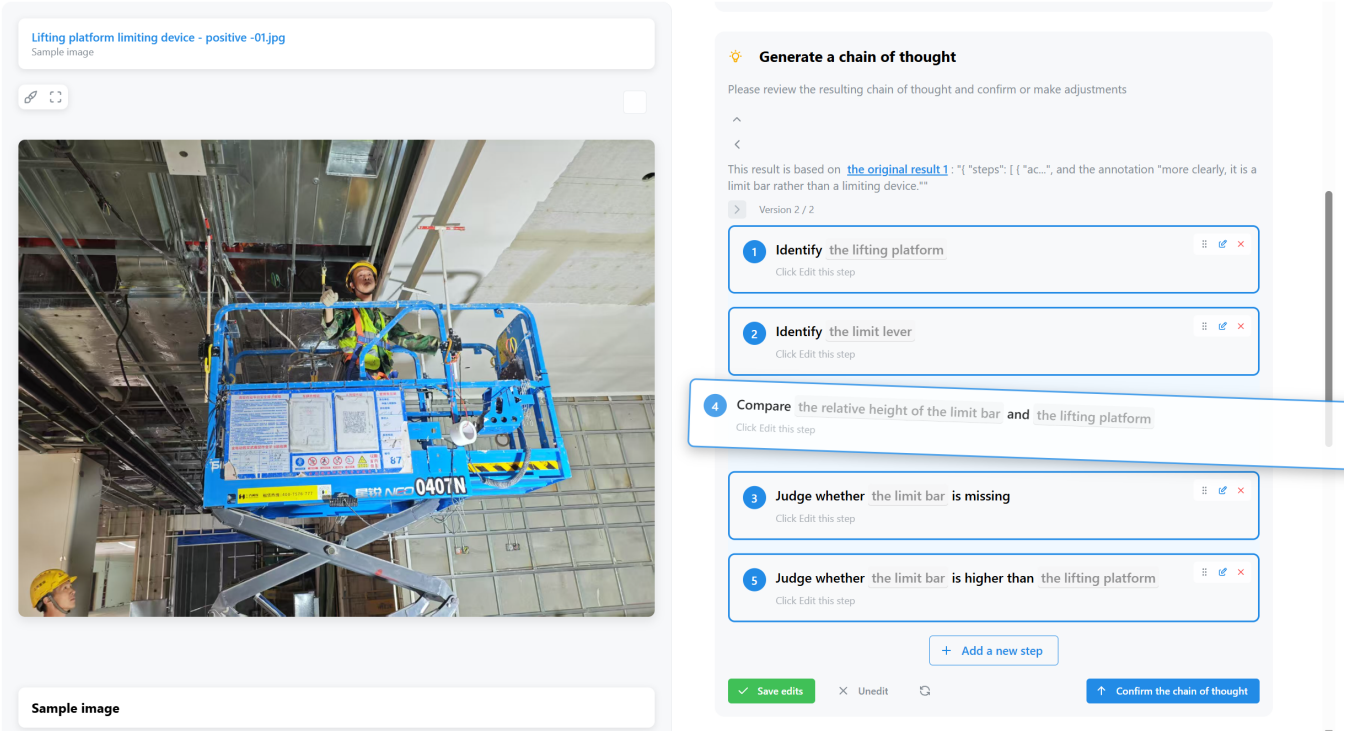


Figure 10: Raw image of Figure 3D

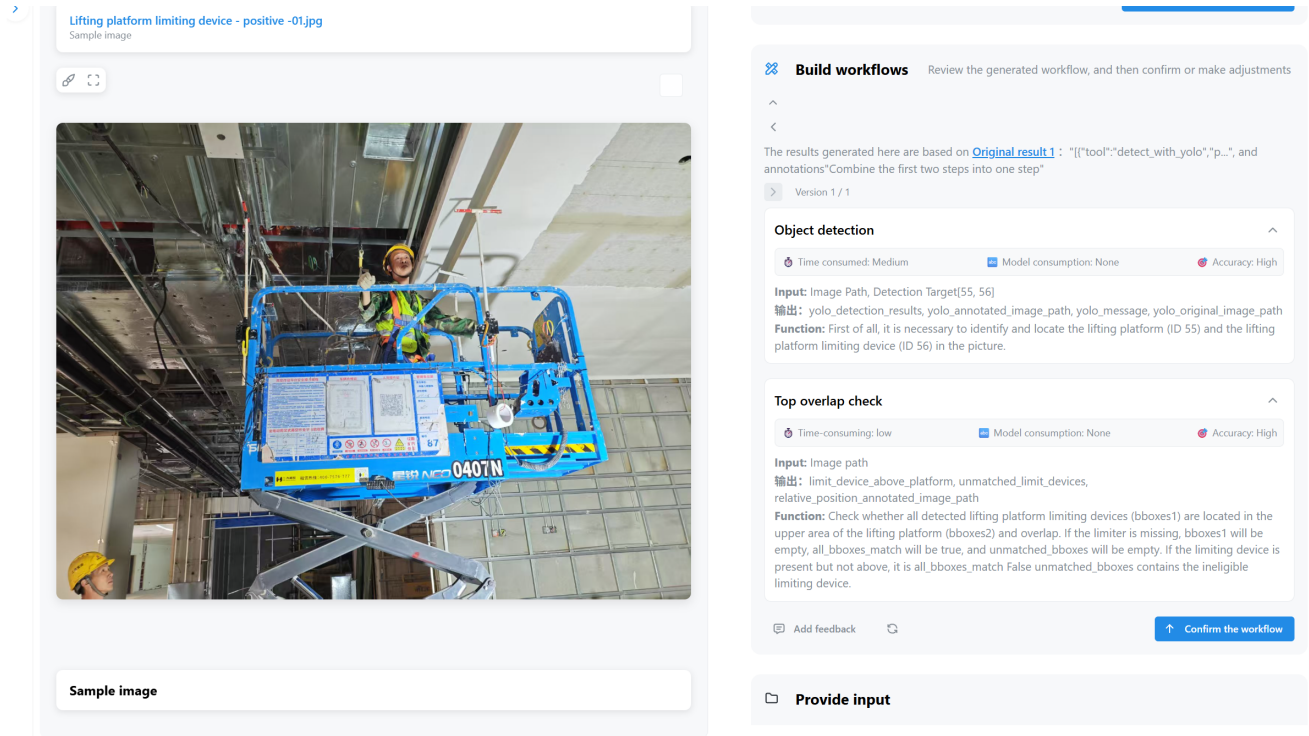


