ExplaiNavility: Interpretable Vision-and-Language Navigation with Reasoning Supervision

William Raccagni[1,2], Niyati Rawal[1], Marcella Cornia[1], Lorenzo Baraldi[1], and Rita Cucchiara[1]

[1] University of Modena and Reggio Emilia, Italy
{name.surname}@unimore.it
[2] University of Pisa, Italy
{name.surname}@phd.unipi.it

Abstract. Vision-and-Language Navigation (VLN) requires embodied agents to follow natural language instructions while grounding their decisions in complex visual environments. Despite recent progress, most VLN agents remain difficult to interpret, providing limited insight into why a navigation action is selected. In this work, we introduce ExplaiNavility, an interpretable VLN framework that augments navigation with explicit reasoning supervision. Our model combines a visual scene encoder with a multimodal LLM, using structured prompts to jointly condition on the instruction, the navigation history, and the candidate viewpoints available at each step. Beyond predicting the next navigable location, the model is trained to produce human-readable explanations of its decisions and descriptive low-level action labels, such as turning left, turning right, moving forward, or stopping. To supervise these outputs, we augment R2R trajectories with speaker-style action explanations generated from consecutive observations, together with action labels obtained from the simulator. By fine-tuning the embodied multimodal agent with these auxiliary reasoning objectives, ExplaiNavility encourages more transparent decision-making while preserving strong navigation ability. Experiments on the R2R benchmark show that our approach improves over prior explainable VLN methods on unseen environments, while producing interpretable rationales that make the agent's behavior easier to inspect. These results suggest that explicit reasoning supervision can benefit both navigation performance and interpretability in embodied instruction-following agents.

Keywords: Vision-and-Language Navigation · Explainable Embodied AI · Multimodal Large Language Models

1 Introduction

Vision-and-Language Navigation (VLN) [1, 19, 46] is a central task in embodied AI, lying at the intersection of computer vision, natural language processing,

Fig. 1: Interpretable VLN enhanced by reasoning. Given panoramic observations and a language instruction, ExplaiNavility predicts the next navigation action while generating interpretable outputs. The *Speaker* provides a natural language explanation for the decision, while *Direction* describes the corresponding low-level action. The agent executes the selected action and repeats this process until reaching the goal.

and robotics. Its goal is to develop embodied agents that can perceive complex visual environments and follow natural language instructions to navigate them. This requires the integration of multimodal perception, semantic understanding, spatial reasoning, and sequential decision-making, making VLN a challenging benchmark for grounded intelligence. Beyond its academic relevance, VLN has strong potential for real-world applications, including assistive home robotics [2, 33], autonomous driving systems [23], and mobile personal assistants [13].

Early progress in VLN was driven by sequence-to-sequence navigation policies, speaker-based data augmentation, and reinforcement learning objectives designed to improve generalization to unseen environments [17, 34, 40]. More recent methods have shifted toward Transformer-based architectures [16, 37] and pretrained multimodal representations [10, 11, 21, 27], enabling stronger fusion of visual observations, language instructions, and navigation history. In this context, Large Language Models (LLMs) [12, 18, 47] and multimodal LLMs [5–7, 28] have emerged as powerful backbones for instruction-driven navigation [44, 48–50], thanks to their ability to process open-ended language, encode commonsense knowledge, and support multi-step reasoning. Recent works further explore long-horizon memory, proactive perception, and explicit reasoning mechanisms for VLN [14, 41–43, 52]. Despite these advances, most existing approaches of-

fer limited insight into the decision-making process behind their predicted actions [48, 51]. As a result, diagnosing failures, assessing robustness, and building user trust in real-world deployments remain difficult.

The lack of interpretability is particularly critical for embodied agents operating in human-centric or safety-sensitive scenarios. In these settings, acting correctly is not sufficient: an agent should also communicate its intentions, justify its actions, and allow users to anticipate its behavior. Interpretability is therefore a key requirement for effective human-agent interaction and reliable deployment in practical applications. However, incorporating reasoning capabilities into VLN systems remains an open challenge.

To address these limitations, we introduce ExplaiNavility, a VLN framework that explicitly integrates structured reasoning into an embodied LLM (Fig. 1). Our approach aims to bridge the gap between navigation performance and interpretability by enabling the agent to reason over visual observations and linguistic instructions. Concretely, we combine a visual scene encoder with an LLM that jointly models perception and decision-making, while producing interpretable outputs that expose the agent's decision process. To supervise these outputs, we further augment R2R trajectories with speaker-style action explanations and descriptive action labels, providing explicit reasoning signals during training.

Our framework is designed to make navigation decisions more interpretable. In addition to predicting the next navigation action, the model generates natural language explanations that justify its decisions and provide insight into its behavior over time. The speaker-style explanations are generated from consecutive trajectory observations using a Vision-and-Language model, while descriptive action labels are obtained from the Matterport3D Simulator. By fine-tuning a multimodal LLM with these supervision signals, we encourage navigation policies that are effective, transparent, and more aligned with human interpretation.

Experiments on the Room-to-Room (R2R) dataset [1], our method outperforms prior explainable VLN models, including NavGPT [50] and AuxRN [51], achieving higher success rates and more precise trajectories on unseen environments. Ablation studies further show that explanation generation contributes to these gains. These results suggest that structured reasoning can improve navigation performance while making the agent's behavior easier to inspect.

