COSMO: Combination of Selective Memorization for Low-cost Vision-and-Language Navigation

Supplementary Material

8. Experimental Setups

8.1. Datasets

All the datasets are built upon the Matterport3D [5] environment which contains 90 photo-realistic houses. Each house is annotated with a navigation graph, wherein the agent's movement is exclusively confined to traversing along the interconnected edges of nodes.

R2R [2] provides step-by-step instructions. The houses are divided into 4 sets: 61 houses for training, which are annotated with 14039 instructions; 11 and 18 houses for validation and testing in unseen environments, respectively, which are annotated with 2,349 and 41,73 instructions. Among the 61 houses for training, 56 houses are also utilized for validation in seen environments. The instructions provided in the R2R dataset describe each action the agent should take during navigation, such as "Walk straight toward the bar with the chairs. Turn left and go straight until you get to three tables with chairs. Turn left and wait near the couch". The R2R dataset consists of 21,567 navigation instructions across 7,189 paths and 10,800 panoramic views within 90 sizable real-world indoor settings. Instructions in this dataset average 29 words in length.

R2R-CE [36] transfers 77% of R2R paths into continuous environments, resulting a total of 5,611 paths and an average path length of 9.89m. Each instruction contains an average of 32 words. Agents have a chassis radius of 0.1m and are allowed to slide along obstacles.

REVERIE (Remote Embodied Referring Expression) [56] provides coarse-grained instructions. The instructions are usually concise and mainly describe the destinations and target objects, such as "Go to the bathroom with a bronze bathtub and bring me the towel above the bathtub". The training set contains 60 houses and 10,466 instructions, the validation seen set contains 46 houses and 1,423 instructions, the validation unseen set contains 10 houses and 3,521 instructions, the test unseen set contains 16 houses and 6,292 instructions. It comprises 21,702 instructions, with each instruction averaging 18 words in length. Although the trajectories in REVERIE align with that in R2R, the task presents a significantly higher level of difficulty due to the absence of explicit action guidance in the instructions, necessitating active exploration but the agent to locate the destination. Given that coarse-grained instructions bear closer resemblance to real-world scenarios, recent research has predominantly focused on the dataset.

	DUET	KERM	COSMO
OSR↑	51.07	55.21	**56.09**
SR↑	46.98	50.44	**50.81**
SPL↑	33.73	35.38	**35.93**
Params(M)↓	181	222	**28**
FLOPs(G)↓	4.95	15.24	**0.46**
MACs(G)↓	4.74	15.04	**0.34**
Inf. Time(s)↓	13.20	526.33	**10.64**
Train Speed (sample/s)↑	29	2	36

Table 6. Comparison of navigation performance and computational costs between DUET, KERM, and COSMO on the validation unseen set of the REVERIE dataset.

	Total	Text	Others
DUET	180.5	87.6	92.9
COSMO	27.6	14.5	13.3

Table 7. Detailed parameter comparison.

8.2. Training Details

R2R. Following previous works [8, 9, 29], we employ augmented data [27] for pre-training. The model undergoes 100k steps of pre-training with a batch size of 64, followed by fine-tuning for 20k steps with a batch size of 8.

R2R-CE. We transfer the model pretrained on the R2R dataset to continuous environments through the Habitat Simulator [63]. The model is finetuned with a batch size of 16 and a learning rate of 1e-5 for 30 epochs.

REVERIE. Following DUET [9], we incorporate augmented data generated by the speaker model during the pre-training phase. The model is pre-trained for 100k steps with a batch size of 32, then finetuned for 20k steps with a batch size of 8.

9. More Comparisons

Tab. 6 presents the values in the radar chart on the left side of Fig. 1 in the paper, along with the comparison with KERM. Considering that all the metrics pertaining to computational cost aim for lower values indicating better performance, reciprocal transformations of these metrics are taken in Fig. 1 in the paper. `FLOPs` is calculated by the fvcore[1] library. `MACs` is calculated by the thop[2] library. As

[1]https://github.com/facebookresearch/fvcore.git
[2]https://github.com/Lyken17/pytorch-OpCounter.git

RSS	NE↓	SR↑	SPL↑
Mamba	3.34	69.39	57.18
Bi-Mamba	3.20	70.80	58.13
✓	3.15	72.58	60.68

Table 8. Ablation results of the RSS module on R2R validation unseen split.

