Cross from Left to Right Brain: Adaptive Text Dreamer for Vision-and-Language Navigation

Supplementary Material

Abstract

This supplementary material provides additional details to support the reproducibility of our proposed method ATD and further demonstrates its effectiveness through extended analyses and experiments.

▷ ***Sec. 1****: Configuration of the data collection process for the State-Estimation LLM and the Imagination LLM, including prompt design, candidate node caption collection, and data visualization.*

▷ ***Sec. 2****: A detailed explanation of the evaluation metrics and additional implementation details of the training process.*

▷ ***Sec. 3****: Additional experiments, including an analysis of model-size effects and experiments conducted on the SOON dataset. We also present visualizations of the navigation process and intermediate feature representations.*

▷ ***Sec. 4****: Discussion of the limitations of our work and potential directions for future research.*

1. Instruction Data Collection

To train the State Estimation LLM and Imagination LLM, we separately collect instruction data for Q-Former fine-tuning of each LLM.

State Estimation Intruction Collection. We employ a more capable LLM [1] to perform text-based state estimation of the navigation process and use its outputs as ground truth to train our left brain. This process is more akin to distilling the state estimation capability into the Q-Former that we fine-tune. The prompt is shown in Fig. 1.

Imagination Intruction Collection. To train the Imagination LLM, we gather the captions of candidate nodes associated with each node to serve as the training ground truth. An illustration of this collection process is shown in Fig. 2. As depicted, for each current node, we first collect images from all candidate nodes and stitch them together into a panoramic view. We then employ the Qwen2.5-VL-7B-Instruct [6] model to generate a caption for the panoramic image. By combining these captions, we produce the imagination ground truth for the current node.

System Prompt. As described in our method section, both the Left Brain and Right Brain use prompts to integrate the visual information generated by the Q-Former with the instruction input before passing it to the LLM. The system prompts for the State Estimation LLM and the Imagination LLM are illustrated in the Fig. 3.

2. Additional Implementation Details

2.1. Evaluation Metrics

In this section, we present a comprehensive overview of the evaluation metrics employed in our study. Consistent with prior research, we utilize five key metrics: Trajectory Length (**TL**), Navigation Error (**NE**), Success Rate (**SR**), Oracle Success Rate (**OSR**), and Success Weighted by Path Length (**SPL**). The following subsections provide a detailed description of each metric along with its corresponding mathematical formulation:

- **TL.** The total length of the predicted trajectory, calculated as the sum of the shortest path distances between consecutive viewpoints along the trajectory.

$$L_{\text{traj}} = \sum_{i=1}^{N-1} d(v_i, v_{i+1}), \tag{1}$$

  where v_i is the i-th viewpoint in the predicted path, $d(a, b)$ is the shortest path distance between viewpoints a and b, and N is the number of viewpoints.
- **NE.** The shortest path distance between the final viewpoint of the predicted trajectory and the goal viewpoint.

$$E_{\text{nav}} = d(v_N, v^*), \tag{2}$$

  where v_N is the last viewpoint in the predicted trajectory and v^* is the goal viewpoint.
- **SR.** Indicates whether the agent successfully reached the goal. For tasks with a set of goal viewpoints, success is achieved if the final viewpoint of the predicted path is within the goal set. Otherwise, success is defined by whether the navigation error is below a certain threshold ϵ.

$$S = \begin{cases} 1 & \text{if } v_N \in G \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

  or, if no goal set exists:

$$S = \begin{cases} 1 & \text{if } d(v_N, v^*) < \epsilon \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

  where v_N is the final viewpoint of the predicted path, G is the goal viewpoint set, v^* is the goal viewpoint, and ϵ is the success threshold.
- **OSR.** Measures whether any viewpoint along the predicted path falls within the goal set, reflecting the best

State Estimation Intruction Collection Prompt

{image}
You are now an agent navigating within an indoor environment. Your current task is as follows: {instruction}. Based on your observations of the surrounding environment, determine your current location and identify which parts of the navigation task have already been completed. Additionally, based on this assessment, decide which direction you should take as your next action in the current situation. Please provide a brief description of your current surroundings and clearly state the direction or action you plan to take next. Combine your observations and state estimations into a clear and concise paragraph.

