ALLMAN: A German Vision-Language-Action Model

Christian Scherer*,[1] Maximilian Tölle*,[1,2] Theo Gruner*,[1,3]
Daniel Palenicek*,[1,3] Tim Schneider*,[1] Patrick Schramowski[1,2,3]
Boris Belousov Jan Peters[1,2,3,4,5]

***Abstract*— Large vision-language-action (VLA) models have shown remarkable capabilities for learning general robot policies. However, the predominance of English in both the large language model (LLM) backbone training data and the robotics data limits the accessibility of these models. Specifically, training such policies for other languages, such as German, extends their usefulness to the non-English-speaking rest of the world. We present ALLMAN, the first German VLA model, built upon LEOLM — a Llama 2-based LLM specifically fine-tuned on large German datasets. To train ALLMAN, we machine-translate several English vision-language and VLA datasets into German. We then adapt the PRISMATIC and OPENVLA training pipelines to create our German VLA model. Through comparative analyses with OPENVLA, we demonstrate the importance of incorporating German language capabilities within the base model. Our findings underscore the importance of training VLAs in other languages beyond English. This work serves as a proof-of-concept for multi-language VLAs, paving the way for broader, more inclusive robotics applications.**

"Alman: [oft scherzhaft] Person, die stereotypisch deutsche Verhaltensweisen und Eigenschaften wie übermäßige Ordentlichkeit, Regelbefolgung und Pünktlichkeit aufweist"[1]

∼ Digital Dictionary of the German Language [1]

I. INTRODUCTION

The vision of integrating adaptable robots into our everyday lives is a fast-approaching reality [3]. Robots can potentially assist us in diverse areas, from elderly care to home management. However, in many potential application areas, robots must collaborate with or receive instructions from humans. To make this collaboration accessible and seamless for non-expert humans, it is important that the robot is able to interpret, act upon, and respond in natural language. Since many people prefer speaking in their native language, multilingualism is critical to making robots accessible and useful in our global, multilingual society.

Today's VLA models enable robots to perceive their environment visually, understand natural language instructions, and transform these into physical robot actions [4], [5], [6], [7], [8], [9], [10]. These models are often based on pre-trained vision-language backbones. However, a common drawback of current VLAs is their limited support for languages other than English. While multilingual LLM are prevalent [11], [12], [13], there are few multilingual vision-language model (VLM) [14], [15], [16] and no multilingual VLA models yet. For example, LLAMA 2's pre-training data is approximately 90% English, with German — despite being the third most supported language — making up only 0.17% of the dataset [17].

We take the first step toward closing this gap by developing a VLA model specifically trained for the German language, named ALLMAN. We base ALLMAN directly on the German LLAMA 2-based model LEOLM [18], which we transform into a VLM before training it to a VLA. We evaluate ALLMAN on the LIBERO [19] evaluation benchmark and show that our model can follow German language instructions. Our contribution marks an essential step in enabling multilingual human-robot interactions, making robots more adaptable to diverse linguistic environments.

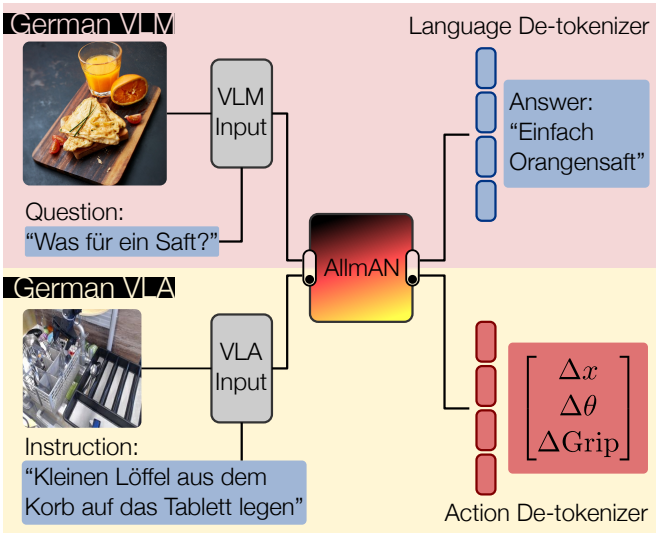

Fig. 1. The ALLMAN training stages from VLM to VLA training.

