PHYSICA 0.1: Self-Learning Foundation Model for Zero-Shot Skilled Task Learning

Nova X Robotics AI Team[1] **O. Kroemer**[2] **R. Lioutikov**[3]
[1]Nova X Robotics [2]Robotics Institute, Carnegie Mellon University [3]Karlsruhe Institute of Technology

Abstract: Industrializing physical AI requires solving new tasks without data or teleoperation campaigns. PHYSICA bootstraps from slow frontier-model demonstrations, then learns from its own attempts: an object-frame contact-physics model guides robot-specific skill policies toward faster execution. It learns 23 simulated tasks without human demonstrations, $\sim$4$\times$ faster than its bootstrap and than human teleoperation. An SO-101 arm reduces a cube pick from 19.7 s to 0.9 s. Project page and video: https://physica-corl.vercel.app/
Keywords: Agentic Robotics, Self-Learning, Foundation Models

1 Introduction & Related Work

We introduce PHYSICA, a model that learns new manipulation tasks without human demonstrations and improves its own execution beyond human teleoperation speed. A frontier model (XLM) acts as a zero-shot controller and produces first successful but often conservative executions, which serve as PHYSICA's only demonstrations. From these, PHYSICA learns an object-frame model of the task's contact physics and robot-specific skill models, which it refines through self-practice. This design follows a semantic task hierarchy (Fig. 1), where frontier models increasingly cover task semantics and task physics (L3/4) and can control robots natively. However, skilled (fast, fluent and robust), execution also requires contact physics (L2) and robot-specific movement (L1), which internet-scale data captures poorly and must be learned through interaction. Diffusion policies [1, 2, 3] and VLAs [4, 5, 6, 7] instead learn all levels jointly from demonstrations, and System 1/2 architectures [8, 9, 10] underrepresent contact dynamics in their action decoders. Their performance scales with demonstration data [11, 12], but even with automatic labeling [13, 14] it remains bounded by the demonstrator, and RL fine-tuning often degrades pre-trained generalization [15, 16]. Given the versatility of HW and target environments, we posit that a stronger focus on robots learning from their own practice is ultimately a more scalable approach better suited for real-world deployments compared to collecting demonstrations for every embodiment and task. *Related Work:* Frontier models control robots through affordances and code [17, 18, 19, 20], and agentic systems [21] compose programs [22, 23]. They improve the program; PHYSICA learns motor skills [24, 25, 26, 27]. Self-improvement [28, 29, 30, 31, 32, 33, 34, 35, 36, 37] typically starts from human demonstrations or dense rewards; PHYSICA bootstraps from frontier-model executions instead.
Contributions: (1) frontier-model control as the sole source of demonstrations, (2) a separation of an object-frame physics model from a robot-specific skill model, and (3) faster-than-human execution from self-practice, $\sim$4$\times$ faster than the bootstrap across 23 tasks.

2 PHYSICA

PHYSICA operates on three layers: a reasoning model ($\sim$0.1 Hz) understands the task, provides demonstrations and sequences subtasks; an object-frame physics model M_θ ($\sim$10 Hz) predicts contact physics; and a robot-specific skill policy g_ϕ ($\sim$100 Hz).

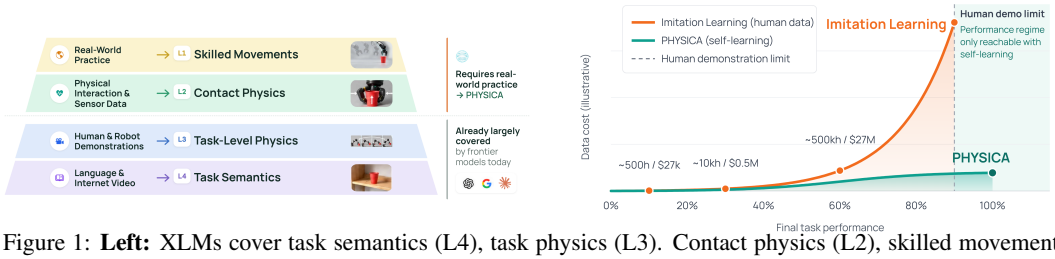


Figure 1: **Left:** XLMs cover task semantics (L4), task physics (L3). Contact physics (L2), skilled movement (L1) need practice. **Right:** data/ performance: IL needs exp. more human data and is capped by demonstrator.

