2026 European Conference on Computing in Construction
Corfu, Greece
July 12-15, 2026

VISION–LANGUAGE MODELS IN CONSTRUCTION: ASSESSING PROMPT ROBUSTNESS IN OPEN-VOCABULARY OBJECT DETECTION

Federica Madaschi[1], Gelare Taherian[2], Marco Lorenzo Agostino Trani[1] and Ehsan Rezazadeh Azar[2]
[1]Department of Architecture Built environment and Construction Engineering,
Polytechnic University of Milan, Milan, Italy
[2]Department of Architectural Science, Toronto Metropolitan University, Toronto, Canada

Abstract

Vision–language models provide an alternative to supervised computer vision, enabling open-set object recognition without retraining or large datasets. Recent advances like GroundingDINO enable open-vocabulary object detection using natural-language prompts, but their sensitivity to prompt formulation in the AEC domain remains underexplored. This study assesses GroundingDINO's reliability in detecting construction objects through analysis of prompt variations, including lexical variation and prompt multiplicity. A custom dataset of 48 objects was tested under eight prompts. Results show lexical variation affects detection performance, with "regular base term" achieving a 56.1% F1 score. Multiple prompting lowers performance by ~2%, highlighting AEC's effectiveness and limits.

Introduction

Computer vision has been increasingly adopted in the Architecture, Engineering, and Construction (AEC) domain to support operational and managerial tasks such as progress monitoring, safety inspections, and quality control. However, deployment faces challenges due to scene changes, object diversity, occlusions, and project-specific visuals. Conventional supervised object detection needs large, labelled datasets and a fixed set of objects, limiting scalability and adaptability across projects and phases. (Duan et al., 2022). Vision–language models (VLMs) provide the foundation for open-vocabulary object detection (OVOD), enabling detection driven by natural-language prompts rather than predefined class sets, and have shown promise in overcoming closed-set limitations in AEC applications (Adil et al., 2025).

Vision–language models (VLMs) support open-vocabulary object detection (OVOD), which uses natural-language prompts instead of fixed labels. This offers an alternative to closed-set detection in AEC (Adil et al., 2025). OVOD models such as *GroundingDINO* (Liu et al., 2023) accept textual queries and provide spatially grounded detections, allowing flexible scene querying without retraining or extensive annotation. However, the quality and relevance of the detections generated by VLMs depend heavily on the prompt's quality and formulation (Minderer et al., 2024; Nebbia and Kovashka, 2024). Prior work shows prompt design greatly influences vision-language model performance and outcomes, underscoring their sensitivity to wording and structure (Gu et al., 2023). Since prompts are key to customizing VLM behavior, their design is crucial to aligning outputs with specific AEC needs.

In safety- and management-critical workflows, sensitivity to prompt formulation may lead to missed detections or false positives, undermining trust in automated systems. While recent studies have demonstrated the feasibility of VLM-based methods for construction-related tasks, the role of textual prompt formulation has not yet been systematically examined in AEC contexts, where scene complexity and operational requirements pose additional challenges. Moreover, prompt-related reliability issues, such as inconsistent detections or hallucinated objects, have not been explicitly analyzed in complex construction scenes. To address this gap, this paper presents a structured evaluation of prompt robustness for open-vocabulary object detection in construction settings. Focusing on *GroundingDINO*, the study investigates two key factors: (1) lexical variation, examining how different linguistic representations of the same construction object affect detection performance, and (2) prompt multiplicity, assessing the impact of single versus multiple prompts on accuracy and false positives. Using a custom dataset of construction-related objects spanning indoor and outdoor scenes, the study aims to benchmark the practical effectiveness and limitations of prompt-driven OVOD for AEC operational and managerial applications.

Related Works

Recent literature on OVOD methods has drawn mainly on pre-trained vision-language models, such as CLIP (Radford et al., 2021), which enable a detector to be queried with natural-language inputs, thereby overcoming the limitations of closed-vocabulary paradigms. In this context, numerous works have adopted hand-crafted textual prompts, in which the class name is inserted into simple linguistic templates (e.g., "a photo of a {class}"), sometimes combining multiple templates and averaging their embeddings to reduce dependence on a single formulation (Minderer *et al.*, 2024). These approaches have demonstrated the effectiveness of textual prompting for querying OVOD models, but they implicitly assume a stable correspondence between the canonical class name and the visual concept. A subsequent line of research challenged this assumption by examining lexical variation in prompts, including the use of synonyms and

