Developments in the Built Environment 27 (2026) 100992

Contents lists available at ScienceDirect

Developments in the Built Environment

journal homepage: www.sciencedirect.com/journal/developments-in-the-built-environment

Enhancing building product inventory automation through a hybrid spatial-semantic modeling framework

Shuting Mi [a], Xiaofei Wu [a], Shunchang Liu [a], Zhao Huang [a], Catherine De Wolf [a], Mennatallah El-Assady [a], Pei-Yu Wu [a,b,*]

[a] ETH Zürich, Rämistrasse 101, Zürich, 8092, Switzerland
[b] Delft University of Technology, Julianalaan 134, Delft, 2628 BL, The Netherlands

ARTICLE INFO

Keywords:
Object detection
Vision-language model
Expert-in-the-loop
Building product dataset
Circular construction

ABSTRACT

Manual building inventories hinder material recovery in circular construction. While vision-language models (VLMs) offer advanced semantic reasoning, spatial grounding limitations in cluttered environments restrict automated auditing. We propose a hybrid framework decoupling spatial localization from semantic assessment via a dual-model consensus layer using weighted boxes fusion, cascading VLM routing to recover low-confidence detections, and a self-evolving active learning pipeline for expert-in-the-loop annotations. Trained and evaluated on a novel multi-modal building product dataset, the Weighted Boxes Fusion ensemble of *YOLOv12m* and *RF-DETR* ($\tau = 0.55$) outperformed standalone detectors across metrics while maintaining competitive recall. Downstream *Qwen3-VL-8B* verification with spatial overlap filtering marginally improved high-precision detections for *YOLOv12m*, though its efficacy was constrained by low-confidence noise in *RF-DETR*. Subsequently, single-exemplar ($k = 1$) VLM prompting achieved a balanced accuracy of 0.79 and an F_1 score of 0.73 for condition classification. Preliminary expert evaluations confirm the system's usability for interactive metadata refinement in unconstrained scenes. This work establishes a scalable paradigm for automated visual inventory to support digital product passports and material reuse.

1. Introduction

The increasing emphasis on circular construction has revitalized interest in the reuse of existing building products (De Wolf et al., 2024). Components that maintain structural integrity throughout a building's life cycle are now recognized as valuable resources for reconstruction and material recovery. This paradigm shift is reinforced by policy frameworks such as the EU Construction and Demolition Waste Management Protocol (European Commission, 2024) and national standards, including the Swiss SIA 430 (Swiss Society of Engineers and Architects, 2023), and DIN SPEC 91484 (DIN German Institute for Standardization, 2024). These initiatives mandate or incentivize standardized building product inventories prior to demolition or renovation, presenting both opportunities and challenges for the construction sector.

Current pre-demolition and renovation audits typically follow a two-stage workflow: an initial screening to identify reusable components, followed by detailed documentation of selected reclaimed products. This manual and labor-intensive process is difficult to scale, particularly under the increasing project volume and compressed timelines faced by industry stakeholders (Wu, 2024). Furthermore, uncertainty regarding recovery costs, compounded by non-standardized documentation practices, often leads to misclassification or disposal of valuable materials, thus undermining circular economy objectives.

Advances in deep learning-based vision recognition and reasoning offer promising pathways to automate building product inventory by generating relevant product-specific visual attributes, thereby enhancing inspection efficiency and data consistency (Keulemans and Adams, 2024; Rao et al., 2025). Nevertheless, three fundamental gaps impede the broad adoption of these approaches within the Architecture, Engineering, Construction, and Operations (AECO) sector. First, a notable scarcity of annotated multi-modal building component datasets that simultaneously capture physical geometry and material condition in situ. Second, contemporary vision-language models (VLMs) exhibit a systematic failure in precise localization and spatial grounding within cluttered indoor environments. Third, the inherent subjectivity of product condition assessment lacks a rigorous methodological framework to reconcile machine-learned probabilities with expert judgment.

* Corresponding author at: Delft University of Technology, Julianalaan 134, Delft, 2628 BL, The Netherlands.
E-mail address: p.y.wu@tudelft.nl (P.-Y. Wu).

https://doi.org/10.1016/j.dibe.2026.100992
Received 17 February 2026; Received in revised form 25 June 2026; Accepted 19 July 2026
Available online 25 July 2026
2666-1659/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Consequently, no individual model is currently sufficient to facilitate automated building inventories within complex real-world audit settings.

To address these challenges, this research proposes a hybrid methodological framework that formalizes the visual auditing process through three integrated modules. The framework begins with *spatial localization* where specialized object detection models establish precise geometric boundaries and initial classification labels within complex indoor environments. This stage is followed by a *semantic reasoning* module that utilizes VLMs to assess physical conditions and material properties through a decoupled inference logic. Finally, the framework incorporates an *expert-in-the-loop refinement* module that applies confidence-driven thresholding to reconcile automated probabilistic outputs with expert revision. This systematic structure ensures a rigorous transition from unstructured visual data to validated digital inventories of in situ building product for real-world circular construction applications.

This work contributes to the formalization and preliminary evaluation of this framework by introducing three primary validation artifacts. These comprise the system architecture integrated with a novel multi-modal building product dataset, a comprehensive benchmarking suite for evaluating specialized computer vision and vision-language models, and an interactive expert-in-the-loop user interface designed to support the generation of structured inventories. The primary methodological contribution of this work extends beyond system-level integration to formalize a dynamic information routing and self-evolving paradigm for unconstrained indoor environments. First, we introduce a dual-model spatial coordination mechanism that applies Weighted Boxes Fusion (WBF) (Solovyev et al., 2021) to ensemble predictions from distinct object detection models. Second, we formalize a decoupled processing approach that establishes operational boundaries between spatial coordinate regression and generative attribute inference, utilizing a secondary VLM routing layer to recover false negatives from low-confidence spatial detections. Third, we formalize this interaction into an active learning pipeline where expert-in-the-loop annotations dynamically expand the building taxonomic vocabulary of the model, establishing a self-evolving loop where classification performance improves over time. By applying multi-model spatial fusion and hybrid optimization in visual inventory tasks for pre-demolition audits, this approach facilitates scalable and data-rich documentation aligned with emerging digital product passports (Honic et al., 2024) and reclaimed material marketplace initiatives.

2. Relevant work

2.1. Visual recognition and reasoning in construction

Visual recognition and reasoning leverage advanced deep learning methods to interpret visual data such as objects, scenes, and their spatial relationships for enhanced decision-making and problem-solving. Traditional computer vision (CV) methods primarily tackle tasks such as object detection and image classification. Visual reasoning further extends these capabilities by incorporating VLMs that enable interpretation of complex spatial relationships, logical reasoning (Du et al., 2025), and causal inference. These algorithmic advancements, including visual question answering, scene understanding, and robotics, are deployed in complex construction environments to various extents.

2.1.1. Computer vision applications

Visual data serve as a vital information source (Kim, 2020) in construction, making CV an ideal domain application. Modern CV techniques facilitate feature extraction, semantic and instance segmentation, pose estimation, action recognition, and scene reconstruction. Notable advancements include the superior performance of vision transformers (ViTs) (Dosovitskiy et al., 2021), the adoption of self-supervised and few-shot learning techniques to alleviate annotation burdens, synthetic data generation through generative AI, and multi-modal or 3D approaches for spatial scene comprehension. Together, these developments enable real-world applications such as operational activity recognition (Li et al., 2024), real-time progress monitoring (Yang et al., 2023), automated inspection, safety monitoring, construction documentation (Paneru and Jeelani, 2021), and digital twin-based infrastructure digitalization (Lv et al., 2026).

Recent work on visual reasoning in construction focuses on geometric modeling and deep learning-driven visual representations. However, such reasoning is not yet deeply integrated into design workflows. For example, Wong et al. (2022) created a deep learning-based framework for geometric construction problems to automate technical drawing tasks and support spatial reasoning in design. Future research moves toward unifying reasoning mechanisms for the stepwise interpretation of complex scenes. In particular, advancing spatio-temporal reasoning and integrating these models with digital twins may enable real-time, multi-modal decision-making, while object-centric neural networks offer strong potential for compositional reasoning.

2.1.2. Vision-language applications

VLMs extend AI capabilities by linking visual data with semantic text through vision encoders (e.g., CNNs or ViTs) and language encoders, connected by cross-attention mechanisms. VLMs can be grouped by primary capability: vision–language understanding (VLU), text generation from multi-modal input, and multi-modal input–output systems (Ghosh et al., 2024). VLU models, including *CLIP* (Radford et al., 2021), *VLMo* (Bao et al., 2022), and *VideoMAE* (Tong et al., 2022), excel at visual interpretation of images and videos. Another category, with models such as *GPT-4V* (OpenAI, 2024), *LLaVA* (Liu et al., 2023), and *Flamingo* (Alayrac et al., 2022), specializes in text generation conditioned on visual input. Multi-modal systems, exemplified by mixture-of-experts architectures like *Gemini* (Team, 2023) and diffusion-based CoDi (Composable Diffusion) (Tong et al., 2023) and CoDi-2 (Tang et al., 2024) models, handle both multi-modal input and output tasks.

Despite their advances, current VLMs continue to struggle with spatial reasoning in complex 3D contexts. To address this, Chen et al. (2025) employed mechanistic interpretability to trace cross-layer attention, dynamically adapting context windows for improved spatial inference. Zhang et al. (2025) introduced pipelines to transform 3D scenes into annotated 2D data, supporting enhanced VLM spatial benchmarks. Hybrid approaches combining conventional CV inference with VLM generalization are increasingly viewed as promising.

Within the construction sector, VLMs have been applied to safety assurance, inspection planning, and waste management. For example, Chen (2024) tested VLMs for construction safety inspection in zero-shot modes using captioning, visual question answering, and element grounding for safety concepts. Kim et al. (2025) created a fine-tuned VLM for hazard recognition with construction-specific datasets, achieving 94.25% accuracy and outperforming GPT-4V in textual hazard reasoning. Chen and Yin (2025) introduced *ChatCH*, a tailored hazard reporting VLM, attaining a precision of 89.4% across 32 hazard categories. Sun et al. developed *ConClip*, a VLM for construction waste recognition, leveraging feature interaction networks for multi-modal performance (Sun et al., 2024). Most recently, Sun et al. (2025) demonstrated a training-free VLM pipeline to generate 3D inspection trajectories for infrastructure robotics, utilizing semantic and spatial alignment from multi-view images for efficient inspection path planning. VLMs have also been applied to historical building digitization, where VLM-driven few-shot segmentation combined with language-embedded 3D Gaussian splatting enables realistic and interactively searchable 3D scene models (Liang et al., 2026).