Figure 11: Raw image of Figure 3E

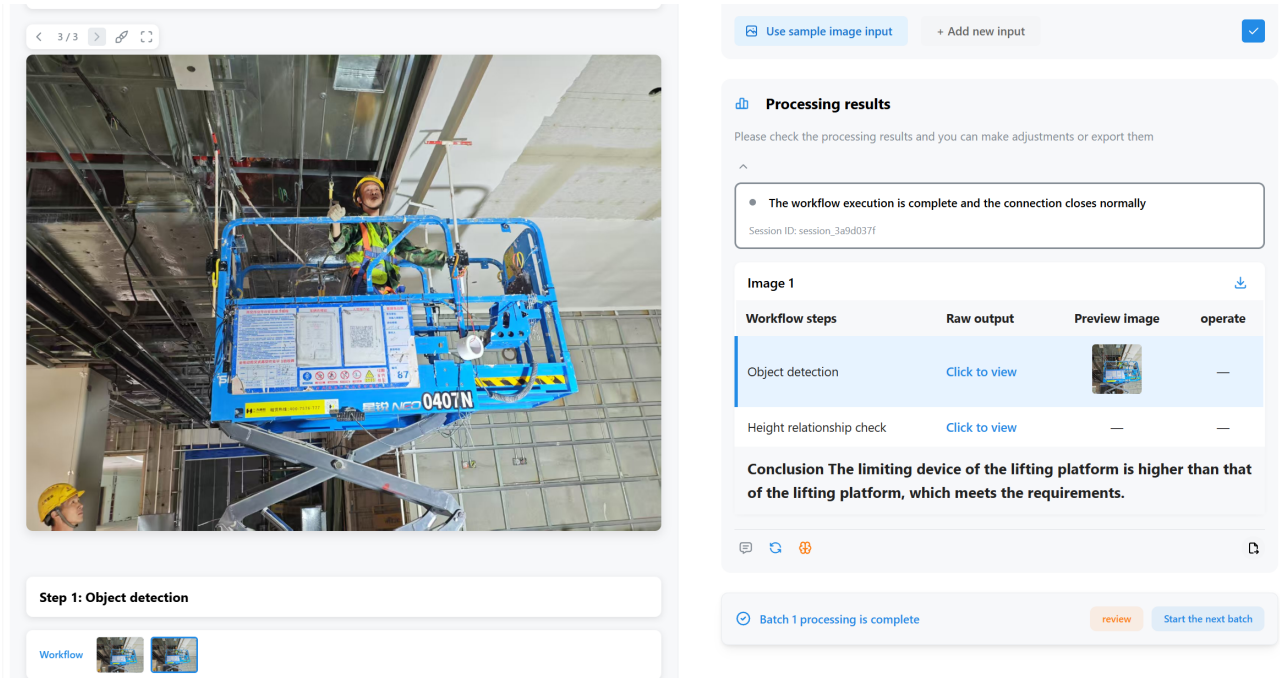