semantically related terms. Recent studies show that replacing the base term with synonyms or near-synonyms can significantly degrade performance, highlighting the limited semantic robustness of OVOD models relative to equivalent linguistic variations (Nebbia and Kovashka, 2024). To mitigate this sensitivity, class-embedding enrichment strategies have been proposed by aggregating multiple prompts that include both the base term and its synonyms, suggesting that prompt design is a structural component of the detection process (Gu *et al.*, 2022). At the same time, other works have extended the use of prompts beyond isolated class names, introducing descriptive and attributive nominal phrases, as in the Open-Vocabulary Attribute Detection task, where textual representations consist of name-adjective combinations or more articulated descriptions (Bravo *et al.*, 2023). Even when such formulations are not explicitly defined as "definitions", they implicitly encode information about the function, material, and context of the object, delineating a continuum from the bare class name to increasingly rich and semantically specific prompts. However, most existing studies focus on the impact of these strategies on average accuracy, without systematically analyzing how the prompt's linguistic complexity affects model stability or the risk of generating false predictions. This is the context for *GroundingDINO*, an advanced OVOD model that accepts free-form natural language sentences and returns spatially anchored bounding boxes aligned with textual queries (Liu *et al.*, 2023). *GroundingDINO* is queried using single object names, short sentences with attributes, or multiple prompts separated by punctuation. However, the literature analyzing its performance tends to keep the textual prompt fixed or generic, focusing primarily on variations in the scene or visual content, as in studies that evaluate the model's robustness to visual perturbations while keeping The present work analyses *Groundingdino*'s prompt formulation, especially in construction, building on existing experimental methods and focusing on lexical variation and multiplicity (Mütze et al., 2025).

Methodology

The methodology systematically generates prompts to evaluate the OVOD model's performance across different texts, as shown in Figure 1. After creating a custom AEC dataset, prompt formulations were organized along two dimensions. The first-dimension concerns prompt complexity, distinguishing between easy and complex prompts. Easy prompts are characterized by minimal linguistic manipulation of the target object and involve simple lexical variations, such as a base term or a synonym. In contrast, complex prompts introduce additional levels of linguistic transformation as descriptive definitions, thereby increasing interpretive complexity while still referring to the same visual target. This dimension was designed to assess the model's sensitivity to wording and semantic expressiveness through controlled lexical variations for each object class. The second-dimension addresses prompt multiplicity, differentiating between single- and multiple-prompt configurations. Single Prompt settings rely on a single lexical call for a single object present in the analyzed image, enabling the evaluation of the model's response to isolated linguistic variations. Based on the prompt type identified in the first stage, a Multiple Prompt configuration was subsequently defined, allowing comparison of OVOD performance when detecting AEC-related objects under single- and multi-prompt conditions.

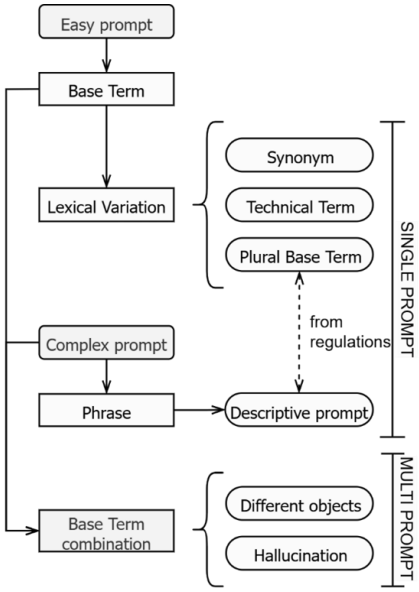


Figure 1: Overview of the prompt formulation strategy

Dataset development

A tailored image dataset was created to facilitate model assessment across various object categories and varying levels of visual complexity. Specifically, 48 construction-related objects were included and mainly organized into indoor and outdoor scenario groups. For each object, three images were selected to represent a range of realistic conditions, resulting in a custom dataset of 144 images. These depict variations in visual appearance: a normal view, where the object is the primary subject; a small-scale view, where the object appears reduced relative to the surrounding context (Lin *et al.*, 2014); and a cluttered view, in which other elements in the scene partially occlude the object (Malaikrisanachalee et al., 2025). A structured naming convention was adopted to organize the dataset. To each object a Base Term was assigned, intended as the canonical object name used for Ground Truth labelling, while a suffix indicated the visual condition (Normal, Small, Cluttered). The Base Terms were formulated using natural, everyday language, following the labelling strategy introduced by Taherian et al. (2025).

