

Rule-VLN: Bridging Perception and Compliance via Semantic Reasoning and Geometric Rectification

Jiawen Wen¹, Penglei Sun¹(✉), Wenjie Zhang¹, Suixuan Qiu², Weisheng Xu¹,
Xiaofei Yang³, and Xiaowen Chu¹(✉)

¹ The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China

{jwen341,psun012,wzhang834,wxu421}@connect.hkust-gz.edu.cn,
xwchu@hkust-gz.edu.cn

² Beijing Normal University, Beijing, China

qiusuixuan@mail.bnu.edu.cn

³ Guangzhou University, Guangzhou, China

xiaofeiyang@gzhu.edu.cn

Abstract. As embodied AI transitions to real-world deployment, the success of the Vision-and-Language Navigation (VLN) task tends to evolve from mere reachability to social compliance. However, current agents suffer from a “Goal-driven trap”, prioritizing physical geometry (“*can* I go?”) over semantic rules (“*may* I go?”), frequently overlooking subtle regulatory constraints. To bridge this gap, we establish Rule-VLN, the first large-scale urban benchmark for rule-compliant navigation. Spanning a massive 29k-node environment, it injects 177 diverse regulatory categories into 8k constrained nodes across four curriculum levels, challenging agents with fine-grained visual and behavioral constraints. We further propose the Semantic Navigation Rectification Module (SNRM), a universal, zero-shot module designed to equip pre-trained agents with safety awareness. SNRM integrates a coarse-to-fine visual perception VLM framework with an epistemic mental map for dynamic detour planning. Experiments demonstrate that while Rule-VLN challenges state-of-the-art models, SNRM significantly restores navigation capabilities, reducing CVR by 19.26% and boosting TC by 5.97%. The project page is available at <https://jeffry-wen.github.io/Rule-VLN/>.

Keywords: Vision-and-Language Navigation · Embodied Agents · Object Insertion

1 Introduction

Vision-and-Language Navigation (VLN) [5] models have demonstrated superior performance across a variety of tasks. Moreover, the emergence of Multimodal Large Models (MLMs) [11] has injected new momentum into the VLN field.

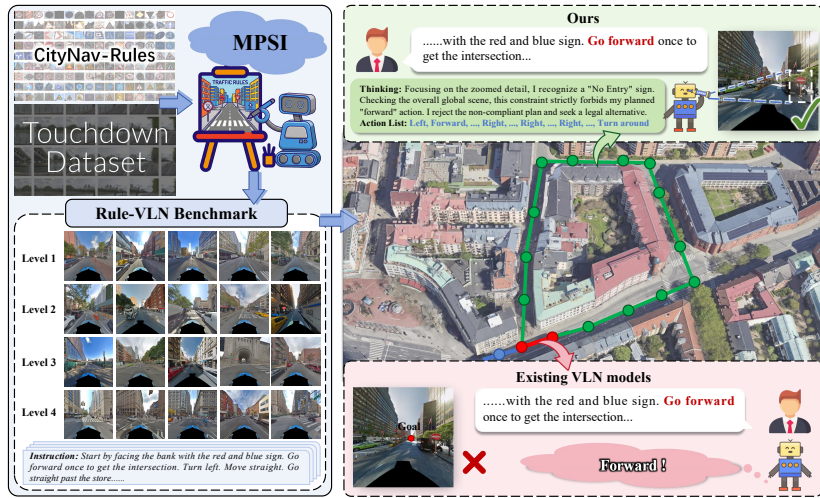


Fig. 1: The Rule-VLN Paradigm. **Left:** Benchmark construction via MPSI pipeline by injecting semantic constraints into urban topologies. **Right:** Unlike standard agents (bottom) violating “No Entry” signs, our method (top) helps the agent detect prohibitions, prunes illegal actions, and executes compliant detours (green path).

By grounding natural language instructions into visual observations, state-of-the-art agents [5, 8, 9, 37, 43, 45, 53, 58] demonstrate remarkable proficiency in goal-oriented planning. However, as embodied AI transitions from simulated testbeds to real-world deployment, the definition of navigation success must evolve beyond mere reachability to encompass social compliance and safety. In complex urban scenarios, valid trajectories are governed strictly by semantic rules (i.e., *may I go?*) rather than just physical geometry (i.e., *can I go?*) [24, 51]. Although physically traversable, a road may be semantically forbidden due to regulatory signage (e.g., “No Entry”); disregarding such constraints can lead to critical safety hazards. Therefore, enabling agents to perceive and follow these rule-based constraints is not just an improvement but a critical requirement for safe urban navigation and human-computer interaction [19].

Despite obvious progress in visual reasoning [27], current VLN models exhibit a critical compliance deficit. Existing works primarily focus on optimizing shortest-path efficiency or maximizing exploration coverage [3, 13], often neglecting the integration of high-level regulatory signals. As shown in Figure 1, standard agents tend to fall into a “Goal-driven trap”, where they follow clear paths but ignore critical semantic restrictions. Existing models over-rely on salient geometry and fail to ground small-scale semantic cues effectively, preventing them from overriding geometric priors [31, 54]. Furthermore, the scarcity of diverse, safety-critical training data in existing datasets [8, 38, 44] exacerbates this alignment problem.

To bridge the gap between idealized paths and real-world constraints, we build upon Touchdown [8] to introduce Rule-VLN. Compared to the existing VLN task, our dataset explicitly incorporates traffic signs, which serve as important navigational cues despite their small visual footprint. Spanning a massive 29,000-node urban environment, Rule-VLN injects 177 diverse regulatory categories into 8,180 specific nodes systematically organized across four curriculum levels. To ensure these generated constraints are visually and geometrically realistic, we employ dynamic graph modification alongside a Mask-Prioritized Semantic Injection (MPSI) pipeline. Evaluations demonstrate that these non-physical obstacles significantly impede the navigation performance of current SOTA models. To achieve the ability of rule navigation, we propose the Semantic Navigation Rectification Module (SNRM), a plug-and-play, zero-shot module that integrates macro-to-micro visual reasoning with local dynamic mapping navigation. This training-free approach enables generic VLN models to adhere to rules, maximally increasing Task Completion (TC) by up to 5.97% and reducing Constraint Violation Rates (CVR) by 19.26%.

Our contributions are summarized as follows: (1) We propose Rule-VLN, the first large-scale benchmark for rule-based urban navigation, presenting a rigorous challenge through high-quality semantic injection and dynamic graph modification. (2) We introduce SNRM, a universal, zero-shot module that equips pre-trained agents with rule compliance by effectively bridging visual perception with topological planning. (3) Extensive experiments validate the significant difficulties posed by Rule-VLN and demonstrate that SNRM effectively restores navigation capabilities and safety in constrained environments.

2 Related Work

2.1 Vision-Language Navigation (VLN)

VLN requires agents to ground natural language into long-horizon actions. In outdoor settings, benchmarks like Touchdown [8] and map2seq [38] introduce complex graph-structured environments. Recently, foundation models have reshaped the field. NaviLLM [62] introduces schema-based instruction tuning to enhance human-like reasoning, while MapGPT [9] employs map-guided prompting for adaptive global planning. NavGPT-2 [63] enables large language models to have visual navigation capabilities. VELMA [40] verbalizes observations for reasoning, VLN-Video [30] exploits temporal video cues, and FLAME [55] leverages synthetic data for MLLM adaptation. NavAgent [34] performs navigation tasks by integrating multi-scale environmental contexts, while MMCNav [61] accomplishes its goals via a multi-agent collaborative framework. Despite these advances, general VLMs prioritize navigational efficiency over constraint adherence, creating a granularity gap by favoring salient geometry over subtle rule signals. While retraining [14] can address this, it requires high training costs and lowers overall navigation performance. Consequently, developing a plug-and-play module has become crucial.