2.2. Interactive machine learning

Interactive machine learning (IML) represents an emerging paradigm in which human expertise is directly integrated into the training and refinement of machine learning models through iterative feedback loops (Mosqueira-Rey et al., 2023). IML prioritizes human-centered interaction, enabling users to actively label data, correct outputs, rank results, and guide the model's learning process (Wondimu et al., 2022). The principal objective is to fuse human judgment and ML automation, resulting in interpretable, adaptive, and context-aware systems that more closely reflect domain-specific requirements.

In the construction domain, the adoption of IML methodologies remains nascent. For instance, Saka et al. (2024) have examined explainability and expert interactivity in predicting the circularity of demolition waste, but their approach lacks a genuine bi-directional feedback mechanism for model retraining based on human input. There is substantial potential for expanding IML across multiple stages of the construction informatics pipeline. Human-in-the-loop annotation can markedly improve both the quality and the diversity of labeled datasets, helping to overcome persistent limitations in data availability and representativeness for construction machine learning applications.

IML interfaces supports interactive visual analytics for construction monitoring, safety inspection, and progress tracking by enabling real-time validation and iterative modification of model outputs (Zhang et al., 2021). Adaptive guidance mechanisms are another promising direction, empowering models to prioritize domain-specific cues and continuously refine predictions in dynamic environments (Holter and El-Assady, 2024). Moreover, integrating multi-modal inputs, such as visual, textual, and sensor data, with expert feedback throughout training and validation could substantially enhance model performance in downstream tasks, including hazard detection and quality assurance (Wu et al., 2022)

3. Data

3.1. Multi-modal dataset curation

In the absence of in situ building product datasets, a custom multi-modal dataset was curated from diverse empirical sources. These included egocentric video recordings from university buildings, empirical pre-demolition audit inventories, and images from online material marketplace platforms. The compiled dataset integrates visual imagery, textual descriptions, and tabular metadata spanning both assembled and disassembled product contexts to ensure a comprehensive representation of building components.

The building product image dataset comprises approximately 900 unique images, each annotated with multi-label building product classification. Primary sources comprise video frames captured via Project Aria smart glasses during institutional building walkthrough (Engel et al., 2023), supplemented with building product listing images from Swiss material resale platforms and product catalogs from Swiss pre-demolition audit inventories. All images underwent automated parsing through a custom script, followed by manual annotation with bounding boxes. Classification adhered to the eBKP-H code (element-based construction cost plan for building construction) (Swiss Society of Engineers and Architects, 2020) to standardize product categories, resulting in 37 defined classes. For the Aria glass footage, bounding box annotation was performed by four co-authors using Roboflow (Ciaglia et al., 2022), with each annotation cross-reviewed by the remaining annotators. Images from the online platforms and pre-demolition inventories were sourced with existing category labels, which were mapped to the 37 eBKP-H classes. Bounding boxes were required to tightly enclose the visible extent of each object, and inter-annotator disagreements were minimal given the visual unambiguity of the 37 common building product classes, resolved through discussion during cross-review.

The dataset comprises 3999 annotated object instances distributed across 37 classes. As illustrated in Fig. 1, the class distribution exhibits a pronounced long-tail characteristic where ceiling light represents the majority with 1366 instances (34.2%). Conversely, marginal categories such as letterbox, bike stand, bench, and bathtub are represented by only a single instance each. This distribution skew originates from two primary factors. The first factor relates to the naturalistic occurrence of objects in university environments where ubiquitous elements such as lighting and doors appear with high frequency while specialized furniture or appliances remain sparse. The second factor involves the integration of online material marketplace platform listings which introduced several unique categories as isolated examples not encountered during physical site walkthroughs. Such long-tail distributions represent a prevalent challenge in scene-specific detection tasks and are consistent with established benchmarks such as LVIS (Gupta et al., 2019).

For condition classification, a multi-modal building product condition subset of 155 cropped images was extracted from material inventories. The subset originates from real-world pre-demolition audit records provided by Swiss environmental services, consulting firms, and associations promoting material reuse, all of which are members of Cirkla, an umbrella organization and network for building component reuse in the Swiss construction industry. Each inventory sheet contains structured fields covering element designation, eBKP-H classification code, photographic documentation, room and location, physical dimensions, quantity, condition rating, dismantlability, and reuse potential. The photographic documentation and condition rating are the fields directly used for the condition classification task, while the element designation and eBKP-H classification code provided the object category labels for the detection dataset. Condition labels were assigned by certified pre-demolition auditors with at least five years of field experience. Each sample pairs a cropped product image with descriptive text. Items are categorized into three condition classes: good, average, and poor, as illustrated in Fig. 2(a). Labels were transformed into one-hot encoded vectors, designating the dominant condition class with a value of 1 and the remainder as 0. The dataset exhibits class imbalance, with 110 samples labeled good, 35 average, and 10 poor, as shown in Fig. 2(b). This imbalance was explicitly addressed during model training and evaluation to mitigate bias.

3.2. Data pre-processing

The dataset was randomly partitioned into 720 training images, 90 validation images, and 90 test images following an 80/10/10 ratio. To enhance model robustness and mitigate overfitting in the object detection task, data augmentation was applied exclusively to the training set using Roboflow (Ciaglia et al., 2022). Each original image was expanded into three augmented variants through a combination of transformations, including hue adjustments ranging between −15° and +15°, saturation modulation from −25% to +25%, brightness variation between −15% to +15%, Gaussian blur with a maximum radius of one pixel, and introduction of random noise affecting up to 0.1% of pixels.

Following augmentation, the training set increased to 2160 samples, yielding a total of 2340 images across the full dataset, with validation and test sets maintained as originally allocated. Subsequent standardization was performed via an automated pre-processing pipeline, which included orientation correction, resizing all images to a uniform resolution of 640 × 640 pixels, and contrast enhancement utilizing adaptive histogram equalization to improve feature delineation and consistency in feature extraction. For the condition classification task, the multi-modal image–condition pairing dataset underwent analogous pre-processing. Given the limited sample size, a stratified three-fold cross-validation scheme was implemented, wherein each fold served once as the test set while the remaining folds composed the training set. This procedure enabled comprehensive and balanced model evaluation using all available labeled data.

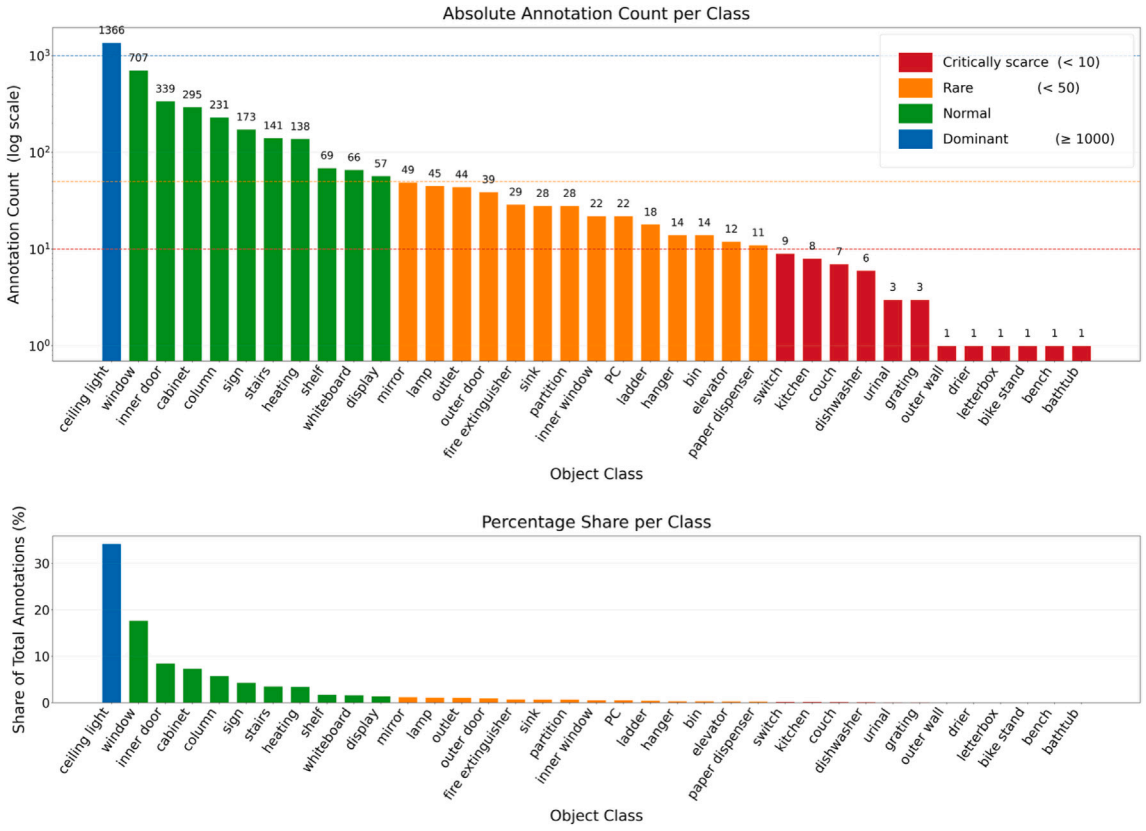
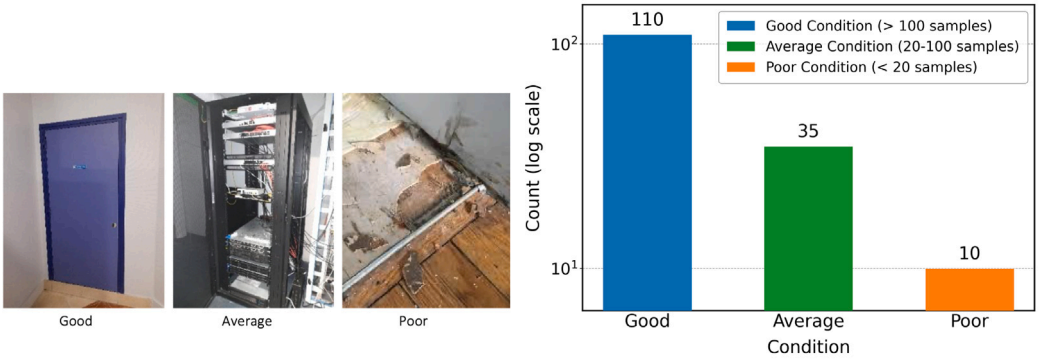


Fig. 1. Per-class instance count and share across the full dataset. Classes are sorted in descending order of instance frequency, illustrating the long-tail distribution.



