ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

Open-Vocabulary Robotic Object Manipulation using Foundation Models

Stig Griebenow, Ozan Özdemir, Cornelius Weber and Stefan Wermter *

University of Hamburg - Knowledge Technology Group
Vogt-Kölln-Straße 30, 22527 Hamburg - Germany

Abstract. Classical vision-language-action models are limited by unidirectional communication, hindering natural human-robot interaction. The recent CrossT5 embeds an efficient vision action pathway into an LLM, but lacks visual generalization, restricting actions to objects seen during training. We introduce OWL×T5, which integrates the OWLv2 object detection model into CrossT5 to enable robot actions on unseen objects. OWL×T5 is trained on a simulated dataset using the NICO humanoid robot and evaluated on the new CLAEO dataset featuring interactions with unseen objects. Results show that OWL×T5 achieves zero-shot object recognition for robotic manipulation, while efficiently integrating vision-language-action capabilities.

1 Introduction

In recent years, artificial intelligence (AI) has emerged as a key methodology in the field of human-robot interaction (HRI). Classical vision-language-action (VLA) models primarily rely on unidirectional communication, limiting their capacity to interact naturally with humans. Recent crossmodal VLA models aim to address this limitation by also generating language. One such model, CrossT5 [1], successfully leverages a pre-trained large language model (LLM) for language-based tasks. The so-called late fusion architecture allows CrossT5 to be trained on a very limited dataset, enabling it to learn action execution while retaining the original language capabilities of the T5 LLM [2].

One of its limitations is a lack of visual generalization, which restricts it to performing actions only on objects seen during training. To address this limitation, this research proposes the OWL×T5 model, integrating the OWLv2 open-vocabulary object detection model [3] into the CrossT5 architecture as a vision capable foundation model alongside T5 to enable zero-shot generalization of robotic actions to novel objects. We train OWL×T5 on a small dataset generated in a simulated environment using the NICO humanoid robot [4]. For evaluation, we introduce a new simulated dataset, Crossmodal Language-Action on Everyday Objects (CLAEO), which features interactions with everyday objects in simulation (see Fig. 1). Results indicate that the model successfully generalizes actions to previously unseen objects while retaining the original language-based capabilities and data efficiency of CrossT5.

*The research was supported by the DFG under the Crossmodal Learning (TRR-169) project and by the Horizon Europe project TERAIS under Grant agreement 101079338.

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933.
Available from http://www.i6doc.com/en/.

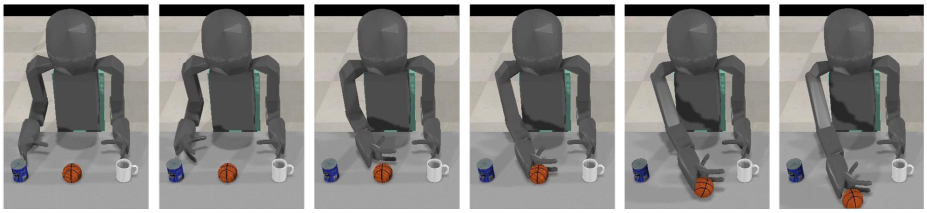


Fig. 1: A successful action execution of OWL×T5 on novel objects in simulation. The language input for this example was *push ball*.

2 Related Work

Recent research into crossmodal vision-language-action models often leverages foundation models, such as LLMs and open-vocabulary object detection models. Cross-modal models include LaMI, ELMiRA, RT-2, and CrossT5. LaMI [5] enhances human-robot interaction by guiding robots with a set of atomic actions generated through high-level language inputs, coordinating movements with speech to create multimodal expressions. Moving beyond traditional designs, it adopts an example-driven approach for HRI. ELMiRA [6] is a modular framework that combines LLMs with domain-specific models, enabling robots to understand commands, describe scenes, and manipulate objects. Integrating components like vision-language models, object detection, and spatial control, ELMiRA achieves robust interaction in tabletop scenarios, showcasing its open-ended HRI capabilities. CrossT5 [1], the focus of this research, builds a multimodal architecture within the T5 language model to facilitate action-language translation. By inserting a Crossmodal Transformer between the encoder and decoder of T5, CrossT5 achieves efficient training, strong language understanding, and effective robotic control. The RT-2 model [7] leverages vision-language models trained on Internet-scale data and fine-tunes them with robotic data, encoding actions as text tokens to support generalization to novel tasks.

