International Journal of Advanced Scientific Research and Engineering Trends

ISSN: 2456-0774

Volume 10 Issue 8 (2026) | Pages: 1-18

A Survey on Predicting the Locations of Unseen Objects in Household Environments Using Vision-Language Models

Nirmala Devi A C[1], C. Rangaswamy[2]

[1]Research center, Department of Electronics and communication Engineering, SJC Institute of Technology, Chikkaballapura, affiliated to VTU belagavi, India

[1]Department of Electronics and communication Engineering, SJC Institute of Technology, Chikkaballapura, India

[2]Department of Electronics and communication Engineering, SJC Institute of Technology, Chikkaballapura, affiliated to VTU belagavi, India.

[1]nirmalajagannath79@gmail.com ,[2]crsecesait@gmail.com

Peer Review Information

***Type**: Article*
***Received**: 11 May 2026*
***Revised**: 07 June 2026*
***Accepted**: 09 July 2026*
***Published**: 03 August 2026*

Abstract

Locating a household object that is not currently in view is a different and harder problem than detecting one that is. A useful domestic robot or assistive system must predict where a misplaced remote, medication box, or set of keys is likely to be, drawing on semantic priors, memory of past observations, and knowledge of how objects move through a home over time. This review surveys the emerging body of work on AI-driven predictive household object localization, with particular attention to systems that combine vision-language models (VLMs) with heuristic and commonsense reasoning. We first formalize the predictive localization problem and distinguish it from object detection, object goal navigation, and semantic search, organizing the literature along three axes: known-object retrieval, unseen-object grounding, and temporal prediction. We then examine the two technical pillars of current frameworks, open-vocabulary perception built on models such as CLIP, Grounding DINO, and multimodal large language models, and reasoning modules that convert commonsense knowledge into ranked location predictions. A consolidated account of datasets, simulators, benchmarks, and evaluation metrics follows, including where present evaluation practice fails to test prediction of out-of-view, relocated objects. We review applications in domestic service robotics, assistive technology for elderly and memory-impaired users, and smart-home systems, and close with a research agenda covering personalization, temporal reasoning, uncertainty calibration, privacy, and deployment on resource-limited hardware.

Keywords: Ambient Assisted Living; Deep Learning; Egocentric Vision; Episodic Memory; Heuristic Reasoning; Multimodal Learning; Object Detection; Object Localization; Predictive Localization; Smart Homes; Vision-Language Models.

How to Cite This Article

Devi, N. A. C., & Rangaswamy, C. (2026). A survey on predicting the locations of unseen objects in household environments using vision-language models. *International Journal of Advanced Scientific Research and Engineering Trends,* 10(8), 1–18.

Introduction

Finding objects in a home is one of those tasks that people perform without conscious effort and machines still find genuinely hard. A person asked to fetch the television remote does not scan every surface in the house, they check the sofa, the coffee table, and perhaps the bedroom, in roughly that order, because experience has taught them where remotes tend to end up. Reproducing this ability in artificial systems has become a central problem for domestic service robots, assistive technologies for elderly users, and smart-home applications, and it now sits at the meeting point of computer vision, natural language processing, and embodied AI research [1], [2]. The economic and social motivation is considerable. Populations in most industrialized countries are ageing, and misplaced items such as spectacles, keys, medication, and phones are a recurring source of frustration and, for users with memory impairment, a genuine safety concern. A system that can predict where an object is likely to be, rather than merely recognize it when it happens to be in view, would change what household robots and ambient assistive systems can actually do for people.

The distinction between prediction and detection deserves emphasis because it defines the scope of this review. Object detection, in its conventional formulation, answers the question of what is present in the current image and where it appears in pixel coordinates. Predictive localization answers a different question: given that the target object is not currently visible, where in the environment is it most likely to be found? Answering this second question requires capabilities that detection pipelines do not possess on their own, including semantic knowledge about typical object placements, memory of past observations, models of how objects move through a household over time, and a way to rank candidate locations under uncertainty [2], [26]. Recent work has begun to combine large vision-language models (VLMs), which supply open-vocabulary perception and a surprising amount of commonsense knowledge, with heuristic and probabilistic reasoning modules that turn this knowledge into ranked location predictions [18], [19], [31]. The pace of this work has accelerated sharply since 2023, and no existing survey treats predictive household object localization as its central subject, which is the gap this review addresses.

The contributions of this review are fourfold. First, it formalizes predictive object localization and situates it relative to neighbouring tasks such as object goal navigation and semantic search. Second, it surveys the two technical pillars of current systems, namely vision-language perception and heuristic or common sense reasoning, and examines how published frameworks combine them. Third, it consolidates the datasets, benchmarks, and metrics used across the field and identifies where evaluation practice falls short of the predictive setting. Fourth, it distils open challenges into a research agenda covering privacy, personalization, temporal reasoning, and deployment on resource-limited hardware.

Review on existing Paper

Kai Zhao et al., [37] proposed a novel framework called Cascaded Open-vocabulary Camouflaged Understanding Network to improve open-vocabulary camouflaged object segmentation (OVCOS). The methodology follows a two-stage architecture consisting of segmentation and classification. In the first stage, the authors integrate a fine-tuned CLIP vision-language model with the Segment Anything Model**,** where CLIP-derived visual and textual embeddings are used as prompts to guide SAM toward accurate localization of camouflaged objects. The SAM decoder is further enhanced using Conditional Multi-Way Attention and an Edge Enhancement module to improve boundary detection. In the second stage, instead of cropping segmented regions, the predicted segmentation mask is used as **a** soft spatial prior through an alpha channel, enabling CLIP to preserve the complete image context while performing open-vocabulary classification. The framework employs multi-modal prompt tuning, cross-modal semantic alignment, BCE, IoU, and Dice loss functions during training. Experimental evaluation was conducted on the OVCamo**,** CAMO**,** COD10K, and NC4K datasets. The proposed method achieved state-of-the-art performance with cSm = 0.668**,** cFwβ = 0.615**,** cIoU = 0.568, and cMAE = 0.265 on OVCamo, outperforming the baseline OVCoser by significant margins. The adapted SAM also ranked first on most evaluation metrics across conventional camouflaged object segmentation benchmarks, demonstrating excellent segmentation accuracy, boundary refinement, robustness, and cross-domain generalization.

A.-Ș. Bulzan et al., [38] presented a comprehensive study on integrating Vision-Language Models into object detection to improve semantic understanding and detection accuracy. The proposed methodology investigates how pre-trained VLMs, particularly those trained on large-scale image-text datasets, can overcome the limitations of conventional object detectors that rely solely on visual features. The authors analyze different strategies for incorporating language information into object detection pipelines, including prompt engineering, feature alignment, cross-modal attention mechanisms, and multimodal feature fusion. They compare the effectiveness of zero-shot, few-shot, and fine-tuning approaches for recognizing both seen and unseen object categories. From a technical perspective, the paper discusses the role of CLIP-based visual and textual embeddings, transformer-based architectures, and contrastive learning techniques in enhancing object localization and classification The study concludes that integrating multimodal knowledge with modern detection frameworks offers a promising direction for next-generation intelligent object detection systems, particularly in applications requiring open-vocabulary recognition, real-world scene understanding, and scalable deployment across diverse visual domains.

Haoran Hou et al., [39] proposed a spatial multimodal knowledge-driven framework for 3D Scene Graph Generation that exploits the inherent hierarchical structure of 3D environments to improve relationship prediction between objects. The methodology first employs a vision-language model with a cross-modal tuning strategy to infer support relationships among objects using limited supervision. These support relationships are then used to construct two complementary representations: a hierarchical visual graph, which captures contextual visual features, and a hierarchical symbolic knowledge graph**,** which models semantic relationships extracted from textual knowledge. To integrate these modalities, the authors design a Graph Reasoning Network that progressively accumulates spatial multimodal knowledge by correlating visual context with symbolic information. Additionally, they introduce a new benchmark dataset, 3DSSG-M, derived from the original 3DSSG dataset by balancing predicate distributions and reducing frequency bias, thereby enabling a fairer evaluation of scene graph generation methods. Extensive experiments and ablation studies demonstrate that the proposed framework consistently outperforms existing state-of-the-art approaches in object relationship prediction.

L. He et al**.,** [40] proposed a vision–language model (VLM)-empowered framework to improve nighttime object detection, where poor illumination, low contrast, and image noise significantly reduce detection accuracy. The proposed methodology incorporates semantic knowledge from a pre-trained vision–language model to enhance feature representation while introducing two novel modules: a Consistency Sampler and a Hallucination Feature Generator. Extensive experiments conducted on benchmark nighttime object detection datasets demonstrate that the proposed approach consistently outperforms existing state-of-the-art methods in detection accuracy while maintaining strong robustness across diverse nighttime environments. The results confirm that leveraging vision–language semantic priors together with consistency-guided sampling and feature hallucination effectively enhances object detection under adverse lighting conditions, making the framework suitable for autonomous driving, intelligent surveillance, and other low-light vision applications.

Sheng Jin et al., [41] presented a comprehensive survey of Vision-Language Models for visual recognition tasks, highlighting their evolution from conventional deep learning models toward large-scale multimodal pre-training. The survey systematically reviews the methodology adopted in modern VLMs, including contrastive vision-language pre-training**,** image-text alignment, prompt learning, transfer learning, and knowledge distillation techniques. The authors analyze widely used VLM architectures, pre-training objectives, benchmark datasets, and downstream applications such as image classification, object detection, semantic segmentation, and optical character recognition. They explain how models like CLIP learn joint visual-text embeddings from massive image-text pairs, enabling zero-shot prediction without task-specific fine-tuning. The paper also categorizes recent research into pre-training methods, transfer learning strategies such as prompt tuning and visual adaptation, and knowledge distillation approaches for lightweight deployment. Furthermore, the survey compares the performance of existing VLMs across multiple public benchmarks and discusses their strengths and limitations. The authors identify several research challenges, including improving multimodal alignment, reducing computational complexity, enhancing generalization, addressing data quality issues, and developing more efficient adaptation techniques. Overall, the survey provides a detailed technical foundation and comprehensive overview of VLM research, serving as a valuable reference for future studies in multimodal learning and computer vision.

