CAREL: Instruction-guided Reinforcement Learning with Cross-modal Auxiliary Objectives

Anonymous Author(s)
Affiliation
Address
email

Abstract

Grounding the instruction in the environment is a key step in solving language-guided goal-reaching reinforcement learning problems. In reinforcement learning, the primary aim is to maximize cumulative rewards, which frequently have sparse values in goal-conditioned settings. However, in goal-reaching scenarios, the agent must comprehend the different parts of the instructions within the environmental context in order to complete the overall task successfully. In this work, we propose **CAREL** (***C****ross-modal* ***A****uxiliary* ***RE****inforcement* ***L****earning*) as a new framework to solve this problem using auxiliary loss functions inspired by video-text retrieval literature. The results of our experiments suggest superior sample efficiency and generalization for this framework in different multi-modal reinforcement learning problems.

1 Introduction

Numerous studies have examined the use of language goals or instructions within the context of reinforcement learning (RL) [30, 10, 19]. Language goals typically provide a higher-level and more abstract representation than goals derived from the state space [31]. While state-based goals often specify the agent's final expected goal representation [18, 9], language goals offer more information about the desired sequence of actions and the necessary subtasks [18]. Therefore, it is important to develop approaches that can extract concise information from states or observations and effectively align it with textual information, a process referred to as grounding [30].

Previous research has attempted to ground instructions in observations or states using methods such as reward shaping [12, 24] or goal-conditioned policy/value functions [40, 14, 1, 8], with the latter being a key focus of many studies. Their approaches incorporate various architectural or algorithmic inductive biases, such as cross-attention [13], hierarchical policies [16, 2], and feature-wise modulation [22, 4]. Typically, these works involve feeding instructions and observations into policy or value networks, extracting internal representations of tokens and observations at each time step, and propagating them through the network. Previous studies have explored auxiliary loss functions to improve these internal representations in RL [35, 36, 39]. However, these loss functions lack the alignment property between different input modalities, such as visual/symbolic states and textual commands/descriptions. Recent studies have suggested contrastive loss functions to align text and vision modalities in an unsupervised manner [21, 37, 29, 38, 17]. Most of these studies fall under the video-text retrieval literature [41, 21], where the language tokens and video frames align at different granularities.

Since these methods require a corresponding textual input along with the video, the idea has not yet been employed in language-informed reinforcement learning, where the sequence of observation might not always match the textual modality (due to action failures or inefficacy of trials). One can leverage the success signal or reward to detect the successful episodes and consider them aligned to

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

the textual modality containing instructions or environment descriptions. Doing so, the application of the abovementioned auxiliary loss functions makes sense.

In this study, we propose a new framework, called **CAREL** (***C****ross-modal* ***A****uxiliary* ***RE****inforcement* ***L****earning*), for the adoption of auxiliary grounding objectives from the video-text retrieval literature [41], particularly X-CLIP [21], to enhance the learned representations within these networks and improve cross-modal grounding at different granularities. By leveraging this grounding objective, we aim to enhance the grounding between language instructions and observed states by transferring the multi-grained alignment property of video-text retrieval methods to instruction-following agents. Our experiments on the BabyAI environment [4] showcase the effectiveness of the idea in improving the systematic generalization and sample efficiency of instruction-following agents.