(a) Material condition examples. (b) Condition label distribution.

Fig. 2. Condition distribution and examples of the labeled building product image and condition dataset.

4. Method

This research presents a hybrid framework for object-centric building product reuse inspection that bridges the semantic reasoning of VLMs with the spatial precision of task-specific vision models, as depicted in Fig. 3.

We formalize this framework across three sequential modules to represent the decoupled spatial grounding and semantic inference architecture as a generalizable approach. Let an indoor scene be represented by an input image $I \in \mathbb{R}^{H\times W\times 3}$. Module 1 addresses spatial localization through a trained object detector f_{det} that maps I to a set of N localized components:

$$D = f_{det}(I) = \{d_1, \ldots, d_N\} \tag{1}$$

where each detection $d_i = (b_i, c_i, p_i)$ comprises bounding box coordinates b_i, an eBKP-H class label $c_i \in C$, and a confidence score $p_i \in [0, 1]$.

Module 2 executes semantic reasoning by decoupling localization from assessment. A cropping function ϕ first extracts the object patch $I_i = \phi(I, b_i)$, which a VLM (f_{vlm}) then processes conditioned on a k-shot textual prompt P_k:

$$R_i = f_{vlm}(I_i, P_k) = (y_i, m_i, t_i) \tag{2}$$

In this formulation, R_i captures the semantic attributes of the component including the physical condition $y_i \in \{Good, Average, Poor\}$, the predicted material m_i, and the descriptive context t_i.

The final module incorporates expert-in-the-loop refinement to validate the automated outputs. The initial automated state is defined as $S_{auto} = \{(d_i, R_i)\}_{i=1}^{N}$. To mitigate the subjectivity of condition assessments, the interactive dashboard introduces an expert correction

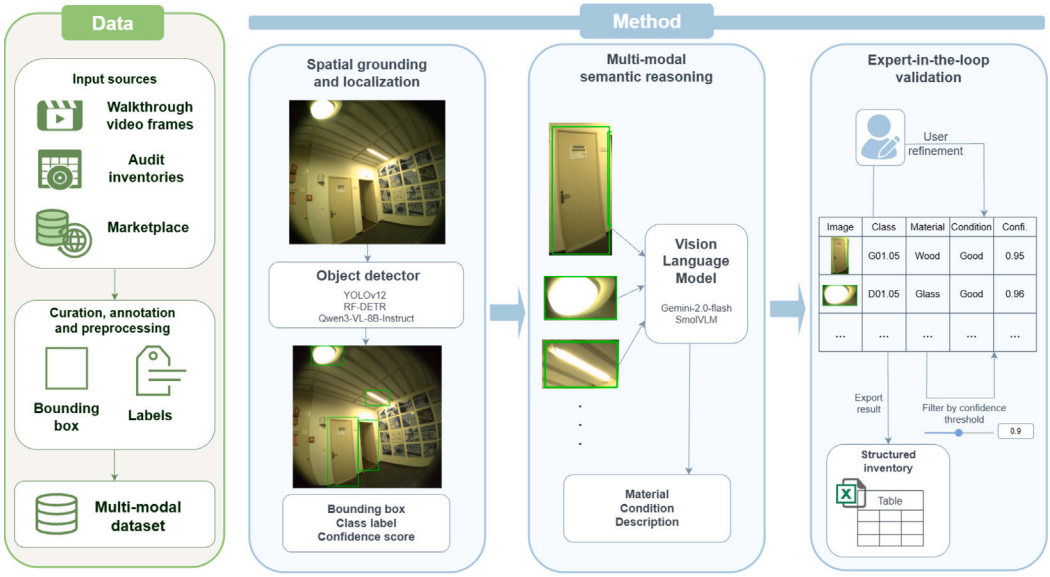


Fig. 3. Schematic representation of the hybrid framework ensembling object detection models with VLMs, coupled with a human-in-the-loop interface for expert evaluation.

function H_{expert} governed by a user-defined confidence threshold τ:

$$S_{final} = H_{expert}(S_{auto}, \tau) \tag{3}$$

Instances where $p_i < \tau$ are flagged for manual review. H_{expert} formalizes the human-AI synergy by integrating manual label corrections and bounding box adjustments to yield the validated inventory S_{final}.

4.1. Module 1: Spatial grounding and localization

Due to the high quality but limited scale of the custom datasets, the *YOLOv12m* (Tian et al., 2025), *RF-DETR* (Robinson et al., 2025), and *Qwen3-VL-8B-Instruct* (Qwen Team, 2025) models were fine-tuned on the annotated building component dataset to specialize them for domain-specific features. To determine the optimal configuration for building product detection and condition assessment, a spectrum of specialized object detection models and VLMs was systematically evaluated. Selection criteria focused on the balance between localization accuracy, computational efficiency, and adaptability to complex indoor built environments.

The evaluation utilized several foundational datasets and models. The YOLO (You Only Look Once) series was selected to represent single-stage detection strategies, which frame object detection as a direct regression problem from pixels to bounding box coordinates and class probabilities. These variants leverage CNN-based backbones and feature aggregation modules for multi-scale fusion. In contrast, the transformer-based *RF-DETR* was included for its anchor-free architecture and *DINOv2* (Oquab et al., 2024) backbone, which facilitate superior spatial reasoning and contextual understanding without reliance on non-maximum suppression. The specific model chosen for this specific object detection task include *YOLOv8m* benchmarked on COCO (80 classes) and OpenImages v7 (600 classes), *YOLOv11m*, *YOLOv12m*, *RF-DETR*, *Gemini-2.0-Flash*, *Qwen3-VL-8B-Instruct*.

To ensure comparative rigor, the VLMs were evaluated using distinct implementation strategies. *Gemini-2.0-Flash* was deployed in a zero-shot configuration utilizing a specialized prompting strategy to enforce structured JSON outputs. To accommodate its native spatial encoding, normalized coordinates were requested $[y_{min}, x_{min}, y_{max}, x_{max}]$, subsequently reordered to the standard format $[x_{min}, y_{min}, x_{max}, y_{max}]$ and rescaled normalized values (0–1000) to absolute pixel values. *Qwen3-VL-8B-Instruct* was hosted locally via the Ollama framework (Ollama Team, 2024) and adapted through a parameter-efficient QLoRA regime (Dettmers et al., 2023). This unified benchmarking framework facilitates a direct quantitative comparison between broad semantic generative models and task-specific detection architectures. Five standardized metrics were employed: precision, recall, mAP@0.5, mAP@0.5:0.95, and F1 score. Since *RF-DETR* outputs metrics differently than *YOLO* models, all metrics were recalculated post-inference through a standardized implementation, ensuring comparability.

4.2. Module 2: Multi-modal semantic reasoning

Following the spatial grounding stage, identified objects were processed to generate material descriptions and condition classifications. Four distinct architectures were selected to represent key methodological paradigms: *ResNet-18* as a CNN-based supervised baseline, *ViT (base-16)* as its transformer-based counterpart, *SmolVLM* as a lightweight open-source VLM amenable to parameter-efficient fine-tuning, and *Gemini-2.0-Flash* as a state-of-the-art proprietary multimodal model. This selection spans the full spectrum from conventional vision classifiers to instruction-following VLMs, enabling a principled evaluation of whether semantic reasoning provides measurable advantages over spatial feature extraction for condition assessment.

Given restricted fine-tuning access for *Gemini-2.0-Flash* (Team, 2023), the model was evaluated in a few-shot, prompt-based configuration to enforce label consistency. Demonstrations were selected with balanced class representation: k exemplars were drawn uniformly from each of the three condition classes (e.g., good, average, poor) using a fixed random seed, yielding $3k$ in-context examples per query. The classification prompt template was held constant across all k evaluations to isolate the effect of demonstration count. *SmolVLM* was adapted via the QLoRA method to optimize parameter efficiency and trained over six epochs. To minimize the computational footprint, 8-bit optimizers and gradient checkpointing were implemented. Condition classification performance was assessed via balanced accuracy and macro-F1 score metrics, selected to contend with dataset imbalance.

4.3. Module 3: Expert-in-the-loop validation and system integration

The proposed hybrid framework was integrated into a web-based interface to evaluate its practical relevance and functional utility. An interactive dashboard was implemented using the *Streamlit* framework for real-time visualization of model outputs. Designed to be

Table 1
Overview of survey participants.

No	Professional profile	Affiliation	Reuse expertise	Years
1	Project engineer	Engineering consultancy	Health and safety in demolition	20
2	Reuse consultant	Engineering consultancy	Circular construction	8
3	Material engineering	Certification company	Material engineering	20
4	Founder	Construction startup	Circular construction	15
5	Reuse consultant	Reuse association	Deconstruction and safety	25
6	Reuse consultant	Reuse association	Reclaimed material redistribution	4
7	Marketplace engineer	Reuse marketplace	Reclaimed material redistribution	5
8	Reuse educator	Upcycling and marketplace	Deconstruction and reuse	5

Table 2
Performance comparison of visual grounding across individual zero-shot, fine-tuned, and ensembled models using Weighted Bounding boxes Fusion (WBF). Bold values denote optimal performance per metric per category.

Model	Precision	Recall	mAP@0.5	mAP@0.5:0.95	F1
Gemini-2.0-Flash	0.3438	0.3005	–	–	0.3207
Qwen3-VL-8B	0.2254	0.2814	–	–	0.2503
Fine-tuned Qwen3-VL-8B	0.4647	0.6475	–	–	0.5411
Fine-tuned YOLOv8m (COCO)	0.7063	0.7295	0.7537	0.4695	0.7177
Fine-tuned YOLOv8m (OIv7)	0.7268	0.7486	0.7518	**0.4869**	0.7375
Fine-tuned YOLOv11m (COCO)	0.7378	0.7459	0.7693	0.4828	0.7418
Fine-tuned YOLOv12m (COCO)	**0.7500**	0.7705	0.7694	0.4791	0.7601
Fine-tuned RF-DETR (COCO)	0.7461	**0.7787**	**0.7782**	0.4747	**0.7621**
WBF (Fine-tuned YOLOv12m+RF-DETR)	**0.8091**	0.7760	**0.7964**	**0.4905**	**0.7922**

platform-agnostic, the dashboard offers core functionalities including input upload, model selection, output verification, and manual revision. An expert evaluation study was conducted during September 2025 in collaboration with reuse actors from the Swiss national reuse association Cirkla to evaluate the framework. To ensure rigorous validation, selection criteria were applied to recruit participants with expertise spanning the circular construction value chain, specifically across architecture, consulting, planning, implementation, and distribution. A cohort of eight domain experts was assembled, ensuring that no participant repeated the session or possessed prior familiarity with the interface. As summarized in Table 1, the panel represented a diverse range of professional profiles and institutional experience, with industrial tenure varying from early-career practitioners to senior experts.

