Analyse des modèles de fondation pour la navigation tout-terrain

Quentin Picard, Louis Dezons, David Filliat

Analyse des modèles de fondation pour la navigation tout-terrain

Quentin Picard[1], Louis Dezons[1,2], David Filliat[2]

[1] AMIAD, Pôle Recherche, Palaiseau
[2] U2IS, ENSTA, Institut Polytechnique de Paris, Palaiseau

quentin.picard@polytechnique.edu

Résumé

La navigation robotique tout-terrain est un sujet important pour la défense qui vise à déployer un robot sur un terrain inconnu et à pouvoir naviguer de manière robuste sur différents types de sol (béton, terre, sable, boue). Ce résumé analyse l'utilisation des modèles de fondation qui démontrent des caractéristiques de généralisation pour la robotique et les opportunités pour les appliquer à la navigation.

Mots-clés

robotique, navigation, modèle de fondation, VLA, VLM

Abstract

Offroad robotic navigation is an important topic for defense, aiming to deploy a robot in an unknown environment with robust navigation across various ground (concrete, soil, sand, mud). This summary analyzes the use of foundation models that demonstrate generalization characteristics for robotics and the opportunities to apply them to navigation.

Keywords

robotics, navigation, fundation model, VLA, VLM

1 Introduction

Le développement de modèles de fondation se basant sur les grands modèles de langage a révolutionné de nombreux domaines, dont l'analyse de scène et son application à la robotique, en apportant notamment des capacités de généralisation sans précédent. Dans un premier temps, ils permettent de comprendre l'environnement dans lequel se trouve le robot, par exemple pour définir les espaces d'intérêts à partir d'observations en suivant des instructions textuelles [18]. Dans un second temps, ils peuvent générer des actions pour commander un robot en fonction des modalités d'entrée du modèle [2, 8, 11].

Ce résumé propose deux principales contributions : 1/ une analyse des récents modèles de vision-langage-action pour la robotique où, à partir d'une instruction et d'observation, une commande est envoyée au robot. 2/ une étude des limitations des modèles actuels ainsi que les opportunités pour la navigation tout-terrain.

2 Des VLM aux VLA

Les modèles vision-langage-action (VLA) [9] visent à étendre les capacités de compréhension de la scène des modèles vision-langage (VLM) en générant directement des commandes de contrôle pour un robot, que ce soit pour la manipulation d'objets ou la navigation en intérieur ou extérieur. Une architecture générique d'un VLA intègre quatre principales parties issues du VLM : l'encodeur d'image, le projecteur, le *tokenizer*, le modèle de langage (LLM) et le module d'action du robot pour avoir un modèle VLA, comme le montre la Figure 1.

2.1 Compréhension de l'environnement

L'encodeur d'image utilise des modèles dont l'architecture se base sur l'utilisation de *transformers*, comme [16, 19, 12]. Ces modèles, pré-entraînés sur une large quantité de données, produisent des caractéristiques d'encodage riches et transférables à différentes tâches sans avoir besoin de ré-entraînement. Le choix de l'encodeur dépend de ses performances pour différentes applications de perception connues, comme la classification, la segmentation ou l'estimation de profondeur. Le projecteur a ensuite pour rôle d'aligner et de faire correspondre la dimensionalité de la sortie de l'encodeur à la dimension des jetons d'entrée du LLM.

Le LLM représente la majeure partie du modèle VLA en termes de calcul et de mémoire. En pratique, n'importe quel modèle de langage pré-entraîné sur une large quantité de données peut être utilisé. De nombreux modèles en libre accès sont disponibles permettant d'être intégré au VLA, comme Llama, Gemma ou Mistral. Le choix du LLM dépend de deux principaux aspects : 1/ la capacité de généralisation à différentes observations et 2/ le nombre de paramètres. Ce dernier se compte en milliard pour la majorité des modèles et apporte une complexité calculatoire importante. Bien que cela permette au modèle de langage de décrire l'observation avec précision, l'exécution temps-réel pour la robotique mobile sur une plateforme embarquée reste limité [17]. Le développement de modèles plus léger comme LLaVA-Mini [20] pouvant être embarqués est un sujet d'intérêt pour la robotique et les systèmes autonomes, mais est hors du champ de ce résumé.