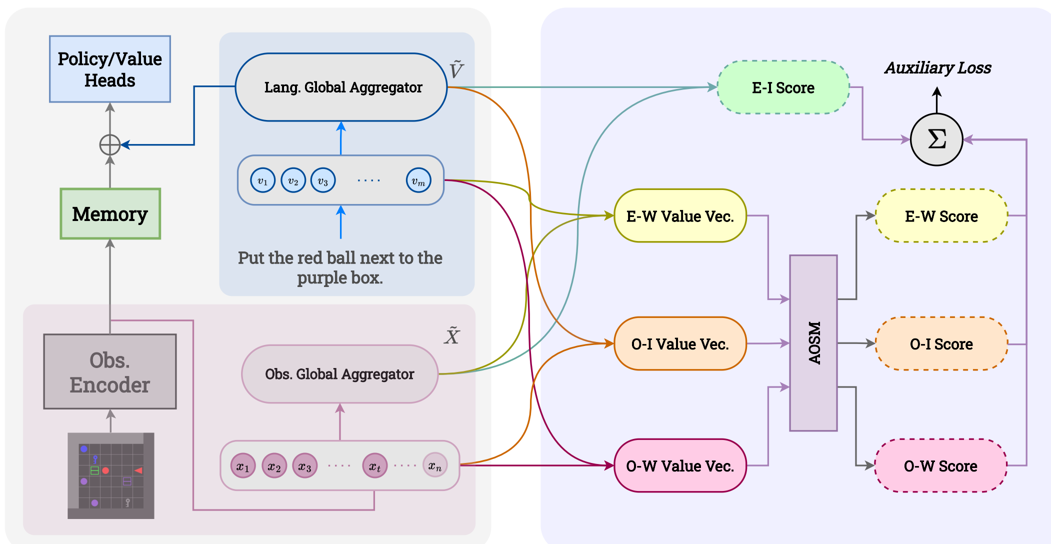

Figure 1: **Overall view of CAREL.** In this figure, we showcase CAREL over a candidate baseline model from [4]. (*Left*) The blue box handles the instruction and its local/global representations, while the pink box contains the components related to observation. (*Right*) The purple box shows the calculation steps for the X-CLIP loss.

2 Related Work

Language-informed RL: There has been a plethora of research on the involvement of natural language [30, 10, 19], either as instructions [29, 22, 24] or descriptions [40] in sequential decision-making [14, 30, 8], especially RL [19]. Besides the fully textual problems where the action/state space is text-based [6], the involvement of language has proven to help agents in visual [16, 12, 23, 34] or symbolic [37, 24, 4, 3] environments, improving their sample efficiency and generalisation [32]. The main approaches to such problems include reward shaping [12, 24], hierarchical RL [16, 2], transfer learning from pre-trained vision-language models [26, 25, 27], or architectural inductive biases in the involvement of language modality as input [40, 14, 1, 8]. One crucial aspect of all of these methods is grounding [30, 40], which enables an embodied agent to understand the language modality in the context of observations [40], reward [12, 24], or dynamics of the environment [32]. This understanding relies on a proper alignment between the language modality and the non-language modalities e.g., visual observations. In this work, we address this problem by means of multi-modal and multi-grained auxiliary unsupervised loss functions borrowed from video-text retrieval literature [21].

Video-text retrieval studies: Across the domain of language-grounding problems, Video-Text Retrieval (VTR), a task involving intricate alignment and abstraction of temporal images (videos), has gained prominence as a fundamental challenge within text-based retrieval. Recent advancements in VTR and Image-Text Retrieval (ITR) research have seen a notable shift towards the adoption of contrastive loss [41], in contrast to the earlier prevalent self/cross-attention mechanisms [11, 28]. Notably, CLIP [29], a Large-scale Vision-Language Pre-training (VLP) model, has successfully leveraged contrastive loss for image-text retrieval, inspiring a wave of video-text retrieval models to follow suit. Among these models, X-CLIP [21] and CLIP4CLIP [20] have emerged as exemplar, yielding remarkable results. Particularly, X-CLIP excels at extracting fine-grained and coarse-grained features from videos, enhancing the alignment between individual frames and the overall video content with textual instructions. However, these ideas have not been employed in RL problems. Inspired by the success of approaches like X-CLIP, we have introduced an auxiliary loss designed to assist RL model encoders in achieving improved representations for both sequences of observations/states and text.

3 CAREL Framework

In this study, we incorporate an auxiliary loss inspired by the X-CLIP model [21] to enhance the grounding between instruction and observations in instruction-following RL agents. This auxiliary loss serves as a supplementary objective, augmenting the primary RL task with a multi-grained alignment property which introduces an additional learning signal to guide the model's learning process. This design choice was motivated by the need to improve the model's ability to extract meaningful information from its observations and align it more effectively with the intended instruction, ultimately enhancing the overall performance of the RL system.

