Review

Spatial Intelligence from a Cognitive Map Perspective: A Survey

Yuxuan Tian [†] , Yuheng Ji [‡,†] , Xiaolong Zheng [*,†] , Ziheng Qin , Yipu Wang , Xinyi Zheng , Yuyang Liu , Shuanghao Bai , Zhe Li , Liang Wang , Daniel Dajun Zeng

Posted Date: 26 May 2026

doi: 10.20944/preprints202605.1782.v1

Keywords: spatial intelligence; cognitive map; 3D scene understanding; spatial reasoning; embodied AI

 doi:10.20944/preprints202605.1782.v1

Review

Spatial Intelligence from a Cognitive Map Perspective: A Survey

Yuxuan Tian [1,2,†], Yuheng Ji [1,2,†,‡], Xiaolong Zheng [1,2,†,*], Ziheng Qin [1,2], Yipu Wang [1,3], Xinyi Zheng [4], Yuyang Liu [1,2], Shuanghao Bai [5], Zhe Li [6], Liang Wang [1,2] and Daniel Dajun Zeng [1,2]

[1] State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences
[4] Beihang University
[5] Xi'an Jiaotong University
[6] Nanyang Technological University
[*] Correspondence: xiaolong.zheng@ia.ac.cn
[†] Equal contribution.
[‡] Project leader.

Abstract

Spatial intelligence requires agents to form and utilize internal representations of the physical world for perception, reasoning, and generation. While recent advances in foundation models, embodied systems, and three-dimensional representation learning have substantially expanded spatial capabilities, existing research remains fragmented across heterogeneous tasks and model paradigms. This survey revisits spatial intelligence from a cognitive map perspective and positions cognitive maps as its representational blueprint. In this view, diverse lines of research can be understood through a shared question: how an internal spatial representation is constructed, maintained, reasoned over, and realized. To make this perspective operational, we define cognitive maps as internal spatial representations characterized by *abstraction*, *globality*, and *persistency*. Based on this definition, we organize the literature into three cognitive-map-centric processes that correspond to the core dimensions of spatial intelligence: *perception* for cognitive map construction, *reasoning* for internal inference with the map, and *generation* for external realization of the map. By adopting a mechanism-centric viewpoint, this survey connects previously isolated research directions into a coherent framework and identifies emerging challenges toward unified spatial intelligence systems. The related resources of this study are accessible at https://github.com/Klingsor-tyx/Awesome-Spatial-Cognitive-Map.

Keywords: spatial intelligence; cognitive map; 3D scene understanding; spatial reasoning; embodied AI

1. Introduction

Spatial intelligence refers to the ability to form and use internal representations of the physical world to support perception, reasoning, and generation. In humans, this capacity enables coherent navigation, object manipulation, and mental simulation of spatial transformations [1–3]. For artificial agents, we organize the literature through three interrelated dimensions: **Perception**, which constructs structured spatial representations from fragmented, local, and often incomplete observations; **Reasoning**, which infers relationships, states, and dynamic changes in the environment with such representations to support planning and decision-making; and **Generation**, which externalizes internal representations into scene synthesis or world simulation.

Recent advances in foundation models, embodied agents, and three-dimensional representation learning have significantly expanded the scope of these capabilities[4–6]. Spatial perception has progressed from passive two-dimensional recognition toward holistic 3D scene understanding, semantic mapping, and cross-view alignment[7–10]. Spatial reasoning has extended from local relational judgment to long-horizon navigation, multi-view inference, and counterfactual simulation[11–17]. Spatial

 doi:10.20944/preprints202605.1782.v1

generation has evolved from object-level synthesis to structured scene construction and dynamic environment simulation[18–23]. As these tasks grow in temporal depth and structural complexity, these seemingly disparate research directions are in fact converging on a more fundamental common requirement: agents need a unified internal representation that persists beyond transient sensory inputs, maintains global consistency under partial observability, and support closed-loop behavior over extended time spans.

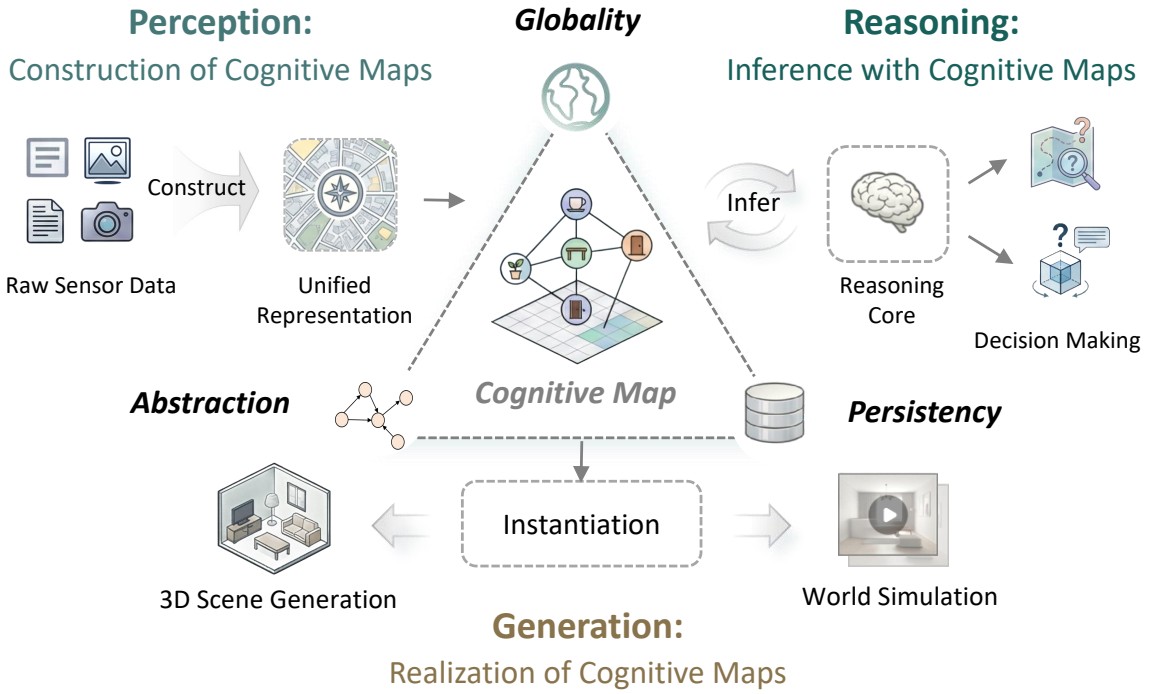

Figure 1. The unified framework of spatial intelligence through the lens of cognitive maps. The cognitive map serves as the central architectural blueprint that bridges three core domains: (1) Perception, which drives the construction of globally consistent representations from fragmented observations; (2) Reasoning, which performs spatial inference on these representations for long-horizon planning and decision-making; and (3) Generation, which guides the realization of the internal map into 3D environments or world simulations.

This further raises a more fundamental question: *what kind of representation can serve as a common foundation across spatial perception, reasoning, and generation?* In this survey, we argue that this question can be understood through the perspective of cognitive maps. In other words, cognitive maps should not be viewed as merely one parallel branch within spatial intelligence, but rather as the underlying blueprint of internal representation that makes spatial intelligence possible.

1.1. Cognitive Maps: Blueprint and Definition

The concept of cognitive maps originally emerged from studies of spatial cognition in biological agents, where it was used to describe the internal spatial representations that agents form in environments [24–26]. In this survey, rather than confine the term to its classical navigational meaning, we generalize the cognitive map as an engineering abstraction for spatial intelligence systems: a unified internal spatial representation framework that underlies the three core processes of perception, reasoning, and generation. Specifically, an agent constructs this representation from local and fragmented observations, performs inference and planning over it, and further instantiates it through environment generation or world simulation. In task settings characterized by partial observability, long temporal horizons, and cross-view reasoning, such a representation serves as the critical intermediary between immediate sensory input and globally coherent spatial behavior. To make this concept operational, we define a cognitive map as an internal spatial representation that simultaneously satisfies abstraction, globality, and persistence, as illustrated in Fig. 2.

 doi:10.20944/preprints202605.1782.v1

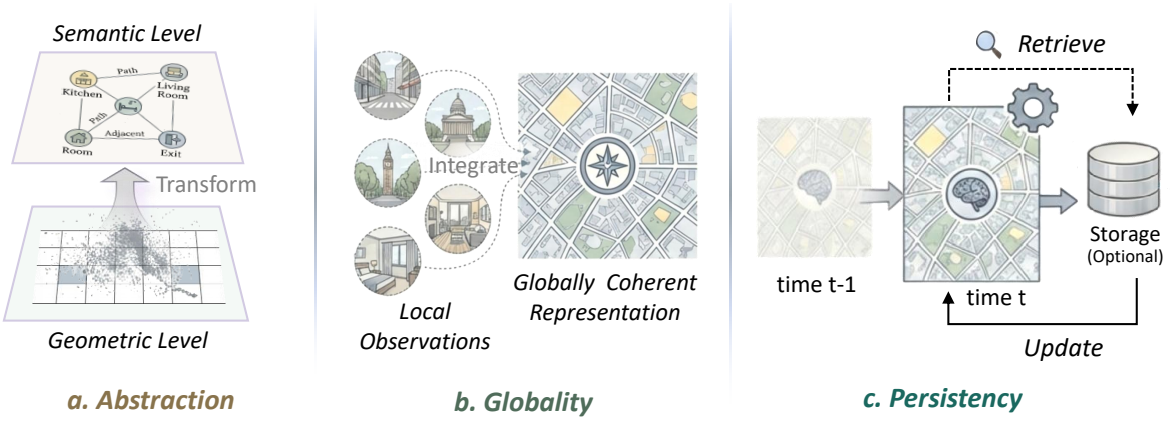

Figure 2. The definition of cognitive map in the context of spatial intelligence. A unified internal representation must simultaneously satisfy three fundamental properties: (a) **Abstraction**, which transforms raw sensory inputs into structured semantic entities and relations; (b) **Globality**, which integrates partial, instantaneous observations into a holistic and cross-view consistent spatial layout; and (c) **Persistency**, which ensures the continuous maintenance and dynamic updating of the internal state over time.

- **Abstraction** requires the encoding of semantically structured concepts derived from raw sensory streams. A cognitive map should transform low-level geometric inputs such as voxels or point clouds into structured entities, attributes, and relations, potentially including topological structures, thereby bridging perception and cognition.
- **Globality** refers to spatial coherence beyond the instantaneous field of view. By integrating partial observations across time and viewpoints, the cognitive map forms a holistic representation whose spatial extent scales from local functional areas to large architectural layouts. This property ensures cross-view consistency, consolidating structures observed from different perspectives into a unified entity.
- **Persistency** enables the continuous maintenance and update of spatial representations through memory mechanisms. The cognitive map functions as an evolving internal state rather than a single-pass byproduct of perception. This property supports dynamic updates like tracking object relocation, without reconstructing the entire environment at every step.

Together, these three properties constitute the operational criteria for a cognitive map. Only when a representation possesses abstraction, globality, and persistence at the same time does it go beyond merely describing space and begin to genuinely support spatial intelligence.

More importantly, from this perspective, the value of a cognitive map lies not in storing an additional layer of spatial information, but in specifying the fundamental mode of operation of a spatially intelligent system. The system must first abstract raw observations into a structured representation, then integrate local cues distributed across time and viewpoints into a globally coherent spatial layout, and continuously maintain this representation so that it can be repeatedly queried, updated, and validated throughout interaction. For this reason, spatial perception, reasoning, and generation are not three isolated classes of tasks, but three consecutive processes organized around the same internal representation: perception for constructing the cognitive map, reasoning for inferring states and making decisions over it, and generation for realizing it into externalized environments or simulated outcomes. Research directions that have often been studied in isolation under different terminologies, including metric-semantic mapping, scene graphs, spatial memory systems, and structured world models, can be reinterpreted as different instantiations of the same core problem: how to construct, maintain, reason over, and realize a cognitive map.

1.2. Related Work and Positioning

Existing surveys on spatial intelligence have largely followed two main paths. One line of work is task-centered, addressing specific problems such as scene graph learning [27,28], Vision-

 doi:10.20944/preprints202605.1782.v1

and-Language Navigation [29], and 3D generation [30,31], and typically organizing the literature by application scenario or task category. The other line is model-centered. Especially with the rise of large language models (LLMs) and vision-language models (VLMs), these surveys place greater emphasis on adaptation strategies such as prompting, fine-tuning, and modality alignment, as well as on the performance boundaries of such models in spatial capabilities [5,32,33]. Other works extend spatial intelligence to interdisciplinary contexts such as urban systems [34] or explore neuroscience-inspired parallels between biological circuits and artificial agents [35].

Taken together, these studies address questions such as what spatial intelligence does, what models can be used to realize it, and with which domains it can be associated. However, they devote relatively limited attention to a more fundamental issue: whether there exists a shared internal representational mechanism underlying different spatial tasks and modeling paradigms. In fact, despite many seemingly heterogeneous approaches including metric-semantic mapping for geometric and semantic grounding [36,37], scene graphs for relational modeling [38,39], spatial memory for long-term maintenance [40,41], and world models for predictive modeling [42,43], differ in form, input-output structure, and application setting, they are all essentially attempting to construct a structured internal spatial representation.

Motivated by this observation, this survey departs from conventional task-centric and model-centric perspectives in favor of a mechanism-centric view. We treat cognitive maps as the computational substrate of spatial intelligence and systematically examine how such internal representations are constructed, maintained, deployed for reasoning, and further realized in generation and interaction. Our goal is not merely to offer another taxonomy for spatial intelligence, but to provide a unified interpretive framework for these fragmented research that reveals their shared principles of internal representational mechanisms.

1.3. Organization and Contributions

The remainder of this survey is organized as follows. Section 2 discusses how cognitive maps are constructed through spatial perception, with emphasis on the formation mechanisms of different types of internal spatial representations. Section 3 examines how cognitive maps support spatial reasoning, focusing on how reasoning modules read, manipulate, and make use of internal spatial representations. Section 4 discusses how cognitive maps are realized in spatial generation and world simulation, characterizing the process through which internal representations are transformed into external spatial forms. Section 5 further connects these capabilities to representative applications in both open-loop and closed-loop interactive settings. Section 6 outlines the open challenges and discusses promising future directions.

This survey makes the following contributions: First, we extend the notion of cognitive maps from its classical role in biological navigation to a unified foundation of internal representation for modern spatial intelligence, proposing an operational definition centered on abstraction, globality, and persistence. Second, we introduce a mechanism-centric unified perspective that reorganizes spatial intelligence into three fundamental processes centered on cognitive maps: construction, reasoning, and realization. Third, under this framework, we systematically review the literature on spatial intelligence, revealing the shared representational mechanisms across different methodologies. We further identify key trends and open problems toward unified spatially intelligent systems.

 doi:10.20944/preprints202605.1782.v1

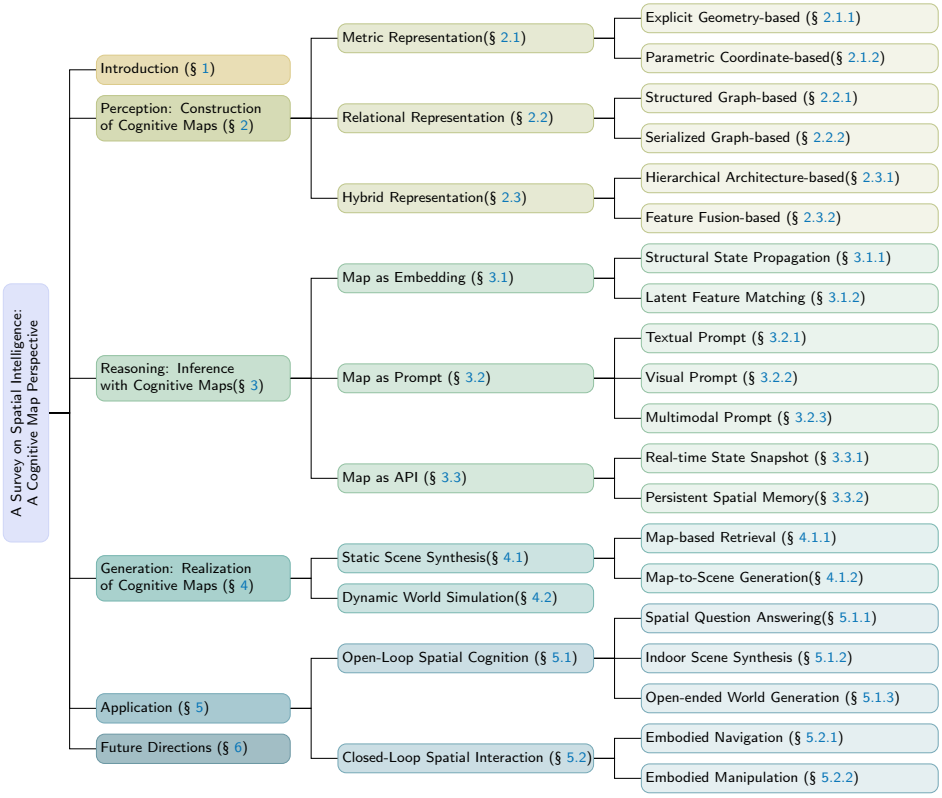

Figure 3. The overall structure of this survey. Based on our proposed taxonomy unifying perception, reasoning, and generation as the construction, inference, and realization of cognitive maps, we deliver a comprehensive review highlighting emerging trends toward spatial intelligence systems.

