3D Scene Graphs: A Survey of Vision-Language Models for Robot Perception and Planning

S7784976 Mamoru Ota
S7899827 Paul Pham Dang

***Abstract*—Robust perception and high-level planning are essential capabilities for autonomous robots operating in complex, unstructured environments. Recent advances in 3D scene graphs have enabled structured representations that unify semantic understanding with spatial geometry, providing a rich framework for integrating perception, memory, and action. Simultaneously, the emergence of vision-language models (VLMs) has introduced open-vocabulary recognition and cross-modal reasoning, dramatically improving generalization and adaptability in robotic perception. This survey presents a comprehensive overview of recent research at the intersection of 3D scene graph construction and VLM-based perception, with a particular focus on applications to robotic planning and interaction. We introduce a novel taxonomy of scene graph systems based on vocabulary openness, spatial representation, and temporal dynamics. We analyze core components including object and relation grounding, hierarchical spatial mapping, dynamic graph updates, and task-level integration for navigation and manipulation. Additionally, we highlight recent innovations in multi-robot systems, topological representations, and language-guided planning. Finally, we discuss key challenges, emerging trends, and promising directions toward general-purpose, language-informed spatial intelligence in robotics.**

***Index Terms*—3D Scene Graphs, Vision-Language Models, Robotic Perception, Semantic Mapping, Open-Vocabulary Recognition, Task Planning, Spatial Reasoning**

I. Introduction

ROBOTIC systems operating in unstructured and dynamic environments must go beyond geometric understanding to reason about semantics, spatial relations, and long-horizon tasks. To support such capabilities, *3D scene graphs* have emerged as a powerful representation, capturing the entities in a scene, their attributes, and the spatial and semantic relationships between them [1]. These graphs offer a structured and interpretable abstraction of the world, enabling tasks such as language grounding, spatial reasoning, and planning.

Recent advances have introduced *open-vocabulary 3D scene graphs*—systems that leverage large-scale vision-language models (VLMs), such as CLIP, to recognize a wide variety of object categories and relations without the need for closed-set, task-specific labels [2]–[4]. This open-set capability significantly improves the adaptability of robotic systems to novel environments and user instructions. VLMs trained with natural language supervision [5] offer zero-shot generalization and semantic richness that were previously difficult to achieve in robotic perception.

These advances have enabled the integration of open-vocabulary scene graphs into a variety of robotic perception and planning pipelines. For instance, *ConceptGraphs* [2] propose a system that builds open-vocabulary 3D scene graphs directly from RGB-D input, grounding object and relationship predictions using VLM embeddings. *Grounded Classical Planners* [6] demonstrate how symbolic planners can be driven by VLM-based scene understanding, effectively bridging vision and high-level task execution. Similarly, *Dynamic Scene Graphs* [7] address the challenge of long-term autonomy by incrementally updating open-vocabulary scene graphs over time for persistent, language-guided manipulation.

Additionally, several works explore scene graph construction for spatial memory [8], egocentric change tracking [9], multi-robot coordination [10], and hierarchical representation of indoor and outdoor spaces [4], [11]–[13]. Collectively, these contributions highlight a growing consensus: open-vocabulary scene graphs, powered by vision-language models, offer a unified framework for perception and planning in robotics.

A. Contributions

In this paper, we present a comprehensive survey of recent work at the intersection of 3D scene graphs, vision-language models, and robotic planning. Our contributions are as follows:

- We propose a taxonomy of 3D scene graph systems based on vocabulary openness, spatial representation, and temporal structure.
- We review methods for constructing open-vocabulary scene graphs using VLMs for perception.
- We analyze how these representations are integrated into planning systems for navigation, manipulation, and interaction.
- We identify open challenges and promising directions, including dynamic scene understanding and long-term spatial memory.

B. Paper Structure

The remainder of the paper is organized as follows: Section II introduces background on scene graphs and vision-language models. Section III presents a taxonomy of 3D scene graph systems. Section IV reviews perception techniques using VLMs. Section V discusses planning and interaction. Section VI presents a discussion of the key challenges and outlines future directions. Section VII concludes the paper.

II. Background

This section reviews the foundational elements relevant to the integration of 3D scene graphs and vision-language models

in robotic perception and planning. We first introduce the concept of scene graphs and their evolution into 3D and open-vocabulary variants. Then, we provide an overview of vision-language models, their capabilities, and their growing role in robotics.

A. Scene Graphs and Their Extensions

Scene graphs were originally proposed as structured representations of visual scenes, encoding objects as nodes and their pairwise relationships as labeled edges [14]. This abstraction allows for reasoning over high-level semantics and spatial configurations, making it particularly useful for tasks such as image retrieval, question answering, and visual reasoning.

While early scene graph research focused on 2D image domains, the need for spatially grounded, three-dimensional representations led to the development of *3D scene graphs*. Armeni et al. [1] first proposed a framework that integrates 3D geometry with semantic relationships and camera pose, enabling holistic spatial reasoning for indoor environments. These graphs encode not just object categories and attributes but also metric positions, part-of relationships, and hierarchical structures.

