www.itcon.org - Journal of Information Technology in Construction - ISSN 1874-4753

FROM VISUAL DETECTION TO TRACEABLE PROGRESS EVIDENCE: A SCHEDULE-DRIVEN ONTOLOGY FRAMEWORK FOR CONSTRUCTION MONITORING

SUBMITTED: May 2026
PUBLISHED: October 2026
EDITOR: Farzad Rahimian
DOI: 10.36680/j.itcon.2026.050

Federica Madaschi, PhD Student
Department of Architecture, Built Environment and Construction Engineering, Politecnico Milano, Milan, ITA
https://orcid.org/0009-0003-6349-0035
federica.madaschi@polimi.it

Jacopo Cassandro, PhD
Department of Architecture, Built Environment and Construction Engineering, Politecnico Milano, Milan, ITA
https://orcid.org/0000-0002-1487-8178
jacopo.cassandro@polimi.it

Marco Lorenzo Trani, Professor
Department of Architecture, Built Environment and Construction Engineering, Politecnico Milano, Milan, ITA
https://orcid.org/0000-0002-5305-7305
marco.trani@polimi.it

SUMMARY*: Construction progress monitoring is increasingly combining Computer Vision (CV) and semantic technologies, yet schedules, site photographs, and monitoring outputs often remain disconnected. This limits the interpretability, traceability and reuse of progress-verification results in construction reporting. This research proposes a schedule-driven pipeline that links project planning data and construction-site visual evidence through open-vocabulary object detection, RDF-based Knowledge Graphs, SPARQL queries and controlled LLM-based verbalisation. The project schedule is used to derive and optimise the labels queried by GroundingDINO, while the resulting detections, photographic metadata and schedule information are integrated into a Knowledge Graph to assign progress-related semantic statuses. The LLaMA 3.1 8B model then reformulates the structured SPARQL output into concise technical statements without altering the rule-based classification. The pipeline is tested on the installation of a 100-metre free-standing tower crane at a hotel construction site in Milan. The results show that schedule-driven CV, semantic modelling, and controlled natural-language generation can support transparent, traceable, and semi-automatic documentation of construction progress. The workflow also contributes to SDG 9, SDG 11, and SDG 12 by supporting digital innovation, sustainable construction management, and traceable, resource-aware decision-making.*

KEYWORDS*: construction progress monitoring, open-vocabulary object detection, linked data, ontology, knowledge graphs.*

REFERENCE*: Madaschi, F., Cassandro, J., & Trani, M. L. (2026). From visual detection to traceable progress evidence: A schedule-driven ontology framework for construction monitoring. Journal of Information Technology in Construction (ITcon), Special issue: 'Digital Twins for Smart and Sustainable Built Environments: Perspectives from SASBE 2025', 31, 1197-1218. https://doi.org/10.36680/j.itcon.2026.050*

COPYRIGHT*: © 2026 The author(s). This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

1. INTRODUCTION

The Architecture, Engineering and Construction (AEC) sector is increasingly relying on digital tools to manage project information across the building lifecycle. Information models, scheduling platforms, cost databases, construction documents and site monitoring systems all contribute to the production of project knowledge. However, these information sources are often managed as separate artefacts, with limited semantic connections between design data, planned activities and construction-site evidence. As a result, information remains fragmented across heterogeneous formats and tools, limiting the ability to query, interpret, and reuse project data in an integrated manner (Han et al., 2015). This fragmentation becomes particularly critical during the construction phase, when project information must be continuously compared with the site's actual state. Progress monitoring methods have traditionally aimed to compare the planned condition of the project with the as-built state observed during construction (Golparvar-Fard et al., 2009). Among these methods, 4D BIM has become one of the most established approaches, as it links geometric information with the construction schedule to support temporal simulation and progress control (Charef et al., 2018). Other technologies, including IoT-based systems, sensing devices and manual surveying procedures, have also been adopted to collect site data, although their practical implementation may be affected by cost, scalability and operational constraints (Oesterreich and Teuteberg, 2016; Rebolj et al., 2017; Teizer et al., 2007). More recently, Computer Vision (CV) has emerged as a promising approach for analysing construction-site imagery and supporting progress monitoring. By detecting objects, resources and site conditions from photographs or videos, CV can reduce reliance on manual inspections and provide more continuous and objective evidence of construction activities (Moragane et al., 2024). At the same time, Knowledge Graphs (KGs) and Semantic Web technologies have gained increasing attention in the AEC domain because they enable heterogeneous data to be represented as interconnected entities and queried through formal semantic languages (Pauwels and Terkaj, 2016; Rasmussen et al., 2019). Therefore, the combination of CV and KGs offers the potential to connect visual evidence from the site with structured project information, such as schedules, resources and task relationships.

Despite these advances, two main limitations remain. First, many progress-monitoring approaches focus primarily on detecting whether activities are temporally consistent with the planned schedule, while offering limited support for interpreting the semantic meaning of the evidence detected. Structured outputs such as detected labels, confidence scores, activity matches, or temporal consistency indicators still require expert interpretation before they can be transformed into actionable construction-site knowledge. Second, even when semantic technologies are used to integrate project and site data, the results are often expressed as tables, graph query outputs, or classification labels, which are not immediately suitable for formal reporting or technical documentation (Rasmussen et al., 2019).

To address this gap, this paper proposes a schedule-driven pipeline that integrates CV, KGs, SPARQL querying and controlled natural-language generation to support the interpretation and documentation of construction progress. The project schedule is used as the semantic reference for visual analysis: activities and resources associated with planned activities are used to define and optimise the labels queried by *GroundingDINO* (Liu et al., 2023; Ren et al., 2024), constraining object detection to the expected construction process rather than to a generic set of visual classes. The resulting detections, together with photographic metadata and schedule information, are represented in RDF (Resource Description Framework) and queried through SPARQL to align observed site evidence with planned activities, consistently with RDF-based approaches for graph-based construction progress verification (Madaschi, Cassandro, et al., 2026).

The SPARQL output is used as structured input for a controlled Large Language Model (LLM)-based verbalisation step. LLaMA 3.1 8B (Grattafiori *et al.*, 2024) is the model used to convert the verification result into a concise technical statement, using only structured evidence such as the detected object, photograph timestamp, related scheduled activity, confidence score, and inferred progress condition. The model does not query the graph or assign the semantic state; it only reformulates the result into a human-readable explanation indicating whether the evidence supports an activity in progress, documents completion, suggests a possible delay or requires expert review.

The proposed pipeline is tested and validated on a real construction case study concerning the installation of a 100-metre free-standing tower crane at a hotel construction site in Milan. The remainder of the paper is organised as follows: Section 2 reviews the relevant literature; Section 3 describes the methodology; Section 4 presents the case study application; Section 5 reports and discusses the results; and Section 6 summarises the conclusions,

limitations and future developments. In addition, this paper extends the previous conference work on the semantic integration of construction schedules and site photographs (Madaschi, Cassandro, et al., 2026). Compared with the conference version, the present study introduces a schedule-driven label-generation procedure, an expanded SPARQL-based semantic classification, and a controlled LLM-assisted verbalisation layer for progress reporting.

2. LITERATURE REVIEW

2.1 Semantic integration and knowledge graphs in AEC

The AEC sector relies on heterogeneous information sources, including design models, construction schedules, cost data, technical documents and site records. Although these data are all relevant to project management, they are often produced and stored in separate systems. This fragmentation limits interoperability and makes it difficult to retrieve and relate information across different project domains.

In response to this issue, KGs and Semantic Web technologies have been increasingly investigated to structure, connect, and query construction information more consistently. KGs represent information as entities and relationships, providing a formal mechanism to link heterogeneous data while preserving their semantic meaning. In the AEC domain, ontology-based modelling has been widely explored to improve interoperability and enable the machine-readable representation of construction knowledge. Foundational contributions, such as ifcOWL, demonstrated how IFC-based building information can be translated into Semantic Web representations, supporting data exchange and reuse beyond proprietary BIM environments (Beetz et al., 2009; Pauwels and Terkaj, 2016). Building on this perspective, AEC KGs have been proposed to manage interrelated project information and support complex queries across models, documents, properties, and external datasets (Rasmussen et al., 2019). From a technical perspective, ontology-based KGs combine a graph data model with explicit conceptual structures. The Resource Description Framework (RDF) represents information as triples, while RDF Schema (RDFS) and the Web Ontology Language (OWL) define classes, properties, hierarchies, and semantic constraints. This allows project entities to be identified through stable identifiers and connected through formally defined relationships.

In AEC, this logic has led to the development of modular ontologies addressing different aspects of the built environment. Examples include ifcOWL for IFC-based building information (Pauwels and Terkaj, 2016), the Building Topology Ontology (BOT) for spatial and topological relationships (Rasmussen et al., 2021), and the Digital Construction Ontologies (DiCon) for construction workflows, resources, agents, and contextual information (Zheng et al., 2021). These models allow building elements, activities, resources, and documents to be queried across domains using SPARQL. However, the effectiveness of ontology-based integration depends not only on the availability of semantic models, but also on their alignment with the information needs of the specific application. Zheng et al. (2021), for instance, emphasise the role of competency questions and SPARQL queries in evaluating whether an ontology can represent the information required to support construction workflow applications. This aspect is particularly relevant when different information sources describe the same project at different levels of granularity. If construction tasks, resources, and site observations are not modelled at compatible levels of detail, query results may be formally correct but remain difficult to interpret from an operational perspective. More broadly, the comprehensive review by Farghaly et al. (2023) traced the evolution of ontology research in the AEC sector over two decades, identifying major application domains including construction processes, cost management, operation and maintenance, health and safety, sustainability, monitoring and control, smart cities, heritage building information modelling, and regulatory compliance. Their analysis highlights both the increasing specialisation of AEC ontologies and the persistent need for improved interoperability, ontology reuse, and integration across heterogeneous project information domains.

