Phy-RFMs: Scalable Physics-Grounded Foundation Models for Robust Robotic Manipulation

Xu Xuezhou
National University of Singapore
xuezhou.xu@u.nus.edu

September 25, 2025

Build up physics-grounded robotic foundation models (Phy-RFMs). Recent Vision–Language–Action (VLA) foundation models—such as OpenVLA (Kim et al., 2024), Gemini Robotics (Team et al., 2025), and NVIDIA's GR00T N1 (Bjorck et al., 2025)—have advanced robot manipulation by unifying multimodal understanding with action control. Microsoft's Magma further demonstrates strong generalist capabilities across both digital and physical domains (Yang et al., 2025). Despite these achievements, their reliance on predominantly vision–language–action data leaves them underprepared for contact-rich tasks where success depends on force regulation, compliance, and physical reasoning—for example, gently handling deformable objects or detecting whether a door is latched. This limitation arises from the scarcity of force/torque and tactile representations in current training corpora. Emerging works such as PhyGrasp (Guo et al., 2024), Physics Context Builders (Balazadeh et al., 2024), and DeliGrasp (Xie et al., 2024) illustrate the potential of incorporating object-level physical attributes into control, pointing toward the need for physics-grounded foundation models that treat physical interaction as a first-class modality.

Simulation Toolchain for Scalable Data. Developing such physics-grounded models requires a robust simulation-based toolchain for scalable training and data collection. GPU-accelerated simulators such as Isaac Gym (Makoviychuk et al., 2021), Orbit (Mittal et al., 2023), ManiSkill3 (Tao et al., 2024), and Genesis (Authors, 2024) already enable large-scale synthetic data generation and demonstrate promising sim-to-real transfer in manipulation tasks. While rigid-body dynamics in these platforms are sufficiently accurate for many benchmarks (e.g., RoboTwin), key limitations remain. Contact modeling is often coarse, with force sensors reporting only joint- or body-level aggregates, while surface properties such as friction or compliance are usually treated as static parameters rather than empirically identified. These simplifications hinder tasks requiring fine-grained physical reasoning. Challenges become even more acute for fluids and deformables: triangular mesh representations, effective for rigid objects, struggle to capture continuous deformation or topological change. Alternative methods—including particle-based dynamics, finite-element meshes, and differentiable material point methods—offer progress but involve trade-offs between fidelity, efficiency, and differentiability, and no unified representation has yet emerged (Xian et al., 2023; Zhang et al., 2025). Differentiable engines such as DiffTaichi (Hu, Anderson, et al., 2019) and ChainQueen (Hu, Liu, et al., 2019) address parts of these challenges, enabling gradient-based co-optimization of simulators and control policies.

Physics Tokens and Representations. Beyond better simulators, structured representations are needed to integrate physics into robotic foundation models. Analogous to tokenization in language models, I propose *physics tokens*: compact encodings of measured or inferred quantities such as mass, friction coefficients, compliance, force/torque vectors, and slip indicators. Prior works hint at this potential—PhyGrasp shows that physical descriptors improve grasp selection (Guo et al., 2024), DeliGrasp infers mass and friction to guide contact-rich manipulation (Xie et al., 2024), and tactile sensors like GelSight extract force and slip cues from high-resolution contact images (Yuan et al., 2017). However, what is

missing is a generalizable tokenization scheme that unifies tactile, proprioceptive, and inferred attributes, enabling networks not only to encode but also to *reason* about physical outcomes.

Physics-Structured Network Architectures. Representations alone are insufficient; network architectures must be designed to exploit them. Simply concatenating physics tokens into a generic transformer risks underutilizing their structure. Instead, architectures inspired by dynamical systems provide a more natural fit. Liquid neural networks embed continuous-time dynamics into neurons and have shown adaptability in out-of-distribution scenarios with fewer parameters than standard recurrent networks (Kumar et al., 2023). Hamiltonian and Lagrangian neural networks enforce conservation principles, improving stability and interpretability (Cranmer et al., 2020; Greydanus et al., 2019). Similarly, physics-informed neural networks (PINNs) penalize violations of governing differential equations during training, improving sample efficiency and robustness (Cai et al., 2021). By combining such architectures with physics tokens, this project aims to create RFMs that generalize not only semantically but also remain robust under contact disturbances.

Algorithms for Physics-Grounded Learning. If representations and architectures determine how physics is encoded, algorithms dictate how it is learned. Active exploration for system identification (ASID) enables efficient estimation of latent parameters such as friction and mass (Memmel et al., 2024). Differentiable physics engines—DiffTaichi (Hu, Anderson, et al., 2019), ChainQueen (Hu, Liu, et al., 2019), and FluidLab (Xian et al., 2023)—allow joint optimization of policies and material parameters, narrowing the sim-to-real gap for deformable and fluid tasks. In parallel, physics-aware planning frameworks such as PhyPlan combine learned skills with Monte Carlo Tree Search to boost robustness in long-horizon manipulation (Vagadia et al., 2024).