J. Chen et al., [42] proposed a unified framework for Human–Object Interaction Detection that integrates modern detection architectures with vision–language foundation models to improve interaction recognition accuracy and computational efficiency. The methodology introduces two progressively enhanced frameworks, SOV-STG-VLA and Hybrid-SOV. The SOV-STG-VLA framework reformulates HOI detection as a Subject–Object–Verb decoding task, enabling the model to jointly predict humans, objects, and their interactions. It further incorporates a Split Target Guided denoising strategy to accelerate convergence and improve structural consistency during training. The framework utilizes DINO-v3 foundation model features to obtain richer visual representations and stronger semantic discrimination. Extensive experiments conducted on the HICO-DET and V-COCO benchmark datasets demonstrate that the proposed methods achieve state-of-the-art detection accuracy with faster convergence and improved inference efficiency compared with existing HOID approaches. The results confirm that combining advanced detection architectures with vision–language foundation models significantly enhances human–object interaction recognition while providing a scalable and effective solution for real-world scene understanding applications.

H. Chang et al., [43] proposed CT-FSOD, a CLIP-based text-guided Few-Shot Object Detection framework that enhances the detection of novel object classes using only a few annotated samples. The proposed methodology addresses the limitations of conventional FSOD methods, which depend heavily on visual prototypes extracted from limited support images and are sensitive to background clutter, pose variations, and intra-class diversity. CT-FSOD integrates the semantic knowledge of the pre-trained Contrastive Language–Image Pretraining model through two complementary branches. The Prototype Distillation Branch employs a frozen CLIP image encoder to generate robust, semantically aligned visual prototypes from support images, while the Text Matching Branch constructs initial class prototypes using CLIP text embeddings generated with CoOp-style learnable context tokens, providing pose-invariant semantic guidance. Query image features extracted by a trainable CNN backbone are jointly aligned with both visual and textual CLIP representations using prototype distillation and a CLIP Alignment Loss, enabling the detector to learn highly discriminative and generalized feature

representations. Extensive experiments on the PASCAL VOC and MS COCO few-shot object detection benchmarks demonstrate that CT-FSOD consistently surpasses existing state-of-the-art methods across multiple few-shot settings. The proposed framework significantly improves novel-class detection accuracy without increasing inference complexity, highlighting the effectiveness of combining vision–language semantic knowledge with few-shot learning for robust and generalized object detection.

J. Lu et al., [44] proposed Vision–Language Model-Supervised Domain Adaptation, a novel framework for cross-domain object detection in remote sensing that addresses the performance degradation caused by variations in sensor type, imaging modality, resolution, and viewing geometry. Unlike conventional domain-adaptive object detection methods that rely solely on noisy target-domain features for feature alignment, the proposed methodology incorporates a frozen Vision–Language Model as a stable semantic reference during domain adaptation. The framework introduces two key technical modules: the Vision–Language Model-Supervised Prototypical Alignment module and the Global Cross-Domain Contrastive Alignment module. VLPA performs a tri-domain adversarial alignment by simultaneously aligning feature distributions among the source domain, target domain, and VLM semantic space, thereby reducing feature inconsistency caused by domain shifts. Meanwhile, GCCA applies supervised contrastive learning to improve intra-class compactness and inter-class separability, resulting in more discriminative feature representations. These results confirm that integrating stable vision–language semantic knowledge significantly enhances robustness and generalization for cross-domain remote sensing object detection.

F. Liu et al., [45] proposed RemoteCLIP**,** the first vision–language foundation model specifically designed for remote sensing**,** to overcome the limitations of conventional self-supervised and masked image modeling approaches that lack semantic understanding and require task-specific fine-tuning. The proposed methodology employs Contrastive Language–Image Pretraining to jointly learn visual and textual representations from large-scale remote sensing image-caption pairs. To address the scarcity of annotated remote sensing data, the authors introduce a data scaling strategy that converts heterogeneous annotations into unified image-caption pairs using Box-to-Caption and Mask-to-Box conversion techniques. Furthermore, UAV imagery is incorporated to construct a pretraining dataset that is 12 times larger than the combination of previously available datasets. Technically, RemoteCLIP learns aligned image and text embeddings through contrastive learning, enabling zero-shot image classification, few-shot learning, k-nearest neighbor classification, linear probing, image–text retrieval, and object counting without extensive task-specific adaptation. The model was evaluated on 16 remote sensing benchmark datasets, including the newly introduced RemoteCount dataset for object counting. Experimental results demonstrate that RemoteCLIP consistently outperforms existing foundation models across different model scales, achieving 9.14% higher mean recall on the RSITMD dataset, 8.92% improvement on the RSICD dataset for image–text retrieval, and up to 6.39% higher average accuracy than the original CLIP model on 12 zero-shot classification benchmarks. These outcomes demonstrate that RemoteCLIP provides superior semantic representation, robust generalization, and excellent performance across diverse remote sensing applications.

Z. Wei et al., [46] proposed a Rewriting-driven Augmentation framework to address the problem of limited annotated data in Vision–Language Navigation by generating diverse observation–instruction pairs without relying on additional simulators or web-collected data. The proposed methodology employs a combination of Vision–Language Models, Large Language Models, and Text-to-Image Generation Models to create realistic and semantically consistent training samples. First, an Object-Enriched Observation Rewriting module uses VLMs and LLMs to rewrite scene descriptions by introducing diverse objects and spatial layouts, which are then synthesized into new observations using text-to-image generation models. Next, an Observation-Contrast Instruction Rewriting module generates new navigation instructions by reasoning about the differences between the original and rewritten observations, ensuring semantic alignment between images and textual instructions. Extensive experiments conducted on the Room-to-Room, REVERIE, Room-for-Room, and R2R-Continuous Environment benchmarks demonstrate that RAM consistently improves navigation success rates and generalization in unseen environments. The experimental results confirm that foundation model-driven observation and instruction rewriting effectively enhances multimodal learning, reduces dependency on simulator-generated data, and significantly improves the robustness of vision–language navigation systems.

J. Ke et al., [47] proposed VLDadaptor, a Vision–Language Model -based domain adaptive object detection framework designed to mitigate performance degradation caused by domain shifts between labeled source-domain data and unlabeled target-domain data. The proposed methodology leverages the semantic knowledge embedded in a pre-trained CLIP vision–language model to guide domain adaptation through a knowledge distillation strategy. Instead of directly aligning source and target feature distributions, VLDadaptor transfers rich cross-modal semantic information from the frozen VLM to the object detector, enabling the detector to learn domain-invariant and semantically discriminative representations. Technically, the framework introduces a Vision–Language Distillation module that aligns visual features extracted by the detector with CLIP-generated image and text embeddings The results show significant improvements in mean Average Precision, enhanced robustness under varying environmental conditions, and superior generalization to unseen target domains. Overall, the proposed framework confirms that integrating vision–language semantic knowledge through distillation effectively bridges domain gaps and improves cross-domain object detection performance.

W. Zhai et al., [48] proposed the Vision–Language Prior Guidance Network for zero-shot object counting**,** aiming to accurately estimate the number of objects belonging to unseen categories without requiring category-specific training samples. The proposed methodology leverages the powerful semantic knowledge of pre-trained vision–language models to bridge the gap between visual features and textual descriptions. Specifically, the framework utilizes language priors to guide the counting process by aligning image features with semantic embeddings of object categories. The model is trained using a combination of density estimation and contrastive learning objectives, allowing it to generalize well across diverse object categories while reducing dependence on annotated counting datasets. Extensive experiments conducted on several benchmark object counting datasets demonstrate that the proposed framework consistently outperforms existing zero-shot and few-shot counting approaches. The experimental results show significant improvements in counting accuracy, lower Mean Absolute Error (MAE), and better generalization to novel object classes. Overall, VLPGNet demonstrates that integrating vision–language priors with density-based counting provides an effective and scalable solution for zero-shot object counting, making it suitable for real-world applications such as crowd analysis, ecological monitoring, and intelligent surveillance.

X. Yang et al., [49] proposed an end-to-end Vision–Language Pretraining framework incorporating a Semantic Visual Loss to improve multimodal representation learning for vision–language tasks. The methodology addresses the limitations of conventional VLP approaches, where visual feature extraction and multimodal learning are often performed separately, preventing effective optimization of the visual encoder during pretraining. The proposed framework enables joint optimization of the visual and textual encoders by introducing the Semantic Visual Loss, which enhances semantic consistency between image representations and corresponding textual descriptions. Technically, the framework employs a transformer-based vision–language architecture that learns aligned image–text embeddings through multimodal contrastive learning. The proposed method consistently outperformed existing end-to-end vision–language pretraining approaches by achieving higher retrieval accuracy, stronger semantic alignment, and improved downstream task performance. The results demonstrate that incorporating Semantic Visual Loss significantly enhances the quality of multimodal feature representations while enabling efficient end-to-end optimization, making the framework suitable for a wide range of vision–language applications requiring robust semantic understanding.

P. Zhang et al.,[50] proposed a Language-Guided Object Localization framework with a Refined Spotting Enhancement module to accurately localize objects in remote sensing imagery using natural language descriptions. The methodology addresses the challenges of language-guided object localization in aerial images, including complex backgrounds, large-scale variations, densely distributed objects, and weak semantic correspondence between textual queries and visual features. The proposed framework first extracts multimodal representations using a vision–language architecture that jointly encodes remote sensing images and textual descriptions.The framework is optimized using localization and multimodal alignment losses to improve both spatial precision and semantic consistency. Extensive experiments conducted on benchmark remote sensing language grounding datasets demonstrate that the proposed approach consistently outperforms existing state-of-the-art methods in localization accuracy and robustness. The experimental results show significant improvements in object spotting precision, intersection-over-union (IoU), and grounding performance, particularly in challenging scenarios involving small objects, cluttered scenes, and complex spatial relationships. Overall, the proposed method effectively combines language guidance with refined visual feature enhancement, providing a robust solution for natural language-based object localization in remote sensing applications.