For construction progress monitoring, this issue becomes critical because planned activities, resources, time intervals, and site observations must be interpreted together. Construction-oriented semantic schemas, such as the Digital Twin Construction (DTC) ontology, provide useful structures for representing tasks, resources, and process-related information in digital twin environments (Borrmann et al., 2022). Ontology-aided querying approaches further demonstrate how semantic models can support the retrieval of project information via formal or natural-language queries (Yin et al., 2023). Recent studies have also highlighted the potential of knowledge-based technologies to support both human and computational interpretation of construction information across the building lifecycle (Shen et al., 2024).

Despite these advances, semantic integration alone does not necessarily yield outputs directly usable for construction reporting or decision-making. KG-based workflows often return graph query results, tables, or semantic triples that still require expert interpretation before they can be translated into actionable site knowledge. This limitation is especially relevant for progress verification, where the objective is not only to retrieve information, but also to explain the relationship between planned activities and observed site evidence.

2.2 Computer vision and open-vocabulary detection for construction monitoring

CV has become a widely studied approach for monitoring construction progress. By analysing site images and videos, CV methods can detect construction elements, equipment, resources, and working conditions, providing visual evidence to support comparisons between planned and actual construction states (Golparvar-Fard *et al.*, 2009). Compared with manual inspections, CV can support more frequent and objective data acquisition, although its performance is affected by occlusions, changing site conditions, visual clutter and the variability of construction objects (Moragane *et al.*, 2022). Deep learning-based object detection has improved the ability to recognise visual patterns in complex site environments and has been applied to progress tracking, activity recognition, quality control and safety monitoring. Nevertheless, conventional object detection systems usually rely on predefined object classes and annotated training datasets. This limits their scalability across construction projects, where the relevant objects vary according to project phase, site organisation and planned activities. Open-vocabulary object detection models address part of this limitation by enabling detection through natural-language prompts rather than fixed class sets. Models such as GLIP, OWL-ViT and *GroundingDINO* have shown that visual regions can be localised through text-conditioned inference, supporting more flexible object detection without retraining on project-specific classes (Li et al., 2022; Liu et al., 2023; Minderer et al., 2022). *GroundingDINO* is particularly relevant in this context because it combines open-set detection with grounded language-image pre-training, allowing arbitrary textual queries to be used for object localisation (Liu et al., 2023; Ren et al., 2024).

From an architectural perspective, open-vocabulary detectors extend conventional object detection by replacing fixed class prediction with text-conditioned localisation. Instead of assigning each detected region to a predefined category, these models compare visual features extracted from the image and textual features derived from natural-language prompts. GLIP introduced this principle by unifying object detection and phrase grounding within a grounded language-image pre-training framework, enabling the model to associate image regions with textual expressions and transfer to unseen object categories (Li et al., 2022). OWL-ViT follows a different strategy by adapting contrastively pre-trained vision-language models to detection, using a Vision Transformer architecture and text queries to localise objects beyond a closed vocabulary (Minderer et al., 2022). *GroundingDINO* builds on this line of research by combining the Transformer-based DINO detector with grounded language-image pre-training. Its architecture includes separate image and text backbones, a feature enhancer for early vision-language fusion, a language-guided query selection module that selects object queries relevant to the input text, and a cross-modality decoder that refines bounding boxes and aligns them with textual tokens (Liu et al., 2023). During inference, the user provides one or more textual prompts; the model extracts visual and linguistic embeddings, estimates correspondences between text tokens and image regions, and returns bounding boxes with confidence scores for the most relevant regions. This mechanism makes *GroundingDINO* suitable for construction monitoring because project-specific objects can be searched for using textual labels derived from schedules, resources, or technical documents, without requiring retraining on a fixed set of construction classes.

However, prompt-based detection introduces a methodological issue: textual labels are not neutral inputs. Their formulation can affect text–region alignment, detection consistency and false-positive behaviour. Recent work on prompt robustness in construction scenarios has shown that *GroundingDINO* is sensitive to lexical formulation, with concise base-term prompts producing more stable results than highly technical, plural or descriptive expressions (Madaschi, Taherian, *et al.*, 2026). This evidence suggests that the labels used in CV-based progress monitoring should be derived and optimised based on both the construction process and the model's grounding behaviour. Schedule-driven label optimisation, therefore, represents a relevant step for linking visual recognition to planned construction resources while improving the reliability and interpretability of object detection outputs.

2.3 From CV–KG integration to interpretable progress documentation

The integration of CV and KGs has been proposed to overcome the limitations of approaches based solely on visual detection or semantic modelling. CV systems can extract evidence from site images, but their outputs

typically consist of detected labels, bounding boxes and confidence values. KGs, in turn, can represent project information and support semantic querying across heterogeneous construction data (Rasmussen *et al.*, 2019). Combining the two approaches enables linking detected objects to planned activities, expected resources, and temporal relationships, thereby supporting progress verification and planned-versus-actual comparison (Han *et al.*, 2015; Madaschi, Cassandro, *et al.*, 2026). This integration supports a richer interpretation of visual evidence. The same detected object may have different meanings depending on the photograph timestamp and the corresponding planned activity. For example, an operational machine observed after the planned completion of a related task may indicate a possible delay, whereas a permanently constructed element may document task completion. RDF-based modelling and SPARQL querying can therefore support progress verification by retrieving, for each image, the detected objects, confidence values, capture timestamp and related scheduled activities. This graph-based representation also improves traceability because each interpretation can be linked back to the original photograph, the detected object, the confidence value, and the planned activity.

Despite this potential, many CV–KG and vision-based progress-monitoring workflows still produce structured computational outputs, such as tables, graph-derived metrics, consistency indicators, visual labels, or classification results, which may require further interpretation before being incorporated into construction reporting and decision-making (Rehman et al., 2022; Pfitzner et al., 2024). These outputs are suitable for computational processing but are not always directly usable in construction reporting, progress documentation or technical decision-making. The interpretation of the result often remains delegated to experts, who must translate structured outputs into explanatory statements. This limitation indicates the need for a workflow that preserves the computational rigour of semantic verification while improving the interpretability of its outputs.

From this perspective, controlled LLM-based verbalisation can serve as a reporting layer positioned after semantic reasoning. The generative model should not replace graph-based querying or rule-based state assignment; rather, it should reformulate structured evidence into concise technical statements. This separation is important because LLMs can produce unsupported content when generation is not constrained by structured inputs (Ji et al., 2023), while recent research on structured outputs further confirms the relevance of controlled generation in technical workflows (Geng et al., 2025). Accordingly, the following methodology defines a schedule-driven pipeline that integrates visual detection, semantic alignment, and interpretable progress documentation into a single traceable workflow.

3. METHODOLOGY

The proposed methodology develops a schedule-driven computational pipeline to link project planning data to construction-site visual evidence. The objective is to support the semi-automated verification and documentation of construction progress by integrating planned activities, site photographs, CV outputs, semantic alignment, and report-oriented verbalisation into a traceable workflow. The methodology assumes that the input schedule provides a sufficiently detailed, complete, and up-to-date representation of the current construction plan. Activity dates, hierarchical relationships, and resource descriptions must be maintained consistently throughout project execution, since they support both the generation of object-detection prompts and the subsequent semantic interpretation of visual evidence.

The pipeline uses two main input sources: the project schedule, exported from Microsoft Project in XML format, and construction-site photographs, provided in standard image formats. These inputs are processed through two sequential modules. Module 1 derives a visual analysis plan from the schedule and applies object detection to the photographic dataset. Module 2 integrates the resulting evidence with schedule information into a KG, performs SPARQL-based semantic alignment and generates controlled technical statements from the structured query results. The overall workflow is shown in Figure 1.

3.1 Module 1: Computer vision and image alignment with the project schedule

Module 1 constructs the visual analysis plan and applies object detection to the construction-site photographic dataset through an automated procedure.

Once the Microsoft Project XML schedule and the site photographs are provided as inputs, the pipeline automatically extracts the information required for the analysis. From the schedule, it retrieves planned activities, start and end dates, task identifiers, hierarchical relations and assigned resources. From the images, it extracts the

available EXIF metadata, including capture date and, when available, capture time. The extracted schedule information is then automatically used to define the visual search context for each photograph. For each image, the capture timestamp is compared with the planned activity intervals to identify the activities active on the same date. When no activity is active, the pipeline reconstructs the nearest temporal context by identifying the latest completed task before the image date and the earliest following task after it. This prevents photographs captured between two planned activities from being forced into an incorrect task assignment, while still preserving their relationship with the schedule.

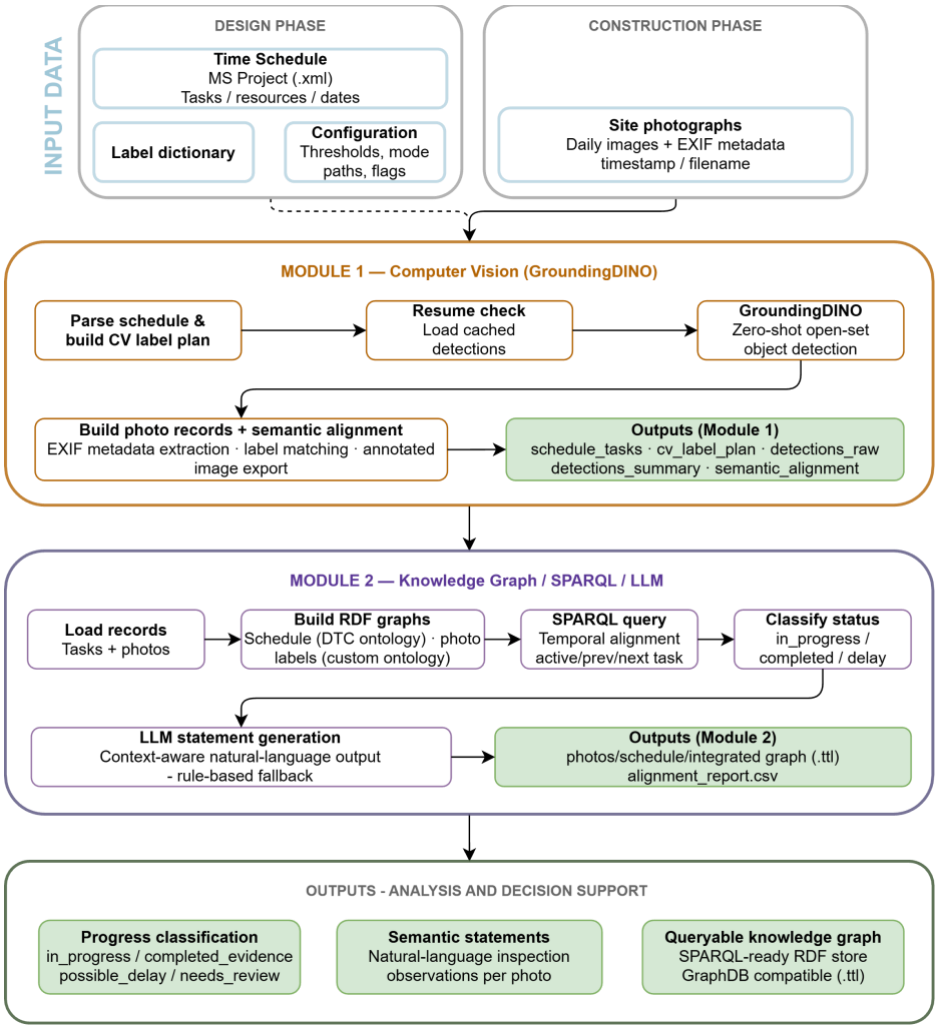

Figure 1: Overall workflow of the schedule-driven CV and semantic interpretation pipeline. For readability, only the principal output statuses are shown.