Contributions. Our main contributions are summarized as follows:

- We propose ExplaiNavility, a VLN framework that integrates structured reasoning into an embodied LLM, enabling the agent to jointly predict navigation actions and generate interpretable explanations that make its decision-making process more transparent.
- We introduce auxiliary modules for action explanation and descriptive action generation, enabling the agent to provide natural language rationales for its decisions and improving the transparency of navigation policies.
- We augment R2R trajectories with speaker-style action explanations and simulator-derived descriptive action labels, providing explicit supervision for training interpretable VLN agents.

– We demonstrate on the R2R dataset that our approach improves over prior explainable VLN methods, surpassing NavGPT and AuxRN on unseen environments. Ablation studies show that explanation generation contributes to trajectory accuracy and generalization, highlighting the role of structured reasoning for interpretable VLN agents.

2 Related Work

2.1 Vision-and-Language Navigation

Vision-and-Language Navigation (VLN) requires an embodied agent to reach a target location by following natural language instructions in a visually grounded environment. Solving this task requires language understanding, visual perception, spatial reasoning, and sequential decision-making.

The standard VLN formulation was introduced by Anderson *et al.* [1], who connected language-guided navigation with realistic indoor environments from Matterport3D and introduced the Room-to-Room (R2R) dataset. This formulation moved beyond earlier navigation settings based on rendered environments [8, 20, 24, 26, 29, 30, 38] or predefined symbolic labels [9, 36], establishing R2R as a widely used benchmark for grounded instruction following. Early VLN agents typically followed sequence-to-sequence architectures with attention, learning to align instructions with visual observations and action sequences. Subsequent work focused on improving generalization to unseen environments. Speaker-Follower [17] introduced speaker-driven data augmentation and pragmatic re-ranking, while RCM [40] used cross-modal matching as an intrinsic reward for reinforcement learning. EnvDrop [34] further improved robustness through speaker-based back-translation and environmental dropout, highlighting the importance of visual diversity during training.

Later methods shifted from recurrent policies to Transformer-based architectures, pretrained multimodal representations, explicit memory, and global planning. PREVALENT [21] introduced a pretraining and fine-tuning paradigm for transferable vision-language-action representations. EGP [15] emphasized contextual global planning through dynamic graph construction, while HAMT [10] modeled instructions, past panoramic observations, and actions with a history-aware multimodal Transformer. DUET [11] further combined global topological reasoning with fine-grained local grounding through dual-scale planning. Together, these methods moved VLN from reactive action prediction toward agents that exploit pretraining, history modeling, and structured planning.

Recently, foundation models have introduced a new paradigm for VLN. LLM-based methods such as NavGPT [50] convert observations, navigation history, and candidate actions into text, enabling zero-shot decision-making through explicit language reasoning. NavGPT-2 [49] further bridges LLM-based navigation and specialist VLN policies by aligning visual observations with a frozen LLM while preserving linguistic navigational reasoning. NaviLLM [48] adapts a multimodal LLM to embodied navigation using schema-based prompts and multi-view

fusion, while NaVid [44] fine-tunes a video-based VLM for end-to-end action prediction from monocular RGB streams. Other recent works address long-horizon navigation by improving memory and context selection: StreamVLN [41] introduces slow-fast context modeling for streaming VLN, DecoVLN [42] decouples observation, reasoning, and correction, and ProFocus [43] performs proactive perception and focused reasoning over high-value waypoints. These advances show the increasing role of foundation models in VLN, but they mainly focus on action prediction, memory management, or planning efficiency rather than direct supervision of human-readable explanations for each navigation decision.

2.2 Reasoning and Explainability in Embodied AI

Reasoning and explainability are increasingly important in embodied AI, especially for navigation agents that operate in human-centric environments. Beyond reaching the goal, an interpretable agent should expose why an action is selected and make its behavior easier to inspect.

Early work on explainable embodied navigation studied the connection between exploration, visual evidence, and language generation. Explore-and-Explain [4] learns to explore unseen environments and recount what the agent observes, while the method proposed in [3] combines efficient exploration and mapping with smart scene description to generate informative and non-repetitive natural language descriptions of relevant scenes. More recently, explainable social navigation methods have integrated visual perception, saliency-based evidence, and language generation to improve transparency in human-robot interaction [31]. These works highlight the importance of producing explanations that are grounded in the agent's perception and behavior.

Reasoning has also become central to modern VLN. AuxRN [51] introduces self-supervised auxiliary tasks that encourage the agent to explain previous actions, estimate progress, predict orientation, and evaluate trajectory consistency. NavGPT [50] and NavGPT-2 [49] exploit LLM reasoning to support high-level planning and interpretable navigation decisions. Recent methods further investigate richer reasoning schemes: EvoNav [14] combines history experience and future thought for zero-shot VLN in continuous environments, Aux-Think [39] trains models with Chain-of-Thought supervision while keeping inference efficient, and FantasyVLN [52] explores multimodal Chain-of-Thought reasoning for long-horizon navigation. Complementary to these approaches, imagination-based methods such as Adaptive Text Dreamer [45] use language to predict future environmental semantics, avoiding costly visual generation while supporting navigation decisions.

Our work follows this line of reasoning-driven VLN, but focuses on explicit step-level interpretability. Rather than using reasoning only for planning, memory selection, or latent action prediction, ExplaiNavility supervises the agent to jointly predict the next navigation action, generate a natural language explanation, and output a descriptive low-level action label. This directly links navigation decisions with human-readable rationales, making the agent's behavior easier to inspect while preserving strong navigation performance.