We calculate the proposed loss function over the successful episodes generated by an arbitrary instruction-conditioned RL model within a batch of online trials. To avoid the model being influenced by goal-unrelated behavioral patterns in unsuccessful trajectories, we exclude those trajectories from consideration and leverage reward values to organize only successful ones into a separate batch for the auxiliary loss.

Each successful episode contains a sequence of observations $ep = (O_1, ..., O_n)$ meeting the instructed criteria and an accompanying instruction $instr = (I_1, ..., I_m)$ with m tokens. Since the X-CLIP loss requires local and global encoders for each modality, we must choose such representations from the model or incorporate additional modules to extract them. To explore the exclusive impact of the auxiliary loss and minimize any changes to the architecture, we use the model's existing observation and instruction encoders, which are crucial components of the model itself. We utilize these encoders to extract local representations for each observation O_t denoted as $x_t \in \mathbb{R}^{d\times 1}$, $t = 1, ..., n$ and each instruction token I_i denoted by $v_i \in \mathbb{R}^{d\times 1}$, $i = 1, ..., m$. The global representations can be chosen from the model itself or added to the model by aggregation techniques such as mean-pooling or attention. We denote the global representations for observations and the instruction by $\tilde{X}$ and $\tilde{V}$, respectively. The auxiliary loss function is then calculated according to [21] as below. We restate the formulas in our context to make this paper self-contained.

To utilize contrastive loss, we first need to calculate the similarity score for each episode (ep) - sequence of observations- and instruction ($instr$) pair denoted as $s(ep, instr)$. To do this, we calculate four separate values; Episode-Instruction (S_{E-I}) score, as well as Episode-Word (S_{E-W}), Observation-Instruction (S_{O-I}) and Observation-Word (S_{O-W}) similarity values.

Episode-Instruction score can be calculated using this formula

$$S_{E-I} = (\tilde{X})^T(\tilde{V}) \tag{1}$$

with $\tilde{X}, \tilde{V} \in \mathbb{R}^{d\times 1}$, $S_{V-T} \in \mathbb{R}$.
Other values are calculated in a similar manner:

$$S_{E-W} = (V\tilde{X})^T \tag{2}$$

$$S_{O-I} = X\tilde{V} \tag{3}$$

$$S_{O-W} = XV^T \tag{4}$$

where $X = (x_1, ..., x_n) \in \mathbb{R}^{n\times d}$, $V = (v_1, ..., v_m) \in \mathbb{R}^{m\times d}$, $S_{E-W} \in \mathbb{R}^{1\times m}$, $S_{O-I} \in \mathbb{R}^{n\times 1}$ and $S_{O-W} \in \mathbb{R}^{n\times m}$ are respectively the local representations for the observations and the instruction tokens, and similarity values. These values are then aggregated with appropriate attention weights via a technique called **A**ttention **O**ver **S**imilarity **M**atrix (**AOSM**). Episode-Word (S'_{E-W}) and Observation-Instruction (S'_{O-I}) scores are calculated from the values as follows:

$$S'_{O-I} = \sum_{i=1}^{n} \frac{\exp(S_{O-I}[i,1]/\tau)}{\sum_{j=1}^{n}\exp(S_{O-I}[j,1]/\tau)} S_{O-I}[i,1] \tag{5}$$

$$S'_{E-W} = \sum_{i=1}^{m} \frac{\exp(S_{E-W}[1,i]/\tau)}{\sum_{j=1}^{m} \exp(S_{E-W}[1,j]/\tau)} S_{E-W}[1,i] \tag{6}$$

For the Observation-Word score a bi-level attention is performed, resulting in two fine-grained similarity vectors. These vectors are then converted to scores similar to the previous part:

$$S_{instr} = \sum_{i=1}^{m} \frac{\exp(S_{O-W}[1,i]/\tau)}{\sum_{j=1}^{m} \exp(S_{O-W}[1,j]/\tau)} S_{O-W}[1,i] \tag{7}$$

$$S_{ep} = \sum_{i=1}^{n} \frac{\exp(S_{O-W}[i,1]/\tau)}{\sum_{j=1}^{m} \exp(S_{O-W}[j,1]/\tau)} S_{O-W}[i,1] \tag{8}$$

where $S_{instr} \in \mathbb{R}^{n\times 1}$ show the similarity score between the instruction and n observations in the episode and $S_{ep} \in \mathbb{R}^{1\times m}$ represents the similarity between the episode and m words in the instruction.

