A Survey of Vision-and-Language Navigation and Related AI Techniques

www.surveyx.cn

Abstract

This survey paper provides a comprehensive examination of Vision-and-Language Navigation (VLN), emphasizing the integration of visual and linguistic data to optimize navigation in dynamic environments. The survey explores advancements in VLN by analyzing training paradigms, evaluation methods, and architectural insights within Vision-Language Models (VLMs), addressing challenges such as generalization in real-world scenarios and the integrity of object information. It systematically compares Vision-Language Architecture (VLA) designs, evaluates semantic-aware techniques for data integration, and discusses innovative architectures that enhance reasoning capabilities. The survey highlights the role of foundational AI concepts, including LSTM and Transformer architectures, in improving navigation performance. It also delves into the integration of Reinforcement Learning (RL) and Imitation Learning (IL) for training agents, actor-critic algorithms like A2C for policy optimization, and the significance of semantic mapping in embodied AI. The paper identifies potential research areas, such as improving Scene Memory Transformers, expanding knowledge-based frameworks, and enhancing training datasets to overcome current limitations. By providing a structured overview of current advancements and future directions, this survey serves as a valuable resource for researchers aiming to develop more sophisticated and reliable navigation systems capable of operating in complex real-world scenarios.

1 Introduction

1.1 Significance of Vision-and-Language Navigation

Vision-and-language navigation (VLN) represents a pivotal area within embodied AI, focusing on the amalgamation of visual and linguistic data to facilitate agents' navigation in complex 3D environments based on natural language instructions. This integration is crucial for developing agents capable of interpreting and acting upon ambiguous language descriptions [1]. Despite notable progress, existing methodologies often fail to fully leverage the alignment between visual and textual sequences, hindering performance [2].

The capacity of VLN agents to adapt to varied environments while executing user instructions in real-time underscores the importance of this integration [3]. Bridging the gap between Success Rate (SR) and Oracle Success Rate (OSR) is vital for enhancing navigation efficacy, an aspect that has received insufficient attention in prior research [4]. The implementation of a hint generator that offers detailed visual descriptions can further enhance navigation performance by providing contextual information to the agent [5].

The importance of merging vision and language extends to examining how different architectures, such as RNNs and Transformers, affect the representational properties of models trained on multimodal data [6]. This integration not only boosts the adaptability of AI systems across various settings but also advances the development of sophisticated agents capable of navigating and interacting

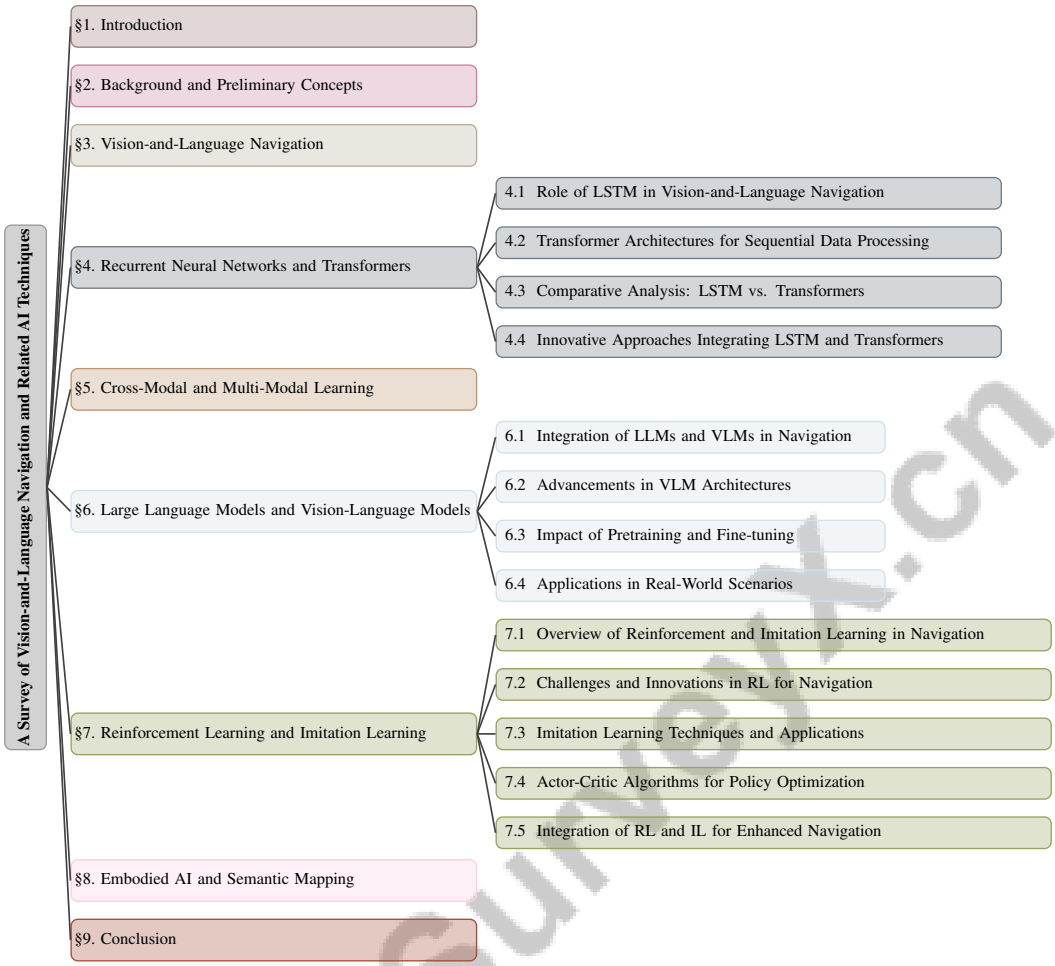

Figure 1: chapter structure

with real-world environments [7]. Addressing spatial route priors and object relationships, often overlooked, is essential for the advancement of robust VLN systems.

The proposed framework demonstrates significant improvements over the state-of-the-art VLN-CE baseline in real-world scenarios, showcasing the potential of utilizing foundation models for effective navigation without prior training in simulation [8]. The intersection of vision and language in navigation tasks is crucial for advancing AI technologies, fostering innovative applications in autonomous systems, and enhancing the generalization capabilities of AI agents in real-world contexts.

1.2 Scope and Objectives of the Survey

This survey encompasses a thorough exploration of Vision-and-Language Navigation (VLN), emphasizing the synthesis of visual and linguistic data to optimize navigation tasks, particularly in dynamically evolving environments [9]. The primary objective is to investigate advancements in VLN by analyzing various training paradigms, evaluation methods, and architectural insights within Vision-Language Models (VLMs), while addressing the challenges of generalization in real-world scenarios [10]. This includes a critical assessment of existing VLN models, which often neglect the integrity of object information or allow irrelevant visual inputs during navigation, thereby affecting overall performance [2].

The survey aims to enhance the integration of visual and linguistic data for navigation tasks, with a focus on generating precise and detailed instructions [5]. By systematically comparing different Vision-Language Architecture (VLA) designs and their performance in various environments, the

survey seeks to establish a robust framework for evaluating diverse design choices and training methodologies. Additionally, it explores semantic-aware techniques for extracting embeddings from mono-modal, multi-modal, and cross-modal data, particularly concerning the temporal aspects of time-series data [11].

Moreover, the survey addresses the challenges of instance-level and attribute-level navigation, aiming to enhance precision in navigation tasks. It incorporates the development of novel architectures that integrate frozen pretrained vision and language backbones to bolster reasoning capabilities across multiple dimensions, including mathematics and logic [12]. This is complemented by the introduction of Multimodal Instruction Navigation with demonstration Tours (MINT), which leverages recorded demonstration videos to improve robot navigation capabilities [13].

Through these objectives and scope, the survey aims to provide a comprehensive overview of current advancements in vision-and-language navigation, identify potential areas for future research, and serve as a valuable resource for researchers to analyze navigation instructions through high-level annotations of navigation concepts. The proposed generative language-grounded policy (GLGP) further emphasizes the need for innovative approaches to enhance performance in unseen environments [14]. Additionally, the integration of a speaker model that generates instructions and a follower model that executes them is examined to improve the navigation process, enabling agents to better comprehend and act on complex instructions.

1.3 Structure of the Survey

This survey is meticulously structured to provide a comprehensive examination of Vision-and-Language Navigation (VLN) and its related AI techniques. The paper commences with an **Introduction** section that elucidates the significance of VLN, its scope, and objectives, setting the stage for further exploration. Following this, the **Background and Preliminary Concepts** section delves into foundational concepts essential for understanding VLN, including discussions on recurrent neural networks and transformer architectures, as well as cross-modal learning techniques.

The core of the survey is divided into focused sections. The **Vision-and-Language Navigation** section investigates the integration of visual and linguistic data, highlighting challenges and recent advancements in navigation techniques. The **Recurrent Neural Networks and Transformers** section examines the roles of LSTM and Transformer architectures, featuring a comparative analysis and innovative approaches that combine both architectures for enhanced navigation.

The survey further explores , emphasizing the integration of diverse data sources, such as text and images, to improve navigation performance. It discusses how attention mechanisms facilitate the interaction between linguistic and visual information, enhancing the effectiveness of navigation tasks in VLN scenarios. By examining the information flow in multimodal large language models and the role of explicit spatial representations, the survey illustrates how these approaches can significantly elevate the performance of navigation agents, addressing existing ambiguities and limitations in traditional methods [15, 16, 17, 18, 19]. This is followed by an analysis of **Large Language Models and Vision-Language Models**, focusing on their impact on navigation tasks and recent architectural advancements.

In the **Reinforcement Learning and Imitation Learning** section, the survey explores strategies for training navigation agents, including actor-critic algorithms and the integration of RL and IL for enhanced performance. The penultimate section on **Embodied AI and Semantic Mapping** discusses AI systems' interactions with the physical world, emphasizing semantic mapping techniques and their incorporation into navigation systems.

Finally, the **Conclusion** section summarizes key findings and insights, discussing future directions and potential research opportunities in the field. This structured approach ensures a thorough exploration of vision-and-language navigation, providing valuable insights and resources for researchers and practitioners in the domain.The following sections are organized as shown in Figure 1.