Figure 12: Raw image of Figure 3F

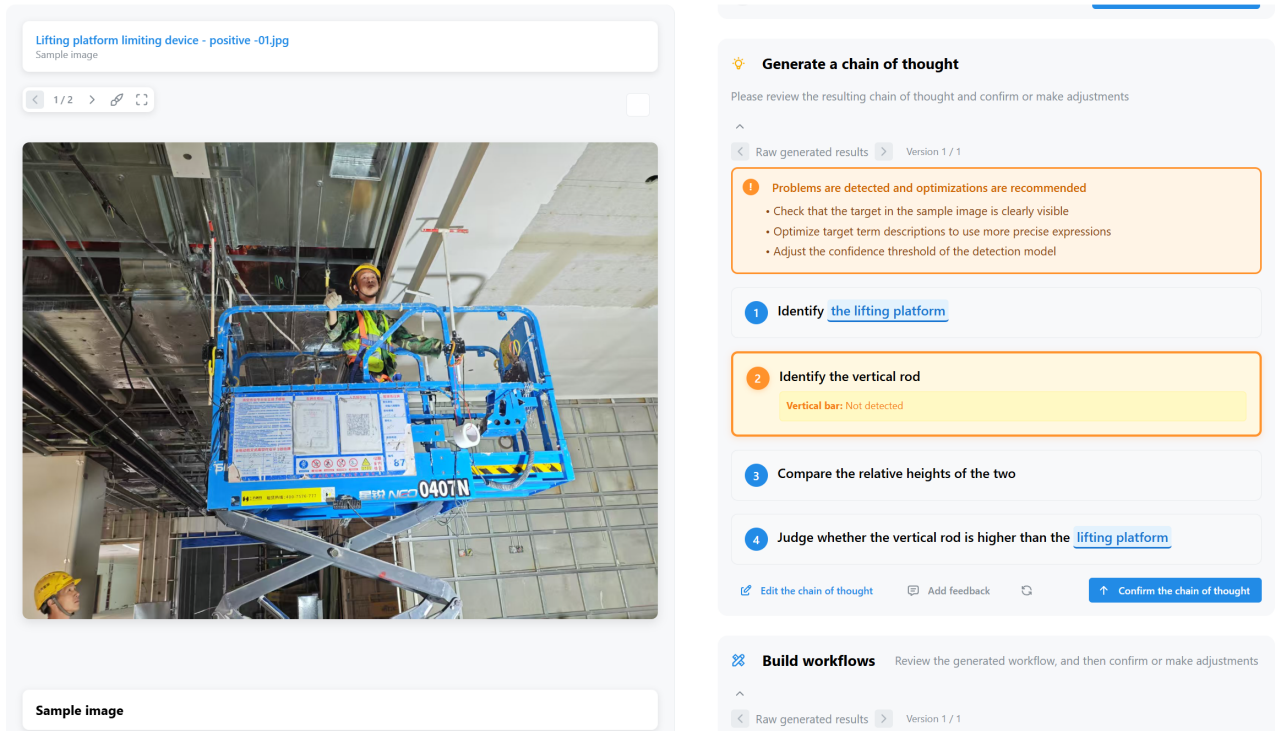


Figure 13: Raw image of Figure 3G