3D Gaussian Map with Open-Set Semantic Grouping for Vision-Language Navigation

Supplementary Material

This supplementary document provides more details of our approach and additional experimental results, which are organized as follows:

- Additional Details (§A)
- Model Details (§B)
- Discussion (§C)

A. Additional Details

List of Symbols. Table A1 concisely lists the symbols, excluding unnecessary subscripts for clarity.

Notation	Description	Index
$\mathcal{X}$	Natural language instructions	§3
$\mathcal{I}$	RGB images	§3.1; Eq. (2)&(3)&(4)&(13)
$\mathcal{D}$	Depth images	§3.1; Eq. (2)&(3)&(6)&(14)
$\mathcal{V}$	Neighboring nodes	§3.1
$\mathcal{V}^*$	Other observed nodes	§3.1
$\mathcal{G}^*$	3D Gaussian primitives	§3.1; Eq. (2)
$\mathcal{A}$	Action space	§3.3
(u,v)	2D position of each pixel	§3.1; Eq. (8)&(9)
z	Distance to the center of Gaussian primitives	§3.1; Eq. (1)&(6)
$\boldsymbol{\mu}$	Position of Gaussian primitives	§3.1
$\boldsymbol{c}$	Color of Gaussian primitives	§3.1; Eq. (4)
$\boldsymbol{s}$	Scale of Gaussian primitives	§3.1
$\boldsymbol{r}$	Rotation of Gaussian primitives	§3.1
σ	Semantic of Gaussian primitives	§3.1; Eq. (8)
α	Opacity of Gaussian primitives	§3.1&3.2; Eq. (4)&(6)&(8)
$\boldsymbol{g}$	Gaussian representation	§3.2
m	2D masks of RGB images	§3.2; Eq. (7)
$\boldsymbol{F}^s$	2D CLIP feature	§3.2; Eq. (7)&(14)
$\hat{F}^s$	Semantic feature of Gaussian primitives	§3.2; Eq. (8)&(14)
$\boldsymbol{F}^e$	Scene feature	§3.2; Eq. (9)
$\boldsymbol{F}^v$	View feature	§3.3
$\boldsymbol{F}^i$	Instance feature	§3.3; Eq. (11)
$\boldsymbol{p}^e$	Scene-level action probabilities	§3.3; Eq. (9)&(12)
$\boldsymbol{p}^v$	View-level action probabilities	§3.3; Eq. (10)&(12)
$\boldsymbol{p}^i$	Instance-level action probabilities	§3.3; Eq. (11)&(12)
$\boldsymbol{p}^c$	Multi-Level action probabilities	§3.3; Eq. (12)

† Subscript t in the paper denotes the navigation step.

Table A1. Notation and Description of Key Symbols.

Visualization. We provide additional visualization results on R2R [1] *val unseen* splits to further illustrate the advantages of our approach in spatial and semantic understanding. In Fig. A1 (a), we highlight the role of geometric priors in navigation. By leveraging the structured spatial information of our 3D Gaussian Map, the agent accurately interprets elevation changes and navigates through a multi-level environment. Specifically, the agent follows the instruction to *"descend three steps"* and *"exit through the doorway"*, demonstrating how our approach effectively utilizes geometric priors to enhance spatial awareness. Moreover, in Fig. A1 (b), we showcase our method's ability to handle open-set semantics. The agent is required to recognize and utilize semantic cues, such as *"the picture hanging on the wall"* and *"the kitchen"*, to navigate through the scene. By integrating OSG (§3.2) into the 3D Gaussian Map, the agent correctly associates these semantic elements with their spatial locations, ensuring accurate decision-making.