2. Perception: Construction of Cognitive Maps

In spatial intelligence systems, the role of perception extends beyond conventional visual or sensory processing. More fundamentally, it concerns how to construct a unified internal spatial representation with globality and abstraction from raw, transient, and typically local sensor observations, such as images or video streams. This representation, referred to as a cognitive map in this survey, embodies the transformation from instantaneous measurements of the physical world to an internalized spatial model, constituting a critical transition from data-driven perception to cognition-level understanding.

From the perspective of internal structure and information organization, existing research on cognitive maps can be categorized into three spatial paradigms, as shown in Fig. 4 : (i) *Metric Representations*, which emphasize precise geometric structure and physical attributes of space; (ii) *Relational Representations*, which focus on modeling topological organization and semantic dependencies within the environment; and (iii) *Hybrid Representations*, which jointly model metric and relational information to enable spatial understanding and reasoning across different levels of abstraction. The representative methods under each paradigm is summarized in Tab. 1.

 doi:10.20944/preprints202605.1782.v1

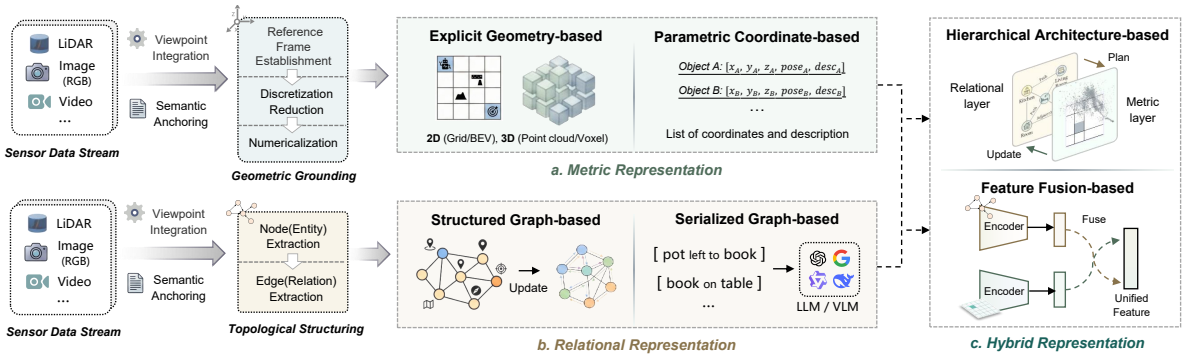

Figure 4. Three spatial representation paradigms for cognitive map construction: (a) Metric representations emphasizing precise geometric structures and physical attributes; (b) Relational representations focusing on topological organization and semantic dependencies; and (c) Hybrid representations jointly integrating both for complementary multi-level spatial reasoning.

2.1. Metric Representation

Metric representations emphasize the construction of cognitive maps that explicitly encode metric information during the perception stage, with the central objective of quantitatively modeling the physical properties of space. These approaches focus on precise geometric characterization of objects and structures within the environment by explicitly incorporating metric elements such as coordinates, distances, scale, volume, and three-dimensional shape. As a result, the internal spatial representations exhibit geometric consistency and computational tractability. The resulting cognitive maps are typically grounded in continuous or discrete geometric structures, emphasizing accurate reconstruction of spatial layouts as well as geometric alignment across viewpoints and over time.

2.1.1. Explicit Geometry-based

Explicit geometry-based methods aim to directly construct geometric representations of the environment, emphasizing the recovery or accumulation of measurable and spatially aligned information from perceptual inputs to form cognitive maps with explicit physical meaning. From the perspective of geometric representation, such methods can be broadly categorized into 2D planar representations and 3D geometric representations.

2D Planar-based methods project 3D spatial observations onto a unified 2D representation, enabling geometric reasoning on a planar manifold [37,44–65]. Such representations are typically instantiated as occupancy grids, cost maps, or bird's-eye-view (BEV) layouts. Representative approaches differ in how semantic information is integrated into the planar map. GridMM [58] projects historical observations into a unified egocentric grid using depth and pose information, and aggregates visual features according to their relevance to the navigation instruction. MapNav [48] constructs annotated semantic maps by projecting point clouds onto a 2D plane and attaching textual labels to regions and objects. GPT4Scene [47] reconstructs a BEV map from video streams, providing a global planar layout that facilitates holistic spatial understanding. OneMap [65] constructs a 2D planar semantic belief map that accumulates multi-view semantic features into grid cells while encoding uncertainty information, thereby maintaining long-term spatial consistency in dynamic environments.

3D Geometry-based methods preserve full 3D scene structure using representations such as point clouds, voxels, or occupancy fields, integrating multi-view observations into a unified volumetric space and mitigating the distortion and occlusion limitations of 2D projections [66–70]. For instance, OccWorld [66] and DOME [68] explicitly model environments with 3D occupancy grids, providing volumetric information for motion planning and collision avoidance.

A growing body of work incorporates the temporal dimension into 3D geometric representations, enabling the construction of dynamic and persistent 3D world models [42,43,71–77]. These approaches model scene evolution over time to maintain spatial consistency under continuous interaction and environmental changes. For example, 3DLLM-Mem [71] builds a long-term spatio-temporal memory

 doi:10.20944/preprints202605.1782.v1

that maps multi-view visual observations into a unified 3D voxel space. In contrast, NeoVerse [74] introduces a Gaussian Splatting-based representation to encode dynamic scene evolution as time-varying geometric primitives, achieving high-fidelity spatial consistency and temporal continuity in monocular video streams. CogniMap3D [78] emphasizes the construction of 3D cognitive map by maintaining an accumulative geometric memory bank for persistent and efficient retrieval and update.

2.1.2. Parametric Coordinate-based

Parametric coordinate-based representations further parse continuous metric space into discrete and structured sets of coordinate parameters [41,79–86]. By explicitly encoding spatial entities as parameter tuples (e.g., object positions, orientations, and scales), these methods transform complex geometric reasoning into symbolic or numerical operations that can be more directly handled by vision-language models. For instance, representative approaches such as Ego3D-VLM [82] convert objects' global 3D coordinates into textualized metric descriptions through agent-centric perception and depth estimation, effectively reducing perspective-aware spatial reasoning to coordinate-based logical manipulation. Thinking with Blueprints [85] constructs an object-centric JSON-style blueprint that explicitly records the positions, sizes, and attributes of relevant objects in a scene. To address memory decay and backtracking requirements in long-horizon tasks, MrSteve [83] introduces a temporal axis to form spatio-temporal parameter tuples, binding precise 3D locations with event annotations to record the dynamic evolution of the environment in a parameterized manner. For more complex scenarios involving object state transitions, Embodied VideoAgent [84] further enriches the parameter space with fine-grained, real-time attribute fields, enabling coordinate maps to capture persistent changes in object properties and supporting long-term modeling of dynamic environments.

2.2. Relational Representation

Relational representations focus on characterizing the topological and semantic relationships among objects in space during the perception stage, rather than encoding precise metric information. These approaches emphasize the organization and constraints of spatial structure, capturing high-level relations such as adjacency, containment, and relative orientation between objects and regions. Existing studies primarily represent such relationships using different forms of data structures, including explicit graph-based structures and serializable symbolic or textual representations.

 doi:10.20944/preprints202605.1782.v1

Table 1. Overview of Different Representations

Category	Method	Venue	Base	Semantic Scope	Input Modality	Construction Mechanism
Metric	GridMM[58]	ICCV'23	Geometry (2D)	Open-vocabulary	RGB-D	G M
	Dynam3D[77]	NeurIPS'25	Geometry (3D)	Open-vocabulary	RGB-D	G E M
	SpNav[45]	AAAI'26	Geometry (2D)	Open-vocabulary	RGB-D	G P
	CogniMap3D[78]	ICLR'26	Geometry (3D)	Instance-specific	Video	G E
	APC[79]	ICCV'25	Coordinate	Open-vocabulary	RGB	G E
	EfficientNav[81]	NeurIPS'25	Coordinate	Open-vocabulary	RGB-D	G E
	VideoAgent[84]	ICCV'25	Coordinate	Open-vocabulary	RGB-D + Video	G E P
	ReMEmbR[41]	ICRA'25	Coordinate	Open-vocabulary	Video	P
Relational	SGM[87]	ICML'23	Structured Graph	Closed-set	RGB	E
	MemoNav[88]	CVPR'24	Structured Graph	Instance-specific	RGB-D	M
	VTSCN[89]	TPAMI'24	Structured Graph	Closed-set	RGB	E
	TB-HSU[90]	AAAI'25	Structured Graph	Closed-set	3D Point Cloud	G E
	SSGVS[91]	CVPR'23	Serialized Graph	Closed-set	RGB	E
	MapGPT[92]	ACL'24	Serialized Graph	Open-vocabulary	RGB	P
	Hi-Dyna Graph[93]	ArXiv'25	Serialized Graph	Open-vocabulary	Video	E P
	PanoNav[94]	AAAI'26	Serialized Graph	Open-vocabulary	RGB	P
Hybrid	Hydra[95]	RSS'22	Graph + Geometry	Closed-set	RGB-D	G E
	PSG-4D[96]	NeurIPS'23	Graph + Geometry	Closed-set	RGB-D + Video	G E
	BSG[97]	ICCV'23	Geometry + Graph	Closed-set	RGB-D	G E
	ConceptGraphs[98]	ICRA'24	Graph + Geometry	Open-vocabulary	RGB-D	G E M
	Sg-CityU[99]	ACM MM'24	Graph + Geometry	Closed-set	3D Point Cloud	E
	CogNav[100]	ICCV'25	Graph + Coordinate	Open-vocabulary	RGB-D	E P
	Struct2D[101]	NeurIPS'25	Geometry + Coordinate	Open-vocabulary	RGB-D + Video	E P
	ASCENT[102]	RA-L'26	Graph + Geometry	Open-vocabulary	RGB-D	G P
	SUSA[103]	AAAI'26	Graph + Geometry	Open-vocabulary	RGB	P
	GeoNav[104]	PR'26	Graph + Geometry	Open-vocabulary	RGB	G E P

Semantic Scope characterizes the breadth of semantic abstraction. **Closed-set**: Limited to a predetermined semantic space with finite categories. **Open-vocabulary**: Enables zero-shot querying of arbitrary concepts via language-driven representations. **Instance-specific**: Bypasses categorical abstraction to target distinct, individual entities. Abbreviations in the **Construction Mechanism** column denote the following: G (Geometric Reconstruction) reconstructs pure metric structures by explicitly projecting 2D pixels into a global coordinate system; E (Entity Extraction) extracts discrete semantic entities with precise boundaries; M (Feature Mapping) maps feature vectors onto continuous 3D spatial structures; P (Foundation Model Parsing) translates visual streams end-to-end into structured scene descriptions leveraging multimodal foundation models.

 doi:10.20944/preprints202605.1782.v1

2.2.1. Structured Graph-based

These methods explicitly construct perceived spatial relational information as graph-structured data. Such graph construction processes include instruction-parsed explicit relational graphs, hierarchically structured graphs and dynamically evolving relational graphs.

Some methods transform external instructions into explicit graph-structured relations by parsing language or visual prompts into semantic nodes and spatial edges in the cognitive map [89,105–110]. GC-VLN [105] uses LLMs to decompose navigation instructions into a directed acyclic graph of waypoints and objects, whose edge attributes are further translated into geometric constraints for path determination. In addition, some works represent cognitive maps as generative spatial-semantic graph priors, with node and edge attributes modeled under learned structural constraints[111–115].

To capture multi-scale spatial structure, another line of work organizes cognitive maps as hierarchical graphs, ranging from local objects to functional regions and global places[38,39,90,116–118]. For example, TB-HSU [90] constructs a three-level hierarchical 3D scene graph, where discrete object nodes are clustered into region nodes based on contextual affordances and further aggregated into room-level nodes at the top hierarchy. HiGS [116] introduces a progressive hierarchical spatial-semantic graph that dynamically organizes spatial relations and semantic dependencies, thereby supporting recursive optimization from global layouts to local details.

For continuous interaction, cognitive maps can also be modeled as dynamically evolving relational graphs that grow and update over time [87,88,119–126]. SGM [87] incorporates timestamp attributes into nodes and edges and employs node and edge predictors to infer missing or outdated topological connections. MemoNav [88] introduces a working-memory-inspired pipeline that maintains map nodes as short-term memory, aggregates them into global nodes as long-term memory, and selects task-relevant subgraphs via graph attention for online maintenance. EPoG [126] maintains an evolving belief graph of objects and their spatial relations, using LLMs to infer unobserved nodes and potential relations during exploration and manipulation.

2.2.2. Serialized Graph-Based

Serialized Graph-based methods serializes complex spatial topological relationships into one-dimensional representations, enabling originally high-dimensional and discrete graph structures to be directly processed by language models. Such serialization preserves the topological constraints among spatial entities while endowing the cognitive map with improved interpretability and semantic reasoning capacity. According to the form of serialization, existing approaches can be broadly categorized into structured symbolic sequences and natural language summaries.

The first category focuses on structured symbolic serialization of spatial topological relations [91,93,127–133]. SayPlan [127] serializes large-scale 3D scene graphs into JSON strings or Cypher query scripts, enabling hierarchical filtering and localization over complex spatial environments. KARMA [128] organizes regions and objects into structured lists containing names, states, and connectivity, constructing stable long-term semantic memory. Open Scene Graphs [129] automatically derive serialization templates based on a predefined OSG schema, mapping detected objects and structures into a unified one-dimensional symbolic framework.

Another category constructs cognitive maps through natural language summaries, transforming spatial topological relations into narrative textual representations that better align with the semantic priors of language models [92,94,134–136]. MapGPT [92] introduces an Online Linguistic-formed Map that incrementally encodes newly observed nodes and topological relations into continuous path-descriptive text during agent exploration. PanoNav [94] generates global semantic scene summaries from panoramic visual perception and stores them in a dynamically bounded queue, forming an implicit cognitive map with self-localization capability.

 doi:10.20944/preprints202605.1782.v1

2.3. Hybrid Representation

Real-world spatial cognition and decision-making rely on both metric information and relational topological knowledge, and a single representation paradigm is often insufficient. Hybrid representations aim to jointly model metric and relational information, enabling complementary reasoning across multiple levels of spatial abstraction. Existing studies follow two technical directions: one employs hierarchical architectures to decouple representation responsibilities across different levels, while the other leverages feature fusion to align heterogeneous spatial information within a shared latent space.

2.3.1. Hierarchical Architecture-Based

Within hierarchical architectures, metric representations and relational representations function as two foundational bases operating at distinct levels of abstraction. Such a layered design enables the system to process information independently at different representational scales, while simultaneously establishing the structural prerequisites for coordination between high-level abstract planning and low-level geometric perception and execution.

Early studies [137–141] typically organized these two levels in a largely decoupled manner. For example, GRAINS [139] used a recursive autoencoder to form abstract scene structures before generating concrete geometric layouts, while SceneHGN [140] employed hierarchical graph networks in which high-level graph structure and low-level geometric modules remained largely independent. Without effective cross-level feedback, such methods were often prone to semantic–geometric misalignment.

Subsequent works [95,96,102,142–148] introduced directional information flow across levels, partially alleviating this fragmentation. FSR-VLN [142] organizes RGB-D observations and pose information into scene graphs for vision-and-language navigation, allowing topological decisions to constrain local path selection, while Stairway to Success [102] builds topological models from floor awareness and obstacle mapping to guide low-level navigation. However, such interactions remain predominantly unidirectional.

More recent approaches [40,100,104,149–157] explore bidirectional coupling between metric and relational representations, allowing geometric perception and relational reasoning to jointly shape cognitive maps. For example, Map2Thought [149] couples continuous 3D metric maps with discrete topological reasoning, CLiViS [153] iteratively updates cognitive maps through the interaction of language-guided planning and visual perception, and GraphEQA [154] integrates 3D semantic scene graphs with multimodal memory and hierarchical planning. Collectively, these methods reflect a shift toward mutually informed, dynamically coordinated representations across abstraction levels, enhancing consistency and adaptability in complex spatial reasoning scenarios.

2.3.2. Feature Fusion-Based

Representations from the metric space and relational topology are first encoded into feature vectors and then fused in a shared feature space to produce a unified representation capable of supporting complex tasks. Depending on the fusion operators and interaction logic, existing methods mainly manifest three categories of strategies: direct algebraic fusion, attention-based dynamic fusion, and graph-operator-based structured fusion.

Many methods [39,97,98,158–160] employ direct algebraic operators, such as feature concatenation, pooling, or linear projection, to align relational features with metric features along spatial dimensions. For example, BSG [97] projects local BEV grid features into scene graph node features, preserving geometric constraints while maintaining topological connectivity. Similarly, SEK [158] directly concatenates relational features from an external knowledge base with geometric features extracted from sketches in the hidden layers of the decoder, encouraging the generated 3D geometry to satisfy physical proportional priors.

Other works [103,161,162] adopt attention-based fusion mechanisms. For instance, SUSA [103] uses hierarchical cross-attention to inject high-level topological relations into low-level visual metric features, improving environmental representation under limited perception. MAPInstructor [161]

 doi:10.20944/preprints202605.1782.v1

projects local observations into a 3D voxel space while introducing a global topological map as structured prompts. Within a shared attention space, metric tokens and map prompt tokens jointly participate in self-attention computations, enabling dynamic interaction between metric and relational modalities.

