A Survey of Spatial Intelligence and Multimodal Learning in Large Language Models

www.surveyx.cn

Abstract

This survey examines the integration of spatial intelligence and multimodal learning within the context of large language models (LLMs), highlighting their transformative potential across various domains. Spatial intelligence, encompassing spatial reasoning and cognition, is fundamental for tasks in AI, robotics, and cognitive science. The paper reviews the historical development and current state of research, emphasizing the role of spatial reasoning in problem-solving and decision-making. It analyzes the capabilities and limitations of LLMs in spatial reasoning tasks, noting the challenges of spatial hallucination and context inconsistency. Vision-language integration and multimodal learning are explored as methodologies to enhance spatial reasoning, with benchmarks such as DriveMLLM and SpatialPIN providing frameworks for evaluation. Real-world applications, including robotics and navigation, demonstrate the practical implications of these technologies. Despite advancements, challenges persist, particularly in achieving human-like performance in complex tasks. Future research directions include refining automated labeling techniques, exploring heuristic combinations, and enhancing multimodal model efficiency while addressing ethical considerations. The survey underscores the importance of integrating spatial intelligence and multimodal learning to advance AI capabilities, paving the way for more sophisticated and adaptable systems capable of nuanced spatial understanding.

1 Introduction

1.1 Overview of Paper Structure

This survey investigates the interplay between spatial intelligence and multimodal learning within large language models (LLMs), focusing on their performance in spatial reasoning tasks, the influence of innovative prompting techniques, and the potential of LLMs to yield new insights in spatial contexts, as supported by recent empirical studies and benchmark datasets [1, 2, 3, 4]. The introduction establishes the significance of spatial intelligence across fields such as artificial intelligence (AI), robotics, and cognitive science, paving the way for a discussion on the integration of spatial reasoning with LLMs and the benefits of vision-language integration and multimodal learning.

Section 2 provides definitions and a comprehensive background on key concepts, including spatial intelligence, spatial reasoning, LLMs, vision-language integration, multimodal learning, and spatial cognition. It reviews the historical development and current state of research in these domains, establishing a foundational understanding for readers.

Section 3 explores the cognitive processes underlying spatial intelligence and cognition, assessing how they can be effectively modeled in artificial systems. It emphasizes the importance of spatial intelligence in fields such as architecture, STEM, and medicine while addressing the limitations of current AI models, including GPT-4, in comprehending spatial transformations, especially in three-dimensional contexts. By evaluating model performance through tasks like the Revised Purdue Spatial Visualization Test: Visualization of Rotations (Revised PSVT:R) and augmented reality applications,

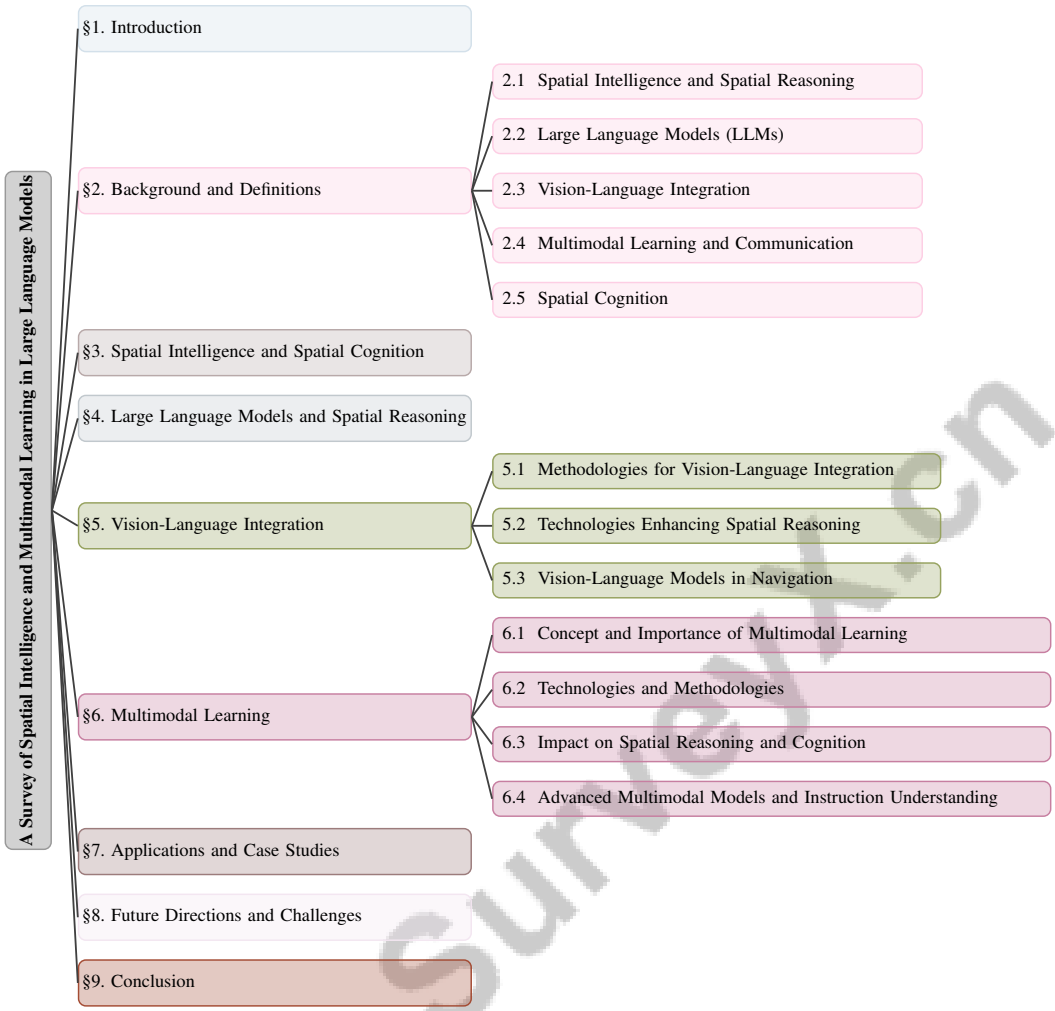


Figure 1: chapter structure

this section illustrates how supplementary information can enhance AI's spatial reasoning capabilities. Innovative approaches, such as disentangling information extraction from reasoning processes, are discussed to improve AI's generalizability and effectiveness in spatial reasoning tasks, ultimately aiming to better support spatial learning in assembly and manufacturing processes [5, 6]. The role of spatial reasoning in problem-solving and decision-making is also examined, highlighting the enhancement potential through AI technologies.

In Section 4, the capabilities and limitations of LLMs in spatial reasoning tasks are analyzed, reviewing existing studies and experiments to evaluate their spatial understanding and reasoning abilities. This section discusses challenges and potential solutions for enhancing spatial reasoning in LLMs, emphasizing the necessity for comprehensive benchmarks to assess model performance, as noted by Wang et al. [7].

Section 5 investigates vision-language integration, detailing methodologies and technologies employed to improve comprehension and reasoning in AI systems. The effectiveness of various approaches in enhancing spatial reasoning capabilities is assessed, particularly in applications that have successfully integrated vision-language models. This evaluation highlights the development of benchmarks like SpatialEval, which addresses vital aspects of spatial reasoning, including relationship understanding and navigation. The impact of disentangling information extraction from reasoning processes on model generalizability is also examined. Furthermore, the integration of LLMs into visual domains, particularly through visual-LLMs, showcases significant advancements in spatial awareness and task performance, such as visual question answering, underscoring the importance of fine-tuning techniques and optimal coordinate representations [8, 5, 9].

Section 6 discusses multimodal learning, emphasizing its role in enhancing spatial intelligence through the integration of diverse data types. The current landscape of multimodal learning approaches and technologies is analyzed, focusing on their effects on spatial reasoning and cognition, especially regarding large multimodal models (LMMs) and vision-language models (VLMs). Recent studies indicate that while LMMs excel in various vision and language tasks, their spatial reasoning capabilities remain underexplored, with findings suggesting that enhancements like bounding boxes and scene graphs can improve spatial understanding. Nonetheless, challenges persist, as these models often struggle with human-perspective queries and complex multi-hop spatial questions. Benchmarks such as Spatial-MM and SpatialEval have been introduced to evaluate spatial reasoning across multiple dimensions, revealing that VLMs frequently underperform compared to LLMs when visual input is available, highlighting the importance of textual context in enhancing model performance. This review aims to inform future research and development in multimodal models to bridge the gap in spatial intelligence relative to human cognition [8, 2].

Section 7 presents real-world applications and case studies where spatial intelligence, spatial reasoning, and multimodal learning have been effectively implemented. Applications such as sMoRe (Spatial Mapping and Object Rendering Environment) introduced by Xing et al. [10] illustrate the practical implications and contributions to advancements in fields like robotics and navigation.

In Section 8, we critically examine the significant challenges and limitations currently faced in the research landscape of multimodal language models, particularly in their application to disciplines such as chemistry, environmental science, and urban planning. Specific shortcomings, including issues with spatial reasoning, cross-modal information synthesis, and multi-step logical inference, are highlighted through evaluations of models like MaCBench and GPT-4V. Promising future research directions are discussed, emphasizing the need for improved training data curation and methodologies to enhance these systems' capabilities across various scientific workflows and interdisciplinary applications [11, 12]. Emerging trends and technologies that could shape the future of spatial reasoning in AI systems are also highlighted, along with considerations for innovative approaches and evaluation frameworks.