The validation procedure followed a standardized physical user study protocol where each individual session lasted approximately 30 min. The session began with a brief 5-minute introduction to the task and interface features. Participants were then asked to perform a standardized indoor environment inventory task consisting of three sequential activities. First, the participant uploaded a indoor environment image and selected an object recognition model alongside a preferred VLM architecture. Second, the participant manipulated a confidence threshold slider to filter low-quality or false-positive object detections. Third, the participant reviewed the generated inference results and executed manual post-edits to correct misaligned bounding boxes, adjust product labels, or update material conditions.

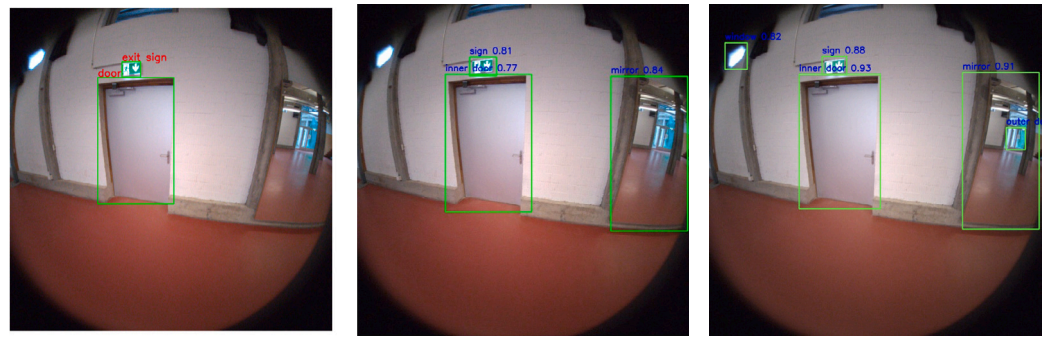
A total of eight unique cases representing complex indoor environments were reviewed across the sessions. Following the completion of the interactive workflow, feedback was gathered through a digital survey containing structured questionnaire in Appendix. This instrument evaluated three dimensions including perceived usefulness, annotation accuracy, and the trustworthiness of prediction outputs. The qualitative and quantitative responses were subsequently aggregated and analyzed using thematic coding to evaluate human-AI synergy and identify system bottlenecks. Because no digital tools currently exist for automated visual inspection and material inventory procedures are manual, no baseline comparative work was conducted. Consequently, this preliminary evaluation serves as an empirical verification of system usability and a proof of concept for the proposed workflow.

5. Results

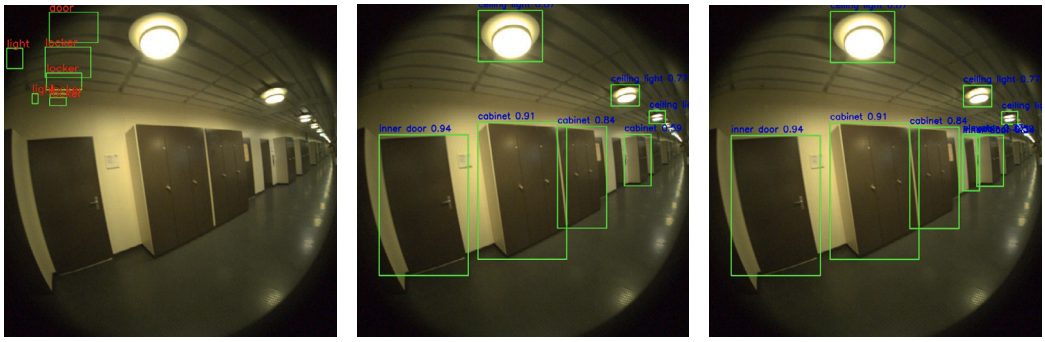
5.1. Evaluation of spatial localization and detection accuracy

The visual comparison of detection outputs across *Gemini-2.0-Flash*, *YOL Ov12m*, and *RF-DETR* is illustrated in Fig. 4, with *Qwen3-VL-8B-Instruct* shown in Fig. 5. While task-specific detectors yielded consistently well-aligned results, *Gemini-2.0-Flash* exhibited pronounced localization errors primarily due to a failure to adhere to specified coordinate constraints. This discrepancy is evident in Figure 4(d), which captures the raw, misaligned output generated by the model's native spatial encoding. Following post-inference coordinate reordering and scale normalization, the detections in Figure 4(a) demonstrate improved spatial alignment, though they remain significantly less precise than those of the specialized architectures. In contrast, the fine-tuned *Qwen3-VL-8B-Instruct* exhibited enhanced localization and a stronger adherence to domain-specific building component categories compared to the zero-shot baselines. These results highlight a pronounced performance hierarchy: *RF-DETR* attained the highest recall per scene, whereas *YOLOv12m* demonstrated superior spatial localization, underscoring the inherent advantages of task-specific architectures over zero-shot VLMs for precise building component detection.

As detailed in Table 2, specialized detection models significantly outperform VLMs across primary metrics, particularly in localization accuracy. Notably, the autoregressive nature of VLM outputs precludes the generation of calibrated confidence scores requisite for ranking-based precision–recall integration. Consequently, mAP values were omitted for these models, and comparative performance was assessed via precision, recall, and F1 scores. Among fine-tuned models, *YOLOv12m* achieved the highest precision (0.7500), reflecting a conservative detection strategy that minimizes false positives. In contrast, the transformer-based *RF-DETR* attained the highest recall (0.7787), mAP@0.5 (0.7782), and F1 score (0.7621), demonstrating superior robustness in identifying occluded or visually subtle instances. This divergence underscores a fundamental architectural trade-off: *YOLOv12m* prioritizes spatial localization precision, whereas *RF-DETR* emphasizes comprehensive object coverage. The fine-tuned *Qwen3-VL-8B-Instruct* exhibited marked improvement over its pre-trained baseline and surpassed *Gemini-2.0-Flash* in domain-specific accuracy. However, detector-based methods maintain a distinct performance lead on the validation set, likely due to their specialized spatial regression layers.



(a) *Gemini-2.0-Flash* scene 1 (b) *RF-DETR* scene 1 (c) *YOLOv12m* scene 1



(d) *Gemini-2.0-Flash* scene 2 (e) *RF-DETR* scene 2 (f) *YOLOv12m* scene 2

Fig. 4. Comparison of object detection performance across representative indoor scenes. Panel (d) illustrates the raw detection output from *Gemini-2.0-Flash* prior to coordinate reordering and scaling post-processing.

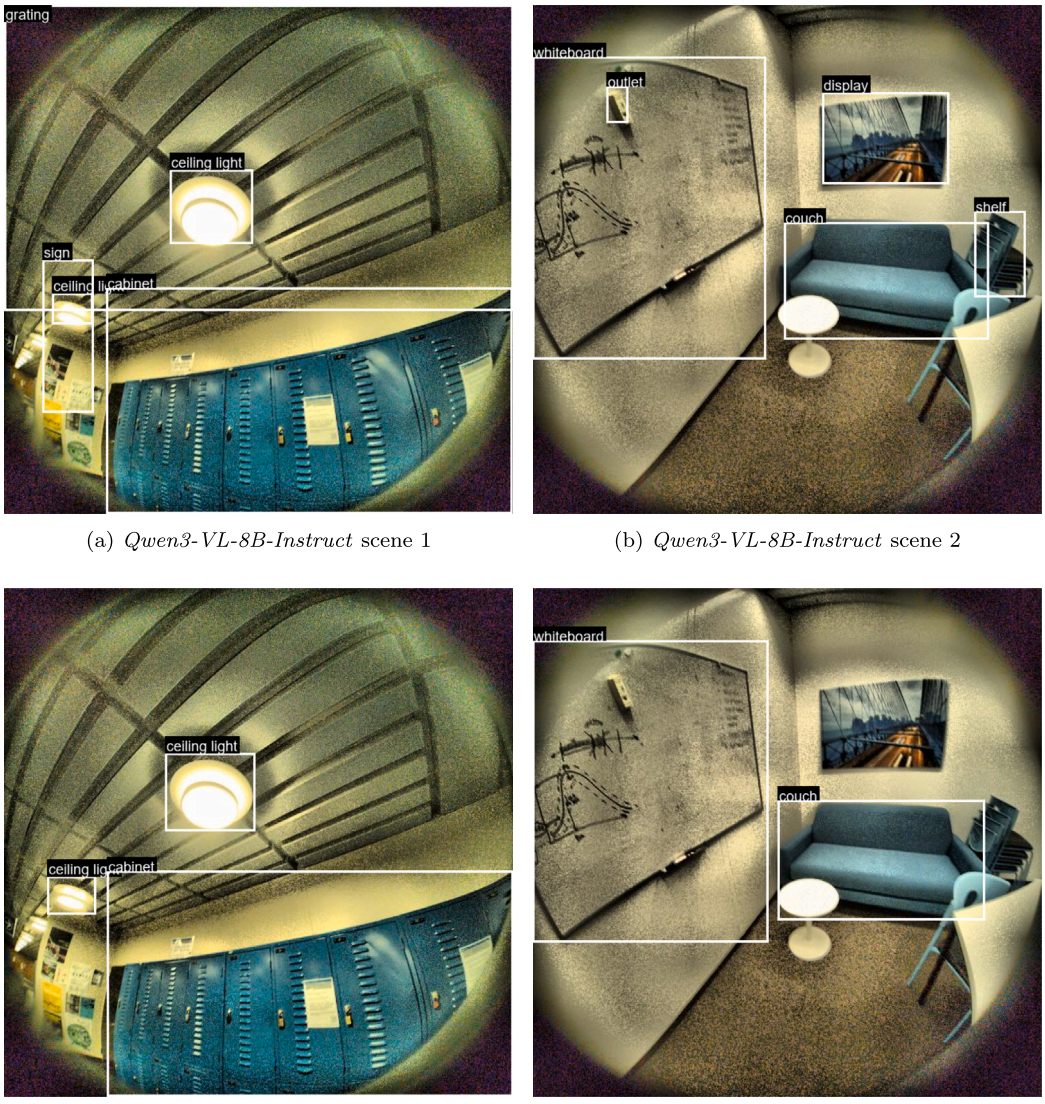
In terms of mAP@0.5:0.95, the *YOLOv8m* model fine-tuned on the Open Images v7 dataset achieved the highest score (0.4869), although the margin over other YOLO variants and *RF-DETR* was minor. The difference in F1 score between *YOLOv12m* and *RF-DETR* remained minimal (0.7601 vs. 0.7621), indicating comparable overall effectiveness. Given these complementary strengths, both *YOLOv12m* and *RF-DETR* were incorporated into the web interface. This dual-model configuration allows experts to alternate between precision-oriented and recall-driven detection modes, accommodating diverse requirements in visual inventory assessment workflows.