La majorité des modèles de fondation prennent en considération deux principales modalités, l'image et le texte. De

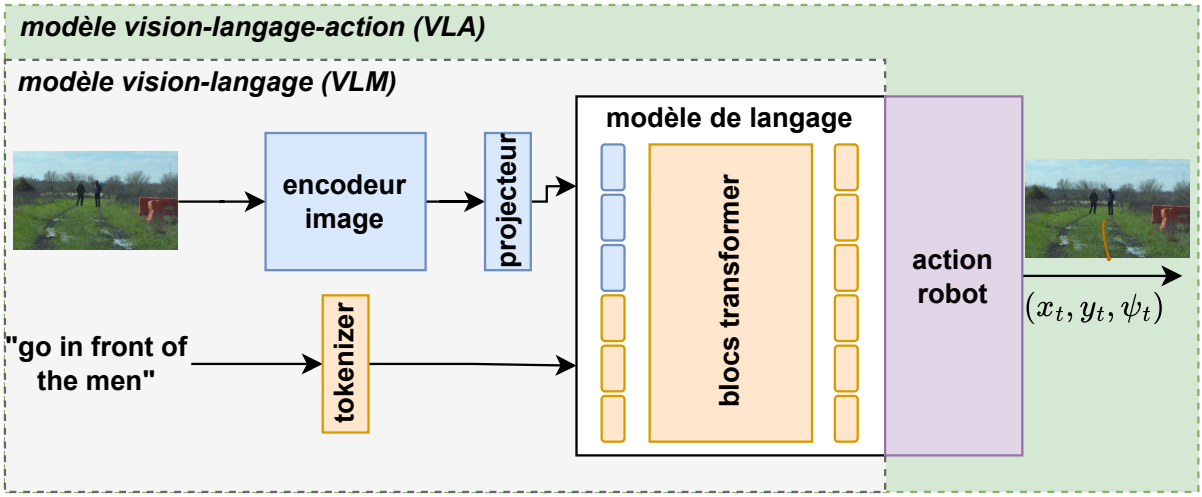

FIGURE 1 – Architecture d'un modèle *vision-langage-action* (VLA).

récents modèles VLM sont développés pour la robotique et notamment sur la capacité d'un modèle à comprendre l'environnement 3D, soit au delà des images RGB [18, 10, 4]. Pour atteindre une précision élevée en navigation dans un espace à trois dimensions, une compréhension fine de la scène 3D est indispensable. Peu d'études abordent cette problématique, mais [21] étend LLaVA pour intégrer au sein d'un LLM des informations 3D (potentiellement issues d'un LiDAR). Concrètement, des patchs 3D sont générés en associant aux patchs 2D extraits par un encodeur CLIP leurs coordonnées spatiales 3D, puis ces jetons sont injectés dans le LLM grâce à des embeddings positionnels dédiés. Par ailleurs, plusieurs travaux proposent d'entraîner des encodeurs à la manière de CLIP pour projeter les nuages de points dans le même espace de caractéristiques que les images et le texte [7, 13, 15]. Si une correspondance rigoureuse est établie entre les modalités d'entrée, les caractéristiques extraites pourraient être utilisées comme entrée d'un LLM sans avoir à l'adapter à la nouvelle modalité.

2.2 Génération d'actions

Les actions du robot peuvent être produites de deux manières différentes : 1/ par décodage [3] des jetons produits par le LLM (Figure 2a) ou 2/ par l'utilisation d'un modèle de diffusion [2] qui apporte des propriétés utiles pour la robotique (Figure 2b).

Le modèle RT-2 [3] est un VLA qui remplace les jetons existants du LLM par des jetons d'action avec un décodage. Par exemple, la réponse "128 91 241 5 101 127" est décodée en action à six degrés de libertés correspondant à la translation et la rotation du système. L'application visée par ce modèle est le contrôle d'un bras robotique pour effectuer des tâches de manipulation. Ce modèle se base sur deux VLM [5, 6] qui sont ré-entraînés à des fins d'action robotique ce qui permet de transférer les propriétés de généralisation au contrôle de différents robots.

