ET tu, CLIP? Addressing Common Object Errors for Unseen Environments

Ye Won Byun* Cathy Jiao* Shahriar Noroozizadeh* Jimin Sun* Rosa Vitiello*

Carnegie Mellon University

{yewonb,cljiao,snoroozi,jimins2,rvitiell}@andrew.cmu.edu

Abstract

We introduce a simple method that employs pre-trained CLIP encoders to enhance model generalization in the ALFRED task. In contrast to previous literature where CLIP replaces the visual encoder, we suggest using CLIP as an additional module through an auxiliary object detection objective. We validate our method on the recently proposed Episodic Transformer architecture and demonstrate that incorporating CLIP improves task performance on the unseen validation set. Additionally, our analysis results support that CLIP especially helps with leveraging object descriptions, detecting small objects, and interpreting rare words.

1. Introduction

Embodied instruction following (EIF) tasks entail executing fine-grained navigation and interaction action sequences according to natural language directives. This requires processing and understanding information from heterogeneous sources to successfully navigate and interact with unseen environments [1, 7]. In the multimodal research community, large-scale pre-trained models have been shown to improve multimodal alignment and generalization performance [2,5,8,14]. In particular, several recent works evaluate the CLIP (Contrastive Language Image Pre-training) [10] model's capabilities for embodied AI tasks, including object navigation [3,4] and vision language navigation [6, 11, 13]. The most common approach in this direction has been simply replacing the visual encoder with CLIP's visual encoder.

In this work, we hypothesize that pre-training on large-scale image-text pairs will induce more generalizable multimodal representations, leading to better performance in unseen environments of the ALFRED task [12]. In contrast to previous literature, we propose a simple model-agnostic method to use CLIP as an auxiliary module to take advantage of CLIP's multimodal alignment capabilities. Concretely, we introduce a novel object detection loss without having to change the model's architecture. We investigate the proposed method through preliminary experiments based on the Episodic Transformer (ET) [9] architecture, a competitive system on the ALFRED leaderboard. Our empirical results suggest that our novel loss objective improves generalization to unseen environments, especially by alleviating the difficulty of detecting small objects and interpreting rare words – which are challenging error conditions in current state-of-the-art models.

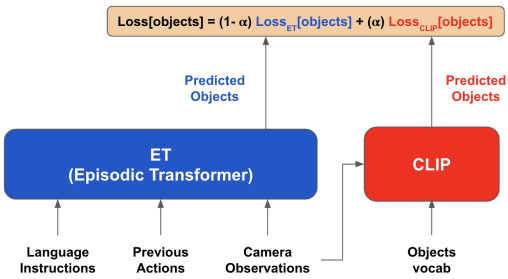

Figure 1. ET-CLIP model as modified from [9]

2. Proposed Approach

We use CLIP [10] as an auxiliary source of information for object detection and interaction by including CLIP as an additional module in ET [9]. During training, we feed camera observation inputs from ET into CLIP along with a list of all ALFRED object words (with "none" also being an option). A predicted object for each camera observation is obtained from both the CLIP module and ET, and we compute their object prediction losses: $\mathcal{L}_{\text{CLIP}}(\text{obj})$ and $\mathcal{L}_{\text{ET}}(\text{obj})$, respectively. The final object loss in ET is as follows:

$$\mathcal{L}(\text{obj}) = \alpha \cdot \mathcal{L}_{\text{CLIP}}(\text{obj}) + (1-\alpha) \cdot \mathcal{L}_{\text{ET}}(\text{obj})$$
$$\text{where } \alpha \in [0,1] \quad \text{(see Figure 1)}$$

Both ET and CLIP weights are updated during training. During inference, the CLIP module is ignored, and object prediction is solely done by ET.

* Everyone Contributed Equally – Alphabetical order

Methods	Success Rate ↑	Goal-Conditioned Success Rate ↑
ET-Baseline [9]	0.1	7.8
ET-CLIP	**1.0**	**7.9**

Table 1. Success rates and goal-conditioned success rates of the baseline Episodic Transformer (ET) model and our ET-CLIP model on the unseen validation set.

Subset	ET	ET-CLIP	Improvement
All	7.8	**7.9**	+ 0.1
Object properties	7.7	**8.0**	+ 0.3
Small objects	5.1	**5.6**	+ 0.5
Rare semantics	5.9	**6.7**	+ 0.8

Table 2. Goal-conditioned success rates on the unseen validation set of the ET-Baseline and ET-CLIP on subsets of instructions.

3. Preliminary Experiments & Results

Experimental setting We run our baseline experiments based on the code released by the authors of the ET paper[1]. More specifically, we use the base ET model, which does not employ the data augmentation strategy. We train both the ET baseline and the ET-CLIP models for 20 epochs, and refer to the original ET model for hyperparameters. The weighting coefficient α of the auxiliary CLIP loss was chosen as 0.5 based on the magnitude of the two loss terms to ensure that the loss ranges are similar in the two models. We note that the discrepancy of our results from [9] stems from different random seeds, as noted by the authors[2].

Results Table 1 shows the results for success rate and goal-conditioned success rate of the ET-Baseline and the ET-CLIP models for the unseen validation splits. As seen in Table 1, the ET-CLIP model performs better in unseen scenes. This suggests that adding CLIP object detection as an auxiliary loss helps with generalization. We further analyze how CLIP aids in performance improvement for specific error conditions, pertaining to task instruction characteristics in Section 4.