We enhance CrossT5 because it offers a promising approach by integrating the T5 LLM in a unique way. Rather than transforming vision and action inputs into tokens fed into the LLM, CrossT5 employs a "late fusion" approach: the T5 model is split into encoder and decoder components, with the hidden representation fed into a Crossmodal Transformer alongside other inputs. This late fusion approach strikes a balance by combining the strengths of vision and language features, requiring significantly less training data and fine-tuning compared to "early fusion" models like RT-2. Additionally, unlike modular approaches such as ELMiRA and LaMI, which rely on a predefined action library where each action must be manually added, CrossT5 can theoretically be trained to generalize actions. This, provided there is sufficient action data, makes it adaptable and efficient.

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

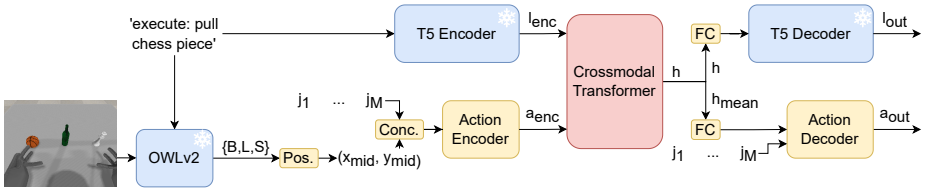


Fig. 2: OWL×T5 architecture in the *execute* case. The j vectors represent joint angle values from NICO's arms. $\{B, L, S\}$ denote {Bounding boxes, Labels, confidence Scores} from OWLv2, and (x_{mid}, y_{mid}) represents the position of the target object. *Conc.* denotes concatenation, *FC* is a fully connected layer, and *Pos.* is the position calculation.

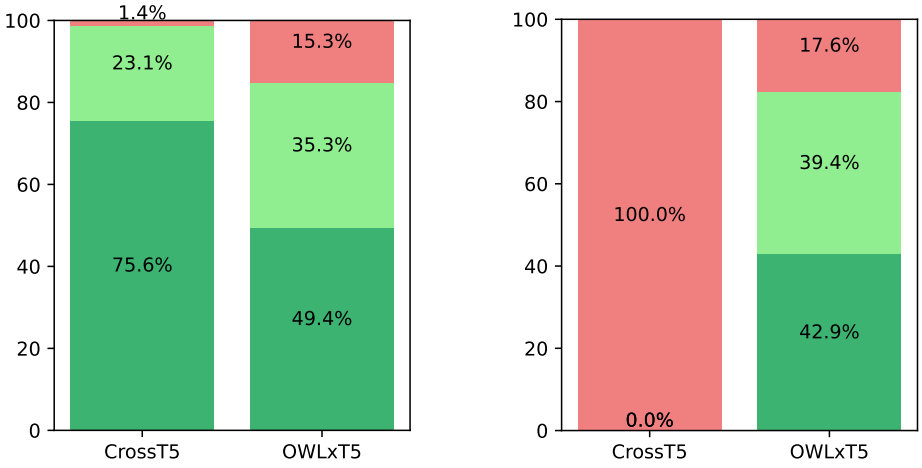
3 OWL×T5: Late Fusion with Object Detection

