Integrating Text-Visual and Task Attention for Language-Guided Robot Learning

Giuseppe Rauso[1], Riccardo Caccavale[1], Vincenzo Lippiello[1] and Alberto Finzi[1]

[1]*Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione, Università degli Studi di Napoli "Federico II"*

Abstract

In this work, we investigate the interaction between text, visual, and task attention models during the learning and execution of structured tasks expressed in natural language. In this direction, we propose an architecture that leverages and combines different attention models at multiple levels. Firstly, a multi-modal attention mechanism is introduced, enabling the agent to map objects in the environment to the words in the given mission, expressed in natural language, to effectively perform the required task. Secondly, an additional attention mechanism is introduced to direct the agent's textual attention to the parts of the sentence relevant to the subtasks yet to be completed. The agent is trained in MiniGrid environments using the Proximal Policy Optimization algorithm, and its performance is evaluated by comparing the proposed architecture with a baseline that excludes attention mechanisms. In addition, an ablation study is conducted on the attention module for task attention.

Keywords

Language Conditioned Reinforcement Learning, Multi-modal Attention, Behavior Transparency, Robot Task Learning

1. Introduction

This work introduces a novel approach to improving robot task learning and execution by integrating text-visual and task-attention models. Attention mechanisms, extensively studied in cognitive neuroscience and widely adopted in artificial intelligence, have proven effective in improving performance and training efficiency, particularly in machine learning. For example, transformers in natural language processing (NLP) have revolutionized the field by enabling models to contextually weigh word relevance, capturing long-range dependencies in text. In this paper, we explore the interaction between text-visual and task attention models within a reinforcement learning framework, where robots are tasked with completing missions specified by natural language instructions. We focus on a language-conditioned reinforcement learning setting, where joint observation and textual representations are used to enhance policy generalization and transferability to novel environments. Here, mission goals are defined textually, and the robotic agent is trained first to develop an attention model that maps task-related words to the corresponding visual features, and second, in the case of composite tasks, to mask those same words when the task they describe is completed, focusing attention on the ones that are still relevant. This model supports the agent's ability to focus on objects relevant to the mission, improving task learning and execution.

Over the years, various attention models, inspired by neuroscience, have been proposed in machine learning, with applications in image and video classification [1, 2], translation [3, 4], and question answering [5, 6]. In reinforcement learning, the use of attention mechanisms to highlight visual features relevant to the task is proposed in works such as [7], which introduces the *Deep Attention Recurrent Q-Network* (DARQN), or [8], where a soft-attention mechanism is used in combination with a *Deep Q-Network* [9]. In other works, different attention mechanisms, such as multi-attention in [10] or self-attention in [11, 12], are used to improve navigation or to learn relationships between entities in a reinforcement learning context. Some studies have explored attention mechanisms that combine different input sources, as seen in [13], where query vectors are generated from the output

11th Italian Workshop on Artificial Intelligence and Robotics (AIRO 2024)

giuseppe.rauso@unina.it (G. Rauso); riccardo.caccavale@unina.it (R. Caccavale); vincenzo.lippiello@unina.it (V. Lippiello); alberto.finzi@unina.it (A. Finzi)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

of an LSTM layer, while key and value vectors are derived from the encoding of visual observations. In language-conditioned reinforcement learning, several works have investigated the use of natural language to define goals, such as in [14, 15, 16], also using gated attention mechanisms, such as in [17], and combining images and text in attention calculations [18]. Our approach aligns with multi-modal attention as in [18], but with a different aim: we focus on mapping task-related words to observed environmental features, creating attention maps that enhance both performance and interpretability, and, at a higher level, considering only the relevant words based on the tasks completed. The framework introduced here extends that of [19], where text and visual attention models are combined. In this work, we further develop the approach to show how task attention and structured tasks can also be incorporated using a similar method.

Our proposed framework, which integrates combined attention mechanisms, is trained using the *Proximal Policy Optimization* (PPO) [20] algorithm in *BabyAI* [21] environments, a platform based on *MiniGrid* [22] that enables the creation of grid-based environments with objects, obstacles, and rooms, where tasks can be defined in natural language using a synthetic language called *Baby language*. We detail the architecture and learning process, highlighting the interaction between textual and visual attention models, as well as the process of suppressing words related to completed tasks. The approach is evaluated against a baseline lacking attention mechanisms. Experimental results confirm the efficacy of our approach, demonstrating its advantages in both performance and behavior transparency.