2.2 Rule-Compliant and Safe Navigation

As VLN advances toward real-world deployment, safety has transitioned from an implicit metric to an explicit objective. Existing approaches primarily address safety through two lenses: physical traversability and social compliance. Methods like Safe-VLN [57] and others [25, 28] focus on collision avoidance by predicting navigable areas, while VLM-Social-Nav [41] incorporates social norms by treating them as soft cost scores during planning. In terms of constraint modeling, recent works like CA-Nav [10] and GC-VLN [56] decompose instructions into subgoals to guide heuristic search. However, a critical limitation remains: these systems typically model constraints as optimization objectives (soft rewards) rather than strict prohibitions. They struggle to handle semantic safety scenarios where a path is geometrically traversable but logically forbidden (e.g., “No Entry”). To address this, our SNRM framework introduces an Epistemic Mental Map that treats semantic rules as hard constraints. Unlike soft-penalty approaches, SNRM enables zero-shot, plug-and-play topological rectification, strictly enforcing rule adherence without altering the navigation backbone.

2.3 Data Synthesis for Embodied AI

To address the high annotation costs and long-tail disturbances in embodied AI, mechanism-oriented data synthesis has become a pivotal strategy. In VLN, specialized benchmarks like R2R-UNO [21], VLN-ChEnv [33], and RAM [50] target specific failure modes ranging from path obstructions to environmental changes. General-purpose 2D editing has progressed with AnyDoor [12] and SmartEdit [23]. However, applying generic synthesizers to Rule-VLN presents two distinct challenges. First, standard models operating in perspective space typically ignore the equirectangular distortion inherent in panoramic views. Second, they often fail to generate verifiable regulatory signals. This lack of reliability stems from weak spatial-relation consistency [18, 22], severe text unreadability [35, 60], and attribute-binding violations [26]. Our MPSI pipeline addresses these limitations via a dual-mask conditioning strategy with explicit panoramic projection, ensuring injected rules are geometrically rectified and semantically legible.

3 Rule-VLN Benchmark

3.1 Task Formulation

In Rule-VLN, the agent is tasked with navigating a discrete environment $G = (V, E_{\text{geo}})$ following a natural language instruction $X = (x_1, \dots, x_L)$, where V denotes navigable nodes and E_{geo} represents intrinsic geometric connectivity. At each step t , the agent at node v_t perceives a panoramic observation $O_t = \{o_{t,k}\}_{k=1}^K$, consisting of K discrete visual slices. The navigation policy selects a neighbor $v_{\text{next}} \in \mathcal{N}(v_t)$ by identifying the slice $o_{t,k}$ that visually aligns with the target direction, governed by a geometric projection mapping $k = \Pi(v_t, v_{\text{next}})$.

Semantic Connectivity vs. Geometric Connectivity. Standard VLN assumes an identity mapping between geometry and traversability, *i.e.*, any $e \in E_{\text{geo}}$ is navigable. Rule-VLN challenges this by introducing a Dynamic Semantic Constraint. We posit that edge traversability is conditional on rule-compliance, modeled by a binary validity mask $\mathcal{M}$:

$$\mathcal{M}(v_t, v_{next} \mid X, \mathcal{R}) = \mathbb{I}[\mathcal{C}(o_{t,\Pi}(v_t, v_{next}) \mid \mathcal{R}) = 1] \quad (1)$$

where $\mathbb{I}[\cdot]$ is the indicator function and $\mathcal{C}$ evaluates whether the visual slice corresponding to the edge $e = (v_t, v_{next})$ contains regulatory prohibitions (e.g., “Do not enter” signs) defined in the rule set $\mathcal{R}$.

Consequently, the agent operates on a Semantically Pruned Graph $G' = (V, E_{\text{sem}})$, where $E_{\text{sem}} = \{e \in E_{\text{geo}} \mid \mathcal{M}(e) = 1\}$. The objective is to find a trajectory $\tau = \langle v_0, \dots, v_n \rangle$ that reaches the target defined by X , subject to the strict constraint that $\tau \subseteq E_{\text{sem}}$. When the intended geometric path is obstructed by $\mathcal{M}(e) = 0$, the agent must suppress geometric shortest-path heuristics and infer a latent feasible detour compliant with $\mathcal{R}$.

3.2 City Navigation Rules Dataset

To bridge the perception-compliance gap, we introduce **CityNav-Rules-73K** (Figure 2a), the first large-scale rule-semantic dataset explicitly coupling visual signals with fine-grained actionable constraints. Comprising 73,937 samples across 177 categories from different regions around the world, the dataset features two core innovations designed for navigation logic rather than mere classification:

LLM-Driven Discrete Action Mapping. To translate abstract rules into rigorous control constraints, we employ GPT-5 to map each visual category to a Permissible Action Subspace $\mathcal{A}_{\text{valid}}$ derived from the global discrete action set (Straight, Left, Right, U-turn). This process converts semantic classifications into precise geometric lookup tables (e.g., “no-right-turn” yields $\mathcal{A}_{\text{valid}} = \{\text{Straight, Left, U-turn}\}$), directly linking perception to graph traversal.

Visual-Semantic Hybrid Representation. We construct a Hybrid Representation interleaving Visual Descriptors (e.g., “Red Circle”) and Semantic Imperatives (e.g., “No Entry”). This challenges agents to decouple appearance from intent—for instance, learning that a “Red Circle” implies prohibition while a “Blue Circle” implies mandatory action. This design penalizes superficial pattern matching and enforces robust causal links between visual cues and geometric navigability.

3.3 Touchdown-Semantic Constraint Benchmark Construction

To strictly evaluate rule compliance, we construct Rule-VLN by injecting semantic constraints into the Touchdown graph (Fig. 2b). We identify strategic nodes via a criticality score aggregating four metrics: (1) Degree Centrality [17] for local intersection complexity; (2) Betweenness Centrality [16] for global bottlenecks; (3) Path Dependence [2] to assess detour robustness; and (4) Path Frequency [7]

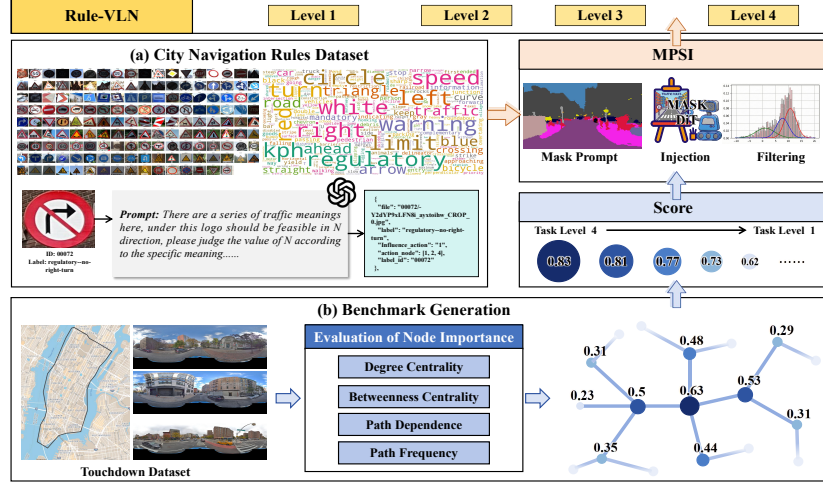


Fig. 2: Rule-VLN Construction Pipeline. (a) CityNav-Rules Dataset: Translates visual signals into permissible action constraints via LLM. (b) Benchmark Generation: Filters strategic nodes via topological metrics and injects constraints via MPSI to construct curriculum environments.