The CrossT5 model exhibits several limitations that restrict its versatility in complex human-robot interaction scenarios. The model object interaction capabilities are restricted to objects it has encountered during training. It was trained using colored cubes, constraining the model's ability to operate in more diverse environments with a wider range of objects. To lift this restriction we introduce the open vocabulary object detection model OWLv2 (base patch16) into the architecture as seen in Fig. 2. The T5 model (T5 small) is split into its encoder and decoder components. The output of the encoder $l_{\text{enc}} \in \mathbb{R}^{512}$ is fed into a crossmodal Transformer. The visual input is a coordinate $(x_{\text{mid}}, y_{\text{mid}})$ of the target object, which is generated by OWLv2 using both visual and language input. The visual coordinates and the action inputs $((j_1 \dots j_M), \quad j_i \in \mathbb{R}^{10}$, sequence length $M \leq 85)$ are concatenated and then fed into the action encoder, which is a Long Short-Term Memory (LSTM) network [8]. The output of the action encoder $a_{\text{enc}} \in \mathbb{R}^{512}$ is also fed into the crossmodal Transformer. The output of the Transformer $h \in \mathbb{R}^{512}$ is then passed through a fully connected (FC) layer before being fed into the T5 decoder, which produces the language output l_{out}. The mean (over the temporal dimension) of this output h_{mean}, along with the action inputs, is fed into the action decoder (also an LSTM), which generates the action output a_{out}.

The model is trained to perform in four distinct modes: *translate*, *execute*, *describe* and *repeat action*. In *translate* mode, it translates the language input from English into German (one of the capabilities of the T5 model). In *execute* mode, the model performs actions described by the text prompts. The *describe* mode lets the model generate language descriptions of the actions. In *repeat action* mode the model's goal is to reproduce a given action.

For training we utilize the mixed loss from Caesar et al. [1]. This approach uses the Mean Squared Error (MSE) between the predicted joint sequence and the target joint sequence for the action output a_{out}, the cross-entropy loss between the predicted and target tokens for the language output l_{out}, except in *translate* mode where the MSE between the hidden vector h, produced by the

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.



(a) Eval on cubes (CLANT) (b) Eval on novel objects (CLAEO)

Fig. 3: Comparison of the actions in the *execute* mode between CrossT5 and OWL×T5 on two datasets, cubes (CLANT) and novel objects (CLAEO). An execution is considered **perfect** if the correct object is moved in the intended direction beyond a specified threshold and no other object was moved. It is considered successful if all criteria are met except the threshold distance. An action is deemed a **failure** if one or more criteria are unmet.

Crossmodal Transformer, and a target hidden vector produced by T5 is used.

We train OWL×T5 on the CLANT (Crossmodal Language-Action and Natural Translation) dataset, introduced by Caesar et al. [1]. This dataset consists of 1440 action execution samples, with 1080 samples (75%) used for training and 360 samples (25%) for testing. Each sample in the action dataset includes a sequence of images, corresponding joint values, and a brief textual description of an action performed by the humanoid NICO robot. NICO interacts with three colored cubes positioned on a table in front of it, within a simulated environment generated using CoppeliaSim [9]. The robot utilizes either its left or right arm to perform one of 12 actions to move one of the cubes, with the image sequence captured by a camera mounted in NICO's eyes. Depending on the sample, sequences of 40, 60, or 85 images and joint angles of both arms are recorded. In total, the data set consists of $12 \cdot 120 = 1440$ samples (12 different actions and 120 possibilities to arrange 6 different coloured cubes on 3 positions). Those sequences are used for training the *execute*, *describe* and *repeat action* mode. For *translate* mode the Tilde RAPID 2019 German-to-English dataset from the ACL 2019 Conference [10] was integrated into CLANT. It comprises of sentence pairs derived from European Commission press releases. The first 1440 samples from RAPID 2019 were sufficient to re-establish the original capabilities of the split T5. We train the model with the Adam optimizer, for 10000 epochs, with a learning rate of 10^{-5} and a batch size of 64 .

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

4 Experiments in Simulation with NICO

We assess the performance of CrossT5 and OWL×T5 on all four modes: *execute*, *translate*, *repeat action*, and *describe*. The *execute* mode is further evaluated in simulation on the cubes seen during training (CLANT) and on novel objects to determine whether the new model effectively transfers its learned capabilities and demonstrates zero-shot learning on unseen objects. Designed to mirror the actions found in CLANT, we created a new evaluation dataset, CLAEO, applying these actions to a set of novel objects which are visually distinct from the original colored cubes. These novel objects were selected from RLBench and were not present in our training data [11]: a *mustard bottle*, a *can*, a *mug*, a *white chess piece*, a *ball*, and an *empty wine bottle*. Due to instabilities in simulation, we let them retain the simple physical properties of the cubes.