Lexical variations definition

In the single-prompt setting, lexical variation was employed to systematically evaluate the model's sensitivity to different textual references for the same visual. Each prompt includes a single linguistic

Table 1: Example of Lexical prompt strategies used in the single-prompt evaluation

Prompt Strategy	Indoor	Outdoor
Base Term [Baseline]	Split System	Construction Barrier
Synonym Term	Air Conditioner	Barricade
Technical Term	Ducted Air-Conditioner	Impact Barrier
Plural Term	Ducted Air-Conditioners	Impact Barriers
Definition Term	Encased assembly or assemblies designed primarily to provide ducted delivery of conditioned air to an enclosed space, room or zone	Energy-absorbing device installed in front of a rigid object to reduce the severity of impact of a vehicle

expression linked to one target object in the image, which helps isolate the impact of wording changes from the image's context or composition. For each object class, multiple lexical variants were created, such as a base term, synonyms, technical terms, plural forms, and descriptive definitions. Concrete examples of the prompt construction for each lexical category are reported in Table 1, which illustrates how prompts were instantiated for representative indoor and outdoor object classes. All variants were evaluated independently under identical but different visual conditions, allowing a controlled comparison of how lexical choices influence detection performance while maintaining a one-to-one correspondence between the prompt and the object.

Base Term

The Base Term labels adopted in this study are based on the labelling strategy introduced by Taherian et al. for evaluating *GroundingDINO* as an OVOD model on construction job sites under complex visual scenarios. In that work, object class labels were intentionally formulated using clear, unambiguous natural-language expressions, with each image assigned to a single target object. This approach ensured a one-to-one correspondence between the text prompt and the visual instance. Following the same principle, Base Term in this study is defined as the most direct and commonly used words for each object class, reflecting standard naming practices used by professionals in the AEC field. These terms act as the linguistic baseline for prompt formulation and show the least linguistically altered way of naming objects, against which the effects of lexical differences are later assessed. However, the work by Taherian et al. did not explicitly analyze the effects of lexical variation on model performance. Building on this gap, the present work further examines the influence of inflectional variants by evaluating whether the use of singular or plural forms of a term affects object recognition accuracy (Minderer *et al.*, 2024).

Synonym Term

Lexical variation using Synonyms Terms was analyzed to verify the model's ability to recognize the same object when indicated using semantically equivalent terms (Nebbia and Kovashka, 2024).For each object class, synonyms were assigned through a cross-dataset analysis of naming conventions in related datasets, including *COCO* (Lin *et al.*, 2014), *Objects365* (Shao *et al.*, 2019), *Open Images* (Open Images V7), and *SODA* (Duan *et al.*, 2022). For missing class, *WordNet* served as a complementary lexical reference (Miller et al., n.d.), and for remaining domain-specific classes not supported by either dataset taxonomies or *WordNet*, additional terminology was derived from online annotation repositories such as *Roboflow* to reflect common usage. This cross-sectional analysis of multiple datasets and resources not only underscored domain gaps, as several object classes in the developed dataset were not well represented in existing datasets, but also allowed for mitigating biases associated with individual taxonomies and evaluating the model's robustness to synonym-based lexical substitutions.

Technical Term and Plural Term

Lexical variation using "related technical terms" was examined, with a focus on terminology in construction sector regulations. The Technical Terms in tests came from international standards in construction and plant sectors, including the *International Organisation for Standardisation* (ISO), the *International Electrotechnical Commission* (IEC), and the *European Norms* (EN/CEN). This choice is motivated by the need to evaluate the model's behavior in the presence of specialized terminology consistent with the application domain's regulatory practice. The main regulations analyzed are listed in Table 2.

In addition, these Technical Terms were also inflected according to standard English grammatical rules to generate Plural Forms, enabling the assessment of the model's response to number variation within specialized terminology.

Descriptive Term

Finally, tests based on descriptive prompts (Descriptive Technical Term) have been established, as many open-vocabulary object detection benchmarks use textual representations based on descriptive sentences rather than individual tokens (Bianchi *et al.*, 2023). Technical regulations often provide descriptive textual definitions of objects, which have been utilized to construct descriptive technical terms. These definitions are regarded as formal and unambiguous linguistic representations of objects.