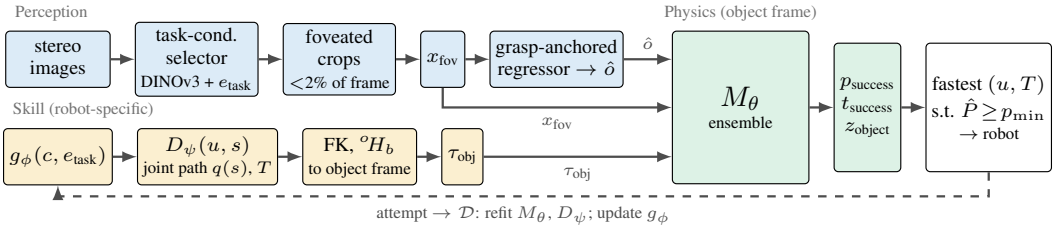


Figure 2: PHYSICA's perception, physics and skill layers. Perception fixates the task object and reads only foveated crops; candidate skills are mapped into the object frame, where the physics model M_θ scores them.

Bootstrap. The XLM receives the instruction, camera images and a control interface (teleop or code) and attempts the task. All attempts, with failures as negative examples, initialize dataset $\mathcal{D}$.

Foveated Perception. A task-conditioned selector fixates the target object in a low-res stereo view. High-res features x_{fov} (DINOv3 [38]) are computed only for that crop as inputs to (M_θ,g), and prediction targets for M_θ and yield target obj pose $\hat{o}$.

Skill policy. A skill is a joint-space trajectory with motion code $u \in \mathbb{R}^{d_u}$ and explicit duration T,

$$g_\phi(c, e_{\text{task}}) \to (u, \log T), \qquad q(s) = D_\psi(u, s), \quad s = t/T \in [0, 1], \tag{1}$$

where c is the context (foveated image, tactile sensor readings when available), e_{task} a frozen VLM task embedding, and D_ψ the decoder of a β-VAE trained on $\mathcal{D}$ with a smoothness penalty,

$$\mathcal{L}_D = \mathbb{E}\big[\|q - D_\psi(u, s)\|^2\big] + \beta\,\mathrm{KL}\big(q_\psi(u \mid q)\,\|\,\mathcal{N}(0, I)\big) + \lambda_s \big\|\partial_s^2 D_\psi\big\|^2. \tag{2}$$

The skill is robot-specific, the architecture allows for open and closed-loop control.

Physics model. M_θ evaluates skills in the object frame, so that the same interaction seen from different robot poses or on different hardware becomes one learning problem. Joint positions at phases s_k are mapped to fingertip paths $\tau_{\text{obj}} = \{{}^oH_b\,\mathrm{FK}(D_\psi(u, s_k))\}_{k=1}^K$ via forward kinematics and $\hat{o}$. Instead of predicting pixels per step, an ensemble predicts the outcome of the whole trajectory,

$$M_\theta(\tau_{\text{obj}}, \hat{o}, x_{\text{fov}}, e_{\text{task}}) \to (\hat{p}, \hat{t}, \hat{z}), \tag{3}$$

i.e., success probability, time to success and object-trajectory code. It is trained with $\mathcal{L}_M = \mathrm{BCE}(\hat{p}, y) + \lambda_t\, y\,|\hat{t} - t| + \lambda_z \|\hat{z} - z\|^2$ on the measured outcomes y, t, z. In simulation, it also predicts tactile readings for closed-loop tactile skills.

Self-improvement. g_ϕ minimizes $\hat{\mathbb{E}}[\hat{t}]$ subject to $\hat{P}(\text{success}) \geq p_{\min}$, where $\hat{P}$ is the ensemble mean minus κ times its spread, so it only exploits regions where M_θ is confident. Practice alternates between executing proposed skills, adding the outcomes to $\mathcal{D}$, refitting M_θ, and updating g_ϕ via gradients through M_θ and on successful and failed motions.