The conclusion synthesizes key findings and insights from the survey, reflecting on the importance of integrating spatial intelligence and multimodal learning in AI research and applications. It discusses the significant implications and future opportunities presented by advanced technologies, particularly regarding retrieval-augmented multimodal reasoning and spatial intelligence, which are crucial for enhancing AI capabilities in complex reasoning tasks and improving educational tools for spatial learning [6, 13].The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Spatial Intelligence and Spatial Reasoning

Spatial intelligence is the capacity to navigate and reason about spatial environments, essential in cognitive science and AI [14]. It supports tasks such as grounding UI elements in natural language and interpreting neural signals through brain data integration. Enhanced knowledge graph embedding methods using geometric features improve link prediction accuracy in geographic contexts [15]. Spatial reasoning, a core component of spatial intelligence, involves understanding and manipulating spatial relationships. It is crucial for comprehending 3D objects from 2D representations and is defined as the NP-hard task of reasoning about topological relationships in the RCC-8 calculus, vital for robotic applications like fetch-and-delivery and object search [16]. However, AI's spatial reasoning is often limited by inadequate qualitative spatial representations, which lack the granularity needed for precise reasoning in dynamic environments [17].

The challenge of generating images that accurately reflect spatial relationships specified in natural language highlights the importance of spatial reasoning in visual tasks [18]. Tasks like spatial localization and view prediction are often treated separately, despite their inherent connections [19]. This complexity is further complicated by models analyzing human posture using non-intuitive sensor data [20]. In AI, spatial reasoning is critical for visual reasoning tasks, yet MLLMs continue to face challenges despite advancements in NLP [21]. Developing models that isolate and evaluate spatial relationship understanding is essential, as existing benchmarks often conflate spatial reasoning with other tasks [22]. Benchmarks for spatial reasoning in multi-agent environments, like those by Kulinski et al., are crucial for addressing these challenges [23].

Enhancing AI capabilities in fields like architecture, engineering, and medicine requires accurate interpretation of spatial expressions. Generative AI models like GPT-4 struggle with spatial transformations, but performance improves with supplementary information, such as augmented reality visualizations or mathematical representations. Advances in disentangling information extraction from reasoning processes show promise in enhancing spatial reasoning over textual data, addressing challenges posed by implicit spatial relationships. Integrating spatial intelligence with advanced reasoning techniques could lead to more effective AI applications across domains requiring spatial understanding [5, 6].

2.2 Large Language Models (LLMs)

LLMs are pivotal in AI, recognized for their ability to understand and generate human-like text. Their relevance to spatial reasoning is growing, as they frequently process natural language instructions requiring integration with visual perception [24]. However, LLMs face significant challenges in dynamic environments that require long-term path planning, with issues like spatial hallucination and context inconsistency hindering their effectiveness in complex scenarios [25]. The integration of LLMs into spatial reasoning is further complicated by the need for zero-shot reasoning capabilities, with many methods relying on resource-intensive fine-tuning [16]. This limitation is notable in tasks where LLMs must navigate and retrieve multiple target objects in unfamiliar environments, demanding high-level spatial intelligence [14].

The advent of MLLMs has expanded LLM applications beyond traditional text-based tasks, incorporating modalities like clinical decision support and medical imaging [26]. These advancements highlight the potential of LLMs to enhance spatial reasoning through diverse data integration, though challenges remain in effectively aligning these modalities.

2.3 Vision-Language Integration

Vision-language integration is a major advancement in AI, enabling the synthesis of visual and textual data to enhance comprehension and reasoning. This integration is crucial for tasks requiring a nuanced understanding of spatial relationships and dynamic environments, allowing AI systems to interpret and interact with the world similarly to human perception. Combining semantic knowledge from language models with heuristic-based planning significantly improves navigation capabilities in continuous 3D environments based on natural language instructions [12]. VLMs are central to this integration, providing robust multimodal information processing and contextual awareness essential for visual reasoning tasks [13].

Despite advancements, challenges persist in accurately representing spatial relationships in generated images. The inclusion of 3D data representations, like point clouds and NeRFs, with LLMs enhances the potential for 3D scene understanding, captioning, and dialogue [27]. Recent benchmarks like DriveMLLM provide comprehensive evaluation frameworks for both absolute and relative spatial reasoning tasks, improving MLLM assessment compared to prior benchmarks [28]. The SpatialPIN framework enhances VLMs' spatial reasoning by leveraging 3D priors from foundational models [29]. These benchmarks are crucial for evaluating and improving VLMs' spatial reasoning capabilities, which often struggle to capture relational information critical for understanding human cognition and language [30].

Methods like AT-ERL incorporate auxiliary tasks that improve spatial reasoning by predicting spatial metrics related to target objects [14]. The effective combination of visual and textual modalities, as demonstrated in video understanding and instruction-following capabilities, leverages LLM strengths to enhance performance in these domains [31].

2.4 Multimodal Learning and Communication

Multimodal learning is crucial in AI, enhancing spatial intelligence by integrating diverse data types, such as natural language commands and visual UI elements [32]. This integration improves spatial reasoning capabilities, enabling AI systems to process and interpret complex spatial information more effectively. The dual operators proposed by Aiguier et al. exemplify enhancements in logical reasoning within spatial contexts, contributing to a more nuanced understanding of spatial relations [33].

The PCA-Bench framework, introduced by Chen et al., formalizes multimodal decision-making within a partially observable Markov decision process, assessing models' ability to take correct actions based on multimodal observations and reasoning [34]. This framework highlights the critical role of integrating multiple data types in decision-making processes. Additionally, integrating text, images, and structured data in datasets, as emphasized by Wang et al., underscores the importance of multimodal learning in enhancing spatial intelligence [7].

The StarCraftImage dataset, discussed by Kulinski et al., illustrates the integration of various data types, including unit IDs, locations, and metadata, to facilitate multimodal learning in spatial reasoning [23]. This dataset serves as a comprehensive platform for evaluating and developing AI's spatial reasoning abilities. In healthcare, the integration of diverse data types is crucial for informed clinical decision-making, as explored by Niu et al., showcasing the broader applicability of multimodal learning beyond traditional AI domains [26].

The symbolic representation of human posture with terms like 'bent' and 'raised', proposed by Freedman et al., aligns closely with human language and understanding, highlighting the potential of multimodal learning to bridge machine interpretation and human cognition [20]. Moreover, using maximal tractable subsets of RCC-8 relations to enhance backtracking search algorithms, as described by Nebel, illustrates the application of multimodal learning in optimizing spatial reasoning processes [35].

The Valley model, proposed by Luo et al., exemplifies a multi-modal foundation capable of comprehending video, image, and language within a unified framework. This model demonstrates the potential of multimodal learning to process and integrate diverse data streams, enhancing AI's ability to perform complex spatial tasks [31]. Multimodal learning significantly advances AI systems' spatial intelligence, enabling more effective navigation and interaction within their environments.

2.5 Spatial Cognition

Spatial cognition involves the mental processes enabling individuals to perceive, interpret, and interact with spatial information. This cognitive ability is fundamental to both human and artificial systems, supporting navigation, understanding spatial layouts, and performing tasks requiring spatial reasoning. In humans, spatial cognition is linked to various intelligence aspects, such as integrating sensory information and making decisions based on spatial relationships [36]. Platforms like PlanetSense exemplify the application of spatial cognition in interpreting geospatial intelligence by combining real-time and historical data to analyze spatial patterns [37].

In artificial systems, spatial cognition is essential for developing models capable of understanding and manipulating spatial concepts. The limitations of DCNNs in learning basic spatial concepts highlight challenges in replicating human-like spatial reasoning in AI [38]. To address these challenges, benchmarks like SPACE evaluate AI models' spatial cognition capabilities, focusing on tasks requiring an understanding of spatial relationships [36].

Integrating geometric properties like direction, distance, and topology into knowledge graph embeddings significantly enhances spatial cognition within geographic contexts [15]. This integration allows AI systems to better interpret and respond to spatial queries involving spatial language and relationships [39]. Benchmarks like CityGPT emphasize the importance of evaluating LLMs on urban spatial tasks, requiring a nuanced understanding of urban layouts and mobility patterns [40].

The theory integrating qualitative characterizations of space and motion, proposed by Suchan et al., highlights the role of spatial schemas in enhancing spatial cognition [41]. These schemas facilitate spatial information interpretation, enabling informed decisions based on spatial reasoning. Moreover, map learning in environments with directional and recognition uncertainties, as discussed by Basye et al., underscores the need for more robust approaches to spatial cognition [42].

Visual thinking plays a critical role in spatial cognition, enhancing the explanatory power of spatial reasoning tasks. The distinction between verificational and inventional simplicity in proofs, as explored by Cain, illustrates the significance of visual thinking in understanding and communicating complex spatial concepts [43]. This capability is crucial for developing AI systems that can effectively interpret and generate spatially-informed actions.

In recent years, the exploration of spatial intelligence has gained significant traction within cognitive science and artificial intelligence research. Understanding how spatial reasoning operates in both

humans and machines is crucial for developing effective AI systems that can mimic human cognitive processes. As illustrated in Figure 2, this figure elucidates the hierarchical structure of spatial intelligence and cognition, categorizing various cognitive processes while modeling these functions in artificial systems. It emphasizes qualitative reasoning, outlines AI modeling benchmarks, and presents frameworks for spatial cognition. Furthermore, the figure delineates the challenges and advancements in aligning machine reasoning with human cognition, thereby providing a comprehensive overview of the current state of research in this field. This integration of visual data not only enhances the clarity of the concepts discussed but also reinforces the importance of understanding the interplay between human and machine spatial reasoning.

