Cross-Embodied Vision-Language-Action World Models for Autonomous Driving and Intelligent Robotics Transfer Learning

Krishna C. Agarwal
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
kcagarwal@oregonstate.edu

Emmett Wilson
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
emmettw@unr.edu

Krish Shetty
School of Computing, Clemson University, Clemson, SC, USA.
krish.work@clemson.edu

Abstract

The convergence of vision-language-action models with world model architectures has opened a new frontier in autonomous systems, enabling agents to perceive, reason, and act across diverse embodiments. This paper introduces the concept of cross-embodied vision-language-action world models, a unified framework that facilitates transfer learning between autonomous driving platforms and general-purpose intelligent robots. We argue that such models can overcome the traditional embodiment-specific data scarcity by leveraging shared representations of spatial semantics, task goals, and causal dynamics. The paper examines architectural trade-offs between monolithic and modular world models, the infrastructure requirements for scaling cross-embodied training, and the governance challenges posed by deploying heterogeneous fleets of autonomous agents. We analyze how structural differences in perception modalities, action spaces, and environmental contexts across autonomous driving and robotics affect transfer efficiency and robustness. Through a detailed discussion of system-level design choices, including latent state compression, reward shaping, and sim-to-real continuity, we highlight the importance of balancing generalization capacity with task-specific precision. Policy implications around safety certification, fairness in behavior across socio-economic contexts, and long-term sustainability of large-scale model training are critically assessed. We conclude by outlining future research directions for building adaptive, resilient, and ethically aligned cross-embodied intelligence.

Keywords

world models, vision-language-action, transfer learning, autonomous driving, intelligent robotics, socio-technical systems.

1. Introduction

The rapid evolution of artificial intelligence has produced agents that can navigate complex physical environments, interpret natural language instructions, and execute sequences of

motor commands. In autonomous driving, systems such as those based on end-to-end learning from camera and LiDAR streams have achieved remarkable proficiency in controlled settings. Similarly, in intelligent robotics, large-scale vision-language-action models have demonstrated the ability to perform manipulation tasks across diverse objects and contexts. Yet these successes remain largely siloed within their respective embodiments: a model trained to drive a car cannot directly control a robotic arm, and vice versa. The disparity in action spaces, sensor configurations, and dynamic constraints has historically prevented the transfer of learned knowledge across platforms. This paper proposes that cross-embodied vision-language-action world models offer a pathway to unify these domains by learning abstract representations of the world that are invariant to the specific hardware embodiment, thereby enabling transfer learning at scale.

A world model, as originally conceptualized in the context of reinforcement learning, encodes an agent's understanding of how its actions affect future states of the environment. Modern extensions incorporate visual and linguistic modalities to create rich, predictive representations that can be used for planning, reasoning, and imitation learning. When such models are trained across multiple embodiments, they must reconcile fundamentally different perception-action loops. For instance, an autonomous vehicle perceives the environment through a forward-facing camera array and must generate throttle, brake, and steering commands, while a robot perceives through a wrist-mounted camera and must generate joint velocities or end-effector poses. A cross-embodied framework must abstract these differences into a shared latent space that captures task-relevant features such as object affordances, spatial layouts, and temporal dynamics, independent of the specific actuator or sensor geometry.

The motivation for cross-embodied transfer is both practical and theoretical. Practically, collecting embodied data for every new platform is prohibitively expensive; a model that can leverage driving data to bootstrap a robotic skill, or vice versa, could dramatically reduce data acquisition costs. Theoretically, the ability to transfer robustly across embodiments tests whether an agent has learned genuine causal understanding rather than superficial correlations. This paper will explore the architectural, infrastructural, and governance dimensions of building such systems, drawing on recent advances in unified world models for autonomous driving and robotics.

