Physically Consistent Humanoid Loco-Manipulation using Latent Diffusion Models

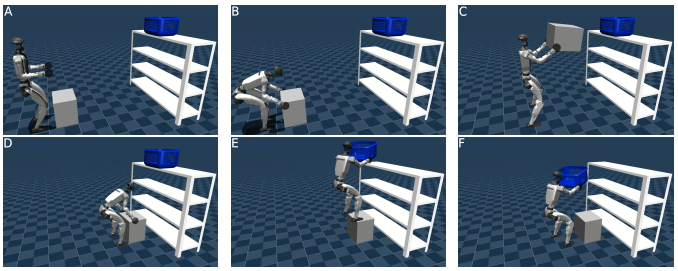


Fig. 1: A long-horizon loco-manipulation task generated with our proposed method; the robot moves a box to enable reaching for a laundry basket.

***Abstract*—This paper uses the capabilities of latent diffusion models (LDMs) to generate realistic RGB human-object interaction scenes to guide humanoid loco-manipulation planning. To do so, we extract from the generated images both the contact locations and robot configurations that are then used inside a whole-body trajectory optimization (TO) formulation to generate physically consistent trajectories for humanoids. We validate our full pipeline in simulation for different long-horizon loco-manipulation scenarios and perform an extensive analysis of the proposed contact and robot configuration extraction pipeline.**

I. Introduction

It has been long argued that humanoids are the best platform to replace humans in repetitive and dangerous tasks, because of the similarities in their morphologies. However, the complexity of these platforms poses significant challenges that have hindered the progress and we still do not see humanoid robots reliably doing real-world tasks. In particular, humanoids are high-dimensional systems with highly unstable dynamics, and performing any reasonable loco-manipulation task requires long-horizon reasoning that existing methods cannot scale to. The similarity between the human and humanoid morphologies can come to rescue, as the robot can imitate the behavior of humans doing the same task. Thanks to the recent advances in generative models, it is nowadays possible to generate a desired human behavior from text prompts. While the outputs of these models do not respect the geometrical and physical constraints of the real world, they can guide existing optimization frameworks to find physically consistent motions quickly.

In this paper, we develop a framework to rapidly synthesize plausible 3D human-object interaction scenes using latent diffusion models (LDMs) [1] for 2D image generation, without the need for ad hoc heuristics or 3D richly annotated data, and use the retargeted motion inside a whole-body trajectory optimization (TO) formulation to generate physically consistent motions for complex long-horizon tasks. The main contributions of this work are as follows:

- We introduce, to the best of our knowledge, the first pipeline that plans both contacts and robot configurations for humanoid loco-manipulation using LDMs.
- We integrate our proposed planner within a whole-body TO formulation to generate physically consistent trajectories.
- We validate our approach in simulation on two challenging long-horizon scenarios, with an extensive analysis.

II. Related Work

Classical approaches for planning and control of loco-manipulation for humanoids consider the effect of manipulated objects on the locomotion system as a disturbance [2]–[5]. To reduce the complexity of the holistic planning problem, more advanced approaches relied on splitting the system into coupled dynamical subsystems [6], separating locomotion and manipulation zones [7], splitting object and locomotion planning [8], or using predefined contact sequences [9]. [10] used a hierarchy of optimal controllers, augmenting the locomotion problem with logic predicates for manipulation [11], but demonstrated only quadrupedal loco-manipulation with a single arm, which is simpler than a humanoid with two arms. More recent efforts use Deep Reinforcement Learning (DRL) for loco-manipulation in the real world [12]–[15], but again only for simple quadrupedal manipulation, and cannot reason about complex, long-horizon humanoid tasks. Imitation learning from teleoperation demonstrations [16], [17] is another option. However, generating such demonstrations for humanoids is extremely difficult [18], as the system is highly unstable. [19] used TO to generate demonstrations that are then imitated using DRL. However, TO is a local approach and fails to generate long-horizon trajectories that require reasoning.