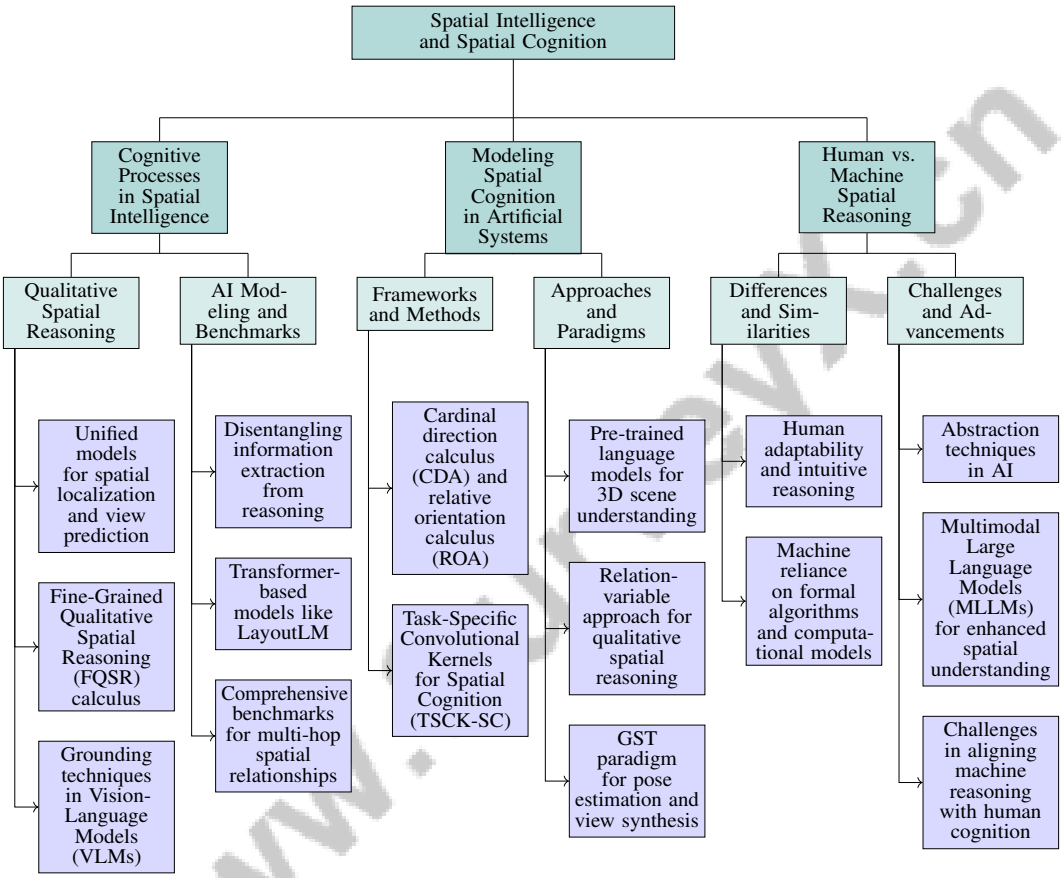


Figure 2: This figure illustrates the hierarchical structure of spatial intelligence and cognition, categorizing cognitive processes, modeling in artificial systems, and comparing human and machine spatial reasoning. It highlights qualitative reasoning, AI modeling benchmarks, frameworks for spatial cognition, and the challenges and advancements in aligning machine reasoning with human cognition.

3 Spatial Intelligence and Spatial Cognition

3.1 Cognitive Processes in Spatial Intelligence

Cognitive processes are fundamental to spatial intelligence, enabling both humans and AI to interpret and manipulate spatial information. These processes facilitate understanding spatial relationships and executing spatially-informed actions, as seen in unified models of spatial localization and view prediction inspired by human reasoning [19]. As illustrated in Figure 3, which highlights the key components of cognitive processes in spatial intelligence, these unified models are complemented by qualitative spatial reasoning, allowing AI to mimic human reasoning through advanced qualitative calculi. The Fine-Grained Qualitative Spatial Reasoning (FQSR) calculus enhances the complexity and efficiency of spatial relations processing [17].

Grounding techniques in Vision-Language Models (VLMs) significantly boost spatial reasoning performance. The compositional approach by Rajabi et al. highlights the importance of grounding in refining VLM capabilities [30]. These cognitive processes are essential for AI modeling, forming the basis for systems capable of complex spatial reasoning and interactions. Recent research emphasizes disentangling information extraction from reasoning to improve models' generalizability. Advancements in transformer-based models, such as LayoutLM, demonstrate their ability to ground natural language in spatial contexts, crucial for user interface applications. Comprehensive benchmarks reveal the strengths and limitations of language models in processing multi-hop spatial relationships and diverse viewpoint interpretations [32, 5, 44, 4]. Incorporating advanced qualitative reasoning and grounding methodologies enables AI systems to perform complex spatial tasks with improved accuracy and adaptability.

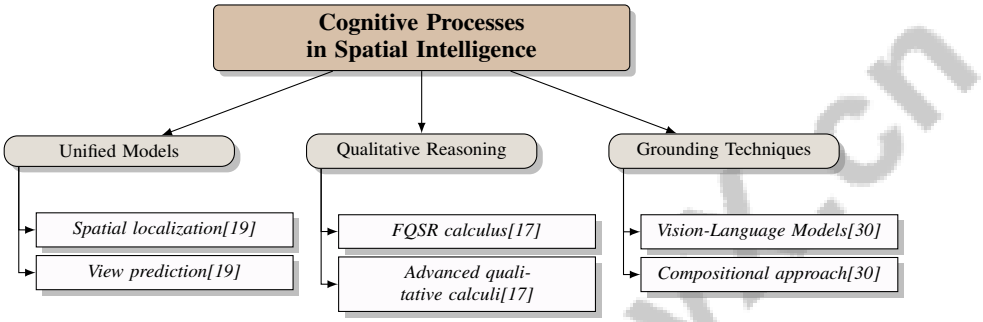


Figure 3: This figure illustrates the key components of cognitive processes in spatial intelligence, highlighting unified models for spatial localization and view prediction, qualitative reasoning with advanced calculi, and grounding techniques in Vision-Language Models.

3.2 Modeling Spatial Cognition in Artificial Systems

Modeling spatial cognition in artificial systems requires understanding human cognitive processes and replicating them computationally. Integrating cardinal direction calculus (CDA) and relative orientation calculus (ROA) through the cCOA method provides a robust framework for spatial relations reasoning [45]. This integration allows for a thorough representation of spatial relationships akin to human reasoning.

The development of Task-Specific Convolutional Kernels for Spatial Cognition (TSCK-SC) enhances DCNNs' ability to generalize spatial cognition tasks beyond training data, improving adaptability and performance [38]. Utilizing pre-trained language models for 3D scene understanding, as shown in SceneGPT, demonstrates the potential of language models for spatial cognition tasks without extensive 3D pre-training [46].

The relation-variable approach simplifies qualitative spatial reasoning integration by treating relations as variables, facilitating AI models to process spatial information similarly to human strategies [47]. The GST paradigm employs a unified training approach to optimize tasks like pose estimation and view synthesis concurrently, reflecting the interconnected nature of spatial cognition tasks in humans and enhancing AI's spatial task performance [19].

3.3 Human vs. Machine Spatial Reasoning

Human and machine spatial reasoning exhibit significant differences and similarities, reflecting the complexities of replicating human cognitive processes in AI. Human spatial reasoning is adaptable and intuitive, shaped by sensory experiences and cognitive development, enabling the interpretation of spatial cues and decision-making based on incomplete information [36]. In contrast, machine reasoning relies on formal algorithms and computational models for spatial information processing, excelling in data processing but struggling with nuanced understanding [17]. While frameworks like RCC-8 enhance AI's spatial reasoning, they lack the intuitive capabilities of humans [35].

Both humans and AI employ abstraction to simplify complex spatial problems. AI uses techniques like dimensionality reduction and hierarchical models, mirroring human abilities to distill essential spatial information [38]. Advancements in AI, including Multimodal Large Language Models (MLLMs),

integrate diverse data types to enhance spatial understanding, bridging gaps between human and machine reasoning [26].

Despite advancements, challenges remain in aligning machine reasoning with human cognition. Machines often lack seamless integration of sensory inputs and contextual knowledge, crucial for human-like reasoning. Developing AI systems that emulate holistic human reasoning is a major research area with implications for robotics, navigation, and human-computer interaction [14].

4 Large Language Models and Spatial Reasoning

4.1 Capabilities and Limitations of LLMs in Spatial Reasoning

Method Name	Spatial Challenges	Computational Complexity	Multimodal Integration
S2RCQL[25]	Spatial Hallucination	Context Inconsistency	Entity Relations
HCM-RCC8[35]	-	Computational Complexity	-
FQSR[17]	Coarse Representation	Computational Resources	-
CSCR[30]	Poor Grounding Ability	Ranking-based Method	Vision-and-language Models
AT-ERL[14]	Suboptimal Navigation Performance	Additional Training Time	Visual Navigation Integration
WoT[21]	Spatial Navigation Tasks	Significant Performance Gaps	Multimodal Capabilities Effectively
SpC-NAV[24]	Ineffective Navigation	Not Explicitly Mentioned	Visual Representations
VAL[31]	-	-	Visual And Textual

Table 1: This table presents a comparative analysis of various spatial reasoning methods, highlighting their specific challenges and capabilities. It examines the spatial challenges, computational complexity, and multimodal integration aspects of each method, providing insights into their effectiveness in addressing the limitations of Large Language Models (LLMs) in spatial reasoning tasks.

Large Language Models (LLMs) have advanced natural language processing but face challenges in spatial reasoning. They frequently misinterpret spatial relationships, resulting in navigation paths that overlook obstacles and lead to dead ends [25]. This issue is compounded by the computational complexity of current spatial reasoning methods, which struggle with large instances [35]. Incorporating frameworks like the Fine-Grained Qualitative Spatial Reasoning (FQSR) calculus can improve granularity and reasoning efficiency, aiding applications in robotics and spatial cognition [17]. However, Vision-Language Models (VLMs) often lack grounding abilities, impairing their capacity to accurately correlate images with text and understand spatial clauses [30]. Table 1 offers a comprehensive overview of the capabilities and limitations of different spatial reasoning methods, which are crucial for understanding the challenges faced by Large Language Models (LLMs) in spatial reasoning.

