Proceedings of the 7th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2025.AST27134

Integrating Motion Planning in Vision Language Action Agents

Jianfeng Pang

School of Mathematical and Computational Sciences, Massey University, Aukland, New Zealand
Jianfeng.Pang.1@uni.massey.ac.nz

Abstract. Vision-Language-Action (VLA) models integrate visual perception, natural language understanding, and embodied control into a unified framework, enabling end-to-end task execution from multimodal instructions. While such models have demonstrated impressive generalization across tasks and environments, their direct outputs—often in the form of discrete action tokens or waypoint sequences—frequently overlook key physical constraints, such as trajectory feasibility, collision avoidance, and dynamic consistency. This limitation hinders deployment in safety-critical and dynamic real-world settings. Integrating motion planning into VLA systems offers a principled solution, embedding geometric and dynamic constraints into the control pipeline to transform high-level semantic goals into safe, smooth, and executable trajectories. This work examines representative integration strategies alongside the trade-offs between discrete tokenized outputs and continuous control policies. Applications are analyzed highlighting performance gains in generalization, safety, and execution efficiency. A discussion of current challenges—such as the balance between planning speed and precision, and generalization across embodiments—is followed by prospective research directions, including continuous prediction with hierarchical control, low-resource edge deployment, and multi-robot collaborative planning. The study underscores motion planning as a critical enabler for reliable, adaptable, and scalable embodied intelligence.

Keywords: Vision-Language-Action models, motion planning, hierarchical control, continuous prediction, embodied intelligence

1. Introduction

Vision-Language-Action (VLA) models constitute a class of multimodal foundation models that tightly integrate visual perception, natural language comprehension, and embodied action generation into a unified computational architecture [1]. In response to a camera view (or video) of the environment and an associated language instruction, a VLA model encodes these inputs into a shared latent representation and directly outputs low-level robot actions, bypassing modular pipelines and enabling end-to-end task execution. This end-to-end design removes the need for separate modules such as object detection, symbolic planning, or hand-crafted policies, fostering generalization across visual, linguistic, and control domains. Existing VLA models often produce action tokens or waypoint sequences without accounting for trajectory feasibility, collision

DOI: 10.54254/2755-2721/2025.AST27134

avoidance, or dynamic consistency, limiting their reliability in safety-critical settings. Integrating motion planning into the VLA pipeline addresses these gaps by enforcing physical constraints and enabling real-time translation of language instructions into safe, executable trajectories [1]. Emphasis is placed on hierarchical control architectures, which can mitigate cumulative execution errors and enhance responsiveness in long-horizon tasks. This analysis clarifies the functional role of motion planning within VLA systems, compares multiple integration paradigms, and outlines directions for future research aimed at robust, efficient, and generalizable embodied agents.

2. Integrate motion planning into VLA

2.1. Waypoint based planning

MoManipVLA adapts pre-trained fixed base VLA models to mobile manipulation through the generation of end effector waypoints that generalize across diverse environments and tasks. Motion planning objectives are formulated to ensure reachability, trajectory smoothness, and collision avoidance, thus coordinating the mobile base and robotic arm into a unified, physically feasible trajectory. A bi-level optimization framework underlies this system: the upper level predicts base waypoints to expand manipulator workspace, whereas the lower level refines end effector trajectories to accomplish manipulation tasks precisely. Extensive evaluations in both simulated OVMM and real-world setups demonstrate a 4.2 % improvement in success rate over the prior state of the art, alongside a significant reduction in required real world finetuning (only 50 expert episodes), thanks to the generalization power of the underlying VLA models [2]. Moreover, MoManipVLA exhibits platform-agnostic adaptability, supporting diverse robotic embodiments without retraining.

2.2. Closed loop & hierarchical mechanisms

LoHoVLA presents a unified architecture designed to handle long-horizon embodied tasks by leveraging a shared pretrained VLM backbone that jointly generates both sub-task language tokens and robot action tokens. This unified representation facilitates generalization across tasks while enabling simultaneous high-level task decomposition and low-level motion execution. Crucially, a hierarchical closed-loop control mechanism is employed, in which high-level planning stages generate sub-task sequences that inform low-level execution, and execution feedback is then used to dynamically adjust high-level plan parameters. This feedback loop mitigates cumulative errors common in long-horizon tasks and reduces drift between intention and action [3].

