

Improving Embodied Instruction Following with Deterministic Methods

Dahyun Kim*, Yong-Ju Lee**

* Dept. of Electrical Engineering, Korea University

** Electronics and Telecommunication Research Institute, Daejeon, Korea

dahyunkeem@korea.ac.kr, yongju@etri.re.kr

Abstract—We propose a modified module structure of FILM, which increases its performance on ALFRED benchmark. Despite methods like imitation learning or reinforcement learning can be used, we will show it is possible to improve the performance by simply modifying the deterministic policy module. our approach aims to use the semantic segmented visual observation when navigating to the target location. Additionally, we propose a method of adaptive decision making for active search. Through active search, the agent can search for the target object by opening closed receptacles. Also, we provide some minor adjustments which helps the agent overcome some semantic segmentation errors. We evaluate the modified FILM with the ALFRED benchmark and achieved the performance of 25.37% with valid seen split and 20.46% with valid unseen split, thus succeeding in improving the performance of the original architecture with a small margin. Additionally, the modified FILM showed a significant improvement in the result of task type “Cool & Place” with a large margin (8.82%), which shows that active search is effective.

Keywords— Embodied Instruction Following, FILM, ALFRED benchmark, Embodied AI

I. INTRODUCTION

The development of robots that can help household tasks always have been a long-standing topic of interest in robotics [1]. Thanks to the recent progress of Deep Learning, it has been possible for agents to process multimodal data and complete complex tasks. Thus, challenges of Embodied AI using various benchmarks [6, 12, 13, 16, 17, 18] are growing more attention. In the Field of Embodied Instruction following (EIF), the agent needs to understand instructions and execute sequential of subtasks involving interaction with objects and state changes of environment. ALFRED (The Action Learning from Realistic Environments and Directives) Benchmark [13] is one of the embodied AI datasets that provides numerous language instruction data and interactive simulation environment, therefore allowing developers to test their agent’s ability to follow instruction in known and unknown environment (**Figure 1**).

Early works [5, 11, 13, 15] approached to the problem with end-to-end method or relied on neural memory, however failed to achieve proper performance showing that EIF remains a very challenging task for end-to-end approaches. Hence, the starter code of the ALFRED challenge, FILM (Following Instructions in Language with Modular methods)

[9] tackled the problem with modular method. A notable point of this approach is that the action module, which determines the agent’s both navigation action and object interaction action, is not learned but deterministic. Yet, FILM’s architecture poses several problems. The navigation action relies too much on the semantic map, which is very likely noisy due to the agent’s incomplete perception of the environment. Also, the agent does not make adaptive decision making, and the exploration is not “active” enough, so the agent often fails the task if the target object is not immediately observable. Besides the architecture drawbacks, the performance of semantic segmentation is also essential since the semantic segmentation determines the agent’s visual perception.



Figure 1. ALFRED(Action Learning From Realistic Environments and Directives) benchmark. It consists of 25k language directives corresponding to expert demonstrations of household tasks.

In this paper, we propose a modified method to increase FILM’s performance on ALFRED benchmark. Despite methods like imitation learning or reinforcement learning can be used [10], we will show it is possible to improve the performance by simply modifying the deterministic policy module. Our approach aims to use the semantic segmented visual observation when navigating to the target location to reduce the excessive reliance on semantic map. Additionally, we propose a method of adaptive decision making for active search. Through active search, the agent is able to search for the target object by opening closed receptacles. Also, we

provide minor adjustment to overcome some semantic segmentation errors.

We evaluate the modified FILM with the ALFRED benchmark and achieved the performance of 25.37% with valid seen split and 20.46% with valid unseen split, thus succeeding in improving the performance of the original architecture with a small margin. Additionally, the modified FILM showed a significant improvement in the result of task type “Cool & Place” with a large margin (8.82%), which shows that active search is effective.