There are also methods [99,146,163] based on graph operators, where information is propagated through graph convolution or recursive graph updates. For example, MMGDreamer [146] uses graph convolutions on a mixed-modality graph to propagate object geometry under semantic constraints, while SG-Bot [99] iteratively refines object 6D poses through coarse-to-fine graph updates, enabling tighter coupling between metric detail and topological structure.

Summary. The perception stage in spatial intelligence centers on constructing cognitive maps from local observations. Metric representations support semantic attributes through geometrically grounded spatial structures, providing precise physical grounding and strong spatial consistency, while being less expressive for high-level structural relations. In contrast, relational representations organize semantic information through topological and entity-level dependencies, making them more suitable for abstract structural reasoning, although they typically offer weaker metric fidelity. Hybrid representations integrate these two organizational forms, linking geometric accuracy with relational abstraction to support more comprehensive spatial understanding. Overall, the development from single-form representations toward hybrid cognitive maps reflects a broader shift toward more unified internal representations for spatial intelligence.

3. Reasoning: Inference with Cognitive Maps

The value of cognitive maps lies in its role as a substrate for reasoning, through which spatial information is read, manipulated, and ultimately transformed into actionable decision-making signals. Spatial reasoning can be viewed as an inference process mediated by cognitive maps, where the essential factor is how the reasoning module accesses, interprets, and exploits the spatial information encoded within it. Accordingly, we focus on the reasoning aspect of spatial intelligence and systematically analyze the different ways in which cognitive maps are utilized during inference. Specifically, we categorize existing approaches according to the technical paradigms by which the reasoning module reads from and operates on cognitive maps, as illustrated in Fig.5: (i) *Map as Embedding*, where the cognitive map is treated as feature representation directly serves as part of the decision-making state; (ii) *Map as Prompt*, where the cognitive map is transformed into prompts in different modalities and injected as external conditioning into sequential reasoning models; and (iii) *Map as API*, where the cognitive map is abstracted into a set of callable operations that directly participate in spatial reasoning at runtime through queries, updates, or constraints. Through this categorization, this section aims to elucidate the design motivations underlying different reasoning paradigms and their implications for spatial reasoning capabilities.

 doi:10.20944/preprints202605.1782.v1

Table 2. Reasoning Paradigms for Inference with Cognitive Maps

Category	Method	Venue	Type	Representation	Backbone	Characteristic
Embedding	CMP [64]	CVPR'17	Propagation	Geometry	ResNet + Iteration Network	I G
	NRNS [141]	NeurIPS'21	Propagation	Graph + Geometry + Coordinate	GAT	I
	VGM [119]	ICCV'21	Matching	Graph	GCN + RNN	G
	CM2 [62]	CVPR'22	Matching	Geometry	Transformer	-
	WS-MGMap [63]	NeurIPS'22	Matching	Geometry	CNN	-
	TSGM [120]	CoRL'23	Matching	Graph	Transformer	-
	SGC [125]	ICCV'23	Matching	Graph	Transformer	G
	BEVBert [145]	ICCV'23	Matching	Geometry + Graph	Transformer	G
	BSG [97]	ICCV'23	Matching	Geometry + Graph + Coordinate	Transformer	-
	EGO2Map [164]	ICCV'23	Matching	Geometry	Transformer	G
	GridMM [58]	ICCV'23	Matching	Geometry	Transformer	-
	MemoNav [88]	CVPR'24	Matching	Graph	GAT + Policy Network	-
	ECL [162]	ACM MM'24	Propagation	Graph	CNN + Transformer	G
	Sg-CityU [39]	ACM MM'24	Matching	Graph + Geometry	VoteNet + GCN	G
	SG-Bot [99]	ICRA'24	Propagation	Graph	CNN + GCN + Generative	-
	BevNav [163]	KBS'25	Matching	Graph + Geometry	Transformer	G
	MapNav [165]	ACL'25	Matching	Geometry	Transformer	G
	OVL-Map [50]	RAL'25	Matching	Geometry	Transformer + LSTM	G
	ObjReact [144]	CoRL'25	Propagation	Graph + Geometry + Coordinate	Policy Network	G
	BrainyMP [123]	TITS'25	Propagation	Graph	GNN	-
	HSAN [38]	NeurIPS'25	Propagation	Graph	Transformer + Policy	I G
	HTSCN [38]	TNNLS'25	Propagation	Graph	GCN	-
	MTU3D [159]	ICCV'25	Matching	Geometry	Transformer	G
	VPN [106]	AAAI'26	Matching	Geometry	Transformer	-
	SeqWalker [137]	AAAI'26	Matching	Geometry	CLIP + Policy Network	-
	HETT [44]	AAAI'26	Matching	Geometry	Transformer	-
	SUSA [103]	AAAI'26	Matching	Geometry	Transformer	G

 doi:10.20944/preprints202605.1782.v1

Table 2. *Cont.*

Category	Method	Venue	Type	Representation	Backbone	Characteristic
Prompt	NLMap [61]	ICRA'23	Textual	Geometry	LLM	I T
	SayPlan [127]	CoRL'23	Textual	Serialized Graph	LLM	I T
	MapGPT [92]	ACL'24	Textual	Serialized Graph	LLM / VLM	I G T
	KARMA [128]	ICRA'25	Textual	Serialized Graph	LLM	I T
	TP-MDDN [54]	NeurIPS'25	Textual	Geometry	LLM	I G
	Struct2D [101]	NeurIPS'25	Multimodal	Geometry + Coordinate	LLM	I G
	See&Trek [166]	NeurIPS'25	Multimodal	Geometry	VLM	I G T
	SpatialMind [138]	NeurIPS'25	Multimodal	Geometry	VLM	I G
	3D-Mem [46]	CVPR'25	Visual	Geometry	VLM	I T
	APC [79]	ICCV'25	Multimodal	Coordinate	VLM	T
	CLiViS [153]	ArXiv'25	Textual	Graph + Coordinate	LLM + VLM	I G T
	CogNav [100]	ICCV'25	Textual	Graph + Geometry	LLM + VLM	I G T
	ReasonNav [52]	CoRL'25	Multimodal	Geometry	VLM	I G T
	SpaceR [51]	ArXiv'25	Textual	Distance	VLM	I G
	Ego3D-VLM [82]	ArXiv'25	Textual	Coordinate	VLM	I G
	GraphEQA [154]	CoRL'25	MultiModal	Graph + Geometry	VLM	I G T
	Dynam3D [77]	NeurIPS'25	Visual	Geometry	VLM	G
	FSR-VLN [142]	ArXiv'25	Multimodal	Graph + Geometry	VLM	I G T
	Video2Layout [86]	ArXiv'25	Textual	Geometry	VLM	I
	GeoNav [104]	PR'26	Multimodal	Graph + Geometry + Coordinate	LLM	I T
	Blueprints [85]	ArXiv'26	Textual	Geometry	VLM	I
	Log-Nav [167]	AAAI'26	Textual	Graph + Coordinate	LLM	I T
	ASCENT [102]	RAL'26	Multimodal	Geometry	LLM	I G T
	OmniNav [150]	ICLR'26	Multimodal	Geometry	VLM	G

 doi:10.20944/preprints202605.1782.v1

Table 2. *Cont.*

Category	Method	Venue	Type	Representation	Backbone	Characteristic
	CAPEAM [60]	ICCV'23	Memory	Geometry	LLM	-
	TopoNav [122]	IROS'24	Snapshot	Graph	DQN	I G T
	BeliefMapNav [69]	NeurIPS'25	snapshot	Graph	VLM+Transformer	I G T
	BSC-Nav [152]	ArXiv'25	Snapshot	Graph + Geometry + Coordinate	VLM	G T
	ReMEmbR [41]	ICRA'25	Memory	Coordinate	LLM	I G
	VideoAgent [84]	ICCV'25	Memory	Coordinate	LLM / VLM	I G
	MG-Nav [147]	ArXiv'25	Snapshot	Graph + Geometry	Diffusion Policy + A*	G
	GC-VLN [105]	CoRL'25	Snapshot	Graph	Constraint Solver	I G T
API	LagMemo [168]	ArXiv'25	Memory	Geometry	3DGS + VLM	G
	RoboMemory [134]	ArXiv'25	Memory	Graph	VLM	I G
	Meta-Memory [80]	ArXiv'25	Memory	Coordinate	LLM + VLM	I G
	RoboOS [40]	ArXiv'25	Memory	Serialized Graph	VLM	I G T
	MrSteve [83]	ICLR'25	Memory	Serialized Graph	LLM	-
	CausalNav [169]	RAL'26	Snapshot	Graph	LLM	I G
	SpNav [45]	AAAI'26	Snapshot	Geometry	VLM	G
	EPoG [126]	ICRA'26	Snapshot	Graph	LLM / VLM	G T

Abbreviations in the **Characteristic** column denote the fundamental properties of the reasoning modules: I (Interpretability) denotes the generation of explicit, human-readable intermediate representations for decision transparency; G (Generalizability) indicates zero-shot transferability to unseen environments or open-vocabulary support; T (Training-Free) refers to plug-and-play deployment without task-specific parameter updates.

 doi:10.20944/preprints202605.1782.v1

3.1. Map as Embedding

Map as Embedding refers to a class of approaches in which the cognitive map is encoded into a latent representation to provide spatial understanding. In this paradigm, the map is treated as an internal state space for inference, whose representation is continuously updated and consumed by the reasoning module. Depending on how the spatial structure of the map is preserved and exploited, this paradigm can be further divided into two representative forms.

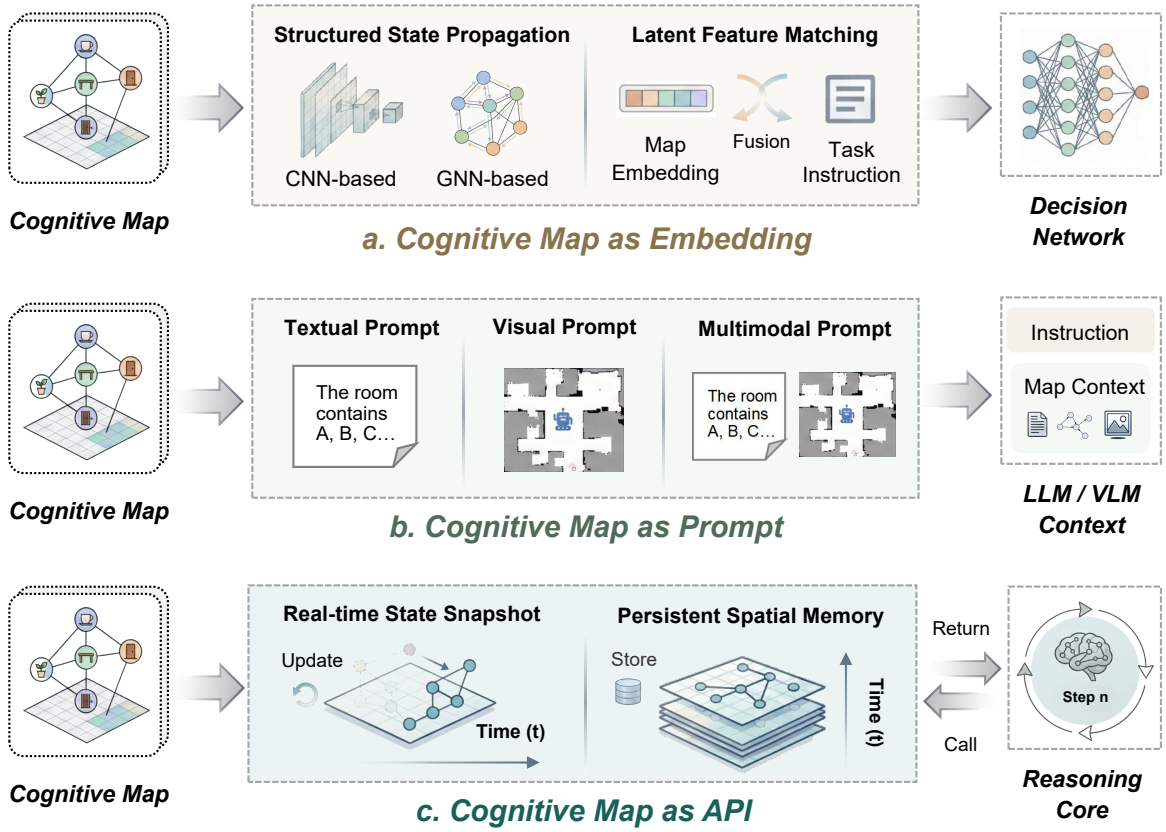

Figure 5. Illustration of three paradigm of inference with cognitive maps. (a) Map as embedding: reasoning operates directly over representation in structural state or latent state. (b) Map as prompt: maps are injected as textual, visual, or multimodal context to guide inference, (c) Map as API: maps function as external, callable memory modules supporting iterative query and state update during decision-making.

3.1.1. Structural State Propagation

A straightforward way to exploit cognitive maps is to represent space in a structured form encoded reachability, uncertainty, cost, or value. Early work[64] embeds the value iteration process into an ego-centric, multi-scale belief map, where local convolutional propagation is used to approximate Bellman updates and directly generate spatial action policies from propagated value functions. Subsequent studies [123,144], extend this idea to structural graph spaces, such as object-level or configuration-space graphs, where shortest path computation or cost propagation is performed over graph nodes, and the resulting propagation outcomes are explicitly transformed into executable cost fields or policy signals.

Another series of work exploits cognitive maps as structured relational memories that provide priors for reasoning and decision making. For instance, NRNS [141] propagates contextual information over an online topological graph to estimate target distance and select exploration subgoals, ECL [162] uses semantic maps as priors for representation learning and policy modules, and Sg-Bot [99] employs a structured world model to imagine target scenes, which are combined with geometric execution modules to produce action decisions.

 doi:10.20944/preprints202605.1782.v1

3.1.2. Latent Feature Matching

Compared with structural state propagation approaches, these methods treat the cognitive map as a latent memory. The map preserves only the regions, landmarks, objects, and contextual cues useful for downstream grounding and reasoning. Spatial reasoning therefore depends less on explicit state propagation, but rather on matching the current observation, goal, or instruction with the most relevant memory unit in the cognitive map [119,145,155]. From this view, latent features are primarily represented in two forms: sparse associative memory and dense fields.

The first route treats the cognitive map as a sparse associative memory. Early work such as VGM [119] shows that a topological graph of visually distinctive places can already support spatial understanding: the map need not reconstruct the full scene, but only store a set of retrievable landmarks and their adjacency relations. This establishes the basic formulation of latent feature matching, in which the current observation queries a structured memory for matching and aggregation. Later work extends this formulation primarily by changing the granularity of the memory unit. One direction enriches place-level memory with stronger semantics, such that retrieval is conditioned not only on visual similarity but also on task intent and object-level context [88,155,165]. Another replaces landmark nodes with richer structured entities, such as object graphs or scene graphs, so that matching operates over objects, relations, and subgraphs rather than isolated viewpoints [39,103,120,125,159,163].

The second route treats the map as a dense latent spatial field. Compared with sparse memory methods, these approaches retain stronger metric structure by organizing the environment into egocentric semantic maps, BEV layouts, or grid memories. Early semantic-map-based methods adopt top-down representations as an intermediate space in which spatial coordinates and semantic categories are jointly encoded, enabling instruction-conditioned reasoning through attention over map regions rather than only retrieval over graph nodes [44,50,62,106,137]. A transition in this route is to turn such maps into transferable spatial embeddings. BEVBert [145] is representative in this regard: it treats topological nodes and local BEV regions as spatial tokens and uses BERT-style pretraining to learn cross-view and cross-location consistency. Later work extends this direction by accumulating longer-horizon grid memories, reorganizing BEV regions with graph structure, and introducing auxiliary pretraining or self-supervised objectives to improve the stability of spatial matching across environments [58,63,97,164]. Cognitive maps have been converted from task-specific map to representation learning over space which locations, viewpoints, and semantics can remain consistently aligned.

3.2. Map as Prompt

The rapid progress of vision–language models (VLMs) has enabled strong semantic understanding and language-based reasoning over visual observations, making them a natural candidate for high-level spatial inference in spatial intelligence systems [170–179]. However, VLMs typically reason over local and view-dependent observations and lack explicit access to global geometry, long-term spatial memory, limits their effectiveness in complex spatial intelligence task. Map as Prompt was proposed to mitigate these limitations, translating structured cognitive map into interpretable conditioning signals that guide VLM reasoning by exposing global and persistent spatial context. Accordingly, existing approaches can be categorized into three types based on how cognitive maps are injected into the reasoning process.

3.2.1. Textual Prompt

Textual prompting treats the cognitive map as a world state serialized into language. A common design is to encode entities including spatial attributes (e.g., coordinates and relative relations) and semantic attributes (e.g., object types and affordances). Early attempts often follow a retrieval-augmented generation paradigm, in which only the objects or regions most relevant to the current task are retrieved from the map and verbalized as reasoning context [61,157]. Subsequent work moves from retrieved snippets to explicit structured prompting over scene graphs, topological graphs, landmarks,

 doi:10.20944/preprints202605.1782.v1

adjacency relations, and routes. SayPlan [127] exemplifies this transition by serializing a 3D scene graph into textual JSON for high-level planning, while later map-guided prompting frameworks encode topological nodes, landmark relations, and route structure as textual prompts, enabling the model to reason over a global world model rather than local observations alone [92,100,180]. More recent systems further introduce hierarchical memory and state decomposition, separating global from local, long-term from short-term, or static from dynamic information so that prompting can support both long-horizon planning and situated action selection [93,128,167]. Across this route, the central trend is that textual maps evolve from retrieved evidence into explicit planning interfaces, and eventually into hierarchical world models that mediate both deliberation and execution.

