SAME: Learning Generic Language-Guided Visual Navigation with State-Adaptive Mixture of Experts

Supplementary Material

In this supplementary material, we aim to provide additional details to support the main content of our paper:

- **Comparison of Training Methods:** In Section A, we compare and discuss the training methodologies employed in *ObjectGoal Navigation (*OBJECTNAV*)* and *Vision-and-Language Navigation (VLN)* research. This comparison highlights the motivation behind the chosen training strategies and the data selection for SAME.
- **Illustration of the DUET Method:** Section B offers a comprehensive explanation of the DUET [3] approach utilized in our study, elucidating its design and integration within our framework.
- **Datasets:** In Section C, we provide a detailed introduction to the additional datasets we used to evaluate SAME.
- **Full Results:** The complete results of SAME are provided in Table 1, showcasing the performance of our method across various metrics and datasets.

A. OBJECTNAV and VLN Training

Besides learning shareable knowledge and task-specific skills from the model design of SAME, another challenge under the unified language-guided navigation framework is to determine the most effective approach to facilitate the learning of agents' language comprehension capacity and grounding it in action prediction. Analyzing the rationality within the contrasting research focuses on ObjectGoal Navigation and VLN offers insights into this challenge. Specifically, we observe that the primary differences lie in the ***training data*** and ***training methods*** used. In this section, we discuss and make strategic decisions in SAME training regarding these two aspects, to address the above challenge.

When language instructions are minimal, the task reduces to OBJECTNAV [2], where the learning objective is the semantic affinity between the target object and the visual perception, and leveraging episodic memory for strategic exploration without redundant revisits, since no extract information is provided from the language instruction. From the ***data*** aspect, it is proven to be effective to learn strategical exploration through human demonstration, and data collection is done in several works [10, 11]. From the ***training*** aspect, OBJECTNAV combines learning "where" and "how" to move, incorporating semantic perception and low-level collision avoidance (`FORWARD`, `TURN_LEFT`, `TURN_RIGHT`, `STOP`) within a continuous environment [12].

On the contrary, VLN requires higher-level language understanding, where the agents not only need to understand the visual semantics of the environment but also need to align past observation and action sequence with the language description. From the ***data*** aspect, VLN agents learn higher-level language comprehension capacity from human-annotated instructions for navigation episodes. From the ***training*** aspect, such alignment is hard to learn directly in a continuous environment, evident by the low performance ($\sim$ 35% SR) on VLN-CE benchmark [6] of the methods that directly operate in continuous environments. Therefore, VLN research typically decouples Vision-Language alignment from collision avoidance by learning to navigate in a discretized environment [1], where the navigable viewpoints are densely sampled from the environment at an average separation of 2.25m to form a navigation graph. The learned multimodal navigation policy performs high-level action by selecting view directions that contain a navigable viewpoint and teleporting between viewpoints on the graph. A waypoint predictor [4, 5, 7] is employed to bridge action space discrepancies in continuous settings. Decoupling the learning of Vision-Language-Action alignment with low-level action control significantly benefits the learning of language understanding capacity, improving VLN-CE success rates by $\sim$ 20%.

To bridge action space discrepancies in continuous settings, modular designs employ waypoint predictors to propose navigable waypoints based on current observation, while the multimodality navigation policy performs view selection conditioned on these waypoints, with a heuristic controller executing low-level actions to move to the waypoint. Decoupling the learning of vision-language-action alignment with low-level action control significantly benefits the language understanding capacity, improving VLN-CE success rates by approximately 20%. In this work, we hypothesize such modular setups optimize the learning of language understanding capacity, which guides us to perform unified policy training in the discrete environment.

Building on the aforementioned discussion, this work concentrates on solving the high-level decision-learning problem by decoupling it from tasks such as collision avoidance and low-level control. This direction motivates the adoption of ***VLN training methods*** for SAME training within a discrete environment. Regarding ***training data***, we combine OBJECTNAV human demonstration data with VLN human-annotated instructions to capture and learn distinct navigation behaviors.

B. DUET Revisit

SAME builds upon the design of the Dual-scale Graph Transformer (DUET) [3]. DUET incorporates a text encoder to process instructions and employs both global and local branches to facilitate cross-modal reasoning at coarse and fine scales.

B.1. Text and Visual Embedding

DUET's text encoder leverages a 12-layer transformer, initialized with LXMERT [13]. For visual embedding, each node's visual observation comprises 36 view images, covering 12 horizontal and 3 vertical directions. To differentiate between these views, a directional embedding E^{ang} is added to the visual features $\hat{\mathcal{O}}_t$, which are extracted by the vision encoder. Since DUET incorporates all 36 view images to construct the spatial observation, navigable adjacent nodes are only visible in a subset of these views, referred to as navigable views. To account for this, a navigable embedding E^{nav} is also included. The final visual embedding is processed by a 2-layer transformer to encode spatial relationships between views, producing panoramic view embeddings:

$$\mathcal{O}^{\text{pano}} = \text{SelfAttn}\left(\hat{\mathcal{O}}_t + E^{\text{ang}} + E^{\text{nav}}\right). \tag{1}$$

B.2. DUET Local Branch

This section focuses on the local branch of DUET, which predicts actions based on the current node's instruction and egocentric observation. Unlike the global branch, no graph-level information is utilized beyond local observations.

B.2.1. Local Visual Embedding

The panoramic view embedding $\mathcal{O}^{\text{pano}}$ is augmented with two types of location embeddings. The first represents the relative location of the current node with respect to the starting node, encoding long-distance directional relationships. The second represents the egocentric directions of adjacent views at the current node, enabling actions such as "turn right."

B.2.2. Local Cross-Modal Encoding

The local branch employs a standard 4-layer cross-modal transformer to capture relationships between vision and language. During action prediction, a mask is applied to exclude unnavigable views and action logits are computed only for the navigable views at the current node.