L'autre technique vise à utiliser un VLM pré-entraîné et y associer un modèle de diffusion. Cela permet de générer des actions à haute fréquence (jusqu'à 50Hz sur NVIDIA GeForce RTX 4090 [2]) et de modéliser une distribution d'actions possibles [1, 8, 11].

Une des principales bases de données utilisées pour ré-entraîner un VLM à l'action robotique est *Open X-Embodiment* (OXE) [14] qui contient plus d'un million de trajectoires acquises par 22 robots différents. L'entraînement d'un modèle VLA, comme RT-2 avec ce jeu de données permet de généraliser les tâches à réaliser à différents robots. Cette analyse montre également que l'utilisation de données qui couvrent un maximum de scénarios est cruciale pour la généralisation à plusieurs tâches et plusieurs robots.

3 Opportunités pour la navigation tout-terrain

L'utilisation de modèles VLA pour la navigation robotique ouvre de nouvelles possibilités pour la compréhension de l'environnement et l'interaction homme-robot par le biais du langage naturel.

Ce dernier apporte plusieurs intérêts pour la robotique : 1/ la possibilité de donner des instructions spécifiques au robot pour faciliter la coordination et pouvoir le guider plus précisément dans le cas d'exploration de zones. 2/ La possibilité de donner du contexte sur l'environnement ambiant, que ce soit pour s'assurer d'éviter certains obstacles ou pour définir un but lointain et global qui doit être atteint grâce à des points d'intérêts intermédiaires.

La compréhension de l'environnement se fait grâce à l'utilisation de VLM pré-entraînés sur une large quantité de données. Utiliser ces modèles en libre accès est bénéfique en y associant le module d'action qui est contraint par les jetons du grand modèle de langage. Un des axes important qui permet d'améliorer la compréhension de la scène est de travailler au delà des images RGB en utilisant les données de profondeur, comme le LiDAR dont l'utilisation reste limitée avec les modèles de fondation.

A partir des modèles VLA, plusieurs limitations ont été identifiées pour la navigation. Les modèles sont généralement utilisés pour la manipulation d'objets à partir d'un

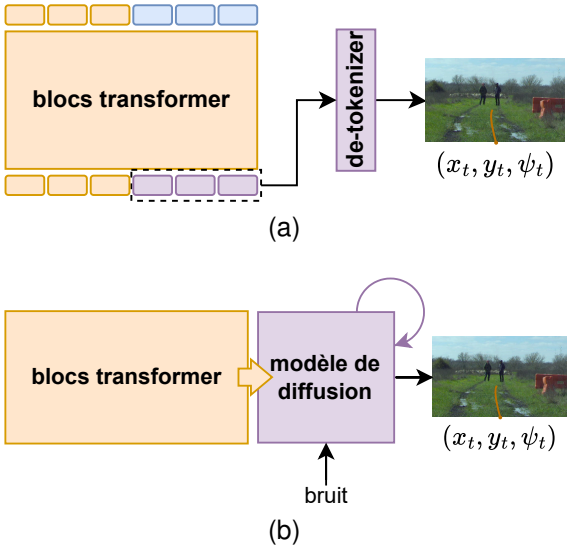

FIGURE 2 – Génération d'actions avec (a) le décodage des jetons et (b) l'utilisation d'un modèle de diffusion.

bras robotique, grâce à l'utilisation d'un jeu données dédié à ce cas d'usage. Un des besoins est donc de générer des données pour la navigation à partir de différents robots avec de nombreuses trajectoires sur plusieurs terrain, que ce soit en extérieur et en intérieur et de longueurs différentes. L'objectif est de pouvoir entraîner un modèle robuste et généralisable à différentes plates-formes.

Un autre point important concerne l'embarquabilité du modèle sur du matériel restreint en calcul et en mémoire. Les modèles qui sont analysées dans ce résumé font des centaines de millions, voire des dizaines de milliards de paramètres, ce qui rend la complexité calculatoire élevée. Un des axes à approfondir est alors de diminuer cette complexité en calcul et en mémoire pour avoir une exécution temps-réel sur plates-formes embarquées pouvant être supportée par le robot tout en assurant une précision et une robustesse de navigation élevée.

4 Conclusion

Le résumé présente une analyse des modèles vision-langage-action appliqués à la robotique. L'architecture générique de ce type de modèle est décrite, de la compréhension de l'environnement avec l'utilisation des VLM à la génération d'action qui étend ces modèles à des VLA.

