Vision-and-Language Navigation with Analogical Textual Descriptions in LLMs

Anonymous ACL submission

Abstract

Integrating large language models (LLMs) into embodied AI models is becoming increasingly prevalent. However, existing zero-shot LLM-based Vision-and-Language Navigation (VLN) agents either encode images as textual scene descriptions, potentially oversimplifying visual details, or process raw image inputs, which can fail to capture abstract semantics required for high-level reasoning. In this paper, we improve the navigation agent's contextual understanding by incorporating textual descriptions that facilitate analogical reasoning across images from multiple perspectives. By leveraging text-based analogical reasoning, the agent enhances its global scene understanding and spatial reasoning, leading to more accurate action decisions. We evaluate our approach on the R2R dataset, where our experiments demonstrate significant improvements in navigation performance.

1 Introduction

With the LLMs being applied across diverse domains, their integration into VLN agents has emerged as a promising development. Zero-shot LLM-based VLN agents represent a significant shift from traditional navigation agents that rely on extensive task-specific training, demonstrating greater adaptability and generalizability to a wide range of environments (Zhang et al., 2024b).

Early approaches for zero-shot LLM-based VLN agents interpret the visual environment by utilizing offline Vision-Language Models (VLMs) (Li et al., 2023; Liu et al., 2023; Wang et al., 2022) to convert visual images into the corresponding textual descriptions (Zhou et al., 2024b; Long et al., 2024a; Qiao et al., 2023). However, as shown in Fig. 1, these textual descriptions often provide very similar information when candidate images contain overlapping views, even if they are captured from different angles. More recently, MapGPT (Chen et al., 2024) processes multiple images simultaneously,

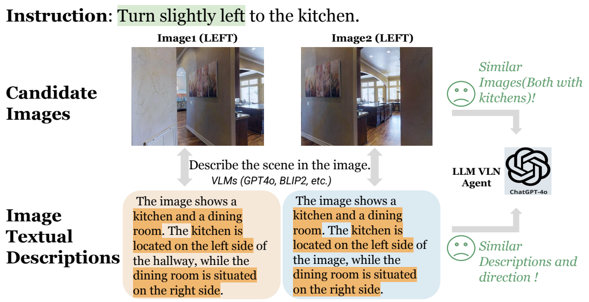

Figure 1: Challenges in current LLM-based VLN Agent. The highlighted orange text shows similar descriptions.

directly feeding them into LLMs as input. This approach reduces redundancy in textual descriptions by leveraging visual differences, but it remains limited when handling highly similar images—such as when both images depict *"a kitchen"* in Fig. 1. Motivated by these challenges, we hypothesize that incorporating additional reasoning processes is necessary to help the agent distinguish key features within the visually similar images while leveraging spatial information to discern their positional differences (*e.g.*, *"slightly left"*).

To address the aforementioned challenges, we propose enhancing the navigation agent's **contextual understanding** by generating textual descriptions of the visual observations, focusing on both *scene understanding from images* and *spatial reasoning within the environment*. Our approach fosters the agent's analogical reasoning and utilizes the power of language to describe differences between images, capturing higher-level scene understanding and spatial relationships. Specifically, instead of treating candidate images as isolated inputs and prompting LLMs to generate independent visual descriptions, we leverage VLMs to compare multiple images and generate contextualized scene descriptions that highlight each image's distinctive features. Furthermore, to strengthen the agent's spatial reasoning, we encourage it to systematically organize and interpret the spatial relationships between images, enabling it to distinguish subtle spatial concepts, such as *"slightly left"* versus *"further left"*. To achieve this, we generate a detailed descriptive paragraph that explicitly captures the spatial relationships among the images based on raw spatial attributes, including rotation angles and distances. In summary, our proposed method bridges the agent's perception and reasoning, enhancing its ability to make more accurate action decisions.

We evaluate our method on the VLN mainstream benchmark Room-to-Room (R2R) (Anderson et al., 2018). Experimental results demonstrate that incorporating our proposed analogical reasoning and spatial descriptions significantly improve navigation performance compared to using raw text or images alone. Furthermore, combining images with our proposed textual descriptions yields the best performance, highlighting the effectiveness of our descriptions in enhancing the agent's reasoning.