Figure 1. **State Estimation generation prompt.**

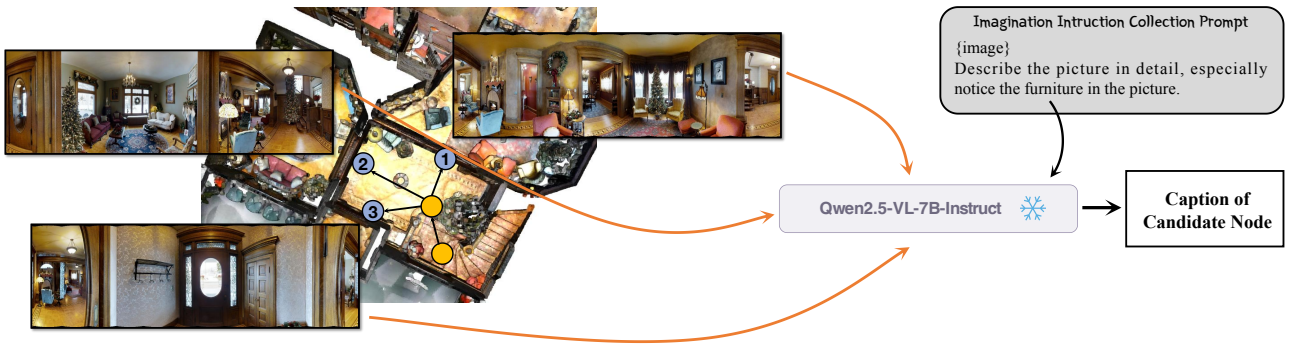

Figure 2. **Imagination Ground Truth Collection.**

possible success if the agent stopped at the closest point to the goal.

$$S_{\text{oracle}} = \begin{cases} 1 & \text{if } \exists v_i \in G \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

or, if no goal set:

$$S_{\text{oracle}} = \begin{cases} 1 & \text{if } \min_i d(v_i, v^*) < \epsilon \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where v_i is the i-th viewpoint along the path.

- **SPL.** A metric combining success and efficiency, penalizing longer trajectories relative to the shortest possible path length.

$$\text{SPL} = S \times \frac{L_{\text{gt}}}{\max(L_{\text{traj}}, L_{\text{gt}})}, \tag{7}$$

where L_{gt} is the ground-truth path length, L_{traj} is the predicted path length.

2.2. Training Details

Following NavGPT2 [7], our best-performing model is trained with additional synthetic data from PREVALENT [4]. We conducted an ablation study by excluding the PREVALENT data and observed that this synthetic data is vital in preventing our method from overfitting. Without incorporating the synthetic data, the validation loss plateaus prematurely during the early stages of training. The parameter size of our model is 1.5B because only the encoder was used during policy training, resulting in half of the parameters of the Flant5-XL model.

3. Additional Experiment Results

Table 1. **Compare FLOPs with visual-imagination methods**.

	PathDreamer [5]	NWM [2]	Ours
GFLOPs ↓	6622.08	813.92	**113.30**

Comparison of FLOPs. We compute the per-step inference FLOPs of visual imagination methods using their released code and checkpoints, evaluated on an A800 GPU. The results in Tab. 1 show that ATD requires substantially lower computational cost. In addition, NWM [2] generates images at a low resolution of 224×224, while PathDreamer [1] generates panoramic images. This suggests that higher-resolution imagination generally incurs higher computational overhead.

State Estimation System Prompt

You are navigating in an indoor environment given the instruction: {instruction}; The navigable locations are listed below: {'Candidate {num}, facing {angle} degrees, front': <image tokens>...}; Please estimate current state and choose the next direction.

Imagination System Prompt

You are navigating in an indoor environment given the instruction: {instruction}; The navigable locations are listed below: {'Candidate {num}, facing {angle} degrees, front': <image tokens>...}; Please imagine the possible objects or furniture in each candidate direction and choose the one that best matches the instructions you should follow.