Integration and Outlook. In sum, this research aims to build robotic foundation models that explicitly leverage physics as a core modality—from high-fidelity simulators and structured physics tokens to physics-informed architectures and learning algorithms. By grounding robots in physical reasoning, we can move beyond purely semantic generalization toward models that understand and predict real-world physical interactions, making them more robust, safe, and efficient in contact-rich environments.

References

Authors, G. (2024). Genesis: A generative and universal physics engine for robotics and beyond. https://github.com/Genesis-Embodied-AI/Genesis

Balazadeh, V., Ataei, M., Cheong, H., Khasahmadi, A. H., & Krishnan, R. G. (2024). Physics context builders: A modular framework for physical reasoning in vision-language models. *arXiv preprint arXiv:2412.08619*.

Bjorck, J., Castañeda, F., Cherniadev, N., Da, X., Ding, R., Fan, L., Fang, Y., Fox, D., Hu, F., Huang, S., et al. (2025). Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*.

Cai, S., Mao, Z., Wang, Z., Yin, M., & Karniadakis, G. E. (2021). Physics-informed neural networks (pinns) for fluid mechanics: A review. *Acta Mechanica Sinica*, *37*(12), 1727–1738.

Cranmer, M., Greydanus, S., Hoyer, S., Battaglia, P., Spergel, D., & Ho, S. (2020). Lagrangian neural networks. *arXiv preprint arXiv:2003.04630*.

Greydanus, S., Dzamba, M., & Yosinski, J. (2019). Hamiltonian neural networks. *Advances in neural information processing systems*, *32*.

Guo, D., Xiang, Y., Zhao, S., Zhu, X., Tomizuka, M., Ding, M., & Zhan, W. (2024). Phygrasp: Generalizing robotic grasping with physics-informed large multimodal models. *arXiv preprint arXiv:2402.16836*.

Hu, Y., Anderson, L., Li, T.-M., Sun, Q., Carr, N., Ragan-Kelley, J., & Durand, F. (2019). Difftaichi: Differentiable programming for physical simulation. *arXiv preprint arXiv:1910.00935*.

Hu, Y., Liu, J., Spielberg, A., Tenenbaum, J. B., Freeman, W. T., Wu, J., Rus, D., & Matusik, W. (2019). Chainqueen: A real-time differentiable physical simulator for soft robotics. *2019 International conference on robotics and automation (ICRA)*, 6265–6271.

Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al. (2024). Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.

Kumar, K., Verma, A., Gupta, N., & Yadav, A. (2023). Liquid neural networks: A novel approach to dynamic information processing. *2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)*, 725–730.

Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Rudin, N., Allshire, A., Handa, A., et al. (2021). Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*.

Memmel, M., Wagenmaker, A., Zhu, C., Yin, P., Fox, D., & Gupta, A. (2024). Asid: Active exploration for system identification in robotic manipulation. *arXiv preprint arXiv:2404.12308*.

Mittal, M., Yu, C., Yu, Q., Liu, J., Rudin, N., Hoeller, D., Yuan, J. L., Singh, R., Guo, Y., Mazhar, H., Mandlekar, A., Babich, B., State, G., Hutter, M., & Garg, A. (2023). Orbit: A unified simulation framework for interactive robot learning environments. *IEEE Robotics and Automation Letters*, *8*(6), 3740–3747. https://doi.org/10.1109/LRA.2023.3270034

Tao, S., Xiang, F., Shukla, A., Qin, Y., Hinrichsen, X., Yuan, X., Bao, C., Lin, X., Liu, Y., Chan, T.-k., et al. (2024). Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai. *arXiv preprint arXiv:2410.00425*.

Team, G. R., Abeyruwan, S., Ainslie, J., Alayrac, J.-B., Arenas, M. G., Armstrong, T., Balakrishna, A., Baruch, R., Bauza, M., Blokzijl, M., et al. (2025). Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*.

Vagadia, H., Chopra, M., Barnawal, A., Banerjee, T., Tuli, S., Chakraborty, S., & Paul, R. (2024). Phyplan: Compositional and adaptive physical task reasoning with physics-informed skill networks for robot manipulators. *arXiv preprint arXiv:2402.15767*.

Xian, Z., Zhu, B., Xu, Z., Tung, H.-Y., Torralba, A., Fragkiadaki, K., & Gan, C. (2023). Fluidlab: A differentiable environment for benchmarking complex fluid manipulation. *arXiv preprint arXiv:2303.02346*.

Xie, W., Valentini, M., Lavering, J., & Correll, N. (2024). Deligrasp: Inferring object properties with llms for adaptive grasp policies. *arXiv preprint arXiv:2403.07832*.

Yang, J., Tan, R., Wu, Q., Zheng, R., Peng, B., Liang, Y., Gu, Y., Cai, M., Ye, S., Jang, J., et al. (2025). Magma: A foundation model for multimodal ai agents. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 14203–14214.

Yuan, W., Dong, S., & Adelson, E. H. (2017). Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, *17*(12), 2762.

Zhang, Y., Luck, K. S., Verdoja, F., Kyrki, V., & Pajarinen, J. (2025). Modesuite: Robot learning task suite for benchmarking mobile manipulation with deformable objects. *arXiv preprint arXiv:2507.21796*.