Another approach, on the contrary, treats the cognitive map as a textual anchor for explicit spatial reasoning. Here the goal is not only to provide planning context, but also to expose an intermediate representation on which reasoning can operate directly. One line of work externalizes object-centric relations and scene structure as symbolic text, allowing models to reason over entities and their interactions rather than raw observations [85,153]. Another line constructs egocentric or coordinate-aligned textual maps from multi-view observations, projecting objects from different views into a unified frame and thereby converting distributed multi-view reasoning into inference within a single coordinate system [82,86]. A further extension internalizes this idea: SpaceR [51] no longer injects a fully explicit map as prompt, but encourages the model to form an implicit cognitive map in a unified coordinate frame by using localization outputs as supervision for reasoning. In this route, the technical progression is from explicit symbolic map injection to increasingly abstract map-based reasoning scaffolds, while the cognitive map remains the organizing structure that stabilizes spatial reasoning.

3.2.2. Visual Prompt

Visual prompting methods externalize cognitive maps as model-readable visual contexts, allowing foundation models to reason over retrieved exemplars, structured visual tokens. 3D-Mem [46] adopts a retrieval-based strategy, clustering similar images and filtering the most relevant ones at inference time to provide spatial relationship. [77] introduces a pretrained zone Encoder to encode cognitive maps by aggregating multiple semantic entities within a region, forming zone tokens that are both instance-level and layout-aware. Representing cognitive maps through intermediate visual states provides an even more direct mechanism for spatial reasoning: VoT [56] outputs intermediate images annotated with textual markers to support visual logical reasoning, while GPT4Scene[47] incorporates a rendering pipeline that generates unified BEV images, constructing cognitive maps that explicitly capture global spatial information.

3.2.3. Multimodal Prompt

Multimodal prompting arises because no single modality suffices for spatial reasoning in some case. The primary form is combining heterogeneous structural cues into a unified prompt for reasoning, as explored in works [104,142,150,154]. These approaches encode heterogeneous information within scene graphs, including visual features, geometric attributes, geographic cues, and BEV representations. OmniNav [150], in particular, constructs reasoning context by combining BEV encodings with image-based memory representations, and further introduces a fast–slow thinking mechanism to balance global planning and local decision-making. Multimodal prompt can also enhance spatial reasoning by exposing actionable cues. Methods like ResonNav [52] and ASCEND [102] leverage additional scene instruction to expand the effective reasoning space.

Recent work pushes multimodal prompting toward fine-grained spatial understanding. Instead of only fusing scene-level graph and map cues, these methods construct intermediate multimodal states that directly expose relative geometry, viewpoint transformations, or motion trajectories to the model. One direction renders intermediate views or layouts conditioned on object-centered language prompts [79]. Struct2D [101] enhances reasoning by integrating a BEV rendering pipeline. See&Trek [166] focuses on long-video understanding by modeling camera trajectories and selecting raw RGB frames to construct cognitive maps that explicitly encode spatiotemporal motion cues. SpatialMind

 doi:10.20944/preprints202605.1782.v1

[138] further decomposing scenes and mapping salient objects to two-dimensional coordinates to represent spatial layouts, which are jointly reasoned over with video inputs.

3.3. Map as API

As spatial intelligence tasks grow in complexity, an agent must frequently query spatial information and actively manipulate its cognitive map during execution. Cognitive map as application programming interfaces (APIs) arise to support multi-round reasoning and action. The map is exposed as a set of callable operations (e.g., `query()`, `update()`) for structured interaction. Under this paradigm, the map is a stateful interface through which reasoning modules can inspect, modify, and propagate changes to the underlying spatial state during inference. We categorize these approaches according to whether the map is designed to represent only the current state of the world or to accumulate historical spatial knowledge across interactions.

3.3.1. Real-time State Snapshot

In spatial intelligence, agents are often required to operate under strong partial observability, continuously changing environments, and dynamically evolving goals, which necessitates the real-time maintenance of cognitive representations tightly coupled with the current execution state. In such settings, real-time snapshots of the environment provide a natural interface for grounding reasoning and decision-making. VLMaps [59] constructs a spatially aligned semantic feature field by jointly performing semantic segmentation and language–space alignment within a scene, enabling agents to localize open-vocabulary concepts in the environment. During planning, the agent can flexibly index object locations through this structured representation. Some works [45,147], instead, maintain snapshots of the world by constructing geometric buffers that preserve hybrid semantic–geometric representations.

A complementary approach conceptualizes spatial reasoning as belief modeling, in which spatial belief states are explicitly maintained and continuously updated through interaction with real-time observations and decision-making processes [57,69]. Within this perspective, cognitive representations take the form of probabilistic spatial belief maps that jointly account for expected observation gain, traversal cost, and path length to minimize anticipated search distance [69]. Extensions of this framework introduce dual-state belief representations that decouple spatial search dynamics from semantic task uncertainty, enabling coordinated reasoning across perceptual and task abstractions [57]. To improve scalability, belief structures are further sparsified into topological graphs, where instruction-relevant subgraphs are retrieved through retrieval-augmented mechanisms and leveraged for structured path planning [122,169]. More recent formulations incorporate landmark-based memory together with regional surprise signals as implicit belief estimators, yielding multi-level state representations that support robust spatial reasoning and decision-making in large-scale environments [152].

3.3.2. Persistent Spatial Memory

Persistent spatial memory extends reasoning beyond the current state by enabling agents to accumulate and reuse spatial experience over time. Early work designs memory as an independent module that preserves instance-level memories [60], aiming to improve efficiency of spatial planning by avoiding unnecessary exploration and aligning action planning with the current environmental state. In particular, the memory module compensates perception using historical masks when the risk of interaction failure arises. Subsequent works [84] extend by maintaining richer attributes such as 3D bounding boxes, object states, relations, and temporal indices. A series of later studies[41,80,168**?**] construct memory through accumulated historical observations, where these observations point to navigable regions. In this view, the cognitive map functions as a repeatedly invoked long-term memory resource that supplements reasoning systems with temporal context, realized through mechanisms such as natural-language-based retrieval [168], attention-based implicit reading [71], or explicit function calling [41,80].

As research progress, cognitive maps are extended into integrated memory systems that jointly encode spatial and temporal information. These systems not only encode where and what, but

 doi:10.20944/preprints202605.1782.v1

also support reasoning about past events and their relevance to current decisions. RoboMemory [134] and Mr.Steve[83] organizes historical observations, interaction outcomes, and spatial relations into multi-level memory structures maintained through symbolic or semi-symbolic representations, enabling repeated querying to support long-horizon task planning and experience transfer. RoboOS series [40,156] elevate cognitive maps to a system-level shared state representations, defining Spatial-Temporal–Embodiment memory to support alignment and collaboration across multiple agents with respect to spatial states.

Summary. Cognitive maps occupy a central role in spatial reasoning. As embeddings, they support efficient retrieval, alignment, and grounding across perception, language, and action, but their reasoning process is often implicit and less interpretable. As prompts, they offers greater flexibility and compatibility with foundation models, since the map can be externalized as textual, visual, or multimodal context that directly conditions downstream reasoning and planning, though this comes with information compression bottlenecks. By contrast, Map as API provides the strongest degree of controllability and closed-loop interaction, allowing agents to query, update, and constrain spatial knowledge at runtime, which is especially valuable in dynamic and long-horizon tasks. Its cost is higher system complexity, such as state management and tool use. Taken together, these paradigms represent different trade-offs among compactness, expressiveness, and operational control in spatial reasoning based on cognitive maps.

4. Generation: Realization of Cognitive Maps

Generation represents the inverse of perception, characterizing a fundamental capability of spatial intelligence where information flows from internal representations to external or simulated environments. It leverages the abstract, globally unified, and persistently maintained cognitive map as a structural blueprint and a conditional prior for the instantiation of concrete, visible, or interactive spatial manifestations. Specifically, the system projects its internalized spatial knowledge, encompassing topological layouts, geometric relationships, and semantic attributes, inversely back onto observable data distributions. As illustrated in Fig. 6, we categorize existing methodologies into two paradigms based on their spatiotemporal characteristics: (i) *Static Scene Synthesis*, which focuses on the restoration of spatial layouts and asset instantiation, and (ii) *Dynamic World Simulation*, which addresses spatiotemporal evolution and consistency maintenance.

4.1. Static Scene Synthesis

Static scene synthesis aims to leverage the structured priors provided by cognitive maps including spatial layouts object semantics and topological relationships to transform abstract internal representations into concrete three-dimensional environments. The central challenge of this process involves addressing the ill-posed mapping problem inherent in translating sparse symbols into dense geometry. Depending on the specific function of the cognitive map within the generation pipeline whether it operates as a blueprint for retrieval planning or serves as a condition for generative models we categorize existing methodologies into two primary paradigms namely *map-based retrieval* and *map-to-scene generation*.

4.1.1. Map-Based Retrieval

In this paradigm the cognitive map primarily functions as a spatial planner. Systems prioritize parsing the semantic and topological constraints embedded within the map to instantiate scenes by retrieving existing high-quality three-dimensional assets. The research focus lies in transforming complex cognitive maps into executable retrieval instructions and layout parameters.

Recognizing that cognitive maps are not flat lists of objects but imply complex hierarchical and grammatical structures, early works endeavored to explicitize these structures to guide scene synthesis [130,131,139]. GRAINS [139] pioneered the modeling of indoor scenes as recursive tree structures where rooms contain functional areas that further contain furniture employing recursive autoencoders to unroll cognitive maps into concrete layout hierarchies. Meta-Sim [130] and Meta-

 doi:10.20944/preprints202605.1782.v1

Sim2 [131] instead adopt probabilistic scene grammars to capture scene rules from visual data under unsupervised or weakly supervised settings, highlighting the role of cognitive maps as compositional priors in scene generation.

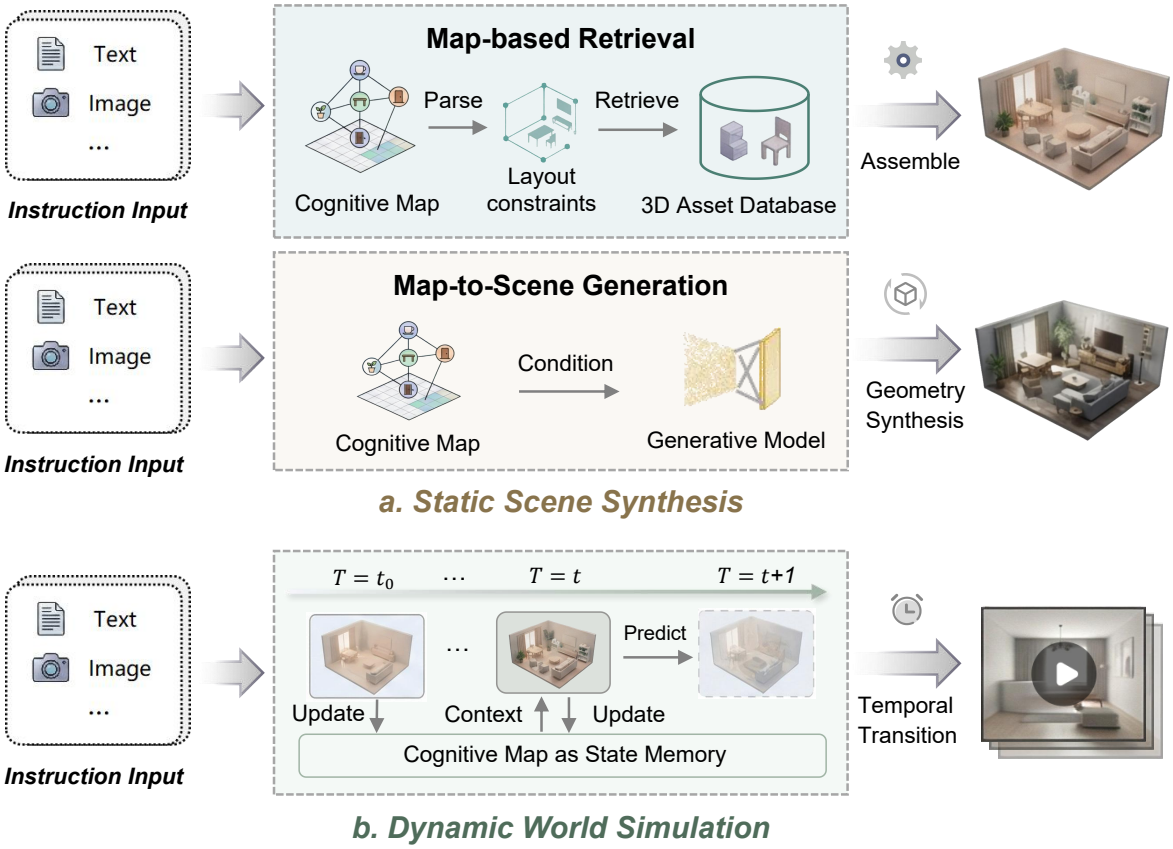

Figure 6. Two spatial generation paradigms conceptualized as the realization of cognitive maps. (a) Static scene synthesis emphasizing structural priors via retrieval or end-to-end geometry generation; and (b) Dynamic world simulation focusing on continuous state evolution for spatiotemporal consistency.

To ensure the physical plausibility and functional consistency of synthesized scenes, many research focuses on relation-oriented layout planning, utilizing explicit relation graphs to constrain object retrieval and placement [117,118,181]. PlanIT [117] proposed a planning-instantiation framework which first generates a relation graph containing constraints such as support and alignment and subsequently utilizes this graph to guide specific model retrieval and pose optimization. Adaptive Synthesis [118] further introduced priors of activity association encoding human activities into the cognitive map to deduce topological relationships between objects based on functional requirements.

With the advancement of multimodal technologies, constructing or manipulating cognitive maps via natural language has emerged as a prominent research direction, facilitating language-driven instruction mapping and agent reasoning. LLMs and VLMs have been introduced as reasoning cores to transform ambiguous linguistic instructions into precise spatial layouts [111,132,182–186]. Early methods [182] parsed natural language into scene graphs for retrieval. HOLODECK [132] further uses LLMs to generate spatial constraint code and invokes downstream solvers to determine object placements in simulators. INSTRUCTSCENE [111] and LayoutVLM [183] combine semantic map priors with multimodal reasoning, using diffusion-based completion or differentiable optimization to align layouts with textual descriptions. FOREST2SEQ [184] further introduced serialization priors enhancing the logic and coherence of layout generation by predicting the sequence of object placement.

doi:10.20944/preprints202605.1782.v1

Table 3. Generation: Realization of Cognitive Maps

Category	Method	Venue	Map Representation	Technique	Scene Type	Granularity
Synthesis	Fu et al.[118]	TOG'17	Graph + Geometry	Retrieval	Indoor	Layout-level
	Ma et al.[182]	TOG'18	Graph	Retrieval	Indoor	Layout-level
	GRAINS[139]	TOG'19	Graph	VAE + Retrieval	Indoor	Layout-level
	PlanIT[117]	TOG'19	Graph	GCN + Retrieval	Indoor	Layout-level
	3D-SLN[114]	CVPR'20	Graph	VAE + Retrieval	Indoor	Layout-level
	SCENEHGN[140]	TPAMI'21	Graph	VAE	Indoor	Geometry-level
	Graph-to-3D[115]	ICCV'21	Graph	VAE	Indoor	Geometry-level
	CommonScenes[187]	NeurIPS'23	Graph	VAE + Diffusion	Indoor	Geometry-level
	SceneDreamer[188]	TPAMI'23	Geometry	GAN	Outdoor	Observation-level
	SEK[158]	ECCV'24	Geometry + Graph	Diffusion	Indoor	Geometry-level
	InstructScene[111]	ICLR'24	Graph	Diffusion	Indoor	Layout-level
	GraphDreamer[136]	CVPR'24	Serialized Graph	Diffusion + SDS	Indoor	Geometry-level
	CityDreamer[189]	CVPR'24	Geometry	GAN	Outdoor	Observation-level
	MagicDrive[190]	ICLR'24	Graph + Geometry	Diffusion	Outdoor	Observation-level
	HOLODECK[132]	CVPR'24	Serialized Graph	LLM + Retrieval	Indoor	Observation-level
	GaussianCity[191]	CVPR'25	Geometry	3D Gaussian Splatting	Outdoor	Geometry-level
	Planner3D[113]	TPAMI'25	Graph	VAE + Diffusion	Indoor	Geometry-level
	MMGDreamer[146]	AAAI'25	Graph + Geometry	Diffusion	Indoor	Geometry-level
	Liu et al.[192]	ICCV'25	Graph + Geometry	Diffusion	Outdoor	Geometry-level
	SpatialGen[193]	3DV'26	Geometry	VAE + Diffusion	Indoor	Observation-level
Simulation	SSGVS[91]	CVPR'23	Serialized Graph	VQ-VAE	Cross	Observation-level
	OccWorld[66]	ECCV'24	Geometry	VQ-VAE	Outdoor	Geometry-level
	DOME[68]	ArXiv'24	Geometry	Diffusion	Cross	Layout-level
	InfiniCube[43]	ICCV'25	Geometry	Diffusion	Outdoor	Observation-level
	VerseCrafter[73]	ArXiv'25	Geometry	Diffusion	Outdoor	Observation-level
	Wu et al.[42]	NeurIPS'25	Geometry	Diffusion	Outdoor	Observation-level
	Spatia[75]	ArXiv'25	Geometry	Diffusion	Cross	Observation-level
	Zhou et al.[72]	NeurIPS'25	Geometry	VAE + Diffusion	Indoor	Observation-level
	Memory Forcing[76]	ArXiv'25	Geometry	Diffusion	Outdoor	Observation-level
	NeoVerse[74]	ArXiv'26	Geometry	Diffusion	Outdoor	Observation-level

Granularity denotes the output level of realization. **Layout-level**: Object-level metric placement. **Geometry-level**: Dense geometry or surfaces. **Observation-level**: Pixel-level observation output.

 doi:10.20944/preprints202605.1782.v1

4.1.2. Map-to-Scene Generation

Diverging from the retrieval paradigm Map-to-Scene Generation leverages generative models including Diffusion Models GANs and Neural Fields to directly synthesize three-dimensional geometry and appearance from latent space. Here the cognitive map functions as a potent conditional signal guiding the model to learn and sample distributions that conform to specific semantic structures.

