Sabek, M., Mei, Q., Lee, G., Golabchi, A., & Gonzalez, V. A. (2026). The Lean Construction Visual Taxonomy (LCVT): bridging the semantic gap. In F. Hamzeh, M. Poshdar, N. P. Garcia-Lopez, & V. Gan (Eds.), *Proceedings of the 34th Annual Conference of the International Group for Lean Construction* (IGLC34) (pp. 14–25). https://doi.org/10.24928/2026/0151

THE LEAN CONSTRUCTION VISUAL TAXONOMY (LCVT): BRIDGING THE SEMANTIC GAP

Mohamed Sabek[1], Qipei Mei[2], Gaang Lee[3], Ali Golabchi[4], & Vicente Gonzalez[5]

ABSTRACT

The architecture, engineering, and construction (AEC) industry faces productivity stagnation due to ineffective production flow management. Although Lean Construction (LC) aims to minimize waste, manual monitoring lacks the high-frequency data required for timely control. Computer Vision (CV) offers automated monitoring but suffers from a "Semantic Gap," where models detect low-level objects but fail to interpret high-level Lean states (e.g., "waiting"). This study proposes the Lean Construction Visual Taxonomy (LCVT), a three-level hierarchical framework–Category, Indicator, Visual Definition grounded in Transformation-Flow-Value (TFV) theory. Crucially, the LCVT provides standardized class definitions to guide "zero-shot" prompt engineering in Vision-Language Models (VLMs). By injecting formal L3 definitions that address entity types, temporal thresholds (e.g., stationary >60 s), and spatial context into VLM models such as GPT-4o and Gemini 2.5, the framework enables sophisticated, lean reasoning without the need for massive custom-labeled datasets. Pilot validation achieved a 0.946 mAP in distinguishing state-dependent equipment loads. By formalizing the visual signatures of waste, the LCVT establishes the data infrastructure necessary for proactive, VLM-driven decision support in construction AI.

KEYWORDS

AI, transformation-flow-value, computer vision, taxonomy, visual management.

INTRODUCTION

The global construction industry is currently facing a productivity crisis, with Sveikauskas et al. (2016) finding that labor shifts between construction sub-industries reduce productivity growth by 0.4% per year, and that overall productivity gains lag significantly behind manufacturing. Projects frequently suffer from cost overruns, schedule delays, and persistent safety hazards, making them among the most dangerous workplaces in the world (Oswald et al., 2020). In response, Lean Construction (LC) has emerged as a robust alternative in

[1] PhD Candidate, Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Canada, sabek@ualberta.ca, orcid.org/0009-0005-2906-9874
[2] Assistant Professor, Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Canada, qipei@ualberta.ca,
[3] Assistant Professor, Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Canada, gaang@ualberta.ca,
[4] Adjunct Professor, Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Canada, alireza1@ualberta.ca,
[5] Professor, Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Canada, vagonzal@ualberta.ca, orcid.org/0000-0003-3408-3863

production management, aimed at maximizing customer value and systematically eliminating waste (Muda) (Alarcón, 1997).

Lean-based tools, such as the Last Planner System (LPS), operationalize these principles by focusing on reliable planning and constraint removal (Ballard, 2000). However, the efficacy of such systems is limited by data quality, as site monitoring still relies heavily on manual, resource-intensive, subjective reporting (Shankaranarayanan & Cai, 2006). This creates a "data gap" that impedes root cause analysis and continuous improvement.

Artificial Intelligence (AI), specifically Computer Vision (CV), offers a transformative solution for automated, continuous, and objective monitoring (Sabek et al., 2024). CV has achieved safety compliance (Seo et al., 2015) and has made progress in monitoring against BIM models (Golparvar-Fard et al., 2009); however, a significant "Semantic Gap" remains (Wang et al., 2024).

While current CV systems are proficient at identifying the presence of resources (e.g., "three workers"), they fail to interpret their state in the production flow, such as distinguishing value-added work from "Waiting Waste". Without a standardized visual language, the definitions of "waste" remain inconsistent and fragmented. This study aims to address this gap by developing a Lean Construction Visual Taxonomy (LCVT). Adopting a Constructive Research approach, we propose the LCVT as a theoretical artifact that translates managerial objectives into machine-readable visual definitions designed to guide supervised learning and unlock the potential of Vision-Language Models (VLMs) through structured prompt engineering. The specific objectives are: (1) to develop a hierarchical, TFV-grounded taxonomy that translates abstract Lean waste categories into formal, machine-readable visual definitions; and (2) to demonstrate the algorithmic feasibility of state-based classification using LCVT-defined classes as an initial proof of concept.