A key automated step in Module 1 is transforming schedule activities and resources into visual prompts for *GroundingDINO*. Since the Gantt chart contains instrumental and human resources, the pipeline filters the extracted resources before object detection. Human resources are excluded from the visual analysis for worker privacy reasons, while organisational and other non-visual resources are excluded or flagged because they cannot be reliably represented as detectable objects in site photographs. Conversely, visually observable instrumental resources are retained and reformulated into concise prompts suitable for open-vocabulary detection. This automated prompt optimisation avoids using schedule entries as raw labels and reduces ambiguity during text-conditioned object localisation. *GroundingDINO* is then applied to the selected photographs using the schedule-derived prompts. For each image–prompt pair, the model returns detected labels, bounding boxes and confidence scores. Detections are subsequently filtered using confidence thresholds and Intersection over Union (IoU)-based post-processing to remove low-confidence predictions and redundant bounding boxes. Both successful and

missing detections are retained: when an expected prompt produces no valid detection above the configured threshold, the case is recorded as a no-detection result rather than discarded. As schematically represented in Figure 2, Module 1 produces an auditable image–activity–resource–label plan and the corresponding object-detection results without requiring manual selection of images, activities or labels during execution.

This intermediate output links each photograph to its timestamp, relevant schedule context, selected resource, visual prompt and detection outcome. At this stage, the module does not assign progress states. It only produces structured visual evidence that can be transferred to Module 2, where detections, missing detections, timestamps and schedule information are integrated into the KG and interpreted through SPARQL-based semantic alignment.

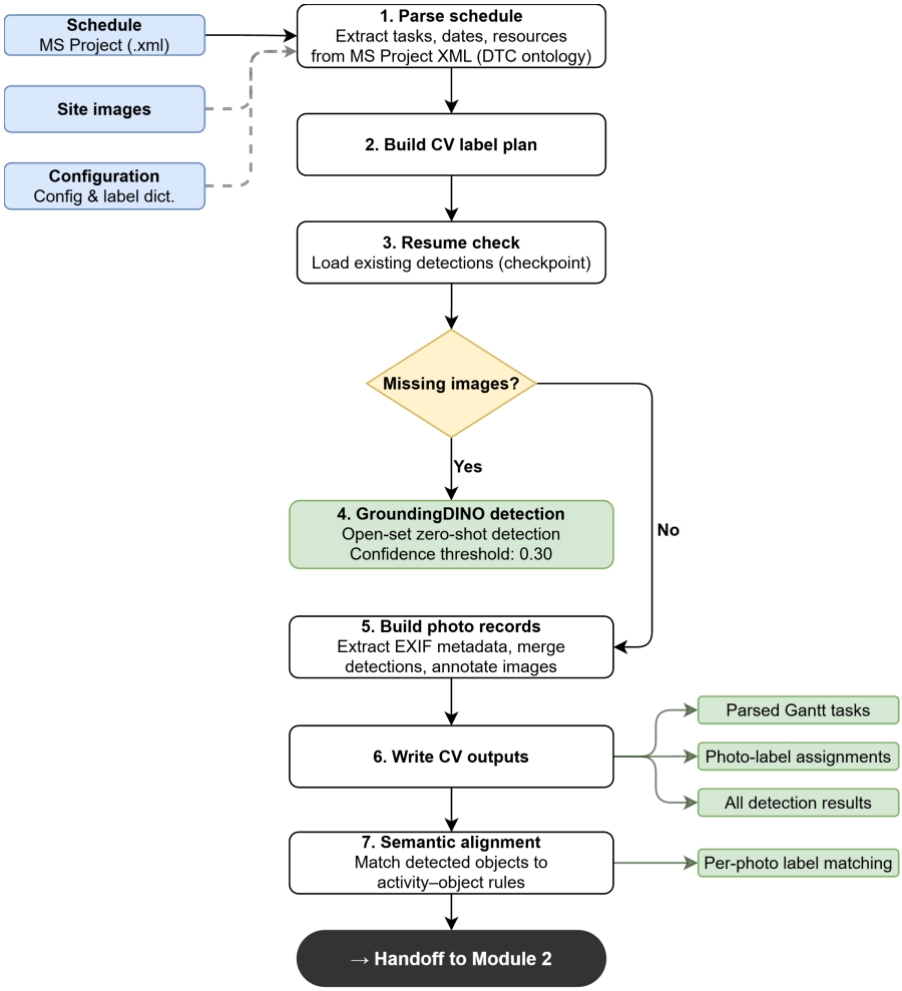

Figure 2: Module 1 Workflow for schedule-based label extraction and GroundingDINO object detection.

3.2 Module 2: Knowledge graph, SPARQL queries & LLM-assisted interpretative generation

Module 2 transforms the structured outputs from Module 1 into a semantic, interpretable representation of construction progress. The module receives the parsed schedule, image metadata, visual analysis plan, detection results, confidence scores and no-detection records produced by the previous stage. These data are automatically converted into RDF and integrated into a multi-domain KG, rather than being treated as independent tabular outputs (Madaschi, Cassandro, *et al.*, 2026). The KG connects three complementary information layers. The schedule layer represents planned activities, time intervals and associated resources through the DTC ontology (Borrmann et al., 2022, Schlenger et al., 2024). The photographic layer represents site images, capture metadata, detected objects, prompts and confidence scores through a project-specific `photo:` vocabulary. The alignment

layer stores the relationships generated by the pipeline between photographs, expected or detected resources, schedule context and semantic status.

This structure preserves the provenance of each result, since every interpretation can be traced back to the corresponding photograph, timestamp, detection output and planned activity. SPARQL queries are then executed to align visual evidence with the project schedule.

For each photograph, the query retrieves the capture timestamp, detected or expected resources, confidence scores, active task when present, nearest preceding and following activities, and the selected schedule context. The alignment does not force each photograph into a single task. Instead, it compares the visual evidence with the temporal and semantic context reconstructed from the schedule, allowing photographs captured during active tasks, between activities or after task completion to be interpreted consistently.

Based on this integrated representation, the pipeline assigns semantic statuses through explicit rule-based logic. The classification depends on the relationship between the photograph timestamp, the detected or missing object, the expected resource and the schedule context. A photograph can therefore be classified as `in_progress`, `completed_evidence`, `possible_delay`, `no_detection`, `needs_review` or `not_classifiable`. These categories constitute the complete set of semantic statuses implemented in the pipeline. Depending on the characteristics of the analysed dataset, only a subset of them may occur in a specific application. For graphical clarity, the workflow figures report only the principal or case-relevant statuses, as shown in the previous figures, while the complete classification logic remains implemented in the code.

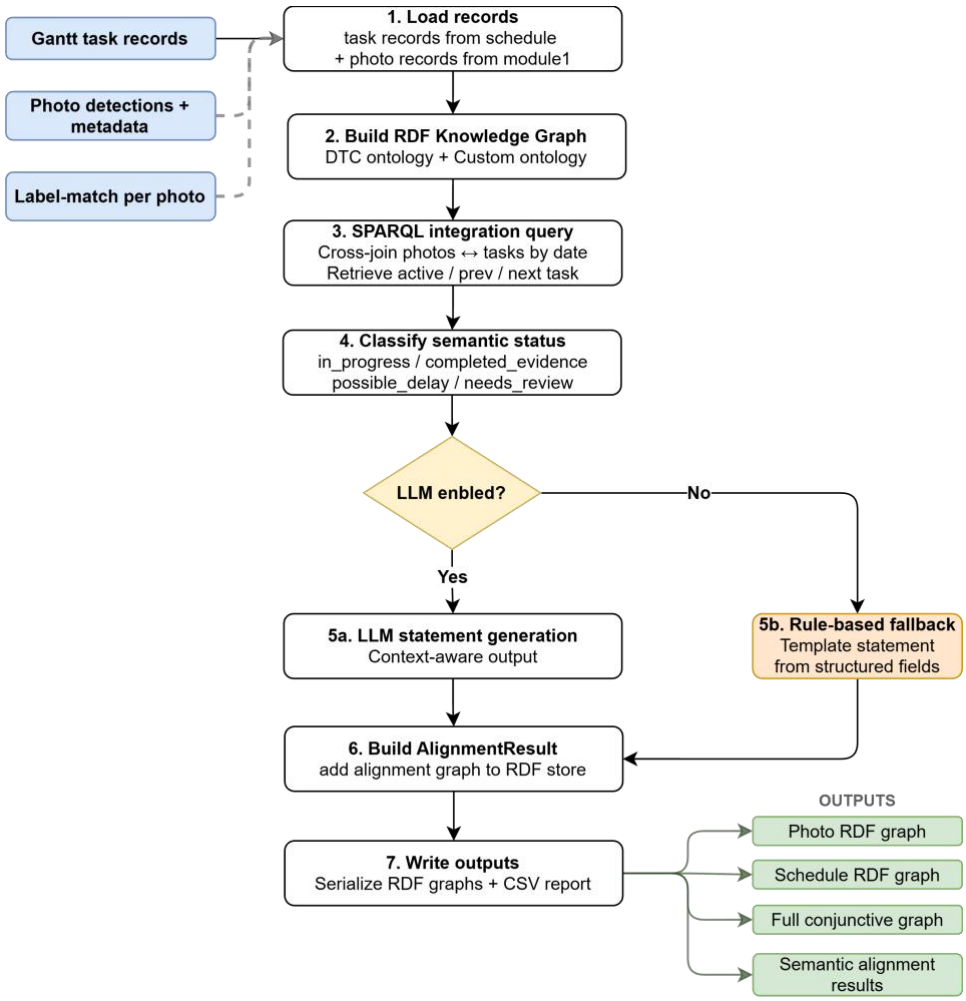

Figure 3: Module 2 workflow for KG construction, SPARQL-based alignment, and LLM-assisted output generation.

