The Sword, Shield, and Achilles' Heel: Characterizing the Linguistic Inductive Bias of Large Language Models for Spatial Reasoning in Navigation Planning

Xudong Zhang[1] , **Jian Yang**[2*] , **Shengkai Wang**[3] , **Jiangpeng Tian**[2] , **Shaowen Chen**[1] , **Xian Wei**[1] , **Ke Li**[2] , **Xiong You**[2]
[1]East China Normal University, China
[2]Information Engineering University, China
[3]Zhengzhou University, China
71275900012@stu.ecnu.edu.cn,
{jian.yang, shengkai.wang, xian.wei}@tum.de,
{tjpeng2011, 15952613852, like19771223, youarexiong}@163.com

Abstract

Large Language Model (LLM)-based navigation systems commonly construct explicit spatial representations (e.g., topological graphs, semantic raster maps) and translate them into textual descriptions as LLMs' inputs. However, the linguistic structures of such text-based spatial representations and the choices of contextual features (e.g., topology, geometry) they contain are often treated as neutral engineering decisions rather than key factors that shape LLMs' behavior. To fill the gap, we propose a dual-interventional framework that disentangles linguistic structures from different contextual cues to evaluate the linguistic inductive bias of LLMs for navigation planning. In the framework, representation intervention varies the linguistic format and the degree of linguistic compression, clarifying when linguistic representations support or inhibit navigation planning. Context intervention, combined with contextual feature combination and conflict probing, explicitly clarifies the preferences and weaknesses of LLMs when processing different contextual cues. Experiments across diverse spatial reasoning tasks and multiple model scales reveal a consistent pattern: topological information is a sturdy shield and the backbone of robust planning; linguistic format is a double-edged sword whose effect depends on model size, task demands, and the compression level; and semantic information is a fatal Achilles' heel—incorrect semantic cues can systematically derail the planning process. Overall, our study shows that effective text-based spatial representations in LLM-based navigation should preserve topological integrity, calibrate representational compression to model capacity, and ensure semantic correctness, rather than simply adopting a single representation. Our code is publicly available at https://github.com/jonesdong150/LLM-Navigation-Inductive-Bias.

*Corresponding author

1 Introduction

Spatial intelligence is a central capability of embodied artificial intelligence, particularly in navigation tasks where agents must plan goal-directed paths rather than rely on reactive obstacle avoidance [Deitke *et al.*, 2020; Savva *et al.*, 2019]. Recently, Large Language Models (LLMs) have been adopted as high-level navigation planners, responsible for reasoning over symbolic spatial representations and generating navigation sub-goals [Wei *et al.*, 2022; Yao *et al.*, 2022; Wang *et al.*, 2023a]. Most existing systems follow a hierarchical architecture, in which LLMs infer abstract navigation decisions from explicit spatial inputs—such as topological graphs or semantic maps—while lower-level controllers execute actions [Shah *et al.*, 2023; Huang *et al.*, 2022a; Rana *et al.*, 2023].

Despite rapid progress, a fundamental question remains insufficiently explored: *How does the linguistic representation of serialized spatial inputs shape LLMs' navigation reasoning behavior?* Prior work largely treats serialization as a neutral engineering step, converting structured spatial representations into text to facilitate planning, with limited systematic analysis of how linguistic structure and contextual cues influence reasoning outcomes [Chen *et al.*, 2024].

From the perspective of LLMs, Transformer-based reasoning operates over token sequences by selectively attending to and integrating evidence across long contexts [Vaswani *et al.*, 2017]. Different linguistic organizations—such as flat descriptions, hierarchical groupings, or clustered representations—induce distinct attention patterns and evidence salience. Moreover, spatial cues differ in their *sequential compatibility*: some cues, such as topological connectivity, align naturally with sequential processing, whereas others, such as coordinate-based geometric descriptions, do not. When combined with representational choices, these differences can systematically bias navigation reasoning even under information-equivalent spatial conditions [Liu and others, 2023; Fatemi and others, 2023].

Cognitive map theory provides complementary methodological guidance. Spatial cognition research emphasizes that navigation relies on multiple spatial information channels and that hierarchical and clustered organizations are fun-

damental properties of spatial knowledge [Tolman, 1948; Montello, 1998; Montello, 2001; Montello and Sas, 2006; Hirtle and Jonides, 1985]. Rather than attributing biological cognition to LLMs, we adopt these principles strictly as a design perspective to guide controlled interventions over linguistic structure and contextual cue accessibility.

Guided by these perspectives, we propose a controlled experimental framework consisting of *representation intervention* and *context intervention* (Fig. 1). Unlike prompt engineering approaches that primarily target performance optimization [Kojima *et al.*, 2022; Zhou *et al.*, 2023], our framework enforces *information equivalence*: the underlying spatial reality is held fixed while linguistic format, compression level, contextual cue combinations, and cue conflicts are systematically manipulated. We evaluate these factors across five spatial reasoning tasks, disentangling the effects of linguistic structure from those of contextual cues and exposing models' implicit prioritization strategies and vulnerabilities.

