Research Proposal

Jingxiang (Boris) Guo* *National University of Singapore*
Singapore, Singapore
* jingxiangguo@u.nus.edu

Abstract

This proposal introduces **EgoVTLA**, a Dexterous VLA (Vision–Language–Action) foundation model that integrates ego-centric vision, tactile sensing, language reasoning, and action generation. The system explicitly targets **cross-embodiment generalization** across robots, with a primary focus on the **Unitree G1 humanoid** platform. EgoVTLA is driven by generative world models, enabling intuitive physics, contact-rich dexterity, and future-state prediction. Two scientific axes motivate this work: (1) **Aligning Human–Robot Perception** through multi-modal sensory fusion and cross-modal causal reasoning, and (2) **Aligning Human–Robot Intuition** through image-foresight world models. Building on these principles, I propose five research directions involving hierarchical WA benchmarks, tactile world models, 4D Gaussian representations, discrete-action dexterous VLA policies, and world-model-in-the-loop distillation.

I. Introduction and Motivation

Recent advances in VLA models—such as RT-2 [1], OpenVLA [2], and Spear [3]—have shown impressive generalization across robotic tasks. Yet these systems remain fundamentally limited in their ability to perform dexterous, contact-rich manipulation, adapt their policies across different embodiments, and reason about causal relationships that involve vision, touch, and proprioception simultaneously. Most VLA systems still behave as powerful imitators rather than perceptually grounded, intuition-driven agents capable of human-like reasoning.

I argue that achieving a **Dexterous VLA Foundation Model** requires addressing two tightly coupled challenges. First, robots must achieve **Human–Robot Perception Alignment** by learning to fuse and cross-reference modalities—vision, tactile signals, sound, and proprioception—in the same causal manner humans use to interpret physical interactions. Second, robots must develop **Human–Robot Intuition Alignment**, acquiring internal world models that allow them to anticipate the consequences of their actions without explicit physics computation. These two principles, illustrated in Fig. 1, form the conceptual foundation of the proposed EgoVTLA architecture.

A. Align Human–Robot Perception

Humans excel at dexterous manipulation because their perception is inherently multi-modal and deeply intertwined. Visual observation informs expectations about tactile feel; tactile feedback refines estimates about geometry and material properties; proprioception anchors interaction to the body; and even auditory cues can signal subtle shifts in contact. Importantly, these modalities are not processed in isolation—they are blended into a coherent causal understanding that allows humans to predict stability, friction, deformation, and failure modes.

Robots, on the other hand, often operate with incomplete or poorly integrated sensory representations. Existing VLA models rarely incorporate tactile sensing [4], and when additional modalities are used, they are typically fed through separate encoders and fused only at a late stage. Such designs prevent the model from discovering cross-modal dependencies and inhibit any form of synesthetic reasoning that is crucial for dexterity.

EgoVTLA adopts an explicitly ego-centric, multi-modal approach. Visual inputs from wrist and head cameras, tactile images when available, proprioceptive signals, and other sensory streams are jointly modeled within a unified representation. Instead of concatenating encoders, the model learns cross-modal