2 Background and Preliminary Concepts

2.1 Foundational Concepts in Vision-and-Language Navigation

Vision-and-Language Navigation (VLN) leverages AI techniques to enable agents to interpret natural language instructions for navigation. Central to this is the use of Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, which excel in handling sequential data and retaining historical context, crucial for informed decision-making in navigation tasks. This enhances semantic parsing of user commands, improving robotic interactions [20].

Transformer architectures, known for parallel processing and self-attention, effectively capture contextual relationships, addressing cross-modal alignment and feature localization challenges in VLN [21]. The Scene Memory Transformer (SMT) exemplifies this by optimizing memory use in navigation, paralleling natural language processing advancements [21].

Cross-modal learning is pivotal in VLN, integrating visual and linguistic data to exploit complementary information. This enables effective navigation by unifying visual and textual representations, and generating natural language instructions from navigational demonstrations, where the Speaker agent creates coherent instructions from action sequences and visual observations [22].

Adapting large language models (LLMs) to diverse input modalities in resource-constrained environments is vital for embodied AI applications. Despite Vision-Language Models' (VLMs) success in 2D image understanding, precise spatial comprehension remains challenging [7]. The visual domain gap between simulations and real-world environments further complicates policy transferability [8].

Recent studies on VLN provide a robust framework for developing agents capable of interpreting and executing navigation tasks in real-world settings. These concepts highlight the necessity of grounding instructions in visual contexts, leveraging knowledge for enhanced reasoning, and employing LLMs to bolster navigational understanding. By integrating skill-specific competencies and pre-training methodologies, these frameworks foster agents that adapt to dynamic scenarios, improving performance in complex navigation tasks [23, 24, 25, 26, 13]. The synthesis of visual and linguistic data enhances navigation capabilities, advancing VLN research and applications.

2.2 Datasets and Evaluation Metrics

Benchmark	Size	Domain	Task Format	Metric
R2R[27]	21,567	Navigation	Natural Language Instruction Following	Navigation Error
R4R[28]	233,613	Navigation	Path Following	CLS
VLN-CE[29]	5,611	Navigation	Instruction Following	Success Rate, Success Weighted by Inverse Path Length
R2R-UNO[30]	229,982	Navigation	Pathfinding	Success Rate, Trajectory Length
RxR[31]	9,800,000	Vision-and-Language Navigation	Navigation Instructions	NDTW, SDTW
ScaleVLN[32]	4,900,000	Visual Navigation	Instruction-Trajectory Pair	Success Rate, Navigation Error
VLN[33]	9,326	Navigation	Instruction-Path Matching	AUC
M-Gib[34]	4,200,000	Navigation	Instruction Following	NDTW, SR

Table 1: This table presents a comprehensive overview of key benchmarks utilized in Vision-and-Language Navigation (VLN) research. It details the size, domain, task format, and evaluation metrics for each benchmark, highlighting their diverse applications and contributions to advancing VLN model development. The benchmarks range from foundational datasets like R2R to large-scale datasets such as ScaleVLN, providing diverse resources for evaluating navigation accuracy and instruction-following capabilities.

Datasets are crucial in VLN research for training and evaluating models that integrate visual and linguistic information. The Room-to-Room (R2R) dataset is foundational, offering 21,567 crowd-sourced instructions averaging 29 words, simulating realistic navigation scenarios [27]. The R4R dataset expands this with longer paths and 233,613 training samples to evaluate instruction fidelity [28]. The VLN-CE subset from R2R provides 5,611 trajectories for continuous environments [29]. The R2R-UNO dataset adds complexity by including diverse obstructions, presenting panoramic views with obstacles [30].

The IR2R and IR2R-CE datasets are critical for comparing advanced methods like OVER-NAV against baselines such as HAMT and MAP-CMA in discrete and continuous environments [35]. The RxR dataset, with 126,000 instructions across 16,500 paths, supports multilingual VLN research [31].

A large-scale dataset introduced by [32] includes 4.9 million instruction-trajectory pairs, surpassing previous datasets in size. Created using over 1200 photo-realistic environments from the HM3D and Gibson datasets, it supports a wide range of vision-language tasks and facilitates model training with extensive visual and linguistic data.

VLN research evaluation metrics assess model effectiveness and generalization capabilities. Key metrics include Success Rate (SR), Oracle Success Rate (OSR), and Path Length (PL) for evaluating navigation accuracy and efficiency. Additional metrics like Task Completion (TC), Shortest-Path Distance (SPD), and Normalized Dynamic Time Warping (nDTW) further gauge performance, alongside new metrics for assessing reasoning capabilities [36]. The AUC metric evaluates the discriminator's ability to differentiate high-quality from low-quality instruction-path pairs based on alignment [33].

Integrating novel multilingual visual-text datasets and advanced evaluation metrics enables the development of sophisticated VLN models. These models adeptly interpret complex natural language instructions within dynamic visual environments, enhancing navigation capabilities. This progress advances the field and lays the groundwork for innovative real-world navigation applications, guiding autonomous agents through unfamiliar terrains while effectively processing diverse linguistic inputs [37, 33]. Table 1 provides a detailed overview of the primary datasets and evaluation metrics employed in Vision-and-Language Navigation (VLN) research, underscoring their significance in training and assessing VLN models.

In recent years, Vision-and-Language Navigation (VLN) has emerged as a dynamic field, characterized by the interplay between visual perception and linguistic understanding. This integration is crucial for developing systems that can navigate complex environments based on both visual cues and verbal instructions. To elucidate the foundational elements of this domain, Figure 2 illustrates the hierarchical categorization of key concepts in VLN. This figure highlights various integration techniques, challenges, and advancements within the field. The structure delineates primary categories such as the integration of visual and linguistic data, the challenges faced by researchers, and recent advancements in navigation techniques. By providing this visual overview, we can better appreciate the current landscape and future directions of VLN research.

3 Vision-and-Language Navigation

3.1 Integration of Visual and Linguistic Data

Integrating visual and linguistic data is essential in Vision-and-Language Navigation (VLN), enabling the translation of natural language instructions into navigational actions by aligning linguistic inputs with visual perceptions [6]. Techniques such as NavHint generate sub-instructions and identify landmarks and distinctive objects, enhancing the agent's processing capabilities [5]. Advanced methods like Temporal Object Relations (TOR) and Spatial Object Relations (SOR) modules improve environmental understanding by learning object relations from temporal and spatial perspectives, crucial for effective navigation [1]. The Grid Memory Map (GridMM) structures the environment into a top-down grid format, capturing visual clues relevant to instructions [7].

The importance of these integration techniques is further illustrated in Figure 3, which depicts the various data integration methods, instruction generation approaches, and benchmark evaluations in VLN. It highlights NavHint, TOR-SOR, and GridMM as key methods for data integration, while the Spatially-Aware Speaker (SAS), NavCoT, and MLSTM-ATT are central to generating instructions. Additionally, the VLN-MP benchmark and multilingual datasets are pivotal for evaluating and assessing the performance of Vision-Language Models (VLMs).

The Spatially-Aware Speaker (SAS) model uses an encoder-decoder architecture to generate instructions from sequences of viewpoints and actions, facilitating data integration [22]. Similarly, the NavCoT method leverages large language models (LLMs) for self-guided reasoning, enhancing decision-making [10]. Multi-layer LSTM networks with attention mechanisms further improve dependency handling, enhancing visual-linguistic integration [20]. Transforming the VLN problem into a node classification task exemplifies diverse data source integration [14].

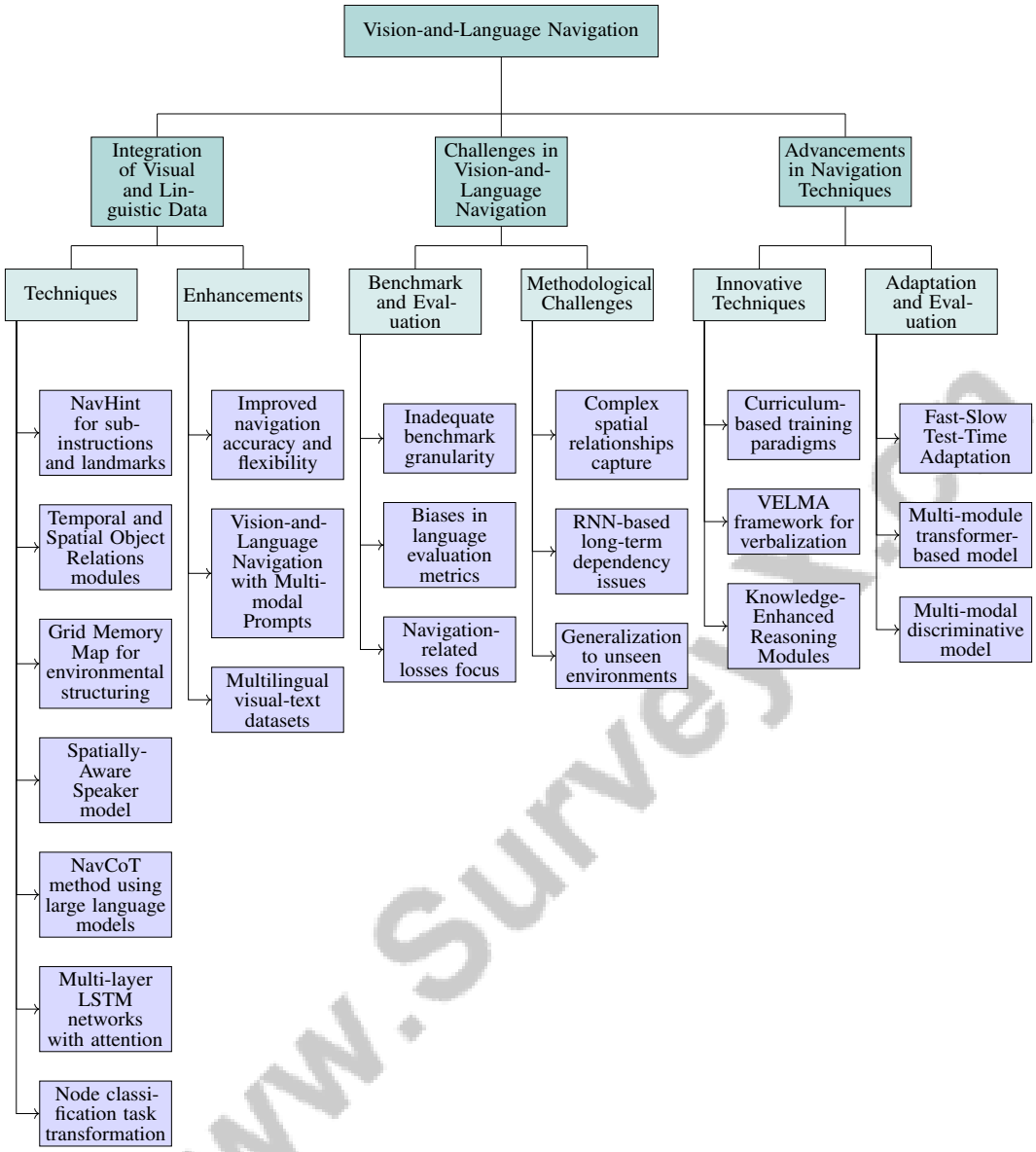

Figure 2: This figure illustrates the hierarchical categorization of key concepts in Vision-and-Language Navigation, highlighting integration techniques, challenges, and advancements. The structure delineates primary categories such as integration of visual and linguistic data, challenges faced, and recent advancements in navigation techniques, providing a clear overview of the field's current landscape and future directions.