II. CREATING ALLMAN

Translating VL(A) datasets into german. Training large models like ALLMAN requires enormous amounts of training data. We used the DeepL [2] translation service to translate existing datasets from English to German. In total, we created three German datasets listed in Table I containing 200M tokens (587M characters) for the different training stages of ALLMAN. We translated LLaVA-1.5 mix665k [20] for VLM training, the Open X-Embodiment

*Equal contribution [1]TU Darmstadt [2]German Research Center for Artificial Intelligence (DFKI) [3]hessian.AI [4]Robotics Institute Germany (RIG) [5]Centre for Cognitive Science
Correspondence:
{maximilian,theo,palenicek}@robot-learning.de

[1] *"Alman: [often jokingly] person who exhibits stereotypically German behavior and characteristics such as excessive neatness, rule-following and punctuality"* (translated by the DeepL translation service [2])

TABLE I

OVERVIEW OF THE MACHINE-TRANSLATED DATASETS.

Dataset	#Tokens (DE)	Function
`LLaVA-1.5 mix665k DE`	197 569 913	VLM Training
`Open X-Embodiment DE`	2 459 185	VLA Training
`LIBERO DE`	714	VLA Finetuning
Total	200 029 812	

(OXE) dataset [5] for VLA training, and LIBERO [19] for VLA finetuning and evaluation.

The ALLMAN model training pipeline. The main backbone of ALLMAN is the German Llama 2-based model LEOLM [18]. Following [21], we transform the LLM into a VLM by adding the ensemble of DINO V2 [22] and SigLIP [23] as visual backbones. The patch features are projected into the embedding space of LEOLM using a small MLP projector. We train the newly obtained architecture by using the Prismatic training pipeline [21] on our German translated `LLaVA-1.5 mix665k DE` dataset. Training of the German VLM was performed on 16 Nvidia A100 GPUs for 3 full epochs (250 GPU hours).

Using the OPENVLA framework and the OCTO-weighted [8] data mixture of our German translated `Open X-Embodiment DE` dataset, we train ALLMAN for a total of 17,545 GPU hours on 64 Nvidia A100 GPUs. Out of time and resource constraints, we prematurely stopped training after 18 epochs (OPENVLA was trained for 27 epochs) and can report an obtained training action token accuracy of 0.75, compared to 0.95 reported by OPENVLA. We plan to continue training from this checkpoint.

Finally, we finetune ALLMAN on our translated `LIBERO DE` dataset for evaluation against OPENVLA. The `LIBERO` dataset [19] consists of four task suites. Following the OPENVLA protocol, we finetune ALLMAN separately on each LIBERO task suite until convergence at 100% action token accuracy. Finetuning takes 50k-80k gradient steps depending on the task suite and requires 19 hours on 32 Nvidia A100 GPUs.

III. INVESTIGATING ALLMAN'S AND OPENVLA'S PERFORMANCE ON GERMAN TASKS

Based on the LIBERO benchmark [19], we evaluate the performance of ALLMAN and OPENVLA on German tasks. Since ALLMAN is a compute-intensive approach, we also explore less demanding options: *(i)* naively evaluating the OPENVLA model on German instructions, *(ii)* using a translation model to convert German instructions into English and querying the original OPENVLA model, and *(iii)* directly fine-tuning the OPENVLA model on the translated `LIBERO DE` dataset.