The confusion matrices in Fig. 6 reveal a consistent failure pattern across both *YOLOv12m* and *RF-DETR* in per-class analysis. Three classes, *bin*, *inner window*, and *paper dispenser*, yield zero correct detections, with all corresponding instances misattribute to the background. This failure is likely driven by severe class imbalance: each of these classes constitutes fewer than 25 training samples (0.28–0.55% of the dataset). The fact that both architectures fail identically suggests data scarcity is more likely the limiting factor than the models themselves. Although *RF-DETR* achieves higher overall recall, its confusion matrix (Fig. 6(a)) exhibits more off-diagonal activations than that of *YOLOv12m* (Fig. 6(b)), indicating a higher rate of inter-class confusion. A representative example is the *switch* and *outlet* pair, where *RF-DETR* misclassifies 33% of *switch* instances as *outlet*, an error far less prevalent in *YOLOv12m*. This suggests that *RF-DETR*'s higher recall comes partly at the cost of semantic precision, particularly among visually similar categories. *YOLOv12m* adopts a more conservative detection strategy, resulting in fewer inter-class errors at the expense of some missed detections.

Given this precision–recall trade-off between the two fine-tuned detectors, we tested whether combining their outputs into a single consensus prediction could improve on both and mitigate false-positive detections in cluttered indoor scenes. We further implemented a dual-model consensus protocol using Weighted Boxes Fusion (WBF) (Solovyev et al., 2021) and reported the results in Table 2. Restricting final bounding box proposals to regions of cross-model agreement yielded consistent gains across both frame-level splits: on the validation set, WBF improved the F_1 score to 0.792, up from 0.760 (*YOLOv12m*) and 0.762 (*RF-DETR*). On the independent test set, the fusion architecture achieved an F_1 score of 0.785, clearly ahead of *YOLOv12m* (0.769) and marginally ahead of *RF-DETR* (0.784).

This gain follows directly from how boxes are fused. Both detectors run down to a low confidence floor ($c \geq 0.1$), and predictions are grouped into clusters by IoU overlap ($\tau_{\text{iou}} = 0.5$). Each fused box takes the average confidence of its cluster, $\hat{c} = \frac{1}{2}\sum_{m=1}^{2} c_m$, so a box seen by only one detector is capped at $\hat{c} \leq 0.5$, while one confirmed by both can reach $\hat{c} = 1.0$. We keep a fused box only if $\hat{c} \geq \tau^*$, chosen by sweeping the validation set in steps of 0.025; F_1 peaks between $\tau \approx 0.50$ and 0.60 (Fig. 7(a)), maximizing at 0.792 when $\tau^* = 0.55$. Since this exceeds the 0.5 ceiling a single detector can reach, every surviving box must come from *both* models, making the threshold a consensus gate rather than an arbitrary cutoff. This corresponds to a real drop in false positives (Fig. 7(b)): 94 (*YOLOv12m*) and 97 (*RF-DETR*) down to 67.

In parallel, we evaluated a complementary VLM verification step: passing low-confidence detector proposals to a downstream vision-language model (*Qwen3-VL-8B*) to recover missed detections in Table 3. Detections with confidence ≥ 0.5 are accepted automatically, and detections with confidence < 0.1 are dropped. For detections in the $[0.1, 0.5)$ band, we first check IoU overlap against already-accepted boxes ($\tau_{\text{iou}} = 0.5$); a box that duplicates an accepted detection is dropped without a model call. The remaining ambiguous boxes are cropped and sent to *Qwen3-VL-8B-Instruct* with the detector's proposed label and confidence, prompted to return a revised class and a calibrated confidence level (high, medium, or low). A detection is kept only if the response is high confidence with a valid class.



(a) *Qwen3-VL-8B-Instruct* scene 1 (b) *Qwen3-VL-8B-Instruct* scene 2

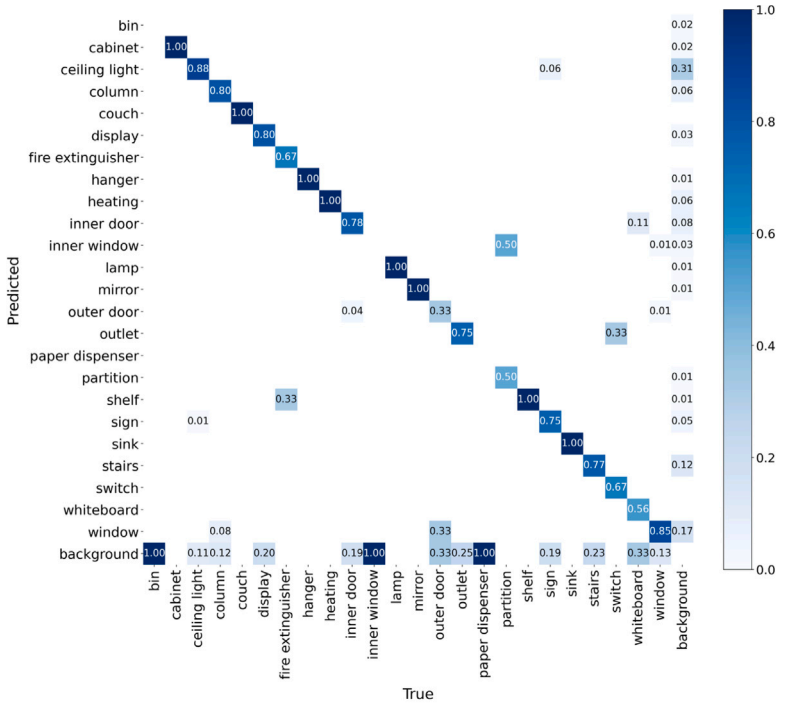
(c) Fine-tuned *Qwen3-VL-8B-Instruct* scene 1 (d) Fine-tuned *Qwen3-VL-8B-Instruct* scene 2

Fig. 5. Visual assessment of VLM object detection performance, highlighting the transition from the zero-shot baseline (a, b) to the fine-tuned architecture (c, d).

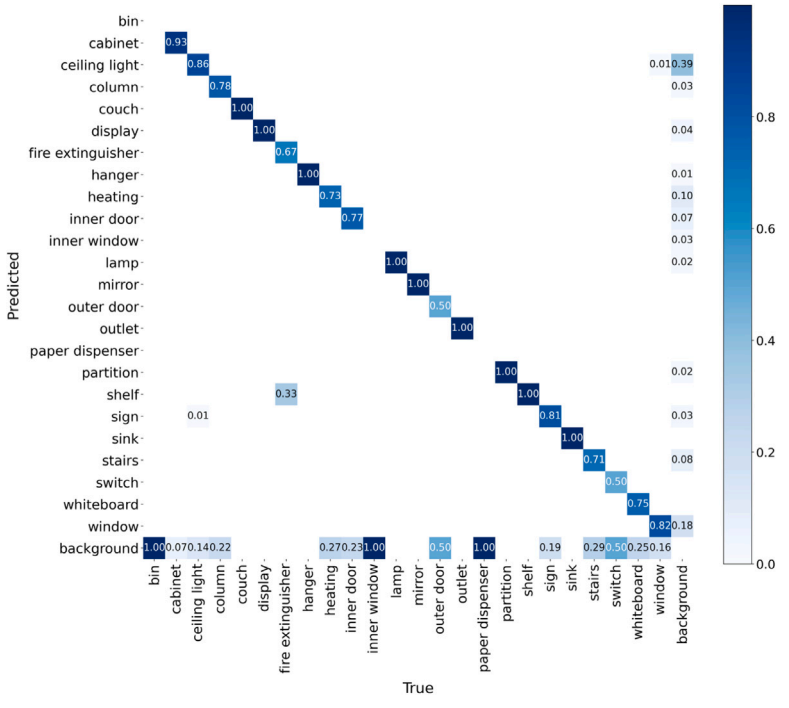
Table 3
Effect of Qwen3-VL verification and overlap filtering on detection quality for each base detector.

(a) YOLO + Qwen3-VL							
Method	Precision	Recall	F1	TP	FP	FN	ΔF_1
YOLOv12m baseline (conf≥0.5)	0.7500	0.7705	0.7601	282	94	84	–
+ Qwen verifier (no overlap filter)	0.7513	0.7760	0.7634	284	94	82	+0.0033
+ Overlap filter + Qwen verifier	0.7513	0.7760	0.7634	284	94	82	+0.0033
(b) RF-DETR + Qwen3-VL							
Method	Precision	Recall	F1	TP	FP	FN	ΔF_1
RF-DETR baseline (conf≥0.5)	0.7461	0.7787	0.7621	285	97	81	–
+ Qwen verifier (no overlap filter)	0.7193	0.7842	0.7503	287	112	79	−0.0118
+ Overlap filter + Qwen verifier	0.7359	0.7842	0.7593	287	103	79	−0.0028

The efficacy of the downstream verification is governed by the precision density of the detector's low-confidence proposals. For *YOLOv12m*, *Qwen* recovers two false negatives without degrading precision, raising F_1 from 0.7601 to 0.7634 (+0.33%); the overlap filter yields negligible variance because the mid-confidence band contains minimal background noise. Conversely, applying the verifier to *RF-DETR* yields a net reduction in performance. Although it recovers the identical two true positives, the high volume of low-confidence noise generated by *RF-DETR* causes *Qwen* to introduce more false positives than it removes. The overlap filter substantially limits this damage, cutting the F_1 loss from −1.18% to −0.28%, a roughly fourfold improvement, while simultaneously decreasing computational overhead by filtering duplicates prior to *Qwen* inference. Ultimately, VLM verification enhances performance exclusively when a detector's low-confidence



(a) *RF-DETR*



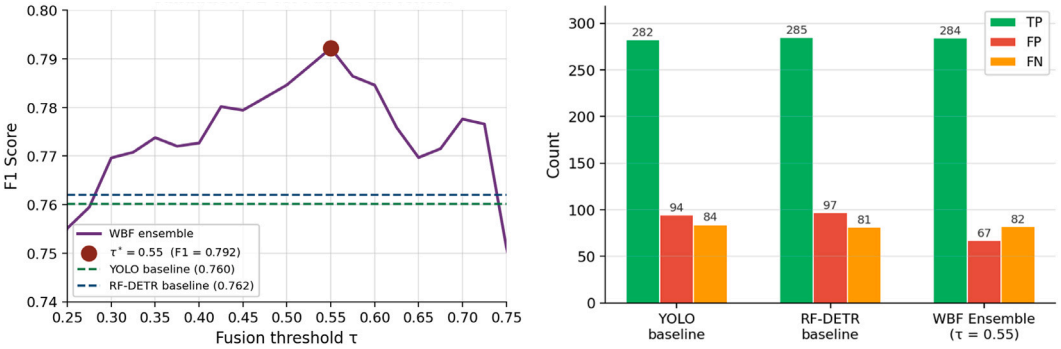
(b) *YOLOv12m*

Fig. 6. Normalized confusion matrices for the fine-tuned detection models. The *background* column captures missed detections, while off-diagonal entries indicate inter-class confusions.

proposals represent genuine missed targets, whereas the spatial overlap filter serves as a critical constraint layer when processing noisy proposal pools.