3 Results

Simulation (LIBERO [39] objects, 23 lift/place tasks). Claude's bootstrap is correct but slow: median 5.05 s per lift, 1.4× human teleoperation (3.60 s). After practice, median time to success falls to 1.23 s (3.96×), passing human median after 393 attempts (Fig. 3). Held-out success reaches 0.84–0.88 on 11 tasks vs. 0.71 for nearest-demonstration replay.



Figure 3: **Left:** Simulated Franka lift time from Claude bootstrap (4.95 s) past human teleoperation (3.80 s, attempt 393) to 0.88 s. **Right:** real SO-101, cube pick: 19.7 s at first exploration, 0.9 s after practice.

Real robot SO-101 Arena with 5 objects. After Claude bootstrap, the loop learns from real attempts. Over ∼7k attempts the median cycle falls from 3 s to 1.4 s at 0.92–1.00 success, the pick itself takes 0.9 s after practice. On 173 held-out attempts, M_θ predicts success with AUC 0.97 and the cube's path to 0.70/1.24 cm (success/failure) vs. 1.55/1.82 cm for a static-object baseline.

4 Conclusion

PHYSICA turns slow frontier-model executions into faster-than-human skills through self-practice; since g_ϕ is specialized to individual skills, future versions may leverage hierarchical policies.

References

[1] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems (RSS)*, 2023.

[2] M. Reuss, M. Li, X. Jia, and R. Lioutikov. Goal-Conditioned Imitation Learning using Score-based Diffusion Policies. In *Robotics: Science and Systems XIX*. Robotics: Science and Systems Foundation, July 2023. doi:10.15607/rss.2023.xix.028.

[3] M. Reuss, Ö. Yağmurlu, F. Wenzel, and R. Lioutikov. Multimodal Diffusion Transformer: Learning Versatile Behavior from Multimodal Goals. In *Robotics: Science and Systems XX*. Robotics: Science and Systems Foundation, July 2024. doi:10.15607/rss.2024.xx.121.

[4] B. Zitkovich et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning (CoRL)*, 2023.

[5] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn. OpenVLA: An open-source vision-language-action model. In *Conference on Robot Learning (CoRL)*, 2024.

[6] M. Reuss, H. Zhou, M. Rühle, Ö. E. Yağmurlu, F. Otto, and R. Lioutikov. FLOWER: Democratizing Generalist Robot Policies with Efficient Vision-Language-Flow Models. In *Proceedings of The 9th Conference on Robot Learning*, Sept. 2025.

[7] P. Vanjani, Z. Li, J. Suliga, M. Reuss, G. Geraci, X. Jiang, and R. Lioutikov. DAM-VLA: Decoupled Asynchronous Multimodal Vision Language Action model. In *10th Annual Conference on Robot Learning*, Oct. 2026.

[8] K. Black et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[9] Physical Intelligence. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

[10] Figure AI. Helix: A vision-language-action model for generalist humanoid control. https://www.figure.ai/news/helix, 2025.

[11] Open X-Embodiment Collaboration. Open X-embodiment: Robotic learning datasets and RT-X models. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[12] M. Reuss, J. Pari, P. Agrawal, and R. Lioutikov. Efficient Diffusion Transformer Policies with Mixture of Expert Denoisers for Multitask Learning. In *Thirteenth International Conference on Learning Representations*, Apr. 2025.

[13] N. Blank, M. Reuss, M. Rühle, Ö. E. Yağmurlu, F. Wenzel, O. Mees, and R. Lioutikov. Scaling Robot Policy Learning via Zero-Shot Labeling with Foundation Models. In *Proceedings of The 8th Conference on Robot Learning*, pages 4158–4187. PMLR, Nov. 2024.

[14] N. Blank, P. Mattes, M. X. Li, J. Suliga, T. Roth, M. Reuss, P. Vanjani, and R. Lioutikov. SPARC: Reliable Spatial Annotations from Robot Demonstrations at Scale. In *10th Annual Conference on Robot Learning*, Oct. 2026.

[15] S. Grover, A. Gopalkrishnan, B. Ai, H. I. Christensen, H. Su, and X. Li. Enhancing generalization in vision-language-action models by preserving pretrained representations. *arXiv preprint arXiv:2509.11417*, 2025.