We compare the models' performances across four datasets, as shown in Fig. 2. These consist of the `LIBERO DE` dataset as well as a paraphrased version of it to assess the generalization ability across linguistic variations. In addition, we test the models with random instructions from the `LIBERO DE` dataset and with no prompts at all to ensure

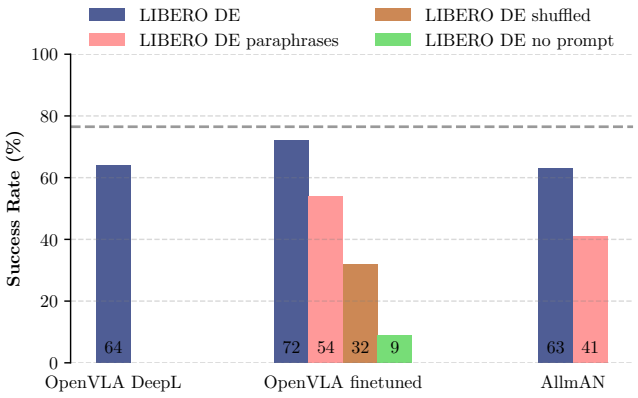

Fig. 2. Evaluation results of the different versions of OpenVLA as well as AllmAN. We report success rates averaged over four `LIBERO DE` task suites on 3 seeds each. For reference, we report the average OPENVLA success rate over the (English) task suits of the original paper [9] by the grey dotted line.

the VLAs responds appropriately to language prompts and does not overfit on the visual input.

We first evaluated the original OPENVLA on German instructions, achieving only 1% success (not displayed). Translating German instructions back to English using DeepL[2] restores performance to 64%, confirming the model's issue is language understanding. Fine-tuning OPENVLA on German instructions improves success to 72%. The performance drops when faced with misleading or missing German instructions but is still much higher than expected, suggesting overfitting to visual inputs. In fact, in the VLM literature, some works have reported that multimodal models can learn to ignore modalities completely [24], [25], [26]. So far, the ALLMAN model underperforms due to early stopping in training.

IV. CONCLUSION

We introduce ALLMAN, the first VLA based on a German backbone capable of understanding and executing German language instructions. Our approach involves machine-translating existing VL(A) datasets to the German language and adapting the training protocol of PRISMATIC and OPENVLA to the LEOLM backbone model. ALLMAN provides proof of concept for the feasibility of non-English VLAs, encouraging further research into multilingual models that can accommodate a broader range of linguistic and cultural contexts.

Limitations and future work. While ALLMAN is a promising step towards multilingual VLAs, several limitations exist. First, ALLMAN is a German-specific model, not a truly multilingual one. Expanding this approach to cover additional languages requires further adaptation and the creation of new datasets. Additionally, we prematurely stopped ALLMAN's training due to time and resource constraints, which likely impacted its performance, especially compared to fine-tuned OPENVLA. We evaluate the model solely on the `LIBERO DE` dataset in a simulated environment; future work focuses on evaluating in real-world robotic settings.

ACKNOWLEDGMENT

This research was supported by the "Third Wave of AI", funded by the Excellence Program of the Hessian Ministry of Higher Education, Science, Research and Art. We also acknowledge the grant "Einrichtung eines Labors des Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) an der Technischen Universität Darmstadt" of the Hessian Ministry of Science and Research, Arts and Culture.

REFERENCES

[1] "DWDS - Der deutsche Wortschatz von 1600 bis heute." Oct. 2024, [Online; accessed 15. Oct. 2024]. [Online]. Available: https://www.dwds.de/d/about.en

[2] "DeepL Übersetzer: Der präziseste Übersetzer der Welt," Oct. 2024, [Online; accessed 15. Oct. 2024]. [Online]. Available: https://www.deepl.com/de/translator

[3] N. M. M. Shafiullah, A. Rai, H. Etukuru, Y. Liu, I. Misra, S. Chintala, and L. Pinto, "Dobb-E: On Bringing Robots Home," 2023.

[4] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, *et al.*, "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," in *7th Conference on Robot Learning (CoRL)*, 2023.