Prompt Multiplicity

In contrast to single-prompt inference, the multiple-prompt approach enables the model to detect and localize several object categories within a single image. This is achieved by constructing a prompt configuration

composed of multiple textual queries, each corresponding to a target object class. The input prompt is formatted as a sequence of object descriptors separated by delimiters, for example:

Prompt = {"object 1.", "object 2.", ..., "object n."}.

Within this setup, the prompt configuration is further extended to evaluate model hallucination. This is done by augmenting the multi-object prompt with an additional query that corresponds to an object does not present in the image. The absent object is selected from the dataset object categories based on its visual similarity to the AEC-related object(s) depicted in the scene, enabling systematic evaluation of the model's susceptibility to visually induced false detections under multi-prompt conditions. An example of prompt format is as following with further demonstration provided in Figure 5.f:

Prompt = {"object 1", "object 2", ..., "object n", "absent object"}.

Table 2: Regulations analyzed for defining the Single prompt-Technical Term

Standard	Domain
ISO 13253:2017	Ducted air conditioners & air-to-air heat pumps
ISO 6707-1:2020	Building construction
ISO 7240-1:2014	Fire safety
ISO 8421-4:1990	Fire safety
ISO 6165:2022	Earth-moving machinery
ISO 30500:2018	Non-sewered sanitation systems
ISO 18650-1	Construction machinery
ISO 45001:2018	Occupational health and safety management systems
ISO 8528-1:2018	Generating sets
ISO 7010:2019	Safety signage & fire prevention symbols
ISO 830:2024	Logistics
ISO 14122-3:2016	Machinery safety
ISO 19085-9:2024	Machinery safety
ISO/CIE 8995-1:2025	Workplace lighting
IEC 61439-2:2020	Low-voltage switchgear and controlgear assemblies
IEC 60364-1:2025	Low-voltage electrical installations
IEC 60050 (series)	Electrotechnical terminology
EN 442-2:2014	Heating systems

Experiment Setup

Based on the proposed framework, prompt elements were systematically defined and combined to create experimental test cases, enabling a controlled analysis of how different forms of lexical and compositional variation influence the OVOD model's object recognition performance.

OVOD Model

In this research, *GroundingDINO* is used as the OVOD model. *GroundingDINO* is designed to operate in scenarios where the categories to be detected are not fixed in advance but are defined dynamically through textual prompts. In this paradigm, text is the primary means for semantic conditioning of inference, making prompt formulation a key factor in the model's behavior, as discussed in works on grounded detection and phrase grounding (Li *et al.*, 2021; Liu *et al.*, 2023). The model is built upon the Transformer-based DINO detector, a member of the DETR family, and extends it by incorporating grounded language–image pre-training for open-set object detection (Carion *et al.*, 2020; Li *et al.*, 2021). In this approach, categories are not treated as separate labels but as context-aware linguistic representations generated by a Transformer-based text encoder, enabling connections between visual regions and textual descriptions, including lexical variants and free phrases (Devlin *et al.*, 2018; Li *et al.*, 2021). From an architectural perspective, *GroundingDINO* combines a multi-scale visual backbone with a pre-trained language encoder and employs a deep multimodal fusion strategy in which the textual prompt affects multiple stages of the detection process, from visual feature enhancement to query initialisation and final prediction refinement. (Liu *et al.*, 2023). Consequently, variations in prompt formulation can affect text–region alignment and prediction consistency, making the model sensitive to linguistic choices at inference time. For inference in this study, GroundingDINO was implemented using the Hugging Face Transformers via zero-shot-object-detection pipeline (Hugging Face, 2025). Given the input prompts, the model returns bounding boxes, associated labels, and confidence scores for detected regions. Each confidence score is computed by selecting the maximum logit across all candidate labels and applying a sigmoid function to obtain a normalized value between 0 and 1, reflecting the alignment between a region and its most relevant textual label. A single confidence threshold is then applied to filter the final detection outputs. In this study, a threshold of 0.25 was used to retain predictions for further evaluation (Raoofi *et al.*, 2024). Future work will investigate the impact of varying this threshold on detection performance and its interaction with prompt formulation.

Evaluation Metrics

To evaluate the performance, detection outputs were filtered using a confidence threshold of 0.25 prior to ground-truth (GT) association and metric computation. *GroundingDINO* operates on free-form input prompts and the evaluation involves the verification of region–prompt alignment. A predicted bounding box was considered a correct detection if its Intersection-over-Union (IoU) with a GT instance exceeded the threshold, and the assigned class label correctly corresponded to the ground-truth class label. IoU quantifies spatial overlap and is defined as the ratio of the intersection area $|A \cap B|$ to the union area $|A \cup B|$ between the predicted (A) and GT (B) bounding boxes.

