Offer #2025-09022

PhD Position F/M Embodied AI

Contract type : Fixed-term contract

Level of qualifications required : Graduate degree or equivalent

Fonction : PhD Position

Level of experience : Recently graduated

Context

The Phd will be done at Inria in the Willow research team.

Assignment

Recent progress in embodied AI has enabled significant advancements in robots' ability to understand the physical world. While learning-based robotic policies often struggle with generalization, foundational models impart rich scene understanding to robotics, thereby opening the path for addressing various challenges. Robotic manipulation is one such challenge that involves handling diverse objects and novel placements.

Generalization: The project proposes to use vision-language models (VLM) to improve object grounding capabilities of manipulation systems. Previous approaches have used Mask-RCNN [1], Segment Anything Model (SAM) [2] to segment out novel objects. In contrast, the proposed method would use Vision-language Models (VLMs) to handle novel object instances as well as generalize to unseen object categories by grounding objects from multi-view images to obtain precise poses.
Successful manipulation also requires an in-depth understanding of object affordances. Thus, the project would also explore fine-tuning VLMs to predict spatial affordances of objects by leveraging large-scale internet video demonstrations [3, 4]. These video demonstrations provide multiple instances and scenarios to grip a particular object. These demonstrations can be used to generate datasets with heatmaps or designated grasp points, and introduce task variations and diversity. Vision-language benchmarks such as [2, 1, 5] will be used to evaluate the effectiveness of vision-language-guided manipulation in constrained settings, including changes to environment and camera setup.

Long-Horizon Tasks: Another challenge in manipulation involves generalizing to long-horizon tasks. Initial attempts break the tasks to a hierarchy of skills and perform skill-chaining [6, 7]. However, they operate on a limited category of sub-tasks and assume a pre-defined hierarchy. The project proposes using imitation learning to train foundation models or VLMs on video demonstrations (collected via simulation or teleoperation) of long-horizon tasks. The objective is to enable these models to generate low-level, fine-grained actions by learning reference trajectories from human demonstrations [8, 9, 10]. Comparative evaluations will be conducted against traditional diffusion-based or RL-based policies. The project would also explore decomposing a long horizon task into a sequence of intermediate tasks intertwining multi-modal prompts and task descriptions for the VLM.

Inference Speed: A major drawback of foundation models is inference speed. I intend to tackle hardware constraints limiting the deployment of multimodal learning in the real-world. To address this limitation, the project aims to develop specialized, quantized models that are trained for specific categories of tasks. By incorporating vision-language action models such as TinyVLA [11, 12], the approach seeks to achieve real-time performance on physical robotic manipulators, thus facilitating faster and more efficient execution of tasks.

Dexterous Manipulation: In addition to generalization and long-horizon tasks, the project focuses on dexterous manipulation. Current learning-based methods, such as ViViDex [13], employ imitation learning from video demonstrations combined with trajectory-guided rewards to train RL policies. The project seeks to extend these methods by introducing novel objects and unfamiliar tasks, enabling vision-language models to acquire robust dexterous manipulation skills. Moreover, these models are expected to support long-horizon tasks and recovery from failure scenarios in highly cluttered environments [14].

In summary, the project explores vision-language guided robotic manipulation and advances object grounding and spatial affordance estimation for novel objects. The project also proposes decomposing long-horizon tasks into fine-grained actions, and specialized, quantized models for real-time performance. These components collectively contribute to generalization, efficient task execution, and improved dexterous manipulation.

