GS-LOMA: Gaussian-Splatting-Guided Long-horizon Legged Loco-Manipulation

Rui Jin

College of Control Science and Engineering, Zhejiang University, China
bbbbigrui@gmail.com

1 Summary of Research

Legged manipulators, resembling human loco-manipulation capabilities, have attracted significant attention. Following exciting advancements in motion controllers and manipulation performance, recent researches have leveraged large language models to enhance the autonomy of legged manipulators for long-horizon tasks. However, these approaches have some limitations in broad manipulation capabilities, active perception and understanding capabilities, and safety assurance.

To address these challenges, our proposal introduces GS-LOMA, a gaussian-splatting-guided long-horizon legged loco-manipulation system. It focuses on harnessing quadrupedal manipulators to accomplish long-horizon loco-manipulation tasks, extending the 3D Gaussian Splatting (3DGS) feature field to legged manipulators to improve traversability and understanding of the physical world, and constructing a neural field to predict collision probabilities for continuous obstacle avoidance.

2 Introduction

Legged manipulators, combining a legged robot and a robotic arm into a unified system, have garnered widespread attention due to their resemblance to human capabilities. Researchers have made exciting strides in developing legged manipulators' motion controllers [1–3] to enhance traversal ability and manipulation performance [4, 5] to facilitate physical interaction with the world at challenging indoor and outdoor terrains where wheeled manipulators may find inaccessible. Although several significant advancements have been made for legged manipulators, enhancing their autonomy to carry out long-term tasks remains a problem.

Research [6, 7] employs the large language model (LLM) to address the challenges in long-horizon loco-manipulation for quadrupeds. However, these works lack many of the required characteristics, as outlined below: (1) Restricted manipulation capabilities: Leveraging the basic quadruped robot platform only supports a limited set of low-level skills; (2) Lacking active perception and understanding capabilities: Obtaining real-world parameters with language descriptions and AprilTags ignores the scenario of unseen environments; (3) Neglecting

safety assurance: Relying solely on an RL-based policy overlooks the safety of the robot.

To tackle these problems, this proposal aims to propose **GS-LOMA**, a gaussian-splatting-guided long-horizon legged loco-manipulation system. (1) Utilize the quadrupedal robot platform attached with a robotic arm to enhance manipulation capabilities for complex tasks. (2) Extend the 3D gaussian splitting [8] (3DGS) feature field [9, 10] to legged manipulators by designing interfaces between LLMs and the feature map, utilizing semantic information from the map to improve the robot's traversability and understanding of the physical world. (3) Construct a neural field to predict the collision probability between the swept volume trajectory of legged manipulators and the 3DGS over a certain period of time, aiming to achieve continuous obstacle avoidance for robot safety.

3 Objectives and Hypotheses to Be Test

1. We aim to utilize the legged manipulator platform to enhance the robot's manipulation capabilities for long-horizon tasks.
2. We aim to use LLMs to endow the robot with reasoning ability, allowing it to decompose long-range tasks into step-by-step plans in natural language.
3. We aim to apply a unified policy for whole-body control of a legged manipulator using reinforcement learning to overcome modeling challenges and enhance the coordination between the robotic arm and legs.
4. We aim to design interfaces connecting LLMs and the 3DGS feature field to obtain positions of target objects and identify the optimal interaction points on these objects.
5. We aim to deploy the 3DGS feature field with semantic information onto legged manipulators to enhance their traversability and understanding capability of the physical world.
6. We aim to propose a neural field to predict collision probabilities between the robot swept volume trajectory and 3DGS for continuous obstacle avoidance.

4 Literature Review

4.1 Long-horizon Locomotion and Manipulation.

Task and Motion Planning (TAMP) methods are commonly used to make plans for long-horizon tasks, which decomposes the planning process of long-term tasks into discrete symbolic states and continuous motion generation [11–13]. TAMP relies on manually specified symbolic rules, thereby requiring known physical states with high dimensional search space in complex tasks and resulting in heavy computational burden [14]. Recently, learning-based methods have been integrated into TAMP framework to accelerate feasible plans [15, 16].

LLMs, such as GPT-4 [17], Gemini [18], LLaMA [19], demonstrate strong understanding and ability to solve complex tasks. For robotics, many works leverage LLMs to enable robots to understand and solve tasks from language descriptions.

Additionally, researchers also consider using Python APIs completed by LLMs to connect robot commands [20, 21].

The most relevant work to this proposal is [6] which employs LLMs to achieve long-horizon locomotion and manipulation on a quadrupedal robot. However, [6] only supports a set of low-level skills relying on the basic quadruped robot platform and predefined AprilTags, lacking active perception and understanding of the physical world. This proposal aims to utilize a legged manipulator as a platform and leverage a GS-based 3D feature field to enhance the robot's understanding capabilities, enabling it to accomplish more complex tasks.