To transcend the limitations of asset libraries, certain studies explore graph-driven end-to-end geometry synthesis methods for generating 3D geometric meshes directly from abstract graph structures [115,140]. Graph-to-3D [115] utilizes Graph Convolutional Networks (GCNs) to extract scene graph features while simultaneously predicting object layout bounding boxes and shape encodings thereby achieving context-aware generation of geometric shapes. Similarly, SCENEHGN [140] introduces hierarchical graph networks to manage generation from the room level down to fine-grained furniture components, ensuring multi-scale geometric consistency.

While the introduction of diffusion models has enhanced generative realism, it has simultaneously introduced challenges regarding multi-object control, prompting researchers to leverage cognitive maps as a disentanglement mechanism for diffusion-based compositional generation [112,136,146,193–195]. CommonScenes [112] and GraphDreamer [136] inject scene graphs into diffusion models as structured guidance, with the latter decomposing prompts into node-level and edge-level components to reduce attribute confusion. MMGDreamer [146] proposed mixed-modal graphs allowing nodes to contain textual or image modalities simultaneously. The layout and appearance decoupling strategy proposed in Ctrl-Room [194] generates layout codes prior to panoramic images, enabling more independent control over scene structure and visual detail. CC3D [195] and SPATIALGEN [193] utilize 2D layouts as conditions to drive 3D GANs or multi-view diffusion models achieving 3D consistent generation under single-image training settings.

To address complex tasks, the generation process integrates agent-assisted and iterative optimization strategies [113,116,135,151,158]. Dynamic cognitive maps composed of scene profiles and semantic point clouds are used to support iterative asset generation and layout refinement by VLM agents [151]. LayoutAgent [135] adopts a planning-guidance mode where a VLM plans the layout to guide diffusion models. HiGS [116] mimics human design thinking by progressive hierarchical spatial semantic graphs employing a coarse-to-fine strategy to recursively refine local regions. Planner3D [113] utilizes LLMs to enhance scene graph nodes and introduces explicit physical regularization losses, reinforcing the physical plausibility of the generated results.

Synthesizing unbounded environments imposes stringent demands on memory efficiency and topological continuity, necessitating a shift where cognitive maps function as compact BEV scaffolds to anchor generative models for continuous rendering [188–192]. SceneDreamer [188] and CityDreamer [189] utilize BEV representations containing semantic and height information to query generative neural hash grids realizing the generation of unbounded natural landscapes and urban architectures. GaussianCity [191] leverages the efficiency of 3D Gaussian Splatting to propose a BEV-Point representation achieving low-memory infinite generation of city-level scenes. Other works focus on autonomous driving scenarios by encoding lane graphs and 3D bounding boxes to precisely control street view generation directly serving perception and simulation tasks [190,192].

4.2. Dynamic World Simulation

While static scene synthesis represents the spatial unfolding of cognitive maps, dynamic world simulation constitutes their temporal evolution. Under this paradigm, the generative system transcends a mere single-frame renderer to become a world model capable of predicting state transitions. Here the cognitive map assumes the critical role of persistent state memory for maintaining spatiotemporal consistency throughout long-term interactions and evolutions thereby addressing the issues of catastrophic forgetting and non-Euclidean spatial hallucinations commonly encountered in video generation.

 doi:10.20944/preprints202605.1782.v1

A central challenge in dynamic generation is that limited context windows often prevent models from preserving stable world structure across long sequences. As a result, previously observed regions may become inconsistent when revisited. To address this issue, recent work introduces persistent memory mechanisms that resemble cognitive maps, enabling the model to retain and retrieve global spatial states throughout generation. One line of research relies on explicit geometric memory. WorldMem [76] maintains an incremental 3D point cloud together with historical frames, and retrieves relevant observations through geometric indexing to enforce consistency with prior world states. Similarly, Spatia [75] incorporates a Visual SLAM module to register generated frames into a global 3D point cloud online, so that dynamic content remains grounded in a stable static map. Beyond explicit point-cloud-based memory, persistent world states can also be maintained through structured latent representations. In one representative approach, video features are aggregated into a voxelized 3D feature grid, allowing unobserved regions to remain consistent with the global scene representation [72]. Another method further refines memory structure by distinguishing between spatial memory consisting of persistent static point clouds and episodic memory consisting of sparse collections of keyframes [42]. This dual design simulates the distinct encoding methods of biological cognitive systems for spatial structures and episodic details preserving both rigid structure and rich visual texture.

Beyond consistency, the central challenge of dynamic simulation lies in modeling the evolution of the physical world and enabling precise control over it. In this setting, the cognitive map becomes a dynamic container that supports 4D world operations. A first requirement is to maintain stable spatiotemporal anchoring in large and continuously evolving environments. To this end, InfiniCube [43] builds a sparse voxel world as the underlying cognitive map, first rendering static sparse voxels into semantic and depth buffers to guide video generation, and then lifting the generated video into a 4D Gaussian field to preserve geometric stability over long horizons. NeoVerse [74] further strengthens 4D world modeling by leveraging in-the-wild monocular videos to improve spatiotemporal construction in complex real-world scenes.

To achieve fine-grained control over generated content, researchers leverage occupancy grids and scene graphs to precisely manipulate semantic and physical states [66,68,73,196]. DOME [68] and OccWorld [66] transform generated objects from RGB pixels into 3D occupancy grids. DOME utilizes diffusion models to predict the future evolution of occupancy grids directly simulating geometric changes in the physical world. This metric-level simulation is crucial for planning and obstacle avoidance in autonomous driving. VerseCrafter [73] introduces the concept of 4D geometric control utilizing static background point clouds in conjunction with object-level 3D Gaussian trajectories to represent world states allowing for precise control over the motion trajectories of specific objects. Semantically, SSGVS [196] employs dynamic scene graph sequences as conditional inputs enabling the system to simulate complex semantic changes such as a person standing up from a chair thereby validating the temporal expressiveness of relational cognitive maps.

Summary. From the perspective of cognitive maps, generation can be understood as the realization of internal spatial representations into external scenes or simulated worlds. Static scene synthesis and dynamic world simulation share the same representational foundation, but differ in their primary objectives. The former focuses on transforming structured spatial priors into concrete layouts, geometry, or observations, whereas the latter further endows these representations with persistence across time, so that the cognitive map functions not only as a spatial blueprint but also as a state memory for long-horizon evolution and controllable state transition. Overall, the literature shows a clear progression from static realization to dynamic simulation, and from scene construction to world modeling. This trend suggests that cognitive maps are evolving from spatial organization priors into persistent generative substrates that support not only the realization of structured scenes, but also the simulation of temporally coherent and interactable worlds.

 doi:10.20944/preprints202605.1782.v1

5. Application

To systematically review the applications of cognitive maps in spatial intelligence, this section categorizes existing methodologies based on the interaction paradigms between the agent and the system. As illustrated in Fig. 7, the subsequent discourse unfolds along two primary trajectories: *Open-Loop Spatial Cognition* and *Closed-Loop Spatial Interaction*.

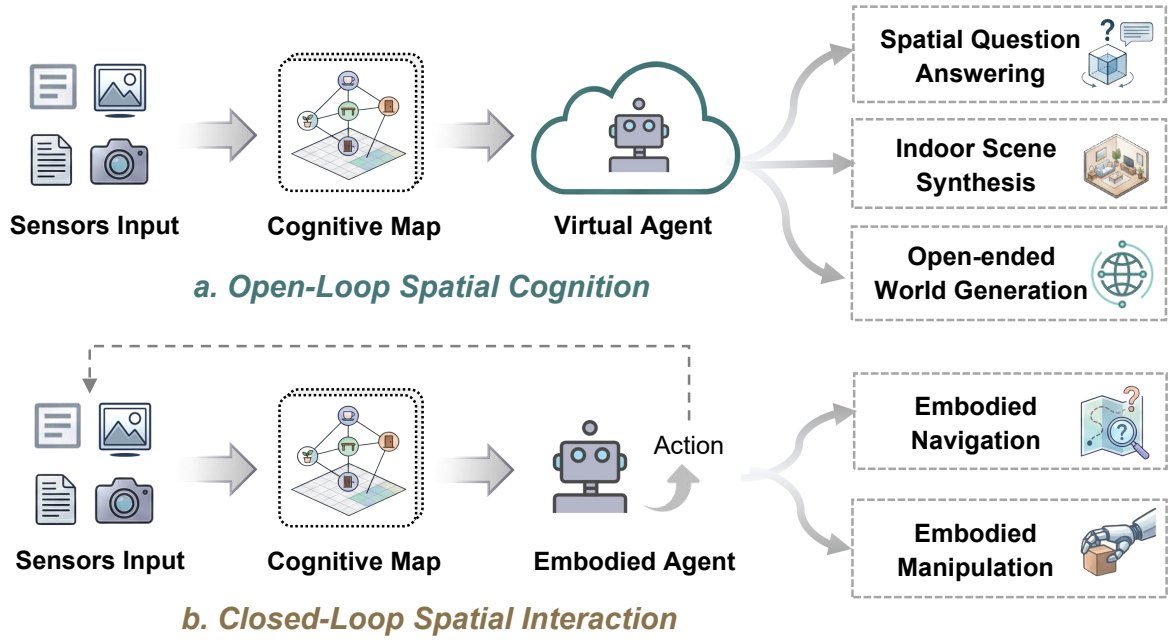

Figure 7. Application paradigms of spatial intelligence based on agent-system interaction: (a) Open-Loop Spatial Cognition: The agent engages in passive observation and offline processing without real-time environmental alteration. (b) Closed-Loop Spatial Interaction: The agent maintains active participation driven by a continuous perception-action loop to dynamically update environmental states.

5.1. Open-Loop Spatial Cognition

Open-loop spatial cognition maps information streams to 3D representations without requiring active physical intervention. It emphasizes deep scene comprehension and generation. Under such circumstances, the cognitive map serves as a critical bridge linking user intent with 3D geometry. It underpins the hierarchical progression of the agent's capabilities, advancing from the logical parsing of existing environments such as spatial question answering, to indoor scene synthesis, and ultimately extending to open-ended world generation.

5.1.1. Spatial Question Answering

This task advances visual perception to complex spatial reasoning. In open-loop paradigms, cognitive maps bridge linguistic queries and observed environments by aggregating discrete visual inputs into globally consistent spatiotemporal representations, resolving information fragmentation and viewpoint limitations.

A first line of work constructs allocentric map-like structures from partial observations, for example by deriving BEV layouts or object-centered spatial maps from point clouds and video streams [47,86,101]. Closely related approaches focus on reconstructing trajectories and scene layouts over time through keyframe-based spatial accumulation combined with VLM reasoning [138,149,166]; SpaceR further improves layout recovery from sparse video cues through reinforcement learning [197]. Another line addresses viewpoint variation by mapping observations from one or multiple egocentric views into a unified global 3D coordinate frame, thereby enabling viewpoint-invariant reasoning [79,82]. As temporal and spatial horizons expand, cognitive maps also become memory structures. CLiViS organizes VLM reasoning through navigation and relational graphs [153], while Sg-CityU extends this idea to urban-scale environments with long-range topological scene graphs [39]. Beyond

 doi:10.20944/preprints202605.1782.v1

representation and memory, cognitive maps can further act as explicit reasoning substrates, which use structured spatial constraints to support active Chain-of-Thought (CoT) reasoning rather than passive visual parsing [85,198].

5.1.2. Indoor Scene Synthesis

While spatial question answering focuses on semantic parsing of existing environments, indoor scene synthesis maps abstract user intentions into geometrically consistent and rationally structured 3D spaces, where the cognitive map serves as a prior structural scaffold to constrain generation. It establishes geometric order and encodes semantic logic alongside user intent, ensuring precise translation from sparse instructions to dense scenes.

For this task, cognitive maps support generation at three complementary levels. First, they provide a structural hierarchy that organizes rooms, objects, and functional areas into coherent layouts, enabling the synthesis of geometrically consistent indoor spaces from sparse user intentions [116,139,140,184]. Second, they serve as relational constraints that encode object co-occurrence, functional dependencies, and activity-oriented semantics, helping generation remain aligned with plausible indoor usage patterns, as exemplified by CommonScenes [187], GraphDreamer [136], and MMGDreamer [146]. Third, cognitive maps can be represented as structured parameters or intermediate scene specifications that directly support end-to-end generation from instructions, allowing systems to translate user intent into executable layouts or scene descriptions with stronger controllability and adaptability [132,135, 183,194]. Taken together, in indoor scene synthesis, cognitive maps function not merely as auxiliary representations but as the organizational substrate that links user intent, spatial structure, and scene realization.

5.1.3. Open-ended World Generation

Extending to unbounded open-ended world generation, cognitive maps maintain logical coherence and physical realism during infinite spatiotemporal expansion by providing rigid constraints against geometric distortions in long-sequence generation.

For real-world simulation, MagicDrive [190] and CityDreamer [189] resolve chaotic perspectives by employing BEV representations to calibrate road topologies. OccWorld [66] and DOME [68] deduce traffic evolution through voxel flow. To overcome large-scale rendering bottlenecks, GaussianCity [191] introduces 3D Gaussian Splatting technology, whereas InfiniCube [43] achieves the coherent generation of infinitely long driving scenes by combining sparse voxels with high-definition maps.

Conversely, for virtual world construction, Memory Forcing [76] dynamically maintains a global 3D memory bank, to align viewpoints and states upon revisiting locations. Building upon this, Agentic 3D [151] utilizes a scene hypergraph to endow the map with profound semantic attributes. Simultaneously, SceneDreamer [188] employs BEV-based height and semantic fields to generate continuously undulating infinite natural landscapes. In general video synthesis, VerseCrafter [73] and NeoVerse [74] combine static background point clouds with dynamic object Gaussians to enable precise control over camera trajectories and multi-object behaviors. To mitigate deformations in extended videos, Spatia [75] and Video World Models [42] utilize dynamically updated global point clouds as explicit 3D caches. Additionally, SSGVS [91] incorporates temporal scene graph constraints, ensuring causal coherence in object interactions over time.

5.2. Closed-Loop Spatial Interaction

Distinct from open-loop observation, closed-loop spatial interaction requires agents to execute long-horizon decisions via continuous perception-action loops in noisy dynamic environments. Here cognitive maps bridge the gap between abstract planning and concrete control. They sustain long-term memory and encode physical affordances to facilitate embodied navigation and manipulation.

 doi:10.20944/preprints202605.1782.v1

5.2.1. Embodied Navigation

As the fundamental prerequisite for agent-environment interaction, embodied navigation encompasses mobility, survival, and exploration. This domain primarily comprises sub-tasks including zero-shot object navigation, vision-language navigation, and long-horizon reasoning and question answering.

In Zero-Shot Object Navigation, explicit cognitive maps balance long-term memory of explored regions with efficient inference of unexplored areas. Many approaches construct explicit topological or scene graphs to provide landmark indexing and generate intuitive obstacle avoidance strategies [88,119,120,122,141,144]. To navigate more complex scenarios, hierarchical cognitive maps decouple abstract topological planning from low-level control [143,147,152], while other methods optimize map representational density and computational efficiency through 3D Gaussian Splatting, multi-task weighting mechanisms, or KV caching [54,81,168]. For exploring unknown targets, maps function as probabilistic belief maps [64,69] or dynamic memory queues [94] to quantify exploration value. In semantic environments, LLMs or VLMs leverage these cognitive maps to deduce target locations using common sense and symbolic logic [52,100,102,199,200]. Conversely, some frameworks utilize explicit maps exclusively during training for geometric supervision to enforce implicit cognitive mapping for map-free spatial perception [125,162,164].