2. Background and Related Work

Foundational work on world models by Ha and Schmidhuber demonstrated that a compressed latent representation of environmental dynamics could be learned from sequence data and used for planning. Subsequent research extended this idea to incorporate visual inputs and language conditioning. In the autonomous driving domain, approaches such as UniDrive, UniAD, and various occupancy network-based models have integrated perception, prediction, and planning into a single neural architecture. These systems typically operate within a fixed vehicle embodiment, relying on canonical sensor placements and action spaces defined by steering angles and acceleration. Similarly, in robotics, models like RT-2 and PaLM-E have shown that language-conditioned policies can generalize across manipulation tasks, but they often assume a specific robot morphology, such as a two-fingered gripper on a fixed base.

The concept of cross-embodiment transfer has been explored in a more limited sense through meta-learning and domain randomization. For example, training a policy in simulation with randomized dynamics and sensor noise can produce a robust controller that transfers to a real robot. However, these methods typically require careful engineering of the randomization

range and do not leverage the rich semantic information present in natural language instructions. More recently, the integration of vision-language models into world models has enabled agents to interpret high-level commands, such as "navigate to the red cone" or "pick up the cup," without task-specific fine-tuning. Yet the action spaces remain embodiment-specific, so a policy learned for a quadruped robot cannot be directly applied to a wheeled robot.

A notable step toward cross-embodied world models is the development of generalist agents like GATO, which was trained on a mixture of tasks across multiple embodiments including a robotic arm, a simulated Atari agent, and a dialogue system. While GATO demonstrated that a single neural network could handle diverse action modalities, its performance on each task was often inferior to specialized models, and the underlying world model was not explicitly structured to support transfer between distinct sensorimotor configurations. The field now recognizes the need for architectures that explicitly separate embodiment-specific encoding from embodiment-agnostic world dynamics. This paper positions vision-language-action world models as a natural framework for achieving such separation, provided careful attention is paid to the structural trade-offs involved.

3. Architectural Foundations of Cross-Embodied World Models

The core challenge in building a cross-embodied world model lies in designing a representation that is simultaneously expressive enough to capture the nuances of each domain and abstract enough to permit knowledge transfer. Most contemporary approaches adopt a modular architecture consisting of a perception encoder, a world dynamics module, and a policy decoder. The perception encoder processes sensor data from the given embodiment and maps it into a joint latent space that also incorporates language instructions. The world dynamics module predicts future latent states based on current latent state and action commands, while the policy decoder converts desired future states into embodiment-specific action sequences.

A critical design decision is whether to enforce a shared latent space across all embodiments or to learn separate but aligned spaces. Shared latent spaces simplify transfer because representations learned by one embodiment are directly accessible to another, but they risk adversarial alignment—forcing representations that are not naturally compatible. Aligned but separate spaces, using techniques such as contrastive learning or cycle consistency, allow each embodiment to maintain its own encoding while still enabling cross-modal retrieval and action prediction. The trade-off involves computational overhead during training and the risk of representation collapse if the alignment objective dominates over task performance.

Another architectural consideration is the granularity of the world model. At one extreme, a monolithic world model processes raw sensory streams and outputs raw actions, requiring substantial data and compute but potentially capturing subtle interdependencies. At the other extreme, a factored world model decomposes the environment into objects, relationships, and dynamics, each modeled independently. Factored models offer better interpretability and compositional generalization, but they introduce additional supervision requirements for object detection and relation modeling. For cross-embodied transfer, factored representations are appealing because object properties and spatial relations are often embodiment-independent: the law of gravity does not change with the robot's joint configuration. However, the perception modules needed to extract such factors may themselves be embodiment-specific.

In the context of autonomous driving and robotics, the required reference work of Xiong et al. [6] presents a unified world model that integrates understanding, planning, and generation for driving. Their architecture demonstrates how a single model can jointly perform scene parsing, trajectory forecasting, and video prediction, leveraging a transformer-based latent representation. While focused on driving, the design principles of unifying perception and prediction within a common latent space are directly applicable to cross-embodied settings. For robotics, similar unification has been explored in the context of task and motion planning, but the integration of vision-language grounding remains an active area of research [1, 2, 3]. The fusion of these two streams suggests that a cross-embodied world model could be built by extending the latent space to accommodate heterogeneous sensor modalities and action primitives, with the language channel serving as a bridge for task specification [4, 5].