	Components		REVERIE Val Unseen		
#	RSS	CS3	OSR↑	SR↑	SPL↑
1	Self-Attn	✓	53.71	48.82	32.38
2	✓	Cross-Attn	50.21	44.42	30.96
3	✓	✓	**56.09**	**50.81**	**35.93**

Table 9. Ablation results of the RSS and CS3 module.

Size	OSR↑	SR↑	SPL↑
8	56.01	50.24	34.36
16	**56.09**	**50.81**	**35.83**
32	54.27	48.54	33.57

Table 10. Ablation on the state space size of RSS and CS3.

illustrated in Section 5.3, inference time denotes the time required for one-step navigation on the REVERIE validation unseen split. We also report training speed (referred to as Train Speed in the table). It denotes the number of samples trained per second by the model with a batch size of 32 on a single A6000 GPU. It prefers higher value. In Tab. 7, we present a detailed comparison of the parameters between DUET and COSMO. After excluding the embedding layer and the text encoder, COSMO contains 13.3M parameters, amounting to merely 14.3% of DUET's parameters (92.9M).

10. More Ablations

Effectiveness of RSS. To further validate the effectiveness of the RSS module, we conduct ablation studies on the R2R validation unseen split, as shown in Tab. 8. RSS is replaced with vanilla Mamba and Bi-Mamba layers. Results show that RSS yields about 2% increase in both SR and SPL over Bi-Mamba.

Superiority of RSS and CS3. Tab. 4 presents the ablation results of the RSS and CS3 modules without altering the hybrid architecture. To further demonstrate the efficacy of these two modules, additional ablation results are provided in Tab. 9. RSS is replaced with self-attention in row #1, resulting in a decrease of 2.0% in SR and 3.6% in SPL. This indicates that RSS effectively captures contextual relationships among tokens while efficiently compressing information into the class token. CS3 is replaced with cross-attention in row #2, leading to a significant decrease of 6.39% in SR and 4.97% in SPL. This not only highlights the necessity of employing a hybrid architecture but also demonstrates the proficiency of CS3 in modeling the interaction between modalities and their mutual selection.

State space size. Table 10 compares different state space sizes in our RSS and CS3. It can be observed that inadequate state space leads to insufficient retention of navigation history, while an excessively large state space results in the inclusion of redundant or noisy information during selection process.

	R2R Val Unseen			RVR Val Unseen		
Instr len	< 20	$20 - 40$	> 40	< 15	$15 - 30$	> 30
GT path len	5.73	6.03	6.24	5.85	5.99	6.12

Table 11. Average ground-truth path length associated with instructions across different length intervals.

	R2R Val Unseen				RVR Val Unseen			
GT path len	4	5	6	7	4	5	6	7
Instr len	16.7	23.9	25.8	29.2	16.9	17.6	18.7	19.4

Table 12. Average instruction length associated with varying ground-truth lengths.

11. Discussions

Correlation between Instruction Length and Complexity of Navigation Tasks. As the length of instructions increases, navigation tasks tend to become increasingly complex. We quantify the complexity of navigation through the length of ground-truth path. The ground-truth paths in both R2R and REVERIE exhibit a length distribution ranging from 4 to 7. Table 11 presents the average ground-truth path lengths associated with instructions of varying lengths. Table 12 presents the average instruction length associated with varying lengths of ground-truth path.

Complementary to Existing Methods. Given our focus on enhancing the fundamental model structure, we employ DUET [9] as the baseline model, which is widely recognized as a strong benchmark in recent literature. Our proposed RSS and CS3 can be integrated with other SoTA methods such as advancing local perception in BEVBert [1]. Results of the integration of COSMO and BEVBert are reported in Tab. 1.

Advantages of CS3 over Mamba. As illustrated in Table 4, the performance enhancement of CS3 over Bi-Mamba is substantial ($+3.86\%$ in SR and $+4.53\%$ in SPL). This is attributed to the effective modal alignment facilitated by the dual-stream architecture of CS3. Although Mamba has been effectively utilized in the multimodal domain, such

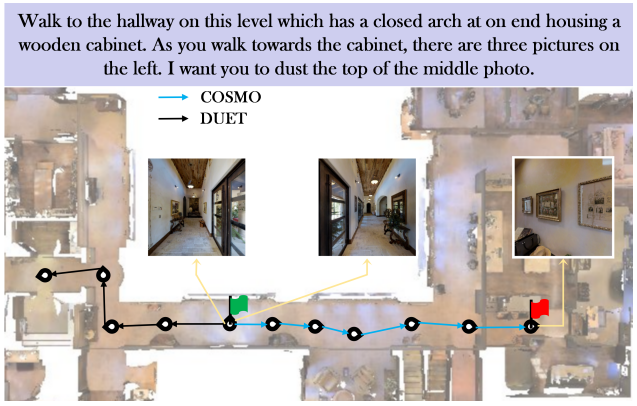

Figure 6. Predicted paths of DUET and COSMO on REVERIE validation unseen set. The red flag denotes the correct endpoint.