2 Related Works

Vision-and-Language Navigation (VLN) is a challenging embodied AI task that requires an agent to navigate in a photo-realistic environment by following natural language instructions (Anderson et al., 2018; Ku et al., 2020; Qi et al., 2020). With the rise of foundation models, most VLN agents focus on integrating pre-trained models and generating large-scale datasets to enhance multimodal representations (Li et al., 2020, 2019; Chen et al., 2021; Qiao et al., 2022; Tan et al., 2019; Li et al., 2022; Wang et al., 2023, 2024; Guhur et al., 2021; Li and Bansal, 2024). Recently, incorporating contemporary LLMs and VLMs into VLN offers a promising solution to mitigate domain-specific training constraints, particularly for zero-shot VLN agents (Zhou et al., 2024b,a; Chen et al., 2024; Long et al., 2024b; Zhang et al., 2024a; Zheng et al., 2024; Qiao et al., 2024). However, current LLM-based VLN agents struggle with distinguishing visually similar scenes and exhibit limited spatial understanding. Our goal is to improve these agents by addressing both challenges.

Analogical Reasoning is a cognitive process that involves comparing different entities to identify underlying structural similarities, particularly in visual domains (Lovett et al., 2009; Lovett and Forbus, 2017; Huang et al., 2021). Rather than relying on surface-level features, it captures spatial and semantic relationships between objects across images, facilitating deeper understanding, abstraction, and generalization. Recent advancements in deep learning have leveraged analogical reasoning to align images with textual descriptions, such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), to establish robust semantic mappings. Building on this foundation, our work extends analogical reasoning to VLN tasks, enabling agents to compare discrete images, discern similarities and differences, and develop a global understanding of the environment.

3 Methods

In this section, we introduce our method, which builds upon MapGPT. Our approach incorporates novel prompting strategies to refine visual observations and integrates additional spatial descriptions of the environment. The model architecture has been shown in Fig 2.

3.1 Task Formulation

In the VLN task, an agent receives a natural language instruction, denoted as I. At each navigation step, the agent perceives visual observations consisting of n discrete images and selects one of these images as its action. The objective is to generate a trajectory (a sequence of images) that follows the given instruction. To achieve this, the LLM-based VLN agent takes multiple sources of information as input, including instruction I, history H_t, topological map M_t, observation O_t, and action space A_t. The agent's decision-making process at step t is formulated as:

$$a_t = LLM(I, H_t, M_t, O_t, A_t), \quad (1)$$

where $a_t \in A_t$. As shown in Fig. 2, the history includes previous step actions, capturing the sequence of movements. The map shows the connectivity graph between places (images). The action space is defined as a combination of direction and image (place), where the direction is determined based on both heading and elevation, including: *go forward*, *turn left/right/around*, and *go up/down*.

3.2 Scene Descriptions for Images

For different LLM-based VLN agents, one of the primary differences lies in how observations O are represented. For instance, NavGPT (Zhou et al., 2024b) and DiscussNav (Long et al., 2024a) utilize VLMs (*e.g.* BLIP-2 (Li et al., 2023)) to convert visual images into corresponding textual descriptions.