3 Problem Formulation

We consider the task of VLN, where an embodied agent is placed in a photorealistic 3D environment and is required to reach a target location by following a natural language instruction. At each discrete time step t, the agent observes the surrounding environment from its current viewpoint and receives a set of navigable candidate viewpoints. The objective is to select the next action that best progresses toward the goal while maintaining consistency with the instruction and the previously visited trajectory.

Formally, let $\mathcal{I} = \{w_1, \ldots, w_m\}$ denote the natural language instruction describing the desired path. At time step t, the agent has access to a navigation history

$$\mathcal{H}_t = \{h_1, h_2, \ldots, h_t\}, \tag{1}$$

where each h_i represents the visual observation associated with a previously selected viewpoint. The current observation is represented by a set of multi-view images

$$\mathcal{O}_t = \{I_t^1, I_t^2, \ldots, I_t^n\}, \tag{2}$$

capturing the panoramic visual context around the agent. These views are encoded into scene representations

$$\mathcal{S}_t = \{s_t^1, s_t^2, \ldots, s_t^n\}, \tag{3}$$

which describe the available candidate directions at the current position. In addition, a special candidate corresponding to the `STOP` action is included, allowing the agent to terminate navigation when it estimates that the target has been reached.

The navigation policy can then be formulated as a conditional action prediction problem:

$$a_t = \arg\max_{a \in \mathcal{A}_t} p(a \mid \mathcal{I}, \mathcal{H}_t, \mathcal{S}_t), \tag{4}$$

where $\mathcal{A}_t$ denotes the set of candidate actions available at step t, including both navigable viewpoints and the `STOP` action. After executing the selected action a_t, the agent moves to the corresponding viewpoint, updates its history, and repeats the process until `STOP` is predicted.

Unlike standard VLN formulations that only require action prediction, our setting additionally requires the agent to produce interpretable outputs that describe and justify its behavior. Specifically, at each step the model predicts: (i) the next navigable direction, (ii) a descriptive low-level action label, such as *turn left*, *turn right*, *move forward*, or *stop*, and (iii) a natural language explanation that grounds the selected action in the perceived visual context and the instruction. This leads to the following multi-output formulation:

$$(a_t, d_t, e_t) = f_\theta(\mathcal{I}, \mathcal{H}_t, \mathcal{S}_t), \tag{5}$$

where a_t is the selected candidate action, d_t is the descriptive action, and e_t is the generated explanation.

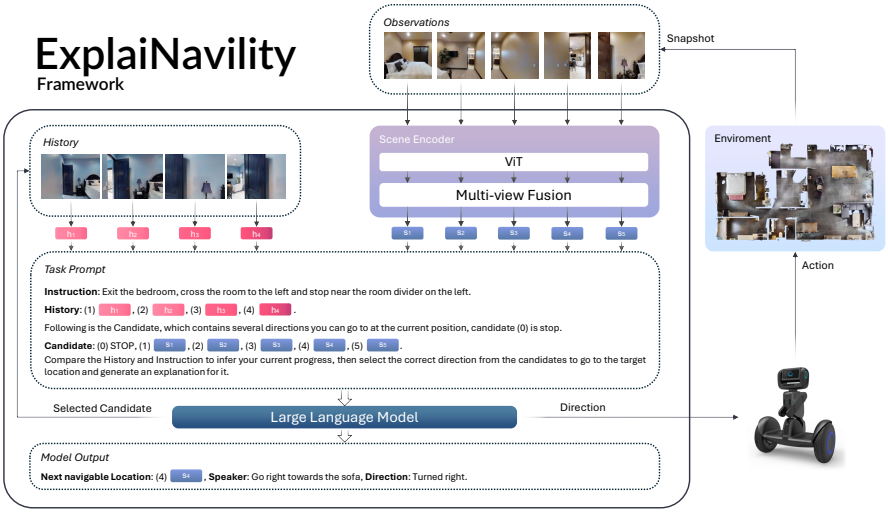


Fig. 2: Overview of ExplaiNavility. Given the current panoramic observation, ExplaiNavility first encodes the available views with a scene encoder and combines them with the navigation history and instruction through a structured multimodal prompt. The LLM then selects the next navigable candidate and generates interpretable outputs, including a natural language explanation and a descriptive low-level action. The selected action is executed in the environment, and the process is repeated until the agent predicts the `STOP` action.

The goal is therefore twofold: learning a navigation policy that successfully reaches the instructed target location, and exposing the agent's decision process through human-readable outputs. This formulation encourages the model to jointly reason over language, visual observations, navigation history, and candidate actions, enabling embodied navigation that is both effective and interpretable.

4 Methodology

We introduce ExplaiNavility, a VLN framework that integrates structured reasoning into an embodied multimodal agent. The framework combines visual scene encoding, structured multimodal prompting, and reasoning supervision to support both action prediction and interpretable output generation. At each step, the agent reasons over the instruction, its navigation history, and the available candidate viewpoints to select the next action, while also producing natural language explanations and descriptive low-level action labels.

The overall framework consists of four main components. First, a visual scene encoder extracts spatial and semantic representations from the multi-view observations surrounding the agent (Sec. 4.2). Second, these visual representations are combined with the navigation instruction and the agent's history through a

structured multimodal prompt, which provides the LLM with an explicit description of the current navigation state (Sec. 4.3). Third, we introduce interpretable reasoning modules that supervise the model to produce both action explanations and descriptive low-level action labels, in addition to the next navigation action (Sec. 4.4). Finally, the resulting embodied multimodal LLM is fine-tuned to jointly optimize navigation and reasoning outputs (Sec. 4.5). An overview of the proposed framework is shown in Fig. 2.