Recent advances have extended this concept to *open-vocabulary 3D scene graphs*, which allow the representation of previously unseen object categories and relationships. ConceptGraphs [2] exemplify this approach by leveraging large vision-language models to embed object and relation descriptions into a shared semantic space. These open-vocabulary scene graphs remove the constraint of fixed taxonomies, allowing robots to perceive and reason about a much broader set of real-world entities.

3D scene graphs have since been utilized for dynamic, long-term representations [7], [9], topological map abstractions [12], [13], multi-robot coordination [10], and hierarchical spatial modeling [4], [11].

B. Vision-Language Models (VLMs)

Vision-language models (VLMs) are neural networks trained to align visual and textual modalities in a shared embedding space. Notable examples include CLIP [5], which learns to match images and natural language descriptions through contrastive training on large-scale web data.

VLMs are particularly powerful for open-vocabulary tasks, as they do not rely on a closed set of object labels. Instead, they enable zero-shot recognition: given a textual prompt (e.g., "a red fire extinguisher"), the model can retrieve or identify the matching visual input without retraining. This property makes VLMs especially appealing for robotics, where environments often contain novel or unlabelled objects.

In recent works, VLMs have been employed as core perception modules for 3D scene graph construction [3], semantic map generation [8], and language-grounded planning [6]. Their generalization capabilities have enabled systems to move beyond rigid object taxonomies, toward flexible and adaptive representations compatible with natural language commands.

By combining the structured reasoning of scene graphs with the semantic richness of VLMs, current approaches aim to build robust, scalable, and general-purpose perception systems for robotic agents operating in open-world environments.

III. Taxonomy of 3D Scene Graph Systems

To understand the diversity of approaches in the field, we propose a taxonomy of 3D scene graph systems along four primary dimensions: (1) vocabulary openness, (2) spatial representation, (3) temporal structure, and (4) multi-agent and hierarchical extensions. This taxonomy enables us to categorize existing methods and highlight the trade-offs between generalization, expressiveness, and computational complexity.

A. Vocabulary Scope: Closed vs. Open

Traditional 3D scene graph systems rely on a fixed vocabulary of object and relation labels, often aligned with standard datasets (e.g., NYU-Depth, SUN RGB-D). These closed-vocabulary approaches enable precise learning but struggle with generalization to unseen categories. Representative examples include IMP [15] and MOTIFNET [16], which both rely on fixed sets of object and predicate categories (e.g., 150 object classes and 50 predicates) derived from the Visual Genome dataset. These closed-vocabulary models perform well in structured settings, but their reliance on predefined labels limits their scalability to open-world environments.

In contrast, open-vocabulary systems leverage vision-language models to recognize an unbounded set of categories and relationships. These systems, such as ConceptGraphs [2], From Pixels to Graphs [3], and OpenGraph [4], use textual prompts and CLIP-style embeddings to flexibly interpret novel scenes. Open-vocabulary graphs are critical for long-term autonomy in dynamic, real-world environments where closed taxonomies are insufficient.

B. Spatial Representation: Metric, Topological, and Hybrid

3D scene graphs differ in how they represent space. The spatial representation in a scene graph determines how the system models the location, layout, and spatial relationships of objects and regions in a 3D environment. Depending on the intended application—such as precise manipulation or high-level navigation—different spatial encodings are adopted.

Some systems employ full metric maps, storing accurate 3D positions of objects and regions. In metric representations, each object or region is associated with exact geometric coordinates (e.g., x, y, z in meters). These maps are typically constructed using RGB-D sensors or LiDAR, and enable operations that require precise spatial reasoning. For example, ConceptGraphs [2] and MR-COGraphs [10] maintain precise spatial graphs built from RGB-D input and LiDAR, enabling detailed manipulation and localization.

In contrast, topological or abstracted representations encode spatial relationships in terms of connectivity or adjacency between regions or entities, rather than using absolute coordinates. RoboHop [13] uses a segment-based topological structure where semantically meaningful image segments serve as graph nodes. Edges are defined based on visual similarity

across images and spatial proximity within images. TopoField [12] models space using hierarchical layout-object-position fields inspired by cognitive neuroscience, abstracting spatial relationships at multiple semantic levels (e.g., object-in-room, room-in-building) to support robust planning in complex environments. Topological graphs require less memory and are more robust in large-scale or uncertain settings, though they lack fine-grained spatial precision.

Some systems, like OpenGraph [4], use a hybrid metric-topological design. These hybrid approaches combine the strengths of both types: topological layers model global layout and connectivity, while metric subgraphs provide local geometric detail. This design allows systems to efficiently navigate large spaces while retaining sufficient precision for tasks like object manipulation.

C. Temporal Structure: Static vs. Dynamic

Most early 3D scene graph systems are static—they build a scene graph once per scan or episode and do not account for temporal change. However, many real-world scenarios require dynamic scene graphs that evolve over time.

Recent work has sought to overcome this limitation by introducing dynamic 3D scene graph representations. Dynamic Open-Vocabulary 3D Scene Graphs [7] maintain temporal consistency and support incremental updates as the robot explores and interacts with its environment. This representation defines three key spatial relationships between objects: ”on”, ”belong”, and ”inside”, where the ”inside” relation explicitly models occlusion by indicating that a previously observed object is now contained within another.