Fig. 3 shows the results of the execution evaluation for both the baseline (CrossT5) and OWL×T5 across two different object categories: *cubes* and *novel objects*. The results indicate that while the CrossT5 model performs slightly better on the familiar cubes, the OWL×T5 model demonstrates improved generalization to novel objects. Specifically, OWL×T5 shows a marked improvement in successfully executing actions on novel objects, where CrossT5 fails to generalize.

Table 1 assesses the language production quality of both models using BLEU2 scores. Despite being trained on language translation less extensively, our model surprisingly outperforms Cross-T5 in the translation task, achieving a higher BLEU2 score. Important to note is that for *translate* mode we are not comparing its output to the dataset's target translations, but directly to the output of the original T5 model, because OWL×T5 is trained to reproduce the language encodings from the T5 model. Thus its translation performance will inherently be limited by that of T5 (T5: BLEU2 of 42.12% on CLANT vs. ours: 40.42%). Our model also generalizes to novel objects in *describe* mode, unlike CrossT5.

We also evaluate the action output for *repeat action* mode using the Normalized Root-Mean Squared Error. CrossT5 achieves 0.25% on CLANT and 6.88% on CLAEO. OWL×T5 achieves 0.37% on CLANT and 0.63% on CLAEO. These results reaffirm that our model can transfer the learnt tasks to new objects and CrossT5 cannot. Overall, our evaluation shows a zero-shot generalization across all 4 tasks.

	CLANT		**CLAEO**	
	CrossT5	**OWL×T5**	**CrossT5**	**OWL×T5**
translate	86.75%	91.68%	—	—
describe	100.00%	100.00%	0.00%	100.00%

Table 1: Language output accuracy (BLEU2) of CrossT5 and OWL×T5 in *translate* and *describe* mode on both datasets. Since CLAEO has the same translation samples as CLANT, the corresponding fields are left empty.

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

5 Conclusion

We have introduced OWL×T5, a novel crossmodal VLA architecture. OWL×T5 integrates the OWLv2 open-vocabulary object detection model into the T5 LLM, enabling the generalization of actions to previously unseen objects in a zero-shot manner. Evaluation with the CLAEO dataset demonstrates OWL×T5's strong performance, particularly in handling new objects, underscoring its enhanced generalization potential for real-world applications. This research contributes to human-robot interaction by supporting more natural and adaptable interactions without requiring extensive training data.

References

[1] Anton Caesar, Ozan Özdemir, Cornelius Weber, and Stefan Wermter. Enabling action crossmodality for a pretrained large language model. *Natural Language Processing Journal*, 7:100072, Jun 2024.

[2] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[3] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36, 2023.

[4] Matthias Kerzel, Erik Strahl, Sven Magg, Nicolás Navarro-Guerrero, Stefan Heinrich, and Stefan Wermter. NICO–Neuro–Inspired COmpanion: A developmental humanoid robot platform for multimodal interaction. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 113–120. IEEE, 2017.

[5] Chao Wang, Stephan Hasler, Daniel Tanneberg, Felix Ocker, Frank Joublin, Antonello Ceravola, Joerg Deigmoeller, and Michael Gienger. LaMI: Large language models for multi-modal human-robot interaction. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–10, 2024.

[6] Connor Gäde, Ozan Özdemir, Cornelius Weber, and Stefan Wermter. Embodying language models in robot action. In *Proceedings of the 32nd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2024)*, pages 625–630, Oct 2024.

[7] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR, 06–09 Nov 2023.

[8] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[9] Eric Rohmer, Surya PN Singh, and Marc Freese. V-REP: A versatile and scalable robot simulation framework. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1321–1326. IEEE, 2013.

[10] ACL 2019 Fourth Conference on Machine Translation (WMT19), shared task: Machine translation of news. `http://www.statmt.org/wmt19/translation-task.html`, 2019. Accessed: 2024-05-12.

[11] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. RLBench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.