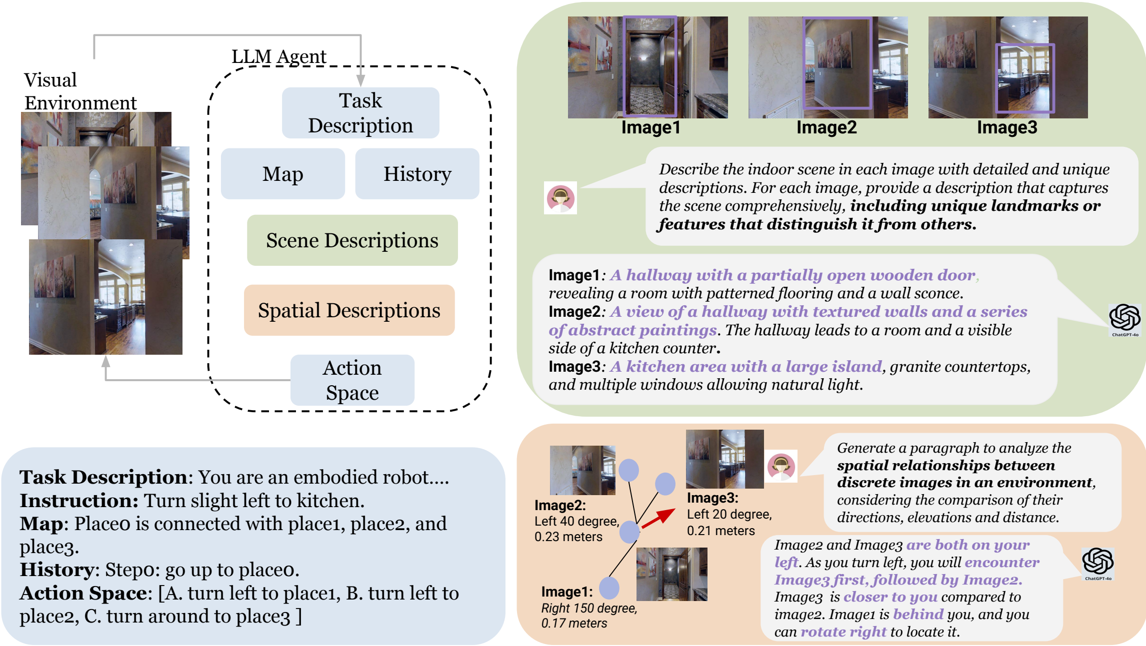

Figure 2: LLM-based VLN Model Architecture. represents the inputs to the Map-GPT agent, while and denote our proposed analogical scene and spatial descriptions, respectively.

While this approach enables language-driven navigation, it has a critical limitation: these textual descriptions treat each discrete image independently, disregarding contextual information across frames. However, a robust VLN agent should not only generate textual descriptions but also ensure that these descriptions encode contextual and relational differences across observations. To achieve this, we propose prompting LLMs to generate detailed visual descriptions while explicitly emphasizing the distinguishing features between different observations, which is formally denoted as follows:

$$OT_1, OT_2, \ldots, OT_n = \mathtt{LLM}(\mathtt{Prompt}(O_1, O_2, \ldots, O_n)), \quad (2)$$

where Prompt is instructions designed to guide the LLMs in generating an analogical analysis of the input. OT_i represents the textual description of the corresponding image O_i.

We illustrate our approach with an example in Fig. 2, where the prompts are demonstrated alongside the corresponding textual descriptions generated for the given images. Our method strategically prompts LLMs to identify distinguishing landmarks that differentiate each image from the others. As a result, the opening sentence of each visual description explicitly highlights these unique features, ensuring a clear comparative distinction. For instance, in Image 1, the description emphasizes a hallway featuring a wooden door, whereas in Image 2, the focus shifts to a hallway with paintings, leading to a room and a kitchen counter. Meanwhile, Image 3 directs attention to a kitchen area centered around a large island. By emphasizing analogical attributes rather than describing each image in isolation, our approach enhances contextual understanding and strengthens the model's ability to discern subtle yet critical differences between visually similar scenes.

3.3 Spatial Descriptions within Environment

A key challenge for the LLM-based VLN agent is effectively representing the spatial structure of its visual environment. In MapGPT, the action space is highly discretized, allowing only a generic *"turn left"* action without differentiating between subtle and significant turns, such as turning 5 degrees versus 30 degrees. This coarse granularity poses a significant limitation when processing instructions like *"turn slightly left"* as the agent lacks the ability to interpret the environment with sufficient details to execute the command precisely. A straightforward approach to addressing this limitation is to directly provide raw heading and elevation values. For example, rather than the ambiguous instruction *"turn left"* the action space could specify *"turn left 5 degrees"*. However, our experiments (Appendix A.3) reveal that the agent struggles to effectively comprehend and integrate this fine-grained spatial information, suggesting that merely providing numerical orientation values is insufficient for enhancing its spatial reasoning.