THEORETICAL FOUNDATIONS

To ensure that the LCVT serves as a robust management tool rather than a mere catalog of objects, it is grounded in established production theory and current AI capabilities.

LEAN PRODUCTION THEORY: THE TFV VIEW

The theoretical backbone of the LCVT is the (TFV) theory of production proposed by Koskela (2000). Traditional construction management views production through the transformation view, which focuses on converting inputs into outputs; current CV applications mirror this by detecting active work, such as "Bricklaying". However, Lean theory emphasizes two additional critical views. The Flow View focuses on minimizing non-transformation activities, such as inspection, waiting, moving, and inventory. The LCVT makes these activities visible by defining visual states for "Waiting," "Motion," and "Inventory". The Value View emphasizes ensuring that products meet customer requirements, supported in the LCVT through visual indicators for "Defects" and "Rework" to ensure that work is "done right".

THE SEMANTIC GAP IN CONSTRUCTION AI

The "Semantic Gap" refers to the disconnect between the low-level features extracted by an AI model (pixels, edges, object bounding boxes) and the high-level concepts required by a user (waste, efficiency, safety risk) (Asadollahi et al., 2025). In the construction context, this gap manifests as a difference between observational and interpretive AI (Wu et al., 2021).

- **Observational AI:** "There is a worker at coordinates (x, y)." (Current State of the Art).
- **Interpretive AI:** "The worker at (x, y) is experiencing a type of waste, 'Waiting', due to 'Poor Material Staging'." (The Goal of LCVT).

In the absence of a formal taxonomy to map low-level observations to high-level concepts, AI remains a passive observer. Without such systematic categorization, AI cannot bridge the gap between detection and interpretation.

THE EMERGENCE OF VISION-LANGUAGE MODELS (VLMS)

The timing of this framework is critical, given the rapid emergence of VLMs such as GPT-4o, Gemma 3, and Florence-2 (Gemma 3 Team et al., 2024; Xiao et al., 2023). Unlike traditional CV models, which require thousands of labeled images to learn specific classes, VLMs interpret visual content using "zero-shot" natural-language prompts. However, because VLMs are generic and do not inherently understand "Lean Construction," they are susceptible to generating inaccurate outputs (hallucinations) when prompted with underspecified queries about site efficiency. The LCVT provides the necessary domain logic for structured prompt engineering, enabling these powerful models to perform lean analysis by feeding them precise definitions (e.g., "Identify 'Idle Equipment' defined as...") without the need for massive custom-labeled datasets. While VLMs enable zero-shot inference, which provides great results in general cases, fine-tuning on construction-specific datasets, such as SODA (Duan et al., 2022), ACID (Xiao & Kang, 2021), and CIVAD (Sabek et al., 2025), can optimize their performance and outputs to be more specific to construction definitions and naming schemes.

METHODOLOGY: CONSTRUCTIVE RESEARCH APPROACH

This research follows the Constructive Research (Design Science Research) methodology, as defined by March and Smith (1995) and operationalized by Hevner et al. (2004). Unlike descriptive research, which seeks to understand existing phenomena, constructive research creates innovative artifact constructs, models, methods, or instantiations that solve identified practical problems. The validity of an artifact is determined by its utility and relevance rather than by empirical generalization alone (Kasanen et al., 1993). The LCVT addresses a specific problem: the inability of current CV systems to interpret Lean Construction concepts. The development followed a three-phase protocol that combined systematic literature analysis, principled taxonomy design, and operationalization for AI implementation. This approach ensures that the artifact is both theoretically grounded and practically implementable. It should be noted that the current validation focuses on algorithmic feasibility; practitioner-based evaluation, which would further strengthen the artifact's validity under Design Science principles, remains an important direction for future work.

PHASE 1: LITERATURE-BASED INDICATOR IDENTIFICATION AND HIERARCHICAL DESIGN

The LCVT development required the systematic extraction of observable waste indicators from the lean literature and the establishment of a taxonomic structure suitable for AI implementation. This phase involved literature synthesis, a comparative analysis of existing classification systems, and justification of the hierarchical design.

Literature synthesis process

A targeted literature review focused on three domains: (1) foundational Lean texts (Ohno, 1988; Shingo, 1989), (2) Lean Construction applications (Koskela, 2000; Ballard, 2000), and (3) digital monitoring studies (Barbosa & Costa, 2021), to identify the physical manifestations of waste. For example, Ohno's (1988) abstract definition of "waiting" lacks the specificity required for visual detection. By examining work sampling studies (Gouett et al., 2011) and time-motion analyses, we identified that waiting manifests visually through specific states: workers stationary with no tools in hand, equipment parked with engines running but controls disengaged, and crews standing without active engagement. This iterative abstraction-to-

observation process was repeated for each waste category, yielding a comprehensive list of preliminary visual indicators.