II. RELATED WORKS

Numerous works have been published in the field of embodied vision and language tasks, Visual Language Navigation (VLN) for example. A lot of progress has been made since VLN was first introduced, SOTA models showing success rate of about 61% on unseen test environments [4].

However, VLN is somewhat unreal and artificial compared to the tasks in real life. The task is mostly focused on navigation, and the environment is relatively static and does not interact with the agent. Consequently, the concept of Embodied Instruction Following was introduced to reduce the gap between tasks in simulation environment and tasks in real life. ALFRED [13] is a benchmark developed for checking the progress of EIF. In ALFRED, the language instruction is given by both high-level goal and low-level instruction. With these language instructions, the agent needs to complete various and complicated household tasks involving interaction with the environment.

Early approach [11, 13, 15] used end-to-end method that included either concatenating language, visual features, and previous actions into one input or training a large end-to-end model. Later works [5, 14, 19] tackled the problem with more modular methods, modularly processing raw language and visual inputs into structured form. However, they still involved training large end-to-end module. Recent method [2, 9] became more modular, and used semantic mapping to represent the environment. Especially FILM [9], the starter code of ALFRED challenge, consists of four modules: Language Processing, Semantic Mapping, Semantic Search Policy, and Deterministic policy. All modules except the Deterministic Policy is learned. The Deterministic policy is the action module of FILM that decides the agent’s next action; thus, it is crucial to the model’s performance. Despite FILM is a significant model in the ALFRED challenge, its architecture poses some drawbacks.

1) Navigation relies too much on the semantic map: The agent determines if it’s arrived at the destination by the coordinates in the semantic map. This may cause error in navigation because the semantic map is often noisy, especially when multiple receptacles and target objects are located close to each other.

2) Search is not active: FILM does not open closed receptacles when navigating. Thus, the agent fails the task if the target object is not immediately observable, especially when it is located inside close receptacles.

3) Semantic segmentation error: In the task of EIF, including the ALFRED challenge, the agent relies heavily on semantic segmentation when navigating. Thus, accurate semantic segmentation result of Visual observation is essential for following instruction.

To migrate the mentioned drawbacks, we propose the modified deterministic policy. With the modified deterministic policy, the agent uses visual observations for locating the target object and performs “active” search if the target object is not observed in the current environment by changing the list of subtasks. Additionally, we provide minor changes that can help the semantic segmentation error.

III. TASK EXPLANATION

A. Task Background

Our work is done in the context of the ALFRED benchmark. In ALFRED, the agent needs to follow the natural language instruction that is given by the form of both high-level goal and low-level step-by-step instruction. The ALFRED challenge contains 7 task types: Pick & Place, Stack & Place, Pick Two & Place, Clean & Place, Heat & Place, Cool & Place and Examine in Light. Each task types are parameterized by 84 object classes in 120 scenes. The task takes place in simulated home environments, which is provided by the AI2-THOR simulator [6].

For the data splits, the validation set, and test set are split into seen and unseen folds. Scenes in the seen folds are subset of those in the train set, and scenes in the unseen fold are distinct from the train set.

B. Task Description

At each timestamp t , the agent gets a new visual observation O_t , and it needs to take an action A_t . The agent can choose among 13 actions. These actions include 5 navigation actions: RotateRight, RotateLeft, MoveAhead, LookUp, LookDown, and 7 interaction actions: PickupObject, PutObject, OpenObject, CloseObject, ToggleObjectOn, ToggleObjectOff, and SliceObject. The interaction actions need a pixelwise mask to denote the target object in the agent’s first-person view (Egocentric RGB). To end an episode, the agent needs to choose a special Stop action.

IV. METHOD

Our proposed method consists of three main parts. Use of visual observation when navigating, active search and modification of subtasks, and semantic segmentation error fixes. The modified structure is illustrated in **Figure 2**.