To address this challenge, we draw inspiration

Methods	NE↓	SR↑	SPL↑
NavGPT (with GPT-4)	6.46	34	29
MapGPT (with GPT-4)	6.29	38.8	25.8
MapGPT (with GPT-4V)	5.63	43.7	34.8
MapGPT (with GPT-4o)	5.31	43.8	36.5
Ours (with GPT-4o)	**4.79**	**49.5**	**42.5**

Table 1: Results on the validation unseen set of the R2R dataset. We implement our method solely on GPT-4o (OpenAI, 2024), as GPT-4V has been deprecated.

Methods	#	Image	Text	GPT	SR↑	SPL↑
NavGPT	1	-	BLIP-2	GPT-3.5	16.7	13.0
MapGPT	2	-	BLIP-2	GPT-4	41.2	25.4
	3	-	BLIP-2	GPT-4o	38.5	26.9
	4	-	GPT-4o	GPT-4o	45.6	36.2
	5	✔	-	GPT-4v	47.7	38.1
	6	✔	-	GPT-4o-05-13	41.2	35.1
	7	✔	-	GPT-4o	47.7	38.7
Ours	8	-	GPT-4o(SI)	GPT-4o	48.2	36.2
	9	-	GPT-4o (SI+SP)	GPT-4o	**50.0**	36.4
	10	✔	GPT-4o (SI+SP)	GPT-4o	**50.0**	**40.2**

Table 2: Results on 72 diverse scenes from the R2R dataset. All GPT-4o versions are from the 08-06 release, except GPT-4o-05-13, which is from the 05-13. SI: scene descriptions for images; SP: spatial descriptions.

from the approach of obtaining analogical scene descriptions from images and extend it to spatial understanding. Our approach focuses on constructing a structured contextual representation that captures spatial relationships across discrete images. Fig. 2 illustrates our designed prompts for describing spatial relationships. We begin by computing the spatial relation, including the relative rotational angle (*e.g.*, "left by 20 degrees") and the relative distance (*e.g.*, "0.21 meters"; note that MapGPT ignores distance). These computed attributes are then incorporated into a structured prompt that guides the LLMs to generate a detailed paragraph analyzing the spatial relationships. The generated description explicitly considers directional comparisons, elevation differences, and distance variations, ensuring a comprehensive understanding of the spatial context. We provide full prompts in the Appendix A.4. We denote the generated spatial description as S, and our enhanced LLM agent's decision-making process is finally defined as follows:

$$a_t = LLM(I, H_t, M_t, \{O_t, OT_t\}, S_t, A_t), \quad (3)$$

where $\{O_t, OT_t\}$ indicates that our agent can flexibly take either the image, its corresponding scene description, or both as input.

4 Experiments

Datasets and Evaluation Metrics. We evaluate our method on the R2R dataset (Anderson et al., 2018), a standard benchmark for VLN. Our primary evaluation metrics include Success Rate (SR), Success weighted by Path Length (SPL), and Navigation Error (NE). We follow MapGPT conducting evaluations on a sampled subset of the R2R dataset, consisting of 72 scenarios and 216 examples. We also report our results on the R2R unseen dataset ($\sim$ 2000 examples). We provide details of evaluation metrics and implementation in the Appendix A.1.

Results Table 1 shows the final performance results on the R2R unseen dataset, demonstrating that our method significantly enhances the baselines, achieving around 6% improvement in both SR and SPL. Table 2 presents our results on 72 diverse scenes. We compare our approach against other LLM-based agents, varying the image input, text input, and GPT backbones. Our findings highlight the importance of using a more advanced captioner for scene descriptions, as BLIP-2 (#3) significantly underperforms compared to GPT-4o (#4). Additionally, the latest GPT-4o (#7) demonstrates a notable improvement over its previous version (#6). Rows 8 to 10 show our method's results. Comparing #4 and #8, we observe that our scene descriptions enhance navigation performance, particularly in SR, with an improvement of nearly 3%. Row 9 shows that incorporating spatial descriptions further boosts SR by an additional 2%. Notably, our results using only text input surpass the baseline results that take image as input (#7). Finally, in #10, we integrate both analogical scene and spatial descriptions while also including the image as input, resulting in an around 4% improvement in SPL. This result indicates that our analogical reasoning descriptions also enhance reasoning over images, suggesting that while images inherently contain all necessary information, our text-based analogical descriptions compensate for the lack of high-level reasoning in visual understanding.