2. Multi-modal Attention and Language Grounding

We address a robot task learning problem where goals are given in natural language. Our approach uses multi-modal attention mechanisms to align observed features with the words describing the task. This method has two objectives: to improve both task learning and execution while grounding each word to relevant visual features (such as object properties like color) through per-word attention maps. Additionally, it enhances transparency by aligning the attention on text and features with the task at hand. The architecture is end-to-end, with both task execution and attention map learning driven solely by environmental rewards. The environments are based on MiniGrid and are goal-augmented *Partially Observable Markov Decision Processes* (POMDPs).

2.1. System Architecture

Our proposed system takes the agent's observed features O and the natural language mission g, as input, and generates the corresponding policy by utilizing task-relevant attention maps. We define $\tilde{O} = Encoding(O)$ as the encoding of the observation, and $\tilde{g} = Embedding(g)$ as the embedding of the task tokens. Building on the scaled dot-product attention mechanism with query, key, and value from [4], we compute the attention matrix as:

$$A = \operatorname{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{1}$$

where Q and K are, respectively, the projection of the task embedding and the observation encoding onto a space of dimension d_k. Each row of the matrix A highlights the cells related to the corresponding word in the portion of the grid observed by the agent (see Figure 1). To adjust the signal in relevant positions of the observed feature map based on mission words, we propose an alternative to directly multiplying A by a value matrix (as in [4]). Our goal is to derive attention weights for individual words based on the agent's observations, identifying which words are most salient for the grid portion being observed. To achieve this, we compute the *Shannon entropy*, H, on the rows of matrix A and, to obtain the attention weights for individual words, we apply the softmax to the negated entropy vector:

$$w = \operatorname{softmax}\left(\begin{bmatrix} -H(A_1) \\ \vdots \\ -H(A_m) \end{bmatrix}\right) \tag{2}$$

Finally, the attention map M is obtained by calculating the weighted sum of the m rows of matrix A using the weights w. This map is then applied to each feature map of $\tilde{O}$ to produce F, highlighting only the cells corresponding to elements mentioned in the mission text. The filtered feature maps are then passed through an LSTM layer, allowing the agent to function in a partially observable environment, and the output is concatenated with the output of a GRU recurrent layer that processes the mission text.

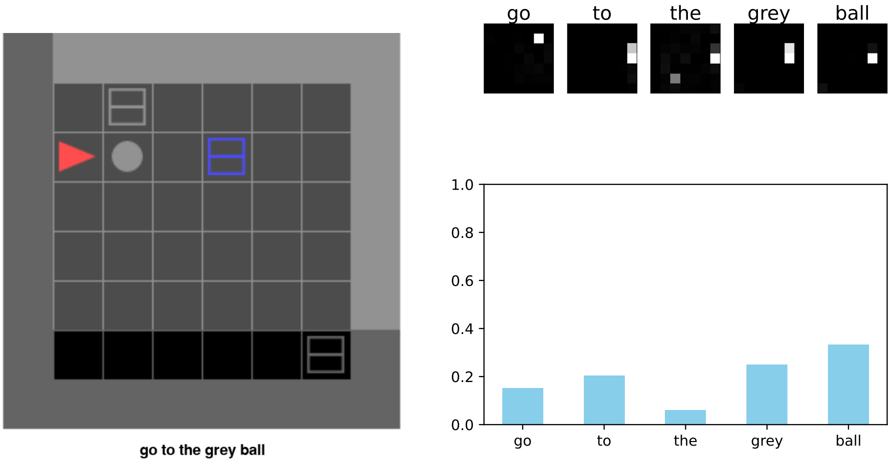

Figure 1: Snapshot of the grid scenario (left) with the associated per-word attention maps (Top right) and the word weights w (Bottom right). The per-word maps are agent-centric with the agent positioned in the middle of the right side, facing left, with the positions of objects relative to the agent mirrored along the vertical axis. In this case, the agent must complete the task described by the phrase ”go to the grey ball”.