2.3. Prediction aware motion planning

In changing environments, where obstacles and conditions shift over time, predictive motion planning is key to staying reliable. VLMPC frameworks handle this by using video or trajectory prediction to look ahead and choose actions that are safer and more efficient. Predicted changes can be turned into geometric constraints—like where obstacles might be—and built into the planning process to avoid collisions. This shift from reacting in the moment to planning improves long-horizon tasks. By factoring in what's likely to happen next, these methods help robots act more smoothly and succeed more often, even in uncertain situations [4-5].

DOI: 10.54254/2755-2721/2025.AST27134

2.4. Tradeoffs: quality vs speed

A good number of VLA models, like RT-2 and OpenVLA for instance, represent actions by means of discrete tokenization. They frame motion generation as an autoregressive task that is similar to language modeling, as seen in references [6-7]. This approach does simplify training as it maps actions to symbolic tokens. However, it also restricts spatial accuracy and smooth temporal control, particularly when it comes to high frequencies, as pointed out in reference [8]. On the other hand, continuous policy outputs, which are utilized in models such as π_0, present an alternative. They directly generate joint trajectories, which can go up to 50 Hz, by using diffusion or flow-matching decoders. This enables motions to be more fluid, yet it also boosts the computational load and latency, as noted in reference [9]. Compact models like TinyVLA strive to strike a balance among these trade-offs. They do this by pairing lightweight multimodal encoders with diffusion-based decoders, thereby achieving faster inference and better data efficiency. Even though they are of a smaller size, they still manage to remain competitive with larger models like OpenVLA when it comes to accuracy, generalization, and responsiveness, as shown in reference [10]. Such designs are highly appropriate for edge deployment and low-resource robotic systems that demand real-time control.

3. Case studies

3.1. Mobile base & manipulator

MoManipVLA adapts fixed-based VLA models for mobile manipulation by generating end-effector waypoints that guide whole-body planning for both the mobile base and robotic arm. Motion planning objectives, including reachability, smoothness, and collision avoidance—are jointly optimized through a bi-level framework, enabling zero-shot generalization to new tasks and environments. Extensive evaluations on the OVMM benchmark and real-world robotic platforms demonstrate a 4.2 % improvement in task success rates over prior mobile manipulation approaches, necessitating only 50 expert trajectories for fine-tuning due to the generalization capability of the pre-trained VLA models. This approach supports deployment across various robot embodiments without additional retraining [2].

3.2. Aerial robotics & humanoid

The UAV VLA system integrates satellite imagery processing with a visual-language model and GPT-based planning to generate flight paths and action sequences for aerial vehicles. Experimental results indicate a 22 % reduction in trajectory length and a mean mapping error of 34.22 m when identifying target objects, while flight plan generation is completed in approximately 5 minutes, making it 6.5 times faster than human operators [11]. This demonstrates the viability of VLA systems for large-scale aerial mission planning.

In humanoid robotics, the RT-H model extends the RT series of VLA architectures to bipedal humanoid platforms, combining multimodal perception with a transformer-based policy network for whole-body manipulation and locomotion [12].

3.3. Long horizon tasks

LoHoVLA introduces a unified architecture for long-horizon embodied tasks, jointly generating sub-task instructions and action tokens from a pretrained vision-language model. Training on the

DOI: 10.54254/2755-2721/2025.AST27134

LoHoSet dataset—20 distinct Ravens simulator tasks, each supported by 1,000 expert demonstrations—LoHoVLA significantly outperformed both flat VLA and hierarchical baselines in unseen tasks. These results underscore the benefit of integrated planning and control within a shared representational framework for improving generalization and execution accuracy on multi-step tasks [3].