5 Conclusion

In this paper, we propose enhancing the contextual understanding of LLM-based VLN agents by generating analogical scene and spatial descriptions. We encourage the agent to compare images from different perspectives and help the agent construct a structured spatial understanding of the environment. We evaluate our method on the R2R dataset and demonstrate that our approach significantly improves navigation performance.

6 Limitation

Despite the significant improvement in navigation performance achieved by our analogical reasoning descriptions, several limitations remain. First, the quality of the generated descriptions heavily depends on the underlying language model, which may introduce biases or hallucinations that could impact decision-making. Second, the process of generating analogical descriptions adds an additional computational step, potentially increasing processing costs compared to direct image-based navigation.

References

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683.

Jiaqi Chen, Bingqian Lin, Ran Xu, Zhenhua Chai, Xiaodan Liang, and Kwan-Yee Wong. 2024. 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)*, pages 9796–9810.

Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. 2021. History aware multimodal transformer for vision-and-language navigation. In *Advances in Neural Information Processing Systems*, pages 5834–5847.

Pierre-Louis Guhur, Makarand Tapaswi, Shizhe Chen, Ivan Laptev, and Cordelia Schmid. 2021. Airbert: In-domain pretraining for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1634–1643.

Ziqi Huang, Hongyuan Zhu, Ying Sun, Dongkyu Choi, Cheston Tan, and Joo-Hwee Lim. 2021. A diagnostic study of visual question answering with analogical reasoning. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2463–2467. IEEE.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. 2020. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4392–4412.

Jialu Li and Mohit Bansal. 2024. Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 36.

Jialu Li, Hao Tan, and Mohit Bansal. 2022. Envedit: Environment editing for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15407–15417.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

Yuxing Long, Xiaoqi Li, Wenzhe Cai, and Hao Dong. 2024a. Discuss before moving: Visual language navigation via multi-expert discussions. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 17380–17387. IEEE.

Yuxing Long, Xiaoqi Li, Wenzhe Cai, and Hao Dong. 2024b. Discuss before moving: Visual language navigation via multi-expert discussions. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 17380–17387. IEEE.

Andrew Lovett and Kenneth Forbus. 2017. Modeling visual problem solving as analogical reasoning. *Psychological review*, 124(1):60.

Andrew Lovett, Emmett Tomai, Kenneth Forbus, and Jeffrey Usher. 2009. Solving geometric analogy problems through two-stage analogical mapping. *Cognitive science*, 33(7):1192–1231.

OpenAI. 2024. Hello gpt-4o.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. 2020. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991.

Yanyuan Qiao, Qianyi Liu, Jiajun Liu, Jing Liu, and Qi Wu. 2024. Llm as copilot for coarse-grained vision-and-language navigation. In *European Conference on Computer Vision*, pages 459–476. Springer.

Yanyuan Qiao, Yuankai Qi, Yicong Hong, Zheng Yu, Peng Wang, and Qi Wu. 2022. HOP: history-and-order aware pre-training for vision-and-language navigation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 8524–8537.

Yanyuan Qiao, Yuankai Qi, Zheng Yu, Jing Liu, and Qi Wu. 2023. March in chat: Interactive prompting for remote embodied referring expression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15758–15767.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Hao Tan, Licheng Yu, and Mohit Bansal. 2019. Learning to navigate unseen environments: Back translation with environmental dropout. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2610–2621.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. 2022. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*.

Zun Wang, Jialu Li, Yicong Hong, Songze Li, Kunchang Li, Shoubin Yu, Yi Wang, Yu Qiao, Yali Wang, Mohit Bansal, et al. 2024. Bootstrapping language-guided navigation learning with self-refining data flywheel. *arXiv preprint arXiv:2412.08467*.

Zun Wang, Jialu Li, Yicong Hong, Yi Wang, Qi Wu, Mohit Bansal, Stephen Gould, Hao Tan, and Yu Qiao. 2023. Scaling data generation in vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12009–12020.

Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and Wang He. 2024a. Navid: Video-based vlm plans the next step for vision-and-language navigation. In *Robotics: Science and Systems (RSS)*.