Figure 3. **State Estimation System Prompt and Imagination System Prompt.**

Table 2. **Comparison of different LLMs.**

Methods	#	Val Seen					Val Unseen				
		TL	NE↓	OSR↑	SR↑	SPL↑	TL	NE↓	OSR↑	SR↑	SPL↑
$ATD_{FlanT5\text{-}XL}$	1	13.02	3.34	74.24	69.44	61.72	13.68	3.37	74.37	67.52	56.01
$ATD_{FlanT5\text{-}XXL}$	2	13.08	2.98	79.43	73.65	65.25	13.25	3.18	79.61	71.31	60.07

Table 3. **Evaluation on SOON `val unseen` split.**

Methods	**SOON**		
	OSR↑	**SR↑**	**SPL↑**
GBE [8]	28.54	19.52	13.34
DUET [3]	**50.91**	**36.28**	22.58
ATD	37.91	29.43	**23.41**

Effect of the LLM model. To demonstrate the transferability of our method, we also implemented the ATD approach on a different LLM, as shown in the Tab. **??**. FlanT5-XL has 1.5 billion parameters, while FlanT5-XXL has 5 billion. It can be observed that $ATD_{FlanT5\text{-}XL}$ achieves slightly higher SR on both `val seen` and `val unseen` sets, whereas $ATD_{FlanT5\text{-}XLL}$ shows a significant advantage in SPL. This might be because $ATD_{FlanT5\text{-}XXL}$ has higher confidence in reaching the destination and thus stops early, resulting in a higher SPL. However, if the decision is incorrect, it could lead to a lower SR.

More Results on SOON. As shown in the Tab. 3, we train and evaluate ATD on the SOON [8] dataset. Our method achieves strong performance in SPL, but remains inferior to the best-performing approaches in terms of SR and OSR. Notably, we do not incorporate the localization head; instead, we directly apply the original ATD architecture to the SOON dataset. Consequently, the RGSPL metric cannot be computed for our model, which also helps explain its suboptimal performance on SR and OSR. Nevertheless, the relatively high SPL score indicates that our imagined key-semantic strategy enables the agent to more efficiently infer the direction and path toward the target object, reflecting strong instruction-following capabilities.

More Visualization of Navigation Progress. During the navigation stage, our LLM operates in a decoder-free manner. To demonstrate the role of the LLM in predicting key semantics throughout navigation, we project all selected features into a shared latent space and compute their cosine similarity. Specifically, we extract the key semantics generated during navigation and feed the corresponding text into the LLM encoder to obtain the latent representation ($feature1$). Meanwhile, we obtain $feature2$ from the latent outputs of the imagination LLM during navigation and compute its cosine similarity with the key-semantic representation ($feature1$). This procedure ensures that all representations reside in a common latent space, allowing cosine similarity to meaningfully capture their relative distances. The detailed computation process is illustrated in Fig. 4. Notably, the two sets of features shown in the figure are computed independently, with no interaction between their respective inputs within the model. As shown in Fig. 5, we illustrate how the cosine similarity of different key semantics evolves throughout the navigation process. We manually select specific key semantics as anchors to examine how the latent representations change over time. The bar chart below reflects the relative similarity values; due to the modality gap, the computed cosine similarities remain relatively low, with the highest value around 0.30. Moreover, because the agent frequently encounters hallways during navigation, the similarity associated with 'hallway' remains consistently high throughout the process.

Visualization of Attention Matrix. As shown in Fig. 6, we visualize the attention matrices $\mathbf{A}$ produced by the SGCA at each layer. For this visualization, the number of SGCA layers is set to 4. As shown in **Row 1** of Fig. 6, a region within the red box gradually gains emphasis during the information interaction between state estimation and imagination. By the final layer (Layer 4), its attention weight becomes significantly higher compared to earlier layers. This suggests that SGCA may indeed play a role in filtering and highlighting the more important information within the imagination. In **Row 2**, the region within the orange box shows a noticeable decrease in attention weight as it propagates through the layers, becoming minimal in the final output. This may indicate that, under the supervision of state estimation, SGCA learns to suppress certain parts of the infor-

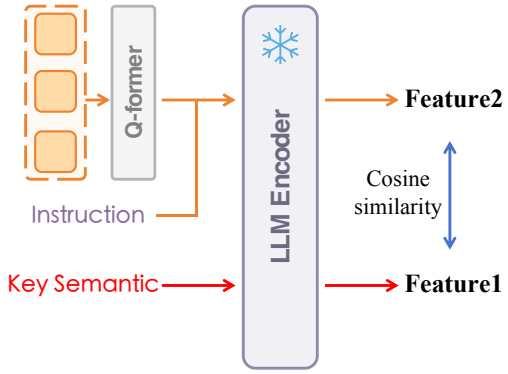

Figure 4. **Computing the similarity between the latent features of the Imagination LLM and the key semantics.**

mation from imagination—potentially those related to navigation steps that have already been completed.