This rule-based step avoids a binary interpretation of visual evidence and makes uncertainty explicit when the available information is incomplete or ambiguous. In the current workflow, as illustrated in Figure 3, the semantic status is assigned using SPARQL-based rules rather than by the LLM. The SPARQL-derived records are subsequently used as structured input for the LLM-assisted verbalisation step. LLaMA 3.1 8B receives only graph-derived fields, including the photograph identifier, capture date, detected or expected resource, confidence score, selected task, semantic status and alignment note. Its role is limited to reformulating these structured records into concise technical statements suitable for progress reporting. The model does not query the KG, does not assign semantic statuses and does not introduce additional project information.

This separation between semantic reasoning and linguistic formulation is central to the proposed workflow. SPARQL rules perform the alignment and classification, while the LLM improves the readability of the final output under controlled input constraints. A deterministic template-based fallback is also included to generate reproducible statements when LLM generation is unavailable or disabled.

The final outputs of Module 2 include the integrated KG, the SPARQL result table, the semantic alignment table, the LLM-generated statements and the final status report.

3.3 Computational environment and local deployment

The proposed pipeline was implemented and executed on a local workstation to ensure reproducibility, data control and operational independence from cloud-based services.

The experimental setup was equipped with an Intel Core i9-14900HX processor, 64 GB of RAM, a 1.9 TB NVMe SSD and an NVIDIA GeForce RTX 4090 Laptop GPU with 16 GB of dedicated video memory. This configuration provided the computational resources required to run both the object-detection and LLM-assisted verbalisation stages locally.

The software environment was configured in Ubuntu 26.04 LTS through Windows Subsystem for Linux (WSL) and managed in Visual Studio Code using a dedicated Python 3.10.20 kernel. WSL enabled the execution of the Linux-based dependencies required by GroundingDINO within the local Windows workstation. The CV stage was executed using PyTorch 2.5.1 with CUDA 12.1 support, while the NVIDIA RTX 4090 Laptop GPU was used as the CUDA device for local inference.

The verbalisation stage was also deployed locally. LLaMA 3.1 8B was executed through Ollama version 0.24.0 using the llama3.1:8b model. In this configuration, the model was used only after SPARQL-based semantic alignment, serving as a controlled component to generate concise technical statements from structured records. This local deployment avoided dependence on external API-based services, including authentication requirements, usage quotas, rate limits and request-based costs. It also ensured that construction-site images, intermediate semantic records and generated statements remained within the same controlled computing environment.

4. SEMANTIC MODELLING AND ONTOLOGY INTEGRATION

To support the semantic integration of schedule information, construction-site photographs, object-detection results, and interpretative outputs within a unified KG, this research combines the reuse of existing Semantic Web ontologies with a project-specific vocabulary for representing visual evidence.

The ontology network was designed to establish explicit semantic relationships between planned construction activities, expected resources, photographic observations, detected objects and confidence values, thereby enabling cross-domain querying and traceable interpretation of construction progress.

At the core of the semantic structure, schedule information is represented through the DTC ontology, which provides classes and properties for modelling construction tasks, temporal intervals and associated resources. Through this ontology, planned activities are represented as semantically connected entities rather than isolated schedule records, enabling their integration with construction-site observations and subsequent SPARQL-based reasoning.

The visual domain is represented through a lightweight project-specific ontology identified by the `photo:` namespace. This vocabulary was introduced because existing construction ontologies do not provide a sufficiently detailed representation of image-based object-detection evidence and related confidence metrics. Within this

structure, each construction-site image is modelled as an instance of `photo:Photo`, identified through a stable URI and enriched with metadata such as `photo:photoId`, `photo:filename`, `photo:captureDate`, `photo:captureTime`, and `photo:captureDateTime`. This modelling explicitly links each visual observation to a precise temporal context, allowing the comparison between photographic evidence and the planned schedule.

The objects detected by *GroundingDINO* are represented as instances of `photo:Object` and are connected to the corresponding image via the `photo:hasObject` property. Semantic labels associated with the detections are expressed through `rdfs:label`, enabling both human-readable interpretation and semantic querying. Rather than representing confidence values as isolated numerical attributes, the proposed model introduces explicit `photo:Detection` and `photo:Accuracy` entities. Each detection is therefore associated with a dedicated confidence instance linked through `photo:hasAccuracy`, while the numerical score is represented as a typed literal (`photo:value^^xsd:decimal`). This modelling strategy improves traceability because each recognition result can be queried together with its originating image, detected object, associated label and confidence level.

The resulting ontology network establishes semantic connections between the planning and visual monitoring domains. Planned activities represented through the DTC ontology are linked, through SPARQL-based alignment procedures, to the visual evidence represented in the `photo: vocabulary`. This integration enables the construction of a multi-domain KG in which schedule data, photographic metadata, detected objects, semantic states and interpretative outputs can be queried together within GraphDB. The graph structure also preserves provenance relationships among the original site image, the detected resource, the associated construction activity, and the resulting semantic interpretation, thereby supporting transparency, auditability, and future extensibility toward more standardised ontology alignments.

The ontologies and namespaces employed in this research are reported in Table 1. These vocabularies support the representation of the schedule, photographic evidence, detected objects, confidence values and typed literals used throughout the semantic integration process.

Table 1: Ontologies and namespaces used in this research.

Namespace	URI	Description
`photo:`	<https://example.org/project/photos#>	Project-specific ontology for representing photographs, detected objects and confidence values.
`dtc:`	<https://dtc-ontology.cms.ed.tum.de/ontology#>	Digital Twin Construction ontology
`rdf:`	<http://www.w3.org/1999/02/22-rdf-syntax-ns#>	RDF core vocabulary
`rdfs:`	<http://www.w3.org/2000/01/rdf-schema#>	RDF Schema vocabulary
`xsd:`	<http://www.w3.org/2001/XMLSchema#>	XML Schema datatypes

5. CASE STUDY APPLICATION

The proposed pipeline was applied to a real hotel construction site in Milan, Italy, monitored for approximately one year across several construction phases, including diaphragm walls, foundation piles, tower crane installation and basement slab construction. For this research, the analysis was restricted to the installation of a 100-metre free-standing tower crane, carried out between November and December 2024. This phase was selected because it is a technically critical and schedule-sensitive operation that directly affects the execution of subsequent construction activities.

Figure 4: Overview of the case-study construction site during the monitored construction phases.

Figure 4 provides an overview of the case-study construction site and situates the photographic dataset within the broader monitoring campaign, while Figure 5 schematically represents the crane installation schedule and the associated resources used to derive the object detection labels.

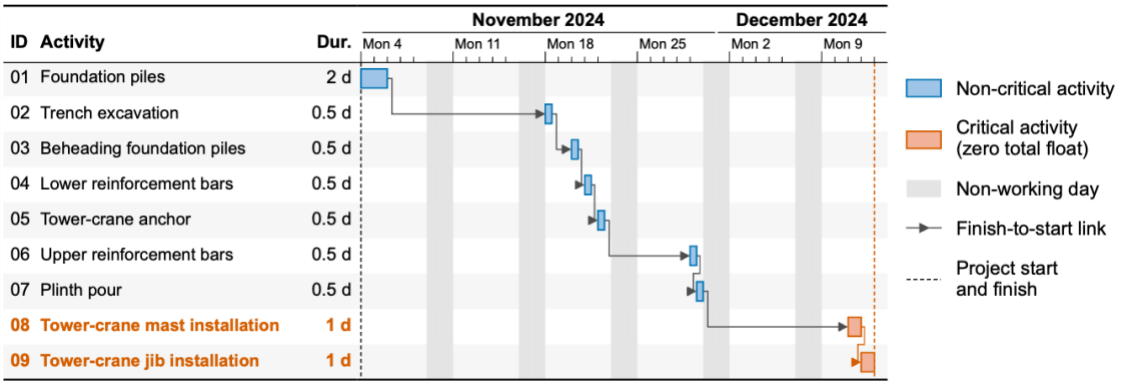

Figure 5: Tower crane installation time schedule and associated construction resources used for schedule-driven object detection.

The starting point of the application was the project schedule developed in Microsoft Project, which outlines the tower crane installation process as a sequence of planned activities, from foundation piling and excavation to reinforcement placement, plinth pouring, and crane assembly. Each activity is associated with construction resources, including drilling equipment, excavators, reinforcing bars, crane foundation components, concrete delivery equipment, mobile cranes, and tower crane components.

6. RESULTS AND DISCUSSIONS

This section presents the results obtained from applying the proposed schedule-driven pipeline to the tower crane installation case study. In line with the methodological structure, the analysis follows the progressive transformation of construction-site data from visual detection to semantic interpretation and report-oriented verbalisation.

6.1 Results of the schedule-driven computer vision process

The application of Module 1 produced a structured set of visual evidence for the tower crane installation phase. The photographic dataset included ten construction-site images. At least one valid object was detected in nine images, while one image produced no detection above the configured thresholds. Overall, the detection process generated twelve valid object detections.