As illustrated in Figure 4, the challenges faced by LLMs in spatial reasoning are multifaceted, encompassing issues such as spatial misinterpretation and the inherent computational complexity of the tasks at hand. The figure also outlines potential improvement methods, including the FQSR calculus and auxiliary tasks, while highlighting promising models like Valley and SpC-NAV that demonstrate potential in overcoming these limitations. Reinforcement learning approaches also fall short in spatial reasoning and mapping, leading to suboptimal performance [14]. Multimodal Large Language Models (MLLMs) face difficulties in generating and processing visual representations, resulting in performance discrepancies [21]. Additionally, a lack of correspondence between visual elements and instructions hinders effective navigation, highlighting the need for better alignment between visual and textual data [24]. Despite these challenges, models like Valley, which excels in video understanding tasks, demonstrate the potential of LLMs in zero-shot and few-shot learning scenarios [31].

4.2 Refining Spatial Reasoning in Large Language Models

Enhancing spatial reasoning in LLMs requires innovative techniques for processing spatial data. The Spatial-Configuration-Based-Navigation (SpC-NAV) method improves spatial reasoning by modeling spatial semantics within instructions and aligning them with visual representations, aiding navigation in complex environments [24]. Disentangled Spatial Reasoning Models (DSRM) use a pipeline model to extract spatial information and apply symbolic reasoning, enhancing clarity and generalization [5]. Integrating 3D geometric features with open-vocabulary object detectors, as shown by the Structured Probabilistic Approach for Spatial Relationship Detection (SPASRD), further advances spatial reasoning in intricate environments [16].

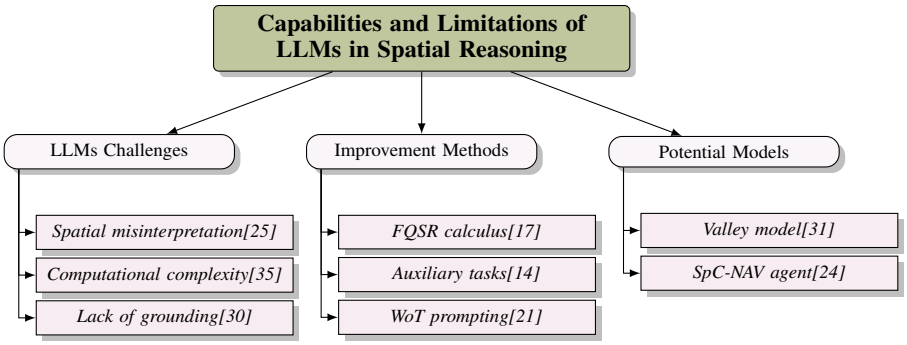


Figure 4: This figure illustrates the challenges faced by Large Language Models (LLMs) in spatial reasoning, highlighting key issues such as spatial misinterpretation and computational complexity. It also outlines potential improvement methods like the Fine-Grained Qualitative Spatial Reasoning (FQSR) calculus and auxiliary tasks, alongside promising models like Valley and SpC-NAV that demonstrate potential in overcoming these limitations.

Method Name	Methodology Approach	Spatial Data Processing	Application Context
SpC-NAV[24]	Spc-NAV	Spatial Configurations	Navigation Problem
DSRM[5]	Dsrm	Symbolic Spatial Reasoner	Spatial Question Answering
SPASRD[16]	Structured Probabilistic Approach	3D Geometric Cues	Robotic Perception Tasks
S2RCQL[25]	Spatial-to-Relational	Transforming Spatial Prompts	Complex Maze Environments
SP[29]	Progressive Prompting Interactions	3D Scene Representations	Robotics Tasks
CSCR[30]	Ranking-based Method	Grounding Module	Spatial Reasoning Tasks

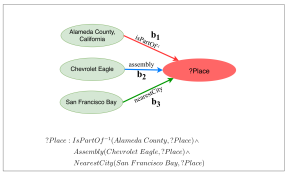
Table 2: Overview of various methodologies for refining spatial reasoning in large language models, highlighting their approach to spatial data processing and specific application contexts. The table includes methods such as SpC-NAV, DSRM, SPASRD, S2RCQL, SP, and CSCR, each contributing uniquely to the enhancement of spatial reasoning capabilities.

The Spatial-to-Relational Transformation and Curriculum Q-Learning (S2RCQL) method refines spatial reasoning by transforming spatial prompts into entity relations and using Q-learning to address context inconsistencies, improving path planning in dynamic settings [25]. The DriveMLLM benchmark is essential for identifying the limitations of current MLLMs, offering a comprehensive framework to evaluate spatial tasks in autonomous driving contexts [28]. The SpatialPIN framework introduces a zero-shot approach by combining prompting and interaction with 3D foundation models, enhancing spatial reasoning capabilities [29]. The Compositional Spatial Clause Ranking (CSCR) method decouples the grounding of subjects and objects from the prediction of spatial relationships, improving accuracy in spatial reasoning tasks [30]. Table 2 provides a comprehensive summary of different methodological approaches employed to enhance spatial reasoning in large language models, detailing their spatial data processing techniques and application contexts.

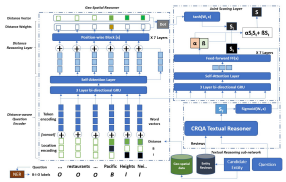
4.3 Expanding Datasets and Evaluation Techniques

Expanding datasets and refining evaluation techniques are crucial for improving the spatial reasoning capabilities of LLMs. These resources offer comprehensive frameworks for assessing and enhancing LLMs' performance in spatial reasoning tasks. The Visual Spatial Reasoning benchmark by Rajabi et al. emphasizes the importance of comparing proposed methods against baseline models using various metrics to evaluate effectiveness [30]. The evaluation of the SpC-NAV method used metrics like Navigation Error (NE), Success Rate (SR), and Success Rate weighted by normalized Path Length (SPL), highlighting the need for robust evaluation techniques that reflect real-world challenges faced by LLMs [24].

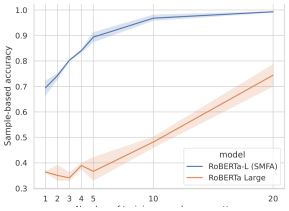
Integrating diverse datasets and evaluation frameworks is vital for advancing LLMs' spatial reasoning capabilities. These resources enable the development of sophisticated models capable of navigating and interpreting complex spatial environments. This advancement contributes to creating robust AI systems with enhanced spatial intelligence, allowing them to extract and reason about complex spatial information, comprehend nuanced spatial transformations, and engage with three-dimensional environments, especially when augmented by interactive visualization techniques like Augmented Reality [5, 6].



(a) A diagram illustrating relationships between entities[48]



(b) CRQA Textual Reasoner: A Deep Learning Approach for Geo-Spatial Reasoning[49]



(c) Comparison of Sample-Based Accuracy for RoBERTa-L (SMFA) and RoBERTa Large Models across Different Numbers of Training Epochs[50]

Figure 5: Examples of Expanding Datasets and Evaluation Techniques

As illustrated in Figure 5, the exploration of LLMs and their spatial reasoning capabilities is a burgeoning research area focused on enhancing datasets and evaluation techniques. The first example visualizes intricate relationships between entities, such as "Alameda County, California," "Chevrolet Eagle," and "San Francisco Bay," using directional arrows to denote connections like "isPartOf" and "assembly," emphasizing the importance of understanding spatial hierarchies within datasets. The second example features the CRQA Textual Reasoner, a deep learning model for geo-spatial reasoning, integrating multiple layers, including a Geo-Spatial Reasoner and a Joint Scoring Layer, to improve contextual understanding for complex question-answering tasks. Lastly, the third example compares sample-based accuracy between RoBERTa-L (SMFA) and RoBERTa Large models across various training epochs, highlighting the significance of optimizing training processes to enhance model performance. Collectively, these examples underscore ongoing efforts to expand datasets and refine evaluation techniques to better leverage the potential of LLMs in spatial reasoning tasks [48, 49, 50].

5 Vision-Language Integration

5.1 Methodologies for Vision-Language Integration

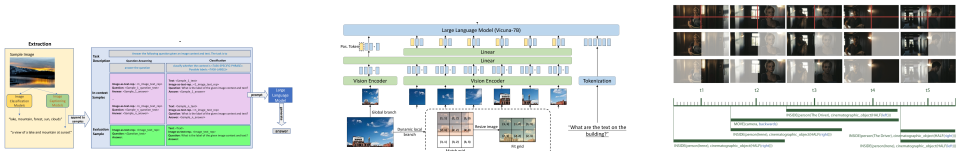
Vision-language integration in AI leverages advanced methodologies to enhance interpretative and interactive capabilities. Datasets like SpatialVLM enrich Vision-Language Models (VLMs) by training them on spatial reasoning questions from real-world images, thereby enhancing spatial reasoning [27]. The SpC-NAV methodology integrates spatial configurations with visual data, improving navigation decision-making [24]. RoboTool exemplifies synergy between language understanding and visual data by processing natural language instructions into executable robot control code [51].

ManipVQA fine-tunes Multimodal Large Language Models (MLLMs) with manipulation knowledge, underscoring domain-specific enhancements in visual-language interaction [52]. ViT3D integrates visual semantics with brain data through a unified network, facilitating deeper vision-language interaction understanding [53]. The SPRIGHT dataset, emphasizing spatial phrases in image captions, serves as a critical resource for training Text-to-Image (T2I) models, refining AI systems' spatially-informed visual content generation [18].

Figure 6 illustrates key methodologies in vision-language integration. Image captioning and question answering demonstrate the interplay of visual and textual data, processed by language models for coherent responses. The Vicuna-7B model's collaboration with a vision encoder enhances text-image retrieval through linguistic frameworks. Lastly, a cinematic scene illustrates dynamic visual storytelling and language interaction, where visual elements convey narrative interpretable through language [54, 55, 41].

5.2 Technologies Enhancing Spatial Reasoning

Innovative technologies enhance spatial reasoning through vision-language integration, enabling AI systems to interpret and interact with environments effectively. As depicted in Figure 7, the



(a) A diagram illustrating the process of image captioning and question answering[54]

(b) Large Language Model (Vicuna-7B) and Vision Encoder for Text-Image Retrieval[55]

(c) Scene from a film featuring a woman and a man in a dark room with red crosshairs indicating camera movements[41]

Figure 6: Examples of Methodologies for Vision-Language Integration

hierarchical organization of these technologies is categorized into three main areas: vision-language integration, probabilistic approaches, and 3D geometric information. This figure highlights key methodologies and datasets that play a crucial role in advancing spatial reasoning capabilities.