III. Method

In this section, we present our approach to plan contacts and robot configurations to guide a TO procedure for arbitrarily long-horizon humanoid loco-manipulation tasks. Our approach does not rely on task-specific heuristics or 3D interaction datasets; instead, we propose a pipeline that leverages LDMs to generate realistic human-object interaction 2D scenes given a high-level description of the desired interactions. These 2D RGB scenes are used to extract the contact locations and robot configurations that are later used by TO (Sec. IV). The pipeline is illustrated in Fig. 2.

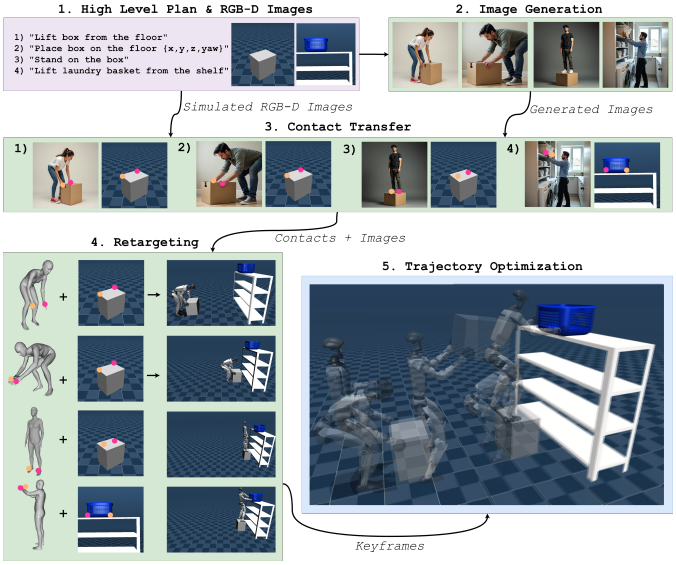


Fig. 2: Pipeline overview.

A. Planning Contacts & Robot Configurations

The planner receives high-level instructions P (that can come from a language model) and RGB-D images $\{R_s, D_s\}$ of the objects in the scene as input. P is an ordered sequence of text prompts describing how to break down the long-horizon task, and for tasks involving placement we assume to receive the target 3D location and yaw of the object. The output of the planner is a sequence of 3D contact locations L and associated robot configurations C, produced in three main steps.

1) Image Generation: Given P, we use a state-of-the-art LDM [20] to generate a collection of images R_g demonstrating how to accomplish the task. Using the task prompts directly leads to images that do not depict a full-body person, which is essential for our pipeline since we need the vast majority of the human body to be visible to extract hand/feet contacts and the respective robot configuration. We therefore automatically append a static appearance description to each prompt, yielding "*A scene of a person* {*predicate*}+*ing* {*subtask prompt without predicate*}. *The person has dark hair and is wearing casual clothes such a shirt, jeans, and sneakers*".

2) Contact Transfer: The contact transfer stage maps the 2D contact information from the images in R_g to 3D contact locations on the real scene objects. First, we compute the 2D semantic masks of the objects in R_g and R_s. To do so, we use a Vision Language Model (VLM) [21] to perform open-vocabulary object detection, followed by a segmentation foundation model [22] that refines the VLM output into a per-pixel segmentation.

Second, we lift the masked pixels to 3D point clouds using the depth information. For the simulated images R_s we have the ground truth depth D_s from the simulated RGB-D camera in MuJoCo [23] and its correct camera intrinsics. For R_g we are missing both, as LDMs only output RGB images without adhering to a specific camera model. We therefore leverage a zero-shot metric depth foundation model [24] to estimate D_g, and set the intrinsics using an empirical trial and error approach: using the LDMs' image resolution as the focal lengths and half the focal lengths for the principal point offsets leads to a reasonable point cloud geometry, without too much distortion.