to prevent shortcut memorization. Based on this score, we filter impactful nodes to establish four curriculum difficulty levels, injected via MPSI (Sec. 4.1). As shown in Tables 1 and 4, constraints escalate from 31.44% (Level-1) to 91.13% (Level-4). Unlike limited indoor or synthetic datasets, Rule-VLN provides a real-world urban environment with 177 explicit rule types, of which a subset is injected into the node network following a 60% action-impacting rule distribution, enforcing compliant detours through progressive difficulty levels rather than simple shortest-path heuristics.

Table 1: Comparison of various Vision-and-Language Navigation benchmarks in constrained environments.

Benchmark	Scene	Data Source	Prog. Diff.	Rule Cat.	Path Optimization
ImperfectVLN [33]	Indoor	Matterport3D	✗	0	Shortest Path
DynamicVLN [46]	Urban	Synthetic Simulator	✗	<10	Shortest Path
HA-VLN 2.0 [15]	Mixed	3D Scans + Avatars	✗	0 (Implicit)	Shortest Path
Safe-VLN [57]	Indoor	Matterport3D	✗	1 (Collision)	Safe Reselection
Rule-VLN (Ours)	Urban	Real-world images	✓ (4 Levels)	177	Compliant Detour

4 Semantic Injection and Navigation Rectification Module

4.1 Mask-Prioritized Semantic Injection (MPSI)

To mitigate diffusion hallucinations, we propose MPSI (Fig. 3), a pipeline that injects node-action-compliant signs into panoramas.

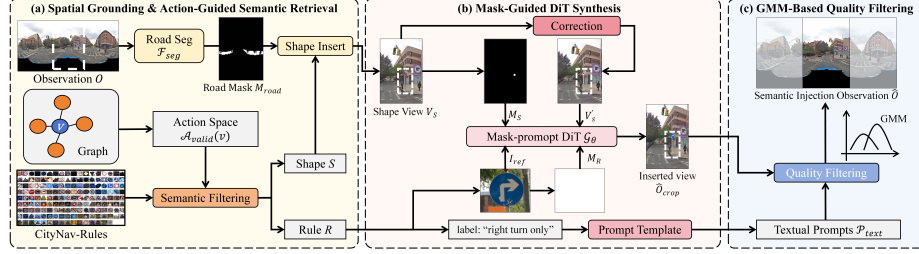


Fig. 3: MPSI Pipeline. (a) Boundary extraction via M_{road} and prior retrieval. (b) Synthesis via dual-mask-conditioned DiT. (c) GMM-based filtering and stitching.

Spatial Grounding and Rule Decoupling. We decouple regulatory signals into geometric shape S and semantic rule R . Given a target node v with permissible action subspace $\mathcal{A}_{valid}(v)$, we retrieve a semantically aligned instance $(S, R) \sim \mathcal{D}_{insert}$. To ensure global quality, we apply a cropping operator $\mathcal{C}$ to the panoramic observation $O \in \mathbb{R}^{H \times W \times 3}$, yielding $O_{crop} = \mathcal{C}(O) \in \mathbb{R}^{H_c \times W_c \times 3}$. Simultaneously, we employ a segmentation prior $\mathcal{F}_{seg}$ to extract a binary road mask $M_{road} = \mathcal{F}_{seg}(O_{crop}) \in \{0, 1\}^{H_c \times W_c}$, physically anchoring the insertion region for S before semantic injection to avoid the position offset.

Mask-Guided DiT Synthesis. Inspired by Insert Anything [42], we employ a mask prompt Diffusion Transformer (DiT) [36] $\mathcal{G}_\theta$ with dual-mask conditioning to prevent feature bleeding. First, the shape prior S is projected into O_{crop} via M_{road} to form V_S , which is optically rectified into V'_S to yield a precise boundary mask M_S . Unlike text-driven generation, we utilize a reference rule image I_{ref} for visual-semantic guidance. We concatenate the binary masks (M_S and rule-specific M_R) and spatial features V'_S with the noisy latent z_t , while injecting semantic features $\mathcal{E}_{ref} = \text{CLIP}_{img}(I_{ref})$:

$$\hat{O}_{crop} = \mathcal{G}_\theta(z_t \oplus V'_S \oplus M_S \oplus M_R, t \mid \mathcal{E}_{ref}) \quad (2)$$

where $\oplus$ denotes channel-wise concatenation and t is the diffusion timestep. This strictly confines the latent trajectory within explicit boundaries, effectively eliminating semantic hallucinations.

GMM-Based Quality Filtering and Stitching. To ensure semantic legibility, we evaluate the CLIP alignment $s_{align} = \cos(E_{img}(\hat{O}_{crop}), E_{text}(\mathcal{P}_{text}))$ between the synthesized crop and the rule category $\mathcal{P}_{text}$. Instead of rigid heuristics, we model the distribution of s_{align} using a bimodal Gaussian Mixture Model

(GMM) ($K = 2$) to adaptively prune low-confidence outliers:

$$p(s_{align}) = \sum_{k=1}^2 \pi_k \mathcal{N}(s_{align} | \mu_k, \sigma_k^2) \quad (3)$$

Candidates belonging to the high-confidence mode are accepted and seamlessly blended back into the original panorama via a stitching operator $\hat{O} = \mathcal{S}(\hat{O}_{crop}, O)$, ensuring full-scale visual fidelity and geometric consistency.

4.2 Semantic Navigation Rectification Module (SNRM)

To bridge the granularity gap, we propose SNRM, a plug-and-play, model-agnostic module that acts as a semantic reflection module for any frozen VLN policy π_{base} . SNRM decomposes compliance into a dual-stage perception framework (Fig. 4).

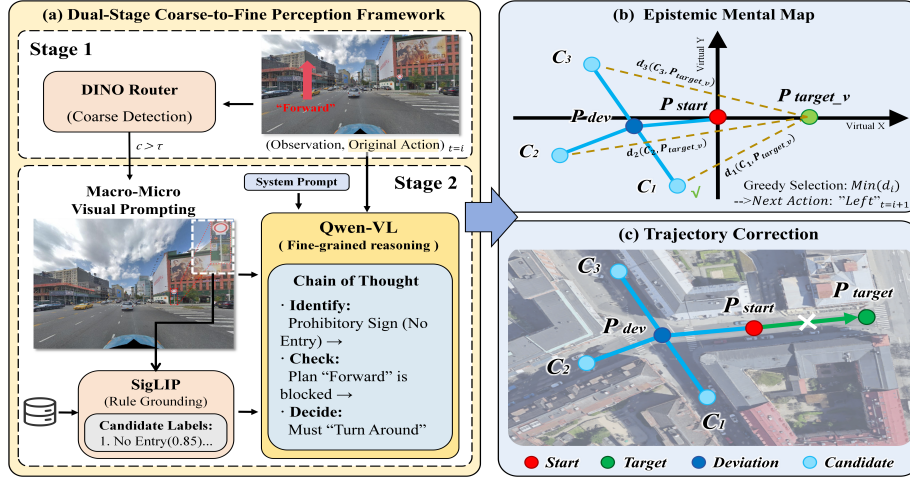


Fig. 4: The SNRM Framework. (a) Illustrating the dual-stage perception mechanism for rule grounding. (b-c) showing the local mental map for trajectory correction.