Across experiments and model scales, we observe a consistent pattern of linguistic inductive biases. Topological information acts as a robust *shield* supporting stable planning. Linguistic format functions as a *double-edged sword*, whose benefit depends on model capacity, task demands, and compression. Semantic information emerges as an *Achilles' heel*: incorrect semantic cues can systematically derail navigation reasoning. Together, these findings highlight the importance of preserving topological integrity, calibrating representational compression, and ensuring semantic correctness in text-based spatial representations for LLM-based navigation.

2 Related Work

2.1 Explicit Spatial Representations for LLM-Based Navigation

LLMs are increasingly integrated into navigation as high-level planners, using topological graphs or semantic maps for environment serialization [Huang *et al.*, 2022b; Ahn *et al.*, 2022; Chen *et al.*, 2024; Zhang *et al.*, 2025; Zhou *et al.*, 2023]. Most prior methods adopt an engineering-oriented perspective, treating these descriptions as neutral data carriers. However, empirical findings remain mixed: linearizing complex structures—such as occupancy grids or hierarchical maps [Thrun *et al.*, 2005; Kuipers, 2000; Latombe, 1991; Liang and others, 2022; Wang *et al.*, 2023b]—into token sequences can introduce inductive biases that conflict with Transformer pre-training [Fatemi and others, 2023; Zhang and others, 2024], potentially degrading performance for larger models in complex environments [Valmeekam and others, 2023].

Unlike these performance-centric studies, we treat linguistic organization and contextual cues as explicit experimental variables. This allows us to systematically analyze how specific representational choices shape LLM-based navigation reasoning rather than merely optimizing for system-level metrics.

2.2 Cognitive Maps

Cognitive science suggests that navigation relies on multiple spatial channels (topology, geometry, landmarks) organized through hierarchical and clustered structures [Tolman, 1948; Montello, 2001; Hirtle and Jonides, 1985; Montello, 1998]. While recent AI research invokes cognitive maps as a metaphor to probe LLMs' world models [Gurnee and Tegmark, 2023; Bubeck *et al.*, 2023], most studies remain descriptive and fail to operationalize these principles for controlled analysis. Consequently, it remains unclear how specific spatial cues influence navigation reasoning.

In contrast, our work operationalizes key cognitive map principles into a dual-interventional framework. By translating these principles into manipulable variables—through representation intervention and context conflict probing—we enable a systematic analysis of how LLMs prioritize, integrate, and override spatial cues in language-mediated navigation.

3 Methodology

This section introduces a dual-intervention framework for examining reasoning behavior in LLM-based navigation planning. Rather than optimizing system performance, our goal is to characterize how LLMs respond to controlled variations in spatial representations and contextual cues. We treat *linguistic format*, *degree of linguistic compression*, *contextual cue combination*, and *contextual cue conflict* as externally controllable variables. These variables are manipulated through *representation and context interventions*, which constitute the primary interface through which serialized spatial information is supplied to the LLM reasoning engine (Figure 1).

3.1 Representation Intervention: Linguistic Format and Compression

To analyze how the *representational structure* of serialized language influences navigation reasoning, we design representation interventions along two fundamental dimensions: *linguistic format* (s) and *degree of linguistic compression* (c). This design choice is motivated by the observation that language-based representations differ primarily in (i) how spatial information is *structurally organized*, and (ii) how much information is *compressed* into a fixed-length token sequence. Moreover, the choice of linguistic format is informed by principles from cognitive map theory, where hierarchical and clustered organizations are regarded as fundamental structures for organizing spatial knowledge.

Formally, let E denote a spatial environment and $I(E)$ its ground-truth *contextual spatial cues*. A representation intervention is defined as a mapping

$$\mathcal{F}_{s,r} : I(E) \rightarrow \mathcal{T}_{s,r}, \tag{1}$$

where $s \in \{\text{Flat}, \text{Hier}, \text{Clus}\}$ specifies the linguistic format (as shown in the upper middle part of Figure 1) and $r \in \{100\%, 50\%, 25\%\}$ denotes the degree of information retention rate (100% indicates the original, uncompressed text). (as illustrated in the compression branch of Figure 1 for the Flat example).