[5] O. X.-E. Collaboration, A. O'Neill, A. Rehman, A. Gupta, A. Maddukuri, A. Gupta, *et al.*, "Open X-Embodiment: Robotic Learning Datasets and RT-X Models," in *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition at 7th Conference on Robot Learning (CoRL)*, 2023.

[6] X. Li, M. Liu, H. Zhang, C. Yu, J. Xu, H. Wu, *et al.*, "Vision-Language Foundation Models as Effective Robot Imitators," in *12th International Conference on Learning Representations (ICLR)*, 2024.

[7] Andrew Sohn, Anusha Nagabandi, Carlos Florensa, Daniel Adelberg, Di Wu, Hassan Farooq, *et al.*, "Covariant - Introducing RFM-1: Giving robots human-like reasoning capabilities," Tech. Rep., 2024.

[8] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, *et al.*, "Octo: An Open-Source Generalist Robot Policy," in *Robotics: Science and Systems (RSS)*, 2024.

[9] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, *et al.*, "OpenVLA: An Open-Source Vision-Language-Action Model," in *8th Conference on Robot Learning (CoRL)*, 2024.

[10] C. E. Mower, Y. Wan, H. Yu, A. Grosnit, J. Gonzalez-Billandon, M. Zimmer, *et al.*, "Ros-llm: A ros framework for embodied ai with task feedback and structured reasoning," in *arXiv preprint arXiv:2406.19741*, 2024.

[11] B. Workshop, T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, *et al.*, "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model," 2023.

[12] Llama Team and AI @ Meta, "The Llama 3 Herd of Models," 2024.

[13] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, *et al.*, "GPT-4 Technical Report," 2024.

[14] X. Chen, X. Wang, S. Changpinyo, A. J. Piergiovanni, P. Padlewski, D. Salz, *et al.*, "PaLI: A Jointly-Scaled Multilingual Language-Image Model," in *11th International Conference on Learning Representations (CVPR)*, 2023.

[15] X. Chen, J. Djolonga, P. Padlewski, B. Mustafa, S. Changpinyo, J. Wu, *et al.*, "PaLI-X: On Scaling up a Multilingual Vision and Language Model," in *Conference on Computer Vision and Pattern Recognition*, 2024.

[16] X. Chen, X. Wang, L. Beyer, A. Kolesnikov, J. Wu, P. Voigtlaender, *et al.*, "PaLI-3 Vision Language Models: Smaller, Faster, Stronger," 2023.

[17] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models," 2023.

[18] LAION, "LeoLM/leo-hessianai-7b · Hugging Face," https://huggingface.co/LeoLM/leo-hessianai-7b, 2023.

[19] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone, "Libero: Benchmarking knowledge transfer for lifelong robot learning," in *Advances in Neural Information Processing Systems*, 2024.

[20] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved Baselines with Visual Instruction Tuning," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[21] S. Karamcheti, S. Nair, A. Balakrishna, P. Liang, T. Kollar, and D. Sadigh, "Prismatic VLMs: Investigating the Design Space of Visually-Conditioned Language Models," in *41st International Conference on Machine Learning (ICML)*, 2024.

[22] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, *et al.*, "DINOv2: Learning Robust Visual Features without Supervision," *Transactions on Machine Learning Research (TMLR)*, 2024.

[23] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid Loss for Language Image Pre-Training," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[24] R. Cadene, C. Dancette, M. Cord, D. Parikh, *et al.*, "Rubi: Reducing unimodal biases for visual question answering," *Advances in neural information processing systems*, vol. 32, 2019.

[25] R. Hu, D. Fried, A. Rohrbach, D. Klein, T. Darrell, and K. Saenko, "Are you looking? grounding to multiple modalities in vision-and-language navigation," *arXiv preprint arXiv:1906.00347*, 2019.

[26] C. Clark, M. Yatskar, and L. Zettlemoyer, "Don't take the easy way out: Ensemble based methods for avoiding known dataset biases," *arXiv preprint arXiv:1909.03683*, 2019.