Object detection was performed with *GroundingDINO* using a box confidence threshold of 0.350 and a text threshold of 0.250. The box threshold was used to retain detections with sufficient localisation confidence, while the text threshold controlled the semantic correspondence between the textual prompt and the detected visual region. Overlapping detections were filtered using an IoU threshold of 0.700 to reduce redundant bounding boxes for the same object. This configuration was selected to limit low-confidence detections before transferring the results to the semantic alignment stage.

The detected objects included drilling equipment, excavators, reinforcement bars, a tower crane anchor template, a concrete mixer truck, a mobile crane and a tower crane jib. These outputs should be interpreted as visual observations extracted from the photographs, rather than as direct progress-verification results. Their meaning depends on their relationship with the photograph timestamp and the planned schedule, which is evaluated in the following KG and SPARQL-based alignment stage.

The results show that large and visually distinctive resources, such as drilling equipment, mobile cranes and the tower crane jib, were generally localised with more stable confidence values. Conversely, elements characterised by repetitive geometry, partial visibility or dense construction-site textures, such as reinforcement bars or excavation-related evidence, produced lower confidence scores. This behaviour is consistent with previous

findings showing that *GroundingDINO* performs more reliably on objects with clear geometry and strong visual–textual correspondence, while performance decreases for small-scale, ambiguous or cluttered objects (Taherian et al., 2026). Construction-site conditions such as occlusions, scale variation, clutter and uncontrolled illumination may further affect object localisation and increase the risk of false positives or missed detections (Ding et al., 2024; Taherian et al., 2026).

Representative detection outputs are reported in Figure 6 and summarised in Table 2. The selected examples include successful detections (4, 5), lower-confidence detections (1, 3) and one no-detection case (2). These outputs constitute the visual evidence transferred to Module 2, where detections, missing detections, timestamps and schedule information are integrated into the KG and interpreted through SPARQL-based semantic alignment.

(1) (2) (3) (4) (5)

Figure 6: Representative schedule-driven object detection results obtained with GroundingDINO.

Table 2: Representative GroundingDINO outputs selected for semantic alignment.

Case	Detection label	Conf.	Visual evidence	Semantic use
(1)	Drilling rig	0,56	Equipment for foundation pile execution.	Supports `in_progress` if aligned with the active scheduled task.
(2)	Excavator	-	No object detected above threshold.	Recorded as `no_detection` or `needs_review`.
(3)	Reinforcement bars	0,43	Reinforcement system in the crane foundation	Supports reinforcement-related progress evidence.
(4)	Tower crane anchor template	0,64	Physical component of the tower crane foundation	Supports `completed_evidence` after installation.
(5)	Tower crane jib	0,67	Crane jib during assembly operations.	Supports `in_progress` during jib installation.

6.2 Knowledge graph and SPARQL-based semantic alignment results

The outputs generated by Module 1 were converted to RDF and integrated into a multi-domain KG comprising three interconnected semantic layers: the schedule graph, the photographic graph, and the alignment graph. As shown in Figure 7, this structure enabled construction-site photographs, detected objects, confidence values, scheduled activities and temporal relationships to be represented as connected semantic entities rather than as isolated tabular records.

The integrated graph contained 1,333 RDF triples, including scheduled activities, construction-site photographs, detected object entities, confidence-value entities and semantic alignment records. As summarised in Table 3, the graph included 10 scheduled tasks, 10 construction-site photographs, 13 detected object entities and 12 confidence entities. These results show that the workflow successfully transformed heterogeneous construction data into a queryable semantic structure that links planned activities, site photographs, detected resources, and temporal information.

The semantic alignment process was implemented through SPARQL queries executed in GraphDB. Rather than treating object detections as isolated outputs, the queries reconstructed the temporal and semantic context of each construction-site photograph by combining image timestamps extracted from EXIF metadata, detected objects and confidence values generated by *GroundingDINO*, scheduled tasks represented through the DTC ontology, temporal relationships between photographs and planned activities.

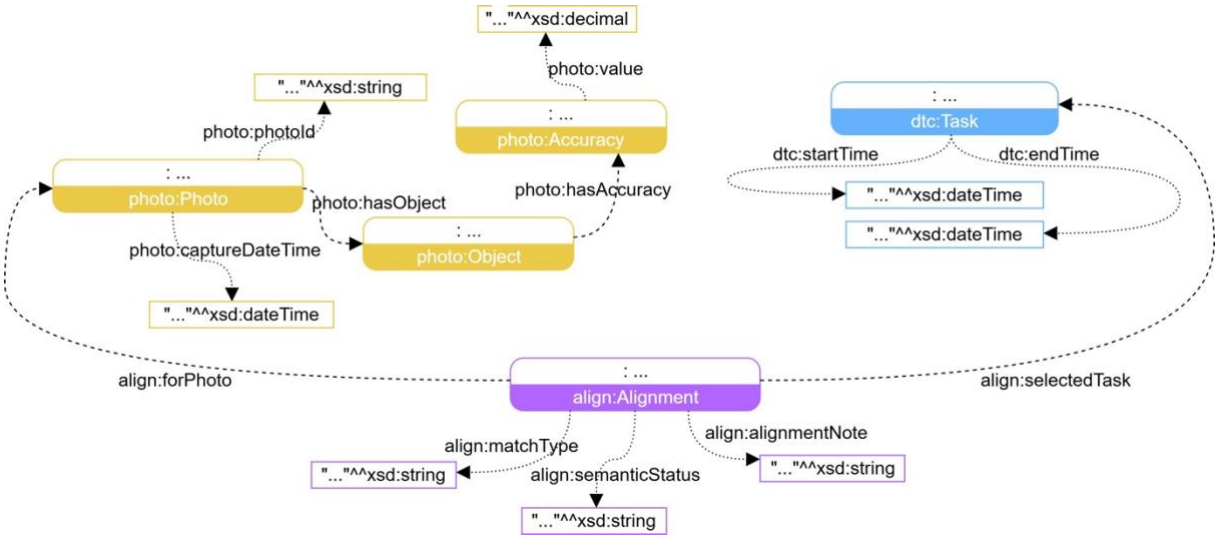

Figure 7: Excerpt of the integrated RDF KG showing semantic relations between photographs, detected resources, scheduled activities and semantic alignment entities.

Table 3: Composition of the integrated KG.

Semantic component	Number
Scheduled tasks (`dtc:Task`)	9
Construction-site photographs (`photo:Photo`)	10
Detected objects (`photo:Object`)	13
Confidence entities (`photo:Accuracy`)	12

For each photograph, the SPARQL query retrieved the capture date and detected visual resources, the active task on the photograph date, when present, the closest preceding and following scheduled tasks, the matched construction resource, the semantic alignment status assigned through rule-based logic (Table 4). The query normalised the property `photo:captureDateTime` to the day level (`?pDate`) and compared it with the task intervals represented through `dtc:startTime` and `dtc:endTime`. Active tasks were identified through temporal filtering conditions equivalent to `FILTER(?pDate >= ?taskStart && ?pDate <= ?taskEnd)`.

Table 4: Representative SPARQL-based semantic alignment results.

Photo ID	Detected object	Active task	Previous task	Next task	Semantic status
4fbf028325001281	drilling rig	Foundation piles	-	Excavation	`in_progress`
4fbf028325001281	excavator	-	Excavation	Beheading foundation piles	`possible_delay`
833e29d38239ca4c	tower crane anchor template	-	Tower crane anchor	Upper reinforcement bars	`completed_evidence`
2a42c10887396399	mobile crane	Tower crane jib installation	Tower crane mast installation	-	`in_progress`
5d9635bacd335d80	-	Upper reinforcement bars; Plinth pour	Tower crane anchor	Tower crane mast installation	`no_detection`

When no task was active on the image date, the nearest schedule context was reconstructed through two temporal subqueries based on `MAX(?prevEnd)` and `MIN(?nextStart)`, thereby identifying the latest completed task and the earliest following task associated with the photograph timestamp.

Aggregation functions, such as `GROUP_CONCAT`, were also used to combine multiple detections associated with the same image into a unified semantic representation of the visual content. As a result, the SPARQL output provided a structured semantic record that integrated visual, temporal, and schedule-related information into a single query result.

Overall, the alignment graph contained 23 semantic alignment records. Among these, 13 records were associated with an active scheduled task, while 10 were positioned between planned activities. Although the pipeline implements six semantic statuses. The case-study application generated records in five categories: `in_progress`, `possible_delay`, `completed_evidence`, `no_detection` and `needs_review` and their distribution is reported in Table 5. No record was classified as `not_classifiable` in the case study dataset.

Table 5: Distribution of semantic alignment statuses.

Semantic status	Records	Percentage
`in_progress`	7	30.4%
`possible_delay`	3	13.0%
`completed_evidence`	2	8.7%
`no_detection`	10	43.5%
`needs_review`	1	4.3%

The number of semantic alignment records does not directly correspond to the number of photographs because a single image may generate multiple alignment instances associated with different detected resources, expected labels, or semantic conditions. Therefore, the alignment process operates at the level of image–resource relationships rather than at the image level alone. For example, a single photograph containing both a mobile crane and a tower crane jib generated two independent semantic alignment records because each detected resource was evaluated separately against the schedule context.

The relatively high proportion of `no_detection` records demonstrates that the workflow preserves unsuccessful or uncertain detections as explicit semantic evidence rather than filtering them out from the semantic interpretation process. These categories demonstrate that the KG was not limited to storing object-detection outputs but also supported an explicit semantic reasoning layer that could interpret relationships between visual evidence and planned construction activities. Importantly, the semantic status was not generated by the LLM but inferred through explicit SPARQL-based temporal and semantic rules operating over the integrated KG.