A. Use of Visual Observation when navigating

To reduce the reliance on the semantic map, the visual observation is used when navigating as illustrated in Figure 1. If the current goal object is observed in current visual observation frame, the search goal coordinates is changed to current location. This prevents the agent from neglecting the goal object in front of it.

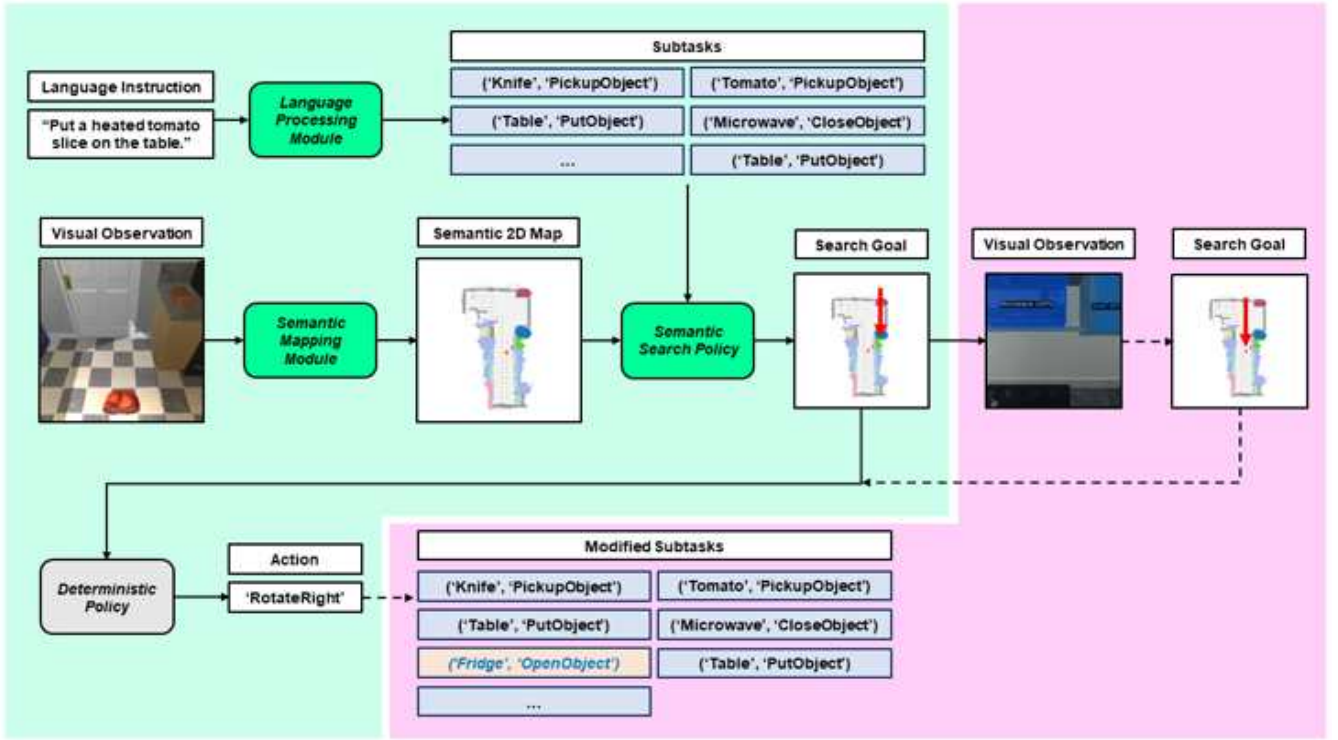


Figure 2. Model Overview. The background color denotes the original and the modified architecture of FILM (green: original architecture, pink: modified architecture). The Search Goal is selected as the coordinates of the target object in the semantic map or selected by the Semantic Search Policy for every 25 steps. But if the target object mask is observed in the Visual Observation, the Search Goal is changed to current location. Additionally, if the goal object is not found after enough steps, the list of subtasks is modified, and the agent opens receptacles for search.