TOUCHDOWN uses real-world visual data to improve navigation and spatial understanding by integrating vision-language data [56]. The tag map method generates navigation plans from user task specifications by seamlessly integrating textual and visual information [57]. Neuro-symbolic training methods incorporate logical rules and symbolic reasoning to bridge symbolic logic and sensory data [58].

Algorithms combining qualitative spatial reasoning with probabilistic methods enhance navigation in sensory-deprived agents, highlighting the importance of integrating reasoning paradigms [59]. The MaCBench framework evaluates spatial reasoning in multimodal contexts, identifying AI models' strengths and weaknesses [12]. The POLY MATH dataset presents challenges requiring reasoning over visual and textual data, crucial for assessing AI cognitive capabilities in multimodal reasoning [60].

Structured probabilistic approaches use 3D geometric information for accurate spatial relationship representation, essential for detailed spatial reasoning [16]. SpatialPIN enhances spatial reasoning by leveraging rich 3D information from foundational models, allowing VLMs to generalize to unseen tasks [29]. GST's camera tokenization method models 2D projections and spatial perspectives simultaneously, improving spatial reasoning performance [19]. FQSR uses ternary relations for precise spatial analysis and decision-making [17].

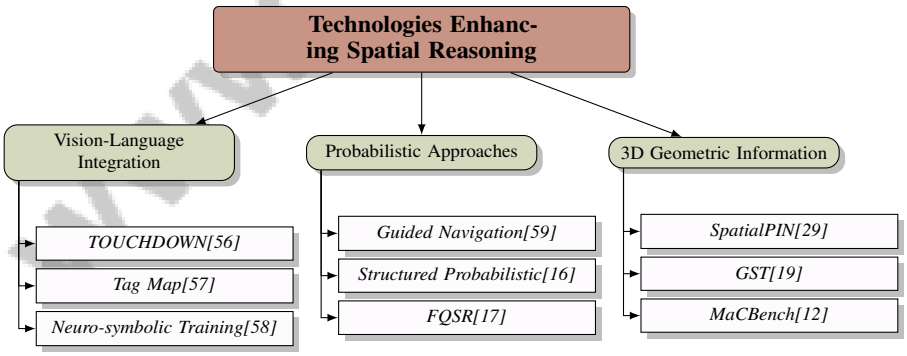


Figure 7: This figure illustrates the hierarchical organization of technologies enhancing spatial reasoning, categorized into vision-language integration, probabilistic approaches, and 3D geometric information, highlighting key methodologies and datasets.

5.3 Vision-Language Models in Navigation

Vision-language models significantly enhance navigation by integrating visual and textual data for improved spatial reasoning. The OPRAm method applies qualitative reasoning about relative directions to better interpret spatial relationships in navigation contexts [61]. The DCIP framework uses natural language instructions for navigation, demonstrating how linguistic and visual inputs enhance AI systems' understanding and execution of complex tasks [62].

Transforming navigation into a question-answering task, systems like VLMnav streamline navigation by selecting actions based on visual inputs and task prompts [63]. In practical applications, PIVOT demonstrates the versatility of vision-language models across various robotic navigation tasks [64]. The approach by Deguchi et al. generates accurate paths with minimal user input, enhancing navigation efficiency in environments requiring rapid decision-making [65]. Models trained with SAT data show significant performance improvements on dynamic spatial questions, highlighting the impact of targeted training on enhancing spatial reasoning [66].

6 Multimodal Learning

Multimodal learning is a cornerstone of artificial intelligence (AI), facilitating the integration of varied data types to enrich learning and comprehension. This section delves into the core concepts and significance of multimodal learning, emphasizing its transformative role in crafting AI systems proficient in processing complex information. By analyzing the interaction among different modalities, we can better understand how this approach enhances spatial intelligence and decision-making, with subsequent subsections detailing its applications and implications.

6.1 Concept and Importance of Multimodal Learning

Multimodal learning revolutionizes AI by integrating diverse data types or sensory inputs, thus improving understanding and decision-making. This approach is vital for developing AI that can interpret complex information from multiple sources, thereby enhancing spatial intelligence. For example, combining geo-coordinates with textual reviews demonstrates how multimodal learning enriches the understanding of spatial relationships [49]. The CLIPORT framework illustrates this by utilizing pre-trained semantic representations for spatial manipulation tasks [67].

Multimodal frameworks, such as those by Suchan et al., allow seamless integration of perceptual inputs to generate natural language descriptions, boosting AI interpretative abilities [41]. Hierarchical embodiment datasets like EMMA-X exemplify advanced strategies that enhance task planning and spatial reasoning in robotic control [68]. The Spatial-MM dataset, including subsets such as Spatial-Obj and Spatial-CoT, is invaluable for evaluating AI's spatial reasoning across various relationships [2].

The sMoRe framework shows practical applications of multimodal learning by enabling users to manipulate virtual objects with voice or text commands, reducing cognitive and physical effort in mixed reality environments [10]. This is further supported by the situational grounding from the VoxML modeling language, which encodes knowledge about objects and events, facilitating deeper understanding of multimodal interactions [69]. Additionally, integrating multiple perspectives from autonomous agents enhances guided navigation for sensory-deprived agents, highlighting the role of multimodal learning in spatial reasoning and navigation [59].

Li et al. introduced a dataset with multiple-choice questions based on top-view maps, aiding AI systems in comprehending spatial information [70]. The Spoken Dialogue Spatial Question Answering System (SQAS) employs multimodal learning to integrate computer vision and dialogue management, improving answer accuracy [71].

6.2 Technologies and Methodologies

Technological and methodological advancements in multimodal learning have greatly enhanced AI systems' ability to process and integrate diverse data types for improved spatial reasoning. Datasets from realistic 3D simulations, such as those from the ProcTHOR framework, provide a rich context for evaluating spatial reasoning in AI models [44]. In geospatial intelligence, platforms like PlanetSense support scalable big data architectures crucial for multimodal learning, enabling extensive spatial data integration and analysis [37].

The DHLM framework integrates deep learning with traditional statistical methods to model complex interactions and provide nuanced understandings of spatial relationships [72]. Datasets like those in DriveMLLM, which pair images with linguistically diverse natural language questions, challenge AI systems to perform spatial reasoning in dynamic contexts [28]. ViT3D combines 3D brain structures

with visual and linguistic inputs, representing a cutting-edge approach in multimodal learning that deepens spatial cognition understanding [53].

Rozanova et al. introduced a dataset with numerous commands paired with UI elements, foundational for multimodal learning in spatial reasoning tasks, emphasizing the grounding of natural language instructions in visual contexts [32]. Addressing LLMs' limitations, datasets with varying difficulty levels enable quantitative evaluations of LLM performance in spatial reasoning tasks [73]. Ahrens et al. describe datasets including images of abstract objects and tasks related to existence and relation prediction, providing a comprehensive framework for assessing AI's spatial reasoning capabilities [74].

Technologies and methodologies in multimodal learning, including 3D geometric information integration and advanced relation prediction modules, are essential for enhancing AI systems' spatial reasoning capabilities. This is evidenced by their impact on performance in vision and language tasks such as Text-based Visual Question Answering (TextVQA) and the Spatial-MM dataset, which reveals critical insights into LMMs' strengths and limitations in understanding complex spatial relationships [75, 2].

As illustrated in Figure 8, the hierarchical categorization of technologies and methodologies in multimodal learning highlights key frameworks, innovative datasets, and spatial reasoning benchmarks that enhance AI systems' spatial reasoning capabilities. By leveraging diverse datasets and innovative approaches, these systems achieve a more comprehensive understanding of their environments, leading to more effective interactions.

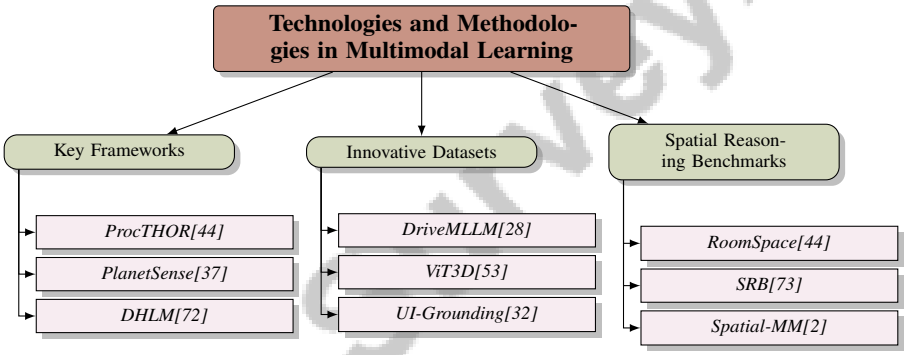


Figure 8: This figure illustrates the hierarchical categorization of technologies and methodologies in multimodal learning, highlighting key frameworks, innovative datasets, and spatial reasoning benchmarks that enhance AI systems' spatial reasoning capabilities.

6.3 Impact on Spatial Reasoning and Cognition

Multimodal learning significantly enhances spatial reasoning and cognitive processes by integrating diverse data types such as textual descriptions, visual representations, and real-time sensory inputs. The ROBOSPATIAL dataset exemplifies this impact, with models trained on it showing superior performance in spatial reasoning tasks compared to baseline models, indicating improved understanding of spatial relationships [76]. The 3DSR-Bench, including visual question-answer pairs related to height, location, orientation, and multi-object reasoning, offers a comprehensive framework for evaluating spatial reasoning capabilities with diverse entities [77].

Datasets like TOUCHDOWN, which integrate navigation instructions and spatial descriptions, improve learning outcomes in spatial reasoning, underscoring the importance of multimodal learning in cognitive processes related to navigation and spatial understanding [56]. The SpatialVLM framework provides a robust platform for Vision-Language Models (VLMs) to perform both qualitative and quantitative spatial reasoning tasks effectively, enhancing their applicability in real-world scenarios [27].