IoU is highly sensitive to minor localization errors, where even small pixel-level shifts can reduce the overlap,

despite successful detection. Since the dataset in this study exhibits scale variation and includes particularly small objects per class, the IoU threshold was set to 0.25 ensuring that the detection of small-scale indoor components is not unfairly penalized while maintaining the focus on reliable object presence verification (Eggert *et al.*, 2017). Precision, Recall, and F1-score were computed for each object class and experimental scenario according to Eq. (1). A detection was counted as a true positive when the IoU threshold was met. GT instances without matched predictions were labelled as false negatives, while incorrectly associated predictions were treated as false positives.

$$P = \frac{TP}{TP+FP}, R = \frac{TP}{TP+FN}, F1 = \frac{2\times P\times R}{P+R} \quad (1)$$

Results and Discussion

This section shows *GroundingDINO*'s evaluation results in construction scenarios. The analysis focuses on two main questions: (i) how prompt formulation impacts the reliability of OVOD; (ii) whether using multiple prompts enhances performance.

An initial overview of the results compares the Single-Prompt and Multiple-Prompt configurations for indoor and otdoor scenes, as shown in Figure 2. The Multiple Prompt setup does not significantly outperform the Single Prompt baseline: recall drops slightly from 0.752 to 0.716, while precision (0.448 vs. 0.446) and F1 score (0.561 vs. 0.550) are nearly the same or slightly lower with the multiple-prompt configuration. These results suggest that, in the present experimental setting, increasing the number of prompts does not yield a clear overall improvement in OVOD performance and may introduce additional ambiguity.

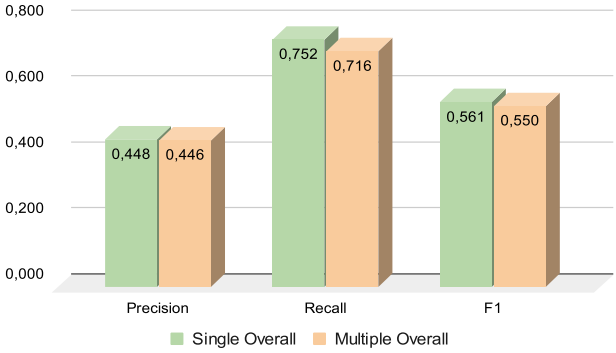


Figure 2: Overall performance comparison between Single Prompt and Multiple Prompt.

A more detailed analysis categorized by scene type indicates that the effect of prompt formulation is significantly influenced by contextual factors. In indoor environments (Figure 3), the utilization of multiple prompts results in a moderate enhancement of the F1 score, rising from 0.547 in the Single Prompt condition to 0.577 in the Multiple Prompt arrangement. This enhancement is chiefly attributable to an increase in recall (from 0.706 to 0.750), with a slight improvement observed in precision (from 0.447 to 0.469). These results suggest that, in structured indoor construction scenes with constrained geometry and reduced background clutter, prompt multiplicity can help the model retrieve a broader set of relevant objects without substantially increasing false positives.

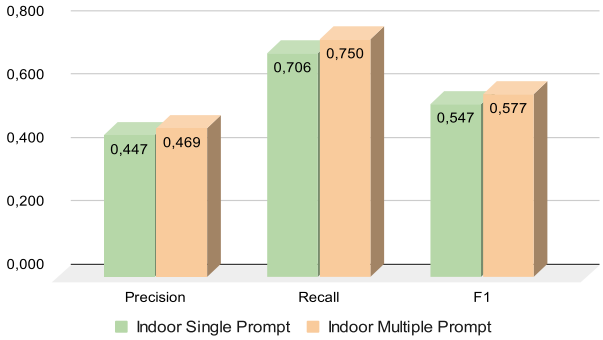


Figure 3: Performance comparison by scene type: Indoor

In contrast, outdoor construction scenes exhibit markedly different behavior (Figure 4). In this case, the Multiple Prompt configuration leads to a simultaneous reduction in recall (from 0.800 to 0.680) and precision (from 0.449 to 0.422), resulting in a more noticeable drop in F1 score from 0.575 to 0.520.