Similarly, Lost & Found [9] tracks semantic and geometric changes over time, including object additions, removals, and displacements. It employs relations such as ”close to”, ”part of”, and ”contains”, with the ”contains” relation serving an analogous purpose to ”inside” in modeling object containment.

These dynamic frameworks are critical for enabling long-term spatial reasoning and memory in persistent robotic agents, ensuring robust environmental understanding despite changes over time.

D. Multi-Agent and Hierarchical Extensions

Some systems extend scene graphs to multi-agent settings or hierarchical structures. MR-COGraphs [10] enables distributed mapping and scene graph sharing between robots with limited bandwidth. Intelligent Spatial Perception [11] and OpenGraph [4] construct hierarchical graphs with multiple levels of abstraction (e.g., object–room–building), enabling reasoning at different spatial scales.

These extensions are especially relevant for large-scale deployments, such as warehouse automation, smart buildings, and autonomous vehicles operating in open-world environments.

IV. Perception with Vision-Language Models

The integration of vision-language models (VLMs) into robotic perception pipelines has catalyzed significant advances in open-vocabulary 3D scene graph generation. This section reviews methods that leverage VLMs to achieve rich semantic understanding beyond fixed label sets, enabling robots to interpret complex environments from raw sensory data.

A. Open-Vocabulary Object and Relation Recognition

Traditional 3D perception pipelines rely on supervised detectors trained on closed sets of categories, limiting their ability to generalize. In contrast, recent works use VLM embeddings to recognize a wide variety of objects and relationships without task-specific retraining. For instance, ConceptGraphs [2] utilize CLIP embeddings to classify segmented 3D regions with arbitrary textual prompts, enabling recognition of novel categories in real-time. Similarly, From Pixels to Graphs [3] generates open-vocabulary scene graphs by grounding visual regions in VLM embedding spaces, which significantly improves flexibility in object and predicate recognition.

B. Multi-Modal Fusion for 3D Understanding

Constructing 3D scene graphs requires fusing multi-modal data—typically RGB images, depth, and pose information. VLM-based approaches incorporate semantic cues from images and natural language with spatial geometry to form coherent graphs. For example, MR-COGraphs [10] integrates multi-robot RGB-D and LiDAR data with open-vocabulary labels to produce consistent semantic maps across agents. Language-Embedded Gaussian Splats (LEGS) [8] incrementally build room-scale 3D representations by fusing geometric splats with language embeddings, enabling fine-grained semantic labeling with spatial context.

C. Incremental and Long-Term Scene Graph Construction

Robots operating in dynamic, real-world environments require continuous updates to their semantic maps. Dynamic Open-Vocabulary 3D Scene Graphs [7] address this by incrementally refining object and relation labels as new observations arrive, maintaining temporal consistency. Lost & Found [9] extends this idea by explicitly tracking changes in scene graphs from egocentric perspectives, enabling detection of moved, added, or removed objects over time.

These incremental approaches leverage the generalization capabilities of VLMs to accommodate unseen objects and evolving scenes, facilitating robust long-term autonomy.

D. Leveraging Large Language Models for Scene Graph Enrichment

While vision-language models can generate object labels and relationships, large language models (LLMs) provide broader world knowledge and commonsense reasoning capabilities, which are increasingly being leveraged to enrich scene graphs beyond direct visual input.

Intelligent Spatial Perception [11] demonstrates how LLMs can guide hierarchical 3D scene graph generation by inferring object affordances, relationships, and task-relevant attributes.

TABLE I
COMPARISON OF KEY PAPERS ON 3D SCENE GRAPHS FOR ROBOTIC PERCEPTION AND PLANNING

Paper	Year	Vocab	Graph	Spatial	Temp.	VLM	LLM	Method / Contribution
ConceptGraphs [2]	2023	Open	3D SG	Metric	Static	LLaVA-7B	GPT-4	Open-vocab 3D graphs using CLIP for object grounding and planning.
From Pixels to Graphs [3]	2023	Open	2D/3D SG	Projective	Static	BLIP	-	Generates scene graphs directly from images using CLIP.
MR-COGraphs [10]	2023	Open	Multi-Robot 3D	Metric	Static	CLIP	-	Multi-agent open-vocab mapping via efficient CLIP-based 3D graphs.
Grounding Classical Task Planners [6]	2023	Open	Symbolic SG	Metric + Symbolic	Static	ViLBERT	-	Connects VLM-based perception with symbolic planners.
OpenGraph [4]	2024	Open	Hier. 3D Graph	Hybrid	Static	RAM, TAP, Grounding DINO	SBERT, LLaMA	Scalable hierarchical 3D graph for large-scale outdoor environments.
Dynamic 3D SGs [7]	2024	Open	Incr. 3D SG	Metric	Dynamic	ecognize-Anything,SAM-2, Grounding DINO,CLIP	GPT-4o	Builds dynamic 3D graphs incrementally for long-horizon manipulation.
Lost & Found [9]	2024	Open	Dyn. 3D SG	Metric	Dynamic	-	-	Tracks object-level changes over time in egocentric views.
LEGS [8]	2024	Open	3D Splat Graph	Metric	Incr.	CLIP	-	Builds spatial graphs from VLM-labeled Gaussian splats in real time.
Intelligent Spatial Perception [11]	2024	Open	Hier. 3D SG	Metric + Topo	Static	CLIP	ERNIE 3.5-8K	Uses LLMs to annotate and structure semantic indoor maps.
RoboHop [13]	2024	Open	Topo Graph	Topo	Static	CLIP	GPT-4	Object-centric topo maps for language-based visual navigation.
Topo-Field [12]	2024	Open	Topo-Field Graph	Topo-Metric	Static	CLIP	Sentence-BERT	Brain-inspired layout-object topology for map abstraction.
IMP [15]	2017	Closed	2D SG	Projective	Static	-	-	Image-based scene graph generation using message passing.
MOTIFNET [16]	2018	Closed	2D SG	Projective	Static	-	-	Captures object and relation patterns from Visual Genome.