The second attention operation is performed on these vectors to calculate the Observation-Word similarity score (S'_{F-W}), which represents the similarity between all observations and words:

$$S'_{instr} = \sum_{i=1}^{n} \frac{\exp(S_{instr}[i,1]/\tau)}{\sum_{j=1}^{n} \exp(S_{instr}[j,1]/\tau)} S_{instr}[i,1] \tag{9}$$

$$S'_{ep} = \sum_{i=1}^{m} \frac{\exp(S_{ep}[1,i]/\tau)}{\sum_{j=1}^{m} \exp(S_{ep}[1,j]/\tau)} S_{ep}[1,i] \tag{10}$$

Where $S'_{instr}, S'_{ep} \in \mathbb{R}^1$ are instance-level scores. We average the two scores to find the Observation-Word score:

$$S'_{O-W} = \frac{S'_{ep} + S'_{instr}}{2} \tag{11}$$

The final similarity score between an episode and an instruction is computed using the previously calculated scores:

$$s(ep, instr) = \frac{S_{E-I} + S'_{E-W} + S'_{O-I} + S'_{O-W}}{4} \tag{12}$$

This method takes into consideration both fine-grained and coarse-grained contrasts. Considering N episode-instruction pairs in a batch of successful trials, the auxiliary loss is calculated as below:

$$\mathcal{L}_{aux} = -\frac{1}{n} \sum_{i=1}^{N} (\log \frac{exp(s(ep_i, instr_i)}{\sum_{j=1}^{N} exp(s(ep_i, instr_j))} + \log \frac{exp(s(ep_i, instr_i)}{\sum_{j=1}^{N} exp(s(ep_j, instr_i))}) \tag{13}$$

The total objective is calculated by adding this loss to the primary RL loss, $\mathcal{L}_{RL}$, with a coefficient of λ_C.

$$\mathcal{L}_{total} = \mathcal{L}_{RL} + \lambda_C . \mathcal{L}_{aux} \tag{14}$$

The overall architecture of a base model [4] and the calculation of the auxiliary loss is depicted in Figure 1. If the shape of the output representations from the observation and instruction encoders does not align, we employ linear transformation layers to bring them into the same feature space. This transformation is crucial as it facilitates the calculation of similarity between these representations within our loss function.

4 Experiments

In our experiments, we conducted a comparative analysis to assess the impact of X-CLIP [21] auxiliary loss on generalization and sample efficiency of instruction-following agents. We try to answer the following questions:

- Does the proposed CAREL approach actually help instruction-following agents (Section 4.1)?
- Is it possible to apply CAREL to other multi-modal settings in the context of RL agents (Section 4.2)?

Two series of experiments are performed to answer the abovementioned questions. In the following parts, we explain the experimental settings for each set of experiments and state the results to showcase the efficacy of CAREL.[1]