Comparative analysis of existing classification systems

To ensure that the LCVT represents a theoretical advancement rather than an ad hoc list, we analysed existing taxonomic structures in adjacent domains. In manufacturing, the Toyota Production System organizes waste into categories and subcategories. In computer vision, ImageNet and COCO use hierarchical class structures. In construction safety, Haslam et al. (2005) developed a multi-level taxonomy that organizes accident causation from high-level influences on specific factors. However, none of these systems bridges the specific semantic gap in lean construction. Models trained on ImageNet can detect the class "person" but not "idle person." Safety taxonomies focus on hazards rather than production efficiency. Existing Lean Construction frameworks remain conceptual rather than providing operational visual definitions. This gap justifies a purpose-built taxonomy.

Hierarchical design rationale: the three-level structure

The three-level hierarchy (Category to Indicator to Visual Definition) is a proposed structure synthesized by the authors to align requirements emerging from TFV theory and AI implementation constraints:

1. **L1 - Category (Managerial Alignment):** Decision-makers think in terms of high-level objectives aligned with TFV theory: "Reduce Waiting Waste" (Flow view), "Eliminate Defects" (Value view). L1 categories provide a conceptual anchor, ensuring that AI outputs remain interpretable by non-technical stakeholders.
2. **L2 - Indicator (Observable Phenomena):** Computer vision operates on observable states, not abstractions. L2 indicators translate conceptual categories into specific phenomena that cameras capture: "Idle Worker," "Site Clutter," "Blocked Access." This level serves as the "vocabulary" for the dataset annotation and model class definitions.
3. **L3 - Visual Definition (Unambiguous Operationalization):** Each indicator requires a formal definition specifying visual conditions, temporal constraints, and contextual parameters. For example, "Idle Worker" is defined as *"a worker in an active work zone who is stationary for greater than 60 seconds, has no tools in hand, and is not interacting with others."* This specificity enables both human annotators and AI systems to make consistent judgments.

Drawing on knowledge representation theory, this structure mirrors ontology engineering practices (Guarino, 1998), in which taxonomies organize domain knowledge from abstract concepts to operational definitions. This hierarchical organization is designed to facilitate scalability, enabling new indicators to be added without restructuring the framework. It also supports interpretability by providing a logical framework for aggregating AI outputs from L3 detections to L2 metrics to L1 management reports.

PHASE 2: OPERATIONALIZATION AND FORMAL SPECIFICATION

Phase 2 focused on transforming abstract indicators into formal visual definitions (L3) that are suitable for direct implementation in computer vision systems and dataset annotation protocols. To ensure consistency across the L3 definitions, we developed a standardized specification template synthesized from the analysis of visual ambiguities in Phase 1. This template requires each definition to address six attributes to systematically eliminate the vagueness found in previous studies.

This attribute-based approach aligns with ontology engineering principles (Guarino, 1998), which define entities by their properties and relationships.

- **Entity Type:** Primary object(s) of observation (worker, equipment, material, space)

- **Visual State Condition:** Observable physical characteristics (stationary, moving)
- **Temporal Constraint:** Duration thresholds distinguishing states (>60 seconds vs. brief pause)
- **Spatial Context:** Location parameters (active work zone, walkway, laydown area)
- **Relational Condition:** Interactions with other entities (tool in hand, blocking path)
- **Exclusion Criteria:** Explicit negations preventing misclassification (not on break, not inspecting)

For example, "Idle Worker" might be operationalized as:

Entity: Worker; **Visual State:** Stationary; **Temporal:** >60 seconds; **Spatial:** Active work zone; **Relational:** No tools AND not interacting; **Exclusion:** Not performing inspection

This contrasts with informal definitions in prior CV research, where "idle" has been variously defined as "not moving" (no temporal threshold), "standing still" (no spatial context), or "not performing recognized activity" (a circular definition). The LCVT's explicit six-attribute specification eliminates such ambiguity.

Temporal Threshold Calibration

Establishing appropriate temporal thresholds requires a literature-informed approach. Work sampling studies (Gouett et al., 2011) typically use 30-60 second observation intervals to distinguish productive from non-productive time. We adopted 60 s as the default threshold, a conservative value that minimizes false positives from brief legitimate pauses while remaining within established work sampling intervals.

All temporal parameters in the L3 definitions are explicitly noted as configurable (denoted by variables X, Y, and Z), allowing practitioners to adapt the definitions to specific contexts while maintaining a consistent logical structure. Phase 2 produced the complete LCVT specification, including L2 indicators, each with formal L3 definitions following the six-attribute template, calibrated temporal thresholds, and documented inter-indicator relationships, directly implemented for dataset annotation and AI model training.