Specifically, LLMs are prompted with region-level and object-level descriptions to generate scene semantics, which are then structured into hierarchical graphs spanning layout, objects, and actions. This enables the graph to represent not just 'what is where,' but also 'what can be done with it'—for example, identifying a 'table in the dining room' as suitable for placing items or hosting activities.

E. Challenges and Limitations

Despite their promise, VLM-based perception methods face challenges including domain gaps between training data and robotic environments, ambiguity in open-vocabulary predictions, and computational demands for real-time deployment. Balancing accuracy, efficiency, and generalization remains an active area of research. Moreover, grounding natural language in 3D space to ensure spatially consistent and actionable scene graphs is a non-trivial problem requiring further advances. To provide a unified comparison of the main contributions, methods, and system characteristics of recent works, Table I summarizes key aspects of surveyed papers.

F. Comparison of ConceptGraphs and DovSG Architectures

ConceptGraphs [2] and DovSG [7] represent two influential systems in the development of open-vocabulary 3D scene graphs for robotic perception and planning.

ConceptGraphs emphasizes the construction of a rich semantic scene graph by combining RGB-D perception with open-vocabulary segmentation and vision-language model (VLM)-based relational reasoning. As shown in Figure 1, the pipeline incrementally fuses multi-view features into a unified 3D map and uses VLMs to both caption objects and infer their interrelations. This results in a graph structure well-suited for symbolic reasoning and high-level planning.

DovSG, on the other hand, prioritizes long-term robustness and adaptability in dynamic environments. Figure 2 illustrates a complete architecture integrating perception, dual-level memory, planning, and execution. Its scene graph is continuously updated in real time, allowing the system to detect manual changes (e.g., moved objects) and revise its action plan accordingly. This reflects a shift from static mapping toward persistent semantic memory, enabling robust performance in open-world scenarios.

Both architectures use LLMs or vision-language models for semantic grounding and planning. However, DovSG places stronger emphasis on temporal consistency and memory updates, while ConceptGraphs provides a cleaner abstraction for downstream graph-based reasoning tasks. Their complementary focuses highlight the emerging need to unify semantic perception, graph-based world modeling, and LLM-informed decision making in mobile robotics.

V. PLANNING WITH 3D SCENE GRAPHS

3D scene graphs provide structured, semantically rich representations of the environment that are well-suited to support various levels of robotic planning, from high-level task execution to low-level motion control. This section surveys approaches leveraging 3D scene graphs and vision-language models (VLMs) to enable flexible, language-guided, and context-aware planning.

A. Semantic Task Planning

Traditional task planners such as those based on PDDL operate on symbolic world models using pre-defined predicates and object categories. However, grounding these symbolic models in perceptual data remains a key challenge, due to the semantic gap between raw sensor observations and high-level symbolic representations. Grounding Classical Task Planners via Vision-Language Models [6] addresses this issue by proposing TPVQA, a vision-based symbolic planning framework that leverages Vision-Language Models (VLMs) to bridge this gap. As illustrated in Fig. 3, TPVQA exploits the preconditions and effects defined in classical planners as a source of domain knowledge, and formulates grounding as a series of Visual Question Answering (VQA) tasks. By querying VLMs to verify whether preconditions are satisfied or whether the intended effects of actions have been achieved,