Third, we use a sampling-based optimization to transfer the 2D contact locations from R_g to the 3D world. Finding correct semantic correspondences across R_g and R_s is challenging, because we have limited control during image generation over object properties such as viewpoint, shape, and texture, leading to significant intra-class variation between the generated and the simulated objects. We use a semantic-aware foundation model [25] to obtain semantic matches, but depending exclusively on it is not reliable, as the large intra-class variation leads to incorrect mappings that deteriorate the output trajectory (Sec. V-B). We therefore refine these correspondences using the point cloud geometries: the underlying idea is that correct semantic matches should result in a good geometrical overlap. Hence, we generate a pool of the top N candidate correspondences per sampled 2D point on the objects' masks in R_g, and search within this pool. For each candidate set, we solve for the rigid-body transform with SVD and refine it with ICP. We repeat this process for 10 iterations and pick the transform with the highest overlapping score; in practice, we found that within 3 iterations the best transform is already found. Finally, we obtain L by applying the transform to the 3D lifted hand and feet 2D locations from a human pose estimator [26] running on R_g, and project L to the closest object's surface to avoid penetrations.

3) Retargeting: In the retargeting stage, we use the depicted humans in R_g and L to obtain the robot configuration R, a 35D vector describing the robot's 6D base state and joint angles. Since we cannot map the human configuration directly to the robot due to differences in degrees of freedom, limb length, and height, we formulate an Inverse Kinematics (IK) based retargeting to a kinematically feasible robot configuration. We use WHAM [26] on R_g to obtain the human's joint angles, foot positions, and base orientation. For the joints, following [27], we only consider those that have a corresponding match on the robot. Foot positions are taken as the planar distance between pelvis and ankles (foot height is set by the task). The base orientation is obtained by applying the rigid-body transform from Sec. III-A2 to the 3D body model and computing the relative orientation between the pelvis and the simulated object. The final IK uses the hand/feet contacts and feet pitch angle as constraints, with the joint angles acting as a regularizer towards a human-like configuration.

IV. Trajectory Optimization

In this section, we outline our TO formulation and how the extracted contact locations and robot configurations are used within it. We use the centroidal dynamics coupled with whole-body kinematics formulation similar to [28], and the same dynamics is used for the manipulated objects. Contacts are modeled as patch contacts on surfaces perpendicular to gravity, with a linearized friction cone, a zero tangential velocity

constraint, and a patch orientation constraint that aligns the contact with the surface. For moving robot-object contacts, these constraints become relative and Newton's third law is enforced on the contact wrenches. A nonholonomic constraint is also used for the trolley chassis, with pure rolling contacts and zero lateral velocity on the rear wheels.

A. Task-generic Cost

To avoid task-specific tuning, the stagewise cost is $L_{\text{stage}} := b_{\text{st}} L_{\text{st}} + (1 - b_{\text{st}}) L_{\text{wk}} + L_{\text{reg}} + L_{\text{slack}}$, where b_{st} flags stance, $L_{\text{st}}, L_{\text{wk}}$ are the stance and walking phase costs, L_{reg} is a regularizer, and L_{slack} penalizes slack variables. All costs are in quadratic form $w\|\cdot\|_2^2$.

B. Keyframe Cost

The full configuration of the robot and the contact locations from the planning module constitute a *keyframe*, which provides waypoints for a long-horizon task. We add the following cost on the keyframe robot pose:

$$L_{\text{kf}}^{\text{b}} := W_{\text{kf}}^{\text{b}} [(r_{\text{base},z} - r_{\text{base},z}^{\text{kf}})^2, \|\boldsymbol{\Theta} - \boldsymbol{\Theta}^{\text{kf}}\|_2^2]^\top, \tag{1}$$

where $W_{\text{kf}}^{\text{b}} = [100, 10]$ and $(\cdot)^{\text{kf}}$ denotes the keyframe value. Keyframes also indicate a desired relative foot position w.r.t. the object, giving a global reference $\mathbf{r}_f^{\text{des}}$, which we add to the stance phase cost as