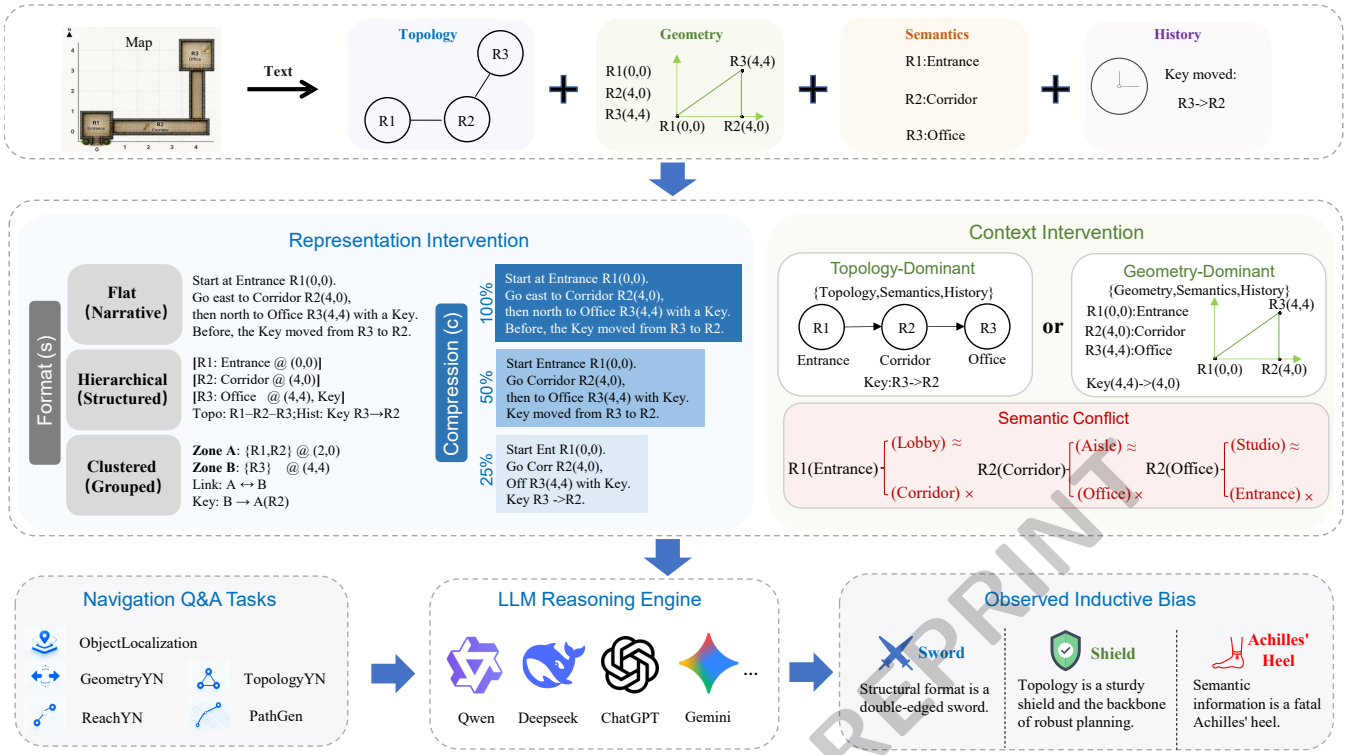

Figure 1: Unified framework overview. Representation intervention manipulates linguistic organization and compression under information-equivalent settings through Flat, Hierarchical, and Clustered formats. Context intervention controls the accessibility and consistency of Topology, Geometry, Semantics, and History cues through dominance and conflict probing. The framework reveals three characteristic inductive-bias signatures in LLM-based navigation reasoning: the Sword (structure-dependent representation effects), the Shield (topological robustness), and the Achilles' Heel (semantic vulnerability).

As illustrated in Figure 1, we enforce strict Information Equivalence: for any combination of (s, c), the resulting token sequence $\mathcal{T}_{s,c}$ encodes an identical set of *contextual spatial cues*, spanning Topology,Geometry,Semantics and History. Thus, representation intervention modifies only the *structure* and *textual density* of the language interface, while keeping the underlying spatial information invariant.

Linguistic Format Intervention (Regime R1). To isolate the independent effect of *linguistic format*, we vary the structural organization of spatial information while holding spatial complexity and contextual spatial cues constant (as illustrated in the upper part of Figure 1):

- *Flat*: A continuous natural language narrative that describes the environment in a linear prose form. This format aligns with dominant pre-training data distributions (e.g., web text and documents) of contemporary LLMs such as LLaMA [Touvron *et al.*, 2023; Grattafiori and others, 2024] and Qwen [Bai *et al.*, 2023; Yang and others, 2024].
- *Hierarchical*: A structured representation that organizes spatial information into nested levels (e.g., regions, rooms, and objects) with explicit structural markers.
- *Clustered*: A functional grouping strategy that partitions entities and locations into localized zones based on relational or functional proximity.

Crucially, all three formats contain exactly the same underlying spatial facts—namely, identical contextual spatial cues in Topology,Geometry,Semantics and History. The only difference lies in the representational structure of the interface. This controlled setup transforms serialized language into a measurable variable, enabling precise quantification of how structural organization affects LLM-based navigation planning.

Linguistic Compression Intervention (Regime R2). To examine how textual density interacts with representational structure, we manipulate the *information retention rate* ($r \in \{100\%, 50\%, 25\%\}$) within each format:

- *Full Retention (100%)*: Fully articulated descriptions with complete syntactic scaffolding.
- *Moderate Retention (50%)*: Semi-structured representations that remove redundant linguistic fillers while preserving all contextual spatial cues.
- *Low Retention (25%)*: Highly condensed symbolic encodings that maximize information density within the token sequence.

Importantly, compression intervention preserves all contextual spatial cues and modulates only the *textual density* of

these cues within a fixed attention budget, thereby enabling controlled analysis of how compression interacts with model capacity during planning.

3.2 Context Intervention: Cue Dominance and Semantic Conflict

Beyond representational form, we further decompose spatial information into a set of explicit contextual cue dimensions:

$$\mathcal{C} = \{\text{Topology}, \text{Geometry}, \text{Semantics}, \text{History}\}.$$

Here, Topology refers to node connectivity; Geometry involves coordinate sets; Semantics represents object/room labels; and History tracks the temporal sequence. As illustrated in Figure 1, context intervention selectively manipulates the availability and consistency of these dimensions to probe the model's reasoning pillars.