as VL-Mamba [61] and Cobra [81], these models perform modal alignment prior to inputting visual and textual features into Mamba language models. In contrast, VLN models, like HAMT [8] and DUET [9] along with their variants, employ cross-modal attention mechanisms for both multi-modal alignment and interaction. This distinction is a key reason why SSMs cannot be directly applied to VLN tasks. As illustrated in Equ(3) and Equ(4), the influence of the input token x_t at time t on the state space is controlled by matrix $\mathbf{B}_t$, while the resolution of the input is determined by $\mathbf{\Delta}_t$. When tokens in the input sequence originate from different modalities and are not aligned, it is evidently inappropriate to apply the same strategy (S_B and S_Δ) for controlling their impact on the state space and the sampling frequency. In this context, we propose CS3 as a dual-stream selective SSM. As illustrated in Algorithm 1, for instance, x represents visual features, and y represents textual features. Now the objective is to utilize the textual features to update the visual features. Thus, the input to the state space is y, while the output of state space acts upon x. That is to say, input matrix $\mathbf{B}$ and resolution matrix $\mathbf{\Delta}$ should be derived from y, whereas the output matrix $\mathbf{C}$ should be derived from x. This design facilitates effective alignment and interaction between multi-modal features.

Limitations. The limitation of COSMO lies in its longer navigation trajectory. On REVERIE val unseen split, the Trajectory Length (TL) of DUET is 22.11, whereas COSMO's TL is 23.08. The examples in Fig. 5 show that COSMO requires extensive exploration and backtracking to identify the correct direction. In future work, we plan to incorporate common-sense knowledge to assist COSMO in more rapidly determining the correct path and thereby reducing its TL.

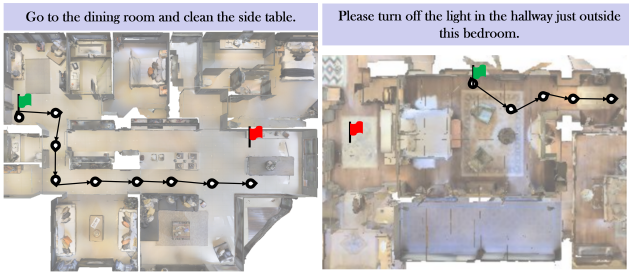

Figure 7. Failure cases of COSMO on REVERIE validation unseen set.

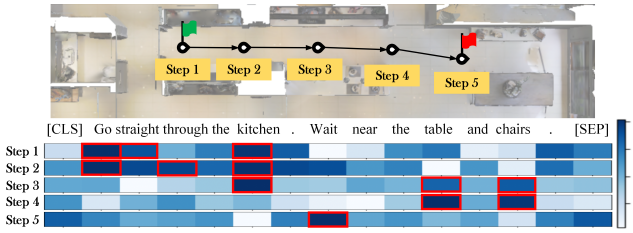

Figure 8. Visualization of attention maps.

12. More Qualitative Examples

We visualize the predicted paths of DUET and COSMO in Figure 6. The instruction requires navigating to the end of a hallway featuring a closed arch. Given that the starting point is midway along a lengthy corridor, it is crucial to accurately identify the closed arch. DUET failed to select the correct direction, which ultimately led to its inability to locate the target destination. In contrast, COSMO successfully identified the direction containing the closed arch, demonstrating its ability to accurately ground the objects as described in the instruction within the environment. Consequently, COSMO finds the three paintings mentioned.

In Figure 7, we visualize two failure cases. In the left example, COSMO successfully located the dining room. However, it was unable to navigate closer to the side table. In the right example, COSMO was able to identify the light in the hallway, but the presence of two corridors surrounding the room led to ambiguity in the instructions, resulting in an error.

We analyze the attention weights from the last cross-attention layer of the global cross-modal encoder for the node ultimately selected by COSMO. As shown in Fig.8, COSMO attends to the first sentence in steps 1–2, highlights landmarks in step 3, and shifts focus to the second sentence in steps 4–5. These observations demonstrate COSMO's ability to align language with visual input.