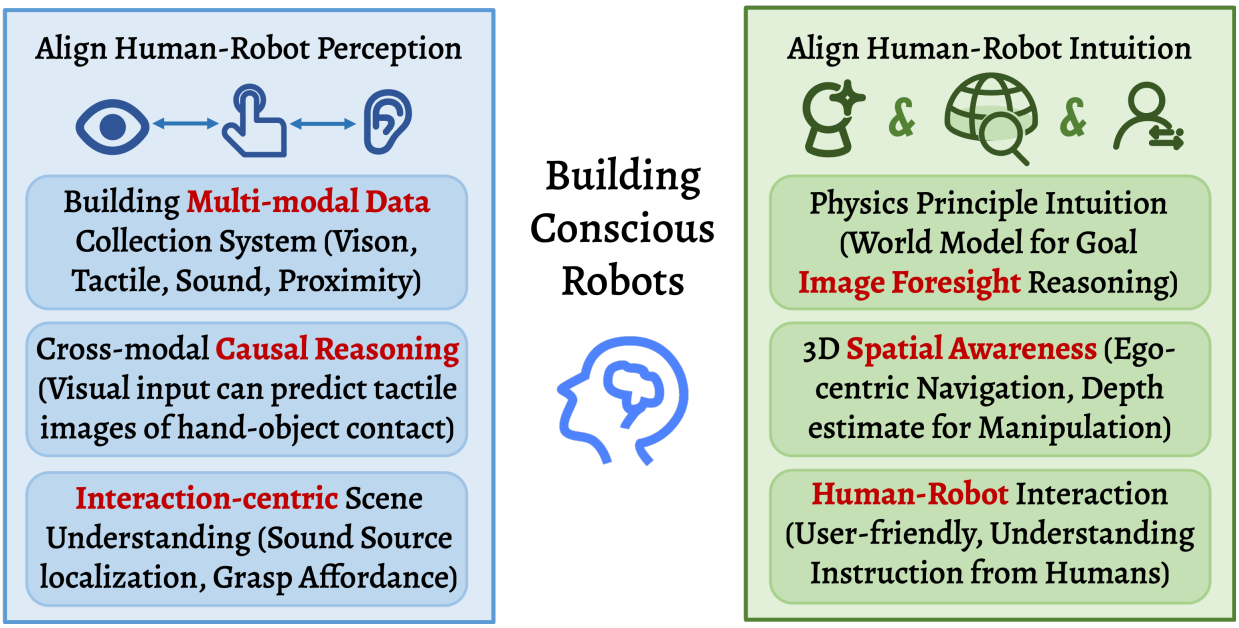

Fig. 1: Two pillars of EgoVTLA: aligning perception and intuition.

predictive relationships—for instance, predicting tactile deformation solely from visual contact geometry. This perceptual grounding is essential not only for dexterity but also for **cross-embodiment transfer**, enabling skills learned on manipulation arms (e.g., Franka or XArm) to generalize to the **Unitree G1 humanoid** without retraining from scratch.

B. Align Human–Robot Intuition

Humans possess an intuitive understanding of physics built from years of interacting with the world. We do not compute forces or torques explicitly; rather, we rely on internal mental models that predict how objects will move, how contact will propagate, and how stable a grasp will be. The emergence of powerful generative world models—such as Genie [5], DynamiCrafter [6], and Marble [7]—provides a pathway toward endowing robots with similar foresight capabilities.

A robot equipped with a world model can simulate future visual frames conditioned on actions, enabling it to "imagine" the outcomes of its decisions before acting. This image-foresight acts as an implicit physics engine, supporting multi-step planning, re-grasping strategies, and material-aware interactions. Fig. 2 illustrates how EgoVTLA incorporates world-model-driven prediction to guide dexterous behavior. By combining vision, scene geometry, depth, and 3D point flow, the system builds an internal representation of the future that is far richer than what traditional model-based control provides.

II. Proposed Research: EgoVTLA

EgoVTLA unifies ego-centric perception, tactile prediction, linguistic reasoning, foresight-based decision making, and cross-embodiment action control into a single dexterous VLA foundation model. The ultimate testbed of this system will be the **Unitree G1 humanoid**, whose human-like structure makes it ideal for evaluating generalizable dexterous policies.

A. Hierarchical WA Benchmark for Dexterous VLA

To systematically study the interaction between world models and action policies, I propose building the first benchmark dedicated to hierarchical WA (World–Action) architectures. Instead of evaluating manipulation end-to-end, this benchmark decouples high-level world modeling from low-level control