H. Lin et al., [51] proposed RS-MoE**,** a Vision–Language Model based on a Mixture of Experts architecture to improve remote sensing image captioning and remote sensing visual question answering**.** The methodology addresses the limitations of existing vision–language models, which often fail to capture the diverse semantic characteristics of remote sensing imagery due to complex land-cover patterns, varying object scales, and limited domain-specific annotations. The proposed framework incorporates a Mixture of Experts mechanism that dynamically activates specialized expert networks according to the input image and language query, enabling adaptive learning of multimodal representations. The visual encoder extracts discriminative spatial features from remote sensing images, while the language encoder processes textual descriptions or questions. A cross-modal fusion module integrates visual and textual embeddings through attention mechanisms, allowing efficient interaction between the two modalities. During training, the expert routing strategy selectively engages the most relevant experts, improving model capacity without significantly increasing computational cost. The results confirm that combining the Mixture of Experts architecture with vision–language learning effectively enhances semantic understanding and multimodal reasoning, making RS-MoE a robust solution for advanced remote sensing image interpretation tasks.

R. Zhang et al., [52] proposed Det-Agent**,** an open-vocabulary object localization and detection framework that integrates reinforcement learning with vision–language models to improve the detection of both seen and unseen object categories. The proposed methodology addresses the limitations of conventional open-vocabulary object detectors, which often rely on exhaustive region proposals and static feature matching, resulting in high computational cost and inaccurate localization. Instead, Det-Agent formulates object localization as a sequential decision-making process**,** where a reinforcement learning agent progressively adjusts the object bounding box based on interactions with the visual environment. The experimental results show improved generalization to unseen object categories, reduced

computational overhead through adaptive search, and enhanced robustness in complex scenes. Overall, the proposed framework demonstrates that combining reinforcement learning with vision–language semantic guidance provides an effective and scalable solution for open-vocabulary object localization and detection.

S. Yuan et al., [53] proposed SPENav**, a** Vision–Language Navigation framework that enhances navigation performance through Dynamic Object Filtering and Spatial Perception Enhancement**.** The methodology addresses a major challenge in VLN, where navigation agents are often distracted by irrelevant objects and fail to effectively capture spatial relationships between landmarks and navigation instructions. The proposed framework first employs a Dynamic Object Filtering module that identifies and suppresses non-essential objects in the visual scene while preserving objects relevant to the navigation instruction. This adaptive filtering mechanism enables the agent to focus on informative visual cues and reduces semantic noise during navigation. To further improve spatial reasoning, the framework introduces a Spatial Perception Enhancement module that explicitly models spatial relationships between detected objects and textual instructions using cross-modal attention and spatial feature encoding. The experimental results indicate that dynamic object filtering effectively reduces visual distractions, while enhanced spatial perception significantly improves landmark understanding and navigation accuracy. Overall, SPENav provides a robust and efficient framework for intelligent embodied navigation by combining adaptive object selection with spatially aware vision–language reasoning.

H. Zhang et al., [54] proposed a One-Stream Vision–Language Memory Network for object tracking, aiming to improve target localization by jointly exploiting visual appearance and language descriptions within a unified framework. The methodology addresses the limitations of conventional object tracking approaches that rely solely on visual information and often struggle under occlusion, illumination changes, background clutter, and target deformation. Instead of processing visual and textual information through separate branches, the proposed framework adopts a one-stream vision–language architecture that simultaneously learns multimodal representations from image features and natural language descriptions of the target. The model is trained end-to-end using joint tracking and multimodal alignment objectives to optimize localization accuracy and feature consistency. Extensive experiments conducted on standard object tracking benchmarks demonstrate that the proposed approach consistently outperforms existing state-of-the-art trackers in terms of tracking precision, success rate, and robustness under challenging conditions such as occlusion, fast motion, scale variation, and background interference. The results confirm that integrating vision–language learning with a memory-based architecture significantly enhances long-term object tracking performance and improves generalization across diverse tracking scenarios.

S. Sadhwani et al., [55] proposed a Vision–Language Model based framework for the real-time detection of mixed-critical events, enabling intelligent systems to identify and prioritize multiple events with different severity levels in dynamic environments. The methodology addresses the limitations of traditional vision-based event detection systems, which primarily rely on object recognition and often fail to capture the semantic relationships required to distinguish critical from non-critical events. The proposed framework integrates a pre-trained Vision–Language Model with a real-time event detection pipeline to jointly analyze visual content and natural language descriptions of critical scenarios. A cross-modal semantic alignment mechanism is employed to associate visual observations with textual event prompts, allowing the model to recognize both known and previously unseen events through semantic reasoning. The framework further incorporates a priority-aware decision module that classifies detected events according to their level of criticality, enabling rapid response in time-sensitive situations. The results show that leveraging vision–language semantic knowledge significantly enhances the recognition of complex mixed-critical events, making the framework suitable for applications such as intelligent surveillance, smart cities, autonomous systems, disaster monitoring, and public safety.

J. Yang et al., [56] proposed Zone-YOLO, a vision–language object detection framework that enhances object detection performance by incorporating Zone Prompts into the YOLO detection architecture. The methodology addresses the limitations of conventional object detectors, which mainly rely on global visual features and often fail to accurately detect small, occluded, or densely distributed objects in complex traffic environments. The proposed framework introduces Zone Prompts, which encode region-specific semantic information and spatial context as language-guided cues to direct the detector toward areas of interest. Extensive experiments conducted on intelligent transportation and object detection benchmark datasets demonstrate that Zone-YOLO consistently outperforms existing YOLO-based and vision–language detection methods in terms of mean Average Precision, localization accuracy, and robustness under challenging conditions such as occlusion, varying illumination, and crowded scenes. The results confirm that integrating zone-specific language guidance with efficient object detection significantly enhances semantic understanding while maintaining real-time inference performance, making Zone-YOLO highly suitable for autonomous driving and intelligent transportation systems.

Y. Ma et al., [57] proposed Highlighting Fine-Grained Attributes for Open-Vocabulary Object Detection, a vision–language framework designed to improve the recognition and localization of unseen object categories by explicitly modeling fine-grained semantic attributes. The methodology addresses a major limitation of conventional Open-Vocabulary Object Detection methods, where object representations are typically derived from coarse category names, making it difficult to distinguish visually similar classes. The proposed framework introduces an Explicit Linear Composition strategy that decomposes object semantics into multiple fine-grained attributes, such as color,

shape, texture, and functional characteristics. These attributes are encoded using a pre-trained vision–language model and linearly combined to generate more discriminative text embeddings for object categories. In addition, the framework incorporates an Attribute Highlighting module that strengthens the correspondence between visual features and attribute-aware textual representations through cross-modal alignment and attention mechanisms. The experimental results show that explicit fine-grained attribute composition significantly enhances semantic discrimination, improves localization accuracy, and increases the robustness of open-vocabulary detection. Overall, the proposed framework demonstrates the effectiveness of integrating attribute-aware language representations with vision–language learning for scalable and accurate open-vocabulary object detection.

P. P. Ninh et al., [58] proposed a Lightweight Source-Free Unsupervised Domain Adaptation framework for object detection using stereo vision, aiming to improve detector performance in target domains without requiring access to labeled source-domain data. The methodology addresses two key challenges of conventional domain adaptation methods: dependence on source-domain datasets during adaptation and the high computational cost of existing adaptation frameworks. The proposed approach utilizes stereo vision to exploit geometric consistency between left and right images, enabling the model to learn robust domain-invariant features directly from unlabeled target-domain data. The detector is optimized through a combination of detection loss and consistency constraints without relying on source-domain supervision. Extensive experiments on benchmark stereo object detection datasets demonstrate that the proposed framework achieves superior adaptation performance compared with existing source-free domain adaptation methods while requiring fewer computational resources. The experimental results show improved mean Average Precision, enhanced robustness under varying environmental conditions, and faster inference. Overall, the proposed lightweight stereo-based SFUDA framework provides an efficient and practical solution for cross-domain object detection in autonomous driving, robotics, and intelligent transportation systems.

Y. Zhu et al., [59] proposed ChatNav**,** a Large Language Model -driven framework for zero-shot object navigation, enabling embodied agents to navigate toward unseen target objects without task-specific training. The methodology addresses the limitations of conventional object navigation approaches, which rely heavily on large annotated navigation datasets and exhibit poor generalization to novel environments. ChatNav exploits the semantic reasoning capability of Large Language Models to infer navigation strategies from natural language descriptions and scene semantics. The framework first converts visual observations into structured semantic information using object detection and scene understanding modules. These semantic observations, together with navigation goals, are provided as prompts to the LLM, which performs high-level reasoning to generate navigation decisions and intermediate action plans. To further improve navigation efficiency, ChatNav incorporates semantic memory that stores previously explored regions and recognized objects, reducing redundant exploration and enabling more informed decision-making. Technically, the framework combines vision-language perception**,** prompt engineering**,** semantic reasoning**,** and memory-based planning within a unified navigation pipeline. The results indicate that integrating large language models with semantic scene understanding significantly enhances navigation efficiency, reasoning capability, and adaptability. Overall, ChatNav demonstrates the effectiveness of LLM-based semantic reasoning for robust and scalable zero-shot object navigation in embodied artificial intelligence.