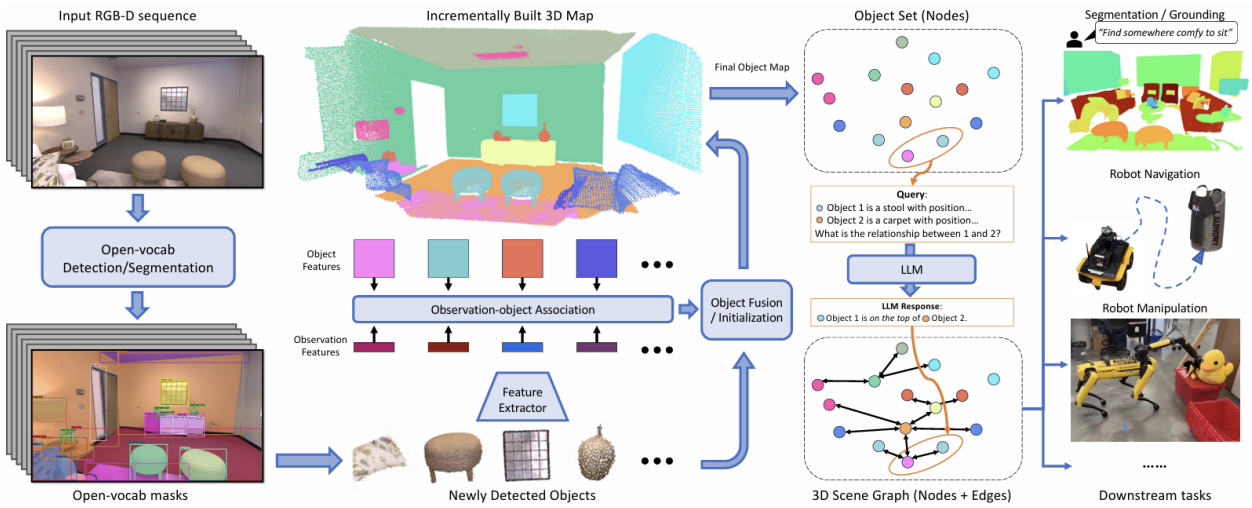

Fig. 1. Overview of the ConceptGraphs system [2]. ConceptGraphs builds an open-vocabulary 3D scene graph from a sequence of posed RGB-D images. Generic instance segmentation models segment RGB frames, semantic features are extracted and fused into 3D space to form a map of objects. Vision-language models then generate captions and inter-object relations, resulting in a structured graph useful for high-level reasoning and planning.

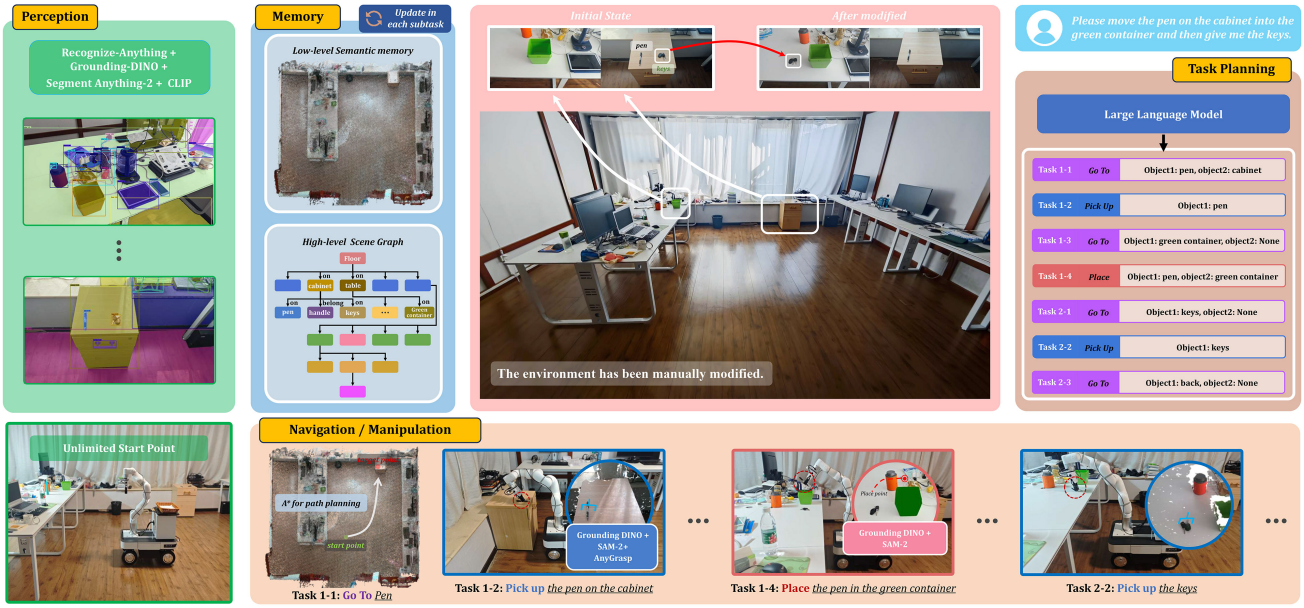

Fig. 2. Overview of the DovSG system. DovSG is a mobile robotic system designed to perform long-term tasks in real-world environments. It detects changes in the scene during task execution, ensuring subtask accuracy. It consists of perception, memory, planning, navigation, and manipulation modules. Its memory system includes both a low-level semantic map and a high-level scene graph, continuously updated to support robust, reactive planning.

the robot can either re-execute failed actions or dynamically re-plan based on the updated state of the world. This enables robust execution of open-ended tasks in perceptually grounded environments by combining the generalization ability of VLMs with the structured reasoning of symbolic planners.