These methodologies showcase advanced integration techniques, enhancing navigation accuracy and flexibility. Vision-and-Language Navigation with Multi-modal Prompts (VLN-MP) addresses textual ambiguity by incorporating visual prompts, significantly improving performance across benchmarks. Multilingual visual-text datasets facilitate comprehensive evaluation of Vision-Language Models (VLMs), enhancing understanding of their capabilities in processing image-text pairs across languages. These integrations equip VLN systems to handle real-world complexities, advancing research and practical applications [15, 37, 18].

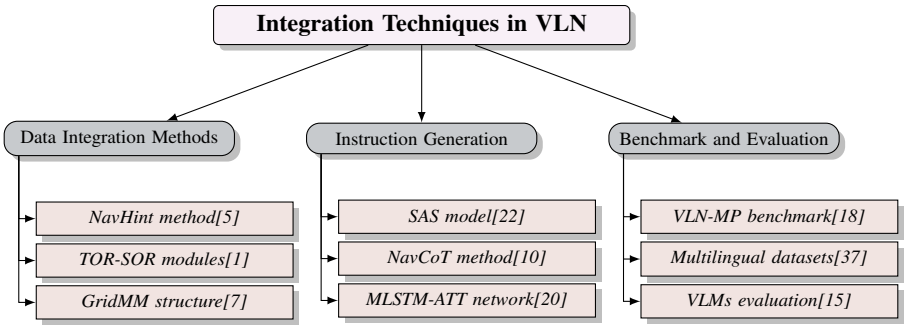

Figure 3: This figure illustrates the integration techniques in Vision-and-Language Navigation (VLN), focusing on data integration methods, instruction generation, and benchmark evaluation. It highlights the NavHint, TOR-SOR, and GridMM as key data integration methods, while SAS, NavCoT, and MLSTM-ATT are central to instruction generation. The VLN-MP benchmark and multilingual datasets are pivotal for evaluation and performance assessment of Vision-Language Models (VLMs).

3.2 Challenges in Vision-and-Language Navigation

VLN faces critical challenges, including inadequate benchmark granularity, which limits assessment of agents' understanding and adherence to instruction components [2]. Existing methods often produce lower-quality instructions due to biases in language evaluation metrics [22]. Many approaches focus on navigation-related losses, failing to provide a comprehensive understanding of textual and visual semantics [5]. Reinforcement learning methods struggle with the exponential number of potential paths, leading to sub-optimal decisions [14].

Current methods inadequately capture complex spatial relationships and detailed object attributes, resulting in information loss [7]. RNN-based methods, in particular, struggle with long-term dependencies, losing critical information when merging observations into fixed-size vectors [21]. Despite these challenges, innovative solutions are emerging, such as methods that generalize to unseen environments without prior mapping or fine-tuning, demonstrating improved performance [8]. Addressing these challenges is essential for advancing VLN and developing systems capable of reliable navigation in real-world applications.

3.3 Advancements in Navigation Techniques

Recent advancements in VLN have introduced innovative techniques enhancing visual-linguistic integration, improving performance and efficiency. Curriculum-based training paradigms optimize sample management without increasing complexity, significantly enhancing task performance [38]. The VELMA framework achieves state-of-the-art completion rates by integrating visual and linguistic information through verbalization, highlighting language's role as a bridge between perceptions and actions [12].

Knowledge-Enhanced Reasoning Modules (KERM) demonstrate success across datasets, validating that external knowledge incorporation enhances VLN tasks [13]. By embedding knowledge into navigation processes, KERM improves decision-making capabilities. The Fast-Slow Test-Time Adaptation (FSTTA) method introduces a two-phase update strategy, enhancing adaptability in dynamic environments and improving accuracy [3]. A multi-module transformer-based model (MMT) shifts focus from action prediction to trajectory grounding, improving target location identification [4].

Advancements in VLN systems, particularly through multi-modal prompts combining language with visual cues, significantly enhance navigation in complex environments. These innovations address text-only instruction limitations, improving accuracy and efficiency. The development of a multi-modal discriminative model allows better evaluation of instruction-path alignment, leading to improved generalization in unfamiliar settings. Collectively, these advancements create more robust and adaptable VLN systems, driving further innovations and enhancing applicability across navigation scenarios [18, 33].

4 Recurrent Neural Networks and Transformers

The development of Vision-and-Language Navigation (VLN) is heavily influenced by foundational architectures like Recurrent Neural Networks (RNNs) and Transformers. Long Short-Term Memory (LSTM) networks, a type of RNN, are essential for processing sequential data while maintaining contextual information, crucial for interpreting navigational instructions and decision-making in dynamic settings. This section delves into the role of LSTMs in VLN, highlighting their advantages and applications in enhancing navigation performance.

4.1 Role of LSTM in Vision-and-Language Navigation

LSTMs are integral to VLN, effectively processing sequential data and retaining context necessary for interpreting complex instructions. They excel in integrating visual and linguistic data by leveraging historical context to inform decisions. Multi-layer LSTM architectures, for instance, extract semantic frames from user commands, enhancing semantic parsing for robotic tasks [20]. In VLN, LSTMs use dual-level alignment strategies to connect instruction history with landmark observations, as exemplified by the VELMA framework, which verbalizes the agent's trajectory to improve interpretability [12]. Techniques like KERM further enhance action prediction and generalization by systematically integrating knowledge [13], while NavCoT demonstrates LSTMs' role in improving interpretability and scalability through structured reasoning [10].

This is illustrated in Figure 4, which highlights the role of LSTM in Vision-and-Language Navigation (VLN). The figure categorizes key methods and frameworks that leverage LSTMs to process sequential data and enhance context retention in VLN systems, contributing to improved navigation performance. Hybrid-training frameworks, such as those by [39], segment long action sequences into unit-grained instances, promoting active exploration and reducing exposure bias. The FSTTA method shows how balanced model updates enhance adaptability without compromising stability, outperforming traditional methods [3]. Overall, LSTMs significantly advance VLN systems by enhancing sequential data processing and context retention, leading to improved performance and success rates, especially when combined with advanced techniques like snapshot ensembles for better generalization [40, 33].

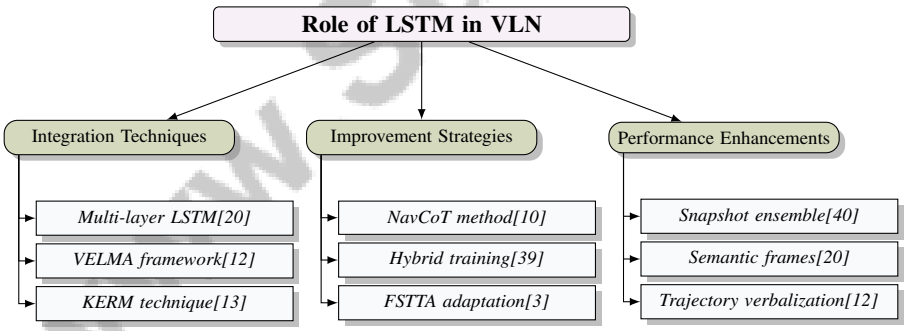

Figure 4: This figure illustrates the role of LSTM in Vision-and-Language Navigation (VLN), highlighting integration techniques, improvement strategies, and performance enhancements. It categorizes key methods and frameworks that leverage LSTMs to process sequential data and enhance context retention in VLN systems, contributing to improved navigation performance.

4.2 Transformer Architectures for Sequential Data Processing

Transformers have revolutionized VLN by providing a robust framework for sequential data processing, surpassing traditional LSTMs. Their self-attention mechanisms enable parallel processing of entire sequences, capturing complex dependencies and contextual relationships across modalities. Models like LLM-BRAIn leverage transformers for structured output generation [41]. The strength of transformers in integrating visual and linguistic data is further evidenced by methods such as OLA-VLM, which enhances visual representation quality within large language models (LLMs) [42]. VLN-BERT emphasizes balanced attention to noun and spatial tokens, facilitating effective navigation through comprehensive contextual understanding [43].

Innovative models like MTVM and Unit-Transformer showcase the versatility of transformers in VLN tasks. MTVM captures temporal context by storing activation outputs from previous steps [44], while Unit-Transformer processes segmented unit instances, enhancing task performance by integrating navigation and interaction phases [39]. Additionally, Zhao et al.'s multi-module transformer processes trajectory data, predicting target locations based on integrated visual and textual inputs [4]. The NavHint method employs a Transformer-based decoder to generate hints from the navigation agent's visual output, refining sequential data processing for improved outcomes [5].

The Scene Memory Transformer (SMT) exemplifies transformers' adaptability by using attention mechanisms to aid navigation in partially observable environments [21]. This adaptability and flexibility across diverse data modalities make transformers indispensable in advancing VLN systems.