B.3. DUET Global Branch

This section introduces the global branch of DUET, which tasks the topological map representation $\hat{\mathcal{G}}_t$ and encoded language instruction $\hat{\mathcal{W}}$ to predict actions by selecting any nodes on the graph.

B.3.1. Node Embedding

For each node on the graph, two additional encodings are applied: a location encoding E^{loc} and a navigation step encoding E^{step}. The location encoding represents the egocentric position of a node on the map, capturing its orientation and distance relative to the current node. On the other hand, the navigation step encoding assigns a value corresponding to the latest visited timestep for previously visited nodes, while unexplored nodes are encoded with a value of 0. This encoding scheme enables the model to differentiate nodes based on their navigation history, thereby enhancing alignment with the provided instructions. Additionally, a special "stop" node is introduced into the graph to signify the stop action. This node is connected to all other nodes in the graph.

B.3.2. Global Cross-Modal Encoding

The encoded node features and word embeddings are processed through a 4-layer graph-aware cross-modal transformer, which is composed of the following two key components, as illustrated in Figure 2.

Cross-Attention Layer This layer models the relationships between the global map and the instruction, enabling cross-modal alignment. SAME examine applying State-Adaptive MoE on the visual query W_q or textual key W_k and value W_v in this layer.

Graph-Aware Self-Attention Layer (GASA) Unlike standard self-attention mechanisms which rely solely on visual similarity, the GASA module incorporates the graph's structural information to refine attention computation, formulated as follows:

$$\text{GASA}(\mathcal{V}) = \text{Softmax}\left(\frac{\mathcal{V}W_q(\mathcal{V}W_k)^T}{\sqrt{d}} + A(\mathcal{E}_t)\right)\mathcal{V}W_v, \tag{2}$$

where $A(\mathcal{E}_t)$ represents the spatial affinity matrix, comprised of pairwise L2 distances among all observed nodes. By incorporating this spatial context, GASA ensures that the model prioritizes spatially or topologically proximate nodes, which are often more contextually relevant than visually similar but distant nodes.

Each block in the global branch concludes with a Feed-Forward Network (FFN). Additionally, SAME explores applying the State-Adaptive MoE mechanism to this FFN, as depicted in Figure 2 of the main paper.

C. Datasets

Besides the R2R, REVERIE, and OBJECTNAV-MP3D, we include 4 other datasets for larger scale training and evaluation.

- RxR-EN [8]: English split of the RxR dataset, which contains longer instructions compared to R2R and non-shortest trajectory from starting point to ending point.

Benchmark	Val Unseen					Test Unseen				
	TL	NE↓	nDTW↑	SR↑	SPL↑	TL	NE↓	GP↑	SR↑	SPL↑
R2R [1]	13.65	2.73	71.05	76.25	66.16	14.80	3.03	–	73.92	64.41
RxR-EN [8]	22.69	6.53	51.20	50.52	42.19	–	–	–	–	–
REVERIE [9]	18.87	5.18	48.54	46.35	36.12	19.47	–	–	48.60	37.10
SOON [15]	34.42	8.12	–	36.11	25.42	37.99	–	–	38.18	27.11
CVDN [14]	30.90	12.72	–	24.48	17.23	–	–	7.07	18.15	12.18

Table 1. Full results of SAME on all VLN benchmarks.

- CVDN [14] requires the agent to comprehend the conversation history and infer the correct next actions based on the dialogue context. For evaluation, we use the standard metric, Goal Progress (GP), which calculates the average difference between the completed trajectory length and the remaining distance to the goal.
- SOON [15]: Similar to REVERIE, the instructions describe target rooms and objects, with an average length of 47 words. The expert paths vary in length from 2 to 21 steps, with an average of 9.5 steps.
- R2R-CE [6]: Transfering the discrete trajectories in R2R to continuous 3D scans rendered by Habitat [12], allowing an agent to navigate freely in open space while requiring interaction with obstacles.

D. Full Results on All VLN Tasks

We show the full results of SAME on all the tested VLN benchmarks in Table 1.

References

[1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 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, 2018. 1, 3

[2] Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. ObjectNav Revisited: On Evaluation of Embodied Agents Navigating to Objects. In *arXiv:2006.13171*, 2020. 1

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16537–16547, 2022. 1, 2

[4] Yicong Hong, Zun Wang, Qi Wu, and Stephen Gould. Bridging the gap between learning in discrete and continuous environments for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15439–15449, 2022. 1

[5] Jacob Krantz and Stefan Lee. Sim-2-sim transfer for vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, pages 588–603. Springer, 2022. 1

[6] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020. 1, 3

[7] Jacob Krantz, Aaron Gokaslan, Dhruv Batra, Stefan Lee, and Oleksandr Maksymets. Waypoint models for instruction-guided navigation in continuous environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15162–15171, 2021. 1

[8] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. 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, 2020. 2, 3

[9] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991, 2020. 3

[10] Ram Ramrakhya, Eric Undersander, Dhruv Batra, and Abhishek Das. Habitat-web: Learning embodied object-search strategies from human demonstrations at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5173–5183, 2022. 1

[11] Ram Ramrakhya, Dhruv Batra, Erik Wijmans, and Abhishek Das. Pirlnav: Pretraining with imitation and rl finetuning for objectnav. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17896–17906, 2023. 1

[12] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9339–9347, 2019. 1, 3

[13] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, 2019. 2

[14] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406, 2020. 3

[15] Fengda Zhu, Xiwen Liang, Yi Zhu, Qizhi Yu, Xiaojun Chang, and Xiaodan Liang. Soon: Scenario oriented object navigation with graph-based exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12689–12699, 2021. 3