4.1 Instruction-following with CAREL

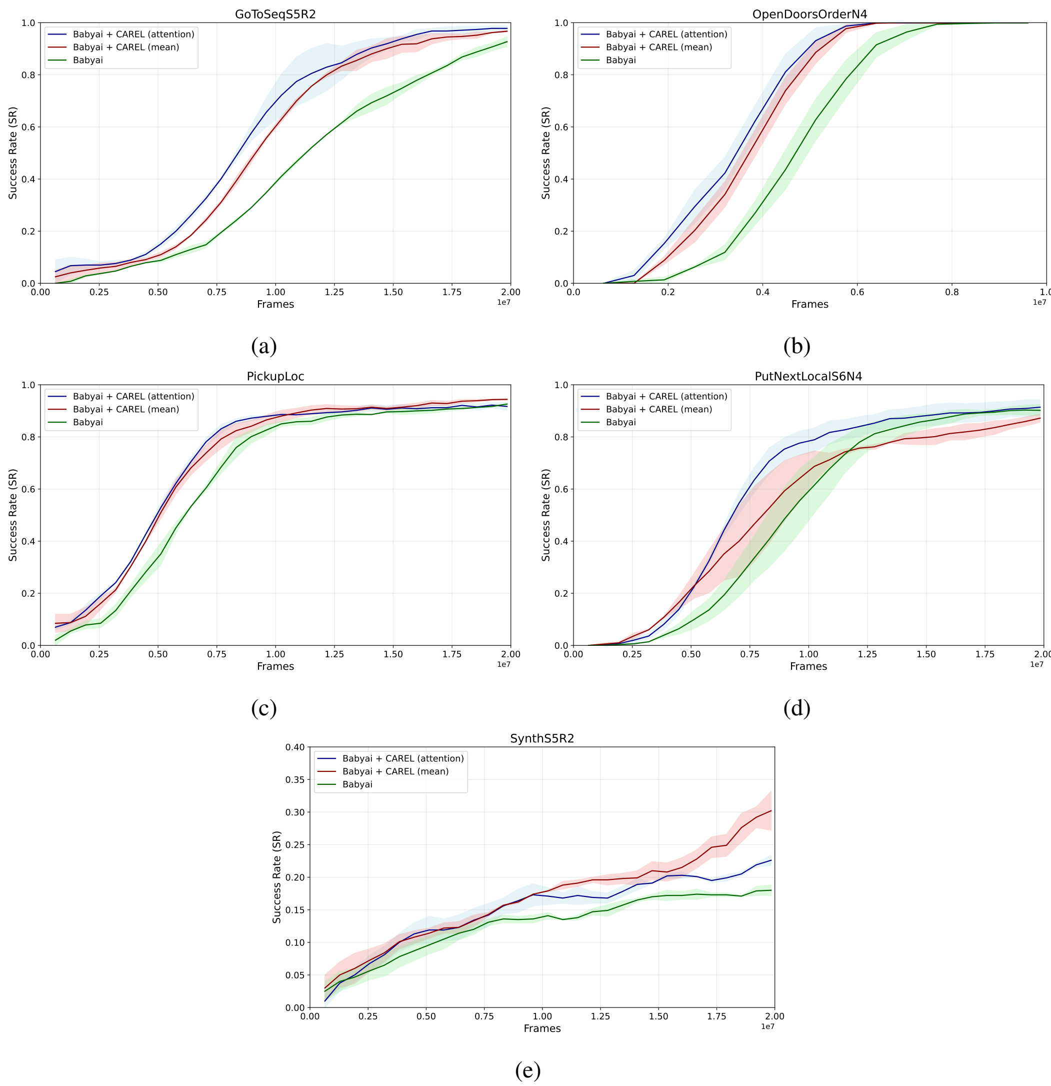

Figure 2: Test time comparison between success rates of the proposed method (CAREL) and the baseline model.

We employ the BabyAI environment [4], a lightweight but logically complex benchmark with procedurally generated difficulty levels, which enables in-depth exploration of grounded language learning in the goal-conditioned RL context. This environment provides a 2D grid-world environment with multiple objects, such as keys, balls, boxes, and doors, which can be distractors at specific

[1]For the experiments reported in this paper, we have used one NVIDIA 3090 GPU and one TITAN RTX GPU over two weeks.

difficulty levels and take one of the six possible colors in the BabyAI environment. The agent is tasked with a synthetic and natural-looking instruction and receives a sparse reward at the end of the episode if all steps of the instruction are accomplished successfully.

We use BabyAI's baseline model as the base model and minimally modify its current structure. Word-level representations are calculated using a simple token embedding layer. Then, a GRU encoder calculates the global instruction representation. Similarly, we use the model's default observation encoder, a convolutional neural network with three two-dimensional convolution layers. All observations pass through this encoder to calculate local representations. Mean-pooling/Attention over these local representations is applied as the aggregation method to calculate the global observation representation. The RL agent is trained using the PPO algorithm [33] and Adam optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is $7e - 4$, and the batch size is 256. We set $\lambda_C = 0.01$ and the temperature $\tau = 1$ as CAREL-specific hyperparameters. To minimize the changes to the baseline model updates, we backpropagate the gradients in an outer loop of PPO loss to be able to capture episode-level similarities. This gradient update with different frequencies has been tried in the literature before [22].