Une des propriétés intéressantes des modèles VLA est la capacité de généraliser à différentes plates-formes robotiques ainsi qu'à différentes tâches. Plusieurs axes de recherche ont été identifiés pour le cas de la navigation robotique dont l'utilisation de plusieurs modalités incluant le LiDAR et le texte pour améliorer la perception. Ainsi que la génération d'un jeu de données conséquent dédié à la navigation tout-terrain mettant en situation plusieurs plates-formes robotiques dans différents environnements.

Références

[1] Florent Bartoccioni, Elias Ramzi, Victor Besnier, Shashanka Venkataramanan, Tuan-Hung Vu, Yihong Xu, Loick Chambon, Spyros Gidaris, Serkan Odabas, David Hurych, Renaud Marlet, Alexandre Boulch, Mickael Chen, Eloi Zablocki, Andrei Bursuc, Eduardo Valle, and Matthieu Cord. Vavim and vavam : Autonomous driving through video generative modeling. *arXiv preprint arXiv :2502.15672*, 2025.

[2] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0 : A vision-language-action flow model for general robot control, 2024.

[3] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alex Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2 : Vision-language-action models transfer web knowledge to robotic control. In *arXiv preprint arXiv :2307.15818*, 2023.

[4] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm : Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024.

[5] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, AJ Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Peter Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby,

and Radu Soricut. Pali-x : On scaling up a multilingual vision and language model, 2023.

[6] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e : an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[7] Georg Hess, Adam Tonderski, Christoffer Petersson, Kalle Åström, and Lennart Svensson. Lidarclip or : How i learned to talk to point clouds. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

[8] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla : An open-source vision-language-action model. *arXiv preprint arXiv :2406.09246*, 2024.

[9] Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision-language-action models for embodied ai, 2025.

[10] Soroush Nasiriany, Fei Xia, Wenhao Yu, Ted Xiao, Jacky Liang, Ishita Dasgupta, Annie Xie, Danny Driess, Ayzaan Wahid, Zhuo Xu, Quan Vuong, Tingnan Zhang, Tsang-Wei Edward Lee, Kuang-Huei Lee, Peng Xu, Sean Kirmani, Yuke Zhu, Andy Zeng, Karol Hausman, Nicolas Heess, Chelsea Finn, Sergey Levine, and Brian Ichter. Pivot : iterative visual prompting elicits actionable knowledge for vlms. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[11] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Charles Xu, Jianlan Luo, Tobias Kreiman, You Liang Tan, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo : An open-source generalist robot policy. In *Proceedings of Robotics : Science and Systems*, Delft, Netherlands, 2024.

[12] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2 : Learning robust visual features without supervision. *arXiv preprint arXiv :2304.07193*, 2023.

[13] Aljoša Ošep, Tim Meinhardt, Francesco Ferroni, Neehar Peri, Deva Ramanan, and Laura Leal-Taixé. Better call sal : Towards learning tonbsp ;segment anything innbsp ;lidar. In *Computer Vision – ECCV 2024 : 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXXIX*, page 71–90, Berlin, Heidelberg, 2024. Springer-Verlag.

[14] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment : Robotic learning datasets and rt-x models : Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.

[15] Gilles Puy, Spyros Gidaris, Alexandre Boulch, Oriane Simeoni, Corentin Sautier, Patrick Perez, Andrei Bursuc, and Renaud Marlet. Three Pillars Improving Vision Foundation Model Distillation for Lidar . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21519–21529, Los Alamitos, CA, USA, June 2024. IEEE Computer Society.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[17] Ahmed Sharshar, Latif U Khan, Waseem Ullah, and Mohsen Guizani. Vision-language models for edge networks : A comprehensive survey. *arXiv preprint arXiv :2502.07855*, 2025.

[18] Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint : A vision-language model for spatial affordance prediction in robotics. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 4005–4020. PMLR, 06–09 Nov 2025.

[19] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[20] Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. LLaVA-mini : Efficient image and video large multimodal models with one vision token. In *The Thirteenth International Conference on Learning Representations*, 2025.

[21] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d : A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv :2409.18125*, 2024.