Yue Zhang, Ziqiao Ma, Jialu Li, Yanyuan Qiao, Zun Wang, Joyce Chai, Qi Wu, Mohit Bansal, and Parisa Kordjamshidi. 2024b. Vision-and-language navigation today and tomorrow: A survey in the era of foundation models. *Transactions on Machine Learning Research*. Survey Certification.

Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. 2024. Towards learning a generalist model for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13624–13634.

Gengze Zhou, Yicong Hong, Zun Wang, Xin Eric Wang, and Qi Wu. 2024a. Navgpt-2: Unleashing navigational reasoning capability for large vision-language models. In *European Conference on Computer Vision*, pages 260–278. Springer.

Gengze Zhou, Yicong Hong, and Qi Wu. 2024b. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7641–7649.

A Appendix

A.1 Experiments

Evaluation Metrics Three main metrics are used to evaluate navigation performance: (1) Navigation Error (NE): the mean of the shortest path distance between the agent's final position and the goal destination. (2) Success Rate (SR): the percentage of the predicted final position being within 3 meters from the goal destination. (3) Success Rate Weighted Path Length (SPL): normalizes success rate by trajectory length.

Implementation Details. We utilize GPT-4o-08-06 as the backbone for our LLM-based agent, given that GPT-4V has been deprecated. In this work, we employ GPT-4o-08-06 as the backbone for our LLM-based agent, as GPT-4V has been deprecated. MapGPT reports its results using GPT-4o-05-03, but our implementation with GPT-4o-08-06 achieves better performance (around 6% on success rate). To ensure deterministic outputs, we set the temperature to 0. Additionally, we constrain the agent's decision-making process by limiting the maximum number of generated actions to 15 and the maximum token output from GPT to 2000.

A.2 Qualitative Examples

Fig. 3 and Fig. 4 present two qualitative examples illustrating the effectiveness of the proposed analogical scene and spatial descriptions. In Fig. 3, the scene descriptions generated by BLIP-2 and GPT-4o are highly similar despite the visual differences between the scenes. Even for GPT-4o, across three images, the descriptions primarily focus on the general scene, referring to an *"ornate chapel interior"* without providing distinguishing details. In contrast, our method emphasizes different aspects of each image: for example, Image 1 highlights *"the confessional booth"*, Image 2 focuses on *"the benches"*, and Image 3 emphasizes *"the grand altar"*. These distinct descriptions enable the agent to accurately select Image 2, which aligns with the given instruction. Furthermore, in Fig 4, We present an example demonstrating the effectiveness of spatial descriptions. In this case, both Image 4 and Image 5 contain an entranceway. However, our approach encourages the agent to infer that less left/right rotation corresponds to a direction closer to forward. As a result, the agent correctly reasons that Image 5 is better aligned with the instruction *"walk to"*.

A.3 Different Strategies for Spatial Reasoning

We conduct experiments to examine how different spatial reasoning strategies impact navigation performance. Intuitively, enabling an agent to understand nuanced spatial concepts can be achieved by explicitly incorporating varying degrees of rotation into its action space. For example, the agent's action space is more precisely defined, such as *"turn 5 degrees left"*. However, our results reveal that introducing fine-grained rotational actions leads to a slight decline in navigation performance (row #2 in Table. 3). This suggests that VLN agents struggle to effectively structure spatial information when relying solely on numerical rotations degrees. To address this, we propose generating descriptive paragraphs that systematically capture spatial relationships between images. Empirical results demonstrate that our approach enhances navigation performance compared to directly using numerical values into the action space (#3 in Table. 3).