Contextual Cue Combination (Regime C1 Results). Drawing from the duality of spatial representations in robotics, we construct two non-conflicting regimes to test *cue sufficiency*. Both regimes provide sufficient information for navigation but rely on distinct reasoning backbones:

- *Topology-Dominant*: The model operates on an explicit connectivity graph (nodes and edges). This setup mirrors the *Spatial Semantic Hierarchy* in cognitive robotics [Kuipers, 2000] and recent LLM-based topological planners [Zhou *et al.*, 2023; Shah *et al.*, 2023], where navigation is treated as a graph search problem.
- *Geometry-Dominant*: The model operates on coordinate sets without explicit links. This mirrors metric Simultaneous Localization and Mapping (SLAM) systems [Thrun *et al.*, 2005] or vector-based neural planners [Zhang *et al.*, 2017; Parisotto and Salakhutdinov, 2017], requiring the LLM to infer adjacency and traversability implicitly via Euclidean distance calculations.

Comparing performance across these regimes reveals the model's intrinsic *inductive bias*: does the LLM reason more effectively over explicit graph structures (Topo) or implicit coordinate spaces (Geom)?

Contextual Conflict Probing (Regime C2 Results). To expose system vulnerabilities, we introduce a targeted *Semantic Conflict* probe (see the semantic-conflict branch in Figure 1). This design simulates *perceptual aliasing*, a critical failure mode in real-world embodied AI where vision systems fail to distinguish between visually similar environments [Cadena *et al.*, 2016].

We inject this conflict specifically into the Topology-Dominant regime—typically considered the robust "shield" of navigation—to conduct an adversarial stress test on this most robust component, asking whether even its topological support harbors exploitable vulnerabilities.

- *Structural Truth (The Shield)*: The topological graph remains perfect, with unique identifiers (IDs) (e.g., $ID_{R2} \neq ID_{R3}$) and correct connectivity.
- *Semantic Ambiguity (The Trap)*: We deliberately inject duplicate semantic labels for distinct nodes (e.g., labeling both the *Guest Bedroom* and *Master Bedroom* simply as "Bedroom").

This probe forces a competition between strict logic and semantic intuition. A robust reasoner should distinguish nodes by their unique IDs or topological context (History), whereas a fragile reasoner will be misled by the semantic ambiguity. We hypothesize that this semantic confusion constitutes the "Achilles' heel" of LLM spatial reasoning.

4 Experiments

To ensure that the observed behaviors reflect systematic trends rather than stochastic noise, our dual-interventional experiments (comprising R1, R2, C1, and C2) generated a total of 57,500 queries. This extensive evaluation scale provides high-fidelity insights into the pillars of navigation reasoning in LLMs.

4.1 Scene Datasets and Navigation Tasks

We evaluate our proposed interventions on three curated synthetic scene sets (Set-A, B, and C). While large-scale open-source navigation datasets such as Matterport3D or Habitat-based benchmarks offer visual diversity, they are unsuitable for our study because their annotations are typically *context-incomplete* and do not allow for the strict *variable isolation* required by our intervention axes [Cadena *et al.*, 2016; Chen *et al.*, 2024]. Specifically, existing datasets often lack synchronized and independent control over Topology, Geometry, and History, making it impossible to enforce the Information Equivalence necessary to isolate representational effects from information availability.

Our synthetic design addresses these gaps:

- *Set-A*: 10 low-complexity scenes for linguistic format intervention (R1).
- *Set-B*: 10 scenes with complexity gradients (G_1–G_5), used for linguistic compression (R2).
- *Set-C*: 10 scenes specifically designed for context interventions (C1: Dominance and C2: Conflict).

The experimental regimes and the detailed complexity schedule for Set-B are summarized in Table 1. Across all sets, we evaluate five tasks spanning core spatial cognitive requirements: (1) ObjectLocation: Local entity grounding [Huang *et al.*, 2022a]; (2) GeometryYN: Validating spatial relations (e.g., "left/right", "A is north of B") through spatial reasoning [Shah *et al.*, 2023]; (3) TopologyYN: Verifying connectivity from linear text [Wang *et al.*, 2023b]; (4) ReachYN: Multi-hop reachability analysis; (5) PathGen: Integrated sequential planning [Ahn *et al.*, 2022].

4.2 Experimental Models

Our model selection spans a range of reasoning capabilities and deployment constraints, covering both open and closed models. This tiered strategy is crucial for characterizing how linguistic inductive biases evolve with model capacity [Kaplan *et al.*, 2020], and is also motivated by practical deployment considerations, offering actionable guidance for embodied systems.

- *Deployment-Friendly (Small)*: Ultra-lightweight models ($\leq$4B) optimized for on-device execution, including Qwen3 (0.6B, 1.7B, 4B) and Llama-3.2 (1B, 3B).

Regime	Scene Set	Complexity	Variables
R1	Set-A	Fixed (Low)	Format (s)
R2	Set-B	G_1–G_5	$s \times c$
C1	Set-C	Fixed (Med)	Cue mix
C2	Set-C	Fixed (Med)	Semantic conflict

Set-B Complexity Details (G_1–G_5)

Grad	#Rms	#Objs	Hist.	Cognitive Load
G_1	6	4	2–3	Basic cognition
G_2	8	6	4–5	Basic cognition
G_3	10	8	6–7	Structured reasoning
G_4	12	10	8–9	Structured reasoning
G_5	14	12	10–11	Extreme reasoning

Table 1: Experimental Overview. (Top) Intervention regimes and dataset assignments. All regimes are evaluated across 5 tasks. (Bottom) Detailed complexity schedule for Set-B used in Regime R2.