Dual-Stage Coarse-to-Fine Perception Framework. To reliably extract subtle regulatory cues from complex observations, SNRM employs a coarse-to-fine pipeline (Figure 4 a). It can be divided into two main stages. **Stage 1: Macro-Micro Visual Prompting.** Motivated by visual prompting [52], we design a human-like attention method that follows a “global-to-local” perception paradigm. To mitigate computational costs, a lightweight detector (e.g., DINO) first scans the panorama O_t for potential cues using generic prompts P_{detect} . Upon detection, we generate a Macro-Micro Visual Prompt: a global view V_{macro} with highlighted bounding boxes to preserve topological context, and a magnified crop V_{micro} to enhance fine-grained symbol resolution. By incorporating explicit visual prompts, this approach enhances the model’s perception of fine-grained

rule signals while maintaining robust global awareness. This ensures that the generated actions are both rule-compliant and consistent with environmental connectivity. **Stage 2: Knowledge-Driven Rule Grounding (KDRG).** To prevent open-ended hallucinations, we anchor VLM inference to a predefined regulatory knowledge base $\mathcal{K}$, a lightweight textual rule-name bank containing the 177 normalized CityNav-Rules categories. We employ SigLIP to compute the zero-shot similarity between V_{micro} and rule descriptions in $\mathcal{K}$, retrieving the top-ranked text prior $T_{rule} = \arg \max_{t \in \mathcal{K}} \text{Sim}(V_{micro}, t)$. Finally, the tuple $(V_{macro}, V_{micro}, T_{rule})$ and the intended action $a_{base} \sim \pi_{base}$ are fed into a VLM (e.g., Qwen-3VL). Through structured Chain-of-Thought (CoT) reasoning, the VLM outputs a safety token $s \in \{\text{Safe}, \text{<Correct Action>}\}$, strictly gating the execution of a_{base} .

Epistemic Mental Map and Trajectory Correction. Upon detecting a conflict, SNRM overrides the base policy via a dynamic 2D virtual topology (Fig. 4 b-c). Defining the conflict origin as $P_{start} = (0, 0)$, we anchor the intended target at $P_{target_v} = (\sin(\Delta\theta_{exp}), \cos(\Delta\theta_{exp}))$ based on the expected heading $\Delta\theta_{exp}$. As the agent executes forced compliant actions to a deviation point P_{dev} (updated via dead reckoning), SNRM replans by evaluating candidates $\mathcal{C} = \{C_1, \dots, C_n\}$ through a penalized greedy heuristic:

$$C_{best} = \arg \min_{C_i \in \mathcal{C}} \left(\|P(C_i) - P_{target_v}\|_2 + \lambda \cdot \mathbb{I}_{backtrack}(P(C_i)) \right) \quad (4)$$

where $P(C_i)$ is the predicted coordinate, and $\mathbb{I}_{backtrack}$ applies a severe penalty λ if $P(C_i)$ falls within a critical radius of visited nodes. To further robustify against spatial drift, we employ a visual memory buffer: if the cosine similarity between current and start node features $\cos(f_{curr}, f_{start}) > \tau_{sim}$, a closed-loop trap is detected, triggering a forced detour correction.

5 Experiment

5.1 Experimental Setup

Evaluation Metrics. We evaluate navigation via Task Completion (TC) [8], Shortest-Path Distance (SPD) [8] and Success weighted by Path Length (SPL) [4]. To rigorously quantify rule adherence, we introduce the Constraint Violation Rate (CVR), which normalizes the number of violations V_i by the agent’s actual exposure to active constraints ($\mathcal{R}_{neighbor}$):

$$\text{CVR} = \frac{1}{N} \sum_{i=1}^N \frac{V_i}{\sum_{t=1}^{L_i} \mathbb{I}(\text{Node}_t \in \mathcal{R}_{neighbor})} \quad (5)$$

where L_i is path length and the indicator function $\mathbb{I}[\cdot]$ identifies steps with proximal regulatory signals. Synthesis quality is measured via PSNR [49], SSIM [49], LPIPS [59], and FID [20].

Implementation Details. We benchmark against ORAR [39], Loc4plan [47], VELMA [40], and FLAME [55] using official hyperparameters. To ensure rule

visibility, we retrained models that take partial views as input, such as FLAME, by turning the FOV from 60° to 120° and fine-tuning. The MPSI was performed on a single NVIDIA RTX 4090 GPU. Training model is conducted on $4 \times$ NVIDIA H100 GPUs, while inference and SNRM deployment utilize a single H100. Notably, our proposed SNRM is training-free.

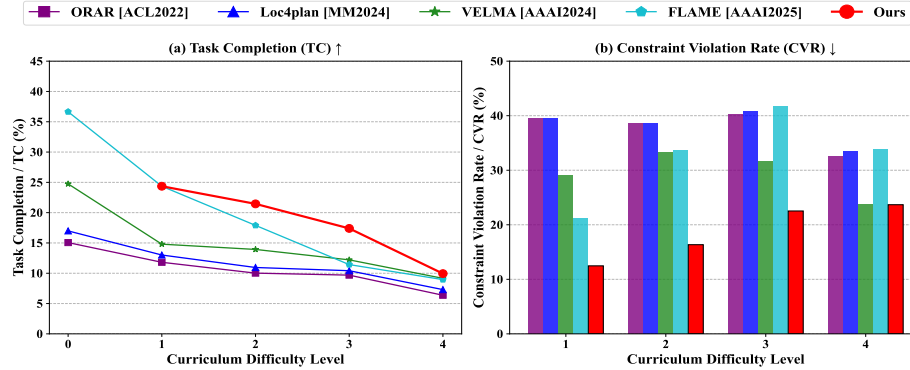


Fig. 5: Performance metrics of SOTA models on the Rule-VLN benchmark.

5.2 Evaluating the Robustness of Current VLN Methods in Rule-VLN

Analysis of Evaluation Results. The algorithmic injection of regulatory symbols establishes a rigorous four-level curriculum (Table 4). To establish baselines, we evaluate state-of-the-art (SOTA) VLN architectures. As illustrated in Fig. 5, the introduction of explicit semantic constraints induces a catastrophic performance collapse across all paradigms, compared to an environment without constraints (Level 0). Even recent foundation models like FLAME exhibit a sharp degradation in Task Completion (TC) alongside alarmingly high Constraint Violation Rates (CVR, often exceeding 30-40%).

This phenomenon highlights a severe granularity gap: current models overfit to idealized topologies and remain entirely blind to fine-grained semantic imperatives (e.g., driving into “No Entry” zones). These baseline results validate Rule-VLN as a necessary testbed and forcefully underscore the need for our proposed SNRM module, which we analyze quantitatively in Sec. 5.3.