A. Rasekh et al., [60] proposed an explainable CLIP-based object recognition framework that simultaneously improves classification accuracy and model interpretability by modeling the joint probability of visual rationales and object categories**.** The methodology addresses a major limitation of conventional Vision–Language Models (VLMs), where CLIP provides strong recognition performance but offers limited insight into the reasoning behind its predictions. Instead of predicting object categories directly, the proposed framework introduces an intermediate rationale generation stage, where the model identifies human-understandable semantic attributes or explanations that justify the predicted category. The model is optimized using a joint learning objective that balances recognition accuracy with rationale consistency, producing predictions that are both reliable and explainable. Extensive experiments conducted on benchmark object recognition datasets demonstrate that the proposed approach consistently outperforms conventional CLIP-based methods in classification performance while generating meaningful and human-interpretable explanations. The experimental results indicate improvements in classification accuracy, explanation fidelity, and semantic consistency without introducing significant computational overhead. Overall, the proposed framework effectively bridges the gap between explainability and recognition performance, making CLIP-based object recognition more transparent and trustworthy for practical applications such as autonomous systems, healthcare, and intelligent surveillance.

M. Zhang et al., [61] proposed a Hierarchical Semantic Knowledge-Based Object Search method to improve object search efficiency for household service robots operating in complex indoor environments. The methodology addresses the limitations of traditional object search techniques, which rely primarily on exhaustive exploration or low-level visual cues without considering semantic relationships between objects and their surroundings. The proposed framework constructs a hierarchical semantic knowledge model that organizes environmental information into multiple levels, including room-level, furniture-level, and object-level semantic relationships. By exploiting prior knowledge about the typical locations and co-occurrence of household objects, the robot predicts the most probable search regions before initiating visual exploration. Technically, the framework integrates semantic knowledge representation**,** hierarchical

reasoning, and probabilistic search planning to guide the robot toward likely object locations. The results indicate that incorporating hierarchical semantic knowledge enables robots to perform more intelligent and efficient object localization in household environments. Overall, the proposed framework provides an effective solution for autonomous service robots by combining semantic reasoning with hierarchical search planning, thereby enhancing object search performance and practical applicability in real-world domestic scenarios.

L. Zhu et al**.,** [62] proposed a Background-Aware Classification Activation Map framework to improve Weakly Supervised Object Localization by explicitly modeling background information during object localization. The methodology addresses a common limitation of conventional Class Activation Map (CAM)-based methods, which often focus only on the most discriminative object regions while ignoring object boundaries and surrounding contextual information. This results in incomplete localization and inaccurate bounding boxes. The proposed BACAM framework introduces a background-aware learning strategy that simultaneously learns foreground and background representations during training The background-aware objective encourages the network to capture complete object regions while minimizing background interference, leading to more accurate localization maps. Extensive experiments conducted on benchmark WSOL datasets, including CUB-200-2011**,** ILSVRC**,** and OpenImages, demonstrate that BACAM consistently outperforms existing state-of-the-art weakly supervised localization methods. The proposed approach achieves significant improvements in localization accuracy, Top-1 localization error, and object coverage while maintaining competitive image classification performance. The results confirm that explicitly incorporating background semantics into activation map generation substantially enhances weakly supervised object localization, making the framework suitable for annotation-efficient object detection and image understanding applications.

M. Meng et al., [63] proposed a Diverse Complementary Part Mining framework to improve Weakly Supervised Object Localization by discovering multiple complementary object regions instead of relying only on the most discriminative parts. The methodology addresses the inherent limitation of conventional Class Activation Map (CAM)-based localization methods, which usually highlight only small, highly discriminative regions, resulting in incomplete object localization. The proposed framework introduces a complementary part mining strategy that progressively identifies diverse object regions through an iterative learning process. Initially, the network detects the most salient object parts, and these regions are subsequently suppressed to encourage the model to discover additional complementary regions that collectively represent the entire object.. The proposed approach achieves significant improvements in localization accuracy, Top-1 localization error, and object coverage while maintaining competitive image classification performance. The results confirm that mining diverse complementary object parts enables more complete object representation and substantially enhances weakly supervised object localization.

C. Tan et al., [64] proposed a Dual-Gradients Localization Framework with Skip-Layer Connections to improve Weakly Supervised Object Localization using only image-level labels. The methodology addresses the limitation of conventional gradient-based localization methods, which often focus on highly discriminative object regions while ignoring complete object structures, leading to inaccurate localization. The proposed framework introduces a dual-gradient learning mechanism that combines gradient information from both classification and localization objectives to generate more comprehensive and accurate activation maps. By jointly exploiting these complementary gradients, the network captures both discriminative semantic features and fine spatial details of target objects. Furthermore, the framework incorporates skip-layer connections, which fuse low-level spatial information with high-level semantic representations across multiple convolutional layers. This multi-scale feature fusion preserves object boundaries and enhances localization precision, particularly for small or partially occluded objects. Technically, the framework integrates gradient-based activation mapping**,** multi-scale feature aggregation, and end-to-end optimization to improve object localization without requiring bounding-box annotations. Extensive experiments conducted on benchmark datasets, including CUB-200-2011**,** ILSVRC, and OpenImages, demonstrate that the proposed approach consistently outperforms existing state-of-the-art weakly supervised localization methods. The experimental results show significant improvements in Top-1 localization accuracy, object coverage, and classification performance. The results confirm that combining dual-gradient supervision with skip-layer feature fusion enables the network to localize complete object regions more accurately while maintaining efficient end-to-end training, making the framework highly effective for weakly supervised object localization tasks.

J. Bai et al., [65] proposed a Self-Directed Weakly Supervised Object Localization framework for remote sensing images, aiming to accurately localize objects using only image-level category labels. The methodology addresses the limitations of conventional weakly supervised localization methods, which typically focus only on the most discriminative object regions and struggle with the unique challenges of remote sensing imagery, such as small object sizes, complex backgrounds, dense object distributions, and varying spatial resolutions. The proposed framework introduces a self-directed localization strategy that progressively transforms image classification knowledge into object localization capability without requiring bounding-box annotations. Extensive experiments conducted on benchmark remote sensing datasets demonstrate that the proposed method consistently outperforms existing state-of-the-art weakly supervised object localization approaches. The experimental results show significant improvements in localization accuracy**,** Top-1 localization performance, and object coverage while maintaining competitive image classification accuracy. The results confirm that the

self-directed learning strategy effectively transfers classification knowledge into robust localization capability, providing an efficient annotation-free solution for remote sensing object localization and image interpretation tasks.

P. Zhou et al., [66] proposed Point Simultaneous Localization and Object Tracking, a real-time framework designed to enable autonomous robots to perform simultaneous localization and dynamic object tracking in changing environments. The methodology addresses the limitations of traditional Simultaneous Localization and Mapping systems, which generally assume static environments and experience degraded localization accuracy when dynamic objects are present. The proposed framework utilizes 3D point cloud data acquired from stereo vision or LiDAR sensors to jointly estimate the robot's pose while tracking moving objects in real time. Technically, PointSLOT introduces a dual-thread architecture consisting of a localization module and an object tracking module that operate collaboratively. The localization module estimates the robot's trajectory by extracting stable geometric features from static regions, while the tracking module detects, associates, and continuously tracks dynamic objects using point cloud clustering, motion estimation, and data association techniques. A dynamic environment filtering mechanism separates static and moving points, preventing dynamic objects from degrading localization performance. The framework further employs efficient optimization strategies to ensure low computational overhead and real-time execution. Extensive experiments conducted in both simulated and real-world dynamic environments demonstrate that PointSLOT achieves higher localization accuracy, more reliable object tracking, and lower trajectory error than existing SLAM-based tracking methods. The experimental results confirm that the proposed framework effectively handles moving objects while maintaining robust robot localization, making it suitable for autonomous driving, service robots, and intelligent mobile robotic applications operating in dynamic real-world environments.

M. Yan et al., [67] proposed a Multimodal Localization Distillation framework to improve 3D Single Object Tracking in intelligent transportation systems by effectively transferring localization knowledge from multiple sensing modalities into a lightweight tracking model. The methodology addresses the limitations of existing 3D object tracking approaches, which often rely on a single sensing modality and suffer performance degradation under occlusion, sparse point clouds, and adverse environmental conditions. The proposed framework employs a teacher–student knowledge distillation strategy, where a multimodal teacher network integrates complementary information from LiDAR point clouds, RGB images, and localization cues to generate accurate spatial representations. A lightweight student network is then trained to mimic the localization capability of the teacher through localization distillation loss, enabling efficient inference while preserving high tracking accuracy. Extensive experiments conducted on benchmark autonomous driving datasets demonstrate that the proposed method consistently outperforms existing state-of-the-art 3D single object tracking approaches. The experimental results show improvements in tracking precision**,** success rate**,** 3D localization accuracy, and robustness under challenging scenarios such as occlusion, viewpoint changes, and sparse sensor observations. Overall, the proposed multimodal localization distillation framework provides an effective balance between tracking accuracy and computational efficiency, making it well suited for real-time intelligent transportation and autonomous driving applications.

X. Li et al., [68] proposed LIO-LOT, a tightly-coupled framework that simultaneously performs LiDAR-Inertial Odometry and Multi-Object Tracking for intelligent transportation systems. The methodology addresses the limitations of conventional autonomous perception systems, where localization and object tracking are treated as independent tasks, leading to inconsistent motion estimation and reduced robustness in dynamic environments. The proposed framework tightly integrates LiDAR point clouds**,** Inertial Measurement Unit data, and dynamic object information into a unified optimization frameworkTechnically, the framework incorporates factor graph optimization**,** data association**,** motion estimation**,** and dynamic object filtering to achieve consistent localization and tracking. Extensive experiments on benchmark autonomous driving datasets demonstrate that LIO-LOT consistently outperforms state-of-the-art LiDAR odometry and multi-object tracking methods in terms of localization accuracy, tracking precision, and trajectory estimation. The experimental results show lower odometry drift, higher object tracking accuracy, and improved robustness in highly dynamic traffic scenarios. Overall, the proposed tightly-coupled architecture provides an efficient and reliable solution for simultaneous localization and multi-object tracking, making it highly suitable for autonomous driving and intelligent transportation applications.