B. Language-Guided Manipulation and Navigation

Open-vocabulary 3D scene graphs facilitate language-guided planning in unstructured environments. Dynamic Open-Vocabulary 3D Scene Graphs [7] demonstrate mobile manipulation planning where a robot incrementally updates its semantic map and executes actions specified via natural language commands. RoboHop [13] uses segment-based topological maps derived from 3D scene graphs to perform open-world visual navigation that adapts to language-guided goals without reliance on fixed object sets.

These approaches enable robots to flexibly plan and replan in response to changing environments and task specifications, improving robustness and autonomy.

C. Hierarchical and Multi-Agent Planning

Hierarchical scene graphs with multiple levels of abstraction support planning at different spatial and semantic scales. OpenGraph [4] employs hierarchical 3D graph representations that enable reasoning from object-level interactions up to large-scale outdoor navigation. Similarly, MR-COGraphs [10]

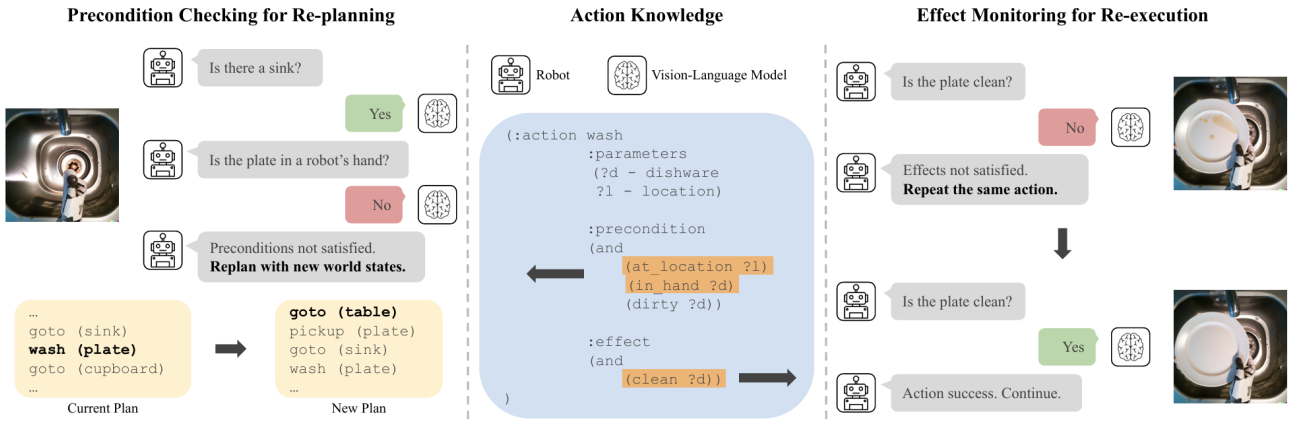

Fig. 3. Overview of TPVQA [6]. By formulating symbolic grounding as a set of VQA tasks, the robot can verify preconditions and effects of actions via VLM queries, enabling dynamic replanning or re-execution when needed.

extend semantic mapping to multi-robot systems, allowing distributed planning and coordination via shared scene graph structures.

Hierarchical and multi-agent frameworks promote scalable planning and effective resource sharing in complex environments such as warehouses, smart cities, and collaborative robotics.

D. Challenges in Planning with Scene Graphs

Despite significant progress, integrating 3D scene graphs into real-world robotic planners poses challenges. The complexity of dynamically updating large-scale graphs, ensuring semantic and geometric consistency, and interpreting ambiguous language commands remain open problems. Additionally, balancing computational efficiency with rich semantic representation is critical for deployment on resource-constrained robotic platforms.

Ongoing research is addressing these issues by developing incremental graph update algorithms, and incorporating large language models for robust command understanding.

VI. Discussion and Future Directions

The integration of 3D scene graphs and vision-language models has significantly advanced semantic perception and planning capabilities in robotics. However, several open research directions remain to fully realize their potential in real-world applications.

A. Towards Lifelong and Continual Scene Understanding

Most current systems assume episodic or short-term interaction with the environment. Lifelong robotic systems must move beyond static representations to maintain scene graphs that evolve continuously—integrating new observations, detecting contradictions, and pruning obsolete or unreliable data. This requires not only mechanisms for long-term memory, spatiotemporal reasoning, and semantic consistency across sessions, but also a form of meta-cognition. Meta-cognitive capabilities allow a robot to monitor and evaluate its own knowledge state: for instance, recognizing when it lacks sufficient information about a scene, when its predictions may be unreliable, or when conflicting information requires resolution.

Such capabilities are especially relevant for open-world settings, where robots encounter novel objects, ambiguous environments, or contradictory cues over time. By embedding uncertainty estimates and novelty detection directly into the scene graph structure (e.g., confidence scores or unknown-node placeholders), robots can explicitly represent what is known, what is uncertain, and what is unknown. Furthermore, meta-cognitive graphs may trigger active learning behaviors, such as re-exploring regions with low-confidence labels or asking for human assistance. Enabling these self-aware mechanisms is a key step toward persistent and trustworthy scene understanding in lifelong deployments.