B. Model Details

2D Action Score. 2D observations provide fine-grained contextual cues, such as object details and textures, which complement our 3D Gaussian Map and play a crucial role in decision-making. Therefore, following prior works [7–9], our agent encodes the panoramic view and detected objects into 2D visual features, denoted as $\boldsymbol{F}^{\text{2D}} \in \mathbb{R}^{768}$, using a multi-layer transformer with feed-forward layers (MLT) [3]. These features are combined with instruction embeddings $\boldsymbol{X} \in \mathbb{R}^{768}$ and processed through another MLT $\mathcal{F}^{\text{MLT}}$ to compute 2D action scores $\boldsymbol{p}^{\text{2D}}$:

$$\boldsymbol{p}^{\text{2D}} = \text{Softmax}(\mathcal{F}^{\text{MLT}}([\boldsymbol{F}^{\text{2D}}, \boldsymbol{X}])) \in [0,1]^{|\mathcal{V}|}, \tag{B1}$$

where $[,]$ denotes concatenation and where $|\mathcal{V}|$ indicates the number of candidate points. To align these scores with the action space $\mathcal{A}$, $\boldsymbol{p}^{\text{2D}}$ is aggregated for the neighboring nodes $\mathcal{V}$ stored in the topological memory. This aggregation employs a nearest neighbor matching function $\mathcal{N}$, which clusters scores from spatially proximate nodes and assigns a unified score to each candidate node:

$$\boldsymbol{p}^{\widehat{\text{2D}}} = \mathcal{N}(\boldsymbol{p}^{\text{2D}}, \mathcal{V}) \in [0,1]^{|\mathcal{V}|}. \tag{B2}$$

This process consolidates scores from nearby nodes, ensuring consistent navigation priorities for each candidate.

Navigation Losses. Following existing methods [3], we adopt a two-stage training regime: pretraining with auxiliary tasks to improve multimodal representations, followed by finetuning with behavior cloning and pseudo-expert supervision to refine the navigation policy.

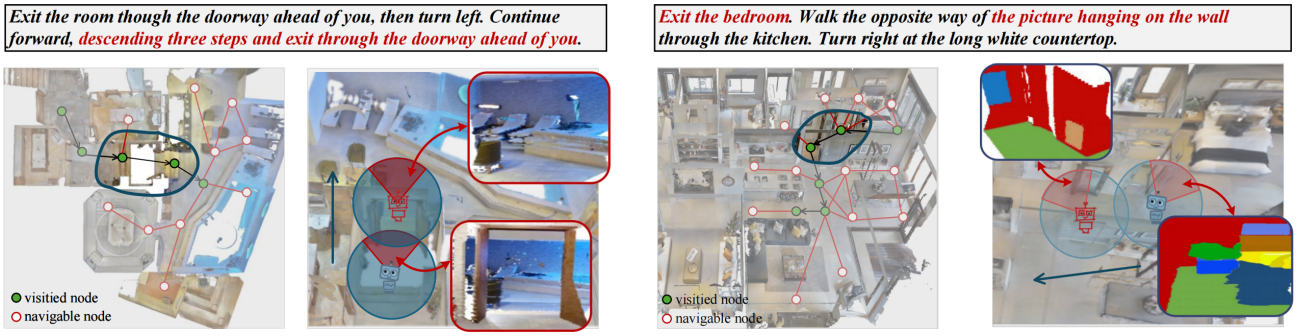
During pretraining, we employ the Masked Language Modeling (MLM) task and the Single-step Action Prediction (SAP) task as auxiliary objectives for R2R [1] and R4R [6], while additionally using the Object Grounding (OG) task for REVERIE [11] to enhance object-level reasoning. The corresponding loss functions are defined as:

$$\mathcal{L}^{\text{MLM}} = -\log p(w_i|\mathcal{X}_{\backslash i}, \mathcal{R}), \tag{B3}$$

$$\mathcal{L}^{\text{SAP}} = \sum_{t=1}^{T} -\log p(a_t^*|\mathcal{X}, \mathcal{R}_{<t}), \tag{B4}$$

$$\mathcal{L}^{\text{OG}} = -\log p(o^*|\mathcal{X}, \mathcal{R}), \tag{B5}$$

where $\mathcal{X}$ represents the natural language instruction, with x_i as the masked word in MLM and $\mathcal{X}_{\backslash i}$ as the remaining context. $\mathcal{R}$ denotes the trajectory, where $\mathcal{R}< t$ indicates the


Exit the room though the doorway ahead of you, then turn left. Continue forward, descending three steps and exit through the doorway ahead of you.