Trajectory Visualization Analysis. Figure 6 visualizes a scenario where a prohibitory sign obstructs the path at $t = 4$. While baselines (Loc4plan, FLAME) exhibit semantic blindness by executing illegal forward actions, SNRM successfully grounds the visual cue and triggers a preemptive U-turn. Leveraging its epistemic mental map, the agent executes a legal detour to bypass the restricted zone, ultimately realigning with the global instruction (“the first left”) at $t = 33$. This visually corroborates SNRM’s ability to transform blind geometric traversal into robust, rule-compliant navigation.

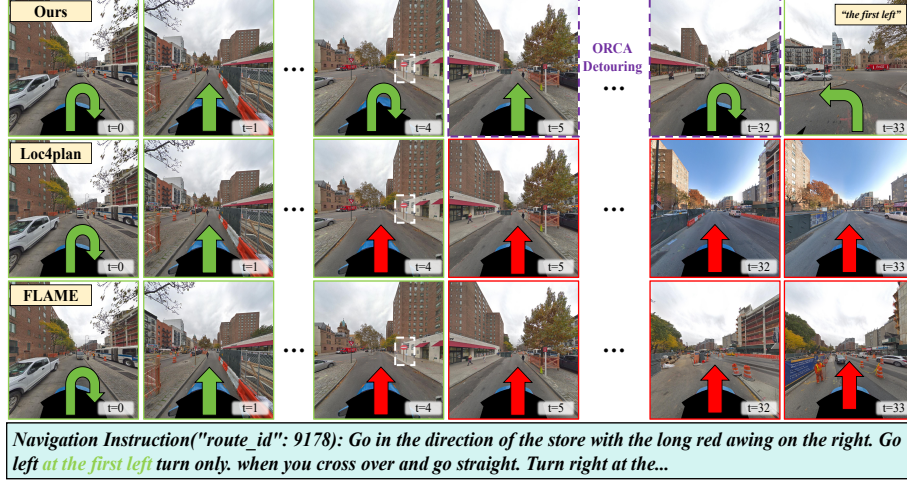


Fig. 6: Visualization results of our method and other baselines on navigation samples. Green arrows indicate strictly compliant and correct actions, while red arrows indicate semantic violations or navigation failures.

5.3 Evaluation of the SNRM

Table 2 demonstrates SNRM’s zero-shot integration with Loc4plan and FLAME across four difficulty levels. Table 3 compares the real-time inference performance after incorporating the SNRM module.

Universal Plug-and-Play Efficacy. SNRM bridges the granularity gap across diverse architectures, yielding significant gains for both traditional (Loc4plan) and MLLM-based (FLAME) agents. For Loc4plan (Level-2), SNRM not only reduces CVR by 8.70% but also improves TC by 1.53%, proving its effectiveness even without heavy reasoning priors. This impact is amplified in FLAME (Level-3), where SNRM slashes CVR by 19.26% while boosting TC by 5.97%, validating its capability to transform blind traversal into semantically constrained navigation regardless of the backbone. Although SNRM inevitably leads to more detours during navigation, it still improves SPL in most experimental settings.

Robustness and Compliance. In Level-1, despite identical TC (24.36%), SNRM improves efficiency (SPD: 16.58 vs. 17.12) and strictly enforces compliance (CVR: 12.47% vs. 21.17%). As constraints intensify in Level-3, the baseline collapses (TC 11.43%, CVR 41.79%), whereas SNRM exhibits superior robustness, recovering TC to 17.40% while suppressing violations. This advantage persists in Level-4, validating SNRM as a reliable safety envelope for complex navigation.

Inference latency and navigation performance. As shown in Table 3, incorporating the VLM introduces additional inference latency, but the event-triggered design of SNRM limits this overhead by relying on the base navigation model in most cases. For SNRM-8B, the trigger rate is only about 26%; taking FLAME as an example, the average per-step latency increases from 0.095 s to

1.158 s, while the CVR decreases by 18.25 percentage points. To further improve practical efficiency, we design a 4B SNRM variant without CoT, which adds only 0.224 s latency per step while still reducing CVR by 10.55 percentage points.

Table 2: Performance comparison across different tasks on the Rule-VLN benchmark. The performance differences brought by integrating SNRM are indicated in subscripts ($\uparrow$ for increases and $\downarrow$ for decreases). Best results within each category are highlighted in **bold**.

Level	Metric	Traditional Models			VLM/LLM-based Models		
		ORAR [39]	Loc4plan [47]	+SNRM	VELMA [40]	FLAME [55]	+SNRM
Level-1	TC $\uparrow$	11.81	13.01	14.07 $\uparrow_{1.06}$	14.80	24.36	24.36 $_{-0.00}$
	SPL $\uparrow$	7.42	7.72	7.76 $\uparrow_{0.04}$	9.87	18.83	17.72 $\downarrow_{1.11}$
	SPD $\downarrow$	21.72	21.81	21.31 $\downarrow_{0.50}$	18.52	17.12	16.58 $\downarrow_{0.54}$
	CVR $\downarrow$	39.48	39.45	36.48 $\downarrow_{2.97}$	29.01	21.17	12.47 $\downarrow_{8.70}$
Level-2	TC $\uparrow$	10.02	10.95	12.48 $\uparrow_{1.53}$	13.93	17.90	21.45 $\uparrow_{3.55}$
	SPL $\uparrow$	6.63	6.78	6.98 $\uparrow_{0.20}$	9.59	13.94	13.91 $\downarrow_{0.03}$
	SPD $\downarrow$	23.04	23.54	23.03 $\downarrow_{0.51}$	19.06	20.04	18.30 $\downarrow_{1.74}$
	CVR $\downarrow$	38.55	38.67	29.97 $\downarrow_{8.70}$	33.30	33.68	16.35 $\downarrow_{17.33}$
Level-3	TC $\uparrow$	9.69	10.42	12.61 $\uparrow_{2.19}$	12.21	11.43	17.40 $\uparrow_{5.97}$
	SPL $\uparrow$	6.54	6.43	7.16 $\uparrow_{0.73}$	7.84	9.03	10.51 $\uparrow_{1.48}$
	SPD $\downarrow$	22.71	23.39	22.23 $\downarrow_{1.16}$	20.08	23.44	19.96 $\downarrow_{3.48}$
	CVR $\downarrow$	40.22	40.80	33.64 $\downarrow_{7.16}$	31.69	41.79	22.53 $\downarrow_{19.26}$
Level-4	TC $\uparrow$	6.37	7.30	7.96 $\uparrow_{0.66}$	9.16	8.95	9.94 $\uparrow_{0.99}$
	SPL $\uparrow$	4.10	4.43	4.54 $\uparrow_{0.11}$	5.87	6.24	6.39 $\uparrow_{0.15}$
	SPD $\downarrow$	23.95	24.50	23.63 $\downarrow_{0.87}$	21.18	26.29	25.61 $\downarrow_{0.68}$
	CVR $\downarrow$	32.50	33.42	27.00 $\downarrow_{6.42}$	23.77	33.89	23.68 $\downarrow_{10.21}$

5.4 Ablation experiments and analysis of SNRM

To dissect the contribution of each module within SNRM, we conduct a comprehensive ablation study on the most challenging Level-4 configuration (Table 5). The baseline setup (#1), devoid of SNRM’s components, exhibits poor navigation success (TC: 8.95) and a high rule violation rate (CVR: 33.89%), acting as a naive geometric navigator.