M. Cui et al., [69] proposed Dense Object Prediction Network, a unified deep learning framework for simultaneous multiclass object counting and localization in remote sensing images. The methodology addresses the limitations of traditional object counting methods, which either estimate object density without precise localization or rely on object detection techniques that perform poorly in densely populated scenes. The proposed framework formulates counting and localization as a joint learning task by predicting dense object representations for multiple object categories. Technically, DOPNet employs a fully convolutional neural network to generate class-specific dense prediction maps**,** where each object is represented by a probability peak corresponding to its spatial location. A dense object prediction module learns discriminative feature representations for different object classes, while a multi-scale feature extraction strategy enables the network to effectively detect objects with varying sizes and spatial distributions. The framework jointly optimizes counting and localization losses, allowing the two tasks to reinforce each other during training. In addition, a peak extraction mechanism converts the predicted density maps into accurate object locations without requiring complex post-processing. Extensive experiments conducted on

benchmark remote sensing datasets demonstrate that DOPNet consistently outperforms state-of-the-art object counting and localization methods. The proposed approach achieves lower Mean Absolute Error for object counting, higher localization accuracy, and better robustness in densely populated and cluttered scenes. The results confirm that integrating dense prediction with multi-class learning provides an effective and scalable solution for simultaneous object counting and localization in remote sensing applications such as urban planning, traffic monitoring, and environmental surveillance.

G. Sena Hocaoğlu et al., [70] proposed a camera–LiDAR sensor fusion framework **for** multiple object localization, aiming to improve the accuracy and robustness of object localization in autonomous driving and intelligent perception systems. The methodology addresses the limitations of single-sensor approaches, where cameras provide rich semantic information but lack reliable depth estimation, while LiDAR sensors offer accurate three-dimensional geometric measurements but limited appearance details. The proposed framework combines the complementary strengths of both sensors to achieve precise localization of multiple objects in complex environments. Technically, the system first performs camera–LiDAR calibration to establish spatial correspondence between image pixels and LiDAR point clouds Extensive experiments conducted on benchmark autonomous driving datasets demonstrate that the proposed sensor fusion framework consistently outperforms camera-only and LiDAR-only localization methods. The experimental results show higher localization accuracy, improved object detection precision, and greater robustness in challenging environments. Overall, the proposed camera–LiDAR fusion approach provides a reliable and efficient solution for real-time multiple object localization, making it well suited for autonomous vehicles, intelligent transportation systems, and mobile robotic applications.

Background and Problem Formulation

Several related tasks are frequently conflated in the literature, and separating them clarifies what predictive localization actually demands. Object detection assigns class labels and bounding boxes to objects visible in a single image. Object localization, as the term is used in this review, extends the question to the whole environment: the system must estimate the position of a target object within a house, whether or not the object is in the current field of view. Object goal navigation (ObjectNav) is the embodied version of this problem, in which an agent must physically travel to an instance of a specified category, and it has become the dominant benchmark task in the area [2], [16]. Semantic search describes the broader family of strategies in which an agent uses knowledge about the environment, rather than exhaustive coverage, to decide where to look next [29], [30]. Predictive localization is the reasoning core shared by all of these: producing a ranked belief over candidate locations for an unseen object, which can then drive navigation, guide a human user, or answer a query directly.

A useful taxonomy divides the problem along three axes. The first is known-object retrieval, in which the system has previously observed the specific target instance and must recall or infer its current position. Episodic memory and semantic maps are the natural tools here, since the last observed location is often, though not always, the best prediction [22], [23]. The second is unseen-object grounding, in which the target has never been observed and the system must rely entirely on priors, for instance predicting that a corkscrew is probably in a kitchen drawer despite never having seen one in this particular home [13], [17]. Recent benchmark work has shown that even strong VLMs perform poorly at inferring the storage locations of concealed items, which confirms that visibility-based training does not transfer automatically to this setting [26]. The third axis is temporal prediction, in which an object that was once observed has since been moved, and the system must reason about household dynamics, human routines, and elapsed time to update its belief. This third setting is the least studied and arguably the most important for real deployments, since homes are rearranged constantly.

The formal problem can be stated as follows. Given an environment E with a set of candidate locations L (rooms, surfaces, and receptacles), a target object description g expressed in natural language, and an observation history H, the system must output a probability distribution P(l | g, E, H) over L, or equivalently a ranked list of locations. Detection quality, prior knowledge, and memory each enter this distribution as separate factors, which is why hybrid architectures dominate the current literature.

Evaluation settings split into simulation and the real world. Simulators such as AI2-THOR [9], Habitat [10], and VirtualHome [11] offer reproducibility, cheap large-scale experimentation, and automatic ground truth, and most published results are produced in them. Procedurally generated environments such as ProcTHOR extend this to tens of thousands of training houses [12]. The cost is a persistent simulation-to-reality gap: rendered scenes are cleaner, object placement follows designer statistics rather than lived-in clutter, and physics simplifications hide manipulation difficulties. Real-world evaluations, typically conducted with mobile manipulators in instrumented apartments, remain small in scale but consistently report lower success rates than simulation, and the discrepancy is largest for small, movable objects, which are precisely the ones people misplace [2], [35]. This tension between reproducible simulation and representative reality runs through the entire field and resurfaces in the discussion of benchmarks in Section 6.

Vision-Language Models for Object Understanding

Perception for household object localization has passed through four recognizable generations. Convolutional detectors trained on fixed label sets, in the lineage of Faster R-CNN and YOLO, dominated until roughly 2021 and still appear as components in deployed robot stacks. Their central limitation for the household setting is the closed vocabulary: a detector trained on 80 COCO categories cannot respond to a request for reading glasses or a specific medication box. The second generation began with CLIP, which learned a joint embedding of images and text from 400 million web pairs and demonstrated that visual recognition could be driven by free-form language rather than fixed labels [3]. CLIP itself produces image-level similarity scores rather than boxes, so a third generation of open-vocabulary detectors adapted the contrastive recipe to localization. OWL-ViT attaches lightweight box and classification heads to a pre-trained vision transformer [4], GLIP reformulates detection as phrase grounding and trains on grounding data at scale [6], and Grounding DINO fuses text and image features throughout the network rather than only at the classification stage, achieving state-of-the-art zero-shot detection on COCO and LVIS [5]. These models allow a robot to be asked for essentially any nameable object, which removed one of the main barriers to open-ended household search.

The fourth generation consists of multimodal large language models such as GPT-4V, LLaVA, PaLI, and Flamingo, which couple a visual encoder to a large language model and can describe scenes, answer questions, and reason about images in dialogue [7], [8]. For localization these models contribute something the pure detectors lack: they carry commonsense knowledge absorbed from text, so they can be prompted to judge whether a mug is more likely in an upper cabinet or under a sink, and they can interpret referring expressions such as the blue folder next to the printer. Referring-expression comprehension and open-vocabulary grounding have consequently become standard subtasks in household perception pipelines, with grounded VLMs resolving which of several visible candidates a user means [1], [5].

Zero-shot and few-shot capability matters more in homes than in almost any other deployment domain, because household inventories follow a long-tailed distribution. Every home contains a handful of common categories and a long tail of idiosyncratic items, and no feasible training set covers a specific family's possessions. Open-vocabulary models handle the head of this distribution well, and few-shot adaptation, either through visual prompting or lightweight fine-tuning, is increasingly used to teach a system particular instances, for example distinguishing one family member's glasses from another's [1], [23].

The limitations are equally well documented. Multimodal LLMs hallucinate: they report objects that are not present and invent plausible but false spatial relations, which is corrosive in a localization system whose entire output is a claim about where something is [1]. Spatial reasoning remains weak, models that name every object in a scene still struggle with left-right relations, containment, and metric distance. Domain gap is a third concern, since web training imagery over-represents catalogue-style photographs and under-represents cluttered drawers, occluded shelves, and poor indoor lighting, and detection accuracy drops accordingly in real homes [26]. Finally, the computational cost of large VLMs conflicts with the latency and power budgets of domestic robots, a point taken up in Section 8. These weaknesses explain why perception alone has not solved the problem and why the reasoning components surveyed next carry so much of the load.

Heuristic and Commonsense Reasoning for Location Prediction

When a target object is out of view, prediction must come from priors, and the oldest and still most effective priors are semantic co-occurrence statistics. Objects associate with rooms (towels with bathrooms, ladles with kitchens) and with receptacles and anchor objects (remotes with sofas, chargers with bedside tables). Early work encoded these relations by hand or mined them from databases and web text, using knowledge graphs such as ConceptNet to score object-location pairs [28], and systems built on such priors achieved usable indirect search a decade ago [29], [30]. The Housekeep benchmark systematized this idea for household rearrangement by collecting human judgements of where hundreds of object categories belong, and showed that language models fine-tuned on text corpora predict human placement preferences surprisingly well [26]. More recently, the extraction step has been delegated to LLMs directly: ESC prompts a language model for soft commonsense constraints linking goal objects to rooms and nearby objects, and converts the scores into a search policy [18], while L3MVN uses an LLM to judge which observed frontier is semantically closest to the target [19]. In effect, the LLM has replaced the hand-built knowledge graph as the source of household commonsense.

Probabilistic formulations give these priors a principled home. Bayesian treatments represent the belief over object locations as a posterior updated with every observation, so that an empty search of the kitchen counter transfers probability mass to the drawers and the dining table [29]. Markov models capture object movement over time, treating locations as states and household activity as the transition process, which allows a system to reason that a coffee mug observed on a desk in the morning has probably migrated to the dishwasher by evening. Occupancy and semantic maps supply the spatial substrate for such reasoning: metric maps annotated with semantic labels, or full semantic grids, let the posterior be expressed over concrete places rather than abstract room names [22], [30]. The practical appeal of the probabilistic framing is that it produces calibrated rankings, supports sequential search that minimizes expected time to find the object, and cleanly absorbs negative evidence, which pure neural scoring does not.