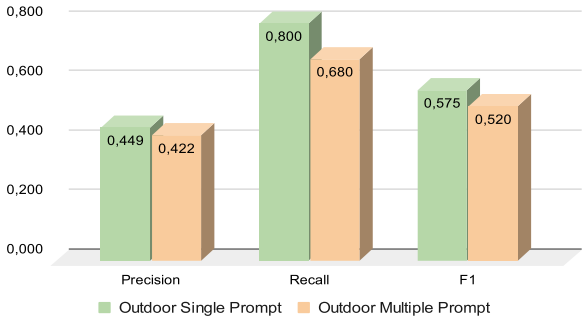


Figure 4: Performance comparison by scene type: Outdoor.

In visually complex, less structured outdoor environments, prompt multiplicity does not yield clear improvements in detecting and discriminating target objects and may be associated with less stable detection outcomes. The presence of heterogeneous materials and temporary structures can raise semantic interference between prompts, leading to both missed detections and false positives. Overall, managing multiple prompts seems more difficult in these environments. To further investigate prompt sensitivity, the following analysis focuses on lexical variation under single-prompt conditions, allowing the effect of wording to be examined independently from prompt multiplicity.

Lexical variation

Table 3 summarizes how lexical variation impacts detection performance. Because this analysis aims to assess the model's sensitivity to lexical differences rather than scene-specific features, the results are aggregated across both indoor and outdoor environments. The Base Term prompt serves as the baseline and achieves the highest overall F1 score (0.561), suggesting that normal and commonly used language for object class names

yields the most reliable alignment between textual prompts and visual representations in this experimental setting. The Synonym Term strategy yields a very similar performance (F1 = 0.555), with nearly identical precision and a slight reduction in recall. This result suggests that *GroundingDINO* exhibits a limited performance variation across closely related lexical forms, although even minor variations in wording can affect recall. A pronounced performance degradation is observed when transitioning to Technical Term prompts. In this case, both precision (0.395) and recall (0.622) decrease, resulting in a lower F1 score (0.483). This behaviour indicates that specialised and domain-specific terminology, while semantically precise from a technical or regulatory standpoint, is less effectively grounded by the model. This reflects a mismatch between the linguistic distributions encountered during large-scale pre-training and the formal vocabulary commonly used in construction standards and technical documentation. The use of Plural Term further degrades performance (F1 = 0.454), with the reduction driven primarily by a drop in precision. This finding suggests that the model is sensitive to inflectional variation and that plurality introduces ambiguity in the mapping between textual prompts and visual instances, even when the underlying semantic concept remains unchanged. This ambiguity increases false detections, thereby reducing detection reliability. The lowest performance is obtained with Definition Term prompts, which exhibit a substantial decline across all metrics (F1 = 0.314). In this configuration, the combination of descriptive modifiers and increased linguistic complexity prevents the model from identifying a stable visual reference for the target object, resulting in both poor recall and a high false-positive rate. Overall, these results suggest a trade-off between lexical variation and detection robustness. While *GroundingDINO* appears to tolerate limited lexical variation, performance tends to decrease as prompts differ from concise normal language object names and incorporate greater technical, inflectional, or descriptive complexity. This emphasizes that in the context of OVOD for construction applications, semantically accurate prompt formulations from a human perspective do not necessarily translate into improved model performance and may instead hinder reliable grounding.

Table 3: Precision, recall, and F1 scores for single-prompt lexical variation strategies.

Prompt Strategy	Precision	Recall	F1
Base Term	0.448	0.752	0.561
Synonym Term	0.451	0.720	0.555
Technical Term	0.395	0.622	0.483
Plural Term	0.347	0.657	0.454
Definition Term	0.230	0.496	0.314

Multiple prompting and Hallucination.

In this section, the analysis focuses on the effects of multiple promoting and hallucination, particularly when prompts include object categories that are not present in the scene. Table 4 reports both absolute performance metrics and relative percentage variations with respect to the single-prompt baseline, enabling a clearer comparison of the impact of hallucination and multi-object prompting configurations on detection performance.

Table 4: Precision (P), recall (R), and F1 scores for SP, MP, and Hall. configurations (Δ% relative to single prompt).

	P	ΔP (%)	R	ΔR (%)	F1	Δ F1 (%)
SP	0.448	–	0.752	–	0.561	–
MP	0.446	−0.4%	0.716	−4.8%	0.550	−2.0%
Hall.	0.447	−0.2%	0.737	−2.0%	0.557	−0.7%

Findings show that Precision remains nearly unchanged across configurations. Moving from single-prompt to multiple-prompt, precision decreases by approximately 0.4%, while the hallucination setting shows only a minor decrease of 0.2%. Recall exhibits a more noticeable change, with a 4.8% decrease for multiple-prompt and 2.0% decrease for hallucination relative to the single-prompt baseline, resulting in modest decreases in F1 score (−2.0% and −0.7%, respectively). These results suggest that hallucination primarily affects the model's ability to reliably retrieve all relevant objects, while having a limited impact on precision.