Impact of the Dual-Stage Perception Framework. The removal of either Macro-Micro Visual Prompting(MMVP) (Setup #3) or Knowledge-Driven Rule Grounding(KDRG) (Setup #2) explicitly degrades performance. Without MMVP (#3), the VLM struggles to simultaneously focus on subtle regulatory symbols and the global topological context, leading to an increased CVR (26.86%). Similarly, removing the KDRG module (#2) deprives the VLM of deterministic text priors, causing it to occasionally hallucinate or misclassify long-tail signs, which marginally drops the TC. Setup #5 demonstrates that these two perception modules act synergistically to reliably translate raw visual cues into actionable semantic constraints.

Impact of Topological Recovery. Setup #4 (w/o Mental Map) achieves the lowest CVR (21.28%) but compromises TC (6.53%), as the agent stops at constraints without a path-generation mechanism. Conversely, the full framework (#5) improves TC to 9.94% and SPD to 25.61, demonstrating that the Epis-

Table 3: Efficiency and performance comparison.

Method	CVR↓	Trigger Rate↓	Avg. Inf. Time/Step↓	Rule-triggered Step Time↓	No-rule Step Time↓
Loc4plan-11M	31.40	–	0.002 s	–	–
+SNRM-8B (CoT)	21.00	26.36%	1.132 s	1.541 s	0.148 s
FLAME-8B	36.86	–	0.095 s	–	–
+SNRM-8B (CoT)	18.61	26.25%	1.158 s	3.852 s	0.217 s
+SNRM-4B	26.31	14.49%	0.319 s	0.941 s	0.234 s

temic Mental Map is essential for translating static rule recognition into active navigation and detour computing.

Table 4: Statistics of affected nodes and instructions across Rule-VLN levels.

Level	# Nodes	Aff. Node(%)	Aff. Instr.(%)
1	1941	9.99	31.44
2	2029	10.45	54.79
3	2163	11.14	74.52
4	2047	10.54	91.13
Total	8180	42.13	99.93

Table 5: Ablation study of SNRM components on Level-4.

Setup	SNRM Components				Metrics		
	MMVP	KDRG	Map	TC	(↑)	SPD (↓)	CVR (↓)
#1	×	×	×	8.95	26.29	33.89	
#2	✓	×	✓	9.02	25.65	25.94	
#3	×	✓	✓	8.88	25.69	26.86	
#4	✓	✓	×	6.53	26.91	21.28	
#5	✓	✓	✓	9.94	25.61	23.68	

Table 6: Synthesis quality and efficiency comparison.

Method	Fidelity		Perceptual		VRAM(G)↓
	PSNR↑	SSIM↑	LPIPS↓	FID↓	
FLUX.1	24.59	0.7518	0.3096	30.31	36.71
MPSI	32.41	0.9280	0.0503	7.72	15.12

Table 7: Performance comparison of different VLMs within SNRM on Level-4.

Method	Param	TC (↑)	SPD (↓)	CVR (↓)
FLAME + SNRM _{Qwen3-vl}	8B	9.94	25.61	23.68
FLAME + SNRM _{Qwen3-vl}	4B	8.88	25.67	24.10
FLAME + SNRM _{LLaVA}	7B	7.74	28.10	21.17
FLAME + SNRM _{Intern}	8B	9.45	25.79	23.45
FLAME + SNRM _{Pixtral}	12B	8.38	26.43	29.74

Impact of VLM Backbones. We instantiate SNRM with Qwen3-VL [6], InternVL-3.5 [48], LLaVA-1.6 [32], and Pixtral [1] to explore the influence of different VLM models on the perception of internal rules. Intra-family scaling (Qwen-VL) confirms that parameter capacity enhances spatial CoT execution. Qwen-VL and InternVL achieve the optimal balance via robust grounding. In contrast, LLaVA-1.6 is overly conservative; its low CVR (21.17%) leads to excessive detouring and navigation failure (low TC). Finally, Pixtral 12B’s underperformance against 8B models confirms that for embodied tasks, visual-spatial grounding is more critical than parameter scale.

5.5 High-Fidelity Semantic Injection via MPSI

Visual Realism and Fidelity. As shown in Table 6, MPSI significantly outperforms the FLUX.1-Fill [29] baseline. High pixel-level fidelity (SSIM 0.9280, PSNR 32.41) confirms that our dual-mask conditioning strictly confines the latent space

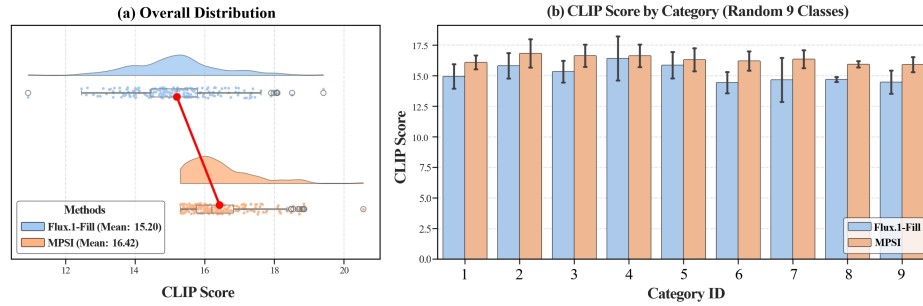


Fig. 7: Quantitative evaluation of semantic alignment using CLIP scores. (a) Overall score distribution. (b) Distribution across randomly sampled categories.

and prevents artifact bleeding into the native panorama. Furthermore, MPSI drastically improves perceptual realism (FID 7.72, LPIPS 0.0503), effectively mitigating the geometric distortions typical of generic diffusion models in localized synthesis. Due to the masking design, MPSI achieves a 58.81% reduction in GPU memory footprint compared to FLUX.1-Fill.

Semantic Alignment and Filtering. To ensure the injected signs serve as actionable constraints, we evaluate cross-modal semantic legibility via CLIP scores (Fig. 7). MPSI achieves a higher overall mean (16.42 vs. 15.20). Crucially, the raincloud plot (Fig. 7a) reveals that while FLUX exhibits a long tail of low-scoring outliers (indicating hallucinations), MPSI’s distribution is sharply truncated at the lower end. This explicitly validates our GMM-Based Quality Filtering in actively pruning corrupted generations. Moreover, MPSI demonstrates consistent superiority across diverse rule categories (Fig. 7b), guaranteeing unambiguous visual cues for downstream tasks.

Qualitative Comparison. Fig. 8 compares MPSI, FLUX.1-Fill, and Google Nano Banana 2. FLUX suffers from severe semantic hallucinations, and Nano Banana 2 yields spatially misplaced injections due to a lack of positional conditioning. MPSI consistently achieves superior visual fidelity and precise geometric grounding.