3.4. Multi-robot & collaboration

Multi-robot collaboration extends VLA systems to scenarios where high-level tasks must be distributed across heterogeneous agents. By parsing natural language instructions into coordinated sub-tasks, VLA models provide semantic consistency in role allocation, while motion planning ensures spatial safety, temporal synchronization, and dynamic adaptability during execution. Recent systems such as Gemini Robotics demonstrate large-scale task orchestration across manipulators and mobile platforms, and Helix provides distributed optimization for real-time trajectory coordination. These results highlight the potential of integrating motion planning into VLA frameworks to enable robust, scalable multi-robot cooperation in domains such as warehouse automation, aerial swarms, and collaborative assembly [13-14].

4. Discussion and future work

The present endeavors to incorporate motion planning into Vision–Language–Action (VLA) systems encounter a crucial trade-off when it comes to the balance between execution speed and control accuracy. High-frequency planning allows for smooth and responsive motion, yet it frequently ends up increasing the computational load and latency. Conversely, giving priority to efficiency may lead to a reduction in precision and an elevation of safety risks, particularly in the context of human–robot collaboration. There is also another hurdle, which is generalization. This involves adapting a policy from one type of robot or task to another, like shifting from a mobile manipulator to a fixed arm, or transitioning from structured tasks to open-ended navigation.

Future research could potentially tackle these issues through the combination of continuous prediction and hierarchical control, much like what is seen in models such as UP-VLA and DreamVLA. This particular approach enhances adaptability within dynamic environments, all the while maintaining the structure of the task as demonstrated in references [15-16]. Also, there is the option of optimizing for edge deployment and execution in low-resource conditions. In such scenarios, lightweight models like TinyVLA and SmolVLA are capable of enabling real-time planning, and this can be achieved without having to sacrifice either safety or the rates of task completion, as indicated in references [10,17]. Moreover, when motion planning integration is extended to collaborative scenarios involving multiple robots, for example, within frameworks like Helix and Gemini Robotics, it can create opportunities for the coordinated execution of large-scale tasks across heterogeneous robotic teams, as detailed in references [13-14].

5. Conclusion

This piece of work highlights the crucial part that motion planning plays in augmenting the capacity, safety aspect, and the extent of generalization of VLA systems. When geometric as well as dynamic constraints are embedded within the planning that is grounded in high-level language, VLA models are enabled to transform semantic intentions into sequences of actions that are both physically viable and efficient. The incorporation of motion planning serves to not only enhance the reliability of

DOI: 10.54254/2755-2721/2025.AST27134

execution but also widen the scope of applicability of VLA architectures across diverse robotic forms and operational settings. Ongoing progress in the areas of joint optimization, continuous action learning, and autonomous planning is of great necessity for bringing about robust and adaptable embodied agents. The research community is urged to seek a more profound integration between VLA frameworks and sophisticated motion-planning approaches, with the ultimate aim of closing the gap that exists between high-level cognitive reasoning and low-level physical execution when it comes to complex real-world tasks.

References

[1] Sapkota, R., Cao, Y., Roumeliotis, K. I., & Karkee, M. (2025). Vision-Language-Action Models: Concepts, Progress, Applications and Challenges (No. arXiv: 2505.04769).

[2] Wu, Z., Zhou, Y., Xu, X., Wang, Z., & Yan, H. (2025). MoManipVLA: Transferring Vision-language-action Models for General Mobile Manipulation.

[3] Yang, Y., Sun, J., Kou, S., Wang, Y., & Deng, Z. (2025). LoHoVLA: A Unified Vision-Language-Action Model for Long-Horizon Embodied Tasks (No. arXiv: 2506.00411).

[4] Zhao, W., Chen, J., Meng, Z., Mao, D., Song, R., & Zhang, W. (2024, July 15). VLMPC: Vision-Language Model Predictive Control for Robotic Manipulation. Robotics: Science and Systems XX. Robotics: Science and Systems 2024.

[5] Zhang, Z., Hess, G., Hu, J., Dean, E., Svensson, L., & Åkesson, K. (2025). Future-Oriented Navigation: Dynamic Obstacle Avoidance with One-Shot Energy-Based Multimodal Motion Prediction (No. arXiv: 2505.00237).