Transformers offer a robust framework for sequential data processing in VLN, significantly improving capabilities over LSTMs in parallel processing, contextual understanding, and adaptability. These advancements enhance VLN systems' ability to navigate intricate environments, integrating multi-modal prompts and generating future view semantics, thereby addressing traditional text-based navigation ambiguities and improving agents' adaptability and overall performance [45, 46, 47, 18, 33].

4.3 Comparative Analysis: LSTM vs. Transformers

In VLN, both LSTMs and Transformers have been pivotal in advancing the integration of visual and linguistic data, yet they exhibit distinct characteristics affecting their performance in navigation tasks. LSTMs excel in processing sequential data through recurrent connections, maintaining contextual information crucial for tasks requiring a deep understanding of temporal dependencies, thus providing semantically rich representations beneficial for semantic relevance and image retrieval [6].

Conversely, Transformers have revolutionized VLN by offering parallel processing capabilities and self-attention mechanisms that capture complex dependencies across sequences, enabling superior translation quality and efficient management of long-sequence tasks compared to LSTMs [6]. Their computational efficiency allows for handling extensive data without the limitations of recurrent structures, making them suitable for high-throughput and scalable tasks [48].

The Snapshot Ensemble Method exemplifies Transformers' strengths by leveraging diverse navigational strategies from multiple model snapshots, compensating for individual limitations and enhancing overall performance [40]. The integration of Vision-Language Model (VLM) outputs into Transformer frameworks, as seen in the FTP framework, enriches action understanding in video-based tasks, enhancing interpretability and accuracy in navigation decisions [49].

While LSTMs are adept at capturing long-term dependencies and producing semantically rich representations, recent findings indicate that Transformers, with their parallel processing abilities and capacity to model complex interdependencies, often outperform LSTMs in intricate navigation tasks. Specifically, Transformers excel in machine translation and multimodal tasks, whereas LSTMs are more effective in tasks requiring semantic relevance, such as understanding user commands in human-robot interactions. This differential highlights the advantages of Transformers in applications demanding rapid processing and integration of diverse data types [20, 6]. The comparative analysis underscores the importance of selecting the appropriate model based on the specific VLN task requirements.

4.4 Innovative Approaches Integrating LSTM and Transformers

Innovative approaches in VLN have explored the integration of LSTMs and Transformer architectures to leverage their strengths for enhanced navigation capabilities. The SASRA model exemplifies this hybrid approach by combining transformers with recurrent neural networks to improve semantically-aware spatiotemporal reasoning, enhancing navigation in complex environments [50].

The MiniVLN framework illustrates the potential of integrating LSTMs and Transformers through a two-stage distillation approach, combining distillation in pre-training and fine-tuning phases. This strategy improves performance while reducing model size, demonstrating the efficiency of a cohesive integration of these architectures [51]. By leveraging the strengths of both models, MiniVLN achieves a balance between computational efficiency and navigational accuracy.

The VLN-PETL approach showcases parameter-efficient transfer learning, focusing on crucial interactions influencing action prediction. This method enhances cross-modal communication by integrating LSTM's sequential processing with Transformer's attention mechanisms, improving navigation outcomes [52].

Furthermore, the Scene Memory Transformer (SMT) employs a flexible memory structure that stores observations separately, deferring aggregation until action computation. This approach enhances adaptability and performance by effectively managing information from both LSTM and Transformer components [21]. The ability to retrieve and manage past observations underscores the utility of hybrid architectures in VLN tasks.

The exploration of hybrid models that merge RNNs and Transformers is a promising area for future research, addressing gaps in understanding how multimodal training influences representation quality [6]. Investigating specific layers aligning with human behavior and exploring new multimodal integration strategies can further enhance VLN systems' alignment and performance [53].

These innovative approaches demonstrate the potential of integrating LSTM and Transformer architectures to create robust VLN systems capable of navigating diverse environments with improved efficiency and accuracy. Research into hybrid models and multimodal integration strategies, such as the NExT-GPT system for any-to-any content generation across text, images, audio, and video, is set to significantly enhance AI capabilities by facilitating sophisticated cross-modal semantic understanding and improving information retrieval in complex data environments. This exploration addresses the limitations of current multimodal large language models (MM-LLMs) while emphasizing the importance of efficiently managing diverse data sources to advance the field toward more human-like AI interactions [54, 19].

5 Cross-Modal and Multi-Modal Learning

5.1 Integration of Diverse Data Sources

Integrating diverse data sources is crucial for enhancing Vision-and-Language Navigation (VLN) systems by synthesizing visual, linguistic, and sensory data to construct a comprehensive environmental understanding. The multi-turn curriculum-based learning approach exemplifies this integration, enhancing Vision-Language Models' (VLMs) in-context learning capabilities and improving performance across benchmarks [55]. Collecting multilingual datasets via machine translation and cross-lingual encoders further illustrates this integration, improving navigation across languages [56].

Incorporating precise spatial understanding into VLMs is critical, as demonstrated by using the original PaLM-E dataset and a spatial VQA dataset for evaluation [57]. This integration addresses gaps in spatial comprehension, facilitating nuanced navigation decisions. Dual-level alignment strategies on datasets like R2R, RxR, R4R, and CVDN emphasize integrating diverse data sources to enhance navigation performance [58].

The OLA-VLM method shows how integrating visual marking techniques with VLMs enhances navigation outcomes while optimizing computational resources [42]. Masking experiments reveal insights into integrating linguistic data, deepening understanding of how diverse sources contribute to navigation performance [43]. By effectively combining visual, linguistic, and sensory data, these systems achieve a more comprehensive understanding of their environments, leading to improved accuracy and efficiency.

Future research should focus on innovative data integration strategies to enhance VLN systems' adaptability and effectiveness in complex environments. This includes leveraging semantic-aware techniques for processing multimodal data and creating comprehensive multilingual visual-text datasets to evaluate and fine-tune VLMs across diverse tasks and languages. Addressing data quality and temporal dynamics challenges within expansive data lakes can improve information retrieval capabilities, ensuring VLN systems' relevance in a rapidly evolving technological landscape [37, 40, 19, 59].

5.2 Attention Mechanisms in Cross-Modal Learning

Attention mechanisms are pivotal in enhancing cross-modal learning by selectively focusing on relevant data portions across modalities, improving visual and linguistic information integration

and interpretation. In VLN, these mechanisms align visual perceptions with linguistic instructions, enabling models to prioritize critical features for accurate navigation. The Scene Memory Transformer (SMT) exemplifies this by enhancing memory utilization, allowing agents to navigate complex environments by attending to pertinent visual and linguistic cues [21].

Models like the Spatially-Aware Speaker (SAS) use attention to generate coherent instructions that align visual observations with linguistic expressions, facilitating effective navigation [22]. The DELA method's dual-level alignment strategy enhances decision-making by ensuring relevant features are attended to across both modalities [58]. In the NavCoT method, attention mechanisms integrate visual and linguistic data, enhancing interpretability and scalability [10].

These examples highlight attention mechanisms' significant role in cross-modal learning, enhancing navigation performance by integrating visual prompts with textual instructions and addressing text-only guidance ambiguities. Recent advancements in multimodal large language models (MLLMs) reveal distinct stages of information flow between visual and linguistic data, leading to more accurate predictions and improved task execution [16, 18]. By focusing on relevant features across modalities, attention mechanisms enhance data integration and interpretation, resulting in improved navigation accuracy and efficiency. Future VLN research should explore innovative attention mechanism applications, further advancing cross-modal learning systems.

5.3 Challenges and Innovations in Cross-Modal Learning

Cross-modal learning presents challenges that must be addressed to enhance diverse data source integration in VLN. Intricate biases can impede model generalization across environments and tasks, often stemming from complexities in aligning visual and linguistic data, leading to sub-optimal navigation performance. Addressing these biases requires innovative approaches that augment learning processes and improve adaptability [60].

Maintaining recall and stability during training, as seen in models like Mamba, is another significant challenge. Instability can hinder effectiveness, especially in dynamic environments where consistent performance is crucial [48]. Overcoming these challenges necessitates robust training paradigms to ensure stability and enhance recall capabilities.

Innovations in cross-modal learning have leveraged natural language to guide agent exploration and improve navigation performance. Language-based augmentation techniques offer promising solutions by addressing biases and enhancing model interpretation of diverse data inputs, facilitating nuanced navigation [60]. Advanced attention mechanisms enhance cross-modal learning by facilitating interaction between linguistic and visual information within MLLMs, particularly in tasks like visual question answering, where models process and combine features from both modalities through distinct representation stages [16, 6, 19].

Continued exploration of these challenges and innovations is crucial for advancing cross-modal learning in VLN. Developing innovative approaches that mitigate biases, improve stability, and integrate natural language with visual prompts for enhanced exploration can significantly advance cross-modal learning systems. This progress will lead to more effective navigation solutions, as evidenced by the Vision-and-Language Navigation with Multi-modal Prompts (VLN-MP) framework, which enhances traditional navigation tasks by combining text and images while demonstrating superior performance across benchmarks. Such advancements pave the way for intelligent agents capable of seamlessly interpreting complex instructions and navigating diverse environments [15, 18].

6 Large Language Models and Vision-Language Models

6.1 Integration of LLMs and VLMs in Navigation

The integration of Large Language Models (LLMs) and Vision-Language Models (VLMs) in navigation systems marks a pivotal advancement in Vision-and-Language Navigation (VLN), enhancing adaptability and interpretability in complex tasks. Dynamic mapping systems that update in real-time, as opposed to static maps, exemplify this integration, enabling more responsive navigation in dynamic environments [8]. The KERM framework underscores the synergy of external knowledge with visual and instructional data, enriching agents' contextual understanding and decision-making processes [13]. Similarly, SayNav illustrates how LLMs can ground navigation strategies in rich linguistic con-

texts, improving task efficiency through dynamic visual data incorporation [61]. This integration also addresses the challenge of catastrophic forgetting in VLMs, as evidenced by the PROOF evaluation across nine benchmark datasets, ensuring navigation systems retain critical knowledge over time [62]. Future VLN research is poised to explore integrating spatially-aware speaker models with multi-modal transformer architectures, enhancing instruction generation capabilities and addressing dataset constraints [22]. The combined capabilities of LLMs and VLMs facilitate developing versatile agents adept at navigating unexpected obstacles, leading to enhanced navigation accuracy and operational efficiency in real-world applications [48, 30, 63, 18]. Continued exploration of innovative integration strategies will likely yield further advancements in VLN systems.