To provide supervision for the reasoning modules, we further construct an augmented speaker-style dataset from R2R trajectories (Sec. 4.1). In particular, we use a multimodal LLM to generate natural language explanations from consecutive trajectory observations, while descriptive action labels are obtained from the Matterport3D Simulator. This augmented supervision allows ExplaiNavility to learn navigation policies that are not only effective, but also more transparent and easier to inspect.

4.1 Speaker Dataset Generation for Explainability

To supervise the reasoning outputs of ExplaiNavility, we augment R2R trajectories with action explanations and descriptive action labels. For each transition along a trajectory, we provide the previous and current observations, together with a textual prompt, to a multimodal LLM (*i.e.*, Gemma-3-12b-it [35] in our implementation). The model is asked to describe the action that explains the visual change between the two observations, using a constrained speaker-style format: `Go [direction] towards the [object/s]`, where `[direction]` is one of `moved forward`, `turned right`, or `turned left`. This produces natural language explanations that connect the agent's motion with visible scene elements.

In addition to these explanations, we derive ground-truth descriptive action labels from the Matterport3D Simulator. Each label belongs to one of four low-level action classes: `turn left`, `turn right`, `move forward`, or `stop`. The resulting augmented dataset provides supervision for both explanation generation and descriptive action prediction, enabling the model to learn navigation behavior together with interpretable reasoning signals.

4.2 Scene Encoding

We follow the scene encoding strategy introduced in NaviLLM [48] to obtain compact visual representations of the agent's surroundings. At each time step, the agent observes a set of views $\{I_i\}_{i=1}^{n}$, where each image corresponds to a candidate viewpoint in the current panoramic observation. Each view is first processed independently by a Vision Transformer (ViT), and the resulting view-level features are then aggregated through a multi-view fusion module to produce scene representations $\{s_i\}_{i=1}^{n}$.

Given an input image I_i, the ViT splits it into a sequence of patches and prepends a learnable `[CLS]` token. The sequence is passed through the Transformer layers, and we retain the final hidden state of the `[CLS]` token as a global

representation of the view:

$$f_i = \mathrm{ViT}(I_i) \quad \text{with } i = 1, 2, \ldots, n, \tag{6}$$

where f_i denotes the visual feature of the i-th view.

To model the spatial and semantic relations among candidate views, we feed the set of view features $\{f_i\}_{i=1}^{n}$ to a Transformer-based multi-view fusion module. This allows each view representation to be contextualized with respect to the other available directions:

$$\{s_i\}_{i=1}^{n} = \text{Transformer-Encoder}(\{f_i\}_{i=1}^{n}), \tag{7}$$

where s_i is the fused scene representation associated with the i-th candidate view.

4.3 Structured Prompting for Multimodal Reasoning

Drawing inspiration from [48], we organize the input to the LLM through a structured multimodal prompt. The prompt explicitly exposes the navigation instruction, the agent's history, and the candidate viewpoints available at the current step, allowing the model to reason jointly over language, past observations, and possible future actions. An example of the resulting prompt is shown in Fig. 3.

Instruction. The instruction describes the sequence of visual-motor actions the agent should follow to reach the target location. For example, an instruction may be: *"Exit the bedroom, cross the room to the left and stop near the room divider on the left."*

History. The history provides temporal context by listing the scene representations associated with the viewpoints previously selected by the agent. Given the chronological history representations $\{\mathbf{h}^i_{\text{scene}}\}_{i=1}^{t}$, each element is preceded by a textual numerical ID indicating its order in the trajectory. The resulting sequence is

$$[Emb(1), \mathbf{h}^1_{\text{scene}}, \ldots, Emb(t), \mathbf{h}^t_{\text{scene}}], \tag{8}$$

where $Emb(i)$ denotes the embedding of the textual identifier associated with the i-th history element. These IDs are not learned visual tokens or task-specific embeddings; they are obtained by encoding textual identifiers such as "(1)", "(2)", and so on, through the LLM tokenizer and embedding layer. They are used to impose an explicit ordering over the trajectory and to make each past observation uniquely referable within the prompt.

Candidates. The candidate block represents the navigable directions available from the current viewpoint. Given the scene representations $\{\mathbf{s}^i_{\text{scene}}\}_{i=1}^{n}$, we assign a textual numerical ID to each candidate view and include a special candidate (0) corresponding to the `STOP` action. The candidate sequence is therefore written as

$$[Emb(1), \mathbf{s}^1_{\text{scene}}, \ldots, Emb(n), \mathbf{s}^n_{\text{scene}}], \tag{9}$$

Instruction: Exit the bedroom, cross the room to the left and stop near the room divider on the left.

History: (1), $\mathbf{h}^1_{\text{scene}}$, ..., ($t$), $\mathbf{h}^t_{\text{scene}}$.

Following are the Candidates, which contain several directions you can go to at the current position; candidate (0) is stop.

Candidates: (0) **STOP**, (1), $\mathbf{s}^1_{\text{scene}}$, ..., ($n$), $\mathbf{s}^n_{\text{scene}}$.

Compare the History and Instruction to infer your current progress, then select the correct direction from the candidates to go to the target location and generate an explanation for it.