Exit the bedroom. Walk the opposite way of the picture hanging on the wall through the kitchen. Turn right at the long white countertop.

(a) (b)

Figure A1. **Visualization of 3D Gaussian Maps** on R2R [1] *unseen* split. (a) Benefiting from the geometric priors in our 3D Gaussian Map, the agent accurately perceives spatial structures and elevation changes, correctly *"descend three steps"* and *"exiting through the doorway"*. (b) Leveraging open-set semantics, the agent correctly associates *"the picture hanging on the wall"* and *"the kitchen"* with their spatial locations, demonstrating fine-grained semantic understanding in navigation. See §A for more details.

partial path up to step t. The variables $a_t^* \in \mathcal{A}_t$ and o^* refer to the expert action and target object, respectively.

For fine-tuning, we adopt DAgger [3] to improve navigation performance. This iterative approach refines the agent's policy by generating trajectories based on its current predictions and dynamically incorporating pseudo-expert feedback to correct suboptimal actions. The pseudo-expert supervises the agent using shortest-path planning from the partially constructed topological memory, enabling robust and adaptive navigation in unseen environments.

C. Discussion

Terms of Use, Privacy, and License. Matterport3D [2], R2R [1], R4R [6], and REVERIE [11] are available for non-commercial research purpose.

Limitations. *i) Real-World Deployment.* Our method is trained and evaluated in the static Matterport3D simulator [2]. Deploying it in dynamic real-world environments may face challenges, such as handling moving objects, which require further research to ensure safe and reliable operation. *ii) Task Generalization.* This work focuses on indoor VLN. Its applicability to other navigation tasks, such as those in [4, 5, 10], remains unexplored and will be investigated in future work. *iii) Environmental Diversity.* Our method is primarily designed and evaluated for indoor environments, which may limit its effectiveness in other scenarios, such as industrial facilities or outdoor spaces.

Broader Impact. We explore the potential of 3DGS-based technology in VLN. Specifically, we propose a 3D Gaussian Map that integrates geometric priors and open-set semantics into a unified representation. Furthermore, we introduce a navigation strategy that incorporates this map into the sequential decision-making process of VLN. We validate the effectiveness of our approach through extensive quantitative and qualitative experiments. Finally, we hope our work inspires new insights and advances in VLN community.

References

[1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian D. Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, 2018. 1, 2

[2] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *3DV*, 2017. 2

[3] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *CVPR*, 2022. 1, 2

[4] Sheng Fan, Rui Liu, Wenguan Wang, and Yi Yang. Navigation instruction generation with bev perception and large language models. In *ECCV*, 2024. 2

[5] Sheng Fan, Rui Liu, Wenguan Wang, and Yi Yang. Scene map-based prompt tuning for navigation instruction generation. In *CVPR*, 2025. 2

[6] Vihan Jain, Gabriel Magalhaes, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. Stay on the path: Instruction fidelity in vision-and-language navigation. In *ACL*, 2019. 1, 2

[7] Rui Liu, Xiaohan Wang, Wenguan Wang, and Yi Yang. Bird's-eye-view scene graph for vision-language navigation. In *ICCV*, 2023. 1

[8] Rui Liu, Wenguan Wang, and Yi Yang. Vision-language navigation with energy-based policy. In *NeurIPS*, 2024.

[9] Rui Liu, Wenguan Wang, and Yi Yang. Volumetric environment representation for vision-language navigation. In *CVPR*, 2024. 1

[10] Rui Liu, Sheng Fan, Wenguan Wang, and Yi Yang. Underwater visual slam with depth uncertainty and medium modeling. In *ICCV*, 2025. 2

[11] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *CVPR*, 2020. 1, 2