B. Uncertainty and Robustness

Open-vocabulary scene graphs rely on embeddings from pre-trained VLMs like CLIP, but these models can yield uncertain or inconsistent predictions in unfamiliar settings or under noisy input. For instance, a robot may confuse a "remote" with a "smartphone" in cluttered scenes, or misinterpret vague commands like "put it near the box" when multiple boxes are present. These ambiguities can lead to unreliable scene graph construction and suboptimal planning.

To mitigate this, incorporating uncertainty estimates—such as confidence scores—into scene graphs can help flag ambiguous predictions and guide more cautious decision-making. Building on this, hybrid symbolic-neural methods can enhance reliability by verifying neural outputs against logical or physical constraints. For example, a neural module might suggest object identities or spatial relationships, which a symbolic layer can then validate (e.g., ensuring a "cup" is placed on a surface, not floating in space). Finally, confidence-aware planning enables robots to adapt their behavior in response to uncertain perception—such as selecting alternative actions, asking for help, or collecting more data before proceeding.

C. Multi-Modal Reasoning and Integration

While most current approaches to open-vocabulary 3D scene graphs rely on RGB-D input and natural language, real-world robotic perception often demands richer sensory integration. Modalities such as audio, tactile feedback, force sensing, or proprioception offer complementary information that can disambiguate perception and enhance robustness in unstructured environments.

For instance, a robot might fail to visually recognize whether a cabinet is open or closed due to occlusion, but tactile or auditory cues (e.g., resistance or a door-click sound) could provide reliable indicators. Similarly, touch feedback can help differentiate between a hard plastic bottle and a soft paper cup, even if their visual features are similar.

Integrating such multi-modal inputs into scene graph representations opens new opportunities for embodied reasoning. Future research should explore extending scene graphs to include haptic properties, auditory signatures, and temporal interactions. A unified representation that fuses VLM-based visual understanding with physical and temporal cues could lead to more grounded and resilient systems, particularly for manipulation and long-horizon interaction tasks.

Moreover, developing cross-modal alignment mechanisms—e.g., grounding linguistic descriptors like ”squishy” or ”buzzing” to tactile or audio features—can enrich the semantic expressiveness of scene graphs and enhance robot understanding of object affordances and dynamics.

D. Interfacing with Large Language Models

Large language models (LLMs) are increasingly being employed to enrich scene graphs with semantic and contextual information, such as object affordances, relationships, and task relevance. Beyond enrichment, deeper integration of LLMs holds the potential for enabling complex task planning, dialogue-based interaction, and knowledge transfer in robotics. Future systems may rely on LLMs not only to interpret user instructions, but also to reason about goals, infer implicit knowledge, or synthesize plans from abstract scene representations. However, despite their improved capabilities, recent LLMs often exhibit a higher tendency for hallucination—generating plausible-sounding but incorrect or nonexistent objects, relationships, or actions. This raises serious concerns when LLMs are used in real-world decision-making pipelines. Ensuring consistency between LLM-generated content and grounded sensor data, possibly through validation modules or hybrid symbolic-neural approaches, will be critical for safe and reliable deployment.

E. Human-Robot Collaboration and Shared Representations

As robots become integrated into daily environments, their ability to collaborate effectively with humans hinges on establishing shared representations of the world. Open-vocabulary 3D scene graphs, with their interpretable structure and semantic richness, offer a promising interface for communication, joint planning, and error recovery in human-robot interaction (HRI).

For example, in a collaborative assembly task, a human might say, “Hand me the small red bowl next to the toaster.” The robot must not only visually ground the query in its scene graph but also verify whether its interpretation aligns with the human’s intent. The structured nature of scene graphs facilitates such referential grounding and enables the robot to ask clarifying questions (e.g., “Do you mean the red plastic bowl or the ceramic one?”).

Future research could explore ways to align human and robot mental models through shared scene graph updates, multimodal dialogue, or visual explanation systems. Scene graphs may also incorporate interaction history, task context, or human gaze and gesture to enhance referential accuracy and collaborative fluency.

Additionally, the development of bidirectional interfaces—where both the robot and the human can modify or annotate the scene graph—could lead to more transparent and adaptive collaboration. This aligns well with the broader vision of explainable and trustworthy robotics.

F. Benchmarks and Standardization

The diversity of representations, task settings, and datasets currently hinders direct comparison between open-vocabulary 3D scene graph approaches. For example, ConceptGraphs [2] builds dense, object-centric graphs from RGB-D scans, while DOVSG [7] emphasizes dynamic updates for long-term manipulation tasks—yet these systems are evaluated using entirely different protocols, making it difficult to assess relative performance.

Furthermore, many datasets, such as ScanNet or Replica, were not originally designed with open-vocabulary grounding or long-horizon planning in mind. As a result, they lack annotated language instructions, action affordances, or scene graph ground truth. This limits the ability to benchmark performance on tasks that require grounding of natural language into perception and planning.

To address this, there is a growing need for standardized datasets and evaluation metrics tailored to open-vocabulary 3D scene graph tasks. These should include:

- Rich language annotations aligned with 3D scenes.
- Ground-truth scene graphs with object relationships and attributes.
- Benchmarks that span perception, symbolic grounding, task planning, and manipulation.
- Metrics that evaluate graph accuracy, grounding precision, planning success, and task completion rate.