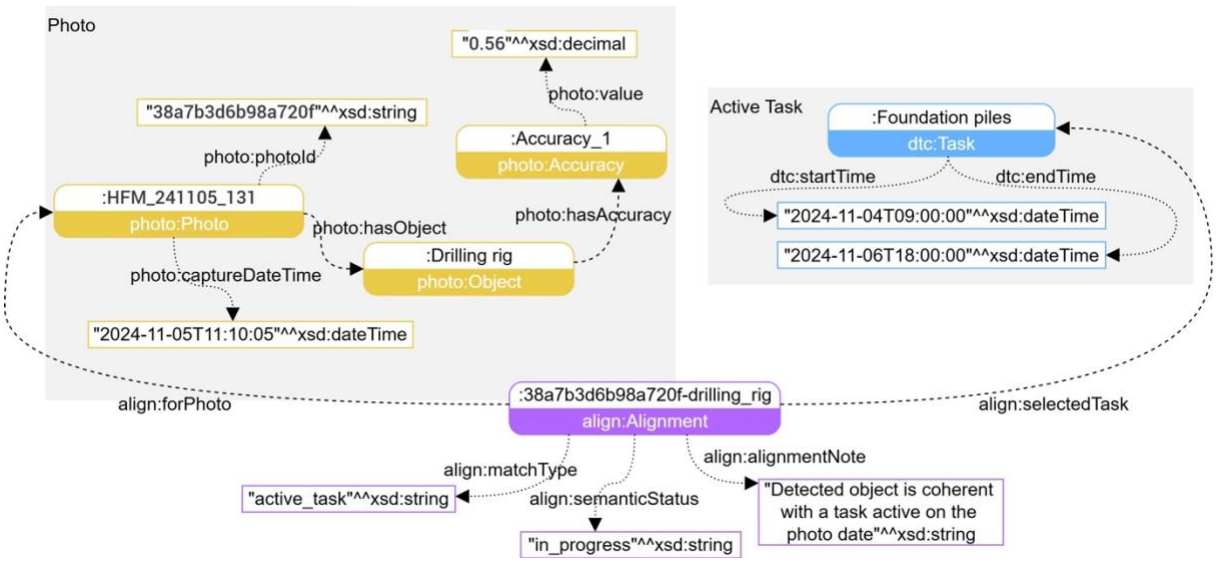

Figure 8: Representative `in_progress` alignment record, where the detected drilling rig is coherent with the active scheduled task Foundation piles on the photograph date.

6.2.1 In-progress semantic alignment

Among the 23 semantic alignment records, seven were classified as `in_progress`. A photograph was classified as `in_progress` when the detected object was coherent with a task active on the image date. Figure 8 presents a representative example.

The photograph captured on 05/11/2024 contains the detected object, a drilling rig for micropiles, associated with a confidence value of 0.56. The SPARQL query identified the task Foundation piles as active on the same date, with a planned interval between 04/11/2024 and 05/11/2024. Since the detected equipment corresponds to the expected resource associated with the active task, the semantic alignment procedure classified the image as `in_progress`. The same reasoning structure was observed for additional records associated with reinforcement activities and crane assembly operations.

For example, the objects mobile crane and tower crane jib detected in the photograph 2a42c10887396399 were both aligned with the task Tower crane jib installation, active on 12/12/2024. Similarly, the detections of reinforcement bars and a concrete mixer truck were associated with the combined activity: Upper reinforcement bars; Plinth pour on 29/11/2024. These results demonstrate that the SPARQL-based alignment successfully connected schedule intervals and detected resources into a coherent semantic interpretation of ongoing construction activities.

6.2.2 Possible-delay semantic alignment

Three semantic alignment records were classified as `possible_delay`. A photograph was classified as `possible_delay` when no scheduled task was active on the image date, but the detected resource remained semantically coherent with a previously completed activity. Figure 9 illustrates this situation.

The photograph captured on 19/11/2024 contains the detected objects "foundation trench" and "excavator", associated with confidence values of 0.43 and 0.49, respectively. The SPARQL query identified no active task on the capture date. However, the preceding task, "Excavation", ended on 18/11/2024, while the subsequent task, "Beheading foundation piles", was scheduled to start on 20/11/2024. Because the detected resources correspond to excavation-related equipment and work areas associated with the previous task, the image was semantically classified as `possible_delay`. This category does not necessarily indicate a schedule failure but highlights a discrepancy between planned and observed site conditions that requires further interpretation or expert validation.

Additional `possible_delay` records were associated with operational resources remaining visible after the planned completion of related construction tasks, including mobile lifting equipment observed outside the corresponding scheduled intervals.

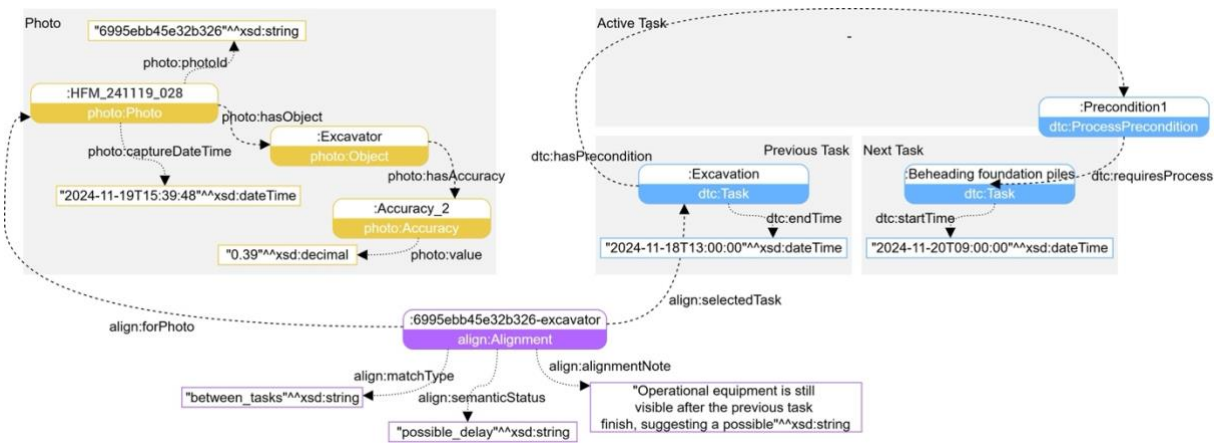

Figure 9: Representative possible delay record: an excavator detected after excavation completion but before Beheading piles, indicating a potential temporal discrepancy needing expert validation.

6.2.3 Completed-evidence semantic alignment

Two records were classified as `completed_evidence`. A photograph was classified as `completed_evidence` when the detected object represented a physical result of a previously completed task rather than operational equipment associated with ongoing work. Figure 10 presents an example of this condition.

The photograph captured on 25/11/2024 contains the detected object tower crane anchor template, associated with a confidence value of 0.64. The SPARQL query identified the task Tower crane anchor as the closest preceding activity, completed on 22/11/2024, while the next planned task (Upper reinforcement bars) was scheduled to start on 29/11/2024.

Since the detected object corresponds to a permanently installed component of the crane foundation system rather than temporary operational machinery, the semantic alignment interpreted the image as documentary evidence of the completion of the preceding task and classified it as `completed_evidence`.

This result highlights that the semantic interpretation depends not only on the detected object itself but also on the resource's semantic role and its temporal relationship with the project schedule.

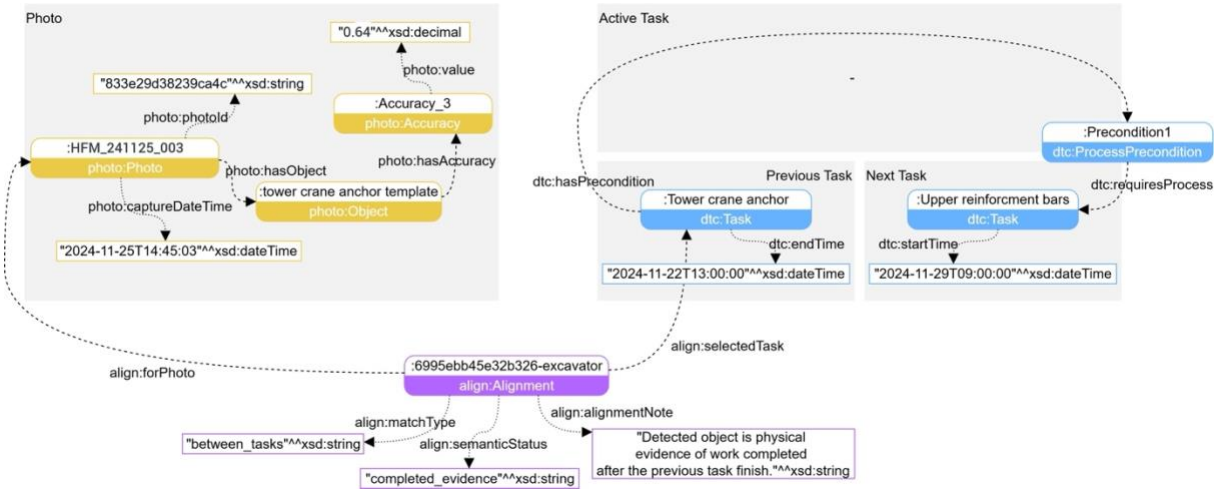

Figure 10: Representative completed_evidence alignment record, where the detected tower crane anchor template is interpreted as physical evidence of the previously completed Tower crane anchor task.

6.2.4 No-detection semantic alignment

Ten semantic alignment records were retained as `no_detection`. A photograph was classified as `no_detection` when an expected visual resource derived from the schedule was searched through *GroundingDINO* prompting, but no object was detected above the configured confidence threshold. Unlike conventional object-detection workflows, the proposed pipeline explicitly retained these negative results within the KG rather than discarding them. This choice allowed the absence of visual evidence to become part of the semantic reasoning process.

Several no-detection cases were associated with excavation-related activities and temporary construction conditions characterised by occlusions, partial visibility or insufficient visual distinction between the searched resource and the surrounding construction context. In these situations, the SPARQL alignment preserved the expected resource, the related scheduled activity and the absence of valid detections as traceable semantic evidence. This result is particularly relevant because it prevents the semantic interpretation process from relying exclusively on successful detections and supports a more transparent representation of uncertainty within construction progress monitoring.