4.2 3D Feature Field For Manipulation.

In long-horizon tasks, both semantics and geometry are equally crucial, as the robot needs to determine the positions of target objects and identify the optimal interaction points on these objects.

Recent works have focused on 3D representations to provide precise 3D positions for robotic manipulation. It is straightforward to fuse the 2D semantics extracted by 2D visual models [22–24] into 3D points or volume. These methods result in inconsistencies in 3D semantics due to the different semantic information from multiple perspectives, conflicting with the motion requirements of mobile manipulation. Researches [25, 26] employ implicit representation to extract the feature field by reconstructing 3D features from 2D images, suffering from high computational costs and an inability to quick updates, making them unsuitable for tasks involving frequent interaction with the environment.

3D Gaussian Splatting (3DGS) [8], which recently emerged as a transformative technique in the explicit radiance field, employing a set of 3D Gaussian primitives to model a scene. Some researchers [10, 27] inject semantic information into the 3DGS map to construct the 3D feature field. 3DGS's explicit and volumetric representation allows for local updates of the constructed field, and its highly parallel "splatting" rasterization enables high frame rate rendering for fast language queries. However, these works are currently limited to robotic arm manipulation. This proposal plans to extend it to mobile manipulators by designing interaction interfaces between LLMs and the 3DGS map, utilizing semantic information from the map to enhance the robot's traversability and understanding of the physical world.

4.3 Swept Volume for Collision Detection.

Geometric representations and computations are pivotal in robotics [28], particularly in whole-body planning. While most research concentrates on employing convex geometric shapes like ellipsoids, polyhedrons, or cylinders to model configuration space or robots for a trade-off between exact representation and collision query efficiency [29–33].

Swept volume [34,35] refers to the 3D space occupied by an object as it moves through its entire range of motion, which has been used for collision detection during motion planning [36–38]. Collision detection based on swept volume can

prevent the issue of missing obstacles caused by collision checking at discrete time instances along a given trajectory [29–31]. Most existing methods cannot achieve real-time computation on a high degree of freedom robots as there is no feasible way to explicitly construct an approximate SDF representing the swept volume for the general case. To tackle this problem, RDF [38], a neural implicit representation for robot safety, is constructed using supervised learning in a tractable fashion to predict the distance between the swept volume of a robot arm and an obstacle. However, this method approximates the swept volume using the convex hull of each link's forward occupancy leading to a loss of geometric accuracy.

[39] derives the closest point loss, a novel regularizer that encourages the network output to be an exact SDF, and utilizes bound loss to obviate the need for exact SDF. Inspired by RDF, this proposal aims to integrate closest point loss and bound loss to construct a neural field to predict the collision probability between the swept volume trajectory of legged manipulators and the 3DGS over a certain period of time without relying on convex hull geometric approximation.

5 Approach

GS-LOMA is planned to consist of three modules including a reasoning & control module (RCM), a 3DGS map module (GSM), and a neural collision detection module (NCM). RCM is planned to utilize a cascade of LLMs for reasoning and RL-based strategies for control implementation. GSM constructs a feature field based on 3DGS. By defining interfaces to facilitate information exchange between RCM and GSM, the agent is enabled to acquire real-world parameters and update the 3DGS map rapidly. NCM aims to predict the robot's collision probability with the 3DGS map to ensure safety.

5.1 RCM: Reasoning & Control Module

To equip the agent with the reasoning and locomotion capabilities required for long-horizon tasks, this module is designed to incorporate high-level LLMs-based reasoning and low-level RL-based control policies.

High-level LLMs-based Reasoning. The high-level LLMs-based reasoning plans to employ a cascade of LLMs [6], enabling the robot to reason from pictures and language descriptions of long-horizon tasks and generate APIs for locomotion and manipulation. Through language queries in the 3DGS feature map to enable the robot to grasp real-world parameters accurately.

Low-level RL-based Control. Modeling a legged robot with an attached arm involves addressing dynamic, high-degree-of-freedom (DoF), and non-smooth control challenges. Utilize an RL-based control policy [4] to effectively alleviate this burden. This proposal plans to use a neural network as a unified policy, where inputs include the robot states, end-effector position and orientation commands, and base velocity commands. Outputs consist of target arm joint angles and target leg joint angles.

5.2 GSM: 3DGS Map Module

Exhibiting strong 3D semantic consistency, the 3DGS feature field can provide precise interactive poses through language queries, aiding the robot in understanding the real world. This proposal plans to divide GSM into three parts: reconstruction, query, and update.

Reconstruction of 3DGS Feature Field. Acquire images and observation poses from RGB-D cameras and locating module. Leverage contrastive learning to efficiently distill CLIP [23] features and augment feature fields with SAM [22] segmentation prior.