6.2 Advancements in VLM Architectures

Recent advancements in Vision-Language Model (VLM) architectures have significantly improved the integration of visual and linguistic data, enhancing performance across various benchmarks. Cross-attention architectures, in particular, have shown superior performance compared to traditional self-attention models when backbone networks remain frozen during training, facilitating effective integration of inputs from both modalities [59]. The introduction of Idefics2, a foundational VLM with 8 billion parameters, marks a milestone in the field, achieving state-of-the-art performance in its category and underscoring the importance of scaling model parameters for enhanced representational capacity and generalization [64]. These advancements reflect the ongoing evolution of VLM architectures, pushing the boundaries of vision-language integration. Enhanced attention mechanisms and expanded model parameters foster the development of advanced VLMs capable of engaging with visual content and textual instructions, allowing for proactive information-seeking behaviors in ambiguous scenarios. This enables tasks ranging from simple caption generation to complex embodied question answering, ultimately leading to robust decision-making in real-world contexts [65, 15, 39, 66]. The pursuit of innovative architectural designs is expected to yield further improvements in VLM performance and versatility.

6.3 Impact of Pretraining and Fine-tuning

Pretraining and fine-tuning are crucial processes that enhance the performance and adaptability of Vision-and-Language Navigation (VLN) models. Pretraining on large-scale datasets fosters robust visual and linguistic representations essential for navigating complex environments. Incorporating pre-trained language models into navigation systems significantly aligns states and actions within the latent space, improving offline reinforcement learning algorithms [67]. Experiments with LLaMA-Adapter and LLaMA 2 models demonstrate the substantial impact of pretraining on model performance [10]. Fine-tuning refines model performance through domain-specific adjustments, enhancing adherence to instructions while maintaining response quality [68]. Moreover, employing candidate waypoint predictors during fine-tuning effectively bridges the discrete-to-continuous gap in navigation tasks, improving overall performance [69]. Evaluation metrics such as Success Rate (SR), Trajectory Length (TL), and Navigation Error (NE) are paramount for assessing pretraining and fine-tuning effectiveness in real-world navigation scenarios. These metrics evaluate a model's proficiency in navigating intricate pathways while minimizing errors, emphasizing the necessity of accurately reaching designated locations based on natural language instructions. Recent critiques of evaluation methods highlight the need for metrics like the Coverage weighted by Length Score (CLS) and complex datasets such as Room-for-Room (R4R) to better understand agents' interpretation and execution of instructions in varied environments [48, 28]. Despite advancements, challenges persist in optimizing model architectures and data quality. Increasing the size of vision transformers (ViT) does not guarantee improved performance in multimodal tasks, necessitating further exploration of model architectures and data quality [70]. Additionally, path-ranking models' reliance on noun tokens, while neglecting spatial and directional cues, hampers navigation performance [43]. This suggests a need for fine-tuning strategies that enhance spatial information processing capabilities. Pretraining and fine-tuning are essential in developing advanced VLN models, significantly influencing performance and efficiency, particularly in integrating diverse datasets and optimizing the selection of supervised fine-tuning (SFT) data for improved outcomes [37, 71]. Leveraging large-scale datasets and strategic data selection ensures models are equipped to handle the intricacies of vision-and-language integration, paving the way for further advancements in the field.

6.4 Applications in Real-World Scenarios

The integration of Large Language Models (LLMs) and Vision-Language Models (VLMs) into navigation systems has significantly broadened the potential for real-world applications, enhancing the capabilities of autonomous agents across various domains. The A2Nav method exemplifies LLM application in facilitating human-robot interaction, enabling action-aware zero-shot learning vital for dynamic environments [72]. In web-based navigation, the WebVLN-Net method enhances navigation and question-answering performance, demonstrating VLM adaptability in digital spaces for efficient information retrieval [45]. Experiments with the HM-EQA dataset reveal how confidence calibration and VLM integration can improve exploration efficiency and decision-making accuracy in simulated human-robot scenarios [66]. BEVBert's performance across multiple VLN benchmarks validates its role in enhancing spatial-aware reasoning, crucial for applications requiring precise navigation [73]. Similarly, the Atlas model advances 3D detection and planning for autonomous driving, establishing the 3D-tokenized LLM as a key development for reliable vehicle navigation [74]. In robotic navigation, the IVLMap method has demonstrated substantial improvements in navigation accuracy, indicating its potential for real-world deployment in instance-aware visual-language integration tasks [75]. VLMs have also proven effective in map parsing for mobile robotics, generating accurate navigation plans from floor plan images and enhancing autonomous navigation efficiency [76]. The CM2 model's superior performance on the VLN-CE dataset showcases the practical applications of cross-modal learning in navigation, outperforming existing methods and adapting to diverse scenarios [17]. Additionally, the MiC model's improvements in navigation and object grounding on the REVERIE benchmark highlight the potential of interactive prompting in refining navigation strategies and enhancing object recognition [77]. Future research should continue to refine the instruction-following dimension and investigate methods to enhance generalization across diverse instruction types [68]. The effectiveness of approaches like MORE-3S in semantically aligning states and actions demonstrates improved decision-making and adaptability in reinforcement learning tasks [67]. Ethical reviews prior to deploying LLMs in autonomous systems are crucial to mitigate safety hazards, bias, and privacy issues [78]. These real-world applications underscore the transformative impact of LLMs and VLMs on navigation tasks, offering enhanced adaptability, accuracy, and efficiency across diverse environments. Ongoing investigations into advanced models, such as the Mamba Transformer and large VLMs, are set to significantly enhance autonomous navigation by integrating improved semantic reasoning and exploration strategies, including building semantic maps and calibrating confidence levels to optimize exploration efficiency and planning [48, 66].

7 Reinforcement Learning and Imitation Learning

7.1 Overview of Reinforcement and Imitation Learning in Navigation

Method Name	Training Methodologies	Integration Techniques	Evaluation Metrics
VLN-ORL[79]	Offline Reinforcement Learning	Reward-conditioned Approach	Success Rate
SLMP[80]	Trial-and-error	Combination OF Methods	Success Rate
SPCL[38]	Curriculum-based Training	Curriculum Learning Strategies	Navigation Error
MORE-3S[67]	Autoregressive Modeling Approach	Multimodal Information Integration	Average Rewards
SX[81]	Auxiliary Learning Task	Strategic Exploration Model	Success Rate
CLIP-MC[82]	Trial-and-error	Combination OF Methods	Task Completion

Table 2: Comparison of various navigation methods employing Reinforcement Learning and Imitation Learning, detailing their training methodologies, integration techniques, and evaluation metrics. The table highlights the diversity in approaches such as offline reinforcement learning, trial-and-error, curriculum-based training, and autoregressive modeling, as well as the different metrics used to assess their performance including success rate, navigation error, and average rewards.

Reinforcement Learning (RL) and Imitation Learning (IL) are pivotal in creating advanced navigation systems, each offering unique training methodologies. RL relies on trial-and-error, optimizing actions via reward signals to enhance decision-making, as demonstrated by the VLN-ORL method in diverse scenarios [79]. Deep RL in motion planners exemplifies its capacity to train agents for autonomous navigation to novel targets [80]. IL, conversely, allows agents to mimic expert behaviors, integrating visual features with language instructions for accurate navigation in varied environments [83, 68, 13, 25]. The synergy of IL and RL leverages expert knowledge and environmental interaction,

enhanced further by curriculum learning for improved adaptability and performance [38]. Table 2 provides a comprehensive comparison of different methodologies and evaluation metrics employed in Reinforcement Learning and Imitation Learning for navigation systems.

Despite RL's benefits, challenges in adapting large-scale models for offline learning persist due to limited interaction [67]. Strategies like strategic novelty-seeking improve recovery capabilities and outcomes [81]. Metrics such as Task Completion and nDTW evaluate RL and IL strategies' effectiveness [82]. Integrating RL and IL fosters robust navigation agents capable of interpreting multimodal instructions and adjusting strategies dynamically, ensuring robust performance in complex environments [25, 84, 30, 85].

7.2 Challenges and Innovations in RL for Navigation

Method Name	Learning Efficiency	Generalization Capability	Integration Challenges
MORE-3S[67]	Reduce Training Time	Unseen Environments	Pretrained Language Models
SEA[86]	Reduce Training Time	Unseen Environments	-
PIVOT[87]	Iterative Visual Optimization	Zero-shot Control	3D Understanding Limitations
HOP[88]	Reduce Training Time	Improved Generalization	Substantial Computational Resources
CN[89]	Zero-shot Navigation	Unseen Environments	-

Table 3: Comparison of various reinforcement learning methods for navigation, highlighting their learning efficiency, generalization capabilities, and integration challenges. The table provides insights into how different approaches address training time reduction, adaptability to unseen environments, and specific integration hurdles with existing models and resources.

Reinforcement Learning (RL) in navigation faces challenges due to the complexity of real-world environments and the need for efficient learning strategies. A major issue is the separation of high-level planning from low-level representations, causing suboptimal performance [67]. Integrating auxiliary tasks enhances learning efficiency, reducing training time and improving performance [86]. Innovative methods like PIVOT enable zero-shot control for spatial reasoning, though current VLMs struggle with 3D understanding and fine-grained control [87]. The black box nature of LLMs complicates integration, exacerbating the sim2real gap [78]. Table 3 presents a comparative analysis of recent reinforcement learning methods in navigation, focusing on their efficiency, generalization potential, and the challenges encountered during integration.

Despite these challenges, innovations like HOP improve generalization and navigation capabilities [88]. CLIPNav demonstrates generalization to unseen environments, crucial for diverse navigation [89]. Recent innovations, such as multi-modal prompts and offline RL, enhance performance by reducing ambiguity and improving knowledge transfer [83, 15, 79, 18, 26]. Addressing current limitations and incorporating innovative solutions will advance RL capabilities, enabling navigation in complex environments with greater accuracy and reliability.