Fig. 3: Structured multimodal prompt. The prompt provides the LLM with the navigation instruction, the ordered history of past scene embeddings, and the candidate scene embeddings available at the current viewpoint. Here, $\mathbf{h}^i_{\text{scene}}$ and $\mathbf{s}^i_{\text{scene}}$ denote scene embeddings produced by the scene encoder, while numerical IDs are textual tokens used to index history steps and candidate viewpoints.

where $Emb(i)$ follows the same textual-ID convention used for the history. This schema allows the LLM to refer to candidate viewpoints by index and select the direction that best matches the instruction and the navigation progress.

Input and Output Hints. We include textual hints before and after the candidate block to make the expected reasoning process explicit. The input hint specifies that the listed candidates correspond to the possible directions available at the current position, with candidate (0) reserved for `STOP`. The output hint asks the model to compare the instruction with the navigation history, infer the current progress, select the correct candidate, and generate an explanation for the decision. These hints provide a consistent interface between the multimodal context and the LLM output, encouraging the model to ground action selection in both trajectory history and visual evidence.

4.4 Interpretable Reasoning Modules

To make the navigation process more transparent, we augment the main action prediction objective with two interpretable reasoning modules. Both modules operate on the same multimodal context used for navigation, namely the instruction, the agent's history, and the candidate scene representations, but provide complementary supervision signals for explaining the agent's behavior.

- **Action Explanation Module.** This module generates a natural language rationale that explains the selected navigation decision in terms of the observed visual context. For example, after selecting a leftward candidate, the agent may generate: *"Go left towards the dining table."*
- **Descriptive Action Module.** This module predicts a low-level textual description of the executed action. The output is selected from the action vocabulary {`turn left`, `turn right`, `move forward`, `stop`}, providing a compact and human-readable summary of the agent's movement.

Together, these modules encourage the model to associate navigation actions with explicit linguistic descriptions, making the predicted behavior easier to interpret and inspect.

4.5 Finetuning Procedure and Inference

At each navigation step, the panoramic observation and the explored navigation graph are encoded into candidate viewpoint representations, including a dedicated `STOP` option. The visual representations of previously selected viewpoints are maintained as navigation history. Both candidate and history representations are injected into the LLM, together with the natural language instruction and the structured prompt described in Sec. 4.3.

For navigation finetuning, the model processes the full multimodal prompt and produces a navigation representation from its final hidden states. A linear classification head maps this representation to scores over the valid candidate viewpoints, and the resulting logits are supervised with a cross-entropy loss against the expert action. In parallel, the LLM is trained with an autoregressive language modeling objective to generate the textual rationale and descriptive action label, providing supervision for the reasoning outputs introduced in Sec. 4.4.

During inference, the agent follows the same step-wise procedure. At each step, the model scores the valid candidate viewpoints and selects the next one either greedily or by sampling when enabled. The selected action is executed in the simulator, the corresponding visual representation is appended to the history, and the explored graph is updated. Navigation is therefore performed through candidate classification, while the textual rationale and action description are generated autoregressively by the LLM.

5 Experiments

5.1 Experimental Setup

Dataset. We evaluate ExplaiNavility on the Room-to-Room (R2R) dataset [1], a standard benchmark for VLN. R2R contains photorealistic indoor environments built from real-world scans and human-annotated navigation instructions, requiring agents to ground natural language in sequential visual decision-making. Following the standard setup, we report results on unseen environments to assess the generalization ability of the proposed method. Specifically, we train our model on the original R2R training split augmented with the generated explanations described in Sec. 4.1, and report results on the validation and test splits with unseen environments.

Evaluation Metrics. We adopt standard VLN evaluation metrics for direct comparison with prior work. Trajectory Length (TL) measures the total distance traveled by the agent. Navigation Error (NE) computes the shortest-path distance between the agent's final position and the goal. Success Rate (SR) reports the percentage of episodes in which the agent stops within 3 meters of the

Table 1: Comparison with explainable VLN methods on the Val Unseen and Test Unseen splits of the R2R dataset.

		Val Unseen				Test Unseen			
Method	Trained on R2R	TL	SPL ↑	SR ↑	NE ↓	TL	SPL ↑	SR ↑	NE ↓
NavGPT [50]	✗	11.45	29	34	6.46	-	-	-	-
AuxRN [51]	✓	-	50	55	5.28	-	51	55	5.15
ExplaiNavility (Ours)	✓	13.48	**59**	**67**	**3.48**	13.71	**58**	**66**	**3.71**

target location. Finally, Success Rate weighted by Path Length (SPL) jointly measures success and efficiency by penalizing successful trajectories that are longer than the shortest path. Together, these metrics capture both the accuracy and efficiency of the navigation behavior.

Implementation and Training Details. Our model uses a scene encoder based on EVA-CLIP-02-Large [32], a 428M-parameter ViT whose weights are kept frozen. The multi-view fusion module is implemented as a two-layer Transformer encoder with hidden dimensionality equal to 1,024, while the language backbone is Vicuna-7B [12]. We fine-tune the model using Adam [25] as optimizer with a learning rate of 5×10^{-4} and a batch size of 64. We perform validation every 2,000 steps and select the checkpoint with the highest SPL on the validation split. To enable parameter-efficient adaptation of the LLM, we use LoRA [22] with rank 64, alpha 16, and dropout 0.1, keeping most of the backbone parameters frozen. Training is conducted on four NVIDIA RTX 6000 GPUs and takes approximately 80 hours.