Queries for Feasible Interaction Pose. Firstly, employ open-vocabulary queries to locate the target object. Secondly, render the depth and normals of 3DGS primitives to obtain the object's detailed geometry. Lastly, use point cloud-based grasping module to generate grasp poses and filter out unfeasible one with the rendered normals.

Rapid Update of 3DGS Map. By employing an object-aware initialization and update strategy, 3DGS with the same semantic labels can quickly update together in response to movements of the robot and objects.

5.3 NCM: Neural Collision Detection Module

Design a DNN-based algorithm to predict the probability between robot swept-volume trajectories and 3DGS . Firstly, compress the swept volume into DNN. The inputs consist of the states of current base state (row, pitch, and base angular velocities), arm state (joint position and velocity of each arm joint), leg state (joint position and velocity of each arm joint), as well as the mean of the 3DGS. The network will output the SDF value between the robot and the 3DGS map over a certain period of time. Secondly, assuming the probability of terminating a light ray provides a strong indication of the probability of terminating a mass particle, calculate the collision probability between each 3DGS and the robot over a certain period of time according to the predicted SDF value and the 3DGS's scale.

The 3DGSs near the trajectory selected by the Axis-aligned Bounding Box (AABB) algorithm will be used for collision detection. Once the sum of selected 3DGS elements' collision probabilities exceeds the safety threshold, an emergency stop signal will be immediately sent, and replanning will be requested.

6 Anticipated outcome and value of the research

1. Enhanced autonomy: The system is expected to significantly improve the autonomy of legged manipulators, enabling them to perform long-term tasks with greater efficiency and independence.
2. Improved traversability and Understanding: The extension of the 3D Gaussian Splatting (3DGS) feature field to legged manipulators is expected to enhance their ability to navigate and understand the physical world through semantic information.

3. Continuous obstacle avoidance: The development of a neural field to predict collision probabilities is expected to contribute to the continuous avoidance of obstacles, ensuring the safety of the robot during operation.
4. Unified policy for whole-Body Control: The application of reinforcement learning for whole-body control is anticipated to overcome modeling challenges and improve the coordination between the robotic arm and legs.
5. Integration of perception and reasoning: The system aims to integrate active perception and understanding capabilities, allowing the robot to reason and plan tasks more effectively using natural language.

References

1. Q. L. J. G. L. C. J. P. Junfeng Long, Zirui Wang, "Hybrid internal model: Learning agile legged locomotion with simulated robot response," in *International Conference on Learning Representations (ICLR)*, 2024.
2. D. Hoeller, N. Rudin, D. Sako, and M. Hutter, "Anymal parkour: Learning agile navigation for quadrupedal robots," *Science Robotics*, vol. 9, no. 88, p. eadi7566, 2024.
3. P. Fankhauser, M. Bjelonic, C. D. Bellicoso, T. Miki, and M. Hutter, "Robust rough-terrain locomotion with a quadrupedal robot," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 5761–5768.
4. Z. Fu, X. Cheng, and D. Pathak, "Deep whole-body control: Learning a unified policy for manipulation and locomotion," in *Conference on Robot Learning*. PMLR, 2023, pp. 138–149.
5. J.-P. Sleiman, F. Farshidian, and M. Hutter, "Versatile multicontact planning and control for legged loco-manipulation," *Science Robotics*, vol. 8, no. 81, p. eadg5014, 2023.
6. Y. Ouyang, J. Li, Y. Li, Z. Li, C. Yu, K. Sreenath, and Y. Wu, "Long-horizon locomotion and manipulation on a quadrupedal robot with large language models," 2024.
7. M. Xu, P. Huang, W. Yu, S. Liu, X. Zhang, Y. Niu, T. Zhang, F. Xia, J. Tan, and D. Zhao, "Creative robot tool use with large language models," *arXiv preprint arXiv:2310.13065*, 2023.
8. B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 1–14, 2023.
9. Y. Zheng, X. Chen, Y. Zheng, S. Gu, R. Yang, B. Jin, P. Li, C. Zhong, Z. Wang, L. Liu, C. Yang, D. Wang, Z. Chen, X. Long, and M. Wang, "Gaussiangrasper: 3d language gaussian splatting for open-vocabulary robotic grasping," 2024.
10. G. Lu, S. Zhang, Z. Wang, C. Liu, J. Lu, and Y. Tang, "Manigaussian: Dynamic gaussian splatting for multi-task robotic manipulation," *arXiv preprint arXiv:2403.08321*, 2024.
11. L. P. Kaelbling and T. Lozano-Pérez, "Hierarchical task and motion planning in the now," in *2011 IEEE International Conference on Robotics and Automation*. IEEE, 2011, pp. 1470–1477.
12. T. Migimatsu and J. Bohg, "Object-centric task and motion planning in dynamic environments," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 844–851, 2020.