References

[1] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan, “Vima: General robot manipulation with multimodal prompts,” in Fortieth International Conference on Machine Learning, 2023.
[2] R. Garcia, S. Chen, and C. Schmid, “Towards generalizable vision-language robotic manipulation: A benchmark and llm-guided 3d policy,” in ICRA 2025.
[3] K. Fang, T.-L. Wu, D. Yang, S. Savarese, and J. J. Lim, “Demo2vec: Reasoning object affordances from online videos,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2139–2147, 2018.
[4] W. Yuan, J. Duan, V. Blukis, W. Pumacay, R. Krishna, A. Murali, A. Mousavian, and D. Fox,
“Robopoint: A vision-language model for spatial affordance prediction in robotics,” in 8th Annual Conference on Robot Learning, 2024.
[5] K. Zheng, X. Chen, O. Jenkins, and X. E. Wang, “VLMbench: A compositional benchmark for vision-and-language manipulation,” in NeurIPS, 2022.
[6] S. Cheng and D. Xu, “League: Guided skill learning and abstraction for long-horizon manipulation,” 2023.
[7] Z. Chen, Z. Ji, J. Huo, and Y. Gao, “SCar: Refining skill chaining for long-

horizon robotic manipulation via dual regularization," in The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024.
[8] F. Ceola, L. Natale, N. S ¨underhauf, and K. Rana, "Lhmanip: A dataset for long-horizon language grounded manipulation tasks in cluttered tabletop environments," 2024.
[9] G. Chen, M. Wang, T. Cui, Y. Mu, H. Lu, T. Zhou, Z. Peng, M. Hu, H. Li, Y. Li, Y. Yang, and Y. Yue, "Vlmimic: Vision language models are visual imitation learner for fine-grained actions," 2024.
[10] X. Li, M. Liu, H. Zhang, C. Yu, J. Xu, H. Wu, C. Cheang, Y. Jing, W. Zhang, H. Liu, H. Li,
and T. Kong, "Vision-language foundation models as effective robot imitators," in The Twelfth International Conference on Learning Representations, 2024.
[11] J. Wen, Y. Zhu, J. Li, M. Zhu, K. Wu, Z. Xu, N. Liu, R. Cheng, C. Shen, Y. Peng, F. Feng, and
J. Tang, "Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation," 2024.
[12] M. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Fos-
ter, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine,
P. Liang, and C. Finn, "Openvla: An open-source vision-language-action model," arXiv preprint
arXiv:2406.09246, 2024.
[13] Z. Chen, S. Chen, E. Arlaud, I. Laptev, and C. Schmid, "Vividex: Learning vision-based dexterous manipulation from human videos," 2025.
[14] H. Liu, S. Guo, P. Mai, J. Cao, H. Li, and J. Ma, "Robodexvlm: Visual language model-enabled task planning and motion control for dexterous robot manipulation," 2025

Main activities

Main activities:

- Analyse and implement related work.
- Design novel innovative solutions.
- Write progress reports and papers.
- Present work at conferences.

Skills

Technical skills and level required : programming skills are required.

Languages : English and possibly French.

Relational skills : Good communication skills.

Benefits package

- Subsidized meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- Possibility of teleworking and flexible organization of working hours
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Social security coverage

General Information

- **Theme/Domain :** Robotics and Smart environments
  Statistics (Big data) (BAP E)
- **Town/city :** Paris
- **Inria Center :** Centre Inria de Paris
- **Starting date :** 2025-09-01
- **Duration of contract :** 3 years
- **Deadline to apply :** 2025-07-24

Contacts

- **Inria Team :** WILLOW
- **PhD Supervisor :**
  Schmid Cordelia / cordelia.schmid@inria.fr

About Inria

Inria is the French national research institute dedicated to digital science and technology. It employs 2,600 people. Its 200 agile project teams, generally run jointly with academic partners, include more than 3,500 scientists and engineers working to meet the challenges of digital technology, often at the interface with other disciplines. The Institute also employs numerous talents in over forty different professions. 900 research support staff contribute to the preparation and development of scientific and entrepreneurial projects that have a worldwide impact.

The keys to success

Essential qualities in order to fulfil this assignment are feeling at ease in an environment of scientific dynamics and wanting to learn and listen. Prior experience in research is a plus.

Warning : you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.

Instruction to apply

Defence Security :
This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating to the protection of national scientific and technical potential (PPST).Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

Recruitment Policy :
As part of its diversity policy, all Inria positions are accessible to people with disabilities.