5.2 Quantitative Results

We compare ExplaiNavility with prior VLN methods that explicitly address explainability, namely NavGPT [50] and AuxRN [51]. These methods represent two different settings: NavGPT uses GPT-4 for zero-shot navigation without VLN-specific fine-tuning, while AuxRN trains a navigation policy on R2R with self-supervised auxiliary reasoning tasks. Table 1 reports results on the Val Unseen and Test Unseen splits of R2R.

ExplaiNavility achieves the best overall performance among explainable VLN methods. On Val Unseen, our model obtains an SPL of 59 and an SR of 67, outperforming AuxRN by 9 SPL points and 12 SR points, and substantially improving over NavGPT. It also reduces NE to 3.48, compared to 5.28 for AuxRN and 6.46 for NavGPT, indicating more accurate goal localization. A similar trend is observed on Test Unseen, where ExplaiNavility reaches 58 SPL and 66 SR, improving over AuxRN while also reducing NE from 5.15 to 3.71. These results show that integrating structured reasoning into an embodied multimodal agent improves navigation performance while preserving interpretable outputs.

To isolate the impact of explanation supervision, we report an ablation in Table 2. Removing explanations decreases SPL from 59 to 57 and SR from 67 to 64, suggesting that explanation generation contributes positively to navigation

Table 2: Ablation study on the effect of explanation supervision on the Val Unseen split of the R2R dataset.

	Val Unseen			
Method	TL	SPL ↑	SR ↑	NE ↓
w/o explanations	11.95	57	64	**2.32**
ExplaiNavility (Ours)	13.48	**59**	**67**	3.48

success and path efficiency. Interestingly, the model without explanations obtains a lower NE, while also producing shorter trajectories. This suggests that explanation supervision changes the navigation behavior, favoring trajectories that more often satisfy the success criterion despite a higher average final distance. Overall, the ablation suggests that reasoning supervision provides a useful training signal for instruction following, leading to higher success rates and improved path-quality scores.

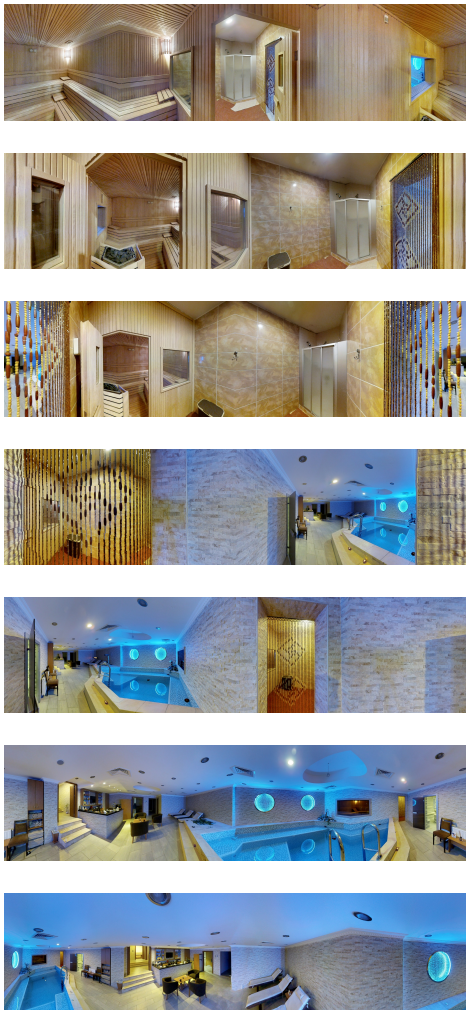
5.3 Qualitative Analysis

We show a qualitative navigation episode produced by ExplaiNavility on the R2R Val Unseen split in Fig. 4. Each step reports the agent's 360° panoramic observation, the generated natural language explanation (**Speaker**), and the corresponding low-level action description (**Direction**). The example illustrates how the model associates its movements with visible landmarks, such as the wall, doorway, pool area, and hot tub, producing explanations that remain aligned with the executed trajectory.

By training on the augmented R2R dataset with both action labels and speaker-style explanations, ExplaiNavility learns to couple navigation decisions with human-readable rationales. These qualitative results show that the model can not only select plausible navigation actions, but also expose its behavior in a form that makes the trajectory easier to inspect.

6 Conclusion

In this work, we presented ExplaiNavility, a VLN framework that integrates structured reasoning into an embodied multimodal agent. By combining scene encoding, structured multimodal prompting, and reasoning supervision, ExplaiNavility jointly predicts navigation actions and produces interpretable outputs as natural language explanations and descriptive action labels. Experiments on R2R show that our approach improves over prior explainable VLN methods on unseen environments, demonstrating the benefit of explanation supervision for navigation performance. Looking forward, we plan to extend this framework to more complex embodied tasks, such as instruction-guided object manipulation and multi-step household planning. Another promising direction is interactive navigation, where agents can ask for clarification when instructions are ambiguous, enabling more effective bidirectional communication with humans.



Initial location

Speaker: Go forward towards the wall.
Direction: Move forwards

Speaker: Go right towards the doorway.
Direction: Turned right

Speaker: Go right towards the doorway.
Direction: Turned right

Speaker: Go right towards the pool and patio area.
Direction: Turned right

Speaker: Go forward towards the hot tub.
Direction: Moved forward

Speaker: Go forward towards the hot tub.
Direction: Stop

Fig. 4: Qualitative example on R2R Val Unseen. The figure shows a navigation trajectory generated by ExplaiNavility. Each row reports the agent's 360° panoramic observation, together with the generated explanation (**Speaker**) and the predicted low-level action (**Direction**).