B. Active Search through Modification of Subtasks

For active search, the list of subtasks is modified during an episode. The agent first starts with the original list of subtasks, which is the output of the Language Processing module.

But if the agent cannot locate the goal object in the environment, then new subtask for opening relevant receptacle is inserted in the subtask list. For example, if the goal object is not found and if it can spawn in “Fridge”, the subtask (“Fridge”, “OpenObject”) is inserted the subtask list as in **Figure 2**.

C. Semantic Segmentations Error Fixes

The pretrained Mask-RCNN module struggles in distinguishing non-sliced object and spliced object. It both classifies both objects as non-sliced object, so it is meaningless to differentiate those two in the subtasks.

Thus, we deleted the “sliced” object from the subtask template. Moreover, in the deterministic policy, there’s a part where the agent rotates the camera angle right and left repeatedly if it is arrived at the goal location but cannot observe any goal mask. The agent often trap itself in a camera rotation loop if it fails in semantic segmentation.

This occasionally happens when the goal object is only observed partially in the visual observation frame. Thus, we

modified the algorithm to move behind to get the whole picture if camera rotation loop happens.

V. EXPERIMENTS AND RESULTS

We evaluated the modified model on ALFRED benchmark (**Figure 3**).

Table 1 shows the modified method’s performance on valid seen split and valid unseen split against the original FILM.

Table 2 shows the performance by task type. Note that among closed receptacles, only “Fridge” is added to active search in these results. As shown in **Table 1**, our approach performs slightly better than the original method both on Valid Seen and Unseen, showing higher SR by 0.74% and 0.36% each.

On the other hand, GC is decreased by 1.42% on Valid Seen split and also decreased by 0.86% on Valid Unseen split. Notably, for the performance by task type shown in **Table 2**, task type “Cool & Place” improved with a significant performance increase (8.82%).

It appears that this most likely improved the overall result. The notable improvement on “Cool & Place” indicates that the active search functioned well, and the navigation of the goal object greatly affects the SR.

TABLE 1. PERFORMANCE ON VALIDATION SEEN AND UNSEEN SPLITS. BOLD NUMBERS ARE TOP SCORES IN EACH SECTION. FOR THE SCORES FOR FILM, WE REFERRED TO THE RESULT OF THE ORIGINAL PAPER.

Method	Valid Seen				Valid Unseen			
	PLWGC	GC	PLWSR	SR	PLWGC	GC	PLWSR	SR
FILM	-	37.20	-	24.63	-	32.45	-	20.10
Our Approach	13.17	35.78	9.86	25.37	9.62	31.59	7.54	20.46

TABLE 2. PERFORMANCE BY TASK TYPE. BOLD NUMBERS ARE TOP SCORES IN EACH TASK TYPE

Method	<u>FILM</u>		<u>Our Approach</u>		<u>FILM</u>		<u>Our Approach</u>	
	Valid Seen				Valid Unseen			
	GC	SR	GC	SR	GC	SR	GC	SR
Examine	50.00	64.41	51.61	37.23	45.06	29.65	46.09	31.21
Pick & Place	27.46	26.92	28.19	28.17	16.67	16.03	11.82	11.00
Stack & Place	23.74	10.71	23.20	10.43	9.90	1.98	10.56	2.75
Clean & Place	58.56	44.04	55.07	41.07	48.89	33.63	46.53	33.63
Cool & Place	27.04	<u>12.61</u>	30.96	<u>21.43</u>	27.41	14.04	27.71	18.35
Heat & Place	40.21	22.02	33.33	23.36	37.77	23.02	33.33	22.79
Pick2 & Place	40.37	23.77	32.28	18.55	29.28	11.84	10.56	2.75