5.2. Benchmark analysis of semantic reasoning and condition assessment

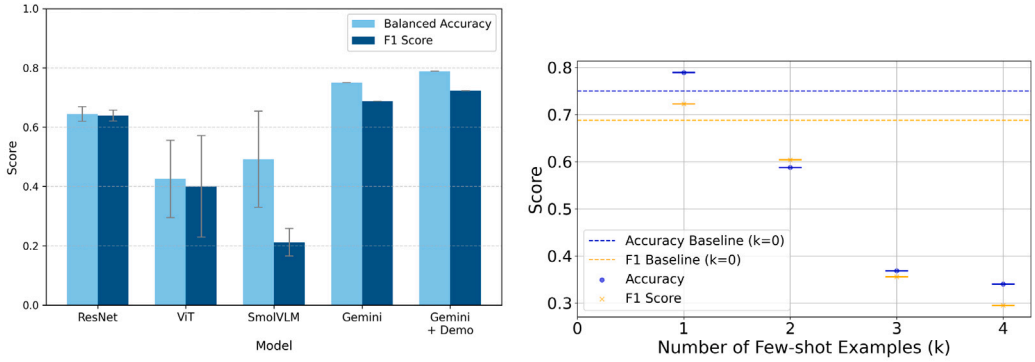
Fig. 8(a) summarizes the performance of the evaluated models on the condition classification task. Three-fold cross-validation was applied to all models with trainable parameters including *ResNet*, *ViT*, and *SmolVLM*. *Gemini-2.0-Flash* was excluded from this procedure as a closed-source inference-only model and was instead evaluated on the full labeled dataset. Overall, *Gemini-2.0-Flash* with few-shot exemplars achieves the superior performance with a balanced accuracy of 0.79 and an F_1 score of 0.72. This is followed by the zero-shot configuration of *Gemini-2.0-Flash* which yielded a balanced accuracy of 0.75 and an F_1 score of 0.69. The trained classifiers showed lower and more variable



(a) F_1 vs. fusion threshold τ (shown for $\tau \in [0.25, 0.75]$; full sweep covers $[0.05, 0.95]$), peaking at $\tau^* = 0.55$ with both single-model baselines (dashed) exceeded across most of the range.

(b) TP/FP/FN at IoU@0.5 per method.

Fig. 7. Weighted Boxes Fusion of YOLOv12m and RF-DETR on the validation set.



(a) Performance comparison among models based on balanced accuracy and F1 score.

(b) *Gemini-2.0-Flash*'s classification performance over few-shot example count (k).

Fig. 8. Condition classification performance among CV models and VLMs.

results across folds. Specifically, the F_1 score for *SmolVLM* remained notably low at 0.58 despite fine-tuning, reflecting a high sensitivity to class imbalance within the multi-modal dataset.

To explore these dynamics in greater detail, Fig. 8(b) examines how classification performance varies with the number of few-shot examples from $k = 1$ to $k = 4$. The results display a non-monotonic response where the highest accuracy and F_1 scores are achieved with a single demonstration ($k = 1$), which exceeds the zero-shot baseline. However, performance gradually declines as k increases further. At the $k = 4$ threshold, both metrics decrease markedly. This trend suggests that excessive demonstrations may introduce context dilution or reduce the salience of relevant semantic cues within the prompt window, likely caused by input sequence saturation.

5.3. Expert-in-the-loop functional usability assessment

To evaluate the perceived usability and collect preliminary feedback on the proposed workflow, we conducted an expert walkthrough with domain specialists in circular construction and building auditing. The visual inference interface shown in Fig. 9, which displays localized imagery, automated counts, and semantic attributes including material and condition, was generally well received. Participants observed that the prototype framework demonstrates functional utility as a preliminary screening tool for cataloging in-situ building products.

Compared to traditional manual workflows based on fragmented photographic logs, participants noted that automated localization and metadata generation have the potential to reduce repetitive data entry by shifting the user role toward expert oversight. Particularly, the automated counting of identical components within a single frame was perceived as beneficial. Participants suggested that extending this unique counting capability to continuous video streams would be necessary to integrate the tool into practical pre-renovation or pre-demolition audit workflows.

Furthermore, the adjustable confidence threshold τ allowed participants to filter uncertain predictions and interactively modify incorrect instances. This interactive capability was associated with a high subjective satisfaction score, yielding an average Likert rating of 4.5 out of 5 regarding user engagement. This flexible threshold adjustment could better accommodate open-vocabulary challenges in heterogeneous indoor scenes containing unfamiliar objects. Participants also indicated that the ability to manually correct condition assessments, which were noted to be less consistent due to variable image quality, is essential for maintaining the integrity of digital building logbooks.

Despite the perceived initial utility, participants highlighted several areas for future improvement beyond visual assessment. The current system remains limited to surface-level visual features, which excludes hidden structural components that often hold the highest reuse potential and carbon-saving impact. Furthermore, the identification of high-granularity attributes, including specific product brands, model numbers, and disassembly potential, were noted as essential variables for practical reuse operations. While these preliminary findings reflect the conceptual viability of the framework, future work should conduct controlled empirical comparisons against manual baselines to quantify objective metrics such as task completion times, error correction rates, and inter-rater agreement.

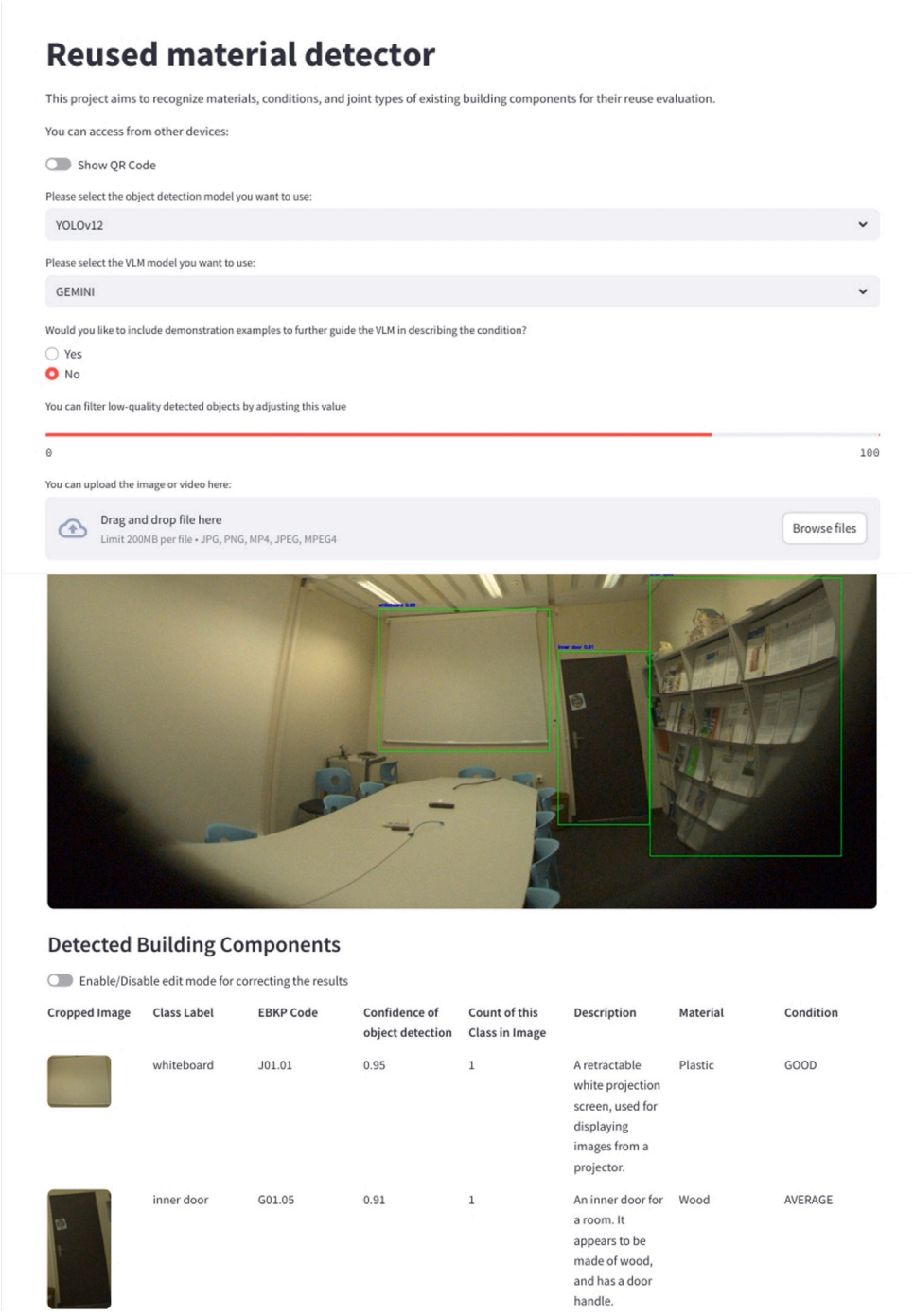


Fig. 9. Interactive web dashboard integrated with the end-to-end visual inference for existing building products assessment.

6. Discussion

6.1. Methodological implications

The results highlight that state-of-the-art VLMs remain limited in spatial precision and object localization, particularly when identifying and counting building products in complex indoor scenes. These shortcomings stem from fundamental architectural and training paradigm differences: VLMs rely on prompting and zero-shot generalization rather than spatially supervised learning. Consequently, their strengths lie in semantic reasoning rather than geometric accuracy. By contrast, specialized real-time object detection models demonstrate robust localization and bounding box alignment due to their dense spatial training regimes.