Acknowledgements

We acknowledge CINECA for the availability of high-performance computing resources under the ISCRA initiative. This work has been supported by the Italian Ministry of University and Research, under the complementary actions to the NRRP "Fit4MedRob - Fit for Medical Robotics" Grant (#PNC0000007), and by the EU Horizon project "ELLIOT" (GA No. 101214398).

References

1. 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/CVF Conference on Computer Vision and Pattern Recognition (2018)
2. Barsellotti, L., Bigazzi, R., Cornia, M., Baraldi, L., Cucchiara, R.: Personalized Instance-based Navigation Toward User-Specific Objects in Realistic Environments. In: Advances in Neural Information Processing Systems (2024)
3. Bigazzi, R., Cornia, M., Cascianelli, S., Baraldi, L., Cucchiara, R.: Embodied Agents for Efficient Exploration and Smart Scene Description. In: Proceedings of the IEEE International Conference on Robotics and Automation (2023)
4. Bigazzi, R., Landi, F., Cornia, M., Cascianelli, S., Baraldi, L., Cucchiara, R.: Explore and Explain: Self-supervised Navigation and Recounting. In: Proceedings of the International Conference on Pattern Recognition (2020)
5. Bucciarelli, D., Moratelli, N., Cornia, M., Baraldi, L., Cucchiara, R.: Personalizing Multimodal Large Language Models for Image Captioning: An Experimental Analysis. In: Proceedings of the European Conference on Computer Vision Workshops (2024)
6. Caffagni, D., Cocchi, F., Barsellotti, L., Moratelli, N., Sarto, S., Baraldi, L., Baraldi, L., Cornia, M., Cucchiara, R.: The Revolution of Multimodal Large Language Models: A Survey. In: Findings of the Annual Meeting of the Association for Computational Linguistics (2024)
7. Caffagni, D., Compagnoni, A., Melis, F., Sarto, S., Dovesi, P.L., Granroth-Wilding, M., Cornia, M., Baraldi, L.: Mind the Heads: Topological Representation Alignment for Multimodal LLMs. arXiv preprint arXiv:2606.23885 (2026)
8. Chaplot, D.S., Sathyendra, K.M., Pasumarthi, R.K., Rajagopal, D., Salakhutdinov, R.: Gated-attention architectures for task-oriented language grounding. In: Proceedings of the AAAI Conference on Artificial Intelligence (2018)
9. Chen, D., Mooney, R.: Learning to interpret natural language navigation instructions from observations. In: Proceedings of the AAAI Conference on Artificial Intelligence (2011)
10. Chen, S., Guhur, P.L., Schmid, C., Laptev, I.: History Aware Multimodal Transformer for Vision-and-Language Navigation. In: Advances in Neural Information Processing Systems (2021)
11. Chen, S., Guhur, P.L., Tapaswi, M., Schmid, C., Laptev, I.: Think Global, Act Local: Dual-scale Graph Transformer for Vision-and-Language Navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
12. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality (2023)
13. Chu, X., Qiao, L., Lin, X., Xu, S., Yang, Y., Hu, Y., Wei, F., Zhang, X., Zhang, B., Wei, X., et al.: MobileVLM : A Fast, Strong and Open Vision Language Assistant for Mobile Devices. arXiv preprint arXiv:2312.16886 (2023)
14. Dai, G., Wang, S., Wang, Z., Xie, G.S., Yang, Y., Pan, J., Sun, Q., Shu, X.: History to Future: Evolving Agent with Experience and Thought for Zero-shot Vision-and-Language Navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2026)

15. Deng, Z., Narasimhan, K., Russakovsky, O.: Evolving graphical planner: Contextual global planning for vision-and-language navigation. In: Advances in Neural Information Processing Systems (2020)
16. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2019)
17. Fried, D., Hu, R., Cirik, V., Rohrbach, A., Andreas, J., Morency, L.P., Berg-Kirkpatrick, T., Saenko, K., Klein, D., Darrell, T.: Speaker-Follower Models for Vision-and-Language Navigation. In: Advances in Neural Information Processing Systems (2018)
18. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The Llama 3 Herd of Models. arXiv preprint arXiv:2407.21783 (2024)
19. Gu, J., Stefani, E., Wu, Q., Thomason, J., Wang, X.: Vision-and-Language Navigation: A Survey of Tasks, Methods, and Future Directions. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (2022)
20. Guadarrama, S., Riano, L., Golland, D., Go, D., Jia, Y., Klein, D., Abbeel, P., Darrell, T., et al.: Grounding spatial relations for human-robot interaction. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (2013)
21. Hao, W., Li, C., Li, X., Carin, L., Gao, J.: Towards Learning a Generic Agent for Vision-and-Language Navigation via Pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
22. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. In: ICLR Workshops (2022)
23. Hu, X., Li, S., Huang, T., Tang, B., Huai, R., Chen, L.: How Simulation Helps Autonomous Driving: A Survey of Sim2real, Digital Twins, and Parallel Intelligence. IEEE Transactions on Intelligent Vehicles (2024)
24. Huang, A.S., Tellex, S., Bachrach, A., Kollar, T., Roy, D., Roy, N.: Natural language command of an autonomous micro-air vehicle. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (2010)
25. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. Proceedings of the International Conference on Learning Representations (2015)
26. Kollar, T., Tellex, S., Roy, D., Roy, N.: Toward understanding natural language directions. In: 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (2010)
27. Landi, F., Baraldi, L., Cornia, M., Corsini, M., Cucchiara, R.: Multimodal Attention Networks for Low-Level Vision-and-Language Navigation. Computer Vision and Image Understanding (2021)
28. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved Baselines with Visual Instruction Tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
29. MacMahon, M., Stankiewicz, B., Kuipers, B.: Walk the talk: Connecting language, knowledge, and action in route instructions. In: Proceedings of the AAAI Conference on Artificial Intelligence (2006)
30. Mei, H., Bansal, M., Walter, M.: Listen, attend, and walk: Neural mapping of navigational instructions to action sequences. In: Proceedings of the AAAI Conference on Artificial Intelligence (2016)