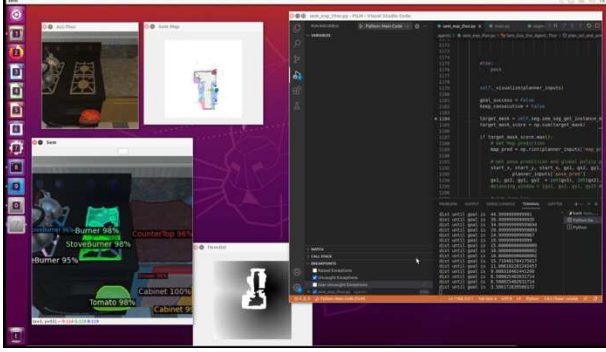


Figure 3. Simulation Result. We evaluated the modified model on valid seen and unseen split of ALFRED benchmark. The agent’s visual observation O_t , the semantic segmentation result of O_t , and the semantic map are shown in the figure. ALFRED benchmark uses AI2-THOR environment that look similar to real world scens and is customizable.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we analyse FILM’s drawbacks of FILM’s architecture, and discuss how to improve the performance by modifying the deterministic policy. We proposed the method of using visual observation while navigating and active search. The use of visual observation reduces the reliance on the semantic map, and through active search, the agent can search for the target object by opening closed receptacles. Also, we provide minor adjustment to overcome some semantic segmentation errors. We succeed in improving the performance by a little margin, showing that modifying deterministic method can improve the model without any learn-based method. Additionally, this also means that deterministic algorithms play an important role in the ALFRED benchmark, giving the AI community a challenge.

As future work, following topics need to be considered. Use of egocentric RGB in the proposed method is naïve. More suitable way of using the vision observation is needed for more human-like and accurate navigation. Reinforced learning may be implied. Furthermore, more closed receptacles need to be added to the active search. Addable receptacles can be Cabinet, Microwave, Drawer, Safe and Box. Moreover, current method adds active search after certain number of steps. This method needs to be improved. Also, changing the list of subtasks could be implemented other ways than active search for more efficient adaptive decision making. Finally, CNN backbone other than similar family of Mask-RCNN should be tested on the ALFRED dataset. Most of the related works use similar models for semantic segmentation: Mask-RCNN and Fast-RCNN. Semantic segmentation problem may be solved by using other CNN models.

ACKNOWLEDGMENT

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00871, Development of AI Autonomy and Knowledge Enhancement for AI Agent Collaboration)

REFERENCES

- [1] Beetz, M., Klank, U., Kresse, I., Maldonado, A., Mosenlechner, L., Pangercic, D., Ruhr, T., & Tenorth, M. (2011). Robotic roommates making pancakes. *2011 11th IEEE-RAS International Conference on Humanoid Robots*
- [2] Blukis, V., Paxton, C., Fox, D., Garg, A., & Artzi, Y. (2021). A Persistent Spatial Semantic Representation for High-level Natural Language Instruction Execution. *Proceedings of the Conference on Robot Learning (CoRL)*. *arXiv:2107.05612v3 [cs.RO]*