PHASE 3: ALGORITHMIC FEASIBILITY VALIDATION

Unlike traditional empirical studies that validate hypotheses through statistical testing, Constructive Research validates artifacts through feasibility assessment and proof-by-demonstration (Hevner et al., 2004). For the LCVT, the critical validation question is as follows: *Can traditional computer vision systems successfully differentiate between lean waste states using the LCVT L3 definitions, moving beyond generic object detection to state-aware classification?*

Validation strategy and indicator selection

To validate this capability, we conducted pilot CV model training focusing on the "Inefficient Transport" indicator within the Transportation Waste category. This indicator was strategically selected because it represents a critical test case: *successful lean monitoring requires distinguishing between different states of the same object type*, not just detecting the presence of an object. The LCVT L3 definition specifies detecting material handling equipment that performs "empty runs" or moves with "partial loads". As demonstrated below, successfully distinguishing "Dump_Truck_Loaded" from "Dump_Truck_Empty" shows that the LCVT definitions provide sufficient visual specificity for the models to learn state-dependent features rather than just generic truck detection. Equipment load states were selected over worker activity states because they present unambiguous physical visual features (visible material in the truck bed), providing a conservative starting point for proof-of-concept validation. Worker

states, which depend on pose interpretation and temporal context, represent a more complex classification challenge addressed in future research agendas.

Dataset and experimental setup

A targeted subset of the Construction Industry Vision Alberta Dataset (CIVAD) (Sabek et al., 2025) was utilized, comprising 800 annotated images of earthmoving operations. Images of dump trucks in both loaded and unloaded states were captured under varying conditions. Each image was labeled according to the LCVT specifications, with bounding boxes and state-based class labels: "Dump_Truck_Loaded" (visible material in the bed), "Dump_Truck_Empty" (empty bed with a higher ride height), and "Excavator" (reference equipment). A YOLOv8-based object detection model was fine-tuned on this LCVT-annotated subset using standard training protocols (70/15/15 train/validation/test split and 300 epochs). The performance was evaluated using the mean average precision (mAP) at an IoU threshold of 0.5.

Results and validation outcome

The trained model achieved a 0.946 mAP (Mean Average Precision), providing an initial feasibility demonstration for state-based classification. Figure 1 illustrates real-time inference on construction site videos, showing the successful differentiation between loaded and empty dump trucks, along with spatial heatmaps that indicate equipment movement patterns to identify inefficient transport cycles. The CV model distinguishes between "Dump_Truck_Loaded" and "Dump_Truck_Empty" on an active earthmoving site. The detection results showed two excavators and multiple dump trucks, each classified by load state. The spatial heatmap overlay (blue-purple gradient) visualizes the cumulative movement patterns, enabling the identification of frequent "empty run" routes. The model achieved mAP of 0.946 on test images, demonstrating that LCVT definitions enable state-based waste classification beyond generic object detection.

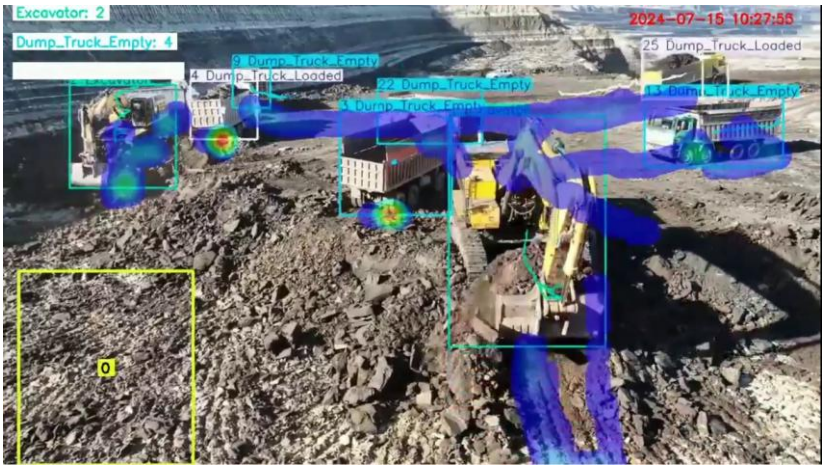

Figure 1: LCVT-based detection of transportation waste states

Key findings

The pilot study results suggest four critical outcomes: (1) The mAP of 0.946 on state-specific classes (Loaded vs. Empty) provides initial evidence that CV models can learn to distinguish visual states defined by the LCVT; (2) The semantic gap bridged model outputs align directly with Lean management objectives rather than producing generic object lists; (3) Initial operational feasibility: performance suggests L3 visual definitions are sufficiently specific for consistent annotation and model training; and (4) Immediate implementability indicate that spatial indicators with clear visual signatures can be deployed using standard object detection architectures. This pilot validation demonstrates proof of concept for spatial indicators with

clear visual signatures; validation of temporal indicators (e.g., Idle Worker) and multi-entity relational indicators remains for future work.