$$L_{\text{kf}}^{\text{f}} := W_{\text{kf}}^{\text{f}} \|(r_{lf,xy} + r_{rf,xy})/2 - \mathbf{r}_f^{\text{des}}\|_2^2, \tag{2}$$

with $W_{\text{kf}}^{\text{f}} = 5\text{e}2$. In addition, each subtask keyframe is used as the nominal state to initialize the shooting nodes of that subtask, which greatly accelerates and robustifies convergence.

C. Collision Avoidance and Overall Problem

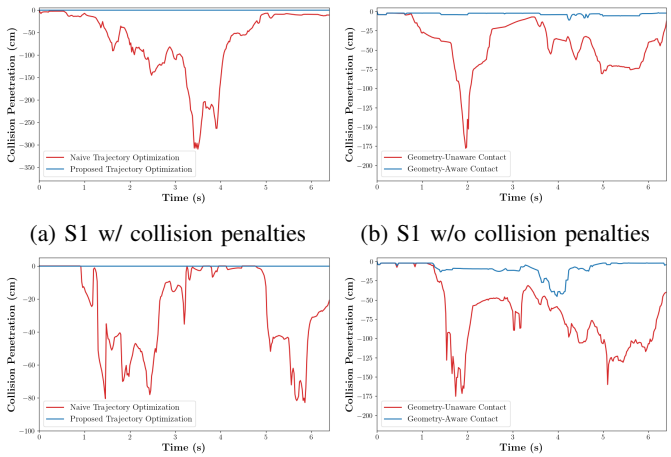
To generate realistic motion, we combine three collision-avoidance techniques: a hard penetration constraint, a non-smooth quadratic penalty on penetrations, and a homotopy-based constraint and penalty that uses a simple smooth surrogate geometry (e.g., an ellipsoid in place of a cube) to guide the solver through non-smooth cases. The overall TO problem is then

$$\begin{aligned} \min. \quad & \frac{1}{N} \sum_{i=0}^{N} \left[L_{\text{stage}}^i + L_{\text{col}}^i \right] + \sum_{j \in \mathcal{K}} L_{\text{kf}}^{\text{b},j} \\ \text{s.t.} \quad & \text{Dynamics, Contacts, Collision,} \end{aligned} \tag{3}$$

where N is the number of shooting nodes, $\mathcal{K}$ is the set of keyframe indices, and L_{col} collects the collision penalties.

V. Experiments

We test our pipeline on two long-horizon scenarios in MuJoCo [23] with the Unitree G1 humanoid. In S1 (basket retrieval), the basket is not easily reachable from the ground, so the robot must move a box close to a shelf and step on top of it to reach the basket. In S2 (trolley), the robot moves a box placed on a table using a trolley and then pushes the trolley. The pipeline takes several minutes per scenario (mainly due to image generation, semantic-aware model inference, and TO), which currently hinders real-time capabilities. Note that without the LDM-derived contacts and configurations, it was impossible for TO to solve the tasks, as both require long-horizon reasoning that is unfeasible for local TO. We refer the reader to the supplementary video for additional qualitative results.



(a) S1 w/ collision penalties (b) S1 w/o collision penalties

(c) S2 w/ collision penalties (d) S2 w/o collision penalties

Fig. 3: Collision penetrations of TO with our pipeline (blue) vs. the naive approach (red) for S1 and S2, with and without collision penalties enabled.

A. Physically Plausible Trajectories

We compare the output of TO when using our full planning pipeline (Sec. III-A) versus a naive baseline that feeds TO only the contacts obtained directly from the semantic-aware foundation model. In both cases, we use a minimal set of collision penalty constraints. Figure 3(a,c) presents the total negative collision penetrations at each timestep. Our pipeline maintains a collision-free behavior throughout the whole trajectory, while the naive approach experiences significant negative penetrations. Enabling all possible collision constraints in TO is not a viable alternative, as TO then fails to solve the task and gets stuck in a local minima.