The PCA-Bench dataset, with 7,510 training examples, plays a crucial role in evaluating multimodal decision-making across domains such as autonomous driving, domestic robotics, and open-world gaming, demonstrating the broad applicability of multimodal learning in enhancing spatial reasoning and decision-making capabilities [34]. Auxiliary tasks, as discussed by Marza, significantly improve

agent performance in multi-object navigation tasks, highlighting the importance of spatial reasoning in complex environments [14].

The benchmark developed by Cohn et al. provides a structured framework for evaluating Large Language Models (LLMs) on qualitative spatial reasoning tasks, contributing to a deeper understanding of their reasoning capabilities and limitations [78]. Insights from Ishida on the impact of multimodal learning on spatial reasoning and cognitive processes highlight the transformative potential of integrating diverse data streams to enhance spatial intelligence and cognitive functions [79].

6.4 Advanced Multimodal Models and Instruction Understanding

Advanced multimodal models are pivotal for understanding and executing complex instructions by integrating diverse data types and leveraging sophisticated learning algorithms. These models synthesize information from multiple modalities, such as text, images, and sensory inputs, to enhance task comprehension and execution. The development of universal prompts for various tasks aligns with the capabilities of advanced multimodal models, facilitating the understanding of complex instructions across different contexts [80].

SceneGPT represents a significant advancement in multimodal learning, with future directions involving its extension to more complex tasks such as navigation and trajectory prediction. By improving object node prediction accuracy, SceneGPT aims to leverage robust multimodal language models to enhance spatial reasoning and instruction understanding [46]. This approach underscores the importance of integrating strong language models with visual data to achieve a deeper understanding of spatial environments and improve task execution.

There is an increasing need to explore additional spatial task categories and refine evaluation methods to enhance chatbot performance in geospatial contexts. Incorporating multimodal approaches is essential for improving chatbot capabilities, allowing for a more comprehensive understanding of spatial tasks and enhancing interactions between AI systems and users [81].

7 Applications and Case Studies

The practical applications of spatial intelligence are vast, particularly in robotics, navigation, and geographic information systems (GIS). This section examines these applications, starting with robotics, where spatial intelligence is crucial for autonomous navigation and interaction in complex environments.

7.1 Applications of Spatial Intelligence

Spatial intelligence is essential in robotics, navigation, and GIS, facilitating the understanding and manipulation of spatial relationships. In robotics, it supports autonomous navigation and object manipulation, as evidenced by the StarCraftImage benchmark, which aids in refining spatial reasoning algorithms for enhanced robot-environment interactions [23]. In navigation, spatial intelligence enables systems to plan and execute paths in dynamic environments. The GSRBench benchmark demonstrates advancements in navigation technologies, crucial for autonomous vehicle performance in complex terrains [22].

In GIS, spatial intelligence enhances GeoAI research, with methods like Geometric Feature-Enhanced Knowledge Graph Embedding (GF-KGE) improving link prediction accuracy for geographic entities, essential for sophisticated GIS applications [15]. This advancement aids spatial data analysis, supporting decision-making in urban planning and environmental monitoring. The mPLUG-Owl model exemplifies the potential of advanced multimodal models in enhancing spatial intelligence through improved instruction understanding, visual comprehension, and reasoning abilities, vital for applications integrating visual and textual data [82].

7.2 Robotic Navigation and Spatial Intelligence

Advanced robotic navigation systems rely on spatial intelligence to effectively interpret and interact with environments. Spatial reasoning enhances robots' navigational abilities, as seen in frameworks

like the RCC-8 calculus [35]. The GSRBench benchmark assesses robots' spatially-informed actions, improving navigational efficiency [22]. Geometric feature-enhanced knowledge graphs further support robots' link prediction in geographic contexts [15].

Multimodal learning integrates visual, textual, and sensory data, enabling robots to process complex spatial information and make informed navigation decisions. The mPLUG-Owl model exemplifies this potential, enhancing instruction understanding and navigational reasoning [82]. Robotic navigation systems utilize spatial intelligence for path planning, obstacle avoidance, and environment mapping, allowing adaptation to dynamic environments and precise task execution. As AI and robotics progress, incorporating spatial intelligence into navigation systems will significantly enhance autonomous capabilities. Current AI models like GPT-4 show promise in understanding spatial transformations, despite inherent limitations in spatial reasoning. Leveraging spatial intelligence alongside innovative navigation techniques fosters the development of adaptable robotic solutions, paving the way for autonomous systems capable of intricate tasks in diverse environments [6, 59, 83, 24, 73].

7.3 Multimodal Learning in Medical and General AI Models

Multimodal learning is increasingly important in AI model development, especially in medical and general applications. It integrates text, images, and structured data to enhance interpretative and decision-making capabilities. In the medical field, multimodal learning combines clinical data, medical imaging, and patient records to improve diagnostic accuracy and treatment planning. Multimodal Large Language Models (MLLMs) in clinical decision support systems illustrate how integrating textual and visual data enhances diagnostic accuracy and reliability [26].

In general AI, multimodal learning improves the processing and interpretation of complex information from diverse sources. By integrating multimodal data streams, AI systems achieve a comprehensive understanding of their environments, enhancing performance in natural language processing, image recognition, and decision-making. Advanced multimodal models like mPLUG-Owl demonstrate the benefits of combining visual and textual data for enhanced instruction understanding and reasoning [82]. This approach fosters robust and adaptable AI models capable of handling diverse tasks by harnessing complementary information. Notable advancements in MLLMs have begun to surpass human expert baselines in specific tasks while addressing data limitations and ethical challenges [55, 13, 3, 5, 26]. This capability is particularly beneficial in scenarios requiring AI models to interpret and respond to complex environments, such as autonomous driving and human-computer interaction.

The application of multimodal learning in medical and general AI models marks a substantial advancement in AI research. By integrating diverse data types, AI systems achieve a holistic understanding of environments, leading to improved decision-making processes. As AI evolves, multimodal learning approaches capable of processing and synthesizing various data types hold great promise for enhancing AI models across domains, particularly in healthcare, where they can improve clinical decision support and patient engagement by providing comprehensive insights into patient health [84, 26].

8 Future Directions and Challenges

8.1 Innovative Approaches and Methods

Advancing spatial intelligence and multimodal learning necessitates innovative methodologies that address existing limitations while leveraging emerging technologies. Future research should prioritize refining automated labeling techniques and expanding structured spatial reasoning methods in diverse robotic contexts [16]. Enhancements in inference speed and frameworks like SpatialPIN are crucial for complex robotics and interactive environments [29]. Symbolic representation research should focus on automating parameter settings and exploring qualitative descriptions of complex human motions [20], alongside optimizing heuristic combinations for broader qualitative spatial reasoning [35].

Further research should enhance auxiliary tasks in navigation scenarios to support advanced spatial reasoning applications [14]. Integrating audio inputs and expanding multilingual capabilities in models like Valley can increase versatility [31]. Scaling techniques such as GST are essential for complex scenarios involving multiple observations and real-world scenes [19]. Incorporating

3D information and leveraging larger datasets with accurate ground-truth annotations will be vital for advancing spatial reasoning capabilities [30]. Addressing fine-grained spatial semantics and challenges with novel objects in zero-shot settings presents innovative pathways for spatial intelligence advancement [24]. Optimizing the computational efficiency of the Fine-Grained Qualitative Spatial Reasoning (FQSR) calculus and its real-world applications will further contribute to this progress [17].

Recent research highlights innovative approaches that significantly enhance spatial intelligence and multimodal learning by addressing AI systems' limitations and leveraging advanced technologies. These include using augmented reality (AR) to improve generative AI's understanding of spatial transformations, employing multimodal models for geospatial applications, and introducing frameworks like Spatial Aptitude Training (SAT) to develop dynamic spatial reasoning capabilities. Techniques such as grid-based visual position encoding have shown promise in improving spatial localization accuracy in multimodal agents, fostering a deeper understanding of spatial relationships and contributing to the development of sophisticated AI systems for complex real-world challenges in architecture, urban planning, and robotics [6, 11, 66, 85, 5].

8.2 Benchmarks and Evaluation Frameworks

Benchmark	Size	Domain	Task Format	Metric
PPNL[86]	160,000	Path Planning	Path Planning	Success Rate, Optimal Rate
RCC-8[87]	392	Qualitative Spatial Reasoning	Compositional Reasoning	Accuracy, Correctness
TOPVIEWRS[70]	11,384	Spatial Reasoning	Multiple-choice Question Answering	EM, PM
GPT-4V[11]	94,986	Agricultural Science	Image Classification	Accuracy, F1-score
VL-Benchmark[54]	42,928	Visual Reasoning	Classification	Accuracy, Macro-F1
RoomSpace[44]	10,000	Spatial Reasoning	Find Relation (fr) And Yes/no (yn) Questions	Accuracy, F1-score
RadVUQA[88]	10,759	Radiology	Visual Question Answering	Response Accuracy, Multiple-Choice Accuracy
Spatial-MM[2]	2,310	Spatial Reasoning	Visual Question Answering	Accuracy, Exact-Match

Table 3: Table summarizing key benchmarks used for evaluating spatial reasoning and multimodal learning in AI models. The table includes information on benchmark size, domain, task format, and evaluation metrics, offering a comprehensive overview of the current landscape in AI evaluation frameworks.

Benchmarks and evaluation frameworks are essential for advancing spatial intelligence and multimodal learning, providing standardized metrics to assess AI models' capabilities. Table 3 provides a detailed overview of the benchmarks and evaluation frameworks utilized to assess the capabilities of AI models in spatial reasoning and multimodal learning tasks. These frameworks are crucial for evaluating large language models (LLMs) and other AI systems in spatial reasoning tasks, ensuring effective interpretation and interaction with complex environments. Metrics must evaluate both navigation success and path optimality to offer a comprehensive understanding of model capabilities [86].