Human-habit modeling addresses the personalization problem. Placement statistics differ between households, and annotator agreement studies confirm that categories such as medicines and packaged food have low cross-household consensus precisely because people organize them differently [26]. Systems that learn per-home preferences, whether by predicting user organization patterns from partial observations [27] or by accumulating placement episodes over weeks of operation, consistently outperform generic priors on the objects that matter most, though they face a cold-start period before enough household-specific data exists. Temporal routine learning extends this to time of day and day of week, so that predictions for keys differ at 8 a.m. and 8 p.m.

The architectural pattern that now dominates is the hybrid neuro-symbolic pipeline, in which an LLM or VLM acts as planner and knowledge source while open-vocabulary detectors act as perceivers, with a probabilistic or rule-based layer mediating between them [18], [19], [31]. CogNav, for instance, models the cognitive process of search explicitly, moving through states of broad exploration, contextual reasoning, and target confirmation under LLM control [31], and TriHelper adds dynamic assistance modules that intervene when the base policy stalls [32]. The division of labour is attractive because it plays to each component's strength: the language model supplies flexible world knowledge, the detector supplies grounded evidence, and the symbolic layer supplies calibration and interpretability. The unresolved question, discussed further in Section 8, is how to make this loop fast and reliable enough for continuous household operation.

AI-Driven Frameworks and System Architectures

Complete systems combine the perception and reasoning ingredients of the previous two sections into architectures that can be grouped into four families. The first is semantic mapping, in which the environment is reconstructed as a queryable spatial index of open-vocabulary features. VLMaps grounds CLIP-style embeddings into a top-down map so that natural-language queries resolve to map coordinates [22], ConceptFusion fuses multimodal features into 3D reconstructions, producing point clouds that can be searched with text, image, or audio queries [23], and OpenScene distils 2D open-vocabulary features into 3D point representations for scene-level understanding [24]. Hierarchical extensions organize the map into a scene graph of floors, rooms, and objects, which matches the way location priors are naturally expressed and allows queries to be answered at the appropriate level of granularity [25]. For predictive localization, semantic maps function as long-term memory: an object seen once is findable later, and the map provides the candidate location set over which predictive distributions are defined.

The second family is object goal navigation and target-driven search agents, which couple prediction to embodied action. Zero-shot approaches have proven remarkably strong: CoWs on Pasture showed that pairing an open-vocabulary detector with classical frontier exploration is a hard baseline to beat [17], VLFM scores exploration frontiers with vision-language similarity to bias search toward semantically promising regions [20], and semantic policy networks learn to map scene semantics directly to search actions [34]. LLM-guided variants such as L3MVN and CogNav insert language-model reasoning into the exploration loop [19], [31], and adaptive strategies such as ApexNav fuse target-centric semantic cues to decide when to explore and when to verify [33]. Across this family the trend is consistent: semantic guidance reduces search time substantially relative to uninformed exploration, and the quality of the location prior is the dominant factor in efficiency.

The third family is memory-augmented systems, which take the temporal dimension seriously. Scene graphs with per-object timestamps, episodic logs of object observations, and retrieval-augmented designs that store placement events allow an agent to answer where an object was last seen and to notice when the world has changed [23], [25]. This family is the natural host for temporal prediction, since transition statistics can be estimated from the accumulated episodes, but published systems remain early: most maintain memory of static placements and few model movement dynamics explicitly.

The fourth division cuts across the others: end-to-end policies versus modular pipelines. End-to-end approaches train a single network from observations to actions, which removes hand-designed interfaces and can discover strategies modular designers would not, but they demand enormous training data, transfer poorly across houses, and offer no inspectable belief over locations. Modular pipelines expose an explicit prediction stage, which aids debugging, allows components to be upgraded independently as better VLMs appear, and produces the ranked location outputs that assistive applications need. The literature currently favours modularity, and surveys of the ObjectNav field reach the same conclusion [2]. A comparative reading of representative systems across their inputs, VLM backbone, reasoning mechanism, evaluation environment, and reported success rates makes the pattern concrete, and a table with exactly these columns is the recommended centrepiece of this section in the full paper, the works cited above [17], [19], [20], [22], [23], [31], [33] form a suitable initial row set for it.

Datasets, Benchmarks, and Evaluation Metrics

Simulation benchmarks anchor evaluation in this field. AI2-THOR provides interactive room-scale scenes with actionable objects [9], and its descendants extend it in two directions: RoboTHOR adds paired physical apartments so that simulation-trained agents can be tested on

real robots in matched layouts [13], and ProcTHOR generates houses procedurally at a scale of ten thousand environments, which transformed training data availability for embodied agents [12]. Habitat contributes photorealistic reconstruction-based scenes and a fast simulator [10], with the HM3D dataset supplying nearly a thousand scanned buildings [14], and the HM3D-OVON extension reframing ObjectNav with open-vocabulary goal categories, which aligns the benchmark with modern VLM capabilities [21]. VirtualHome takes a complementary angle, simulating household activities as executable programs, which makes it useful for modeling the human routines that move objects around [11]. Housekeep occupies a special position for predictive work because its ground truth is human preference about where objects belong rather than a recorded placement, exactly the prior a predictive system must capture [26].

Real-world data remains scarcer. Robot deployments produce small in-house datasets that rarely circulate, and the community consequently borrows from egocentric video. Ego4D, with thousands of hours of daily-life recordings and an episodic memory benchmark that includes visual queries of the form where did I last see X, is the closest existing resource to naturalistic object displacement data, and its query formulation is essentially predictive localization from a human viewpoint [15]. Recent benchmark proposals target the concealed-object case directly, asking models to infer which cabinet or drawer holds a queried item that is not visible, and report that current VLMs sit far below human performance on this task [26].

Metrics follow the task formulations. Embodied search is scored by success rate and Success weighted by Path Length (SPL), which discounts success by the ratio of shortest-path to actual-path distance and has become the field standard [16]. Search efficiency variants report time or distance to find. Perception components are scored with detection metrics such as mean average precision, while prediction quality itself is best captured by top-k accuracy over ranked candidate locations, mean reciprocal rank, and calibration measures on the predicted distribution. Distance-based localization error applies when predictions are metric coordinates rather than discrete places.

The gaps in current benchmarking practice are significant for this review's subject. Almost all established benchmarks evaluate reactive search, in which the object sits in a static scene waiting to be found, whereas the predictive setting requires objects that move between episodes according to plausible household dynamics, and no widely adopted benchmark provides this. Evaluation of ranked predictions independent of navigation is rare, so a system's prior quality is entangled with its exploration policy. Personalization is untested, since benchmark scenes carry no notion of a household's persistent habits. Finally, the field's reliance on simulation means that reported numbers systematically overstate real-home performance, particularly for the small graspable objects that people actually lose [2]. A benchmark combining procedurally generated households, scripted resident routines that relocate objects, and evaluation of top-k location prediction over time would fill the largest of these gaps, and constructing one is among the clearest opportunities the literature currently presents.

Applications

Domestic service robotics is the most direct application. Fetch-and-carry is a staple task in home robot competitions and commercial prototypes, and its bottleneck is rarely grasping alone: before an object can be fetched it must be found, and search time dominates task time whenever the object is not in plain view. Predictive localization converts fetch requests from exhaustive exploration into short, directed searches, and systems that integrate remote multimodal interaction already demonstrate users asking a household robot for objects through language and gesture from another room [35]. Surveys of embodied AI identify household object retrieval as one of the canonical capabilities expected of general-purpose service robots [36], and every improvement in location prediction transfers directly into task throughput for such platforms.

Assistive technology for elderly users and people with memory impairment is arguably the application with the greatest social weight. Misplacing everyday items is among the most commonly reported difficulties in early cognitive decline, and the objects involved, such as spectacles, hearing aids, keys, wallets, and medication, are small, personal, and moved many times a day, which is precisely the regime where generic detectors fail and learned household priors matter. A system that answers where are my glasses with a short ranked list of likely places, ideally with a supporting explanation, functions as an external memory aid and can reduce both caregiver load and user distress. The design constraints here are stricter than in general robotics: predictions must be trustworthy and explainable, false confidence is harmful, and the population served is least tolerant of complicated interfaces, which argues for the calibrated, interpretable prediction stacks discussed in Section 4.

Smart-home integration offers a complementary, robot-free deployment path. Fixed cameras and other ambient sensors observe placement events continuously, so the observation history that mobile robots must gather through exploration arrives for free, and the localization problem reduces to maintaining and querying a persistent object-location belief. This is the setting where habit models and temporal reasoning yield the largest gains, since the sensor stream captures the household's actual routines. It is also the setting where privacy concerns are sharpest, because continuous in-home vision is involved, the tension is examined in Section 8, and edge processing with on-device models is the mitigation the literature most often proposes.

Augmented-reality object finding on mobile and wearable devices closes the loop between prediction and the human user. A phone or headset that has previously scanned the home can overlay guidance toward a predicted location, and egocentric perception research provides the technical basis: the Ego4D visual query task, which asks where an object was last seen in a user's own video history, is effectively AR object finding evaluated at scale [15]. Semantic 3D maps built with open-vocabulary features give such applications their spatial index [22], [23], and the same ranked-prediction interface serves both a robot's search planner and an AR arrow rendered for a person. Across all four application areas the common requirement is a calibrated, queryable belief over object locations, which is a useful reminder that the predictive core, rather than any single embodiment, is the reusable asset.