7.3 Imitation Learning Techniques and Applications

Imitation Learning (IL) is crucial for developing navigation systems, allowing agents to learn from expert demonstrations. IL is particularly useful when direct RL is impractical due to environmental complexity or cost. Techniques leveraging visual inputs guide navigation, enabling agents to interpret cues and translate them into actions [90]. Incorporating auxiliary tasks and supervised pre-training enhances performance by providing a richer set of learning signals, enabling adaptation to diverse environments [90]. Future research will explore auxiliary tasks and pre-training in broader environments, leading to versatile navigation systems capable of managing real-world challenges [15, 30, 91, 78].

7.4 Actor-Critic Algorithms for Policy Optimization

Actor-Critic algorithms, notably Advantage Actor-Critic (A2C), optimize navigation policies by integrating policy-based and value-based approaches, crucial for VLN tasks [72, 83, 92, 90]. A2C combines reinforcement learning with auxiliary tasks, enhancing agents' understanding and execution of navigation commands. The actor proposes actions, while the critic evaluates them, enabling efficient learning in complex environments. A2C's stochastic policy promotes exploration, improving

robustness and adaptability [93]. Integrating image features with text inputs exemplifies A2C's multimodal capabilities.

A2C addresses challenges like sample inefficiency and high-dimensional spaces, fostering exploration and reducing premature convergence. It enhances decision-making through advanced strategies like mistake-aware path planning and semantic mapping [94, 48, 66, 81]. The critic provides feedback to the actor, guiding policy optimization and improving accuracy and efficiency in VLN tasks. By integrating performance insights, A2C enables effective models adaptable to real-world complexities [83, 5, 4]. Continued refinement of these algorithms is expected to yield further advancements in autonomous navigation systems.

7.5 Integration of RL and IL for Enhanced Navigation

Integrating Reinforcement Learning (RL) and Imitation Learning (IL) enhances navigation performance by combining RL's generalization with IL's efficiency. Large-scale synthetic instruction generation improves agents' ability to follow complex instructions across diverse environments. Approaches like Soft Expert Reward Learning (SERL) and reward-conditioned methods enhance performance in unseen scenarios [34, 83, 79, 95]. Incorporating object features into RL frameworks enhances navigation by informing strategies with spatial and semantic relationships [96]. Multimodal and pre-trained sequence models in offline RL transform learning into a supervised task, improving performance in real-world tasks [67].

Combining RL and IL allows systems to benefit from both approaches. IL guides initial policy development, while RL refines policies through interaction. This dual approach enhances adaptability to dynamic environments, leading to improved performance even with unexpected challenges [30, 97, 4, 92, 18]. The integration of RL and IL will significantly enhance autonomous navigation systems, leveraging LLMs for improved reasoning and adaptability, particularly in unpredictable scenarios. Addressing challenges like instruction-reality mismatches and incorporating innovative strategies will achieve greater accuracy and efficiency, bridging the performance gap between state-of-the-art VLN models and LLM-driven approaches [25, 30, 80].

8 Embodied AI and Semantic Mapping

8.1 Introduction to Embodied AI

Embodied AI represents a transformative shift in artificial intelligence, focusing on systems that physically interact with and navigate real-world environments. This paradigm is crucial for navigation tasks, necessitating the interpretation and response to complex environmental cues. Embodied AI systems effectively integrate sensory inputs, such as visual and auditory signals, with advanced decision-making processes to operate in dynamic settings. Techniques like semantic mapping and iterative visual prompting enable these systems to explore surroundings efficiently, adapting to diverse contexts without extensive task-specific training [98, 87, 99, 66].

The role of embodied AI in navigation is pivotal, enhancing the adaptability and robustness of autonomous agents. By incorporating physical embodiment, these systems leverage real-world interactions to refine their understanding of spatial and semantic relationships, crucial for accurate navigation and object interaction. This capability allows agents to formulate sophisticated strategies informed by real-time feedback, significantly improving adaptability to dynamic environments and unforeseen obstacles. For instance, integrating visual prompts in Vision-and-Language Navigation (VLN) tasks clarifies ambiguous textual instructions and fosters a comprehensive understanding of surroundings, thereby enhancing performance in both standard and obstructed navigation scenarios [30, 5, 18].

The advancement of embodied AI is supported by systematic frameworks for evaluating Vision-Language Architectures (VLAs), which illuminate the impact of various design choices on model performance [100]. These benchmarks guide the development of robot policies, ensuring that embodied AI systems are equipped to navigate real-world complexities. Future research may focus on improving model robustness in diverse urban environments and exploring multimodal integration techniques to enhance navigation accuracy and efficiency [36]. By refining the integration of sensory data and decision-making processes, embodied AI holds the potential to significantly advance autonomous systems' capabilities in navigation and interaction.

8.2 Semantic Mapping Techniques

Semantic mapping techniques are crucial for enhancing navigation systems, offering structured representations that integrate spatial and semantic information. These techniques facilitate the extraction of meaningful embeddings from diverse data modalities, improving information retrieval and navigation capabilities. For instance, advancements in vision-and-language navigation (VLN) highlight the importance of merging textual instructions with spatial layouts, allowing agents to navigate unfamiliar settings more effectively. Architectures like Semantic Understanding and Spatial Awareness (SUSA) demonstrate that hybrid semantic-spatial representations can significantly enhance navigation performance, achieving state-of-the-art results across multiple benchmarks [101, 91, 19]. This integration enables autonomous agents to understand and interpret complex environments, leading to more accurate and efficient navigation.

Recent developments in semantic mapping emphasize models capable of integrating and processing diverse data sources. Multi-view learning frameworks, for example, allow for embedding extraction from mono-modal, multi-modal, and cross-modal data, addressing temporal aspects of time-series data [11]. This approach enhances agents' navigational abilities by providing a comprehensive understanding of environments, capturing both static and dynamic elements influencing navigational decisions.

The integration of semantic-aware techniques is exemplified by models leveraging large-scale pre-trained vision and language backbones, enhancing reasoning capabilities across spatial and semantic dimensions critical for effective navigation in complex environments [12]. By incorporating knowledge from diverse domains, these models improve agents' abilities to interpret and act upon navigational instructions, resulting in more accurate and reliable outcomes.

Moreover, combining semantic mapping with reinforcement learning (RL) and imitation learning (IL) strategies has shown promise in enhancing navigation performance. Integrating semantic information into the learning process enables agents to develop nuanced navigation strategies informed by spatial and semantic relationships [67]. This integration fosters contextually aware navigation policies, allowing agents to navigate complex environments with greater precision.

Semantic mapping techniques are essential for improving navigation systems, offering structured representations that integrate semantic understanding and spatial awareness. By employing advanced methods such as topological map generation from natural language instructions and hybrid semantic-spatial representations, these techniques empower navigation systems to interpret and traverse complex, real-world environments, including those with unexpected obstructions. This integration enhances path estimation and action prediction accuracy while enabling adaptive navigation based on dynamic conditions, advancing vision-and-language navigation systems significantly [30, 101, 91, 19]. Continued refinement of these techniques and exploration of innovative strategies for integrating semantic information will further enhance the performance and adaptability of autonomous navigation systems.

8.3 Semantic and Topological Mapping

Semantic and topological mapping are integral to developing advanced navigation systems, providing a dual-layered approach to environmental understanding that enhances agents' navigation capabilities in complex spaces. Semantic mapping creates detailed representations that encompass both spatial layouts and semantic information regarding objects and features. This integration allows navigation systems to recognize and categorize objects while understanding their relationships and relevance to specific navigational tasks, thereby improving overall performance in structured and obstructed settings [30, 63, 102, 18].

Conversely, topological mapping emphasizes the connectivity and adjacency of locations within an environment, generating a network-like representation that facilitates efficient path planning and decision-making. The synergy between semantic and topological mapping enhances navigation systems by combining the strengths of both approaches—semantic mapping offers a rich understanding of environmental features while topological mapping provides a structured representation of spatial relationships. This integration results in a comprehensive environmental understanding, enabling the development of effective and adaptable navigation strategies, particularly in dynamic and complex environments. Recent advancements in VLN leverage topological maps to create interpretable navi-

gation plans from natural language instructions, significantly improving navigation performance in real-world contexts [103, 91, 101, 19, 104].

New explorations have employed multi-view learning frameworks to enhance the extraction of semantic embeddings from diverse data sources, addressing mono-modal and cross-modal information [11]. This approach not only improves agents' interpretations of complex environments but also supports dynamic map updates as new information becomes available.

Moreover, integrating large-scale pre-trained models in semantic mapping has demonstrated potential in enhancing navigation systems' reasoning capabilities. By leveraging powerful vision and language backbones, these models can accurately interpret and act upon complex navigational instructions, particularly beneficial in environments with varying complexity and detail [12].

The combination of semantic and topological mapping also fosters the development of robust navigation policies by incorporating spatial and semantic information into the learning process. This holistic approach enables agents to devise nuanced navigation strategies informed by both the physical layout and the semantic context, leading to improved real-world performance [67].

The integration of semantic and topological mapping signifies a substantial advancement in navigation systems, offering a comprehensive framework for understanding and interacting with complex environments. By continuously advancing techniques and investigating innovative integration strategies, researchers can significantly enhance the performance of autonomous navigation systems, particularly in real-world scenarios that include unexpected obstacles. For instance, the introduction of the R2R with UNexpected Obstructions (R2R-UNO) dataset highlights challenges faced by existing VLN methods, which often struggle to adapt to discrepancies between navigation instructions and actual conditions. The development of the ObVLN (Obstructed VLN) method illustrates how a curriculum training strategy and virtual graph construction can enhance adaptability in obstructed settings. These advancements not only bolster the reliability of autonomous navigation systems but also pave the way for sophisticated applications across various sectors, from indoor navigation to outdoor exploration [48, 30].

9 Conclusion

9.1 Future Directions and Research Opportunities

The Vision-and-Language Navigation (VLN) domain offers a rich landscape for advancing the robustness, adaptability, and efficiency of navigation systems. Future research could focus on enhancing Scene Memory Transformer (SMT) models by improving their resilience to noisy environments and optimizing memory utilization, which would significantly enhance agents' navigational abilities in dynamic settings. Incorporating knowledge-based frameworks like KERM into continuous navigation scenarios could further enrich contextual understanding, thereby enhancing decision-making processes.