However, current benchmarks often fail to capture the nuances of human spatial reasoning. LLMs may struggle with tasks requiring precise logical deductions, highlighting the need for more sophisticated evaluation frameworks that better reflect human cognitive processes [87]. Future improvements should enhance models' abilities to handle diverse input features, increase applicability across domains, and ensure benchmarks comprehensively cover a wide range of spatial reasoning challenges [41].

Evaluating new frameworks, such as the morpho-logic proposed by Aiguier et al., provides valuable insights into the effectiveness of various modal logics and reasoning frameworks, identifying strengths and weaknesses in current methodologies and guiding future research towards more effective evaluation techniques [89].

8.3 Multimodal Model Efficiency and Ethical Considerations

Developing multimodal models that integrate visual, textual, and spatial data significantly advances spatial reasoning capabilities but also presents efficiency challenges and ethical considerations.

Integrating complex qualitative spatial relations, as illustrated by the cardinal direction calculus, raises ethical concerns about potential misinterpretations in AI reasoning, potentially leading to unintended consequences in decision-making processes [45].

The efficiency of multimodal models often depends on high-quality datasets and substantial computational resources. For instance, the TopV-Nav method's reliance on top-view maps underscores the limitations posed by the quality and assumptions of spatial layouts, affecting model performance and generalizability [90]. Similarly, the SpaceNLI benchmark emphasizes the necessity for consistent evaluation metrics to ensure the efficiency and generalizability of Natural Language Inference (NLI) models across diverse spatial contexts [50].

Moreover, developing Vision-Language Models (VLMs) for visual spatial reasoning faces risks such as dataset dependency, computational demands for 3D reconstruction, and potential biases, raising ethical concerns about privacy and AI misuse [91]. Improved model architectures that better integrate visual and textual information are critical for enhancing spatial reasoning capabilities while addressing these ethical challenges [8].

Future research in multimodal model development should prioritize enhancing algorithm efficiency, particularly in integrating moving objects and adapting guided navigation methods for 3D tasks. This approach necessitates careful consideration of ethical implications in algorithm development, ensuring that AI systems are both efficient and ethically sound [59].

9 Conclusion

This survey underscores the pivotal role of integrating spatial intelligence with multimodal learning to advance AI's capabilities across various fields. The foundational nature of spatial reasoning is crucial for interpreting spatial relationships through complex reasoning tasks, as demonstrated by frameworks such as StepGame and TP-MANN, which enhance spatial reasoning in AI systems. The LFG framework exemplifies the necessity of robust spatial reasoning for achieving effective semantic goal-finding in robotic navigation. Furthermore, the potential of embodied multimodal action models like EMMA-X is highlighted, showcasing improvements in robotic task execution, particularly in scenarios requiring extended spatial reasoning.

Nonetheless, significant challenges remain, as Multimodal Large Language Models (MLLMs) continue to fall short of human-like performance in intricate spatial reasoning tasks, indicating the need for ongoing research and development. The integration of spatial intelligence aids in streamlining problem-solving, as evidenced by its application in simplifying proofs. The STUPD dataset illustrates how incorporating dynamic spatial and temporal relations can significantly enhance model performance in visual reasoning.

Moreover, this survey emphasizes the importance of merging spatial intelligence with multimodal learning to enhance AI's ability to process and interpret complex information. The dual logical operators framework, based on mathematical morphology, offers a promising avenue for enriching the expressiveness of spatial reasoning models. The substantial representation of geospatial data within the Common Crawl corpus suggests that LLMs can leverage this data to exhibit spatial reasoning capabilities. Spartun3D further demonstrates the potential of advanced datasets and alignment modules to improve spatial reasoning within 3D-based LLMs, indicating promising directions for future research and applications.

References

[1] Manasi Sharma. Exploring and improving the spatial reasoning abilities of large language models, 2023.

[2] Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Gholamreza Haffari, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models, 2024.

[3] Zichen Zhu, Yang Xu, Lu Chen, Jingkai Yang, Yichuan Ma, Yiming Sun, Hailin Wen, Jiaqi Liu, Jinyu Cai, Yingzi Ma, Situo Zhang, Zihan Zhao, Liangtai Sun, and Kai Yu. Multi: Multimodal understanding leaderboard with text and images, 2025.

[4] Thomas Greatrix, Roger Whitaker, Liam Turner, and Walter Colombo. Can large language models create new knowledge for spatial reasoning tasks?, 2024.

[5] Roshanak Mirzaee and Parisa Kordjamshidi. Disentangling extraction and reasoning in multi-hop spatial reasoning, 2023.

[6] Uttamasha Monjoree and Wei Yan. Ai's spatial intelligence: Evaluating ai's understanding of spatial transformations in psvt:r and augmented reality, 2024.

[7] Kevin Wang, Junbo Li, Neel P. Bhatt, Yihan Xi, Qiang Liu, Ufuk Topcu, and Zhangyang Wang. On the planning abilities of openai's o1 models: Feasibility, optimality, and generalizability, 2024.

[8] Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models, 2024.

[9] Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S. Ryoo, and Tsung-Yu Lin. Learning to localize objects improves spatial reasoning in visual-llms, 2024.

[10] Yunhao Xing, Que Liu, Jingwu Wang, and Diego Gomez-Zara. smore: Enhancing object manipulation and organization in mixed reality spaces with llms and generative ai, 2024.

[11] Chenjiao Tan, Qian Cao, Yiwei Li, Jielu Zhang, Xiao Yang, Huaqin Zhao, Zihao Wu, Zhengliang Liu, Hao Yang, Nemin Wu, Tao Tang, Xinyue Ye, Lilong Chai, Ninghao Liu, Changying Li, Lan Mu, Tianming Liu, and Gengchen Mai. On the promises and challenges of multimodal foundation models for geographical, environmental, agricultural, and urban planning applications, 2023.

[12] Nawaf Alampara, Mara Schilling-Wilhelmi, Martiño Ríos-García, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, N. M. Anoop Krishnan, and Kevin Maik Jablonka. Probing the limitations of multimodal language models for chemistry and materials research, 2024.

[13] Bingshuai Liu, Chenyang Lyu, Zijun Min, Zhanyu Wang, Jinsong Su, and Longyue Wang. Retrieval-augmented multi-modal chain-of-thoughts reasoning for large language models, 2024.

[14] Pierre Marza, Laetitia Matignon, Olivier Simonin, and Christian Wolf. Teaching agents how to map: Spatial reasoning for multi-object navigation, 2023.

[15] Lei Hu, Wenwen Li, and Yunqiang Zhu. Geometric feature enhanced knowledge graph embedding and spatial reasoning, 2024.

[16] Negar Nejatishahidin, Madhukar Reddy Vongala, and Jana Kosecka. Structured spatial reasoning with open vocabulary object detectors, 2024.

[17] Sören Schwertfeger. Fine-grained qualitative spatial reasoning about point positions, 2019.

[18] Agneet Chatterjee, Gabriela Ben Melech Stan, Estelle Aflalo, Sayak Paul, Dhruba Ghosh, Tejas Gokhale, Ludwig Schmidt, Hannaneh Hajishirzi, Vasudev Lal, Chitta Baral, and Yezhou Yang. Getting it right: Improving spatial consistency in text-to-image models, 2024.

[19] Junyi Chen, Di Huang, Weicai Ye, Wanli Ouyang, and Tong He. Where am i and what will i see: An auto-regressive model for spatial localization and view prediction, 2024.

[20] Richard G. Freedman, Joseph B. Mueller, Jack Ladwig, Steven Johnston, David McDonald, Helen Wauck, Ruta Wheelock, and Hayley Borck. A symbolic representation of human posture for interpretable learning and reasoning, 2022.

[21] Sachit Menon, Richard Zemel, and Carl Vondrick. Whiteboard-of-thought: Thinking step-by-step across modalities, 2024.

[22] Navid Rajabi and Jana Kosecka. Gsr-bench: A benchmark for grounded spatial reasoning evaluation via multimodal llms, 2024.

[23] Sean Kulinski, Nicholas R. Waytowich, James Z. Hare, and David I. Inouye. Starcraftimage: A dataset for prototyping spatial reasoning methods for multi-agent environments, 2024.

[24] Yue Zhang, Quan Guo, and Parisa Kordjamshidi. Towards navigation by reasoning over spatial configurations, 2021.

[25] Hourui Deng, Hongjie Zhang, Jie Ou, and Chaosheng Feng. Can llm be a good path planner based on prompt engineering? mitigating the hallucination for path planning, 2024.

[26] Qian Niu, Keyu Chen, Ming Li, Pohsun Feng, Ziqian Bi, Lawrence KQ Yan, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Junyu Liu, Benji Peng, Tianyang Wang, Yunze Wang, Silin Chen, and Ming Liu. From text to multimodality: Exploring the evolution and impact of large language models in medical practice, 2024.

[27] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities, 2024.

[28] Xianda Guo, Ruijun Zhang, Yiqun Duan, Yuhang He, Chenming Zhang, Shuai Liu, and Long Chen. Drivemllm: A benchmark for spatial understanding with multimodal large language models in autonomous driving, 2024.

[29] Chenyang Ma, Kai Lu, Ta-Ying Cheng, Niki Trigoni, and Andrew Markham. Spatialpin: Enhancing spatial reasoning capabilities of vision-language models through prompting and interacting 3d priors, 2024.

[30] Navid Rajabi and Jana Kosecka. Towards grounded visual spatial reasoning in multi-modal vision language models, 2024.

[31] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Da Li, Pengcheng Lu, Tao Wang, Linmei Hu, Minghui Qiu, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability, 2023.

[32] Julia Rozanova, Deborah Ferreira, Krishna Dubba, Weiwei Cheng, Dell Zhang, and Andre Freitas. Grounding natural language instructions: Can large language models capture spatial information?, 2021.

[33] Marc Aiguier and Isabelle Bloch. Dual logic concepts based on mathematical morphology in stratified institutions: Applications to spatial reasoning, 2017.