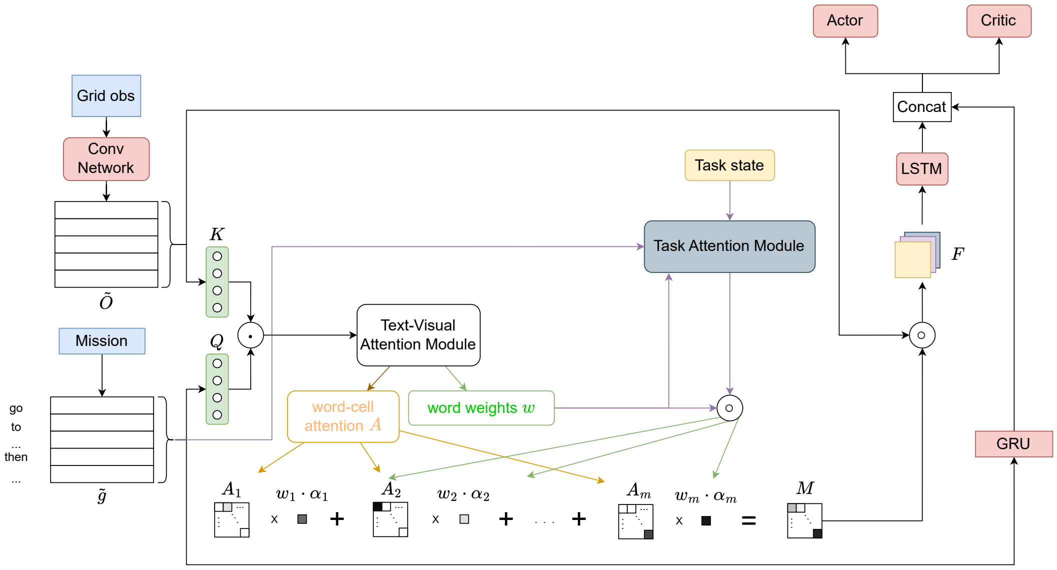

Figure 2: System architecture. The Text-Visual Module is first trained with simple tasks to ground words in visual features. Then, the Task Attention Module is trained to manage attention shifts during the execution of structured tasks. In the single-task case, where the Task Attention Module is not used, we assume the word mask α to be a vector of all ones.

2.2. Training

Agents were trained in a customized 8×8 room environment without internal walls, containing four randomly selected and colored objects. For each task, the goal was to reach or pick up one object, with the other three serving as distractors. Observations were encoded as a 7×7 grid with three channels, representing the agent's field of view, where each grid cell was a tuple *(object id, object color, object state)*. The reward is set to the default for MiniGrid environments: it's a value between 0 and 1 based on the number of steps taken to complete the task, and it's given at the end of the episode. The ”done” action was used to signal task completion, such as reaching or picking up the target object. We compare the effectiveness of the multi-modal attention system against a baseline that lacks attention mechanisms. In the baseline, the attention map generation is removed, and the convolutional network encoding is directly passed to the LSTM network, with its output concatenated with the text encoding from the GRU network (see Figure 2). In the experiments conducted, the model without attention mechanisms converges to a lower reward value and shows significant instability compared to the attention-based model.

2.3. Evaluation

The proposed framework is compared to a no-attention setup across environments of varying sizes and object counts to test robustness in more challenging settings. These larger environments, with additional distractors and a limited field of view, pose greater challenges. The experimental results highlight the robustness of the multi-modal attention agent, which experiences a much smaller performance drop as distractors increase. During testing, the models are evaluated with a number of objects ranging from 4 to 12 and in rooms with dimensions of 8×8, 10×10, and 12×12. As the number of objects increases and across rooms of different sizes, the model with the proposed attentional mechanisms consistently maintains an average reward between 0.8 and 0.9 and a success rate between 90% and 100%. In contrast, the baseline model without attentional mechanisms experiences a drastic performance drop: while it achieves an average reward between 0.8 and 0.85 and a success rate between 90% and 96% with 4 objects across different room sizes, its performance significantly degrades as the number of distractors increases, reaching an average reward between 0.5 and 0.6 and a success rate between 55% and 70% with 12 objects.