THE ARTIFACT: LEAN CONSTRUCTION VISUAL TAXONOMY (LCVT)

The LCVT serves as a semantic layer that makes the construction flow and its invisible inefficiencies machine-readable. Figure 2 illustrates this by positioning the proposed visual indicators within the established production mechanism, highlighting the exact location of the taxonomy intervention for waste detection. The arrows in Figure 2 indicate the 'data extraction' points at which the LCVT -guided analytics modules captured visual evidence of waste from the physical process.

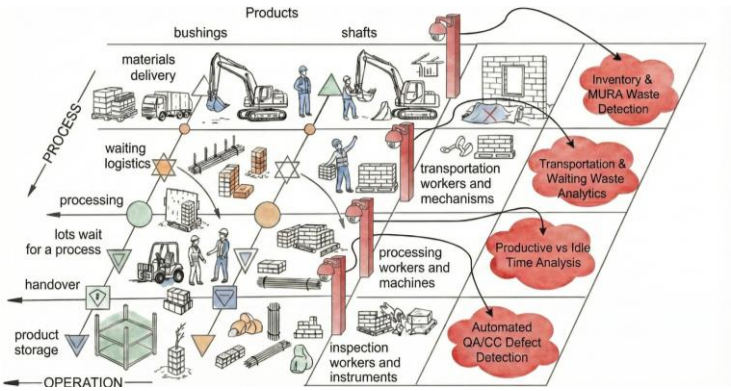

Figure 2: Visualizing waste in the transformation flow framework. adapted from the process-operation mechanism (Shingo, 1989; Koskela, 2000)

COMPREHENSIVE LCVT FRAMEWORK: WASTES, WORKFLOW, AND SAFETY

Figure 3 presents the complete LCVT framework organized into three interconnected sections: Lean Wastes (Muda), Workflow and Flow Indicators, and Safety-Lean Dual-Impact Mappings. These three sections are interconnected through complementary functions in production management. The Lean Wastes section identifies inefficiencies in resource utilization (the 'what' of waste). The Workflow & Flow Indicators section captures the systemic conditions that enable or impede production flow (the 'context' of production).

The Safety-Lean Dual-Impact Mapping section bridges operational efficiency with worker safety, recognizing that many visual states simultaneously indicate both waste and hazard conditions. Together, these sections provide a holistic view: waste indicators identify problems, flow indicators reveal systemic causes, and dual-impact mapping ensures that safety is not compromised in the pursuit of efficiency.

The eight wastes (Muda)

The framework translates each waste into a state-based visual definition. While Figure 3 illustrates representative mappings for primary waste categories, the framework is extensible to other wastes, such as 'Making Do' (Koskela, 2004), or emphasizing temporal and contextual parameters. An "Idle Worker" is defined as a worker stationary for >X seconds (e.g., 60 s) without tools, distinguishing brief pauses from genuine waste. "Idle Equipment" requires temporal thresholds when controls are disengaged. "Motion" waste includes "Poor Tool Staging", which is detected through repetitive movements. "Inventory" manifests as "Site Clutter" in the walkways. "Defects/Rework" appears as reversed work sequences, such as premature wall closing before MEP inspection. "Transportation" waste is captured via

"Inefficient Transport" (partial loads, empty runs). "Overproduction" appears as "Excess Material Buildup" beyond the immediate capacity.

Workflow and flow indicators

Beyond waste identification, the LCVT incorporates flow-state indicators aligned with workflow reliability, a central concept in the Last Planner System (Ballard, 2000). Ballard and Howell (1998) emphasized that a reliable workflow requires both enabling conditions and the absence of constraints. The LCVT operationalizes this by defining both enabling states (e.g., proper material staging) and constraint states (e.g., blocked access) as visual cues. "Material Staging (Good)" represents the optimal state of materials, organized within immediate reach of the active work zone, and precisely staged where needed without impeding access.

"Work-in-Progress (WIP)" provides quantifiable metrics of partially completed work within specific zones (e.g., "Rebar tying 75% complete"), enabling flow measurement. On the impediment side, "Site Congestion" is defined spatially as multiple resources occupying confined areas where movement is visibly restricted. Simultaneously, "Blocked Access" identifies physical obstructions preventing labor or material flow along critical paths. These indicators transform the abstract concept of "flow reliability" into measurable visual states.