This complementary relationship suggests that high-quality spatial inference in building inventories is best achieved through hybrid integration of both approaches. Combining VLMs for semantic interpretation with CNN- and transformer-based detectors for spatial context within a modular pipeline enables dynamic generation of visual inferences with complementary characteristics. Notably, both *YOLOv12m* and *RF-DETR* maintain competitive performance even under the imbalanced long-tail class distribution characteristic of real-world building

inventory datasets. The WBF approach, ensembling *YOLOv12m* and *RF-DETR* at an optimal fusion threshold of $\tau = 0.55$, outperformed standalone models across all metrics while maintaining competitive recall. Furthermore, implementing *Qwen3-VL-8B* verification alongside spatial overlap filtering marginally enhanced the overall performance of *YOLOv12m*, whereas its application to *RF-DETR* resulted in a negligible performance decrease due to low-confidence background noise. Given that the current architecture involves sequential execution between independent models, adopting a more complex agentic framework was deemed unnecessary for this study.

6.1.1. Architectural determinants and trade-offs in spatial inference

The evaluation results demonstrate that model selection for building product detection should be aligned with specific operational requirements. *YOLOv12m* achieved the highest precision, rendering it particularly effective for applications where the mitigation of false positives is critical, such as targeted component localization or hazardous material assessment. Conversely, *RF-DETR* attained superior recall, mAP@0.5, and F1 scores, positioning it as the preferred architecture for comprehensive audits where the identification of all relevant objects is prioritized despite visual ambiguity.

While the performance margin between these fine-tuned models is narrow, *YOLOv12m* exhibited heightened sensitivity to subtle or partially occluded instances, including reflective and distant surfaces that *RF-DETR* occasionally omitted. The performance of VLMs, such as *Gemini-2.0-Flash* and *Qwen3-VL-8B-Instruct*, highlights a distinct trade-off between quantitative precision and semantic flexibility. While the visual grounding capabilities of current VLMs require further optimization to match the spatial accuracy of task-specific vision regressors, their primary utility lies in an unconstrained label space. This open-vocabulary flexibility enables VLMs to generalize across novel scenes and classes without the need for categorical retraining, offering a scalable solution for diverse and unpredictable indoor environments.

6.1.2. Non-monotonicity in few-shot condition classification

For condition classification of in situ products, three-fold cross-validation results show that while fisheye distortion did not significantly affect accuracy, reduced image resolution did. Conventional classifiers, including *ResNet* and *ViT*, underperformed in both balanced accuracy and F1 score. Although the lightweight *SmolVLM* exhibited improved balanced accuracy, its low F1 score indicates poor handling of class imbalance. These findings emphasize that open-source VLMs remain below the performance level of proprietary systems such as *Gemini-2.0-Flash*, which consistently outperformed all baselines.

Anotherfinding in this study is that few-shot prompting with labeled exemplars followed a non-monotonic pattern with respect to the number of demonstrations ($k = 1 - 4$). The strongest results were obtained at $k = 1$, followed by a gradual performance decline with increasing k. Three potential factors may explain this: a primacy effect, where the first example dominates the model's contextual focus; prompt-length saturation, where extended input sequences reduce effective attention allocation; and contextual overfitting, where repetitive exemplars bias the model toward dominant patterns in the demonstration set. To address these behaviors, future work should prioritize systematic curation of representative exemplars and embedding-based similarity retrieval for dynamic selection of contextually aligned demonstrations.

6.2. Limitations and methodological outlook

This section delineates principal limitations of the current study and outlines prospective research directions to overcome methodological challenges in the context of real-world building product inventory.

6.2.1. Expert-in-the-loop for visual interactive labeling

A significant limitation in this domain remains the lack of comprehensive, in-situ building product datasets that are both diverse and independent of particular projects. Currently, the interactive web dashboard developed in this study presents a viable foundation for expert-in-the-loop annotation by serving as a structural basis for an active learning pipeline extension. This collaborative annotation paradigm has the potential to supplement missing material and component data that are frequently absent from legacy building registers, thereby enhancing the overall completeness of circular reuse assessments.

Visual interactive labeling is especially critical for condition classification because existing built-environment datasets suffer from pronounced class imbalances. Expert contributions are imperative for generating high-quality annotations, particularly when evaluating subjective attributes such as material wear, structural damage, or residual performance. The capacity for domain experts to post-edit automated predictions within the dashboard increases annotation granularity, which is highly beneficial for composite products comprising multiple heterogeneous materials. As this annotation database expands, downstream tasks can directly leverage supervised fine-tuning of open-source models. However, the current pilot system lacks the backend infrastructure required for batch processing and automated retraining, both of which are essential to enable scalable deployment in operational environments.

To fully scale the framework against open-domain and open-vocabulary challenges in heterogeneous indoor environments, future work will explore replacing task-specific spatial regressors with Evolving World Object Detection (EWOD). This emerging paradigm couples incremental learning, domain adaptation, and unknown object detection under strict exemplar-free constraints. Recent advances in native open-vocabulary object detection architectures, including the *EW-DETR* model (Monga et al., 2026), the *ODOV* model (Zhang et al., 2026), the *EgoVision* model (e Sadima et al., 2025), and energy-based separation models (Heo et al., 2026), show immense promise for unconstrained, real-world environments in robotics and augmented reality applications. Integrating these methodologies could successfully mitigate the constraints of closed-set object vocabularies while preserving strict spatial precision in the heterogeneous built environment.

6.2.2. Standardized building product ontology

The multi-modal data generated by the dashboard provides a strong foundation for integration with downstream systems, such as digital building logbooks and reclaimed material marketplaces. Visual product data can complement geometric information from point clouds, enriching the fidelity of digital construction records. However, real-time aggregation of object counts across video sequences is not yet supported, limiting applicability in continuous site scanning. Integrating this functionality could substantially improve efficiency for automated building-level inventories.

At present, predicted product classes are automatically mapped to the corresponding eBKP codes, supporting standardization and interoperability with cost planning frameworks. This mapping facilitates robust data exchange with building information modeling (BIM) tools. Future research can extend these capabilities by adopting a building product ontology derived from established engineering taxonomies, enabling more structured and hierarchical annotation of components and their spatial relationships. Incorporating connection types and disassembly attributes into the prediction pipeline would significantly enhance the assessment of reuse potential. Organizing data in this manner would enable structured export into formats such as Industry Foundation Classes 5 (IFC 5), supporting seamless linkage to as-built BIM and digital prodcut passport generation for circular construction workflows.

7. Conclusion

This research formalizes a hybrid spatial-semantic modeling framework for automated building product inventories by reconciling the divergent strengths of task-specific spatial regressors and VLM. We establish a three-module paradigm where the weighted boxes fusion ensemble of *YOLOv12m* and *RF-DETR* reaches an $F1$ score of 0.792 on the validation set when configured at an optimal fusion threshold of $\tau = 0.55$, reducing false-positive detection. Utilizing a complementary *Qwen3-VL-8B* verification layer to recover missed targets enhance standalone *YOLOv12m* performance by +0.33%. Conversely, routing low-confidence proposals from *RF-DETR* yields a net performance decline, though the spatial overlap filter successfully mitigates this degradation by curbing the F_1 loss from −1.18% to −0.28% in environments with high background noise. Simultaneously, *Gemini-2.0-Flash* provide the semantic depth required for open-vocabulary attribute inference. A single-exemplar prompt ($k = 1$) achieves an optimal balanced accuracy of 0.79 for condition classification, whereas increasing the number of demonstrations leads to performance degradation via context dilution and reduced cue salience. Furthermore, the interactive expert-in-the-loop user interface mitigates the data scarcity challenges and long-tail class distributions inherent in real-world building inventory datasets while acting as a foundational pipeline for active learning. By transitioning from manual, fragmented audits to a structured digital product architecture, this study provides a generalizable methodology for the circular economy. The framework establishes a scalable pathway for transforming unstructured visual data into validated attributes for digital product passports and digital building logbooks to enhance material reuse potential and data fidelity within the built environment.

CRediT authorship contribution statement

Shuting Mi: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation. **Xiaofei Wu:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation. **Shunchang Liu:** Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation. **Zhao Huang:** Writing – original draft, Validation, Methodology, Formal analysis, Data curation. **Catherine De Wolf:** Writing – review & editing, Funding acquisition. **Mennatallah El-Assady:** Writing – review & editing, Funding acquisition. **Pei-Yu Wu:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is conducted at the ETH AI Center Projects in Machine Learning Research course and partly funded by Innosuisse Flagship program Swiss Digital Circular Construction Ecosystem (SWIRCULAR) project (ID: 107.235 FS-EE).

Appendix. Expert study survey

I. Qualitative assessment of AI annotations

How effectively does the interface present predicted component labels, material type and condition labels for your assessment? Is there any information missing for you to assess the reuse potential of elements/components?

II. Trust calibration and interaction

To what extent do you agree that being able to revise the prediction output increases your engagement and confidence in collaborating with AI models? (1 = Strongly disagree, 5 = Strongly agree)

III. Comparative operational utility

In your professional opinion, what are the primary advantages or disadvantages of this hybrid AI-driven approach compared to traditional site-walkthrough documentation?

Data availability

Data will be made available on request.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, Antoine, Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K., 2022. Flamingo: a visual language model for few-shot learning. In: 36th Conference on Neural Information Processing Systems. NeurIPS.

Bao, H., Wang, W., Dong, L., Liu, Q., Mohammed, O.K., Aggarwal, K., Som, S., Wei, F., 2022. VLMo: Unified vision-language pre-training with mixture-of-modality-experts. arXiv:2111.02358.

Chen, X., 2024. Construction Safety Inspection with Large Pre-Trained Vision-Language Models (PhD. thesis). University of British Columbia.

Chen, Q., Yin, X., 2025. Tailored vision-language framework for automated hazard identification and report generation in construction sites. Adv. Eng. Inform. 66, 103478. http://dx.doi.org/10.1016/j.aei.2025.103478.

Chen, S., Zhu, T., Zhou, R., Zhang, J., Gao, S., Niebles, J.C., Geva, M., He, J., Wu, J., Li, M., 2025. Why is spatial reasoning hard for VLMs? An attention mechanism perspective on focus areas. arXiv:2503.01773.

Ciaglia, F., Zuppichini, F.S., Guerrie, P., McQuade, M., Solawetz, J, 2022. Roboflow 100: a rich, multi-domain object detection benchmark. arXiv:2211.13523.

De Wolf, C., Byers, B.S., Raghu, D., Gordon, M., Schwarzkopf, V., Triantafyllidis, E., 2024. D5 digital circular workflow: five digital steps towards matchmaking for material reuse in construction. Npj Mater. Sustain. 2, 36. http://dx.doi.org/10.1038/s44296-024-00034-8.

Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L., 2023. QLoRA: Efficient finetuning of quantized LLMs. arXiv:2305.14314.

DIN German Institute for Standardization, 2024. DIN SPEC 91481:2024-02 classification of recycled plastics of polyamides by data quality levels for use and (digital) trading. URL: https://www.dinmedia.de/en/technical-rule/din-spec-91481/375901933.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In: The 9th International Conference on Learning Representations. ICLR.

Du, J., Yu, A., Sugumaran, V., Hu, Y., Yu, G., Zhang, J., 2025. Ontological reasoning for automated tunnel defect diagnosis and root cause identification. Autom. Constr. 178, 106362. http://dx.doi.org/10.1016/j.autcon.2025.106362.

Engel, J., Somasundaram, K., Goesele, M., Sun, A., Gamino, A., Turner, A., Talattof, A., Yuan, A., Souti, B., Meredith, B., Peng, C., Sweeney, C., Wilson, C., Barnes, D., DeTone, D., Caruso, D., Valleroy, D., Ginjupalli, D., Frost, D., Miller, E., Mueggler, E., Oleinik, E., Zhang, F., Somasundaram, G., Solaira, G., Lanaras, H., Howard-Jenkins, H., Tang, H., Kim, H.J., Rivera, J., Luo, J., Dong, J., Straub, J., Bailey, K., Eckenhoff, K., Ma, L., Pesqueira, L., Schwesinger, M., Monge, M., Yang, N., Charron, N., Raina, N., Parkhi, O., Borschowa, P., Moulon, P., Gupta, P., Mur-Artal, R., Pennington, R., Kulkarni, S., Miglani, S., Gondi, S., Solanki, S., Diener, S., Cheng, S., Green, S., Saarinen, S., Patra, S., Mourikis, T., Whelan, T., Singh, T., Balntas, V., Baiyya, V., Dreewes, W., Pan, X., Lou, Y., Zhao, Y., Mansour, Y., Zou, Y., Lv, Z., Wang, Z., Yan, M., Ren, C., De Nardi, R., Newcombe, R., 2023. Project aria: A new tool for egocentric multi-modal AI research. arXiv:2308.13561.

European Commission, 2024. EU Construction & Demolition Waste Management Protocol. Tech. Rep., URL: https://op.europa.eu/en/publication-detail/-/publication/d63d5a8f-64e8-11ef-a8ba-01aa75ed71a1/language-en.

Ghosh, A., Acharya, A., Saha, S., Jain, V., Chadha, A., 2024. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. J. Knowl. Based Syst. arXiv:2404.07214.

Gupta, A., Dollár, P., Girshick, R., 2019. LVIS: A dataset for large vocabulary instance segmentation. arXiv:1908.03195.

Heo, J., Park, K., Park, G.-M., 2026. Detecting unknown objects via energy-based separation for open world object detection. arXiv:2603.29954.

Holter, S., El-Assady, M., 2024. Deconstructing human-AI collaboration: Agency, interaction, and adaptation. Comput. Graph. Forum 43 (3), http://dx.doi.org/10.1111/cgf.15107.

Honic, M., Magalhães, P.M., Bosch, P.V.D., 2024. From data templates to material passports and digital product passports. In: A Circular Built Environment in the Digital Age. Springer, pp. 79–94. http://dx.doi.org/10.1007/978-3-031-39675-5_5.

Keulemans, G., Adams, R., 2024. Emergent digital possibilities for design-led reuse within circular economy. Npj Urban Sustain. 4 (1), http://dx.doi.org/10.1038/s42949-024-00164-x.

Kim, J., 2020. Visual analytics for operation-level construction monitoring and documentation: state-of-the-art technologies, research challenges, and future directions. Front. Built Environ. 6, http://dx.doi.org/10.3389/fbuil.2020.575738.

Kim, T., Kim, S., Chern, W.-C., Park, S., Kim, D., 2025. Optimizing large vision-language models for context-aware construction safety assessment. SSRN. URL: https://ssrn.com/abstract=5200113.

Li, J., Zhao, X., Kong, L., Zhang, L., Zou, Z., 2024. Construction activity recognition method based on object detection, attention orientation estimation, and person re-identification. Buildings 14 (6), 1644. http://dx.doi.org/10.3390/buildings14061644.

Liang, Z., Chan, J.C.F., Zhang, J., Liang, Z., Wang, B., Wang, M., Cheng, J.C.P., 2026. Multimodal LLM-driven language-embedded 3D Gaussian splatting for semantic and realistic digitization of historical buildings. Autom. Constr. 181, 106628. http://dx.doi.org/10.1016/j.autcon.2025.106628.

Liu, H., Li, C., Wu, Q., Lee, Y.J., 2023. Visual instruction tuning. In: 37th Conference on Neural Information Processing Systems. NeurIPS, URL: https://llava-vl.github.io.

Lv, Y., Pang, Y., Ju, L., Xu, W., Liang, Z., He, J., Zhang, L., 2026. Automating digitalization of engineered slopes to support interactive digital twin visualization. Autom. Constr. 181, 106564. http://dx.doi.org/10.1016/j.autcon.2025.106564.

Monga, M., Chudasama, V., Wasnik, P., Jawahar, C.V., 2026. EW-DETR: Evolving world object detection via incremental low-rank detection transformer. arXiv 2602.20985. arXiv:2602.20985.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., Fernández-Leal, Á., 2023. Human-in-the-loop machine learning: a state of the art. Artif. Intell. Rev. 56 (4), 3005–3054. http://dx.doi.org/10.1007/s10462-022-10246-w.

Ollama Team, 2024. Ollama. URL: https://ollama.com/.

OpenAI, 2024. Gpt-4 technical report. arXiv:2303.08774v6.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P., 2024. DINOv2: Learning robust visual features without supervision. Trans. Mach. Learn. Res. arXiv:2304.07193v2.

Paneru, S., Jeelani, I., 2021. Computer vision applications in construction: Current state, opportunities & challenges. Autom. Constr. 132, 103940. http://dx.doi.org/10.1016/j.autcon.2021.103940.

Qwen Team, 2025. Qwen3 technical report. arXiv:2505.09388.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision. arXiv:2103.00020.

Rao, P.A., Rahman, M.M., Duraman, S.B., 2025. Adopting circular economy in construction: a review. Front. Built Environ. 11, http://dx.doi.org/10.3389/fbuil.2025.1519219.

Robinson, I., Robicheaux, P., Popov, M., 2025. RF-DETR. URL: https://github.com/roboflow/rf-detr.

e Sadima, U., Yin, X., Alkhrijah, Y., Hamid, D., Ehatisham Ul Haq, M., Muhammad Usman, S., Khalid, S., A. Alawad, M., 2025. Egovision a yolo-vit hybrid for robust egocentric object recognition. Sci Rep 15, 34723. http://dx.doi.org/10.1038/s41598-025-18341-y.

Saka, A., Taiwo, R., Saka, N., Oluleye, B., Dauda, J., Akanbi, L., 2024. Integrated bim and machine learning system for circularity prediction of construction demolition waste. arXiv:2407.14847.

Solovyev, R., Wang, W., Gabruseva, T., 2021. Weighted boxes fusion: ensembling boxes from different object detection models. Image Vis. Comput. 107, 104117. http://dx.doi.org/10.1016/j.imavis.2021.104117.

Sun, Y., Gu, Z., Yang, S.B., 2024. Probing vision and language models for construction waste material recognition. Autom. Constr. 166, 105629. http://dx.doi.org/10.1016/j.autcon.2024.105629.

Sun, X., Pan, Z., Gao, X., Wu, K., Bera, A., 2025. Text-guided generation of efficient personalized inspection plans. arXiv:2506.02917.

Swiss Society of Engineers and Architects, 2020. eBKP-H SN 506 511 Elementbasierter Baukostenplan Hochbau. URL: https://www.crb.ch/.

Swiss Society of Engineers and Architects, 2023. SIA 430:2023 Vermeidung und Entsorgung von Bauabfällen. Zurich.

Tang, Z., Yang, Z., Khademi, M., Liu, Y., Zhu, C., Bansal, M., 2024. CoDi-2: I:n-context, interleaved, and interactive any-to-any generation. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, URL: https://codi-2.github.io.

Team, Gemini, 2023. Gemini: A Family of Highly Capable Multimodal Models. Tech. Rep., Google., arXiv:2312.11805.

Tian, Y., Ye, Q., Doermann, D., 2025. YOLOv12: Attention-centric real-time object detectors. arXiv:2502.12524.

Tong, Z., Song, Y., Wang, J., Wang, L., 2022. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In: 36th Conference on Neural Information Processing Systems. NeurIPS, URL: https://github.com/MCG-NJU/VideoMAE.

Tong, Z., Yang, Z., Zhu, C., Zeng, M., Bansal, M., 2023. Any-to-any generation via composable diffusion. In: 37th Conference on Neural Information Processing Systems. NeurIPS, URL: https://codi-gen.github.io.

Wondimu, N.A., Buche, C., Visser, U., 2022. Interactive machine learning: A state of the art review. arXiv:2207.06196.

Wong, M.F., Qi, X., Tan, C.W., 2022. EuclidNet: Deep visual reasoning for constructible problems in geometry. http://dx.doi.org/10.54364/aaiml.2023.1152, arXiv:2301.13007.

Wu, P.-Y., 2024. Data-driven Approaches for Predicting Hazardous Substances in the Building Stock (PhD. thesis). Lund University, ISBN: 9789188722829.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., He, L., 2022. A survey of human-in-the-loop for machine learning. Future Gener. Comput. Syst. 135, 364–381. http://dx.doi.org/10.1016/j.future.2022.05.014.

Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., Liu, Z., Liu, C., Zeng, M., Wang, L., 2023. MM-REACT: Prompting ChatGPT for multimodal reasoning and action. arXiv:2303.11381.

Zhang, C., Asce, S.M., Kim, J., Jeon, J., Xing, J., Ahn, C., Asce, A.M., Tang, P., Asce, M., Cai, H., 2021. Toward integrated human-machine intelligence for civil engineering: An interdisciplinary perspective. arXiv:2107.13498.

Zhang, J., Chen, Y., Zhou, Y., Xu, Y., Huang, Z., Mei, J., Chen, J., Yuan, Y.-J., Cai, X., Huang, G., Quan, X., Xu, H., Zhang, L., 2025. From flatland to space: Teaching vision-language models to perceive and reason in 3D. arXiv:2503.22976.

Zhang, Y., Han, R., Zhou, F., Feng, W., Wan, L., 2026. ODOV: Benchmark the open-domain open-vocabulary object detection. arXiv:2508.01253.