4. Discussion

Limitation and Future Work. Currently, the data collected for training the imagination LLM is limited to candidate nodes one step ahead of the current node. This may restrict the model's ability to perform long-horizon imagination. Future work could explore incorporating long-horizon imagination capabilities to more fully leverage the potential of the LLM's imaginative reasoning.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. *arXiv preprint arXiv:2412.03572*, 2024. 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. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16516–16526, 2022. 3

[4] Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. 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*, pages 13137–13146, 2020. 2

[5] Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Pathdreamer: A world model for indoor navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14738–14748, 2021. 2

[6] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 1

[7] Gengze Zhou, Yicong Hong, Zun Wang, Xin Eric Wang, and Qi Wu. Navgpt-2: Unleashing navigational reasoning capability for large vision-language models. In *European Conference on Computer Vision*, 2024. 2

[8] 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

Instruction: Exit theater room and walk straight down the hallway. Continue straight down the hallway pass the kitchen on the right and the bathroom on the left. Turn right into the sitting area and stop near the chair.
Key Semantic: hallway kitchen bathroom chair

Figure 5. **Visualization of Navigation Progress.** The entire navigation sequence consists of five steps. For each step, we compute the similarity between the latent output of the Imagination LLM and the selected key semantics.

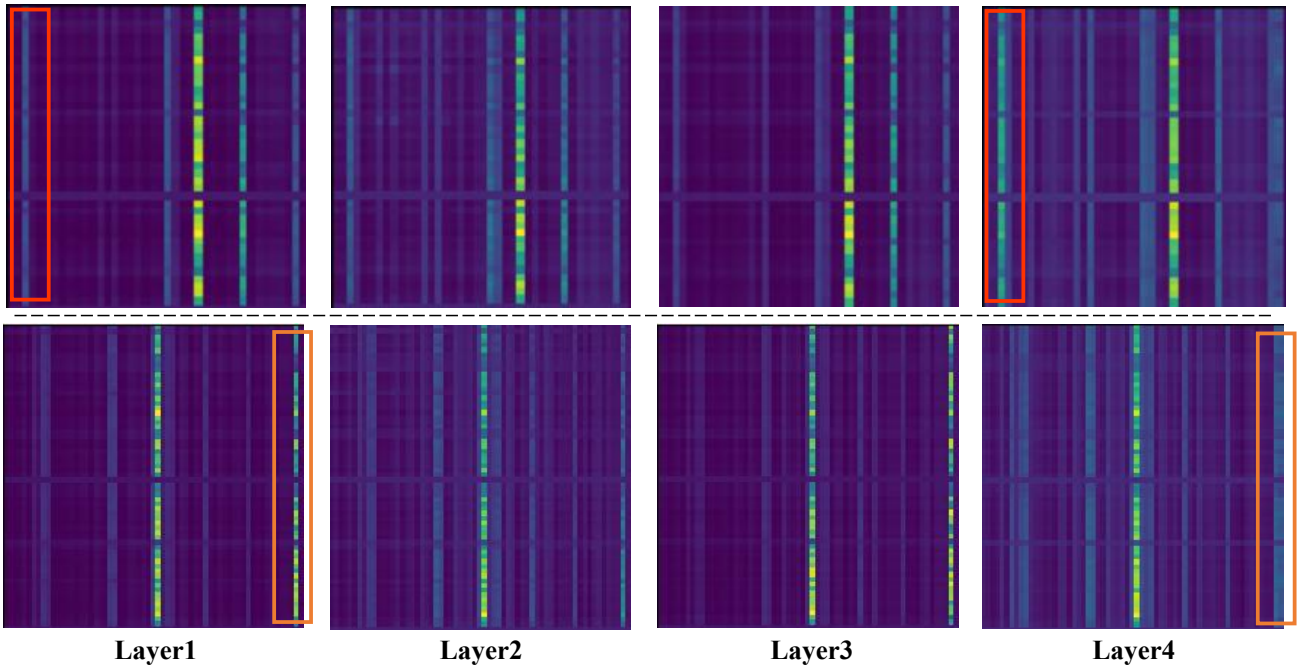

Figure 6. **Visualization of attention matrices across four SGCA layers. Row 1**: The red box shows the increasing emphasis on important information, while **Row 2**: The orange box illustrates the suppression of completed navigation steps.