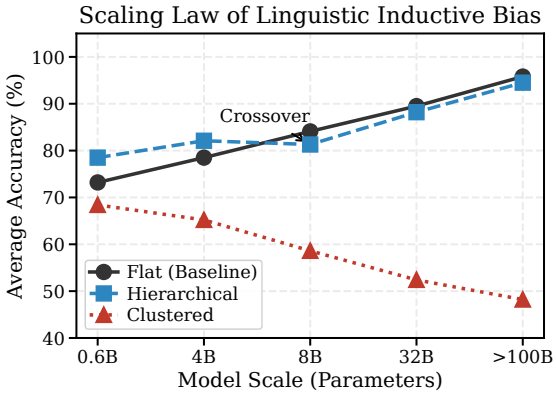

Figure 2: Scaling behavior of linguistic inductive bias (Set-A). A crossover occurs across model scales: small models favor hierarchical formats, while larger models prefer flat descriptions.

- *Deployment-Feasible (Medium)*: Models that typically require high-performance edge hardware or lightweight cloud instances, including Llama-3.1-8B, Qwen3-14B, and Qwen3-32B.
- *High-Intelligence (Large)*: Application Programming Interface (API)-accessible frontier models as capability ceilings, including ChatGPT-5.2 and Gemini-2.5. All experiments use a temperature of 0.0 for deterministic reproducibility.

4.3 Representation Intervention: Linguistic Format and Compression

Following the representation intervention defined in §3.1, we evaluate how varying linguistic format and compression modulates navigation reasoning under controlled, information-equivalent conditions.

Structural Inductive Effects (R1 Results). On Set-A, the linguistic format intervention demonstrates that representational structure functions as a *conditional inductive bias* that scales with model capacity (Figure 2). For deployment-friendly models (≤4B), the `Hierarchical` format consistently outperforms the `Flat` baseline. From a model-mechanism perspective, explicit hierarchical markers pre-organize spatial facts into attention-aligned blocks, reducing the burden on limited-capacity Transformers to infer long-range relational structure and enabling more stable evidence aggregation during navigation planning.

As the parameter scale increases, this structural advantage diminishes and a distinct *crossover* point emerges within the deployment-feasible model interval. At this stage, the `Flat` format begins to match or surpass the `Hierarchical` format. This suggests that larger models possess the emergent capacity to internally reconstruct spatial relations, rendering rigid structural scaffolds unnecessary or even counterproductive due to inductive interference [Fatemi and others, 2023]. At the high-parameter ceiling (e.g., frontier LLMs), the preference for `Flat` descriptions becomes more pronounced.

These findings offer clear engineering guidance: for edge-side deployment where small models are preferred, hierarchical structuring provides essential "scaffolding" for navigation; whereas for cloud-based large models, standard flat descriptions are more effective and avoid structural constraints.

In contrast, the `Clustered` format remains consistently inferior across the entire parameter spectrum. This persistent degradation aligns with a *fragmentation effect* [Liu and others, 2023]: partitioning spatial information into loosely connected clusters impairs the model's ability to integrate evidence for navigation planning. This leads to a practical engineering insight: designers should avoid adopting clustered structures when representing spatial environments for LLM-based navigation planning. By holding spatial complexity and contextual spatial cues constant, R1 isolates linguistic structure as a key variable, revealing how structural biases interact with Transformer scale to shape navigation behavior.

Complexity-Dependent Structural Reversal (R2 Results). R2 complements R1 by introducing a spatial complexity gradient (G_1–G_5) and controlled linguistic compression to probe the operational boundaries of structural inductive biases (Figure 3). Note that based on R1 results, the `Clustered` format is excluded from R2's boundary analysis due to its consistent inferiority. Unless otherwise specified, results report mean accuracy aggregated across scenes and models within corresponding tiers.

(1) Scaffolding-to-Overhead Reversal: Figure 3(1) plots the relative gain $\Delta Acc = Acc_{\texttt{Hier}} - Acc_{\texttt{Flat}}$. For deployment-friendly models (≤4B), ΔAcc is positive at low complexity (G_1–G_2) but reverses sharply beyond G_3, yielding a substantial negative margin at G_5. Mechanistically, this reflects a shift from *structural help* to *structural overhead*: at low spatial load, hierarchical markers align attention to local blocks; as complexity increases, nested structures expand token length and introduce formatting burden, competing with evidence aggregation under a fixed attention budget. In contrast, deployment-feasible and closed-source models (≥4B) show near-zero mean ΔAcc across all levels, indicating they can recover relational structure from natural text more robustly.

Engineering Guidance: For edge deployment with limited-capacity models, use hierarchical formatting only for simple

environments; for complex spatial tasks, switch to flat descriptions to minimize structural overhead.