- [3] Chaplot, D. S., Gandhi, D., Gupta, A., & Salakhutdinov, R. (2020). Object Goal Navigation using Goal-Oriented Semantic Exploration. *arXiv:2007.00643 [cs.CV]*
- [4] Ke, L., Li, X., Bisk, Y., Holtzman, A., Gan, Z., Liu, J., Gao, J., Chli, J., & Srinvasa, S. (2019). Tactical rewind: Self-correction via backtracking in vision-and-language navigation. *Proceedings of the ICCV/CVF Conference on Computer Vision and Pattern Recognition*
- [5] Kim, B., Bhambri, S., Singh, K. P., Mottaghi, R., & Choi, J. (2021). Agent with the Big Picture: Perceiving Surroundings for interactive Instruction Following. *Embodied AI Workshop CVPR 2021*
- [6] Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Deitke, M., Ehsani, K., Gordon, D., Zhu, Y., Kembhavi, A., Gupta, A., & Farhadi, A. (2017). AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv:1712.05474*
- [7] Liu, H., Liu, Y., He, H., & Yang, H. (2022). LEBP-Language Expectation & Binding Policy: A Two-Stream Framework for Embodied Vision-and-Language Interaction Task Learning Agents. *arXiv:2203.04637 [cs.AI]*
- [8] Liu, X., Palacios, H., & Muise, C. (2022). A Planning based Neural-Symbolic Approach for Embodied Instruction.
- [9] Min, S., Chaplot D. S., Ravikumar, P., Bisk, Y., & Salakhutdinov, R. (2022). FILM: Following Instructions in Language with Modular Methods. *ICLR 2022*
- [10] Murray, M., Graduate Student Member, IEEE, & Cakmak, M. (2022). Following Natural Language Instructions for Household Tasks with Landmark Guided Search and Reinforced Pose Adjustment. *IEEE Robotics and Automation Letters*, 7(3), 6870-6877
- [11] Pashevich, A., Schmid, C., & Sun, C. (2021). Episodic Transformer for Vision-and-Language Navigation. *ICCV 2021*
- [12] Ramakrishnan, S. K., Gokaslan, A., Wijmans, E., Maksymets, O., Clegg, A., Turner, J., Undersnader, E., Galuba, W., Westbury, A., Chang, A. X., Savva, M., Zha, Y., & Batra, D. (2021). Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D Environments for Embodied AI. *arXiv:2109.08238v1 [cs.CV]*
- [13] Shridhar, M., Thomason, J., Gordon, Y., Han, W., Mottaghi, R., Zettlemoyer, L., & Fox, D. (2020). ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. *CVPR*
- [14] Singh, K. P., Bhambri, S., Kim, B., Mottaghi, R., & Choi, J. (2021). Factorizing Perception and Policy for Interactive Instruction Following. *arXiv:2012.03208v3 [cs.AI]*
- [15] Suglia, A., Gao, Q., Thomason, J., Thattai, G., & Sukhatme, G. (2021). Embodied BERT: A Transformer Model for Embodied, Language-guided Visual Task Completion. *arXiv:2108.04927 [cs.CR]*
- [16] Talbot, B., Hall, D., Zhang, H., Bista, R. S., Smith, R., Dayoub, F., & Sunderhauf, N. (2020). BenchBot: Evaluating Robotics Research in Photorealistic 3D Simulation and on Real Robots. *arXiv:2008.00635v1 [cs.RO]*
- [17] Weihs, L., Deitke, M., Kembhavi, A., & Mottaghi, R. (2021). Visual Room Rearrangement. *CVPR 2021*
- [18] Xia, F., Shen, W. B., Li, C., Kasimbeg, P., Tchammi, M. E., Toshev, A., Martin-Martin, R., & Savarese, S. (2020). Interactive Gibson Benchmark: A Benchmark for Interactive Navigation in Cluttered Environments. *IEEE Robotics and Automation Letters* 5(2), 713-720
- [19] Zhang, Y., & Chai, J. (2021). Hierarchical Task Learning from Language Instructions with Unified Transformers and Self-Monitoring. *ACL 2021 Findings*



Dahyun Kim. She is currently an undergraduate student in the School of Electrical Engineering at Korea University. She worked as an intern at Electronics and Telecommunication Research Institute from 2022.07–2022.08



Yong-Ju Lee received the B.S. degree and M.S. degree in computer engineering from Chonbuk National University, Jeonju, Republic of Korea, in 1999 and 2001, respectively, and the Ph.D. degree in computer engineering from the Chungnam National University, Daejeon, Republic of Korea, in 2011. Since 2016, he has also been with the ICT Department, University of Science and Technology (UST), where he is currently an Assistant Professor. Since 2020, he has been working at the leader of Visual Intelligence Research Section, Artificial Intelligence Research Laboratory, Electronics and Telecommunications Research Institute (ETRI), Daejeon. His research interests include edge-based AI, Collaborative Learning Agent, Multimodal/Crossmodal Deep Learning.