6 Limitations and Future Work

Despite its effectiveness, our framework has several limitations. On the data synthesis side, although MPSI injects rules under dual-mask constraints, the quality of the synthesized data remains affected by the complexity of the environment and the quality of rule-related images. Moreover, following the discrete-node task formulation of Touchdown may lead to viewpoint-consistency issues when the same rule is observed from adjacent nodes. For SNRM, we design its components, such as QwenVL and SigLIP, in a plug-and-play manner to enhance generalization to diverse real-world regulations. This enables SNRM to benefit from advances in foundation models without retraining the entire module, but it also introduces

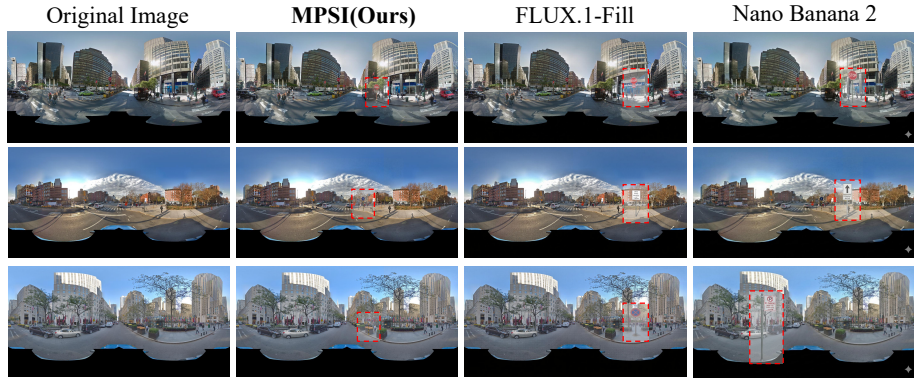


Fig. 8: Qualitative analysis of image inpainting results from MPSI, FLUX.1-Fill, and Google Nano Banana 2 across diverse urban scenarios.

additional inference latency. Future work will explore stronger generative vision models and 3D geometric constraints to narrow the gap between synthesized and real-world scenes. We will also investigate model distillation, feature caching, and quantization to reduce computational overhead while preserving generalization.

7 Conclusion

In this work, we introduce semantic rule constraints into Vision-and-Language Navigation (VLN) to address the issue where existing agents prioritize geometric accessibility over social compliance in real-world environments. We propose Rule-VLN, the first benchmark incorporating such rule compliance challenges into VLN. This is achieved via a novel Mask-Prioritized Semantic Injection (MPSI) pipeline that precisely integrates fine-grained traffic rule signals into urban topology and visual observations. Leveraging Rule-VLN, we demonstrate the limitations of current VLN methods and further introduce the Semantic Navigation Rectification Module (SNRM). This module enables agents to effectively adapt to rule-constrained environments through macro-micro visual prompting and an epistemic mental map method. Our approach achieves state-of-the-art results on Rule-VLN without requiring alterations to the original model architectures. We believe that addressing the shift from physical accessibility to social compliance is critical for the practical deployment of VLN agents and for evaluating their capacity for safe navigation under complex rule constraints.

Acknowledgements

We thank the reviewers for their valuable comments and constructive suggestions. We also gratefully acknowledge the computing resources provided by The Hong Kong University of Science and Technology (Guangzhou) through the Red Bird MPhil Program.

References

1. Agrawal, P., Antoniak, S., Hanna, E.B., Bout, B., Chaplot, D., Chudnovsky, J., Costa, D., De Monicault, B., Garg, S., Gervet, T., et al.: Pixtral 12b. arXiv preprint arXiv:2410.07073 (2024)
2. Albert, R., Jeong, H., Barabási, A.L.: Error and attack tolerance of complex networks. *nature* **406**(6794), 378–382 (2000)
3. An, D., Wang, H., Wang, W., Wang, Z., Huang, Y., He, K., Wang, L.: Etpnav: Evolving topological planning for vision-language navigation in continuous environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **47**(7), 5130–5145 (2025). <https://doi.org/10.1109/TPAMI.2024.3386695>
4. Anderson, P., Chang, A., Chaplot, D.S., Dosovitskiy, A., Gupta, S., Koltun, V., Kosecka, J., Malik, J., Mottaghi, R., Savva, M., et al.: On evaluation of embodied navigation agents. arXiv preprint arXiv:1807.06757 (2018)
5. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., Van Den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 3674–3683 (2018)
6. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
7. Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., Munos, R.: Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems* **29** (2016)
8. Chen, H., Suhr, A., Misra, D., Snively, N., Artzi, Y.: Touchdown: Natural language navigation and spatial reasoning in visual street environments. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 12538–12547 (2019)
9. Chen, J., Lin, B., Xu, R., Chai, Z., Liang, X., Wong, K.Y.: Mapgpt: Map-guided prompting with adaptive path planning for vision-and-language navigation. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 9796–9810 (2024)
10. Chen, K., An, D., Huang, Y., Xu, R., Su, Y., Ling, Y., Reid, I., Wang, L.: Constraint-aware zero-shot vision-language navigation in continuous environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025)
11. Chen, S., Wu, Z., Zhang, K., Li, C., Zhang, B., Ma, F., Yu, F.R., Li, Q.: Exploring embodied multimodal large models: Development, datasets, and future directions. *Information Fusion* **122**, 103198 (2025)
12. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 6593–6602 (2024)
13. Chowda, S.S., Alvi, R., Rahman, S.S., Rahman, M.A., Raiaan, M.A.K., Islam, M.R., Hussain, M., Azam, S.: From language to action: a review of large language models as autonomous agents and tool users. *Artificial Intelligence Review* (2026)
14. Chu, T., Zhai, Y., Yang, J., Tong, S., Xie, S., Schuurmans, D., Le, Q.V., Levine, S., Ma, Y.: Sft memorizes, rl generalizes: A comparative study of foundation model post-training. arXiv preprint arXiv:2501.17161 (2025)
15. Dong, Y., Wu, F., He, Q., Cheng, Z.Q., Li, H., Li, M., Cheng, Z., Zhou, Y., Sun, J., Dai, Q., et al.: Ha-vln 2.0: An open benchmark and leaderboard for human-aware

- navigation in discrete and continuous environments with dynamic multi-human interactions. arXiv preprint arXiv:2503.14229 (2025)
16. Freeman, L.C.: A set of measures of centrality based on betweenness. *Sociometry* pp. 35–41 (1977)
 17. Freeman, L.C.: Centrality in social networks conceptual clarification. *Social networks* **1**(3), 215–239 (1978)
 18. Gokhale, T., Palangi, H., Nushi, B., Vineet, V., Horvitz, E., Kamar, E., Baral, C., Yang, Y.: Benchmarking spatial relationships in text-to-image generation. arXiv preprint arXiv:2212.10015 (2022)
 19. Hamdani, R., Chihi, I.: Adaptive human-computer interaction for industry 5.0: A novel concept, with comprehensive review and empirical validation. *Computers in Industry* **168**, 104268 (2025)
 20. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **30** (2017)
 21. Hong, H., Wang, S., Huang, Z., Wu, Q., Liu, J.: Navigating beyond instructions: Vision-and-language navigation in obstructed environments. In: *Proceedings of the 32nd ACM International Conference on Multimedia*. pp. 7639–7648 (2024)
 22. Huang, K., Duan, C., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **47**(5), 3563–3579 (2025)
 23. Huang, Y., Xie, L., Wang, X., Yuan, Z., Cun, X., Ge, Y., Zhou, J., Dong, C., Huang, R., Zhang, R., et al.: Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8362–8371 (2024)
 24. Islam, C.M., Salman, S., Shams, M., Liu, X., Kumar, P.: Malicious path manipulations via exploitation of representation vulnerabilities of vision-language navigation systems. In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 13845–13852 (2024). <https://doi.org/10.1109/IROS58592.2024.10802618>
 25. Jeong, S., Kang, G.C., Kim, J., Zhang, B.T.: Zero-shot vision-and-language navigation with collision mitigation in continuous environment. arXiv preprint arXiv:2410.17267 (2024)
 26. Kang, W., Galim, K., Koo, H.I., Cho, N.I.: Counting guidance for high fidelity text-to-image synthesis. In: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. pp. 899–908. IEEE (2025)
 27. Ke, F., Hsu, J., Cai, Z., Ma, Z., Zheng, X., Wu, X., Huang, S., Wang, W., Haghighi, P.D., Haffari, G., et al.: Explain before you answer: A survey on compositional visual reasoning. arXiv preprint arXiv:2508.17298 (2025)
 28. Kim, J., Sim, J., Kim, W., Sycara, K., Nam, C.: Care: Enhancing safety of visual navigation through collision avoidance via repulsive estimation. arXiv preprint arXiv:2506.03834 (2025)
 29. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontekst: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
 30. Li, J., Padmakumar, A., Sukhatme, G., Bansal, M.: Vln-video: Utilizing driving videos for outdoor vision-and-language navigation. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 38, pp. 18517–18526 (2024)