Challenges and Open Research Directions

Privacy is the first and least technical-sounding challenge, yet it shapes every deployment decision. Predictive localization improves with observation history, and the richest history comes from continuous in-home sensing, which means cameras watching private living space around the clock. Acceptable systems will need on-device processing so that raw imagery never leaves the home, retention policies that store abstract placement events rather than video, spatial and temporal masking of sensitive areas, and user-facing controls that make the system's memory inspectable and erasable. Almost none of the surveyed literature engages with these requirements, and privacy-preserving formulations of object memory, for example learning placement priors from encrypted or heavily abstracted event streams, are essentially unexplored.

Personalization and cold start form the second cluster. Generic priors carry a new system through its first days, but the objects users care most about follow household-specific patterns, and low cross-household agreement on where such items belong is empirically documented [26], [27]. Open questions include how quickly per-home models can be learned from sparse placement events, how to combine population priors with household evidence in a principled Bayesian way, and how to share statistical strength across homes without sharing data, for which federated learning is an obvious but untested candidate. Temporal reasoning over long horizons is closely related: current memory-augmented systems mostly store static placements, while real prediction requires transition models of how objects circulate through a home across days and seasons, learned from the kind of activity data that VirtualHome-style simulation [11] and egocentric recordings [15] can supply.

Computational cost is the third constraint. The strongest perception and reasoning components are large models with latencies and power draws unsuited to robot and edge hardware, and the field's standard remedy of cloud offloading collides with the privacy requirements above. Distillation of open-vocabulary detectors into edge-sized models, caching of LLM-derived priors so that the language model is consulted rarely rather than per-step, and hybrid designs that reserve heavyweight reasoning for hard cases are all active directions, and edge-oriented variants of open-set detectors have begun to appear [5]. Reported gaps between simulation and on-robot performance suggest that efficiency work will pay off more in deployment than a further point of benchmark accuracy [2].

Conclusion

This review has examined predictive household object localization as a problem in its own right, distinct from the detection and navigation tasks it is usually folded into. The central finding is that the field has converged, largely since 2023, on a hybrid architectural pattern: open-vocabulary vision-language perception supplies what can be seen and named, while heuristic, probabilistic, or LLM-based reasoning supplies where an unseen object is likely to be. Neither component is sufficient alone. Detectors without priors cannot reason about concealed or relocated objects, and priors without grounded perception cannot adapt to a particular home. The surveyed frameworks differ mainly in where they place the boundary between these components and in how they represent the environment, whether as semantic maps, 3D scene graphs, or language-queryable memories.Significant gaps remain. Temporal prediction, reasoning about objects that have moved since last observed, is the least developed of the three problem settings despite being the most representative of real households. Evaluation practice lags the problem definition: no widely adopted benchmark tests ranked prediction of out-of-view objects under dynamic placement, and results in clean simulated scenes overstate real-world performance, especially for the small movable items people actually misplace. Personalization, calibrated uncertainty, privacy-preserving household memory, and inference on embedded hardware are open engineering and research questions rather than solved details. Progress on a shared benchmark with simulated resident routines and ranked-prediction metrics would, in our view, do more to advance the area than further gains on existing navigation leaderboards.

References

1. J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 8, pp. 5625-5644, 2024.
2. J. Sun, J. Wu, Z. Ji, and Y.-K. Lai, "A survey of object goal navigation," IEEE Transactions on Automation Science and Engineering, vol. 22, pp. 2292-2308, 2024.

3. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. International Conference on Machine Learning (ICML), 2021, pp. 8748-8763.
4. Lavanya Vaishnavi D. A. and C. Anil Kumar, “Performance Enhancement of 5G MIMO Antenna Utilizing Verifiable Convolutional Neural Network Optimized With Human Evolutionary Optimization Algorithm,” International Journal of Communication Systems, vol. 39, no. 4, Art. no. e70404, Jan. 2026, doi: 10.1002/dac.70404.
5. H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in Advances in Neural Information Processing Systems (NeurIPS), vol. 36, 2023.
6. J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," in Advances in Neural Information Processing Systems (NeurIPS), vol. 35, 2022, pp. 23716-23736.
7. Lavanya Vaishnavi D. A. and C. Anil Kumar, “Evaluating Supervised Learning Classifier Performance for OFDM Communication in AWGN-Impacted Systems,” Results in Engineering, vol. 26, Art. no. 105178, May 2025, doi: 10.1016/j.rineng.2025.105178.
8. E. Kolve et al., "AI2-THOR: An interactive 3D environment for visual AI," arXiv preprint arXiv:1712.05474, 2017.
9. M. Savva et al., "Habitat: A platform for embodied AI research," in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9339-9347.
10. X. Puig et al., "VirtualHome: Simulating household activities via programs," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8494-8502.
11. M. Deitke et al., "ProcTHOR: Large-scale embodied AI using procedural generation," in Advances in Neural Information Processing Systems (NeurIPS), vol. 35, 2022, pp. 5982-5994.
12. C. Anil Kumar, S. Harish, P. Ravi, M. S. V. N., B. P. Pradeep Kumar, V. Mohanavel, N. M. Alyami, S. Shanmuga Priya, and A. K. Asfaw, “Lung Cancer Prediction from Text Datasets Using Machine Learning,” *BioMed Research International*, vol. 2022, Art. no. 6254177, Jul. 2022, doi: 10.1155/2022/6254177. *(Note: This article has subsequently been retracted by the publisher.)*
13. M. Deitke et al., "RoboTHOR: An open simulation-to-real embodied AI platform," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3164-3174.
14. S. K. Ramakrishnan et al., "Habitat-Matterport 3D dataset (HM3D): 1000 large-scale 3D environments for embodied AI," in Proc. NeurIPS Datasets and Benchmarks Track, 2021.
15. K. Grauman et al., "Ego4D: Around the world in 3,000 hours of egocentric video," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18995-19012.
16. P. Anderson et al., "On evaluation of embodied navigation agents," arXiv preprint arXiv:1807.06757, 2018.
17. S. Y. Gadre, M. Wortsman, G. Ilharco, L. Schmidt, and S. Song, "CoWs on Pasture: Baselines and benchmarks for language-driven zero-shot object navigation," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23171-23181.
18. K. Zhou et al., "ESC: Exploration with soft commonsense constraints for zero-shot object navigation," in Proc. International Conference on Machine Learning (ICML), 2023, pp. 42829-42842.
19. B. Yu, H. Kasaei, and M. Cao, "L3MVN: Leveraging large language models for visual target navigation," in Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2023, pp. 3554-3560.
20. N. Yokoyama, S. Ha, D. Batra, J. Wang, and B. Bucher, "VLFM: Vision-language frontier maps for zero-shot semantic navigation," in Proc. IEEE International Conference on Robotics and Automation (ICRA), 2024, pp. 42-48.
21. N. Yokoyama, R. Ramrakhya, A. Das, D. Batra, and S. Ha, "HM3D-OVON: A dataset and benchmark for open-vocabulary object goal navigation," in Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2024, pp. 5543-5550.
22. C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual language maps for robot navigation," in Proc. IEEE International Conference on Robotics and Automation (ICRA), 2023, pp. 10608-10615.
23. K. M. Jatavallabhula et al., "ConceptFusion: Open-set multimodal 3D mapping," in Proc. Robotics: Science and Systems (RSS), 2023.
24. S. Peng et al., "OpenScene: 3D scene understanding with open vocabularies," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 815-824.
25. Werby, C. Huang, M. Buechner, A. Valada, and W. Burgard, "Hierarchical open-vocabulary 3D scene graphs for language-grounded robot navigation," in Proc. IEEE International Conference on Robotics and Automation (ICRA), 2024.
26. Y. Kant et al., "Housekeep: Tidying virtual households using commonsense reasoning," in Proc. European Conference on Computer Vision (ECCV), 2022, pp. 355-373.
27. N. Abdo, C. Stachniss, L. Spinello, and W. Burgard, "Robot, organize my shelves! Tidying up objects by predicting user preferences," in Proc. IEEE International Conference on Robotics and Automation (ICRA), 2015, pp. 1557-1564.