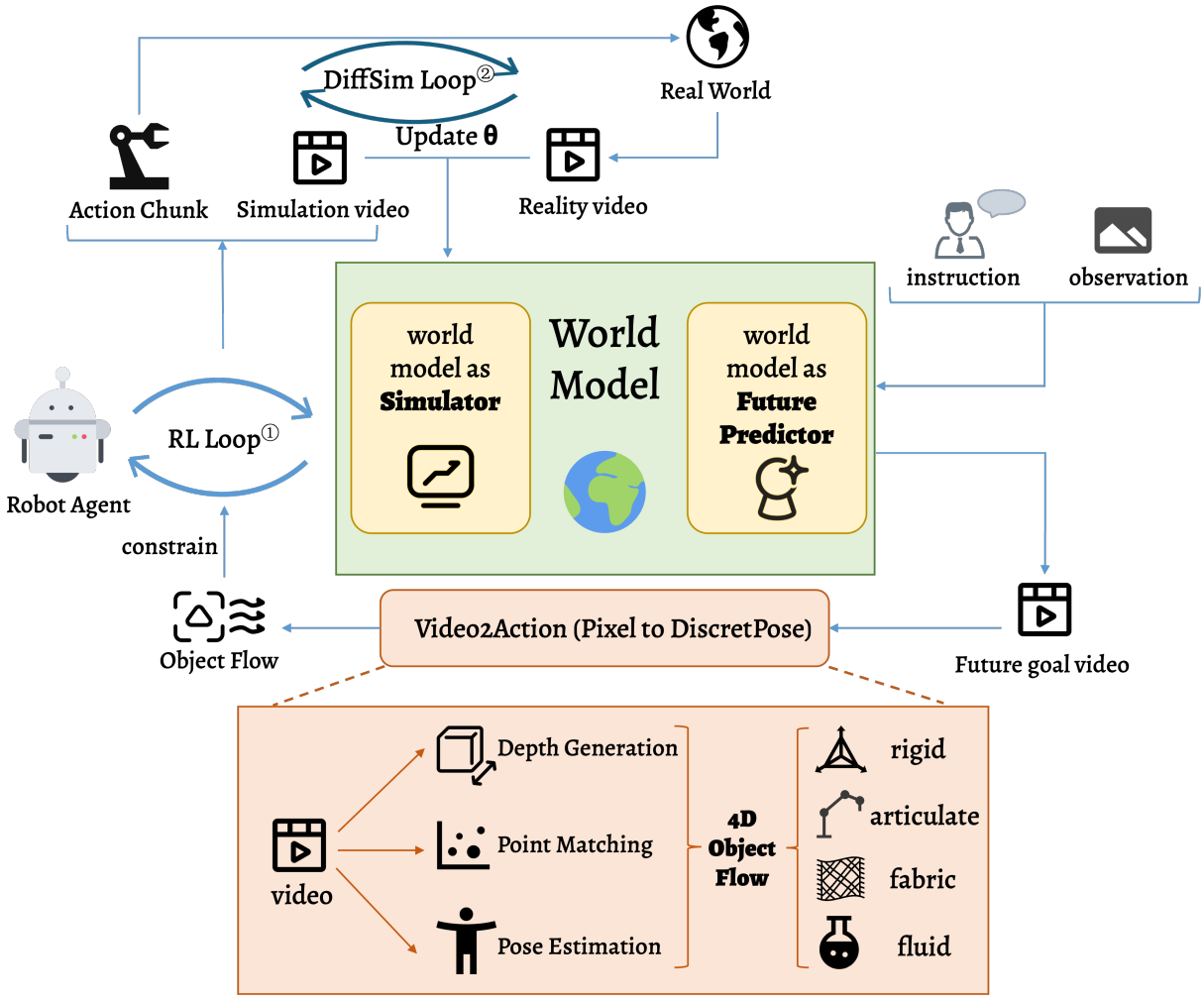

Fig. 2: World-model-driven EgoVTLA: vision → depth/flow → action.

execution. World models are judged by their ability to generate accurate video predictions, depth maps, and 3D point flows, using tools such as CoTracker [8] or DepthAnything [9]. Policies are evaluated on their ability to translate these predicted intermediates into robust actions. This structure allows researchers to identify whether failure arises from perceptual modeling, world-model prediction, or control execution. Furthermore, the benchmark explicitly incorporates **cross-embodiment transfer**, enabling systematic evaluation of how skills migrate from arms to humanoids such as the G1.

B. Tactile World Model

Dexterous tasks rely heavily on tactile feedback. Building upon ViTacFormer [4] and UniFoLM [10], I propose a tactile world model capable of predicting future tactile images given visual observations, historical tactile data, and action sequences. Such a model enables pre-contact reasoning: the robot can estimate whether a grasp will slip, whether a surface is compliant, or whether a tool will rotate stably in hand. This predictive tactile capability is crucial for tasks such as in-hand manipulation, tool operation, and fine assembly.