In Vision-Language Navigation (VLN), cognitive maps align linguistic instructions with continuous visual observations. Early research projected localized visual observations into BEV or grid features to build a globally unified representation for implicit cross-modal fusion within the feature space [44,58,62,63]. To enhance spatiotemporal continuity, dynamically updated global maps serve as explicit memory [48,50,77,155], while others achieve cross-perspective alignment through visual prompting [106]. For long-horizon planning, structured cognitive maps act as spatial databases queryable via language [59]. Methodologies construct hierarchical multimodal scene graphs or dynamic spatial grids to parse instructions into graph constraints, utilizing LLMs to orchestrate low-level control [38,45,49,104,105,124,137]. Furthermore, encoding map data into linguistic prompts equips LLMs with context-awareness and dual-process reflection to conduct temporal long-horizon planning [60,92,142,161,180].

5.2.2. Embodied Manipulation

While navigation addresses the problem of how an agent arrives at destinations, embodied manipulation focuses on interaction mechanics where cognitive maps integrate multi-scale semantics such as object affordances and micro-level states into macroscopic spatial perception for high-precision physical interaction.

For object-level manipulation, BrainyMP [123] instantiates cognitive maps as random geometric graphs, introducing brain-inspired mechanisms to optimize low-level motion planning, guaranteeing the kinematic feasibility of the manipulation. SG-Bot [99] utilizes the imagination mechanism of scene graphs to deduce object dependencies, achieving rearrangement that conforms to semantic logic. Furthermore, ActiveVLA [67] employs active viewpoint strategies to dynamically complete local occlusion-free maps, overcoming visual blind spots during fine-grained manipulation.

As tasks extend to the floor level, SayPlan [127], NLMap [61] and Scene-MMKG [110] abstract environments into natural language-queryable maps and multimodal knowledge graphs, serving as global semantic indices to resolve cross-room anchoring and commonsense reasoning bottlenecks. Meanwhile, KARMA [128] and Hi-Dyna Graph [93] effectively tackle environmental non-stationarity hierarchical dynamic scene maps that decouple global static layouts from local dynamic evolution. MOMAGraph [201] constructs a state-aware unified scene graph, successfully bridging the representational gap between macroscopic navigation and part-level microscopic manipulation. Confronting partially observable and complex dynamic environments, EPoG [126] further integrates dynamic scene graphs with LLM reasoning, thereby realizing the synergistic integrated planning of spatial exploration and long-horizon manipulation. Ultimately, directed towards cross-embodiment collaboration,

 doi:10.20944/preprints202605.1782.v1

frameworks such as RoboOS series [40,156] evolve the cognitive map into a shared spatiotemporal and embodied memory, supporting the unified scheduling and synergy of multi-robot systems.

6. Future Directions

Despite rapid advancements, achieving human-level spatial intelligence remains challenges. Examining current research through the analytical lens of the cognitive map provides unique insights into the bottlenecks of existing systems. Guided by the core properties and mechanism of the cognitive map, this section exposes five essential limitations of current approaches: semantic abstraction, scalable globality, lifelong persistency, generative simulators and perception-action gap.

Deep Semantic Abstraction. Current semantic abstraction in cognitive maps is largely confined to shallow object categorization and basic spatial adjacency, overlooking rich attributes like material, state, implicit functionalities, and causal interaction mechanisms. Future research is expected to deepen these representations. Specifically, map nodes are anticipated to evolve from simple labels into attribute-aware representations encoding object identity, physical properties, and affordances. Simultaneously, edges would transition from descriptive spatial links to explanatory causal mechanisms incorporating physical constraints and interaction logic. Transforming cognitive maps into multidimensional causal knowledge bases would empower them to support complex reasoning and task planning beyond mere localization.

Scalable Globality. Although spatially enhanced VLMs have improved local reasoning about relative positions and geometric relations, they still lack an intrinsic schema for global spatial structure. As a result, they remain limited in inferring the topological layout of an environment from sparse observations and common sense alone. A promising direction is to develop Spatial Foundation Models pre-trained on large-scale 3D scenes and maps to capture general spatial regularities and physical priors, such as room connectivity and object co-occurrence. Used as generative priors for cognitive maps, such models could enable zero-shot completion of global map skeletons from sparse local evidence, helping agents move beyond the physical limits of immediate sensor views.

Lifelong Persistency in Dynamic Environments. While existing spatial intelligence systems can maintain map consistency within a current task cycle through continuous updates, they typically rely on a Static World Assumption, which presumes that object positions in the environment are fixed. When facing frequent object movements or the long-term evolution of environmental layouts in the real world, current systems often struggle to distinguish between localization errors and environmental changes. This failure frequently leads to catastrophic forgetting of old maps or conflicts between new and old information. To transcend the limitations of episodic tasks and achieve lifelong dynamic spatial maintenance, cognitive maps must evolve into 4D spatiotemporal representations that leverage active forgetting and structural consolidation to differentiate between short-term transient entities and long-term persistent backgrounds.

Cognitive Maps as Generative Simulators. At the core of spatial intelligence, cognitive maps need to advance beyond mere spatial repositories to become dynamic engines for forward-looking inference. Currently, spatial reasoning often depends on retrieving existing environmental data, which restricts the agent's ability to extrapolate unobserved or future states. To realize human-like spatial intelligence, future cognitive maps must evolve into generative world models. By serving as internalized mental simulators, cognitive maps can seamlessly close the loop from perception through reasoning to generation. This paradigm shift empowers agents to conduct counterfactual reasoning and simulate future states entirely within a conceptual space. Ultimately, by generating and evaluating spatial structures based on internalized physical laws, the cognitive map transcends representation, becoming the core mechanism that drives complex, anticipatory planning.

The Perception-Action Gap. Despite their widespread adoption in Embodied AI, mapping modules typically function as passive environmental recorders. By treating mapping as an isolated upstream task decoupled from planning, internal map uncertainties rarely influence agent behaviors, depriving agents of proactive environmental understanding. To address this, future research is

 doi:10.20944/preprints202605.1782.v1

anticipated to explore an Action-Centric Mapping paradigm. By leveraging map uncertainties or prediction errors as feedback signals, agents can be motivated to perform exploratory actions, verify hypotheses, and resolve blind spots. Such a transition would elevate cognitive map construction from static data accumulation to a dynamic hypothesis-verification process, ultimately closing the perception-action loop.

7. Conclusion

This survey revisits spatial intelligence from a cognitive map perspective and argues that perception, reasoning, and generation are three interrelated processes organized around a shared internal spatial representation. By defining cognitive maps through Abstraction, Globality, and Persistency, we provide an operational lens for understanding how agents construct structured representations from partial observations, perform inference and decision-making over them, and further realize them into external environments or simulated outcomes. From this viewpoint, diverse paradigms such as mapping systems, scene graphs, spatial memory, and world models can be understood as different instantiations of the same underlying representation, clarifying how their designs influence consistency, inference capability, and realizability. We hope that our cognitive-map framework and mechanism-centric taxonomy provide a principled foundation for analyzing, comparing, and developing future spatial intelligence systems.

Acknowledgments: This work is supported by the Natural Science Foundation of China under Grant Nos. 72225011,72434005 and 72293575.

References

1. Gardner, H. *Frames of mind: The theory of multiple intelligences*; Basic books, 2011.
2. Bellmund, J.L.; Gärdenfors, P.; Moser, E.I.; Doeller, C.F. Navigating cognition: Spatial codes for human thinking. *Science* **2018**, *362*, eaat6766.
3. Epstein, R.A.; Patai, E.Z.; Julian, J.B.; Spiers, H.J. The cognitive map in humans: spatial navigation and beyond. *Nature neuroscience* **2017**, *20*, 1504–1513.
4. Yang, J.; Yang, S.; Gupta, A.W.; Han, R.; Fei-Fei, L.; Xie, S. Thinking in space: How multimodal large language models see, remember, and recall spaces. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 10632–10643.
5. Yu, S.; et al. How far are vlms from visual spatial intelligence? a benchmark-driven perspective. *arXiv preprint arXiv:2509.18905* **2025**.
6. Felicia, G.; et al. From Perception to Action: Spatial AI Agents and World Models. *arXiv preprint arXiv:2602.01644* **2026**.
7. Peng, S.; Genova, K.; Jiang, C.; Tagliasacchi, A.; Pollefeys, M.; Funkhouser, T.; et al. Openscene: 3d scene understanding with open vocabularies. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 815–824.
8. Kerr, J.; Kim, C.M.; Goldberg, K.; Kanazawa, A.; Tancik, M. Lerf: Language embedded radiance fields. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 19729–19739.
9. Wei, Y.; et al. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 21729–21740.
10. Huang, Y.; Zheng, W.; Zhang, Y.; Zhou, J.; Lu, J. Tri-perspective view for vision-based 3d semantic occupancy prediction. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 9223–9232.
11. Hong, Y.; Zhen, H.; Chen, P.; Zheng, S.; et al. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems* **2023**, *36*, 20482–20494.
12. Ma, X.; et al. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474* **2022**.
13. Wang, Y.; Ji, Y.; Liu, Y.; Zhou, E.; Yang, Z.; Tian, Y.; Qin, Z.; Liu, Y.; Tan, H.; Chi, C.; et al. Towards cross-view point correspondence in vision-language models. *arXiv preprint arXiv:2512.04686* **2025**.

 doi:10.20944/preprints202605.1782.v1

14. Anderson, P.; et al. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3674–3683.
15. Chaplot, D.S.; Gandhi, D.P.; Gupta, A.; et al. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems* **2020**, *33*, 4247–4258.
16. Zitkovich, B.; et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In Proceedings of the Conference on Robot Learning. PMLR, 2023, pp. 2165–2183.
17. Ji, Y.; Tan, H.; Shi, J.; Hao, X.; Zhang, Y.; Zhang, H.; Wang, P.; Zhao, M.; Mu, Y.; An, P.; et al. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025, pp. 1724–1734.
18. Höllein, L.; Cao, A.; Owens, A.; Johnson, J.; Nießner, M. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 7909–7920.
19. Fridman, R.; Abecasis, A.; Kasten, Y.; Dekel, T. Scenescape: Text-driven consistent scene generation. *Advances in Neural Information Processing Systems* **2023**, *36*, 39897–39914.
20. Bruce, J.; Dennis, M.D.; Edwards, A.; Parker-Holder, J.; Shi, Y.; Hughes, E.; Lai, M.; Mavalankar, A.; Steigerwald, R.; Apps, C.; et al. Genie: Generative interactive environments. In Proceedings of the Forty-first International Conference on Machine Learning, 2024.
21. Hafner, D.; Lillicrap, T.; Ba, J.; Norouzi, M. Dream to Control: Learning Behaviors by Latent Imagination. In Proceedings of the International Conference on Learning Representations, 2020.
22. Hu, A.; Corrado, G.; Griffiths, N.; Murez, Z.; Gurau, C.; Yeo, H.; Kendall, A.; Cipolla, R.; Shotton, J. Model-based imitation learning for urban driving. *Advances in Neural Information Processing Systems* **2022**, *35*, 20703–20716.
23. Wang, Y.; He, J.; Fan, L.; Li, H.; Chen, Y.; Zhang, Z. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 14749–14759.
24. Tolman, E.C. Cognitive maps in rats and men. *Psychological review* **1948**, *55*, 189.
25. O'keefe, J.; Nadel, L. Précis of O'Keefe & Nadel's The hippocampus as a cognitive map. *Behavioral and Brain Sciences* **1979**, *2*, 487–494.
26. Hafting, T.; Fyhn, M.; Molden, S.; Moser, M.B.; Moser, E.I. Microstructure of a spatial map in the entorhinal cortex. *Nature* **2005**, *436*, 801–806.
27. Chang, X.; Ren, P.; Xu, P.; Li, Z.; Chen, X.; Hauptmann, A. A comprehensive survey of scene graphs: Generation and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2021**, *45*, 1–26.
28. Bae, J.; Shin, D.; Ko, K.; Lee, J.; Kim, U.H. A survey on 3D scene graphs: Definition, generation and application. In Proceedings of the International Conference on Robot Intelligence Technology and Applications. Springer, 2022, pp. 136–147.
29. Zhang, Y.; Ma, Z.; Li, J.; Qiao, Y.; Wang, Z.; Chai, J.; Wu, Q.; Bansal, M.; Kordjamshidi, P. Vision-and-Language Navigation Today and Tomorrow: A Survey in the Era of Foundation Models. *Transactions on Machine Learning Research* **2024**.
30. Tang, X.; Li, R.; Fan, X. Recent Advances in 3D Object and Scene Generation: A Survey. *arXiv preprint arXiv:2504.11734* **2025**.
31. Wen, B.; Xie, H.; Chen, Z.; Hong, F.; Liu, Z. 3d scene generation: A survey. *arXiv preprint arXiv:2505.05474* **2025**.
32. Zha, J.; Fan, Y.; Yang, X.; Gao, C.; Chen, X. How to enable LLM with 3D capacity? a survey of spatial reasoning in LLM. In Proceedings of the Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, 2025, pp. 10817–10825.
33. Zheng, X.; Dongfang, Z.; Jiang, L.; Zheng, B.; Guo, Y.; Zhang, Z.; Albanese, G.; Yang, R.; Ma, M.; Zhang, Z.; et al. Multimodal spatial reasoning in the large model era: A survey and benchmarks. *arXiv preprint arXiv:2510.25760* **2025**.
34. Feng, J.; Zeng, J.; Long, Q.; Chen, H.; Zhao, J.; Xi, Y.; Zhou, Z.; Yuan, Y.; Wang, S.; Zeng, Q.; et al. A survey of large language model-powered spatial intelligence across scales: Advances in embodied agents, smart cities, and earth science. *arXiv preprint arXiv:2504.09848* **2025**.
35. Manh, B.D.; Debnath, S.; Zhang, Z.; Damodaran, S.; Kumar, A.; Zhang, Y.; Mi, L.; Cambria, E.; Wang, L. Mind meets space: Rethinking agentic spatial intelligence from a neuroscience-inspired perspective. *arXiv preprint arXiv:2509.09154* **2025**.

 doi:10.20944/preprints202605.1782.v1

36. Narita, G.; Seno, T.; Ishikawa, T.; Kaji, Y. Panopticfusion: Online volumetric semantic mapping at the level of stuff and things. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2019, pp. 4205–4212.
37. Cartillier, V.; Ren, Z.; Jain, N.; Lee, S.; Essa, I.; Batra, D. Semantic mapnet: Building allocentric semantic maps and representations from egocentric views. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2021, Vol. 35, pp. 964–972.
38. Fang, X.; Fang, W.; Wang, C. Hierarchical semantic-augmented navigation: Optimal transport and graph-driven reasoning for vision-language navigation. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.
39. Sun, P.; Song, Y.; Liu, X.; Yang, X.; Wang, Q.; Li, T.; et al. 3d question answering for city scene understanding. In Proceedings of the Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 2156–2165.
40. Tan, H.; Hao, X.; Chi, C.; Lin, M.; Lyu, Y.; Cao, M.; Liang, D.; Chen, Z.; Lyu, M.; Peng, C.; et al. Roboos: A hierarchical embodied framework for cross-embodiment and multi-agent collaboration. *arXiv preprint arXiv:2505.03673* **2025**.
41. Anwar, A.; Welsh, J.; Biswas, J.; Pouya, S.; Chang, Y. Remembr: Building and reasoning over long-horizon spatio-temporal memory for robot navigation. In Proceedings of the 2025 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2025, pp. 2838–2845.
42. Wu, T.; et al. Video World Models with Long-term Spatial Memory. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.
43. Lu, Y.; Ren, X.; Yang, J.; Shen, T.; Wu, Z.; Gao, J.; Wang, Y.; Chen, S.; Chen, M.; Fidler, S.; et al. Infinicube: Unbounded and controllable dynamic 3d driving scene generation with world-guided video models. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 27272–27283.
44. Ding, X.; Gao, J.; Pan, C.; Wang, W.; Qin, J. History-Enhanced Two-Stage Transformer for Aerial Vision-and-Language Navigation. *arXiv preprint arXiv:2512.14222* **2025**.
45. Zhang, L.; Fu, H.; Hao, X.; Zhang, S.; Zhang, Q.; Liu, R.; Chen, L.; Ding, W. What You See is What You Reach: Towards Spatial Navigation with High-Level Human Instructions **2026**.
46. Yang, Y.; Yang, H.; Zhou, J.; Chen, P.; Zhang, H.; Du, Y.; Gan, C. 3D-mem: 3D scene memory for embodied exploration and reasoning. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 17294–17303.
47. Qi, Z.; Zhang, Z.; Fang, Y.; Wang, J.; Zhao, H. Gpt4scene: Understand 3d scenes from videos with vision-language models. *arXiv preprint arXiv:2501.01428* **2025**.
48. others, L.Z. MapNav: A Novel Memory Representation via Annotated Semantic Maps for Vision-and-Language Navigation, 2025.
49. Li, S.; Hou, J.; Huang, D.; Fu, Y.; Xue, X. Ali-UI: Enhancing Complex Vision-Language Navigation with Alignment of Unified Map and Instruction Parsing. In Proceedings of the Proceedings of the 33rd ACM International Conference on Multimedia, 2025, pp. 3913–3922.
50. Wen, S.; Zhang, Z.; Sun, Y.; Wang, Z. Ovl-map: An online visual language map approach for vision-and-language navigation in continuous environments. *IEEE Robotics and Automation Letters* **2025**.
51. Ouyang, K.; Liu, Y.; Wu, H.; Liu, Y.; Zhou, H.; Zhou, J.; Meng, F.; Sun, X. SpaceR: Reinforcing MLLMs in Video Spatial Reasoning. *arXiv preprint arXiv:2504.01805* **2025**.
52. Chandaka, B.; Wang, G.X.; Chen, H.; Che, H.; Zhai, A.J.; Wang, S. Human-like Navigation in a World Built for Humans. *arXiv preprint arXiv:2509.21189* **2025**.
53. Chapman, N.H.; Lehnert, C.; Browne, W.; Dayoub, F. Enhancing embodied object detection with spatial feature memory. In Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, 2025, pp. 6921–6931.
54. Li, S.; Huang, D.; He, Y.; Fu, Y.; Jiang, Y.G.; Xue, X. TP-MDDN: Task-Preferenced Multi-Demand-Driven Navigation with Autonomous Decision-Making. *arXiv preprint arXiv:2511.17225* **2025**.
55. Li, B.; Lu, R.j.; Zhou, Y.; Meng, J.; Zheng, W.S. Distilling LLM Prior to Flow Model for Generalizable Agent's Imagination in Object Goal Navigation. *arXiv preprint arXiv:2508.09423* **2025**.
56. Wu, W.; Mao, S.; Zhang, Y.; Xia, Y.; Dong, L.; Cui, L.; Wei, F. Mind's eye of LLMs: visualization-of-thought elicits spatial reasoning in large language models. *Advances in Neural Information Processing Systems* **2024**, *37*, 90277–90317.