The evaluation framework for this work is based on systematic generalization to assess the language grounding property of the model. We report agent's success rate and mean return over a set of unseen tasks at each BabyAI level, according to Table 1. These metrics are recorded during validation checkpoints throughout training. We recorded and analyzed the success rate achieved by these models across various levels. Furthermore, Figure 2 illustrates the improved sample efficiency brought about by CAREL. All results are reported over two random seeds.

The results indicate improved sample efficiency of CAREL methods across all levels, especially those with step-by-step solutions that require the alignment between the instruction parts and episode interactions more explicitly, namely `GoToSeq` and `OpenDoorsOrder` which contain a sequence of `Open`/`GoTo` subtasks described in the instruction. The generalization is significantly improved in more complex tasks, e.g., `Synth`.

Table 1: Test splits for BabyAI levels (For more details on the environment, please see [4]).

Level	**Test split**

`GoToSeq`
`PickupLoc`
`PutNextLocalS6N4`
	Instructions containing *"red box"*, *"green ball"*, *"purple key"*, *"yellow box"*, *"blue ball"*, and *"grey key"*.

`SynthS5R2`	*"put the red ball next to the green key"*, *"put the purple box next to the yellow ball"*, *"put the blue key next to the grey box"*, *"go to the red box"*, *"go to the green ball"*, *"pick up the purple key"*, *"pick up the yellow box"*, *"open the blue door"*, *"open the grey door"*,
`OpenDoorsOrderN4`	*"open the blue door, then open the yellow door"*, *"open the green door, then open the grey door"*, *"open the grey door, then open the red door"*, *"open the yellow door, then open the purple door"*, *"open the red door, then open the green door"*, *"open the purple door, then open the blue door"*,

4.2 Multi-modal RL with CAREL

To assess the performance of CAREL in more general multi-modal scenarios of RL, we incorporate the proposed framework in a recently proposed model called SHELM [26], which leverages the knowledge hidden in pre-trained models such as CLIP [29] and Transformer-XL [7]. SHELM uses CLIP to extract textual tokens related to every observation, and then these tokens are passed through the frozen Transformer-XL network to form a memory of tokens throughout the episode. This hidden memory is then concatenated to a local representation of the observation through a CNN network and then passed to actor/critic heads.

For this model, we consider the selected token's representation and the CNN's output as local representations. The global representations for text come from the hidden state of Transformer-XL, and an additional attention aggregator is applied on top of the CNN encoder of observations to obtain the global representations. In order to allow the auxiliary loss to refine local and global representations to the current task with more degrees of freedom, we apply a network similar to adapters [15] consisting of linear layers with ReLU non-linearity in between and a final residual connection. One adapter comes over the Transformer-XL representations and another comes after CLIP for observations. Doing so, we hope the auxiliary X-CLIP loss function will improve the learnable representations to be more suitable for multi-grained alignment. Figure 3 shows the effectiveness of CAREL in the Miniworld environment [5]. We also use a logarithmic scheduler in this experiment to decline λ_C from 0.1 to 0.01. The gradient backpropagation is separated from RL loss similar to section 4.1. These results are reported over two random seeds as well.

Although the model has to train more parameters due to additional adapters, we can observe the improved sample efficiency, which can hint at the improved internal representations by means of the CAREL framework. This can affect the choice of related tokens in CLIP and the hidden representation of Transformer-XL, which corresponds to the memory of tokens and global representation for the textual modality.