(2) Dual Impact of Linguistic Compression: Figure 3(2) reports compression resilience across complexity levels. At low complexity (G_1–G_2), performance degrades moderately as compression increases, showing high redundancy tolerance. In contrast, high-complexity regimes (G_3–G_5) exhibit a *threshold collapse*, where aggressive compression (below 50%) triggers a non-linear drop in accuracy. Mechanistically, aggressive compression removes error-correcting cues essential for attention-based retrieval in complex scenes.

Engineering Guidance: While spatial descriptions can be heavily compressed for simple maps to save tokens, high-complexity environments require maintaining at least 50%–100% of the original text to preserve critical connectivity signals.

(3) Structural Selectivity across Navigation Tasks: Figure 3(3) decomposes task-wise performance at high complexity (G_5). For clarity in visualization, task names are abbreviated as follows: ObjLoc (ObjectLocation), TopoYN (TopologyYN), PathG (PathGen), ReachYN (ReachabilityYN), and GeomYN (GeometryYN).

Results show that `Hier` performs best on *TopoYN* and *ReachYN*, acting as a topological anchor that stabilizes multi-hop evidence aggregation. Conversely, `Flat` is strongest on *ObjLoc* and *GeometryYN*, where lower token overhead favors fine-grained attribute binding and local relation checks.

Engineering Guidance: The choice of representation format should be task-specific: prioritize hierarchical structures for pure connectivity or reachability tasks, and use flat descriptions for attribute-intensive or local geometric checks.

4.4 Context Intervention: Cue Dominance and Semantic Conflict

Unless otherwise specified, all metrics in this section are averaged across all models and scenes under the corresponding intervention conditions, aiming to highlight systemic trends while suppressing confounding variance. Based on the framework defined in Sec. 3.2, we analyze the model's reasoning pillars through two regimes.

Cue Dominance (Regime C1 Results). In this regime, we evaluate the model's preference between topological connectivity and geometric coordinates. We define the Dominance Score (S_{dom}) to quantify this bias:

$$S_{dom} = Acc_{topo} - Acc_{geom} \tag{2}$$

As illustrated in Figure 4, models across all tiers ($\leq$4B, 4B–32B, $\geq$100B) exhibit a consistent and significant preference for the topological regime. Specifically, S_{dom} remains stable between 0.22 and 0.32 regardless of parameter scale. From a Transformer architecture perspective, this suggests that the multi-head self-attention mechanism is more adept at capturing *discrete relational dependencies* (as represented in topological graphs) than performing *implicit numerical mapping* of continuous coordinate spaces. The topological shield acts as a structural prior that aligns with the LLM's pre-training on structured linguistic data.

Engineering Guidance: In embodied AI prompts, developers should prioritize graph-based topological connectivity over raw coordinate sets. Since LLMs' ability to utilize geometric information is significantly limited by their intrinsic inductive bias, explicit relational mapping is essential for reliable navigation.

Contextual Conflict Probing (Regime C2 Results). To expose the "Achilles' heel" of LLM spatial reasoning, we introduce semantic conflict to simulate perceptual aliasing. We define the Semantic Achilles' Heel Index (SAHI) as the relative performance loss under conflict:

$$SAHI = \frac{Acc_{shield} - Acc_{conflict}}{Acc_{shield}} \tag{3}$$

As shown in Figure 5, semantic interference effectively penetrates the topological shield, especially in mid-sized models. Tier 2 models (4B–32B) exhibit the highest vulnerability, with a SAHI score of approximately 0.24 and an average accuracy drop (ΔAcc_{Avg}) exceeding 0.21. This synchronization between relative loss (SAHI) and absolute performance collapse (ΔAcc_{Avg}) indicates a "semantic-logic decoupling". The high-dimensional embeddings of duplicate room labels (e.g., "Bedroom") generate excessive attention scores that override distinct topological tokens, leading to systemic *entity mis-binding* across all core reasoning tasks.

Engineering Guidance: For real-world deployment, semantic disambiguation is mandatory. Relying solely on topological IDs is insufficient to counteract the LLM's inherent semantic bias. Scenes with duplicate labels must be augmented with unique semantic descriptors (e.g., "North Bedroom") to prevent the navigation system from collapsing due to label confusion.

5 Conclusion

LLM-based navigation systems commonly serialize explicit spatial representations into text, yet the *linguistic structure* of this serialization and the *contextual cues* exposed to the model are often treated as neutral engineering choices. This paper challenges this assumption by arguing that serialization itself constitutes a modeling decision that actively shapes attention, evidence weighting, and failure modes in Transformer-based planning.

To make this interface analyzable, we introduce a dual-interventional framework that disentangles (i) representation intervention over *linguistic format* and *degree of compression* under strict information equivalence, and (ii) context intervention that controls cue availability and injects cue conflicts to reveal planning-relevant preferences and vulnerabilities.

Across tasks and model scales, we identify a consistent pattern of linguistic inductive bias captured by three metaphors. First, the Sword: explicit linguistic structure acts as a conditional scaffold, benefiting deployment-friendly models but exhibiting scale-dependent reversals and overhead under high compression or spatial complexity. Second, the Shield: topological connectivity provides the most stable backbone for navigation planning, dominating geometry-based cues even when both are sufficient. Third, the Achilles' Heel: semantic ambiguity can override otherwise-correct structural evidence, penetrating topological robustness and systematically derailing global route construction.