[34] Liang Chen, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Peiyi Wang, Xiangdi Meng, Tianyu Liu, and Baobao Chang. Pca-bench: Evaluating multimodal large language models in perception-cognition-action chain, 2024.

[35] B. Nebel and J. Renz. Efficient methods for qualitative spatial reasoning, 2011.

[36] Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Kraehenbuehl, and Vladlen Koltun. Does spatial cognition emerge in frontier models?, 2024.

[37] Gautam S. Thakur, Budhendra L. Bhaduri, Jesse O. Piburn, Kelly M. Sims, Robert N. Stewart, and Marie L. Urban. Planetsense: A real-time streaming and spatio-temporal analytics platform for gathering geo-spatial intelligence from open source data, 2015.

[38] Xi Zhang, Xiaolin Wu, and Jun Du. Challenge of spatial cognition for deep learning, 2020.

[39] Alexander Kuhnle and Ann Copestake. What is needed for simple spatial language capabilities in vqa?, 2019.

[40] Jie Feng, Yuwei Du, Tianhui Liu, Siqi Guo, Yuming Lin, and Yong Li. Citygpt: Empowering urban spatial cognition of large language models, 2024.

[41] Jakob Suchan, Mehul Bhatt, and Harshita Jhavar. Talking about the moving image: A declarative model for image schema based embodied perception grounding and language generation, 2015.

[42] Kenneth Basye and Thomas L. Dean. Map learning with indistinguishable locations, 2013.

[43] Alan J. Cain. Visual thinking and simplicity in proof, 2018.

[44] Fangjun Li, David C. Hogg, and Anthony G. Cohn. Reframing spatial reasoning evaluation in language models: A real-world simulation benchmark for qualitative reasoning, 2024.

[45] Amar Isli. Integrating cardinal direction relations and other orientation relations in qualitative spatial reasoning, 2004.

[46] Shivam Chandhok. Scenegpt: A language model for 3d scene understanding, 2024.

[47] Sebastian Brand. Relation variables in qualitative spatial reasoning, 2006.

[48] Gengchen Mai, Krzysztof Janowicz, Ling Cai, Rui Zhu, Blake Regalia, Bo Yan, Meilin Shi, and Ni Lao. Se-kge: A location-aware knowledge graph embedding model for geographic question answering and spatial semantic lifting, 2020.

[49] Danish Contractor, Shashank Goel, Mausam, and Parag Singla. Joint spatio-textual reasoning for answering tourism questions, 2020.

[50] Lasha Abzianidze, Joost Zwarts, and Yoad Winter. Spacenli: Evaluating the consistency of predicting inferences in space, 2023.

[51] Mengdi Xu, Peide Huang, Wenhao Yu, Shiqi Liu, Xilun Zhang, Yaru Niu, Tingnan Zhang, Fei Xia, Jie Tan, and Ding Zhao. Creative robot tool use with large language models, 2023.

[52] Siyuan Huang, Iaroslav Ponomarenko, Zhengkai Jiang, Xiaoqi Li, Xiaobin Hu, Peng Gao, Hongsheng Li, and Hao Dong. Manipvqa: Injecting robotic affordance and physically grounded information into multi-modal large language models, 2024.

[53] Guobin Shen, Dongcheng Zhao, Xiang He, Linghao Feng, Yiting Dong, Jihang Wang, Qian Zhang, and Yi Zeng. Neuro-vision to language: Enhancing brain recording-based visual reconstruction and language interaction, 2024.

[54] Sherzod Hakimov and David Schlangen. Images in language space: Exploring the suitability of large language models for vision language tasks, 2023.

[55] Yonghui Wang, Wengang Zhou, Hao Feng, and Houqiang Li. Adaptvision: Dynamic input scaling in mllms for versatile scene understanding, 2024.

[56] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments, 2020.

[57] Mike Zhang, Kaixian Qu, Vaishakh Patil, Cesar Cadena, and Marco Hutter. Tag map: A text-based map for spatial reasoning and navigation with large language models, 2024.

[58] Tanawan Premsri and Parisa Kordjamshidi. Neuro-symbolic training for reasoning over spatial language, 2025.

[59] Danilo Perico, Paulo E. Santos, and Reinaldo Bianchi. Guided navigation from multiple viewpoints using qualitative spatial reasoning, 2020.

[60] Himanshu Gupta, Shreyas Verma, Ujjwala Anantheswaran, Kevin Scaria, Mihir Parmar, Swaroop Mishra, and Chitta Baral. Polymath: A challenging multi-modal mathematical reasoning benchmark, 2024.

[61] Till Mossakowski and Reinhard Moratz. Qualitative reasoning about relative direction on adjustable levels of granularity, 2010.

[62] Pranav Doma, Aliasghar Arab, and Xuesu Xiao. Llm-enhanced path planning: Safe and efficient autonomous navigation with instructional inputs, 2024.

[63] Dylan Goetting, Himanshu Gaurav Singh, and Antonio Loquercio. End-to-end navigation with vision language models: Transforming spatial reasoning into question-answering, 2024.

[64] Soroush Nasiriany, Fei Xia, Wenhao Yu, Ted Xiao, Jacky Liang, Ishita Dasgupta, Annie Xie, Danny Driess, Ayzaan Wahid, Zhuo Xu, Quan Vuong, Tingnan Zhang, Tsang-Wei Edward Lee, Kuang-Huei Lee, Peng Xu, Sean Kirmani, Yuke Zhu, Andy Zeng, Karol Hausman, Nicolas Heess, Chelsea Finn, Sergey Levine, and Brian Ichter. Pivot: Iterative visual prompting elicits actionable knowledge for vlms, 2024.

[65] Hideki Deguchi, Kazuki Shibata, and Shun Taguchi. Language to map: Topological map generation from natural language path instructions, 2024.

[66] Arijit Ray, Jiafei Duan, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. Sat: Spatial aptitude training for multimodal language models, 2024.

[67] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation, 2021.

[68] Qi Sun, Pengfei Hong, Tej Deep Pala, Vernon Toh, U-Xuan Tan, Deepanway Ghosal, and Soujanya Poria. Emma-x: An embodied multimodal action model with grounded chain of thought and look-ahead spatial reasoning, 2024.

[69] James Pustejovsky and Nikhil Krishnaswamy. Situational grounding within multimodal simulations, 2019.

[70] Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić. Topviewrs: Vision-language models as top-view spatial reasoners, 2024.

[71] Georgiy Platonov, Benjamin Kane, Aaron Gindi, and Lenhart K. Schubert. A spoken dialogue system for spatial question answering in a physical blocks world, 2019.

[72] Senqiao Yang, Jiaming Liu, Ray Zhang, Mingjie Pan, Zoey Guo, Xiaoqi Li, Zehui Chen, Peng Gao, Yandong Guo, and Shanghang Zhang. Lidar-llm: Exploring the potential of large language models for 3d lidar understanding, 2023.

[73] He Yan, Xinyao Hu, Xiangpeng Wan, Chengyu Huang, Kai Zou, and Shiqi Xu. Inherent limitations of llms regarding spatial information, 2023.

[74] Kyra Ahrens, Matthias Kerzel, Jae Hee Lee, Cornelius Weber, and Stefan Wermter. Knowing earlier what right means to you: A comprehensive vqa dataset for grounding relative directions via multi-task learning, 2022.

[75] Hao Li, Jinfa Huang, Peng Jin, Guoli Song, Qi Wu, and Jie Chen. Toward 3d spatial reasoning for human-like text-based visual question answering, 2023.

[76] Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics, 2024.

[77] Wufei Ma, Haoyu Chen, Guofeng Zhang, Celso M de Melo, Alan Yuille, and Jieneng Chen. 3dsrbench: A comprehensive 3d spatial reasoning benchmark, 2024.

[78] Anthony G Cohn and Robert E Blackwell. Can large language models reason about the region connection calculus?, 2024.

[79] Shu Ishida. Spatial reasoning and planning for deep embodied agents, 2024.

[80] Zhenfang Chen, Rui Sun, Wenjun Liu, Yining Hong, and Chuang Gan. Genome: Generative neuro-symbolic visual reasoning by growing and reusing modules, 2023.

[81] Hartwig H. Hochmair, Levente Juhasz, and Takoda Kemp. Correctness comparison of chatgpt-4, gemini, claude-3, and copilot for spatial tasks, 2024.

[82] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[83] Stanislav Kikot. Spatial intelligence of a self-driving car and rule-based decision making, 2023.

[84] Soyeon Caren Han, Feiqi Cao, Josiah Poon, and Roberto Navigli. Multimodal large language models and tunings: Vision, language, sensors, audio, and beyond, 2024.

[85] Joongwon Chae, Zhenyu Wang, Lian Zhang, Dongmei Yu, and Peiwu Qin. Grid-augmented vision: A simple yet effective approach for enhanced spatial understanding in multi-modal agents, 2024.

[86] Mohamed Aghzal, Erion Plaku, and Ziyu Yao. Can large language models be good path planners? a benchmark and investigation on spatial-temporal reasoning, 2025.

[87] Anthony G Cohn. An evaluation of chatgpt-4's qualitative spatial reasoning capabilities in rcc-8, 2023.

[88] Yang Nan, Huichi Zhou, Xiaodan Xing, and Guang Yang. Beyond the hype: A dispassionate look at vision-language models in medical scenario, 2024.

[89] Marc Aiguier, Isabelle Bloch, Salim Nibouche, and Ramon Pino Perez. Morpho-logic from a topos perspective: Application to symbolic ai, 2023.

[90] Linqing Zhong, Chen Gao, Zihan Ding, Yue Liao, and Si Liu. Topv-nav: Unlocking the top-view spatial reasoning potential of mllm for zero-shot object navigation, 2024.

[91] Zaiqiao Meng, Hao Zhou, and Yifang Chen. I know about "up"! enhancing spatial reasoning in visual language models through 3d reconstruction, 2024.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.