C. 4D Gaussian Splatting Flow VLA Model

Pixel-based video prediction is insufficient for many dexterous tasks that require structured understanding of geometry. To address this, I propose constructing a 4D Gaussian Splatting representation that

encodes radiance, geometry, and temporal flow in a continuous world model. This representation offers far richer dynamic information than mesh-based geometry and allows EgoVTLA to track object motion, deformation, and contact surfaces with high precision. Conditioning this representation on language instructions further supports high-level planning grounded in structured world knowledge.

D. Discrete "Key-binding" Dexterous Action Space

Instead of generating high-dimensional joint trajectories directly, EgoVTLA explores an abstract key-binding action space inspired by systems such as Genie [5]. Actions such as "twist," "pinch," "press," or "shift grip" are represented as symbolic primitives, which the robot maps to embodiment-specific trajectories. This abstraction reduces policy entropy, simplifies exploration, and enables **embodiment-agnostic dexterous control**. A symbolic "twist" command can translate into different motor behaviors on a Franka arm and the G1 humanoid, yet remain semantically consistent.

E. World-Model-in-the-Loop Distillation

Finally, I propose a closed-loop distillation framework inspired by model-in-the-loop reinforcement learning [11]. The world model generates predictions of future frames, which the policy uses to plan actions. These actions are executed in simulation through systems such as ManiSkill [12]. By comparing the predicted future with the actual executed trajectory, the system obtains a learning signal that simultaneously improves both the world model (RL for WM) and the policy (RL in WM). Iterated over time, this feedback loop produces a self-improving world model capable of supporting increasingly sophisticated dexterous behavior.

III. Expected Contributions

This project will deliver a unified **Dexterous VLA Foundation Model** integrating ego-centric perception, tactile prediction, intuitive physics, and cross-embodiment generalization. It will introduce new hierarchical benchmarks separating world modeling from action generation, a tactile world model for contact-rich tasks, a 4D Gaussian Splatting dynamic representation, a symbolic key-binding action space, and a world-model-in-the-loop distillation system. The final system will be demonstrated on the **Unitree G1 humanoid**, validating the ability of EgoVTLA to generalize dexterous skills across embodiments.

References

[1] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning*. PMLR, 2023, pp. 2165–2183.

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

[3] N. Nikolov, G. Albanese, S. Dey, A. Yanev, L. Van Gool, J.-N. Zaech, and D. P. Paudel, "Spear-1: Scaling beyond robot demonstrations via 3d understanding," *arXiv preprint arXiv:2511.17411*, 2025.

[4] L. Heng, H. Geng, K. Zhang, P. Abbeel, and J. Malik, "Vitacformer: Learning cross-modal representation for visuo-tactile dexterous manipulation," *arXiv preprint arXiv:2506.15953*, 2025.

[5] J. Bruce, M. D. Dennis, A. Edwards, J. Parker-Holder, Y. Shi, E. Hughes, M. Lai, A. Mavalankar, R. Steigerwald, C. Apps *et al.*, "Genie: Generative interactive environments," in *Forty-first International Conference on Machine Learning*, 2024.

[6] J. Xing, M. Xia, Y. Zhang, H. Chen, W. Yu, H. Liu, G. Liu, X. Wang, Y. Shan, and T.-T. Wong, "Dynamicrafter: Animating open-domain images with video diffusion priors," in *European Conference on Computer Vision*. Springer, 2024, pp. 399–417.

[7] W. Labs, "Marble: A multimodal world model," https://marble.worldlabs.ai/, 2025, accessed: 2025-11-29.

[8] N. Karaev, Y. Makarov, J. Wang, N. Neverova, A. Vedaldi, and C. Rupprecht, "Cotracker3: Simpler and better point tracking by pseudo-labelling real videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 6013–6022.

[9] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," 2024. [Online]. Available: https://arxiv.org/abs/2406.09414

[10] Unitree, "Unifolm-wma-0: A world-model-action (wma) framework under unifolm family," 2025.

[11] P. Intelligence, "$\pi^*_{0.6}$: a vla that learns from experience," 2025. [Online]. Available: https://arxiv.org/abs/2511.14759

[12] J. Gu, F. Xiang, X. Li, Z. Ling, X. Liu, T. Mu, Y. Tang, S. Tao, X. Wei, Y. Yao *et al.*, "Maniskill2: A unified benchmark for generalizable manipulation skills," *arXiv preprint arXiv:2302.04659*, 2023.