A.4 Prompts for Spatial Descriptions

> Generate a paragraph to analyze the spatial relationships between discrete images in an environment, considering the comparision of their directions, elevations and distance. The input consists of images with specific angles and distances relative to a central point. Here are some rules to follow: Angles between 120 to 240 degree to the left or right indicate behind or around. Angles equals 180 degrees indicate direct behind. Less angles rotation degrees to the left or right indicate closer to the forward direction. For example, Given places along with their spatial information: Place0 is to my right 180.0 degrees and up 30.0 degrees, positioned 0.21 meters away, Place 2 is to my right 60.0 degrees and up 30.0 degrees, positioned 0.21 meters away Place 3 is to my right 90.0 degrees and up 30.0 degrees, positioned 0.18 meters away. Place 4 is to my right 90.0 degrees, positioned 0.05 meters away. Please generate a descriptive paragraph explaining the spatial relationships and navigation steps to these images. For example: "To navigate to Image0, Image2, and Image3, you need to move upward. As you turn right, you will encounter Image2 first, followed by Image3, and finally Image0, which is directly behind you. Image4 is in the same direction as Image3, but Image3 requires looking up while Image4 does not. Additionally, Image4 is very close to you." Output the response in JSON format with the key 'environmental analysis.' "

	Methods	SR↑	SPL↑
1	MapGPT	47.7	38.7
2	+spatial attributes	46.8	37.9
3	+spatial descriptions	**49.1**	**39.3**

Table 3: Different strategies for spatial reasoning.

Instructions: *Walk forward to the padded benches.*

Image1 Image2 Image3

Scene Descriptions (BLIP-2)

Image1: In the scene, there is an **ornate wooden room with a red curtain**. The room also features two chandeliers hanging from the ceiling, adding a touch of elegance to the space.
Image2: In the scene, there is an **ornate wooden altar with a chandelier** hanging above it. The altar is adorned with intricate carvings and decorations, giving it a sense of grandeur and elegance.
Image3: In the scene, there is an **ornate wooden altar with intricate carvings**. The altar is adorned with two red chairs placed in front of it, creating a cozy and inviting atmosphere.

Scene Descriptions (GPT-4o)

Image1: The image depicts an **interior space with a classic, ornate design, featuring wooden paneling and red drapery.** The room showcases a red velvet bench and curtains, with intricate wooden carvings.
Image2: The image depicts an **ornate, traditional chapel interior with rich red and wooden tones**. Red velvet curtains and cushioned kneelers contrast with the intricately carved wooden altar and paneling.
Image3: **The image depicts an ornate chapel interior with a richly decorated altar and wooden paneling.** The altar features intricate carvings and religious icons.

Analogical Scene Descriptions (GPT-4o)

Image1: The scene depicts an ornate interior space with a focus on a wooden confessional booth. The booth features intricate woodwork and is partially enclosed by a red curtain.
Image2: This image showcases a richly decorated interior with a series of red **velvet benches** facing a wooden altar. The altar is intricately carved, featuring religious iconography. Above, chandeliers hang from the ceiling, adding elegance to the space. The walls are lined with red curtains.
Image3: The focus here is on a grand altar set against a backdrop of detailed wood carvings depicting religious figures. The altar is flanked by red velvet benches, and the ceiling is adorned with ornate chandeliers.

Figure 3: A qualitative example demonstrating the effectiveness of analogical scene descriptions. We collect scene descriptions from both BLIP-2 and GPT-4o. Bold text highlights similar descriptions across images, while text in green boxes represents our generated analogical scene descriptions, each emphasizing different aspects.

Instructions: *Walk to the end of the entrance way. Continue straight until you reach the dining room.*

Image1 Image2 Image3 Image4 Image5

Left 210 degree | Left 180 degree | Left 150 degree | Left 90 degree | Right 30 degree

MapGPT: The instruction is walk to the end of the entrance way. Image 4 shows a hallway that seems to lead further into the building, which aligns with the instruction to walk to the end of the entrance way. I should go to Place 4.

Environmental Analysis: To navigate through the environment, begin by turning to your right to encounter image 5. As you turn left, you will encounter image 4, image 3, image 2, and image 1. Image 2 is directly behind you. Image 5 is closer to the forward direction.

Ours: The instruction suggests that I need to walk through an entrance way, and eventually reach a dining room. The images provided do not directly depict a dining room or a kitchen area, but they do show hallways and rooms with seating areas. Based on the **environmental analysis**, Image 5 are the most straightforward to access, and correspond to the same hallway description. I should go to Place 5.

Figure 4: A qualitative example illustrating the effectiveness of our spatial descriptions. The agent successfully identified Place5 based on its relative position, as it is closer to the forward direction than other images and better aligned with the instruction *"walk to"* compared to Place4, which requires a significant left turn.