The integration of larger vision-language models with systems like NavCoT, supported by innovative training methodologies, holds promise for improving navigation performance. Exploring unit-grained hybrid approaches and extending Fast-Slow Test-Time Adaptation (FSTTA) across diverse tasks and datasets could yield insights into model generalizability and effectiveness, fostering the development of more versatile navigation models.

Addressing the challenges of navigating environments where target locations are missing from trajectories, and improving generalization across varied settings, are critical areas for future research. This requires sophisticated modeling techniques and enhanced training datasets to overcome current limitations and elevate navigation performance. Furthermore, refining sub-instruction modules and developing additional datasets could enhance agents' ability to interpret and execute complex instructions, while the integration of advanced language models and vision representations offers significant potential for performance improvements.

Generalizing methods for more complex navigation scenarios and refining the explore-and-exploit framework are promising research trajectories. Enhancing indoor environment representations and tackling multi-floor navigation complexities are essential steps toward advancing VLN systems. These research directions highlight the potential for significant progress in VLN, supporting the development of sophisticated and reliable navigation systems capable of operating in diverse real-world contexts.

Continued exploration in these areas will significantly enhance the capabilities of autonomous navigation systems, paving the way for innovative applications and improved performance in complex environments.

References

[1] Bowen Huang, Yanwei Zheng, Chuanlin Lan, Xinpeng Zhao, Yifei Zou, and Dongxiao yu. Temporal-spatial object relations modeling for vision-and-language navigation, 2024.

[2] Yicong Hong, Cristian Rodriguez-Opazo, Qi Wu, and Stephen Gould. Sub-instruction aware vision-and-language navigation, 2020.

[3] Junyu Gao, Xuan Yao, and Changsheng Xu. Fast-slow test-time adaptation for online vision-and-language navigation, 2024.

[4] Chongyang Zhao, Yuankai Qi, and Qi Wu. Mind the gap: Improving success rate of vision-and-language navigation by revisiting oracle success routes, 2023.

[5] Yue Zhang, Quan Guo, and Parisa Kordjamshidi. Navhint: Vision and language navigation agent with a hint generator, 2024.

[6] Jindřich Libovický and Pranava Madhyastha. Probing representations learned by multimodal recurrent and transformer models, 2019.

[7] Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, and Shuqiang Jiang. Gridmm: Grid memory map for vision-and-language navigation, 2023.

[8] Chengguang Xu, Hieu T. Nguyen, Christopher Amato, and Lawson L. S. Wong. Vision and language navigation in the real world via online visual language mapping, 2023.

[9] Kunyang Lin, Peihao Chen, Diwei Huang, Thomas H. Li, Mingkui Tan, and Chuang Gan. Learning vision-and-language navigation from youtube videos, 2023.

[10] Bingqian Lin, Yunshuang Nie, Ziming Wei, Jiaqi Chen, Shikui Ma, Jianhua Han, Hang Xu, Xiaojun Chang, and Xiaodan Liang. Navcot: Boosting llm-based vision-and-language navigation via learning disentangled reasoning, 2024.

[11] Qiaolin Xia, Xiujun Li, Chunyuan Li, Yonatan Bisk, Zhifang Sui, Jianfeng Gao, Yejin Choi, and Noah A. Smith. Multi-view learning for vision-and-language navigation, 2020.

[12] Raphael Schumann, Wanrong Zhu, Weixi Feng, Tsu-Jui Fu, Stefan Riezler, and William Yang Wang. Velma: Verbalization embodiment of llm agents for vision and language navigation in street view, 2024.

[13] Xiangyang Li, Zihan Wang, Jiahao Yang, Yaowei Wang, and Shuqiang Jiang. Kerm: Knowledge enhanced reasoning for vision-and-language navigation, 2023.

[14] Xinzhe Zhou, Wei Liu, and Yadong Mu. Rethinking the spatial route prior in vision-and-language navigation, 2021.

[15] Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Eric Wang. Vision-and-language navigation: A survey of tasks, methods, and future directions, 2022.

[16] Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models, 2024.

[17] Georgios Georgakis, Karl Schmeckpeper, Karan Wanchoo, Soham Dan, Eleni Miltsakaki, Dan Roth, and Kostas Daniilidis. Cross-modal map learning for vision and language navigation, 2022.

[18] Haodong Hong, Sen Wang, Zi Huang, Qi Wu, and Jiajun Liu. Why only text: Empowering vision-and-language navigation with multi-modal prompts, 2024.

[19] Pierre Lamart, Yinan Yu, and Christian Berger. Semantic-aware representation of multi-modal data for data ingress: A literature review, 2024.

[20] Martino Mensio, Emanuele Bastianelli, Ilaria Tiddi, and Giuseppe Rizzo. A multi-layer lstm-based approach for robot command interaction modeling, 2018.

[21] Kuan Fang, Alexander Toshev, Li Fei-Fei, and Silvio Savarese. Scene memory transformer for embodied agents in long-horizon tasks, 2019.

[22] Muraleekrishna Gopinathan, Martin Masek, Jumana Abu-Khalaf, and David Suter. Spatially-aware speaker for vision-and-language navigation instruction generation, 2024.

[23] Zijiao Yang, Arjun Majumdar, and Stefan Lee. Behavioral analysis of vision-and-language navigation agents, 2023.

[24] Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. Towards learning a generic agent for vision-and-language navigation via pre-training, 2020.

[25] Gengze Zhou, Yicong Hong, Zun Wang, Xin Eric Wang, and Qi Wu. Navgpt-2: Unleashing navigational reasoning capability for large vision-language models, 2024.

[26] Yue Zhang, Ziqiao Ma, Jialu Li, Yanyuan Qiao, Zun Wang, Joyce Chai, Qi Wu, Mohit Bansal, and Parisa Kordjamshidi. Vision-and-language navigation today and tomorrow: A survey in the era of foundation models, 2024.

[27] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments, 2018.

[28] Vihan Jain, Gabriel Magalhaes, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. Stay on the path: Instruction fidelity in vision-and-language navigation, 2019.

[29] Jacob Krantz and Stefan Lee. Sim-2-sim transfer for vision-and-language navigation in continuous environments, 2022.

[30] Haodong Hong, Sen Wang, Zi Huang, Qi Wu, and Jiajun Liu. Navigating beyond instructions: Vision-and-language navigation in obstructed environments, 2024.

[31] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding, 2020.

[32] Zun Wang, Jialu Li, Yicong Hong, Yi Wang, Qi Wu, Mohit Bansal, Stephen Gould, Hao Tan, and Yu Qiao. Scaling data generation in vision-and-language navigation, 2023.

[33] Haoshuo Huang, Vihan Jain, Harsh Mehta, Jason Baldridge, and Eugene Ie. Multi-modal discriminative model for vision-and-language navigation, 2019.

[34] Aishwarya Kamath, Peter Anderson, Su Wang, Jing Yu Koh, Alexander Ku, Austin Waters, Yinfei Yang, Jason Baldridge, and Zarana Parekh. A new path: Scaling vision-and-language navigation with synthetic instructions and imitation learning, 2023.

[35] Ganlong Zhao, Guanbin Li, Weikai Chen, and Yizhou Yu. Over-nav: Elevating iterative vision-and-language navigation with open-vocabulary detection and structured representation, 2024.

[36] Yunzhe Xu, Yiyuan Pan, Zhe Liu, and Hesheng Wang. Flame: Learning to navigate with multimodal llm in urban environments, 2025.

[37] Jesse Atuhurra, Iqra Ali, Tatsuya Hiraoka, Hidetaka Kamigaito, Tomoya Iwakura, and Taro Watanabe. Constructing multilingual visual-text datasets revealing visual multilingual ability of vision language models, 2024.

[38] Jiwen Zhang, Zhongyu Wei, Jianqing Fan, and Jiajie Peng. Curriculum learning for vision-and-language navigation, 2021.

[39] Ruipu Luo, Jiwen Zhang, and Zhongyu Wei. Breaking down the task: A unit-grained hybrid training framework for vision and language decision making, 2023.

[40] Wenda Qin, Teruhisa Misu, and Derry Wijaya. Explore the potential performance of vision-and-language navigation model: a snapshot ensemble method, 2021.

[41] Artem Lykov and Dzmitry Tsetserukou. Llm-brain: Ai-driven fast generation of robot behaviour tree based on large language model, 2023.

[42] Jitesh Jain, Zhengyuan Yang, Humphrey Shi, Jianfeng Gao, and Jianwei Yang. Ola-vlm: Elevating visual perception in multimodal llms with auxiliary embedding distillation, 2024.

[43] Meera Hahn, Amit Raj, and James M. Rehg. Which way is 'right'?: Uncovering limitations of vision-and-language navigation model, 2023.

[44] Chuang Lin, Yi Jiang, Jianfei Cai, Lizhen Qu, Gholamreza Haffari, and Zehuan Yuan. Multimodal transformer with variable-length memory for vision-and-language navigation, 2022.

[45] Qi Chen, Dileepa Pitawela, Chongyang Zhao, Gengze Zhou, Hsiang-Ting Chen, and Qi Wu. Webvln: Vision-and-language navigation on websites, 2023.

[46] Jialu Li and Mohit Bansal. Improving vision-and-language navigation by generating future-view image semantics, 2023.

[47] Jialu Li, Aishwarya Padmakumar, Gaurav Sukhatme, and Mohit Bansal. Vln-video: Utilizing driving videos for outdoor vision-and-language navigation, 2024.

[48] Yuchen Zou, Yineng Chen, Zuchao Li, Lefei Zhang, and Hai Zhao. Venturing into uncharted waters: The navigation compass from transformer to mamba, 2024.

[49] Hui Lu, Hu Jian, Ronald Poppe, and Albert Ali Salah. Enhancing video transformers for action understanding with vlm-aided training, 2024.

[50] Muhammad Zubair Irshad, Niluthpol Chowdhury Mithun, Zachary Seymour, Han-Pang Chiu, Supun Samarasekera, and Rakesh Kumar. Sasra: Semantically-aware spatio-temporal reasoning agent for vision-and-language navigation in continuous environments, 2021.

[51] Junyou Zhu, Yanyuan Qiao, Siqi Zhang, Xingjian He, Qi Wu, and Jing Liu. Minivln: Efficient vision-and-language navigation by progressive knowledge distillation, 2024.