[6] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., Florence, P., Fu, C., Arenas, M. G., Gopalakrishnan, K., Han, K., Hausman, K., Herzog, A., Hsu, J., Ichter, B., … Zitkovich, B. (2023). RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control (No. arXiv: 2307.15818).

[7] Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., Vuong, Q., Kollar, T., Burchfiel, B., Tedrake, R., Sadigh, D., Levine, S., Liang, P., & Finn, C. (2024). OpenVLA: An Open-Source Vision-Language-Action Model (No. arXiv: 2406.09246).

[8] Pertsch, K., Stachowicz, K., Ichter, B., Driess, D., Nair, S., Vuong, Q., Mees, O., Finn, C., & Levine, S. (2025). FAST: Efficient Action Tokenization for Vision-Language-Action Models (No. arXiv: 2501.09747).

[9] Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., Jakubczak, S., Jones, T., Ke, L., Levine, S., Li-Bell, A., Mothukuri, M., Nair, S., Pertsch, K., Shi, L. X., … Zhilinsky, U. (2024). $π_0$: A Vision-Language-Action Flow Model for General Robot Control (No. arXiv: 2410.24164).

[10] Wen, J., Zhu, Y., Li, J., Zhu, M., Wu, K., Xu, Z., Liu, N., Cheng, R., Shen, C., Peng, Y., Feng, F., & Tang, J. (2025). TinyVLA: Towards Fast, Data-Efficient Vision-Language-Action Models for Robotic Manipulation (No. arXiv: 2409.12514).

[11] Sautenkov, O., Yaqoot, Y., Lykov, A., Mustafa, M. A., Tadevosyan, G., Akhmetkazy, A., Cabrera, M. A., Martynov, M., Karaf, S., & Tsetserukou, D. (2025). UAV-VLA: Vision-Language-Action System for Large Scale Aerial Mission Generation (No. arXiv: 2501.05014).

[12] Belkhale, S., Ding, T., Xiao, T., Sermanet, P., Vuong, Q., Tompson, J., Chebotar, Y., Dwibedi, D., & Sadigh, D. (2024). RT-H: Action Hierarchies Using Language (No. arXiv: 2403.01823).

[13] Mei, Y., Zhuang, Y., Miao, X., Yang, J., Jia, Z., & Vinayak, R. (2025). Helix: Serving Large Language Models over Heterogeneous GPUs and Network via Max-Flow (No. arXiv: 2406.01566).

[14] Team, G. R., Abeyruwan, S., Ainslie, J., Alayrac, J.-B., Arenas, M. G., Armstrong, T., Balakrishna, A., Baruch, R., Bauza, M., Blokzijl, M., Bohez, S., Bousmalis, K., Brohan, A., Buschmann, T., Byravan, A., Cabi, S., Caluwaerts, K., Casarini, F., Chang, O., … Zhou, Y. (2025). Gemini Robotics: Bringing AI into the Physical World (No. arXiv: 2503.20020).

[15] Zhang, J., Guo, Y., Hu, Y., Chen, X., Zhu, X., & Chen, J. (2025). UP-VLA: A Unified Understanding and Prediction Model for Embodied Agent (No. arXiv: 2501.18867).

[16] Zhang, W., Liu, H., Qi, Z., Wang, Y., Yu, X., Zhang, J., Dong, R., He, J., Wang, H., Zhang, Z., Yi, L., Zeng, W., & Jin, X. (2025). DreamVLA: A Vision-Language-Action Model Dreamed with Comprehensive World Knowledge (No. arXiv: 2507.04447).

DOI: 10.54254/2755-2721/2025.AST27134

[17] Shukor, M., Aubakirova, D., Capuano, F., Kooijmans, P., Palma, S., Zouitine, A., Aractingi, M., Pascal, C., Russi, M., Marafioti, A., Alibert, S., Cord, M., Wolf, T., & Cadene, R. (2025). SmolVLA: A Vision-Language-Action Model for Affordable and Efficient Robotics (No. arXiv: 2506.01844).