6.2.5 Needs-review semantic alignment

One record was classified as `needs_review`. This category was assigned when the available information was insufficient to support a reliable semantic interpretation. This condition occurred in cases characterised by missing or ambiguous visual evidence, incomplete metadata or insufficient correspondence between the detected object and the schedule-derived resource.

Rather than forcing an automatic interpretation, the pipeline preserved these cases as explicit review conditions requiring expert validation. The introduction of a dedicated `needs_review` category demonstrates that the proposed semantic alignment workflow does not operate as a fully autonomous black-box classifier but preserves uncertainty and provenance information within the KG. This supports transparency, auditability and human-in-the-loop verification processes.

Table 6: Examples of LLM-assisted verbalisation outputs.

Semantic status	SPARQL-derived semantic record	LLM-assisted verbalisation output
`in_progress`	• Photograph HFM_241105_131.jpg, • captured on 2024-11-05; • detected object: drilling rig; • active task: Foundation piles; • confidence: 0.56.	The detected drilling rig is consistent with the Foundation piles task active on the photograph capture date.
`completed_evidence`	• Photograph HFM_241125_003.jpg, • captured on 2024-11-25; • detected object: tower crane anchor template; • previous task: Tower crane anchor; • confidence: 0.64.	The detected tower crane anchor template provides physical evidence associated with the previously completed Tower crane anchor activity.
`possible_delay`	• Photograph HFM_241119_028.jpg, • captured on 2024-11-19; • detected object: excavator; • previous task: Trench excavation; • confidence: 0.49.	The excavator remains visible after the planned completion of the Trench excavation task, suggesting a possible residual work condition or schedule delay requiring verification.
`no_detection`	• Photograph HFM_241129_001.jpg, • captured on 2024-11-29; • expected visual resource searched for the task Upper reinforcement bars / Plinth pour; • No valid detection above threshold for the selected label.	The expected resource was searched in the photograph, but no valid detection was obtained above the configured confidence threshold.
`needs_review`	• Photograph HFM_241119_028.jpg, • captured on 2024-11-20; • selected task: Beheading foundation piles; • No visual label available from Gantt resources.	The photograph requires expert review because the available information is insufficient for interpretation.

6.3 Results of the LLM-assisted verbalisation

The semantic alignment records were used as structured input for the LLM-assisted verbalisation step. The objective of this stage was to transform SPARQL-derived results into concise technical statements suitable for progress reporting, without generating new classifications. Each input record included the photograph identifier, capture date, detected or expected resource, confidence score when available, selected task, temporal context, semantic status and alignment note.

The generated statements preserved the semantic status assigned by the rule-based alignment procedure:

This is a bullet list:

- `in_progress` cases, the verbalisation described the detected object as evidence coherent with an activity active on the photograph date;
- `completed_evidence` cases, the statements described the detected object as physical evidence of a previously completed activity;
- `possible_delay` cases, they highlighted the presence of operational equipment after the planned end of the related activity, suggesting a potential residual work condition requiring verification;
- `no_detection` cases, the output explicitly reported that the expected resource was searched but not detected above the configured threshold;
- `needs_review`, the statement indicated that the available information was insufficient for automatic interpretation and required expert validation.

Representative examples of the verbalisation outputs are reported in Table 6. The table shows how SPARQL-derived semantic records are converted into report-oriented statements while maintaining the original semantic status and the traceability of the underlying evidence. These examples show that the LLM-assisted step improves the readability of semantic alignment results without altering the underlying rule-based classification. The statements generated provide a controlled textual representation of graph-derived evidence for construction progress documentation.

6.4 Discussion

The results from the case study show that the main value of the proposed pipeline lies in the transition from visual detection to semantically contextualised evidence of progress. The object-detection stage alone produced useful but incomplete information: detected elements such as drilling equipment, reinforcement bars, mobile cranes or tower crane components only become meaningful when interpreted in relation to the photograph timestamp and the planned activity. The integration with the KG and the SPARQL-based alignment provides the necessary context for transforming visual observations into progress-related evidence.

A key outcome is that the pipeline supports a more nuanced interpretation of construction progress than a simple detected/not-detected or planned/as-built comparison. The categories `in_progress`, `completed_evidence`, `possible_delay`, `no_detection` and `needs_review` allow the same type of visual evidence to be interpreted differently depending on its temporal and semantic relation to the schedule. This is particularly relevant for construction monitoring, where the presence of an object does not automatically indicate either progress or delay. The results confirm the importance of distinguishing between temporary resources, permanent physical outcomes and missing evidence.

The relatively high number of `no_detection` records is also significant. Rather than merely reflecting a weakness of the CV model, these cases underscore the importance of preserving negative evidence within the workflow. In construction-site images, missed detections can depend on occlusions, visual clutter, partial visibility or insufficient distinction between the searched object and the surrounding context. By retaining these cases in the KG, the pipeline avoids producing an overly positive interpretation based only on successful detections and makes uncertainty explicit. The `possible_delay` and `needs_review` categories are particularly important from an operational perspective. They do not provide a definitive assessment of non-compliance or delay but identify situations that require expert verification. This confirms that the workflow should be interpreted as a semi-automatic support tool rather than a fully autonomous progress-certification system. Its role is to structure evidence, highlight relevant cases and reduce the interpretative effort required.

The LLM-assisted verbalisation further improves the usability of the results by converting structured SPARQL outputs into concise technical statements. Since the semantic status is assigned before the verbalisation stage, the LLM does not act as a decision-making component, but as a reporting layer. This separation is important because it improves readability without compromising the traceability of the rule-based interpretation. The generated statements are better suited for progress reporting than raw query outputs, while remaining connected to the underlying photograph, detected resource, scheduled task, and semantic status.

Overall, the case study confirms that the proposed pipeline can produce traceable and interpretable evidence of construction progress from heterogeneous projects and site data. Its reliability, however, depends on the quality of the input schedule, the availability of accurate photographic metadata, the visual recognisability of the resources being searched, and the calibration of detection thresholds. For this reason, the workflow should be considered a structured decision-support method for progress documentation, rather than a replacement for technical judgement. Its main contribution is the creation of an auditable chain from the site photograph to the semantic status and the final reporting statement.

7. CONCLUSIONS

This paper presented a schedule-driven semantic pipeline for linking construction-site photographs, project schedules, CV outputs and controlled natural-language reporting. The proposed workflow uses planned activities and associated resources to guide open-vocabulary object detection, integrates the resulting visual evidence into a KG, and applies SPARQL-based rules to assign progress-related semantic statuses. The subsequent LLM-assisted

verbalisation step transforms these structured records into concise technical statements without altering the rule-based classification.

The case study on the installation of a 100-metre free-standing tower crane showed that the pipeline can go beyond isolated object detection by producing traceable evidence of progress linked to photographs, timestamps, detected resources, and scheduled activities. The results demonstrate the potential of combining schedule-driven CV, semantic modelling, and controlled verbalisation to support transparent and semi-automatic documentation of construction progress. But the results also highlight several limitations.

First, the validation was limited to a single construction site, one specific construction phase, ten photographs, and a relatively small set of scheduled activities. Therefore, the findings should be interpreted as evidence of methodological feasibility rather than as a general validation of the workflow across different projects and construction contexts. Second, the reliability of the workflow depends on the quality, completeness, and continuous updating of the input schedule. Since the schedule is the main reference for both object-detection label generation and progress interpretation, incomplete, ambiguous, or outdated activity and resource information may affect not only visual analysis but also the overall interpretation of monitoring results. Construction schedules are frequently revised during project execution; therefore, if the schedule is not regularly maintained and aligned with the current construction plan, the pipeline may systematically identify inconsistencies that reflect schedule obsolescence rather than actual deviations on site. Third, the CV stage remains sensitive to construction-site conditions. Objects with clear geometry and strong visual distinctiveness were detected more reliably, whereas partially occluded, repetitive, or visually ambiguous elements yielded lower-confidence results or no detections. The relatively high number of `no_detection` records confirm that visual evidence cannot be interpreted independently of image quality, object visibility, and threshold calibration.

Finally, the current semantic interpretation is primarily based on the relationships among photographs, detected resources, and schedule information. This may be insufficient in complex construction sites where several parallel activities involve similar or identical resource descriptions. In such cases, temporal alignment alone may not reliably associate a photograph or detected object with the correct ongoing task. The integration of spatial or topological information, such as photograph coordinates, construction zones, building levels, or explicit links between scheduled activities and work areas, would therefore be required to improve the reliability of task–evidence association. More generally, the current pipeline supports progress-oriented documentation, but it does not yet provide comprehensive validation against broader project requirements, construction procedures, or site-safety prescriptions.

Therefore, future work will focus on extending the semantic layer beyond schedule-based progress monitoring. A key development will be the definition of a construction-site ontology capable of structuring project and site information currently contained in predominantly textual documents, such as method statements, safety plans, technical specifications, work procedures, inspection reports and site instructions. Such an ontology would make it possible to represent not only activities, resources and timestamps, but also execution requirements, safety constraints, operational procedures, roles, equipment conditions and compliance criteria. This extension would support more comprehensive validation of on-site visual observations. Detected objects and construction activities could be assessed not only against the planned schedule, but also against the expected execution methods and safety requirements. For example, future versions of the pipeline could verify whether the resources visible in a photograph are consistent with the planned activity, whether the observed configuration complies with prescribed construction procedures, and whether specific safety-related conditions are present or missing.

The long-term objective is to evolve the current progress-monitoring workflow into a broader validation framework for construction sites. By integrating visual evidence, scheduling data, and structured knowledge extracted from project documentation, the system could support a more comprehensive verification of construction progress, execution conformity, and site safety compliance within a single, traceable KG-based environment.

From this perspective, the research supports SDG 9 (Sustainable Development Goal) by advancing digital innovation and interoperable infrastructure monitoring, SDG 11 by promoting more transparent and reliable management of construction processes within sustainable built environments, and SDG 12 by enabling better documentation, traceability, and resource-aware decision-making during construction. Future extensions toward safety and compliance verification may also strengthen the contribution to SDG 8, particularly regarding safer, more accountable working conditions on construction sites.

