Jose Edgar Hernandez Cancino Estrada
Komenskega ulica 22, 1000 Ljubljana, Slovenija
Study programme: Erasmus Mundus Joint Master in Artificial Intelligence
Enrollment number: 63250515

Committee for Student Affairs
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Večna pot 113, 1000 Ljubljana

The master's thesis topic proposal
Candidate: Jose Edgar Hernandez Cancino Estrada

I, Jose Edgar Hernandez Cancino Estrada, a student of the 2nd cycle study programme at the Faculty of computer and information science, am submitting a thesis topic proposal to be considered by the Committee for Student Affairs with the following title:

Slovenian: **Vizualno-jezikovno-akcijski model za dvoročno robotsko manipulacijo s tekstilnimi objekti**
English: **Vision-Language-Action model for bimanual robotic manipulation of textile objects**

This topic was already approved last year: *NO*

I declare that the mentors listed below have approved the submission of the thesis topic proposal described in the remainder of this document.

I would like to write the thesis in English with the following reason: I am not a native speaker, and I am a student of the Erasmus Mundus Joint Master in Artificial Intelligence.

I propose the following mentor:

> Name, surname and title: Prof. dr. Danijel Skočaj
> Institution: Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
> E-mail: danijel.skocaj@fri.uni-lj.si

I propose the following co-mentor:

> Name, surname and title: dr. Domen Tabernik
> Institution: Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
> E-mail: domen.tabernik@fri.uni-lj.si

Ljubljana, December 5, 2025.

Key-words

Artificial Intelligence, Visual-Language-Action Models, Deep Learning, Multimodal Learning, Bimanual Robotic Manipulation, Textile Manipulation

Detailed thesis proposal

Problem & State of the Art. Robotic manipulation of textiles is a challenging task that requires complex action planning based on accurate perception and structured temporal reasoning. Among existing work, *BiFold* [1] introduces temporal context into Vision-Language-Action (VLA) for bimanual folding tasks, *GraphGarment* [2] models cloth dynamics over time using Graph Neural Networks, and *CeDiRNet-3DoF* [3] enables specialized cloth grasp pose prediction from visual data. These models can, however, be computationally demanding, constrained to specific scopes, or lack end-to-end VLA integration. Recent frameworks like *SmolVLA* [4] and *TwinVLA* [5] point toward a clear way of aligning compact VLA models for bimanual manipulation that can potentially be boosted by specialized perception or temporal knowledge.

Expected Contributions / Technical outcome. We will develop a compact VLA model based on SmolVLA, but designed for bimanual textile manipulation, integrating temporal reasoning and *CeDiRNet-3DoF* [3]'s perception capabilities. It will (i) align both arms through shared multimodal context as in *TwinVLA* [5], (ii) incorporate temporal cues for long-horizon planning, and (iii) efficiently generalize across diverse textile tasks, conditioned by natural language instructions.

Methodology & Validation. The model will be developed using PyTorch and will be tested in simulation and on a real SO-ARM101 dual-arm platform [6]. The dataset and benchmark from the *ICRA 2024 Cloth Competition* [7, 8], will be used to evaluate and compare our performance in cloth manipulation (grasp success and coverage proportion for unfolding), alongside the *Benchmarking Bimanual Cloth Manipulation* [9] protocols. To evaluate the language-conditioned behavior, we will use the metrics proposed in *Evaluating Uncertainty and Quality of VLA-enabled* Robots [10] for measuring execution accuracy and model confidence.

References

[1] O. Barbany, A. Colomé, C. Torras, Bifold: Bimanual cloth folding with language guidance, in: 2025 IEEE International Conference on Robotics and Automation (ICRA), 2025, pp. 5652–5659. doi:10.1109/ICRA55743.2025.11127549.

[2] W. Chen, K. Li, D. Lee, X. Chen, R. Zong, P. Kormushev, Graphgarment: Learning garment dynamics for bimanual cloth manipulation tasks (03 2025). doi:10.48550/arXiv.2503.05817.

[3] D. Tabernik, J. Muhovič, M. Urbas, D. Skočaj, Center direction network for grasping point localization on cloths, IEEE Robotics and Automation Letters 9 (10) (2024) 8913–8920. doi:10.1109/LRA.2024.3455802.

[4] M. Shukor, D. Aubakirova, F. Capuano, P. Kooijmans, S. Palma, A. Zouitine, M. Aractingi, C. Pascal, M. Russi, A. Marafioti, S. Alibert, M. Cord, T. Wolf, R. Cadene, Smolvla: A vision-language-action model for affordable and efficient robotics (2025). arXiv:2506.01844.

[5] H. Im, E. Jeong, J. Fu, A. Kolobov, Y. Lee, Twinvla: Data-efficient bimanual manipulation with twin single-arm vision-language-action models (2025). arXiv:2511.05275.

[6] R. Cadene, S. Alibert, A. Soare, Q. Gallouedec, A. Zouitine, S. Palma, P. Kooijmans, M. Aractingi, M. Shukor, D. Aubakirova, M. Russi, F. Capuano, C. Pascal, J. Choghari, J. Moss, T. Wolf, Lerobot: State-of-the-art machine learning for real-world robotics in pytorch, `https://github.com/huggingface/lerobot` (2024).

[7] V.-L. Gusseme, T. Lips, R. Proesmans, J. Hietala, G. Lee, J. Choi, J. Choi, G. Kim, P. Yonrith, D. Tabernik, A. Gams, P. Nimac, M. Urbas, J. Muhovič, D. Skočaj, M. Mavsar, H. Yu, M. Kwon, Y. Kim, F. Wyffels, A dataset and benchmark for robotic cloth unfolding grasp selection: The icra 2024 cloth competition (08 2025). doi:10.48550/arXiv.2508.16749.

[8] Y. Sun, B. Calli, K. Kimble, F. wyffels, V.-L. De Gusseme, K. Hang, S. D'Avella, A. Xompero, A. Cavallaro, M. A. Roa, J. Avendano, A. Mavrommati, Robotic grasping and manipulation competition at the 2024 ieee/ras international conference on robotics and automation [competitions], IEEE Robotics Automation Magazine 31 (4) (2024) 174–185. doi:10.1109/MRA.2024.3481609.

[9] I. Garcia-Camacho, M. Lippi, M. C. Welle, H. Yin, R. Antonova, A. Varava, J. Borras, C. Torras, A. Marino, G. Alenyà, D. Kragic, Benchmarking bimanual cloth manipulation, IEEE Robotics and Automation Letters 5 (2) (2020) 1111–1118. doi:10.1109/LRA.2020.2965891.

[10] P. Valle, C. Lu, S. Ali, A. Arrieta, Evaluating uncertainty and quality of visual language action-enabled robots (07 2025). doi:10.48550/arXiv.2507.17049.