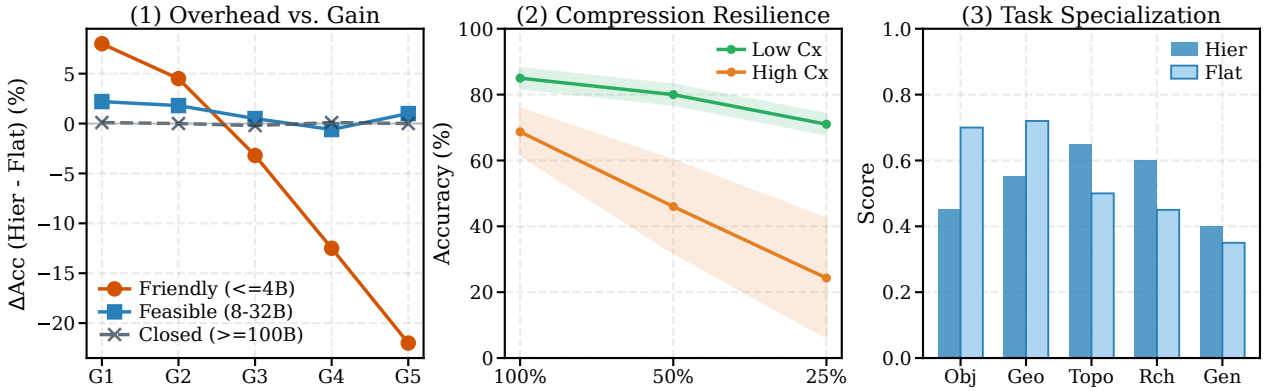

Figure 3: Complexity-dependent boundaries and task-specific utility (R2). (1) Inductive gain ΔAcc reveals a scaffolding-to-overhead reversal for small models. (2) Compression resilience shows a clear "threshold collapse" in high-complexity regimes. (3) Task-wise accuracy at G_5 highlights functional specialization between formats.

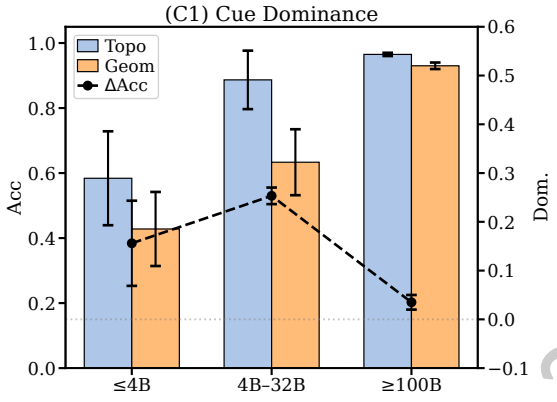

Figure 4: Cue dominance under sufficient information (C1, tier-wise). Bars compare Topo vs. Geom accuracy; dashed line shows dominance score (ΔAcc).

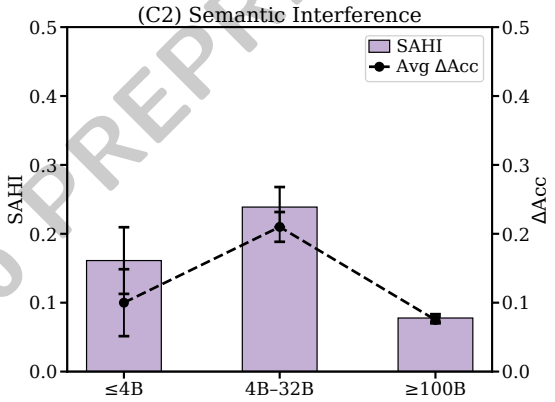

Figure 5: Semantic vulnerability under cue conflict (C2, tier-wise). Bars: SAHI; dashed line: mean performance drop (ΔAcc).

These findings translate into concrete engineering guidance for LLM-based navigation: preserve topological integrity as a first-class constraint, calibrate representational compression to model capacity and spatial load, select linguistic format in a task- and deployment-aware manner, and treat semantic correctness and disambiguation as mandatory rather than auxiliary.

Our study is limited by its focus on controlled synthetic environments designed to ensure information equivalence. An important direction for future work is to extend the proposed intervention framework to real-world benchmarks with perceptual noise and incomplete annotations, testing whether the same inductive-bias signatures persist under realistic embodied settings.

Acknowledgements

We gratefully acknowledge financial support from the National Natural Science Foundation of China (Grant No.42130112, No.42371479); the Ministry of Industry and Information Technology of China through the National Science and Technology Major Project (Grant No.2025ZD1606804, No.2025ZD1601304); the Shanghai Natural Science Foundation (General Program, Grant No.24ZR1419800, No.23ZR1419300); the Science and Technology Commission of Shanghai Municipality (Grant No.22DZ2229004); and the Shanghai Frontiers Science Center of Molecule Intelligent Syntheses.

References

[Ahn *et al.*, 2022] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

[Bai *et al.*, 2023] Jinze Bai, Shuai Bai, Yunfei Chu, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[Bubeck *et al.*, 2023] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, 2023.

[Cadena *et al.*, 2016] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J. Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 32(6):1309–1332, 2016.