3. Structured Tasks and Task Attention

To extend the agent's ability to learn and execute structured tasks expressed in more elaborate mission sentences (e.g., ”A and B”, ”C before D”, and ”E after F”). Preliminary experiments were conducted on a potential extension of the proposed single-task architecture to achieve a higher-level form of attention, this time focused on task execution monitoring. This additional component of the architecture, which extends the framework introduced in [19], is highlighted in the dotted purple box in Figure 2. To train the agent on multiple sub-tasks, we begin with weights from single-task training and use a reduced learning rate. In addition, we introduce a *Task Attention Module*. This module masks sub-sentences corresponding to sub-tasks already accomplished (e.g., in ”before and after” tasks). At time t, it generates a mask to apply to the word weight vector w, directing the agent's focus to the tasks relevant at that step. During training, we employ two learning rates: a lower rate for the pre-trained network and a higher one for the Task Attention Module. In this work, we combine ”go to” and ”pick up” tasks using connectors like ”and”, ”before”, and ”after”, a feature already available in MiniGrid environments. However, we customized the environment to limit combinations to two tasks of these types with the specified connectors. The custom environment for this second phase provides, at each step t, a binary vector c_t indicating task completion status, with 1 in position i if task i is completed, 0 otherwise.

3.1. Task Attention Module

The task attention module takes as input the weighted sum of the mission word embeddings and the vector c_t. These inputs enable the module to generate a mask for mission words related to pending tasks. The module outputs a vector of m values between 0 and 1, which are multiplied by the weight vector w. In this way, words related to a completed task will be ”inhibited”, bringing the weight to 0 or close to 0, thus making the corresponding attention map irrelevant. In the experiments, the Task Attention Module is implemented as a fully connected neural network with ReLU activations and a sigmoid activation in the last layer with m neurons.