[16] J. Zhu, Y. Hong, X. Sun, et al. Can vision-language-action models learn from real-world data continually without forgetting? *arXiv preprint arXiv:2605.26820*, 2026.

[17] M. Ahn et al. Do as I can, not as I say: Grounding language in robotic affordances. In *Conference on Robot Learning (CoRL)*, 2022.

[18] 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 *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[19] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei. VoxPoser: Composable 3d value maps for robotic manipulation with language models. In *Conference on Robot Learning (CoRL)*, 2023.

[20] Gemini Robotics Team. Gemini robotics: Bringing AI into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

[21] M. Y. Seker, S. Aggarwal, R. Wickramarachchi, J. Francis, and O. Kroemer. From language to manipulation: Composing object-centric skills from scratch, 2026.

[22] R. Lu, Y. Wu, E. Kou, H. Lin, L. Fu, W. Xiao, A. Mandlekar, Y. Xu, G. Shi, K. Goldberg, A. Chen, M. Chowdhury, J. Kautz, Y. Zhu, L. Fan, and G. Wang. ASPIRE: Agentic /skills discovery for robotics. *arXiv preprint arXiv:2607.00272*, 2026.

[23] L. Fu, J. Yu, K. El-Refai, E. Kou, H. Xue, H. Huang, W. Xiao, G. Wang, D. Niu, L. Fei-Fei, G. Shi, J. Wu, S. Sastry, Y. Zhu, K. Goldberg, and L. Fan. CaP-X: A framework for benchmarking and improving coding agents for robot manipulation, 2026.

[24] A. J. Ijspeert, J. Nakanishi, H. Hoffmann, P. Pastor, and S. Schaal. Dynamical movement primitives: Learning attractor models for motor behaviors. *Neural Computation*, 25(2):328–373, 2013.

[25] A. Paraschos, C. Daniel, J. Peters, and G. Neumann. Probabilistic movement primitives. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

[26] H. Zhou, W. Liao, X. Huang, Y. Tang, F. Otto, X. Jia, X. Jiang, S. Hilber, G. Li, Q. Wang, Ö. E. Yağmurlu, N. Blank, M. Reuss, and R. Lioutikov. BEAST: Efficient Tokenization of B-Splines Encoded Action Sequences for Imitation Learning. In *Advances in Neural Information Processing Systems*, Dec. 2025.

[27] X. Han, H. Xiong, H. Chen, C. Liu, A. Torralba, Y. Zhu, and Y. Du. B-spline policy: Accelerating manipulation policies via b-spline action representations, 2026. URL https://arxiv.org/abs/2607.09648.

[28] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, and S. Levine. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on Robot Learning (CoRL)*, 2018.

[29] Physical Intelligence. $\pi^*_{0.6}$: a VLA that learns from experience. *arXiv preprint arXiv:2511.14759*, 2025.

[30] K. Bousmalis et al. RoboCat: A self-improving generalist agent for robotic manipulation. *Transactions on Machine Learning Research*, 2024.

[31] Z. Zhou, P. Atreya, A. Lee, H. Walke, O. Mees, and S. Levine. Autonomous improvement of instruction following skills via foundation models. In *Conference on Robot Learning (CoRL)*, 2024.

[32] S. K. S. Ghasemipour et al. Self-improving embodied foundation models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[33] Y. J. Ma, W. Liang, G. Wang, D.-A. Huang, O. Bastani, D. Jayaraman, Y. Zhu, L. Fan, and A. Anandkumar. Eureka: Human-level reward design via coding large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

[34] P. Wu, A. Escontrela, D. Hafner, K. Goldberg, and P. Abbeel. DayDreamer: World models for physical robot learning. In *Conference on Robot Learning (CoRL)*, 2022.

[35] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[36] G. Zhou, H. Pan, Y. LeCun, and L. Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.

[37] M. Assran et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[38] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[39] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

A Nova X Robotics AI Team

The Nova X Robotics AI Team consists of, in alphabetical order: Philip Becker-Ehmck, Christian Daniel, Hamza Durrani and Mohamed Roshdi (Nova X Robotics).