31. Li, Y., Tian, M., Lin, Z., Zhu, J., Zhu, D., Liu, H., Zhang, Y., Xiong, Z., Zhao, X.: Fine-grained evaluation of large vision-language models in autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9431–9442 (October 2025)
32. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), <https://llava-vl.github.io/blog/2024-01-30-llava-next/>
33. Liu, S., Zhang, H., Qiao, Q., Wu, Q., Wang, P.: Vln-chenv: Vision-language navigation in changeable environments. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 3798–3807 (2025)
34. Liu, Y., Yao, F., Yue, Y., Xu, G., Sun, X., Fu, K.: Navagent: Multi-scale urban street view fusion for uav embodied vision-and-language navigation. arXiv preprint arXiv:2411.08579 (2024)
35. Lu, R., Wang, R., Lyu, K., Jiang, X., Huang, G., Wang, M.: Towards understanding text hallucination of diffusion models via local generation bias. In: The Thirteenth International Conference on Learning Representations (2025)
36. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
37. Qiao, Y., Liu, Q., Liu, J., Liu, J., Wu, Q.: Llm as copilot for coarse-grained vision-and-language navigation. In: European Conference on Computer Vision. pp. 459–476. Springer (2024)
38. Schumann, R., Riezler, S.: Generating landmark navigation instructions from maps as a graph-to-text problem. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 489–502 (2021)
39. Schumann, R., Riezler, S.: Analyzing generalization of vision and language navigation to unseen outdoor areas. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 7519–7532 (2022)
40. Schumann, R., Zhu, W., Feng, W., Fu, T.J., Riezler, S., Wang, W.Y.: Velma: Verbalization embodiment of llm agents for vision and language navigation in street view. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 18924–18933 (2024)
41. Song, D., Liang, J., Payandeh, A., Raj, A.H., Xiao, X., Manocha, D.: Vlm-social-nav: Socially aware robot navigation through scoring using vision-language models. IEEE Robotics and Automation Letters **10**(1), 508–515 (2024)
42. Song, W., Jiang, H., Yang, Z., Quan, R., Yang, Y.: Insert anything: Image insertion via in-context editing in dit. arXiv preprint arXiv:2504.15009 (2025)
43. Sun, P., Song, Y., Liu, X., Yang, X., Wang, Q., Li, T., Yang, Y., Chu, X.: 3d question answering for city scene understanding. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 2156–2165 (2024)
44. Sun, P., Song, Y., Zhu, X., Liu, X., Wang, Q., Liu, Y., Xia, C., Li, T., Yang, Y., Chu, X.: City-vlm: Towards multidomain perception scene understanding via multimodal incomplete learning. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 3448–3457 (2025)
45. Sun, P., Tang, S., Wen, J., Liang, Y., Yang, Y., Chu, X.: From terrain to space: A survey on multi-domain data lifecycle for urban embodied agents. Preprints (2026)
46. Sun, Y., Qiu, Y., Aoki, Y.: Dynamicvln: Incorporating dynamics into vision-and-language navigation scenarios. Sensors **25**(2), 364 (2025)

47. Tian, H., Meng, J., Zheng, W.S., Li, Y.M., Yan, J., Zhang, Y.: Loc4plan: Locating before planning for outdoor vision and language navigation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 4073–4081 (2024)
48. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
49. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13**(4), 600–612 (2004)
50. Wei, Z., Lin, B., Nie, Y., Chen, J., Ma, S., Xu, H., Liang, X.: Unseen from seen: Rewriting observation-instruction using foundation models for augmenting vision-language navigation. *IEEE Transactions on Neural Networks and Learning Systems* (2025)
51. Wen, C., Liang, J., Yuan, S., Huang, H., Bethala, G.C.R., Liu, Y.S., Wang, M., Tzes, A., Fang, Y.: How secure are large language models (llms) for navigation in urban environments? (2025), <https://arxiv.org/abs/2402.09546>
52. Wu, J., Zhang, Z., Xia, Y., Li, X., Xia, Z., Chang, A., Yu, T., Kim, S., Rossi, R.A., Zhang, R., et al.: Visual prompting in multimodal large language models: A survey. arXiv preprint arXiv:2409.15310 (2024)
53. Xiang, J., Wang, X., Wang, W.Y.: Learning to stop: A simple yet effective approach to urban vision-language navigation. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 699–707 (2020)
54. Xiao, R., Kim, S., Georgescu, M.I., Akata, Z., Alaniz, S.: Flair: Vlm with fine-grained language-informed image representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24884–24894 (June 2025)
55. Xu, Y., Pan, Y., Liu, Z., Wang, H.: Flame: Learning to navigate with multimodal llm in urban environments. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9005–9013 (2025)
56. Yin, H., Wei, H., Xu, X., Guo, W., Zhou, J., Lu, J.: Gc-vln: Instruction as graph constraints for training-free vision-and-language navigation. arXiv preprint arXiv:2509.10454 (2025)
57. Yue, L., Zhou, D., Xie, L., Zhang, F., Yan, Y., Yin, E.: Safe-vln: Collision avoidance for vision-and-language navigation of autonomous robots operating in continuous environments. *IEEE Robotics and Automation Letters* **9**(6), 4918–4925 (2024). <https://doi.org/10.1109/LRA.2024.3387171>
58. Zhang, J., Wang, K., Xu, R., Zhou, G., Hong, Y., Fang, X., Wu, Q., Zhang, Z., Wang, H.: Navid: Video-based vlm plans the next step for vision-and-language navigation. *Robotics: Science and Systems* (2024)
59. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
60. Zhang, T., Wang, X., Li, L., Tai, Z., Chi, J., Tian, J., He, H., Wang, S.: Strict: Stress-test of rendering image containing text. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 21148–21161 (2025)
61. Zhang, Z., Chen, M., Zhu, S., Han, T., Yu, Z.: Mmcnav: Mllm-empowered multi-agent collaboration for outdoor visual language navigation. In: Proceedings of the 2025 International Conference on Multimedia Retrieval. pp. 1767–1776 (2025)
62. Zheng, D., Huang, S., Zhao, L., Zhong, Y., Wang, L.: Towards learning a generalist model for embodied navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13624–13634 (June 2024)

63. Zhou, G., Hong, Y., Wang, Z., Wang, X.E., Wu, Q.: Navgpt-2: Unleashing navigational reasoning capability for large vision-language models. In: European Conference on Computer Vision. pp. 260–278. Springer (2024)