B. Geometry Improves Contact Transfer

To isolate the effect of our geometry-aware contact refinement (Sec. III-A2), we compare it against a geometry-unaware variant that uses the semantic-aware model output directly. We disable all collision penalties and use the same retargeted robot configuration (Sec. III-A3) in both approaches, measuring self- and robot-object collision penetration. Figure 3(b,d) shows that the geometry-unaware transfer leads to significantly larger penetrations for both scenarios, while our refined contacts leave only small residuals that can be removed with the minimal collision set of Sec. V-A. To remove them otherwise, a geometry-unaware approach would require significantly more collision constraints, which makes the problem extremely non-convex; local TO in such cases gets stuck in local minima and is unable to solve the task.

C. Keyframes Reduce Penetration

We ablate the keyframe components (keyframe base cost (1), keyframe foot position cost (2), and subtask warm-start in Sec. IV-B) against a NoKeyframe baseline. The refined contacts and the same minimal collision set are kept across all settings. Three keyframes are used in each scenario: picking up the box, placing the box, and either grabbing the basket while standing on the box (S1) or the beginning of the pushing (S2); for S2 we also add a cost on the trolley's position to describe the forward pushing goal. We vary the box mass and initial yaw ϕ, which we observed to have a large impact on TO: S1 tests $\{2.5, 5.0\}$ kg and $\phi \in \{0, 0.6, 1.2\}$ rad, and S2 tests $\{2.5, 5.0, 7.5\}$ kg and $\phi \in \{0, \pm 0.6, \pm 1.2\}$ rad. For both scenarios, we observed that using keyframes helps the success rate of TO and leads to lower-penetration solutions. For example, in the $(5.0\,\text{kg}, 0\,\text{rad})$ test of S1, the AllKeyframe case gives zero penetration while removing any part of the keyframe either hinders convergence or results in large penetrations. In particular, settings with the warm-start tend to have better convergence and lower penetration.

VI. Conclusion

We presented a novel approach that generates physically consistent trajectories for long-horizon loco-manipulation tasks by leveraging LDMs to synthesize 2D images demonstrating how a human would accomplish a task, and extracting from them the robot configurations and contact locations that guide a whole-body TO. We evaluated the method in simulation on two challenging scenarios that require long-horizon reasoning. Future work will focus on evaluating the pipeline on a real humanoid robot, developing an LLM-based long-horizon task planner to replace the given high-level plan, and considering human videos instead of 2D images due to the current limited capability of LDMs for complex human-object interactions.

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[2] K. Bouyarmane and A. Kheddar, "Humanoid robot locomotion and manipulation step planning," *Advanced Robotics*, vol. 26, no. 10, pp. 1099–1126, 2012.

[3] L. Penco, N. Scianca, V. Modugno, L. Lanari, G. Oriolo, and S. Ivaldi, "A multimode teleoperation framework for humanoid loco-manipulation: An application for the icub robot," *IEEE Robotics & Automation Magazine*, vol. 26, no. 4, pp. 73–82, 2019.

[4] W. Thibault, F. J. A. Chavez, and K. Mombaur, "A standardized benchmark for humanoid whole-body manipulation," in *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*. IEEE, 2022, pp. 608–615.

[5] J. Li and Q. Nguyen, "Multi-contact mpc for dynamic loco-manipulation on humanoid robots," in *2023 American Control Conference (ACC)*. IEEE, 2023, pp. 1215–1220.

[6] A. Settimi, D. Caporale, P. Kryczka, M. Ferrati, and L. Pallottino, "Motion primitive based random planning for loco-manipulation tasks," in *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2016, pp. 1059–1066.