DATA AND CODE AVAILABILITY

The scripts and source code used to implement the proposed pipeline are available from the corresponding author upon reasonable request. Selected ontology files and anonymised representative data may also be provided, subject to confidentiality constraints.

AUTHOR CONTRIBUTIONS

Conceptualisation, F.M., J.C., and M.T.; Methodology, F.M. and J.C.; Code, F.M. and J.C.; Validation, F.M., and J.C.; Writing—original draft, F.M.; Writing—review and editing, F.M., J.C., and M.T.; Supervision, J.C., and M.T. All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGEMENTS

This paper is an extended version of our previous work, which was presented at the Smart and Sustainable Built Environment Conference, Lille – France | 3rd to 5th November 2025. The authors acknowledge the support and feedback from the conference chairs, Prof. Farzad Rahimian, Professor Abiola Akanmu and Dr M. Reza Hosseini, along with their team, throughout the peer review process of SASBE 2025 and during the conference, which helped improve our submissions.

REFERENCES

Beetz, J., van Leeuwen, J. and de Vries, B. (2009). IfcOWL: A case of transforming EXPRESS schemas into ontologies, Artificial Intelligence for Engineering Design, Analysis and Manufacturing, 23(1), 89–101. https://doi.org/10.1017/S0890060409000122.

Borrmann, A., Sacks, R., Martinez, J., Vilgertshofer, S., Yeung, T. and Schlenger, J. (2022). A Comprehensive Data Schema for Digital Twin Construction, Aarhus Universitet. Retrieved May 25, 2026, from https://openbooks.kb.dk/au/catalog/view/455/312/1842

Charef, R., Alaka, H. and Emmitt, S. (2018). Beyond the third dimension of BIM: A systematic review of literature and assessment of professional views, Journal of Building Engineering, 19, 242–257. https://doi.org/10.1016/j.jobe.2018.04.028

Chen, R., Liang, S., Li, J., Liu, S., Li, M., Huang, Z., Zhang, H. and Cao, X. (2024). Interpreting Object-level Foundation Models via Visual Precision Search, arXiv preprint arXiv:2411.16198. https://doi.org/10.48550/arXiv.2411.16198

Ding, Y., Zhang, M., Pan, J., Hu, J. and Luo, X. (2024). Robust object detection in extreme construction conditions, Automation in Construction, 165, 105487. https://doi.org/10.1016/j.autcon.2024.105487

Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J. and Zisserman, A. (2010). The PASCAL Visual Object Classes (VOC) Challenge, International Journal of Computer Vision, 88(2), 303–338. https://doi.org/10.1007/s11263-009-0275-4

Farghaly, K., Soman, R.K. and Zhou, S.A. (2023). The evolution of ontology in AEC: A two-decade synthesis, application domains, and future directions, *Journal of Industrial Information Integration*, 36, 100519. https://doi.org/10.1016/j.jii.2023.100519

Geng, S., Cooper, H., Moskal, M., Jenkins, S., Berman, J., Ranchin, N., West, R., Horvitz, E. and Nori, H. (2025). JSONSchemaBench: A Rigorous Benchmark of Structured Outputs for Language Models, arXiv preprint arXiv:2501.10868. https://doi.org/10.48550/arXiv.2501.10868

Golparvar-Fard, M., Peña-Mora, F. and Savarese, S. (2009). Application of D4AR – A 4-Dimensional augmented reality model for automating construction progress monitoring data collection, processing and communication, *Journal of Information Technology in Construction (ITcon)*, 14, 129–153. Available at: https://www.itcon.org/2009/13

Grattafiori, A. et al. (2024). The Llama 3 Herd of Models, arXiv preprint arXiv:2407.21783. https://doi.org/10.48550/arXiv.2407.21783

Han, K.K., Cline, D. and Golparvar-Fard, M. (2015). Formalized knowledge of construction sequencing for visual monitoring of work-in-progress via incomplete point clouds and low-LoD 4D BIMs, Advanced Engineering Informatics, 29(4), 889–901. https://doi.org/10.1016/j.aei.2015.10.006

IDEA Research (2025). GroundingDINO. GitHub repository. Available at: https://github.com/IDEA-Research/GroundingDINO (accessed 27 August 2025).

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Dai, W., Chan, H.S., Madotto, A. and Fung, P. (2023). Survey of Hallucination in Natural Language Generation, ACM Computing Surveys, 55(12), Article 248, 1–38. https://doi.org/10.1145/3571730

Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Chang, K.W. and Gao, J. (2022). Grounded Language-Image Pre-training. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10965–10975. https://doi.org/10.48550/arXiv.2112.03857

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L. (2014). Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B. and Tuytelaars, T. (eds.), Computer Vision – ECCV 2014, Lecture Notes in Computer Science, 8693, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J. and Zhang, L. (2023). Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection, arXiv preprint arXiv:2303.05499. https://doi.org/10.48550/arXiv.2303.05499

Madaschi, F., Cassandro, J. and Trani, L. (2026). Cross-checks between project and site: an integrated workflow based on information models and computer vision, Proceedings of the International Conference on Smart and Sustainable Built Environment. Retrieved from https://doi.org/10.1007/978-981-95-8489-5

Madaschi, F., Taherian, G., Trani, M.L.A. and Rezazadeh Azar, E. (2026). Vision–Language Models in Construction: Assessing Prompt Robustness in Open-Vocabulary Object Detection. In: *Proceedings of European Conference on Computing in Construction*, forthcoming.

Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., Wang, X., Zhai, X., Kipf, T., Houlsby, N. (2022). Simple Open-Vocabulary Object Detection with Vision Transformers. In: Computer Vision – ECCV 2022, Lecture Notes in Computer Science, 13670, 728–755. https://doi.org/10.1007/978-3-031-20077-9_42

Moragane, H.P.M.N.L.B., Perera, B.A.K.S., Palihakkara, A.D. and Ekanayake, B. (2024). Application of computer vision for construction progress monitoring: A qualitative investigation, Construction Innovation, 24(2), 446–469. https://doi.org/10.1108/CI-05-2022-0130

Oesterreich, T.D. and Teuteberg, F. (2016). Understanding the implications of digitisation and automation in the context of Industry 4.0: A triangulation approach and elements of a research agenda for the construction industry, Computers in Industry, 83, 121–139. https://doi.org/10.1016/j.compind.2016.09.006

Ontotext (2025). GraphDB Downloads and Resources. Available at: https://graphdb.ontotext.com/ (accessed 26 August 2025).

Pauwels, P. and Terkaj, W. (2016). EXPRESS to OWL for construction industry: Towards a recommendable and usable ifcOWL ontology, Automation in Construction, 63, 100–133. https://doi.org/10.1016/j.autcon.2015.12.003

Pfitzner, F., Braun, A. and Borrmann, A. (2024). From data to knowledge: Construction process analysis through continuous image capturing, object detection, and knowledge graph creation, *Automation in Construction*, 164, 105451. https://doi.org/10.1016/j.autcon.2024.105451

Rasmussen, M.H., Lefrançois, M., Pauwels, P., Hviid, C.A. and Karlshøj, J. (2019). Managing interrelated project information in AEC Knowledge Graphs, Automation in Construction, 108, 102956. https://doi.org/10.1016/j.autcon.2019.102956

Rasmussen, M.H., Lefrançois, M., Schneider, G.F. and Pauwels, P. (2021). BOT: The Building Topology Ontology of the W3C Linked Building Data Group, Semantic Web, 12(1), 143–161. https://doi.org/10.3233/SW-200385

Rebolj, D., Pučko, Z., Babič, N.Č., Bizjak, M. and Mongus, D. (2017). Point cloud quality requirements for Scan-vs-BIM based automated construction progress monitoring, Automation in Construction, 84, 323–334. https://doi.org/10.1016/j.autcon.2017.09.021

Rehman, M.S.U., Shafiq, M.T. and Ullah, F. (2022). Automated Computer Vision-Based Construction Progress Monitoring: A Systematic Review, *Buildings*, 12(7), 1037. https://doi.org/10.3390/buildings12071037

Ren, T., Jiang, Q., Liu, S., Zeng, Z., Liu, W., Gao, H., Huang, H., Ma, Z., Jiang, X., Chen, Y., Xiong, Y., Zhang, H., Li, F., Tang, P., Yu, K. and Zhang, L. (2024). Grounding DINO 1.5: Advance the "Edge" of Open-Set Object Detection, arXiv preprint arXiv:2405.10300. https://doi.org/10.48550/arXiv.2405.10300

Schlenger, J.U., Pluta, K., Mathew, A., Yeung, T., Sacks, R. and Borrmann, A. (2024). Reference Architecture and Ontology Framework for Digital Twin Construction. https://doi.org/10.2139/ssrn.4887436

Shen, X., Sepasgozar, S.M.E. and Ostwald, M.J. (2024). Knowledge-based semantic web technologies in the AEC sector, Automation in Construction, 167, 105686. https://doi.org/10.1016/j.autcon.2024.105686.

Taherian, G., Madaschi, F., Rezazadeh Azar, E. and Trani, M.L.A. (2026). Vision–Language Models in Construction: Evaluating Open-Vocabulary Object Detection under Complex Scenarios. In: *Proceedings of CSCE Québec 2026 / SCGC Québec 2026*, Quebec City, Canada, 3–5 June 2026, forthcoming.

Teizer, J., Lao, D. and Sofer, M. (2007). Rapid Automated Monitoring of Construction Site Activities Using Ultra-Wide Band. In: Proceedings of the 24th International Symposium on Automation and Robotics in Construction (ISARC 2007), Kochi, India, pp. 23–28. https://doi.org/10.22260/ISARC2007/0008

Yin, M., Tang, L., Webster, C., Xu, S., Li, X. and Ying, H. (2023). An ontology-aided, natural language-based approach for multi-constraint BIM model querying, Journal of Building Engineering, 76, 107066. https://doi.org/10.1016/j.jobe.2023.107066

Zheng, Y., Törmä, S. and Seppänen, O. (2021). A shared ontology suite for digital construction workflow, Automation in Construction, 132, 103930. https://doi.org/10.1016/j.autcon.2021.103930