 doi:10.20944/preprints202605.1782.v1

57. Ren, A.Z.; Clark, J.; Dixit, A.; Itkina, M.; Majumdar, A.; Sadigh, D. Explore until confident: Efficient exploration for embodied question answering. *arXiv preprint arXiv:2403.15941* **2024**.
58. Wang, Z.; Li, X.; Yang, J.; Liu, Y.; Jiang, S. Gridmm: Grid memory map for vision-and-language navigation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 15625–15636.
59. Huang, C.; Mees, O.; Zeng, A.; Burgard, W. Visual language maps for robot navigation. *arXiv preprint arXiv:2210.05714* **2022**.
60. Kim, B.; Kim, J.; Kim, Y.; Min, C.; Choi, J. Context-aware planning and environment-aware memory for instruction following embodied agents. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 10936–10946.
61. Chen, B.; Xia, F.; Ichter, B.; Rao, K.; Gopalakrishnan, K.; Ryoo, M.S.; Stone, A.; et al. Open-vocabulary queryable scene representations for real world planning. *arXiv preprint arXiv:2209.09874* **2022**.
62. Georgakis, G.; Schmeckpeper, K.; Wanchoo, K.; Dan, S.; Miltsakaki, E.; Roth, D.; Daniilidis, K. Cross-modal map learning for vision and language navigation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 15460–15470.
63. Chen, P.; et al. Weakly-supervised multi-granularity map learning for vision-and-language navigation. *Advances in Neural Information Processing Systems* **2022**, *35*, 38149–38161.
64. Gupta, S.; Davidson, J.; Levine, S.; Sukthankar, R.; Malik, J. Cognitive mapping and planning for visual navigation. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2616–2625.
65. Busch, F.L.; et al. One map to find them all: Real-time open-vocabulary mapping for zero-shot multi-object navigation. In Proceedings of the 2025 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2025, pp. 14835–14842.
66. Zheng, W.; Chen, W.; Huang, Y.; Zhang, B.; et al. Occworld: Learning a 3d occupancy world model for autonomous driving. In Proceedings of the European conference on computer vision. Springer, 2024, pp. 55–72.
67. Liu, Z.; Gu, Y.; Wang, Y.; Xue, X.; Fu, Y. ActiveVLA: Injecting Active Perception into Vision-Language-Action Models for Precise 3D Robotic Manipulation. *arXiv preprint arXiv:2601.08325* **2026**.
68. Gu, S.; Yin, W.; Jin, B.; Guo, X.; Wang, J.; Li, H.; Zhang, Q.; Long, X. Dome: Taming diffusion model into high-fidelity controllable occupancy world model. *arXiv preprint arXiv:2410.10429* **2024**.
69. Zhou, Z.; Hu, Y.; Zhang, L.; Li, Z.; Chen, S. BeliefMapNav: 3D Voxel-Based Belief Map for Zero-Shot Object Navigation. *arXiv preprint arXiv:2506.06487* **2025**.
70. Wang, H.; Agapito, L. 3d reconstruction with spatial memory. In Proceedings of the 2025 International Conference on 3D Vision (3DV). IEEE, 2025, pp. 78–89.
71. Hu, W.; Hong, Y.; Wang, Y.; Gao, L.; Wei, Z.; Yao, X.; Peng, N.; Bitton, Y.; Szpektor, I.; Chang, K.W. 3DLLM-Mem: Long-Term Spatial-Temporal Memory for Embodied 3D Large Language Model. *arXiv preprint arXiv:2505.22657* **2025**.
72. Zhou, S.; et al. Learning 3D Persistent Embodied World Models. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.
73. Zheng, S.; Yin, M.; Hu, W.; Li, X.; Shan, Y.; Fu, Y. VerseCrafter: Dynamic Realistic Video World Model with 4D Geometric Control. *arXiv preprint arXiv:2601.05138* **2026**.
74. Yang, Y.; Fan, L.; Shi, Z.; Peng, J.; Wang, F.; Zhang, Z. NeoVerse: Enhancing 4D World Model with in-the-wild Monocular Videos. *arXiv preprint arXiv:2601.00393* **2026**.
75. Zhao, J.; Wei, F.; Liu, Z.; Zhang, H.; Xu, C.; Lu, Y. Spatia: Video Generation with Updatable Spatial Memory. *arXiv preprint arXiv:2512.15716* **2025**.
76. Huang, J.; Hu, X.; Han, B.; Shi, S.; Tian, Z.; He, T.; Jiang, L. Memory forcing: Spatio-temporal memory for consistent scene generation on minecraft. *arXiv preprint arXiv:2510.03198* **2025**.
77. Wang, Z.; Lee, S.; Lee, G.H. Dynam3D: Dynamic Layered 3D Tokens Empower VLM for Vision-and-Language Navigation. *arXiv preprint arXiv:2505.11383* **2025**.
78. Wang, F.; Wu, J.; Cai, D.; Hong, Y.; Yan, Y. CogniMap3D: Cognitive 3D Mapping and Rapid Retrieval. In Proceedings of the The Fourteenth International Conference on Learning Representations, 2026.
79. Lee, P.Y.; Je, J.; Park, C.; Uy, M.A.; Guibas, L.; Sung, M. Perspective-aware reasoning in vision-language models via mental imagery simulation. *arXiv preprint arXiv:2504.17207* **2025**.
80. Mao, Y.; Ye, H.; Dong, W.; Zhang, C.; Zhang, H. Meta-Memory: Retrieving and Integrating Semantic-Spatial Memories for Robot Spatial Reasoning. *arXiv preprint arXiv:2509.20754* **2025**.

 doi:10.20944/preprints202605.1782.v1

81. Yang, Z.; Zheng, S.; Xie, T.; Xu, T.; Yu, B.; Wang, F.; Tang, J.; Liu, S.; Li, M. EfficientNav: Towards On-Device Object-Goal Navigation with Navigation Map Caching and Retrieval. *arXiv preprint arXiv:2510.18546* **2025**.
82. Gholami, M.; Rezaei, A.; Weimin, Z.; Mao, S.; Zhou, S.; Zhang, Y.; Akbari, M. Spatial reasoning with vision-language models in ego-centric multi-view scenes. *arXiv preprint arXiv:2509.06266* **2025**.
83. Park, J.; Cho, J.; Ahn, S. MrSteve: Instruction-Following Agents in Minecraft with What-Where-When Memory. In Proceedings of the The Thirteenth International Conference on Learning Representations, 2025.
84. Fan, Y.; Ma, X.; Su, R.; Guo, J.; Wu, R.; et al. Embodied videoagent: Persistent memory from egocentric videos and embodied sensors enables dynamic scene understanding. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 6342–6352.
85. Ma, W.; Sun, S.; Yu, T.; Wang, R.; et al. Thinking with Blueprints: Assisting Vision-Language Models in Spatial Reasoning via Structured Object Representation. *arXiv preprint arXiv:2601.01984* **2026**.
86. Huang, Y.; Xu, W.; Zhang, W.; Zhi, H.; Huang, J.; Xu, Y.; Sun, Y.; et al. Video2Layout: Recall and Reconstruct Metric-Grounded Cognitive Map for Spatial Reasoning. *arXiv preprint arXiv:2511.16160* **2025**.
87. Kurenkov, A.; Lingelbach, M.; Agarwal, T.; Jin, E.; Li, C.; Zhang, R.; Fei-Fei, L.; Wu, J.; Savarese, S.; Martın-Martın, R. Modeling dynamic environments with scene graph memory. In Proceedings of the International Conference on Machine Learning. PMLR, 2023, pp. 17976–17993.
88. Li, H.; Wang, Z.; Yang, X.; Yang, Y.; Mei, S.; Zhang, Z. Memonav: Working memory model for visual navigation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 17913–17922.
89. Zhang, R.; An, G.; Hao, Y.; Wu, D.O. Bridging visual and textual semantics: Towards consistency for unbiased scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2024**, *46*, 7102–7119.
90. Xu, W.; Ila, V.; Zhou, L.; Jin, C.T. TB-HSU: Hierarchical 3D Scene Understanding with Contextual Affordances. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2025, Vol. 39, pp. 8960–8968.
91. Cong, Y.; Yi, J.; Rosenhahn, B.; Yang, M.Y. Ssgvs: Semantic scene graph-to-video synthesis. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2555–2565.
92. Chen, J.; et al. Mapgpt: Map-guided prompting with adaptive path planning for vision-and-language navigation. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 9796–9810.
93. Hou, J.; Xue, X.; Zeng, T. Hi-Dyna Graph: Hierarchical Dynamic Scene Graph for Robotic Autonomy in Human-Centric Environments. *arXiv preprint arXiv:2506.00083* **2025**.
94. Jin, Q.; Wu, Y.; Chen, C. PanoNav: Mapless Zero-Shot Object Navigation with Panoramic Scene Parsing and Dynamic Memory. *arXiv preprint arXiv:2511.06840* **2025**.
95. Hughes, N.; Chang, Y.; Carlone, L. Hydra: A real-time spatial perception system for 3D scene graph construction and optimization. *arXiv preprint arXiv:2201.13360* **2022**.
96. Yang, J.; Cen, J.; Peng, W.; Liu, S.; Hong, F.; Li, X.; Zhou, K.; Chen, Q.; Liu, Z. 4d panoptic scene graph generation. *Advances in Neural Information Processing Systems* **2023**, *36*, 69692–69705.
97. Liu, R.; Wang, X.; Wang, W.; Yang, Y. Bird's-eye-view scene graph for vision-language navigation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 10968–10980.
98. Gu, Q.; et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024, pp. 5021–5028.
99. Zhai, G.; Cai, X.; Huang, D.; Di, Y.; Manhardt, F.; Tombari, F.; Navab, N.; Busam, B. Sg-bot: Object rearrangement via coarse-to-fine robotic imagination on scene graphs. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024, pp. 4303–4310.
100. Cao, Y.; Zhang, J.; Yu, Z.; Liu, S.; Qin, Z.; Zou, Q.; Du, B.; Xu, K. Cognav: Cognitive process modeling for object goal navigation with llms. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 9550–9560.
101. Zhu, F.; Wang, H.; Xie, Y.; Gu, J.; Ding, T.; Yang, J.; Jiang, H. Struct2D: A Perception-Guided Framework for Spatial Reasoning in Large Multimodal Models. *arXiv preprint arXiv:2506.04220* **2025**.
102. Gong, Z.; Li, R.; Hu, T.; Qiu, R.; Kong, L.; Zhang, L.; Zhao, G.; Ding, Y.; Liang, J. Stairway to Success: An Online Floor-Aware Zero-Shot Object-Goal Navigation Framework via LLM-Driven Coarse-to-Fine Exploration. *IEEE Robotics and Automation Letters* **2026**.

 doi:10.20944/preprints202605.1782.v1

103. Zhang, X.; et al. Agent journey beyond rgb: Unveiling hybrid semantic-spatial environmental representations for vision-and-language navigation. *arXiv preprint arXiv:2412.06465* **2024**.
104. Xu, H.; Hu, Y.; Gao, C.; Zhu, Z.; Zhao, Y.; et al. Geonav: Empowering mllms with explicit geospatial reasoning abilities for language-goal aerial navigation. *arXiv preprint arXiv:2504.09587* **2025**.
105. Yin, H.; et al. GC-VLN: Instruction as Graph Constraints for Training-free Vision-and-Language Navigation, 2025.
106. Feng, S.; Wang, Z.; Li, Y.; Kong, R.; Cai, H.; Wang, S.; Lee, G.H.; Li, P.; Jiang, S. VPN: Visual Prompt Navigation. *arXiv preprint arXiv:2508.01766* **2025**.
107. Qian, Z.; Ma, Y.; Ji, J.; Sun, X. X-refseg3d: Enhancing referring 3d instance segmentation via structured cross-modal graph neural networks. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2024, Vol. 38, pp. 4551–4559.
108. Zhang, J.; Zhu, G.; Li, S.; Liu, X.; Song, H.; Tang, X.; Feng, C. Multiview scene graph. *Advances in Neural Information Processing Systems* **2024**, *37*, 17761–17788.
109. Zhang, Y.; et al. EmbodiedVSR: Dynamic scene graph-guided chain-of-thought reasoning for visual spatial tasks. *arXiv preprint arXiv:2503.11089* **2025**.
110. Song, Y.; et al. Scene-driven multimodal knowledge graph construction for embodied AI. *IEEE Transactions on Knowledge and Data Engineering* **2024**, *36*, 6962–6976.
111. Lin, C.; Yadong, M. InstructScene: Instruction-Driven 3D Indoor Scene Synthesis with Semantic Graph Prior. In Proceedings of the The Twelfth International Conference on Learning Representations, 2023.
112. Zhai, G.; et al. CommonScenes: Generating Commonsense 3D Indoor Scenes with Scene Graph Diffusion. In Proceedings of the Advances in Neural Information Processing Systems, 2023, Vol. 36, pp. 30026–30038.
113. Wei, Y.; et al. Planner3d: Llm-enhanced graph prior meets 3d indoor scene explicit regularization. *IEEE transactions on pattern analysis and machine intelligence* **2025**.
114. Luo, A.; Zhang, Z.; Wu, J.; Tenenbaum, J.B. End-to-end optimization of scene layout. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3754–3763.
115. Dhamo, H.; Manhardt, F.; Navab, N.; Tombari, F. Graph-to-3d: End-to-end generation and manipulation of 3d scenes using scene graphs. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 16352–16361.
116. Hong, J.; et al. HiGS: Hierarchical Generative Scene Framework for Multi-Step Associative Semantic Spatial Composition, 2025.
117. Wang, K.; Lin, Y.A.; Weissmann, B.; Savva, M.; Chang, A.X.; Ritchie, D. Planit: Planning and instantiating indoor scenes with relation graph and spatial prior networks. *ACM Transactions on Graphics (TOG)* **2019**, *38*, 1–15.
118. Fu, Q.; Chen, X.; Wang, X.; Wen, S.; Zhou, B.; Fu, H. Adaptive synthesis of indoor scenes via activity-associated object relation graphs. *ACM Transactions on Graphics (TOG)* **2017**, *36*, 1–13.
119. Kwon, O.; Kim, N.; Choi, Y.; Yoo, H.; Park, J.; Oh, S. Visual graph memory with unsupervised representation for visual navigation. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 15890–15899.
120. Kim, N.; Kwon, O.; Yoo, H.; Choi, Y.; Park, J.; Oh, S. Topological semantic graph memory for image-goal navigation. In Proceedings of the Conference on Robot Learning. PMLR, 2023, pp. 393–402.
121. Liang, J.; Wang, Y.; Wang, Z.; Liu, M.; Fu, R.; Wang, Z.; Qin, B. GTR: A Grafting-Then-Reassembling Framework for Dynamic Scene Graph Generation. In Proceedings of the Ijcai, 2023, pp. 1177–1185.
122. Liu, P.; et al. Toponav: Topological graphs as a key enabler for advanced object navigation. *arXiv preprint arXiv:2509.01364* **2025**.
123. Jia, T.; et al. BrainyMP: Enhancing Motion Planning Using Graph Neural Network Inspired by Brain Spatial Relational Memory. *IEEE Transactions on Intelligent Transportation Systems* **2025**.
124. Liu, Q.; et al. Visuomotor navigation for embodied robots with spatial memory and semantic reasoning cognition. *IEEE Transactions on Neural Networks and Learning Systems* **2024**.
125. Singh, K.P.; Salvador, J.; Weihs, L.; Kembhavi, A. Scene graph contrastive learning for embodied navigation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 10884–10894.
126. Yang, H.; Jiao, Z.; Wang, S.; Niu, Y.; Liu, S.; Liu, H. Integrated Exploration and Sequential Manipulation on Scene Graph with LLM-based Situated Replanning. *arXiv preprint arXiv:2602.04419* **2026**.