[52] Yanyuan Qiao, Zheng Yu, and Qi Wu. Vln-petl: Parameter-efficient transfer learning for vision-and-language navigation, 2023.

[53] Anna Bavaresco, Marianne de Heer Kloots, Sandro Pezzelle, and Raquel Fernández. Modelling multimodal integration in human concept processing with vision-and-language models, 2024.

[54] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm, 2024.

[55] Sivan Doveh, Shaked Perek, M. Jehanzeb Mirza, Wei Lin, Amit Alfassy, Assaf Arbelle, Shimon Ullman, and Leonid Karlinsky. Towards multimodal in-context learning for vision language models, 2024.

[56] Kairui Zhou. Accessible instruction-following agent, 2023.

[57] 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.

[58] Mengfei Du, Binhao Wu, Jiwen Zhang, Zhihao Fan, Zejun Li, Ruipu Luo, Xuanjing Huang, and Zhongyu Wei. Delan: Dual-level alignment for vision-and-language navigation by cross-modal contrastive learning, 2024.

[59] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions, 2024.

[60] Dennis Hoftijzer, Gertjan Burghouts, and Luuk Spreeuwers. Language-based augmentation to address shortcut learning in object goal navigation, 2024.

[61] Abhinav Rajvanshi, Karan Sikka, Xiao Lin, Bhoram Lee, Han-Pang Chiu, and Alvaro Velasquez. Saynav: Grounding large language models for dynamic planning to navigation in new environments, 2024.

[62] Da-Wei Zhou, Yuanhan Zhang, Yan Wang, Jingyi Ning, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Learning without forgetting for vision-language models, 2025.

[63] Wanrong Zhu, Yuankai Qi, Pradyumna Narayana, Kazoo Sone, Sugato Basu, Xin Eric Wang, Qi Wu, Miguel Eckstein, and William Yang Wang. Diagnosing vision-and-language navigation: What really matters, 2022.

[64] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024.

[65] Kohei Uehara, Nabarun Goswami, Hanqin Wang, Toshiaki Baba, Kohtaro Tanaka, Tomohiro Hashimoto, Kai Wang, Rei Ito, Takagi Naoya, Ryo Umagami, Yingyi Wen, Tanachai Anakewat, and Tatsuya Harada. Advancing large multi-modal models with explicit chain-of-reasoning and visual question generation, 2024.

[66] Allen Z. Ren, Jaden Clark, Anushri Dixit, Masha Itkina, Anirudha Majumdar, and Dorsa Sadigh. Explore until confident: Efficient exploration for embodied question answering, 2024.

[67] Tianyu Zheng, Ge Zhang, Xingwei Qu, Ming Kuang, Stephen W. Huang, and Zhaofeng He. More-3s:multimodal-based offline reinforcement learning with shared semantic spaces, 2024.

[68] Juyeon Heo, Christina Heinze-Deml, Oussama Elachqar, Shirley Ren, Udhay Nallasamy, Andy Miller, Kwan Ho Ryan Chan, and Jaya Narain. Do llms "know" internally when they follow instructions?, 2024.

[69] Yicong Hong, Zun Wang, Qi Wu, and Stephen Gould. Bridging the gap between learning in discrete and continuous environments for vision-and-language navigation, 2022.

[70] Bozhou Li, Hao Liang, Zimo Meng, and Wentao Zhang. Are bigger encoders always better in vision large models?, 2024.

[71] Yuan Liu, Le Tian, Xiao Zhou, and Jie Zhou. Rethinking overlooked aspects in vision-language models, 2024.

[72] Peihao Chen, Xinyu Sun, Hongyan Zhi, Runhao Zeng, Thomas H. Li, Gaowen Liu, Mingkui Tan, and Chuang Gan. a^2nav: Action-aware zero-shot robot navigation by exploiting vision-and-language ability of foundation models, 2023.

[73] Dong An, Yuankai Qi, Yangguang Li, Yan Huang, Liang Wang, Tieniu Tan, and Jing Shao. Bevbert: Multimodal map pre-training for language-guided navigation, 2023.

[74] Yifan Bai, Dongming Wu, Yingfei Liu, Fan Jia, Weixin Mao, Ziheng Zhang, Yucheng Zhao, Jianbing Shen, Xing Wei, Tiancai Wang, and Xiangyu Zhang. Is a 3d-tokenized llm the key to reliable autonomous driving?, 2024.

[75] Jiacui Huang, Hongtao Zhang, Mingbo Zhao, and Zhou Wu. Ivlmap: Instance-aware visual language grounding for consumer robot navigation, 2024.

[76] David DeFazio, Hrudayangam Mehta, Jeremy Blackburn, and Shiqi Zhang. Vision language models can parse floor plan maps, 2024.

[77] Yanyuan Qiao, Yuankai Qi, Zheng Yu, Jing Liu, and Qi Wu. March in chat: Interactive prompting for remote embodied referring expression, 2023.

[78] Zhenjie Yang, Xiaosong Jia, Hongyang Li, and Junchi Yan. Llm4drive: A survey of large language models for autonomous driving, 2024.

[79] Valay Bundele, Mahesh Bhupati, Biplab Banerjee, and Aditya Grover. Scaling vision-and-language navigation with offline rl, 2024.

[80] Liulong Ma, Yanjie Liu, Jiao Chen, and Dong Jin. Learning to navigate in indoor environments: from memorizing to reasoning, 2019.

[81] Muraleekrishna Gopinathan, Jumana Abu-Khalaf, David Suter, and Martin Masek. Stratxplore: Strategic novelty-seeking and instruction-aligned exploration for vision and language navigation, 2024.

[82] Kanishk Jain, Varun Chhangani, Amogh Tiwari, K. Madhava Krishna, and Vineet Gandhi. Ground then navigate: Language-guided navigation in dynamic scenes, 2022.

[83] Hu Wang, Qi Wu, and Chunhua Shen. Soft expert reward learning for vision-and-language navigation, 2020.

[84] Hao-Tien Lewis Chiang, Zhuo Xu, Zipeng Fu, Mithun George Jacob, Tingnan Zhang, Tsang-Wei Edward Lee, Wenhao Yu, Connor Schenck, David Rendleman, Dhruv Shah, Fei Xia, Jasmine Hsu, Jonathan Hoech, Pete Florence, Sean Kirmani, Sumeet Singh, Vikas Sindhwani, Carolina Parada, Chelsea Finn, Peng Xu, Sergey Levine, and Jie Tan. Mobility vla: Multimodal instruction navigation with long-context vlms and topological graphs, 2024.

[85] Davide Buoso, Luke Robinson, Giuseppe Averta, Philip Torr, Tim Franzmeyer, and Daniele De Martini. Select2plan: Training-free icl-based planning through vqa and memory retrieval, 2024.

[86] Chia-Wen Kuo, Chih-Yao Ma, Judy Hoffman, and Zsolt Kira. Structure-encoding auxiliary tasks for improved visual representation in vision-and-language navigation, 2022.

[87] 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.

[88] Yanyuan Qiao, Yuankai Qi, Yicong Hong, Zheng Yu, Peng Wang, and Qi Wu. Hop: History-and-order aware pre-training for vision-and-language navigation, 2022.

[89] Vishnu Sashank Dorbala, Gunnar Sigurdsson, Robinson Piramuthu, Jesse Thomason, and Gaurav S. Sukhatme. Clip-nav: Using clip for zero-shot vision-and-language navigation, 2022.

[90] Jonáš Kulhánek, Erik Derner, Tim de Bruin, and Robert Babuška. Vision-based navigation using deep reinforcement learning, 2019.

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

[92] Felix Yu, Zhiwei Deng, Karthik Narasimhan, and Olga Russakovsky. Take the scenic route: Improving generalization in vision-and-language navigation, 2020.

[93] Xin Qian, Ziyi Zhong, and Jieli Zhou. Multimodal machine translation with reinforcement learning, 2018.

[94] Seongjun Jeong, Gi-Cheon Kang, Seongho Choi, Joochan Kim, and Byoung-Tak Zhang. Continual vision-and-language navigation, 2024.

[95] Ta-Chung Chi, Mihail Eric, Seokhwan Kim, Minmin Shen, and Dilek Hakkani-tur. Just ask:an interactive learning framework for vision and language navigation, 2019.

[96] Yicong Hong, Cristian Rodriguez-Opazo, Yuankai Qi, Qi Wu, and Stephen Gould. Language and visual entity relationship graph for agent navigation, 2020.

[97] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Talk2nav: Long-range vision-and-language navigation with dual attention and spatial memory, 2020.

[98] Yijun Yang, Tianyi Zhou, Kanxue Li, Dapeng Tao, Lusong Li, Li Shen, Xiaodong He, Jing Jiang, and Yuhui Shi. Embodied multi-modal agent trained by an llm from a parallel textworld, 2024.

[99] Kai Huang, Boyuan Yang, and Wei Gao. Modality plug-and-play: Elastic modality adaptation in multimodal llms for embodied ai, 2023.

[100] Xinghang Li, Peiyan Li, Minghuan Liu, Dong Wang, Jirong Liu, Bingyi Kang, Xiao Ma, Tao Kong, Hanbo Zhang, and Huaping Liu. Towards generalist robot policies: What matters in building vision-language-action models, 2024.

[101] Xuesong Zhang, Yunbo Xu, Jia Li, Zhenzhen Hu, and Richnag Hong. Agent journey beyond rgb: Unveiling hybrid semantic-spatial environmental representations for vision-and-language navigation, 2024.

[102] Bowen Pan, Rameswar Panda, SouYoung Jin, Rogerio Feris, Aude Oliva, Phillip Isola, and Yoon Kim. Langnav: Language as a perceptual representation for navigation, 2024.

[103] Dong An, Hanqing Wang, Wenguan Wang, Zun Wang, Yan Huang, Keji He, and Liang Wang. Etpnav: Evolving topological planning for vision-language navigation in continuous environments, 2024.

[104] Kevin Chen, Junshen K. Chen, Jo Chuang, Marynel Vázquez, and Silvio Savarese. Topological planning with transformers for vision-and-language navigation, 2020.

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.