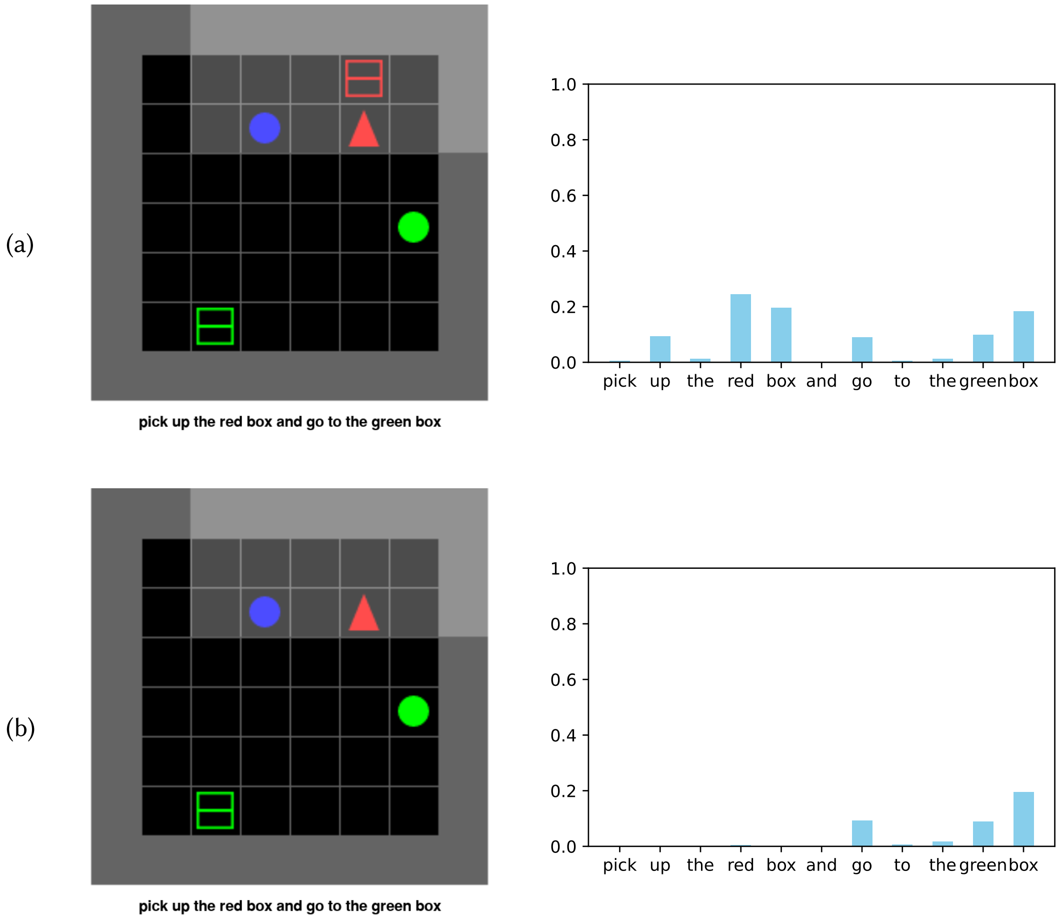

Figure 3: Snapshot of the environment and the word weights w for an ”and” task before picking up the red box (in row (a)) and after picking up the red box (in row (b)). As can be observed, the weights for the words related to the first task, that is, picking up the red box, are set to zero after its completion.

3.2. Training

This module is integrated into the architecture, as shown in Figure 2, and is trained with PPO using the reward from the environment. In this setting, training was carried out in 10x10 environments. To enhance training stability and quality, we calculate a task attention loss function, L_{TA}, with the output from the Task Attention Module and the ground truth masks, obtained during the buffer filling phase. It is added to the general PPO loss, with a negative sign, so that it is minimized during training.

3.3. Evaluation

Preliminary experiments in a 10×10 environment demonstrated the benefits of the proposed architecture with the Task Attention Module, both in enhancing performance and reducing training time. Across structured tasks involving conjunctions and temporal relations (”and”, ”before”, ”after”), the model consistently achieved success rates above 90%, compared to about 70% for versions without the module—whether using only text-visual attention or no attention mechanisms. Moreover, attention-based models stabilized performance in significantly fewer training steps, with the Task Attention Module reaching optimal results within 2 million steps, while other models required longer training and still underperformed. These findings confirm the effectiveness of the Task Attention Module, with further analysis and detailed evaluations planned for future work.

4. Conclusions

We introduced a novel task learning approach where agents, guided by natural language instructions, exploit multi-modal attention mechanisms to align the relevance of mission words with observed features while focusing the agent's attention on task-relevant features during the execution. In the proposed method, the agent is trained in two steps. Firstly, the system is trained with simple tasks to generate per-word attention maps, grounding mission words, and their relevance in environmental observations. In a second phase, we address structured tasks by training a task attention mechanism to suppress words related to already accomplished subtasks, disregarding their textual and visual relevance. We tested the approach in MiniGrid environments to assess its feasibility and performance in simple use cases. The experimental evaluation showed promising results for both single-task and structured task scenarios. Further experiments are already underway with alternative architectures to improve stability and generalize the approach to more complex scenarios. In future work, we aim to investigate the integration of more refined executive attention mechanisms [23, 24, 25], while assessing the scalability of the approach in incrementally structured tasks.

Acknowledgments

This work was partially supported by the projects: EU Horizon INVERSE (grant 101136067) and euROBIN (grant 101070596), Melody (CUP E53D23017550001), SPACE IT UP (PE15 ASI/MUR, CUP I53D24000060005).

References

[1] V. Mnih, N. Heess, A. Graves, k. kavukcuoglu, Recurrent models of visual attention, in: NIPS, 2014, pp. 2204 – 2212.

[2] M. Shan, N. Atanasov, A spatiotemporal model with visual attention for video classification, 2017. arXiv:1707.02069.

[3] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, 2016. arXiv:1409.0473.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017.

[5] J. Andreas, M. Rohrbach, T. Darrell, D. Klein, Learning to compose neural networks for question answering, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1545–1554.

[6] J. Andreas, M. Rohrbach, T. Darrell, D. Klein, Neural module networks, in: CVPR, 2016, pp. 39–48.