[Chen *et al.*, 2024] Jiaqi Chen, Bingqian Lin, Ran Xu, Zhenhua Chai, Xiaodan Liang, and Kwan-Yee K. Wong. Mapgpt: Map-guided prompting with adaptive path planning for vision-and-language navigation. *arXiv preprint arXiv:2401.07314*, 2024. Also appeared in ACL 2024.

[Deitke *et al.*, 2020] Matt Deitke, Winson Han, Álvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, Luca Weihs, Mark Yatskar, and Ali Farhadi. Robothor: An open simulation-to-real embodied AI platform. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, 2020. IEEE/CVF.

[Fatemi and others, 2023] Bahare Fatemi et al. Talk like a graph: Encoding graphs for large language models. *arXiv preprint arXiv:2310.04560*, 2023.

[Grattafiori and others, 2024] Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[Gurnee and Tegmark, 2023] Wes Gurnee and Max Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

[Hirtle and Jonides, 1985] Stephen C. Hirtle and John Jonides. Evidence of hierarchies in cognitive maps. *Memory & Cognition*, 13(3):208–217, 1985.

[Huang *et al.*, 2022a] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. *arXiv preprint arXiv:2210.05714*, 2022. Also appeared in ICRA 2023.

[Huang *et al.*, 2022b] Wenlong Huang, Pieter Abbeel, Deepak Pathak, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022.

[Kaplan *et al.*, 2020] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[Kojima *et al.*, 2022] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022.

[Kuipers, 2000] Benjamin Kuipers. The spatial semantic hierarchy. *Artificial Intelligence*, 119(1–2):191–233, 2000.

[Latombe, 1991] Jean-Claude Latombe. *Robot Motion Planning*. Kluwer Academic Publishers, Norwell, MA, 1991.

[Liang and others, 2022] Jacky Liang et al. Code as policies: Language model programs for embodied control. *arXiv preprint arXiv:2209.07753*, 2022.

[Liu and others, 2023] Nelson F. Liu et al. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023. Also appeared in TACL 2024.

[Montello and Sas, 2006] Daniel R. Montello and Corina Sas. Human factors of wayfinding in navigation. In *International Encyclopedia of Ergonomics and Human Factors*, pages 2003–2008. CRC Press/Taylor & Francis, 2006.

[Montello, 1998] Daniel R. Montello. A new framework for understanding the acquisition of spatial knowledge in large-scale environments. In Max J. Egenhofer and Reginald G. Golledge, editors, *Spatial and Temporal Reasoning in Geographic Information Systems*, pages 143–154. Oxford University Press, 1998.

[Montello, 2001] Daniel R. Montello. Spatial cognition. In Neil J. Smelser and Paul B. Baltes, editors, *International Encyclopedia of the Social & Behavioral Sciences*, pages 14771–14775. Elsevier, 2001. DOI:10.1016/B0-08-043076-7/02492-X.

[Parisotto and Salakhutdinov, 2017] Emilio Parisotto and Ruslan Salakhutdinov. Neural map: Structured memory for deep reinforcement learning. *arXiv preprint arXiv:1702.08360*, 2017.

[Rana *et al.*, 2023] Kanishka Rana, Jack Haviland, et al. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. *arXiv preprint arXiv:2307.06135*, 2023. Also appeared in CoRL 2023.

[Savva *et al.*, 2019] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied AI research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea, 2019. IEEE/CVF. arXiv:1904.01201.

[Shah *et al.*, 2023] Dhruv Shah, Blazej Osinski, Brian Ichter, and Sergey Levine. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Proceedings of the Conference on Robot Learning (CoRL)*, Atlanta, GA, 2023. PMLR. Also available as arXiv:2207.04429.

[Thrun *et al.*, 2005] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics*. MIT Press, Cambridge, Massachusetts, 2005.

[Tolman, 1948] Edward C. Tolman. Cognitive maps in rats and men. *Psychological Review*, 55(4):189–208, 1948.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[Valmeekam and others, 2023] Karthik Valmeekam et al. On the planning abilities of large language models (a critical investigation). In *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, 2023. Curran Associates. arXiv:2310.12397.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[Wang *et al.*, 2023a] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

[Wang *et al.*, 2023b] Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. Can language models solve graph problems in natural language? *arXiv preprint arXiv:2305.10037*, 2023.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

[Yang and others, 2024] An Yang et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[Yao *et al.*, 2022] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

[Zhang and others, 2024] Y. Zhang et al. Can LLM graph reasoning generalize beyond pattern memorization? *arXiv preprint arXiv:2406.15992*, 2024.

[Zhang *et al.*, 2017] Jingwei Zhang, Lei Tai, Joschka Boedecker, Wolfram Burgard, and Ming Liu. Neural SLAM: Learning to explore with external memory. *arXiv preprint arXiv:1706.09520*, 2017.

[Zhang *et al.*, 2025] Lingfeng Zhang, Xiaoshuai Hao, Qinwen Xu, Qiang Zhang, Xinyao Zhang, Pengwei Wang, Jing Zhang, Zhongyuan Wang, Shanghang Zhang, and Renjing Xu. Mapnav: A novel memory representation via annotated semantic maps for vlm-based vision-and-language navigation. *arXiv preprint arXiv:2502.13451*, 2025.

[Zhou *et al.*, 2023] Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. *arXiv preprint arXiv:2305.16986*, 2023.