Given the qualitative examples in Figure 5, the results show that the impact of increased prompt multiplicity is class dependent. For instance, the *sprinkler* (Figure 5-a/5-b) category is not detected under single prompting and remains unaffected by multiple prompting. In contrast, some categories exhibit stable behavior across prompting strategies, such as *ladder* (Figure 5-c/5-d), which is correctly detected under both single- and multiple-prompt configurations. Other categories, however, show marked performance variations when multiple prompts are introduced, reflecting increased semantic ambiguity and class-level confusion. This effect is evident for *red & white tape* (Figure 5-d), where multiple prompting leads to an increase in false detections. Moreover, the results indicate that introducing prompts for object categories absent from the scene can induce false detections without affecting the detection of present targets. An example of this behavior occurs when the "absent duct" (Figure 5-f) category is incorrectly assigned to another object in the image while the present object still is correctly detected (Figure 5-e compared to Figure 5-f). Overall, the comparison between configurations indicates that hallucination impacts OVOD performance by adding false detections. This reveals a trade-off between target detection (recall) and detection accuracy (precision), which should be carefully balanced based on the specificity and requirements of the construction scenario under analysis.

Limitations and Future Works

This work is intended as a focused, exploratory study of prompt sensitivity in OVOD for construction scenarios. The experimental design studies different prompt configurations across visual conditions for various object

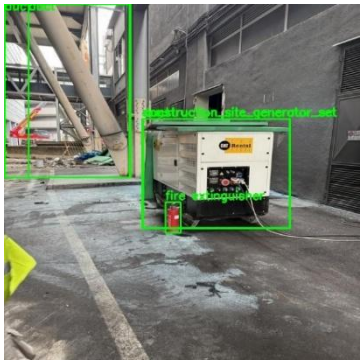
a) "Sprinkler" - Single Prompt
Failure: Target Missed detection & FPs

c) "Ladder" – Single Prompt
Success: Correct detection without FPs

e) "CS generator"- Single Prompt
Succuss: Correct detection without FPs

b) "Sprinkler", "Vent cover", and "lighting" – Multiple Prompt
Failure: Target missed detection for "sprinkler" but correct detections for other classes.

d) "Ladder" and "Red&white tape"- Multiple Prompt
Success: Correct detection for targets but increased FPs



f) "CS generator set", "Fire extinguisher", and "Duct" – Multiple Prompt
Hallucination: FP for absent object (Duct) but correct detections for present Targets

Figure 5: Qualitative detection examples showing representative successes and failure modes under different prompting strategies

categories, yielding numerous evaluation instances. Although this has enabled structured analysis of prompt behavior, some aspects of the evaluation have been limited, given the scope and space constraints. First, while the dataset captures both indoor and outdoor construction environments with diverse object categories and varying visual conditions to reflect real-world scenarios, the number of images per object category remains limited. Second, the evaluation is conducted using a single threshold for IoU. The IoU threshold was intentionally selected to better account for small and partially occluded objects common in construction scenes. However, it limits comparability with standard object detection benchmarks. Future work will address these limitations by expanding the dataset to include a larger and more balanced set of images across object categories. Additionally, subsequent studies will evaluate performance across multiple thresholds to enable more comprehensive and standardized comparisons. Building on the current findings, future research will also further investigate and refine prompting strategies to improve robustness across diverse construction environments.

Conclusions

This study presented an exploratory investigation of prompt robustness in OVOD for construction scenarios, focusing on lexical variation and prompt multiplicity using *GroundingDINO*. The results of this study indicate patterns rather than providing conclusive evidence and show that prompt formulation plays a relevant role in detection performance, influencing both recall–precision trade-offs and the occurrence of hallucinated predictions. In particular, simple base-term prompts provide the most stable and reliable results, while increasing linguistic complexity tends to reduce detection performance. The analysis also shows that multiple-prompt configurations do not consistently improve performance and may introduce context-dependent effects. While some benefits are observed in structured indoor environments, more complex outdoor scenes are associated with reduced stability and increased false detections. Furthermore, the introduction of hallucinated object categories highlights a limitation of OVOD systems, as the model tends to produce false detections rather than null responses, affecting detection completeness and reliability. These findings suggest a gap between human semantic expectations and the model's grounding behavior, which may affect deployment in safety-critical construction applications. Overall, the study emphasizes the importance of careful prompt design in OVOD-based AEC workflows. Future work will extend the evaluation across multiple IoU thresholds and investigate more robust prompting strategies to improve reliability and reduce hallucination effects in real-world construction scenarios.