13. M. Toussaint, "Logic-geometric programming: An optimization-based approach to combined task and motion planning." in *IJCAI*, 2015, pp. 1930–1936.
14. V. N. Hartmann, A. Orthey, D. Driess, O. S. Oguz, and M. Toussaint, "Long-horizon multi-robot rearrangement planning for construction assembly," *IEEE Transactions on Robotics*, vol. 39, no. 1, pp. 239–252, 2022.
15. S. Nair and C. Finn, "Hierarchical foresight: Self-supervised learning of long-horizon tasks via visual subgoal generation," *arXiv preprint arXiv:1909.05829*, 2019.
16. K. Pertsch, O. Rybkin, F. Ebert, S. Zhou, D. Jayaraman, C. Finn, and S. Levine, "Long-horizon visual planning with goal-conditioned hierarchical predictors," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 321–17 333, 2020.
17. OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, at el, W. Zhuk, and B. Zoph, "Gpt-4 technical report," 2024.
18. G. Team, R. Anil, S. Borgeaud, at el, J. Dean, and O. Vinyals, "Gemini: A family of highly capable multimodal models," 2024.
19. H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023.
20. A. Zeng, M. Attarian, B. Ichter, K. Choromanski, A. Wong, S. Welker, F. Tombari, A. Purohit, M. Ryoo, V. Sindhwani, *et al.*, "Socratic models: Composing zero-shot multimodal reasoning with language," *arXiv preprint arXiv:2204.00598*, 2022.
21. J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9493–9500.
22. A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.
23. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
24. M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.
25. S. Kobayashi, E. Matsumoto, and V. Sitzmann, "Decomposing nerf for editing via feature field distillation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 311–23 330, 2022.
26. V. Tschernezki, I. Laina, D. Larlus, and A. Vedaldi, "Neural feature fusion fields: 3d distillation of self-supervised 2d image representations," in *2022 International Conference on 3D Vision (3DV)*. IEEE, 2022, pp. 443–453.
27. Y. Li and D. Pathak, "Object-aware gaussian splatting for robotic manipulation," in *ICRA 2024 Workshop on 3D Visual Representations for Robot Manipulation*.
28. C. D. Toth, J. O'Rourke, and J. E. Goodman, *Handbook of discrete and computational geometry*. CRC press, 2017.
29. Z. Han, Z. Wang, N. Pan, Y. Lin, C. Xu, and F. Gao, "Fast-racing: An open-source strong baseline for SE(3) planning in autonomous drone racing," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 8631–8638, 2021.

30. W. Ding, L. Zhang, J. Chen, and S. Shen, "Safe trajectory generation for complex urban environments using spatio-temporal semantic corridor," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2997–3004, 2019.
31. M. Zhang, C. Xu, F. Gao, and Y. Cao, "Trajectory optimization for 3d shape-changing robots with differential mobile base," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 10 104–10 110.
32. N. Harrison, W. Liu, I. Jang, J. Carrasco, G. Herrmann, and N. Sykes, "A comparative study for obstacle avoidance inverse kinematics: Null-space based vs. optimisation-based," in *Towards Autonomous Robotic Systems: 21st Annual Conference, TAROS 2020, Nottingham, UK, September 16, 2020, Proceedings 21*. Springer, 2020, pp. 147–158.
33. H. Sugiura, M. Gienger, H. Janssen, and C. Goerick, "Real-time collision avoidance with whole body motion control for humanoid robots," in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2007, pp. 2053–2058.
34. K. Abdel-Malek, J. Yang, D. Blackmore, and K. Joy, "Swept volumes: fundation, perspectives, and applications," *International Journal of Shape Modeling*, vol. 12, no. 01, pp. 87–127, 2006.
35. D. Blackmore, M. C. Leu, and F. Shih, "Analysis and modelling of deformed swept volumes," *Computer-Aided Design*, vol. 26, no. 4, pp. 315–326, 1994.
36. T. Lozano-Perez, *Spatial planning: A configuration space approach*. Springer, 1990.
37. T. Zhang, J. Wang, C. Xu, A. Gao, and F. Gao, "Continuous implicit sdf based any-shape robot trajectory optimization," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 282–289.
38. J. Michaux, Q. Chen, Y. Kwon, and R. Vasudevan, "Reachability-based trajectory design with neural implicit safety constraints," *arXiv preprint arXiv:2302.07352*, 2023.
39. Z. Marschner, S. Sellán, H.-T. D. Liu, and A. Jacobson, "Constructive solid geometry on neural signed distance fields," in *SIGGRAPH Asia 2023 Conference Papers*, 2023, pp. 1–12.