 doi:10.20944/preprints202605.1782.v1

127. Rana, K.; Haviland, J.; Garg, S.; Abou-Chakra, J.; Reid, I.; Suenderhauf, N. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. *arXiv preprint arXiv:2307.06135* **2023**.
128. Wang, Z.; Yu, B.; Zhao, J.; Sun, W.; Hou, S.; Liang, S.; Hu, X.; Han, Y.; Gan, Y. Karma: Augmenting embodied ai agents with long-and-short term memory systems. In Proceedings of the 2025 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2025, pp. 1–8.
129. Loo, J.; Wu, Z.; Hsu, D. Open scene graphs for open-world object-goal navigation. *The International Journal of Robotics Research* **2025**, p. 02783649251369549.
130. Kar, A.; Prakash, A.; Liu, M.Y.; Cameracci, E.; Yuan, J.; Rusiniak, M.; Acuna, D.; Torralba, A.; Fidler, S. Meta-Sim: Learning to Generate Synthetic Datasets. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019.
131. Devaranjan, J.; Kar, A.; Fidler, S. Meta-sim2: Unsupervised learning of scene structure for synthetic data generation. In Proceedings of the European Conference on Computer Vision. Springer, 2020, pp. 715–733.
132. Yang, Y.; Sun, F.Y.; Weihs, L.; VanderBilt, E.; Herrasti, A.; Han, W.; Wu, J.; Haber, N.; Krishna, R.; Liu, L.; et al. Holodeck: Language guided generation of 3d embodied ai environments. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 16227–16237.
133. Yu, C.R.; Chae, D.; Seo, D.; Lee, S.; Im, H.; Kim, J. Scene Graph-Guided Proactive Replanning for Failure-Resilient Embodied Agent. *arXiv preprint arXiv:2508.11286* **2025**.
134. Lei, M.; et al. RoboMemory: A Brain-inspired Multi-memory Agentic Framework for Interactive Environmental Learning in Physical Embodied Systems. *arXiv preprint arXiv:2508.01415* **2025**.
135. Fan, Z.; et al. LayoutAgent: A Vision-Language Agent Guided Compositional Diffusion for Spatial Layout Planning, 2025.
136. Gao, G.; Liu, W.; Chen, A.; Geiger, A.; Schölkopf, B. Graphdreamer: Compositional 3d scene synthesis from scene graphs. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 21295–21304.
137. Han, Z.; Wang, X.; Liu, B.; Lyu, Q.; Shang, Z.; Dong, J.; et al. SeqWalker: Sequential-Horizon Vision-and-Language Navigation with Hierarchical Planning. *arXiv preprint arXiv:2601.04699* **2026**.
138. Zhang, H.; Liu, M.; Li, Z.; Wen, H.; Guan, W.; Wang, Y.; Nie, L. Spatial Understanding from Videos: Structured Prompts Meet Simulation Data. *arXiv preprint arXiv:2506.03642* **2025**.
139. Li, M.; Patil, A.G.; Xu, K.; Chaudhuri, S.; Khan, O.; Shamir, A.; Tu, C.; Chen, B.; Cohen-Or, D.; Zhang, H. Grains: Generative recursive autoencoders for indoor scenes. *ACM Transactions on Graphics (TOG)* **2019**, *38*, 1–16.
140. Gao, L.; Sun, J.M.; Mo, K.; Lai, Y.K.; Guibas, L.J.; Yang, J. SceneHGN: Hierarchical Graph Networks for 3D Indoor Scene Generation With Fine-Grained Geometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2023**, *45*, 8902–8919. https://doi.org/10.1109/TPAMI.2023.3237577.
141. Hahn, M.; et al. No rl, no simulation: Learning to navigate without navigating. *Advances in Neural Information Processing Systems* **2021**, *34*, 26661–26673.
142. Zhou, X.; et al. FSR-VLN: Fast and Slow Reasoning for Vision-Language Navigation with Hierarchical Multi-modal Scene Graph. *arXiv preprint arXiv:2509.13733* **2025**.
143. Hou, J.; Xiao, Y.; Xue, X.; Zeng, T. LOG-Nav: Efficient Layout-Aware Object-Goal Navigation with Hierarchical Planning, 2025.
144. Garg, S.; Craggs, D.; Bhat, V.; Mares, L.; Podgorski, S.; Krishna, M.; Dayoub, F.; Reid, I. Objectreact: Learning object-relative control for visual navigation. *arXiv preprint arXiv:2509.09594* **2025**.
145. An, D.; Qi, Y.; Li, Y.; Huang, Y.; Wang, L.; Tan, T.; Shao, J. Bevbert: Multimodal map pre-training for language-guided navigation. *arXiv preprint arXiv:2212.04385* **2022**.
146. Yang, Z.; et al. Mmgdreamer: Mixed-modality graph for geometry-controllable 3d indoor scene generation. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2025, Vol. 39, pp. 9391–9399.
147. Wang, B.; et al. MG-Nav: Dual-Scale Visual Navigation via Sparse Spatial Memory, 2025.
148. Rosinol, A.; Gupta, A.; Abate, M.; Shi, J.; Carlone, L. 3D Dynamic Scene Graphs: Actionable Spatial Perception with Places, Objects, and Humans. In Proceedings of the Robotics: Science and Systems, 2020.
149. Gao, X.; Zhang, Z.; Chen, D.Z.; Xu, S.; Quan, L.; Pérez-Pellitero, E.; Jang, Y. Map2Thought: Explicit 3D Spatial Reasoning via Metric Cognitive Maps. *arXiv preprint arXiv:2601.11442* **2026**.
150. Xue, X.; Hu, J.; Luo, M.; Shichao, X.; Chen, J.; Xie, Z.; Kuichen, Q.; Wei, G.; Xu, M.; Chu, Z. Omninav: A unified framework for prospective exploration and visual-language navigation. *arXiv preprint arXiv:2509.25687* **2025**.

 doi:10.20944/preprints202605.1782.v1

151. Liu, X.; Tai, Y.W.; Tang, C.K. Agentic 3D Scene Generation with Spatially Contextualized VLMs, 2025.
152. Ruan, S.; Wang, L.; Kang, C.; Zhu, Q.; Liu, S.; Wei, X.; Su, H. From reactive to cognitive: brain-inspired spatial intelligence for embodied agents. *intelligence (AGI)* **2025**, *3*, 10.
153. Li, K.; Xu, Q.; Qian, T.; Fu, Y.; Jiao, Y.; Wang, X. CLiViS: Unleashing Cognitive Map through Linguistic-Visual Synergy for Embodied Visual Reasoning. *arXiv preprint arXiv:2506.17629* **2025**.
154. Saxena, S.; et al. Grapheqa: Using 3d semantic scene graphs for real-time embodied question answering. *arXiv preprint arXiv:2412.14480* **2024**.
155. He, L.; et al. Mem4Nav: Boosting Vision-and-Language Navigation in Urban Environments with a Hierarchical Spatial-Cognition Long-Short Memory System. *arXiv preprint arXiv:2506.19433* **2025**.
156. Tan, H.; Chi, C.; Chen, X.; Ji, Y.; Zhao, Z.; Hao, X.; Lyu, Y.; Cao, M.; Zhao, J.; Lyu, H.; et al. Roboos-next: A unified memory-based framework for lifelong, scalable, and robust multi-robot collaboration. *arXiv preprint arXiv:2510.26536* **2025**.
157. Booker, M.; Byrd, G.; Kemp, B.; Schmidt, A.; Rivera, C. Embodiedrag: Dynamic 3d scene graph retrieval for efficient and scalable robot task planning. *arXiv preprint arXiv:2410.23968* **2024**.
158. Wu, Z.; Feng, M.; Wang, Y.; Xie, H.; Dong, W.; et al. External knowledge enhanced 3d scene generation from sketch. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 286–304.
159. Zhu, Z.; et al. Move to understand a 3d scene: Bridging visual grounding and exploration for efficient and versatile embodied navigation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 8120–8132.
160. Zhou, E.; Chi, C.; Li, Y.; An, J.; Zhang, J.; Rong, S.; Han, Y.; Ji, Y.; Liu, M.; et al. RoboTracer: Mastering Spatial Trace with Reasoning in Vision-Language Models for Robotics. *arXiv preprint arXiv:2512.13660* **2025**.
161. Fan, S.; et al. Scene map-based prompt tuning for navigation instruction generation. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 6898–6908.
162. Chen, B.; Kang, J.; Zhong, P.; Liang, Y.; Sheng, Y.; Wang, J. Embodied Contrastive Learning with Geometric Consistency and Behavioral Awareness for Object Navigation. In Proceedings of the Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 4776–4785.
163. Luo, J.; Zhang, J.; Yang, J.; Huang, S.; Cai, B. Learning Bird's Eye View scene graph and knowledge-inspired policy for embodied visual navigation. *Knowledge-Based Systems* **2025**, p. 113959.
164. Hong, Y.; Zhou, Y.; Zhang, R.; Dernoncourt, F.; Bui, T.; Gould, S.; Tan, H. Learning navigational visual representations with semantic map supervision. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 3055–3067.
165. Zhang, L.; Hao, X.; Xu, Q.; Zhang, Q.; Zhang, X.; Wang, P.; Zhang, J.; Wang, Z.; Zhang, S.; Xu, R.M. A novel memory representation via annotated semantic maps for vlm-based vision-and-language navigation. *arXiv preprint arXiv:2502.13451* **2025**.
166. Li, P.; Song, P.; Li, W.; Guo, W.; Yao, H.; Xu, Y.; Liu, D.; Xiong, H. See&trek: Training-free spatial prompting for multimodal large language model. *arXiv preprint arXiv:2509.16087* **2025**.
167. Hou, J.; Xiao, Y.; Xue, X.; Zeng, T. LOG-Nav: Efficient Layout-Aware Object-Goal Navigation with Hierarchical Planning. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2026), 2026.
168. Zhou, H.; et al. LagMemo: Language 3D Gaussian Splatting Memory for Multi-modal Open-vocabulary Multi-goal Visual Navigation. *arXiv preprint arXiv:2510.24118* **2025**.
169. Duan, H.; Luo, S.; Deng, Z.; Chen, Y.; Chiang, Y.; Liu, Y.; Liu, F.; Wang, X. CAUSALNAV: A Long-term Embodied Navigation System for Autonomous Mobile Robots in Dynamic Outdoor Scenarios. *IEEE Robotics and Automation Letters* **2026**.
170. Wang, P.; et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* **2024**.
171. Tan, H.; Ji, Y.; Hao, X.; Chen, X.; Wang, P.; Wang, Z.; Zhang, S. Reason-RFT: Reinforcement Fine-Tuning for Visual Reasoning of Vision Language Models. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems.
172. Team, B.R.; Cao, M.; Tan, H.; Ji, Y.; Chen, X.; Lin, M.; Li, Z.; Cao, Z.; Wang, P.; Zhou, E.; et al. Robobrain 2.0 technical report. *arXiv preprint arXiv:2507.02029* **2025**.
173. Tan, H.; Zhou, E.; Li, Z.; Xu, Y.; Ji, Y.; Chen, X.; Chi, C.; Wang, P.; Jia, H.; Ao, Y.; et al. RoboBrain 2.5: Depth in Sight, Time in Mind. *arXiv preprint arXiv:2601.14352* **2026**.

 doi:10.20944/preprints202605.1782.v1

174. Song, Z.; et al. Maniplvm-r1: Reinforcement learning for reasoning in embodied manipulation with large vision-language models. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2026, Vol. 40, pp. 18558–18566.
175. Ji, Y.; Liu, Y.; Zhang, Z.; Zhang, Z.; Zhao, Y.; Hao, X.; Zhou, G.; Zhang, X.; Zheng, X. Enhancing adversarial robustness of vision-language models through low-rank adaptation. In Proceedings of the Proceedings of the 2025 International Conference on Multimedia Retrieval, 2025, pp. 550–559.
176. Ji, Y.; Wang, Y.; Liu, Y.; Hao, X.; Liu, Y.; Zhao, Y.; Lyu, H.; Zheng, X. Visualtrans: A benchmark for real-world visual transformation reasoning. *arXiv preprint arXiv:2508.04043* **2025**.
177. Tan, H.; Chen, S.; Xu, Y.; Wang, Z.; Ji, Y.; Chi, C.; Lyu, Y.; Zhao, Z.; Chen, X.; Co, P.; et al. Robo-Dopamine: General Process Reward Modeling for High-Precision Robotic Manipulation. *arXiv preprint arXiv:2512.23703* **2025**.
178. Lyu, H.; Chen, C.; Ji, Y.; Xu, C. Egoprompt: Prompt learning for egocentric action recognition. In Proceedings of the Proceedings of the 33rd ACM International Conference on Multimedia, 2025, pp. 2762–2770.
179. Ji, Y.; Liu, Y.; Tan, H.; Huang, X.; Huang, F.; Xu, Y.; Chi, C.; Zhao, Y.; Lyu, H.; Co, P.; et al. PRM-as-a-Judge: A Dense Evaluation Paradigm for Fine-Grained Robotic Auditing. *arXiv preprint arXiv:2603.21669* **2026**.
180. Li, Z.; Lu, Y.; Mu, Y.; Qiao, H. Cog-GA: A Large Language Models-based Generative Agent for Vision-Language Navigation in Continuous Environments. *arXiv preprint arXiv:2409.02522* **2024**.
181. Luo, A.; Zhang, Z.; Wu, J.; Tenenbaum, J.B. End-to-End Optimization of Scene Layout. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.
182. Ma, R.; et al. Language-driven synthesis of 3D scenes from scene databases. *ACM Transactions on Graphics (TOG)* **2018**, *37*, 1–16.
183. Sun, F.Y.; et al. Layoutvlm: Differentiable optimization of 3d layout via vision-language models. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 29469–29478.
184. Sun, Q.; Zhou, H.; Zhou, W.; Li, L.; Li, H. Forest2seq: Revitalizing order prior for sequential indoor scene synthesis. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 251–268.
185. Ji, Y.; Tan, H.; Chi, C.; Xu, Y.; Zhao, Y.; Zhou, E.; Lyu, H.; Wang, P.; Wang, Z.; Zhang, S.; et al. Mathsticks: A benchmark for visual symbolic compositional reasoning with matchstick puzzles. *arXiv preprint arXiv:2510.00483* **2025**.
186. Zhao, Y.; Ji, Y.; Hao, X.; Li, S. FastRSR: Efficient and Accurate Road Surface Reconstruction in Bird's Eye View. In Proceedings of the Proceedings of the 33rd ACM International Conference on Multimedia, 2025, pp. 9080–9089.
187. Zhai, G.; et al. Commonscenes: Generating commonsense 3d indoor scenes with scene graph diffusion. *Advances in Neural Information Processing Systems* **2023**, *36*, 30026–30038.
188. Chen, Z.; et al. Scenedreamer: Unbounded 3d scene generation from 2d image collections. *IEEE transactions on pattern analysis and machine intelligence* **2023**, *45*, 15562–15576.
189. Xie, H.; Chen, Z.; Hong, F.; Liu, Z. Citydreamer: Compositional generative model of unbounded 3d cities. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 9666–9675.
190. Gao, R.; Chen, K.; Xie, E.; HONG, L.; Li, Z.; et al. MagicDrive: Street View Generation with Diverse 3D Geometry Control. In Proceedings of the The Twelfth International Conference on Learning Representations, 2023.
191. Xie, H.; Chen, Z.; Hong, F.; Liu, Z. Generative Gaussian splatting for unbounded 3D city generation. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 6111–6120.
192. Liu, Y.; Li, X.; Zhang, Y.; Qi, L.; Li, X.; Wang, W.; Li, C.; Li, X.; Yang, M.H. Controllable 3D outdoor scene generation via scene graphs. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2025.
193. Fang, C.; et al. SpatialGen: Layout-guided 3D Indoor Scene Generation. In Proceedings of the Thirteenth International Conference on 3D Vision, 2026.
194. Fang, C.; Dong, Y.; Luo, K.; Hu, X.; Shrestha, R.; Tan, P. Ctrl-room: Controllable text-to-3d room meshes generation with layout constraints. In Proceedings of the 2025 International Conference on 3D Vision (3DV). IEEE, 2025, pp. 692–701.

 doi:10.20944/preprints202605.1782.v1

195. Bahmani, S.; Park, J.J.; Paschalidou, D.; Yan, X.; Wetzstein, G.; Guibas, L.; Tagliasacchi, A. Cc3d: Layout-conditioned generation of compositional 3d scenes. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 7171–7181.
196. Cong, Y.; et al. SSGVS: Semantic Scene Graph-to-Video Synthesis. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2023, pp. 2555–2565.
197. Ouyang, K.; et al. SpaceR: Reinforcing MLLMs in Video Spatial Reasoning, 2025.
198. Zhang, Z.; .; et al. Think3D: Thinking with Space for Spatial Reasoning. *arXiv preprint arXiv:2601.13029* **2026**.
199. Jiang, J.; Yang, Y.; Deng, Y.; Ma, C.; Zhang, J. BEVNav: Robot Autonomous Navigation Via Spatial-Temporal Contrastive Learning in Bird's-Eye View. *IEEE Robotics and Automation Letters* **2024**.
200. Li, B.; Lu, R.J.; Zhou, Y.; Meng, J.; Zheng, W.S. Distilling LLM Prior to Flow Model for Generalizable Agent's Imagination in Object Goal Navigation. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems.
201. Ju, Y.; et al. MomaGraph: State-Aware Unified Scene Graphs with Vision-Language Model for Embodied Task Planning. *arXiv preprint arXiv:2512.16909* **2025**.