References

Adil, M., Lee, G., Gonzalez, V.A. and Mei, Q., 2025. Using vision language models for safety hazard identification in construction. *arXiv:2504.09083*.

Bianchi, L., Carrara, F., Messina, N., Gennaro, C. and Falchi, F., 2024. The devil is in the fine-grained details: Evaluating open-vocabulary object detectors for fine-grained understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 22520-22529). DOI: 10.1109/CVPR52733.2024.02125.

Bravo, M.A., Mittal, S., Ging, S. and Brox, T., 2023. Open-vocabulary attribute detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7041-7050). DOI: 10.1109/CVPR52729.2023.00680.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. and Zagoruyko, S., 2020, August. End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213-229). Cham: Springer International Publishing. DOI: 10.1007/978-3-030-58452-8_13.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2019, June. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1* (pp. 4171-4186). *arXiv:1810.04805.*

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

Eggert, C., Zecha, D., Brehm, S. and Lienhart, R., 2017, June. Improving small object proposals for company logo detection. In *Proceedings of the 2017 ACM on international conference on multimedia retrieval* (pp. 167-174). DOI: 10.1145/3078971.3078990

Gu, X., Lin, T.Y., Kuo, W. and Cui, Y., 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv:2104.13921*.

Gu, J., Han, Z., Chen, S., Beirami, A., He, B., Zhang, G., Liao, R., Qin, Y., Tresp, V. and Torr, P., 2023. A systematic survey of prompt engineering on vision-language foundation models. *arXiv:2307.12980*.

Hugging Face (2025) *Zero-shot object detection*, Hugging Face Transformers Documentation, version 5.4.0. Available at: https://huggingface.co/docs/transformers/v5.4.0/en/tasks/zero_shot_object_detection (Accessed: 31 March 2026).

Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N. and Chang, K.W., 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10965-10975). DOI:10.1109/CVPR52688.2022.01069.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L., 2014, September. Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740-755). Cham: Springer International Publishing. DOI:/10.1007/978-3-319-10602-1_48.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H. and Zhu, J., 2024, September. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision* (pp. 38-55). Cham: Springer Nature Switzerland. DOI: 10.1007/978-3-031-72970-6_3.

Malaikrisanachalee, S., Wongwai, N. and Kowcharoen, E., 2025. ESPCN-YOLO: A high-accuracy framework for personal protective equipment detection under low-light and small object conditions. *Buildings*, *15*(10), p.1609. DOI: 10.3390/BUILDINGS15101609.

Minderer, M., Gritsenko, A. and Houlsby, N., 2023. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, *36*, pp.72983-73007.a*rXiv: 2306.09683*.

Mütze, A., Ilyas, S., Dörpelkus, C. and Rottmann, M., 2026. Can We Challenge Open-Vocabulary Object Detectors with Generated Content in Street Scenes?. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 740-750). DOI: 10.48550/arXiv.2506.23751.

Nebbia, G. and Kovashka, A., 2024, June. Synonym relations affect object detection learned on vision-language data. In *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 3770-3776). DOI:/10.18653/v1/2024.findings-naacl.239.

Open Images Dataset V7 (s.d.) Open Images Dataset V7 and Extensions. https://storage.googleapis.com/openimages/web/index.html (Accessed: 5 February 2026).

Raoofi, H., Sabahnia, A., Barbeau, D. and Motamedi, A., 2024. Deep learning method to detect missing welds for joist assembly line. *Applied System Innovation*, *7*(1), p.16. DOI: 10.3390/asi7010016.

Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J. and Sun, J., 2019. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8430-8439). DOI: 10.1109/ICCV.2019.00852.

Taherian, G. et al. (2026), "Vision–Language Models in Construction: Evaluating Open-Vocabulary Object Detection under Complex Scenarios," Proceedings of the Canadian Society for Civil Engineering (CSCE) Annual Conference 2026, Québec City, QC, Canada, June 3–5, 2026, to be published.