31. Sotomi, O., Kodi, D., Arab, A.: Trust Through Transparency: Explainable Social Navigation for Autonomous Mobile Robots via Vision-Language Models. arXiv preprint arXiv:2504.05477 (2025)
32. Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y.: EVA-CLIP: Improved Training Techniques for CLIP at Scale. arXiv preprint arXiv:2303.15389 (2023)
33. Szot, A., Clegg, A., Undersander, E., Wijmans, E., Zhao, Y., Turner, J., Maestre, N., Mukadam, M., Chaplot, D., Maksymets, O., Gokaslan, A., Vondrus, V., Dharur, S., Meier, F., Galuba, W., Chang, A., Kira, Z., Koltun, V., Malik, J., Savva, M., Batra, D.: Habitat 2.0: Training Home Assistants to Rearrange their Habitat. In: Advances in Neural Information Processing Systems (2021)
34. Tan, H., Yu, L., Bansal, M.: Learning to Navigate Unseen Environments: Back Translation with Environmental Dropout. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2019)
35. Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al.: Gemma 3 Technical Report. arXiv preprint arXiv:2503.19786 (2025)
36. Tellex, S., Kollar, T., Dickerson, S., Walter, M., Banerjee, A., Teller, S., Roy, N.: Understanding natural language commands for robotic navigation and mobile manipulation. In: Proceedings of the AAAI conference on artificial intelligence (2011)
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is All you Need. In: Advances in Neural Information Processing Systems (2017)
38. Vogel, A., Jurafsky, D.: Learning to follow navigational directions. In: Proceedings of the 48th annual meeting of the association for computational linguistics (2010)
39. Wang, S., Wang, Y., Li, W., Cai, X., Wang, Y., Chen, M., Su, Z., Li, D., Fan, Z.: Aux-Think: Exploring Reasoning Strategies for Data-Efficient Vision-Language Navigation. In: Advances in Neural Information Processing Systems (2026)
40. Wang, X., Huang, Q., Celikyilmaz, A., Gao, J., Shen, D., Wang, Y.F., Wang, W.Y., Zhang, L.: Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
41. Wei, M., Wan, C., Yu, X., Wang, T., Yang, Y., Mao, X., Zhu, C., Cai, W., Wang, H., Chen, Y., et al.: StreamVLN: Streaming Vision-and-Language Navigation via SlowFast Context Modeling. In: Proceedings of the IEEE International Conference on Robotics and Automation (2026)
42. Xin, Z., Li, W., Jiang, Y., Wang, B., Cong, R., Qin, J., Huang, S.: DecoVLN: Decoupling Observation, Reasoning, and Correction for Vision-and-Language Navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2026)
43. Xue, W., Li, M., Wu, X., Tang, J., Yang, D., Zhang, L.: ProFocus: Proactive Perception and Focused Reasoning in Vision-and-Language Navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2026)
44. 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. arXiv preprint arXiv:2402.15852 (2024)
45. Zhang, P., Su, Y., Wu, P., An, D., Zhang, L., Wang, Z., Wang, D., Zhao, B.: Cross from Left to Right Brain: Adaptive Text Dreamer for Vision-and-Language

Navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2026)
46. Zhang, Y., Ma, Z., Li, J., Qiao, Y., Wang, Z., Chai, J., Wu, Q., Bansal, M., Kordjamshidi, P.: Vision-and-Language Navigation Today and Tomorrow: A Survey in the Era of Foundation Models. Transactions on Machine Learning Research (2024)
47. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al.: A Survey of Large Language Models. arXiv preprint arXiv:2303.18223 **1**(2), 1–124 (2023)
48. 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 (2024)
49. Zhou, G., Hong, Y., Wang, Z., Wang, X.E., Wu, Q.: NavGPT-2: Unleashing Navigational Reasoning Capability for Large Vision-Language Models. In: Proceedings of the European Conference on Computer Vision (2024)
50. Zhou, G., Hong, Y., Wu, Q.: NavGPT: Explicit Reasoning in Vision-and-Language Navigation with Large Language Models. In: Proceedings of the Conference on Artificial Intelligence (2024)
51. Zhu, F., Zhu, Y., Chang, X., Liang, X.: Vision-language navigation with self-supervised auxiliary reasoning tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
52. Zuo, J., Mu, L., Jiang, F., Ma, C., Xu, M., Qi, Y.: FantasyVLN: Unified Multimodal Chain-of-Thought Reasoning for Vision-Language Navigation. arXiv preprint arXiv:2601.13976 (2026)