Safety-Lean dual-impact mapping

The LCVT maps scenarios with dual impacts on safety and productivity: "Site Clutter" creates motion waste and trip/fall hazards, "Poor Material Staging" generates Motion/Waiting waste and struck-by/egress risks, and "Equipment Proximity" triggers Waiting waste and struck-by hazards. Figure 3 synthesizes all LCVT components into a unified conceptual map. Each pathway from L1 categories through L2 indicators to formal L3 visual definitions (e.g., "stationary >60 s," "materials exceeding immediate need") provides the exact specification needed for both human annotators and AI inference, moving beyond simple object detection toward state classification, which is essential for interpreting production flow.

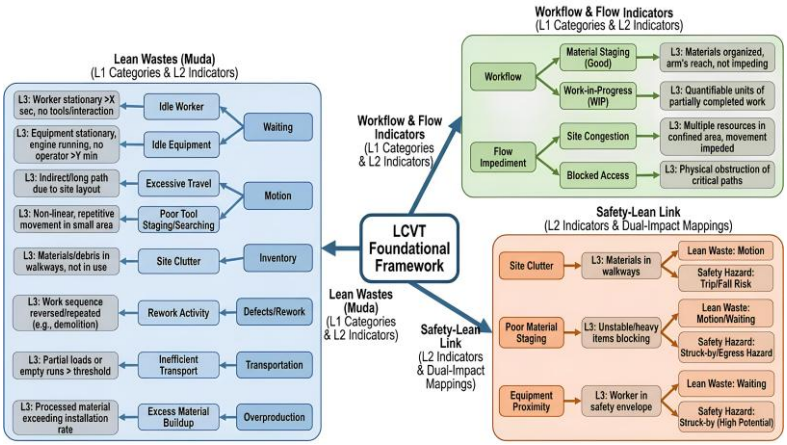

Figure 3: Comprehensive Lean Construction Visual Taxonomy (LCVT) framework

DISCUSSION: EVOLUTION AND IMPLEMENTATION

Figure 4 presents the end-to-end AI ecosystem in which the LCVT acts as the logic core, connecting raw observations to decision-making support through a six-layer pipeline that guides dataset annotation, informs state classification design, and enables domain-aware VLM prompts. The architecture unifies traditional supervised models and VLMs through three analytics engines (Waste Identification, Flow Reliability Analysis, Safety-Lean Dual-Impact Analysis), with outputs interactable with the Last Planner System for make-ready planning.

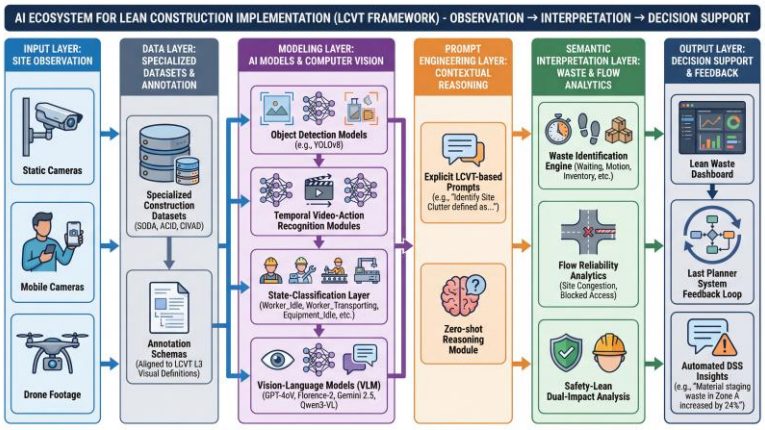

Figure 4: AI ecosystem for Lean Construction implementation using the LCVT framework

Bridging the Semantic Gap

The primary contribution of the LCVT is its function as a "Reference Framework." It provides a translation layer between two distinct domains:

- The Domain of Construction Management: Focused on abstract concepts like "Flow," "Value," and "Reliability."
- The Domain of Machine Vision: Focused on concrete data like "Bounding Boxes" and "Temporal Vectors."

By explicitly defining L3 Visual Definitions, the LCVT enables CV systems to output semantically meaningful data. Instead of a heat map of "worker locations" (observational), an LCVT-enabled system produces a heat map of "waiting hotspots" (interpretive). This direct mapping supports the Transformation-Flow-Value theory by making the "Flow" aspect measurable and visible. However, this pilot demonstration validated only a single spatial indicator with clear visual signatures. Extending to more complex indicators such as worker activity states that depend on pose interpretation, or defects and rework that require correlating images across considerable time spans, presents additional challenges, including contextual variability across activity types, the risk of misclassifying visually idle yet value-adding work (e.g., a worker inspecting or planning), and the need for multi-level qualifiers that associate workers with objects and temporal frame sequences.