28. R. Speer, J. Chin, and C. Havasi, "ConceptNet 5.5: An open multilingual graph of general knowledge," in Proc. AAAI Conference on Artificial Intelligence, 2017, pp. 4444-4451.
29. P. Loncomilla, J. Ruiz-del-Solar, and M. Saavedra, "A Bayesian based methodology for indirect object search," Journal of Intelligent and Robotic Systems, vol. 90, no. 1-2, pp. 45-63, 2018.
30. T. S. Veiga, P. Miraldo, R. Ventura, and P. U. Lima, "Efficient object search for mobile robots in dynamic environments: Semantic map as an input for the decision maker," in Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2016, pp. 2745-2750.
31. Y. Cao et al., "CogNav: Cognitive process modeling for object goal navigation with LLMs," in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2025, pp. 9550-9560.
32. L. Zhang et al., "TriHelper: Zero-shot object navigation with dynamic assistance," in Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2024, pp. 10035-10042.
33. M. Zhang et al., "ApexNav: An adaptive exploration strategy for zero-shot object navigation with target-centric semantic fusion," IEEE Robotics and Automation Letters, vol. 10, no. 11, pp. 11530-11537, 2025.
34. Q. Zhao, L. Zhang, B. He, and Z. Liu, "Semantic policy network for zero-shot object goal visual navigation," IEEE Robotics and Automation Letters, vol. 8, no. 11, pp. 7655-7662, 2023.
35. Xiao et al., "Robi Butler: Multimodal remote interaction with a household robot assistant," in Proc. IEEE International Conference on Robotics and Automation (ICRA), 2025.
36. Y. Liu et al., "Aligning cyber space with physical world: A comprehensive survey on embodied AI," IEEE/ASME Transactions on Mechatronics, 2025.
37. K. Zhao et al., "Open-Vocabulary Camouflaged Object Segmentation with Cascaded Vision Language Models," in Computational Visual Media, vol. 12, no. 2, pp. 473-492, April 2026, doi: 10.26599/CVM.2025.9450512.
38. Ş. Bulzan, C. Cernăzanu-Glăvan and M. -G. Marcu, "Unlocking Object Detection in Vision-Language Models," in IEEE Access, vol. 14, pp. 46178-46191, 2026, doi: 10.1109/ACCESS.2026.3675517.
39. H. Hou, M. Feng, Z. Wu, Y. Guo, Y. Wang and A. Mian, "Spatial Multimodal Knowledge-Driven 3D Scene Graph Prediction With Vision-Language Model," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 36, no. 6, pp. 7483-7498, June 2026, doi: 10.1109/TCSVT.2026.3664122.
40. L. He et al., "Vision–Language Models Empowered Nighttime Object Detection With Consistency Sampler and Hallucination Feature Generator," in IEEE Transactions on Image Processing, vol. 34, pp. 8345-8360, 2025, doi: 10.1109/TIP.2025.3641316.
41. J. Zhang, J. Huang, S. Jin and S. Lu, "Vision-Language Models for Vision Tasks: A Survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 8, pp. 5625-5644, Aug. 2024, doi: 10.1109/TPAMI.2024.3369699.
42. J. Chen and K. Yanai, "Bridging Detection Architectures With Foundation Models: A Unified Framework for Human–Object Interaction Detection," in IEEE Access, vol. 14, pp. 23299-23310, 2026, doi: 10.1109/ACCESS.2026.3659132.
43. H. Chang, W. Lee and M. Kim, "CT-FSOD: CLIP-Based Text-Guided Few-Shot Object Detection," in IEEE Access, vol. 13, pp. 158250-158265, 2025, doi: 10.1109/ACCESS.2025.3604521.
44. J. Lu and H. Chen, "VLSDA: Vision–Language Model-Supervised Domain Adaptation for Cross-Domain Object Detection in Remote Sensing," in IEEE Transactions on Geoscience and Remote Sensing, vol. 63, pp. 1-15, 2025, Art no. 5651515, doi: 10.1109/TGRS.2025.3627226.
45. F. Liu et al., "RemoteCLIP: A Vision Language Foundation Model for Remote Sensing," in IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1-16, 2024, Art no. 5622216, doi: 10.1109/TGRS.2024.3390838.
46. Z. Wei et al., "Unseen From Seen: Rewriting Observation-Instruction Using Foundation Models for Augmenting Vision-Language Navigation," in IEEE Transactions on Neural Networks and Learning Systems, vol. 37, no. 4, pp. 1726-1738, April 2026, doi: 10.1109/TNNLS.2025.3624691.
47. J. Ke, L. He, B. Han, J. Li, D. Wang and X. Gao, "VLDadaptor: Domain Adaptive Object Detection With Vision-Language Model Distillation," in IEEE Transactions on Multimedia, vol. 26, pp. 11316-11331, 2024, doi: 10.1109/TMM.2024.3453061.
48. W. Zhai, X. Xing, M. Gao and Q. Li, "Zero-Shot Object Counting With Vision-Language Prior Guidance Network," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 35, no. 3, pp. 2487-2498, March 2025, doi: 10.1109/TCSVT.2024.3488721.
49. X. Yang, F. Liu and G. Lin, "Effective End-to-End Vision Language Pretraining With Semantic Visual Loss," in IEEE Transactions on Multimedia, vol. 25, pp. 8408-8417, 2023, doi: 10.1109/TMM.2023.3237166.
50. P. Zhang, Y. Zhang, H. Wu, X. Liu, Y. Hou and L. Wang, "Language-Guided Object Localization via Refined Spotting Enhancement in Remote Sensing Imagery," in IEEE Transactions on Geoscience and Remote Sensing, vol. 63, pp. 1-15, 2025, Art no. 5621315, doi: 10.1109/TGRS.2025.3562439.

51. H. Lin et al., "RS-MoE: A Vision–Language Model With Mixture of Experts for Remote Sensing Image Captioning and Visual Question Answering," in IEEE Transactions on Geoscience and Remote Sensing, vol. 63, pp. 1-18, 2025, Art no. 5614918, doi: 10.1109/TGRS.2025.3547988.
52. R. Zhang, X. -J. Wu, C. Wang and C. -L. Liu, "Det-Agent: Open-Vocabulary Object Localization and Detection With Reinforcement Learning Agent," in IEEE Transactions on Multimedia, vol. 28, pp. 5722-5735, 2026, doi: 10.1109/TMM.2026.3668516.
53. S. Yuan, H. Zhang, L. Liu, L. Zhu and X. Dong, "SPENav: Dynamic Object Filtering With Spatial Perception Enhancement for Vision-Language Navigation," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 36, no. 5, pp. 6905-6919, May 2026, doi: 10.1109/TCSVT.2026.3651320.
54. H. Zhang, J. Wang, J. Zhang, T. Zhang and B. Zhong, "One-Stream Vision-Language Memory Network for Object Tracking," in IEEE Transactions on Multimedia, vol. 26, pp. 1720-1730, 2024, doi: 10.1109/TMM.2023.3285441.
55. S. Sadhwani, J. A. Shamsi, M. B. Khan, N. Z. Bawany and H. J. Syed, "Real-Time Detection of Mixed-Critical Events Using Vision-Language Models," in IEEE Access, vol. 13, pp. 181363-181384, 2025, doi: 10.1109/ACCESS.2025.3622638.
56. J. Yang, N. Jia, X. Liu, R. Fan, Y. Sun and W. Zhao, "Zone-YOLO: Vision-Language Object Detection Using Zone Prompt," in IEEE Transactions on Intelligent Transportation Systems, vol. 26, no. 10, pp. 17510-17521, Oct. 2025, doi: 10.1109/TITS.2024.3510117.
57. Y. Ma, M. Liu, C. Zhu and X. -C. Yin, "HA-FGOVD: Highlighting Fine-Grained Attributes via Explicit Linear Composition for Open-Vocabulary Object Detection," in IEEE Transactions on Multimedia, vol. 27, pp. 3171-3183, 2025, doi: 10.1109/TMM.2025.3557624.
58. P. P. Ninh and H. Kim, "Lightweight Source-Free Unsupervised Domain Adaptation for Object Detector Using Stereo Vision," in IEEE Access, vol. 14, pp. 98295-98310, 2026, doi: 10.1109/ACCESS.2026.3706740.
59. Y. Zhu et al., "ChatNav: Leveraging LLM to Zero-Shot Semantic Reasoning in Object Navigation," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 35, no. 3, pp. 2369-2381, March 2025, doi: 10.1109/TCSVT.2024.3485907.
60. Rasekh, S. Kazemi Ranjbar, M. Heidari Sorkhehdizaj, S. Gottschalk and W. Nejdl, "Bridging Explainability and Accuracy in CLIP for Object Recognition via Joint Probability Over Rationales and Categories," in IEEE Access, vol. 14, pp. 13193-13201, 2026, doi: 10.1109/ACCESS.2025.3649809.
61. M. Zhang, G. Tian, Y. Cui, Y. Zhang and Z. Xia, "Hierarchical Semantic Knowledge-Based Object Search Method for Household Robots," in IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 8, no. 1, pp. 930-941, Feb. 2024, doi: 10.1109/TETCI.2023.3297838.
62. L. Zhu et al., "Background-Aware Classification Activation Map for Weakly Supervised Object Localization," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 12, pp. 14175-14191, Dec. 2023, doi: 10.1109/TPAMI.2023.3309621.
63. M. Meng, T. Zhang, W. Yang, J. Zhao, Y. Zhang and F. Wu, "Diverse Complementary Part Mining for Weakly Supervised Object Localization," in IEEE Transactions on Image Processing, vol. 31, pp. 1774-1788, 2022, doi: 10.1109/TIP.2022.3145238.
64. C. Tan, G. Gu, T. Ruan, S. Wei and Y. Zhao, "Dual-Gradients Localization Framework With Skip-Layer Connections for Weakly Supervised Object Localization," in IEEE Transactions on Multimedia, vol. 25, pp. 4933-4942, 2023, doi: 10.1109/TMM.2022.3184486.
65. J. Bai et al., "Localizing From Classification: Self-Directed Weakly Supervised Object Localization for Remote Sensing Images," in IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 12, pp. 17935-17949, Dec. 2024, doi: 10.1109/TNNLS.2023.3309889.
66. P. Zhou, Y. Liu and Z. Meng, "PointSLOT: Real-Time Simultaneous Localization and Object Tracking for Dynamic Environment," in IEEE Robotics and Automation Letters, vol. 8, no. 5, pp. 2645-2652, May 2023, doi: 10.1109/LRA.2023.3256919.
67. M. Yan, Y. Xiong, H. Kurokawa and J. She, "Multimodal Localization Distillation Method for 3D Single Object Tracking Task in Intelligent Transportation Systems," in IEEE Transactions on Intelligent Transportation Systems, vol. 26, no. 10, pp. 17612-17623, Oct. 2025, doi: 10.1109/TITS.2024.3520880.
68. X. Li, Z. Yan, S. Feng, C. Xia, S. Li and Y. Zhou, "LIO-LOT: Tightly-Coupled Multi-Object Tracking and LiDAR-Inertial Odometry," in IEEE Transactions on Intelligent Transportation Systems, vol. 26, no. 1, pp. 742-756, Jan. 2025, doi: 10.1109/TITS.2024.3488288.
69. M. Cui, G. Ding, D. Yang and Z. Chen, "DOPNet: Dense Object Prediction Network for Multiclass Object Counting and Localization in Remote Sensing Images," in IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1-15, 2024, Art no. 5604915, doi: 10.1109/TGRS.2024.3349702.

70. G. Sena Hocaoğlu and E. Benli, "Multiple Objects Localization With Camera-LIDAR Sensor Fusion," in IEEE Sensors Journal, vol. 25, no. 7, pp. 11892-11905, 1 April1, 2025, doi: 10.1109/JSEN.2025.3541431.