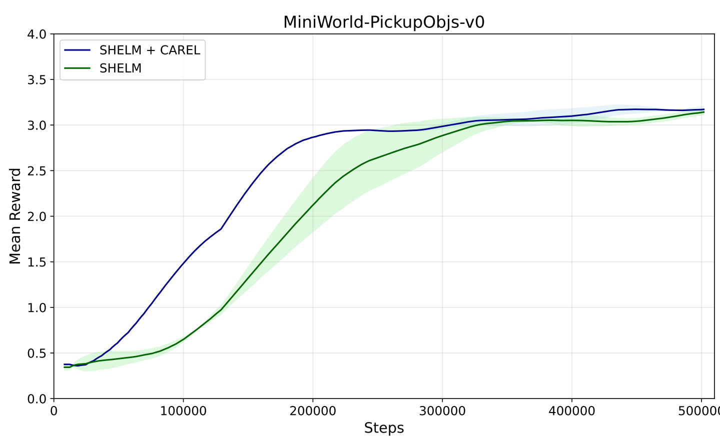

Figure 3: Training time comparison between mean total rewards of the proposed method (CAREL) and the baseline model, SHELM.

5 Conclusion

This paper proposes the CAREL framework to adopt auxiliary cross-modal contrastive loss functions to the multi-modal RL setting, especially instruction-following agents. The aim is to improve the multi-grained alignment between different modalities, leading to superior grounding in the context of learning agents. We apply this method over existing instruction-following agents and multi-modal actor/critic networks. The results indicate the sample efficiency and generalization boost from the proposed framework.

As for the future directions of this study, we suggest further experiments on more complex environments and other multi-modal sequential decision-making agents. Also, there could be various versions of the auxiliary loss, e.g., at multiple levels of granularity with additional modalities such as descriptive text or higher-level information from the image modality. The involvement of the auxiliary signal in the reward function could also be an interesting future direction.

References

[1] Ahmed Akakzia et al. "Grounding language to autonomously-acquired skills via goal generation". In: *arXiv preprint arXiv:2006.07185* (2020).

[2] Jacob Andreas, Dan Klein, and Sergey Levine. "Modular multitask reinforcement learning with policy sketches". In: *International conference on machine learning*. PMLR. 2017, pp. 166–175.

[3] Tianshi Cao et al. "Babyai++: Towards grounded-language learning beyond memorization". In: *arXiv preprint arXiv:2004.07200* (2020).

[4] Maxime Chevalier-Boisvert et al. "Babyai: A platform to study the sample efficiency of grounded language learning". In: *arXiv preprint arXiv:1810.08272* (2018).

[5] Maxime Chevalier-Boisvert et al. "Minigrid & Miniworld: Modular & Customizable Reinforcement Learning Environments for Goal-Oriented Tasks". In: *CoRR* abs/2306.13831 (2023).

[6] Marc-Alexandre Côté et al. "Textworld: A learning environment for text-based games". In: *Computer Games: 7th Workshop, CGW 2018, Held in Conjunction with the 27th International Conference on Artificial Intelligence, IJCAI 2018, Stockholm, Sweden, July 13, 2018, Revised Selected Papers 7*. Springer. 2019, pp. 41–75.

[7] Zihang Dai et al. "Transformer-xl: Attentive language models beyond a fixed-length context". In: *arXiv preprint arXiv:1901.02860* (2019).

[8] Zhiwei Deng, Karthik Narasimhan, and Olga Russakovsky. "Evolving graphical planner: Contextual global planning for vision-and-language navigation". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 20660–20672.

[9] Benjamin Eysenbach et al. "Contrastive learning as goal-conditioned reinforcement learning". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 35603–35620.

[10] Hector Geffner. "Target languages (vs. inductive biases) for learning to act and plan". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 11. 2022, pp. 12326–12333.

[11] Satya Krishna Gorti et al. "X-pool: Cross-modal language-video attention for text-video retrieval". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 5006–5015.

[12] Prasoon Goyal, Scott Niekum, and Raymond J Mooney. "Using natural language for reward shaping in reinforcement learning". In: *arXiv preprint arXiv:1903.02020* (2019).

[13] Austin W Hanjie, Victor Y Zhong, and Karthik Narasimhan. "Grounding language to entities and dynamics for generalization in reinforcement learning". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 4051–4062.

[14] Donald Joseph Hejna III, Pieter Abbeel, and Lerrel Pinto. "Improving Long-Horizon Imitation Through Language Prediction". In: (2021).