[7] P. Ferrari, M. Cognetti, and G. Oriolo, "Humanoid whole-body planning for loco-manipulation tasks," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 4741–4746.

[8] M. Murooka, I. Kumagai, M. Morisawa, F. Kanehiro, and A. Kheddar, "Humanoid loco-manipulation planning based on graph search and reachability maps," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1840–1847, 2021.

[9] H. Ferrolho, V. Ivan, W. Merkt, I. Havoutis, and S. Vijayakumar, "Roloma: Robust loco-manipulation for quadruped robots with arms," *Autonomous Robots*, vol. 47, no. 8, pp. 1463–1481, 2023.

[10] 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.

[11] M. A. Toussaint, K. R. Allen, K. A. Smith, and J. B. Tenenbaum, "Differentiable physics and stable modes for tool-use and manipulation planning," *Robotics: Science and Systems Foundation*, 2018.

[12] 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.

[13] S. Jeon, M. Jung, S. Choi, B. Kim, and J. Hwangbo, "Learning whole-body manipulation for quadrupedal robot," *arXiv preprint arXiv:2308.16820*, 2023.

[14] T. Portela, G. B. Margolis, Y. Ji, and P. Agrawal, "Learning force control for legged manipulation," *arXiv preprint arXiv:2405.01402*, 2024.

[15] S. Ha, J. Lee, M. van de Panne, Z. Xie, W. Yu, and M. Khadiv, "Learning-based legged locomotion: State of the art and future perspectives," *The International Journal of Robotics Research*, vol. 44, no. 8, pp. 1396–1427, 2025.

[16] M. Seo, S. Han, K. Sim, S. H. Bang, C. Gonzalez, L. Sentis, and Y. Zhu, "Deep imitation learning for humanoid loco-manipulation through human teleoperation," in *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*. IEEE, 2023, pp. 1–8.

[17] Y. Ze, Z. Chen, W. Wang, T. Chen, X. He, Y. Yuan, X. B. Peng, and J. Wu, "Generalizable humanoid manipulation with improved 3d diffusion policies," *arXiv preprint arXiv:2410.10803*, 2024.

[18] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," *arXiv preprint arXiv:2401.02117*, 2024.

[19] F. Liu, Z. Gu, Y. Cai, Z. Zhou, S. Zhao, H. Jung, S. Ha, Y. Chen, D. Xu, and Y. Zhao, "Opt2skill: Imitating dynamically-feasible whole-body trajectories for versatile humanoid loco-manipulation," *arXiv preprint arXiv:2409.20514*, 2024.

[20] black forest labs, "Flux," 2014, 2024-07-01. [Online]. Available: https://blackforestlabs.ai/

[21] B. Xiao, H. Wu, W. Xu, X. Dai, H. Hu, Y. Lu, M. Zeng, C. Liu, and L. Yuan, "Florence-2: Advancing a unified representation for a variety of vision tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4818–4829.

[22] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[23] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 5026–5033.

[24] M. Hu, W. Yin, C. Zhang, Z. Cai, X. Long, H. Chen, K. Wang, G. Yu, C. Shen, and S. Shen, "Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation," *arXiv preprint arXiv:2404.15506*, 2024.

[25] J. Zhang, C. Herrmann, J. Hur, E. Chen, V. Jampani, D. Sun, and M.-H. Yang, "Telling left from right: Identifying geometry-aware semantic correspondence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[26] S. Shin, J. Kim, E. Halilaj, and M. J. Black, "Wham: Reconstructing world-grounded humans with accurate 3d motion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2070–2080.

[27] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, "Humanplus: Humanoid shadowing and imitation from humans," *arXiv preprint arXiv:2406.10454*, 2024.

[28] H. Dai, A. Valenzuela, and R. Tedrake, "Whole-body motion planning with centroidal dynamics and full kinematics," in *2014 IEEE-RAS International Conference on Humanoid Robots*, 2014, pp. 295–302.