One possible direction is extending existing datasets with scene graph annotations and open-vocabulary queries. Additionally, simulation environments could support standardized task suites (e.g., pick-and-place, object search) where different systems are evaluated under identical conditions like for example in the same virtual environment. By converging on shared benchmarks, the field can support more rigorous, fair, and reproducible comparisons across systems.

VII. Conclusion

This paper surveyed recent advances in 3D scene graph systems for robotic perception and planning, with a focus

on the integration of vision-language models (VLMs). We categorized representative approaches across key dimensions such as vocabulary openness, 3D spatial structure, and temporal adaptability. We reviewed how VLMs are leveraged for open-vocabulary object recognition, semantic mapping, and grounding of symbolic planners in perceptual inputs. Furthermore, we analyzed challenges related to uncertainty, memory, generalization, and scene dynamics.

In our discussion, we highlighted emerging directions including multi-modal scene graph construction, human-robot shared representations, hybrid symbolic-neural planning, and the need for standardized benchmarks. These perspectives point toward the development of more robust, adaptive, and interactive robotic systems capable of operating in complex, open-world environments.

As robots are increasingly deployed in real-world settings, the synergy between structured 3D representations and the flexibility of language models is poised to enable more general-purpose autonomy, long-term adaptation, and collaborative intelligence.

REFERENCES

[1] I. Armeni, Z.-Y. He, J. Gwak, A. R. Zamir, M. Fischer, J. Malik, and S. Savarese, "3d scene graph: A structure for unified semantics, 3d space, and camera," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5664–5673.

[2] Q. Gu, A. Kuwajerwala, S. Morin, K. M. Jatavallabhula, B. Sen, A. Agarwal, C. Rivera, W. Paul, K. Ellis, R. Chellappa *et al.*, "Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 5021–5028.

[3] R. Li, S. Zhang, D. Lin, K. Chen, and X. He, "From pixels to graphs: Open-vocabulary scene graph generation with vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 28 076–28 086.

[4] Y. Deng, J. Wang, J. Zhao, X. Tian, G. Chen, Y. Yang, and Y. Yue, "Opengraph: Open-vocabulary hierarchical 3d graph representation in large-scale outdoor environments," *IEEE Robotics and Automation Letters*, 2024.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[6] X. Zhang, Y. Ding, S. Amiri, H. Yang, A. Kaminski, C. Esselink, and S. Zhang, "Grounding classical task planners via vision–language models," in *ICRA Workshop on Robot Execution Failures and Failure Management Strategies*, 2023, https://arxiv.org/abs/2304.08587.

[7] Z. Yan, S. Li, Z. Wang, L. Wu, H. Wang, J. Zhu, L. Chen, and J. Liu, "Dynamic open-vocabulary 3d scene graphs for long-term language-guided mobile manipulation," *IEEE Robotics and Automation Letters*, vol. 10, no. 5, pp. 4252–4259, 2025.

[8] J. Yu, K. Hari, K. Srinivas, K. El-Refai, A. Rashid, C. M. Kim, J. Kerr, R. Cheng, M. Z. Irshad, A. Balakrishna *et al.*, "Language-embedded gaussian splats (legs): Incrementally building room-scale representations with a mobile robot," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 13 326–13 332.

[9] T. Behrens, R. Zurbrügg, M. Pollefeys, Z. Bauer, and H. Blum, "Lost & found: Tracking changes from egocentric observations in 3d dynamic scene graphs," *IEEE Robotics and Automation Letters*, 2025.

[10] Q. Gu, Z. Ye, J. Yu, J. Tang, T. Yi, Y. Dong, J. Wang, J. Cui, X. Chen, and Y. Wang, "Mr-cographs: Communication-efficient multi-robot open-vocabulary mapping system via 3d scene graphs," *IEEE Robotics and Automation Letters*, vol. 10, no. 6, pp. 5713–5720, 2025.

[11] Y. Cheng, Z. Han, F. Jiang, H. Wang, F. Zhou, Q. Yin, and L. Wei, "Intelligent spatial perception by building hierarchical 3d scene graphs for indoor scenarios with the help of llms," in *2024 WRC Symposium on Advanced Robotics and Automation (WRC SARA)*. IEEE, 2024, pp. 483–490.

[12] J. Hou, W. Guan, L. Liang, J. Feng, X. Xue, and T. Zeng, "Topo-field: Topometric mapping with brain-inspired hierarchical layout-object-position fields," *IEEE Robotics and Automation Letters*, vol. 10, no. 6, pp. 5385–5392, 2025.

[13] S. Garg, K. Rana, M. Hosseinzadeh, L. Mares, N. Sünderhauf, F. Dayoub, and I. Reid, "Robohop: Segment-based topological map representation for open-world visual navigation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 4090–4097.

[14] H. Li, G. Zhu, L. Zhang, Y. Jiang, Y. Dang, H. Hou, P. Shen, X. Zhao, S. A. A. Shah, and M. Bennamoun, "Scene graph generation: A comprehensive survey," *Neurocomputing*, vol. 566, p. 127052, 2024.

[15] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3097–3106.

[16] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5831–5840.