[7] I. Sorokin, A. Seleznev, M. Pavlov, A. Fedorov, A. Ignateva, Deep attention recurrent q-network, 2015. arXiv:1512.01693.

[8] S. Mousavi, M. Schukat, E. Howley, A. Borji, N. Mozayani, Learning to predict where to look in interactive environments using deep recurrent q-learning, 2017. arXiv:1612.05753.

[9] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. A. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, D. Hassabis, Human-level control through deep reinforcement learning, Nature 518 (2015) 529–533.

[10] J. Choi, B.-J. Lee, B.-T. Zhang, Multi-focus attention network for efficient deep reinforcement learning, 2017. arXiv:1712.04603.

[11] A. Manchin, E. Abbasnejad, A. van den Hengel, Reinforcement learning with attention that works: A self-supervised approach, 2019. arXiv:1904.03367.

[12] V. F. Zambaldi, D. Raposo, A. Santoro, V. Bapst, Y. Li, I. Babuschkin, K. Tuyls, D. P. Reichert, T. P. Lillicrap, E. Lockhart, M. Shanahan, V. Langston, R. Pascanu, M. M. Botvinick, O. Vinyals, P. W. Battaglia, Deep reinforcement learning with relational inductive biases, in: ICLR, 2019, pp. 6826 – 6843.

[13] A. Mott, D. Zoran, M. Chrzanowski, D. Wierstra, D. Jimenez Rezende, Towards interpretable reinforcement learning using attention augmented agents, in: NIPS, 2019, pp. 12318 – 12327.

[14] A. Akakzia, C. Colas, P.-Y. Oudeyer, M. CHETOUANI, O. Sigaud, Grounding language to autonomously-acquired skills via goal generation, in: ICLR, 2021.

[15] F. Röder, M. Eppe, Language-conditioned reinforcement learning to solve misunderstandings with action corrections, in: NIPS Workshop LaReL, 2022.

[16] F. Röder, M. Eppe, S. Wermter, Grounding hindsight instructions in multi-goal reinforcement learning for robotics, in: ICDL, 2022, pp. 170–177.

[17] C. Colas, T. Karch, N. Lair, J.-M. Dussoux, C. Moulin-Frier, P. Dominey, P.-Y. Oudeyer, Language as a cognitive tool to imagine goals in curiosity driven exploration, in: NIPS, 2020, pp. 3761–3774.

[18] S. Peng, X. Hu, R. Zhang, J. Guo, Q. Yi, R. Chen, Z. Du, L. Li, Q. Guo, Y. Chen, Conceptual reinforcement learning for language-conditioned tasks, in: AAAI 2023, 2023, pp. 9426–9434.

[19] G. Rauso, R. Caccavale, A. Finzi, Combined text-visual attention models for robot task learning and execution, in: AIxIA 2024 – Advances in Artificial Intelligence, Springer Nature Switzerland, 2024, pp. 228–240.

[20] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, 2017. arXiv:1707.06347.

[21] M. Chevalier-Boisvert, D. Bahdanau, S. Lahlou, L. Willems, C. Saharia, T. H. Nguyen, Y. Bengio, Babyai: A platform to study the sample efficiency of grounded language learning, in: ICLR, 2019, pp. 4429–4447.

[22] M. Chevalier-Boisvert, B. Dai, M. Towers, R. D. L. Perez-Vicente, L. Willems, S. Lahlou, S. Pal, P. S. Castro, J. K. Terry, Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks, in: NIPS Datasets and Benchmarks Track, 2023, pp. 73383–73394.

[23] R. Caccavale, A. Finzi, Learning attentional regulations for structured tasks execution in robotic cognitive control, Autonomous Robots 43 (2019) 2229 – 2243.

[24] R. Caccavale, M. Saveriano, G. A. Fontanelli, F. Ficuciello, D. Lee, A. Finzi, Imitation learning and attentional supervision of dual-arm structured tasks, in: 2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics, ICDL-EpiRob 2017, IEEE, 2017, pp. 66–71.

[25] R. Caccavale, A. Finzi, A robotic cognitive control framework for collaborative task execution and learning, Topics in Cognitive Science 14 (2022) 327–343.