Guiding "Lean-aware" Computer Vision Models

Current CV object detection models like YOLO are mostly "flat"; they only detect whether specific objects are present in a given scene. The LCVT necessitates the development of Hierarchical and Temporal Models.

- The LCVT implies that identifying an entity (worker, equipment, or material) is insufficient. Models must classify the entity's state. Our pilot validation demonstrated this for equipment states ('Dump_Truck_Loaded' vs. 'Dump_Truck_Empty'), extending this to worker states ('Worker_Idle,' 'Worker_Transporting,' 'Worker_Installing') represents a key area for future validation, requiring training data labeled not only with nouns ('Worker') but also with state-adjectives. Notably, distinguishing worker activities by pose alone presents significant challenges, as visually similar postures may correspond to different productivity states, necessitating richer contextual features.
- Temporal Logic: Many LCVT definitions (e.g., Idle Worker and Idle Equipment) rely on temporal thresholds (X seconds). This suggests that single-frame image analysis is insufficient for lean monitoring; Video Action Recognition models that analyze temporal sequences are required to distinguish between a worker "thinking" and

"waiting" (waste).

- Addressing these temporal requirements, emerging VLMs increasingly incorporate temporal reasoning capabilities that are essential for LCVT implementation.
- Models such as Qwen3-VL (Yang et al., 2025) and Gemini 2.5 (Comanici et al., 2025) demonstrate sophisticated video understanding, enabling the analysis of temporal sequences directly from LCVT-structured prompts. However, video understanding capabilities vary significantly across VLMs (Hong et al., 2025), suggesting that LCVT implementation should prioritize models with demonstrated temporal reasoning for indicators that require duration thresholds.

NOVELTY: ENABLING VISION-LANGUAGE MODELS (VLMS) VIA PROMPT ENGINEERING

The most significant forward-looking implication of the LCVT is its integration with VLMs. Models such as GPT-4o or Florence-2 have demonstrated "zero-shot" capabilities, meaning that they can identify objects they were never explicitly trained on when given a clear description. The L3 definitions of the LCVT served as the optimized prompts for these models.

- Scenario without LCVT: A user uploads an image to a VLM and asks, "Is there any waste here?" The model, which lacks domain context, may provide a generic or hallucinated response.
- Scenario with LCVT: The system injects the LCVT definition into the prompt: "Analyse this image for 'Site Clutter', defined as materials or debris in designated walkways not in active use."
- Results: The VLM leverages LCVT logic to perform domain-informed reasoning. This drastically lowers the barrier to AI adoption, as firms can use VLMs injected with LCVT prompts instead of investing in training expensive custom models.

INTEGRATION WITH DECISION SUPPORT SYSTEMS (DSS)

The LCVT can feed Decision Support Systems (DSS), such as a "Lean Waste Dashboard" (Wang et al., 2024), which aggregates L2 indicator-level waste occurrences and L1 category-level metrics (e.g., "Percentage of Non-Value-Adding Time") for cross-project benchmarking. These metrics provide objective feedback to the Last Planner System during Make-Ready planning, closing the loop between automated monitoring and human decision-making. Scalability in complex, multi-trade site environments and the integration of practitioner feedback remain important areas for future development.

CONCLUSION

The Lean Construction Visual Taxonomy (LCVT) proposes a foundational step toward bridging the "Semantic Gap" between Computer Vision technology and lean principles. By formalizing the visual signatures of waste, flow impediments, and safety interactions, the LCVT establishes an introductory data infrastructure to support AI as an active partner in production control. This framework offers a verifiable standard for researchers and a blueprint for developing semantically rich datasets required for next-generation construction AI. Additionally, encoding domain logic into structured definitions allows practitioners to utilize emerging Vision-Language Models without extensive custom training. Future research should focus on empirical validation across diverse waste categories and activity types, practitioner-based evaluation of the taxonomy's operational utility, operationalizing the LCVT through expert validation, creating large-scale benchmark video repositories, and developing a formal Lean Construction Ontology to capture complex temporal and causal dependencies.

REFERENCES

Alarcón, L. F. (Ed.). (1997). Lean Construction. CRC Press.

Asadollahi, Hamed et al. "Bridging building information systems: A parameter-efficient semantic approach to construction data interoperability." Eng. Appl. Artif. Intell. 155 (2025): 110680.