[15] Neil Houlsby et al. "Parameter-efficient transfer learning for NLP". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 2790–2799.

[16] Yiding Jiang et al. "Language as an abstraction for hierarchical deep reinforcement learning". In: *Advances in Neural Information Processing Systems* 32 (2019).

[17] Juncheng Li et al. "Fine-grained semantically aligned vision-language pre-training". In: *Advances in neural information processing systems* 35 (2022), pp. 7290–7303.

[18] Minghuan Liu, Menghui Zhu, and Weinan Zhang. "Goal-conditioned reinforcement learning: Problems and solutions". In: *arXiv preprint arXiv:2201.08299* (2022).

[19] Jelena Luketina et al. "A survey of reinforcement learning informed by natural language". In: *arXiv preprint arXiv:1906.03926* (2019).

[20] Huaishao Luo et al. "Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning". In: *Neurocomputing* 508 (2022), pp. 293–304.

[21] Yiwei Ma et al. "X-clip: End-to-end multi-grained contrastive learning for video-text retrieval". In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 638–647.

[22] Kanika Madan et al. "Fast and slow learning of recurrent independent mechanisms". In: *arXiv preprint arXiv:2105.08710* (2021).

[23] So Yeon Min et al. "Film: Following instructions in language with modular methods". In: *arXiv preprint arXiv:2110.07342* (2021).

[24] Suvir Mirchandani, Siddharth Karamcheti, and Dorsa Sadigh. "Ella: Exploration through learned language abstraction". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 29529–29540.

[25] Fabian Paischer et al. "History compression via language models in reinforcement learning". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 17156–17185.

[26] Fabian Paischer et al. "Semantic HELM: An Interpretable Memory for Reinforcement Learning". In: *arXiv preprint arXiv:2306.09312* (2023).

[27] Fabian Paischer et al. "Toward Semantic History Compression for Reinforcement Learning". In: *Second Workshop on Language and Reinforcement Learning*. 2022.

[28] Zhengxin Pan, Fangyu Wu, and Bailing Zhang. "Fine-Grained Image-Text Matching by Cross-Modal Hard Aligning Network". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 19275–19284.

[29] Alec Radford et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[30] Frank Röder et al. "The embodied crossmodal self forms language and interaction: a computational cognitive review". In: *Frontiers in psychology* 12 (2021), p. 716671.

[31] Matthias Rolf and Minoru Asada. "Where do goals come from? A generic approach to autonomous goal-system development". In: *arXiv preprint arXiv:1410.5557* (2014).

[32] Laura Ruis et al. "A benchmark for systematic generalization in grounded language understanding". In: *Advances in neural information processing systems* 33 (2020), pp. 19861–19872.

[33] John Schulman et al. "Proximal policy optimization algorithms". In: *arXiv preprint arXiv:1707.06347* (2017).

[34] Mohit Shridhar et al. "Alfred: A benchmark for interpreting grounded instructions for everyday tasks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10740–10749.

[35] Adam Stooke et al. "Decoupling representation learning from reinforcement learning". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 9870–9879.

[36] Haoyu Wang et al. "Constrained Contrastive Reinforcement Learning". In: *Asian Conference on Machine Learning*. PMLR. 2023, pp. 1070–1084.

[37] Lewei Yao et al. "Filip: Fine-grained interactive language-image pre-training". In: *arXiv preprint arXiv:2111.07783* (2021).

[38] Jiahui Yu et al. "Coca: Contrastive captioners are image-text foundation models". In: *arXiv preprint arXiv:2205.01917* (2022).

[39] Ruijie Zheng et al. "TACO: Temporal Latent Action-Driven Contrastive Loss for Visual Reinforcement Learning". In: *arXiv preprint arXiv:2306.13229* (2023).

[40] Victor Zhong, Tim Rocktäschel, and Edward Grefenstette. "Rtfm: Generalising to novel environment dynamics via reading". In: *arXiv preprint arXiv:1910.08210* (2019).

[41] Cunjuan Zhu et al. "Deep learning for video-text retrieval: a review". In: *International Journal of Multimedia Information Retrieval* 12.1 (2023), p. 3.