4. Analysis

We investigate how integrating CLIP helps the ET model's performance on natural language directives. In particular, we look into three subsets of instructions that contain common sources of error: instructions including fine-grained object properties, small objects, and rare semantics. We report our results in Table 2.

Object properties Interestingly, we find that ET-CLIP excels at instructions, noting specific object characteristics such as colors (e.g., "Turn around, walk to the *red* arm chair"), improving the goal-conditioned success rate by 0.3%. The addition of our CLIP module facilitates the model to leverage specific visual cues stated in the language directives more effectively, due to the vision-language alignment learned from pre-training. This is important for correct object detection in embodied interaction tasks, especially when the environment requires semantically disambiguating objects of the same class.

Small objects Existing state-of-the-art models in ALFRED struggle with detecting small objects [7, 9, 15], as they take up a negligible portion of the input image. The range of success rates in this instruction subset (5.1-5.6) is lower compared to the global average (7.8-7.9), which aligns with previous findings. Surprisingly, ET-CLIP improves the goal-conditioned success rate by 0.5% in instructions that involve manipulating smaller objects, such as "pencil" or "keys". As the pre-trained CLIP model is trained with image-caption pairs, it is likely that the resulting representations are conducive to the semantics of the image, even when objects are small in size.

Rare semantics We additionally validate the hypothesis whether CLIP helps ET better understand instructions with rare words, which we define as words that appear less than 30 times in the training set. Since CLIP is trained with numerous captions, it is likely that ET-CLIP can benefit from this knowledge and in turn interpret rare words better than the baseline. Our results show that ET-CLIP improves ET by 0.8% for rare semantics, which affirms our hypothesis.

5. Conclusion

In this work, we explore the potential of incorporating pre-trained CLIP encoders to the ALFRED task. The novelty of our method lies in leveraging CLIP as an additional module through an auxiliary object detection loss. Our approach can be easily applied to other models that employ object detectors. Our modification upon the Episodic Transformer model shows that using CLIP improves task performance especially in unseen environments, enhancing the model's ability to deal with object properties, small objects, and rare semantics. In future work, we hope to validate the effectiveness of our approach on other models in the field of embodied instruction following, to further improve where current models are failing.

[1] https://github.com/alexpashevich/E.T.

[2] https://github.com/alexpashevich/E.T.#et-with-human-data-only

Acknowledgements

We would like to thank Yonatan Bisk for his guidance throughout this project.

References

[1] Valts Blukis, Chris Paxton, Dieter Fox, Animesh Garg, and Yoav Artzi. A persistent spatial semantic representation for high-level natural language instruction execution. In *Conference on Robot Learning*, pages 706–717. PMLR, 2022. 1

[2] Yufeng Cui, Lichen Zhao, Feng Liang, Yangguang Li, and Jing Shao. Democratizing contrastive language-image pre-training: A clip benchmark of data, model, and supervision. *arXiv preprint arXiv:2203.05796*, 2022. 1

[3] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. Clip on wheels: Zero-shot object navigation as object localization and exploration, 2022. 1

[4] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: Clip embeddings for embodied ai. *arXiv preprint arXiv:2111.09888*, 2021. 1

[5] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 1

[6] Xiwen Liang, Fengda Zhu, Lingling Li, Hang Xu, and Xiaodan Liang. Visual-language navigation pretraining via prompt-based environmental self-exploration. *arXiv preprint arXiv:2203.04006*, 2022. 1

[7] So Yeon Min, Devendra Singh Chaplot, Pradeep Ravikumar, Yonatan Bisk, and Ruslan Salakhutdinov. Film: Following instructions in language with modular methods. *arXiv preprint arXiv:2110.07342*, 2021. 1, 2

[8] Simone Parisi, Aravind Rajeswaran, Senthil Purushwalkam, and Abhinav Gupta. The unsurprising effectiveness of pre-trained vision models for control. *arXiv preprint arXiv:2203.03580*, 2022. 1

[9] Alexander Pashevich, Cordelia Schmid, and Chen Sun. Episodic Transformer for Vision-and-Language Navigation. In *ICCV*, 2021. 1, 2

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. 1

[11] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can CLIP benefit vision-and-language tasks? *CoRR*, abs/2107.06383, 2021. 1

[12] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749, 2020. 1

[13] Allison C Tam, Neil C Rabinowitz, Andrew K Lampinen, Nicholas A Roy, Stephanie CY Chan, DJ Strouse, Jane X Wang, Andrea Banino, and Felix Hill. Semantic exploration from language abstractions and pretrained representations. *arXiv preprint arXiv:2204.05080*, 2022. 1

[14] Jesse Thomason, Mohit Shridhar, Yonatan Bisk, Chris Paxton, and Luke Zettlemoyer. Language grounding with 3d objects. In *Conference on Robot Learning*, pages 1691–1701. PMLR, 2022. 1

[15] Yichi Zhang and Joyce Chai. Hierarchical task learning from language instructions with unified transformers and self-monitoring. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4202–4213, Online, Aug. 2021. Association for Computational Linguistics. 2