Ballard, G. (2000). *The Last Planner System of production control*. (Doctoral dissertation, University of Birmingham).

Ballard, G., & Howell, G. (1998). Shielding production: Essential step in production control. Journal of Construction Engineering and Management, 124(1), 11-17.

Barbosa, A. S., & Costa, D. B. (2021). Productivity monitoring of construction activities using digital technologies: A literature review. *Proceedings of the 29th Annual Conference of the International Group for Lean Construction (IGLC)*.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., ... & Mehta, S. V. (2025). Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next-generation agentic capabilities. arXiv preprint arXiv:2507.06261.

Duan, R., Deng, H., Tian, M., Deng, Y., & Lin, J. (2022). SODA: A large-scale open site object detection dataset for deep learning in construction. Automation in Construction, 142, 104499.

Gemma Team (2024). Gemma: Open models based on Gemini research and technology. arXiv preprint arXiv:2403.08295.

Golparvar-Fard, M., Peña-Mora, F., & 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*, 14, 129–153.

Gouett, M. C., Haas, C. T., Goodrum, P. M., & Caldas, C. H. (2011). Activity analysis for direct-work rate improvement in construction. Journal of Construction Engineering and Management, 137(12), 1117-1124.

Guarino, N. (1998). Formal ontology and information systems. Proceedings of FOIS'98, 3-15.

Haslam, R. A., Hide, S. A., Gibb, A. G. F., Gyi, D. E., Pavitt, T., Atkinson, S., & Duff, A. R. (2005). Contributing factors in construction accidents. Applied Ergonomics, 36(4), 401-415.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS Quarterly, 28(1), 75-105.

Hong, W., Cheng, Y., Yang, Z., Wang, W., Wang, L., Gu, X., ... & Tang, J. (2025). Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 8450-8460).

Kasanen, E., Lukka, K., & Siitonen, A. (1993). The constructive approach in management accounting research. Journal of Management Accounting Research, 5, 243-264.

Koskela, L.. (2000). *An exploration towards a production theory and its application to construction*. VTT Publications.

Koskela, L.. (2004). Making-Do: The Eighth Category of Waste. In Bertelsen, S., & Formoso, C. T. (Eds.), *12th Annual Conference of the International Group for Lean Construction*.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems*, 15(4), 251–266.

Ohno, T. (1988). *Toyota production system: Beyond large-scale production*. Productivity Press.

Oswald, D., Ahiaga-Dagbui, D. D., Sherratt, F., & Smith, S. D. (2020). An industry structured for unsafety? An exploration of the cost-safety conundrum in construction project delivery. *Safety science*, *122*, 104535.

Sabek, M., Gonzalez, V., Mei, Q., & Lee, G. (2024). Enhancing construction site safety and efficiency with a YOLO v8-based computer vision model. *CSCE Annual Conference, Niagara Falls*.

Sabek, M., Mei, Q., Lee, G., Golabchi, A., & Gonzalez, V. (2025). Construction Industry Vision Alberta Dataset (CIVAD). *Proceedings of the 42nd International Symposium on Automation and Robotics in Construction (ISARC)*, Montreal.

Seo, J., Han, S., Lee, S., & Kim, H. (2015). Computer vision techniques for construction safety and health monitoring. *Advanced Engineering Informatics*, 29(2), 239–251.

Shankaranarayanan, G., & Cai, Y. (2006). Supporting data quality management in decision-making. *Decision support systems*, *42*(1), 302–317.

Shingo, S. (1989). A Study of the Toyota Production System from an Industrial Engineering Viewpoint. Productivity Press.

Sveikauskas, L., Rowe, S., Mildenberger, J., Price, J., & Young, A. (2016). Productivity growth in construction. *Journal of Construction Engineering and Management*, *142*(10), 04016045.

Wang, Z., Sabek, M., Wu, Y., Mei, Q., Lee, G., & Gonzalez, V. A. (2024). Digital twin-based integrated decision support system for enhanced decision-making in the Last Planner System. *Proceedings of the 32nd Annual Conference of the International Group for Lean Construction (IGLC 32)*.

Wu, H., Zhong, B., Li, H., Love, P., Pan, X., & Zhao, N. (2021). Combining computer vision with semantic reasoning for on-site safety management in construction. *Journal of Building Engineering*, *42*, 103036.

Xiao, B., & Kang, S. C. (2021). Development of an image data set of construction machines for deep learning object detection. *Journal of Computing in Civil Engineering*, *35*(2), 05020005.

Xiao, B., Wu, H., Xu, W., Dai, X., Hu, H., Lu, Y., Zeng, M., Liu, C., & Yuan, L. (2023). Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4818-4829.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., ... & Qiu, Z. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.