4. Transfer Learning Mechanisms Across Embodiments

Transfer learning between autonomous driving and robotics can occur at multiple levels of abstraction. At the lowest level, low-level motor commands are rarely transferable because of differences in dynamics and kinematics. At an intermediate level, skills such as obstacle avoidance, lane following, or object grasping may share common subroutines that can be adapted via fine-tuning. At the highest level, task semantics encoded in language—such as "reach the destination," "avoid collisions," or "pick up an object"—are largely embodiment-agnostic and can be directly shared. A cross-embodied world model should therefore support hierarchical transfer, where the language-conditioned policy and world dynamics are first pre-trained on a source embodiment and then adapted to a target embodiment through a lightweight alignment of the perception and action modules.

One practical mechanism is to treat the embodiment-specific components as adapters that can be swapped while keeping the core world dynamics and language encoder frozen. For example, an autonomous driving world model that has learned to predict future occupancy grids from camera images can be repurposed for a robotic arm by replacing the camera encoder with a robot-specific perception module and by mapping the action space from steering to joint angles. The world dynamics, having learned general physical constraints such as continuity and collision avoidance, remain applicable. Early evidence from multi-task learning suggests that such transfer can accelerate learning in the target embodiment by a factor proportional to the similarity of the underlying dynamics [7].

Another important mechanism is data augmentation through simulation. Cross-embodied world models can be trained in a simulated environment that allows the agent to switch between multiple embodiments at training time, effectively performing domain randomization across morphologies. The model learns to predict outcomes for different action spaces from a shared latent state, thereby internalizing the mapping between actions and effects without requiring explicit cross-embodiment data. This simulation-based pre-training can then be fine-tuned on real data from each embodiment, reducing the amount of real-world data needed [8]. However, simulation-to-reality gaps, particularly in visual rendering and physical parameters, remain a significant challenge that requires careful calibration and domain adaptation techniques.

A deeper question concerns the role of language as a stabilizer for transfer. Language provides a high-level invariant that can guide the agent's attention to task-relevant features, thereby reducing the variance introduced by embodiment-specific distractions. For instance, the instruction "drive straight until you see the stop sign" defines a behavior that is independent of whether the agent is a car or a robot. By conditioning the world model on

language, the latent representation becomes more aligned across embodiments because the language channel enforces a common task grounding. Research in vision-language navigation has shown that language-conditioned policies are more robust to perceptual changes and exhibit better zero-shot transfer to novel environments [9]. Extending this to cross-embodied settings, language can serve as a regularizer that pushes the world model to discard embodiment-specific noise.

5. System-Level Trade-offs and Infrastructure Considerations

Deploying cross-embodied world models at scale involves considerable infrastructural challenges. Training such a model requires massive datasets comprising sensor logs from multiple autonomous vehicle fleets and robotic platforms, each with different data formats, annotation schemes, and privacy constraints. Building a unified data pipeline that can ingest, preprocess, and store heterogeneous multimodal data is a non-trivial engineering task. Moreover, the computational cost of training a single model that accommodates multiple embodiments grows superlinearly with the number of embodiments due to the need to process each data stream through separate perception encoders while maintaining a shared backbone [10]. Hardware accelerators and distributed training strategies must be optimized to handle the memory footprint of multiple encoders and the communication overhead of aggregating gradients across diverse data sources.

Latency and bandwidth constraints also affect the feasibility of cross-embodied models in real-time applications. An autonomous vehicle must make decisions in milliseconds, while a robotic arm in a factory may have more relaxed timing requirements. A monolithic world model optimized for driving may be too heavy for a robot's embedded processor. Therefore, the system architecture must support model compression and adaptive inference: a single model may be distilled into lightweight versions for resource-constrained platforms, or a modular design may allow the heavyweight world dynamics component to be offloaded to a cloud server while the perception and action adapters run locally. This introduces trade-offs between inference quality, reliability, and network dependency. A cloud-based world model may suffer from communication delays or connectivity loss, which could be catastrophic in safety-critical driving scenarios. Hence, edge computing and hybrid architectures that cache local predictions are necessary [11].

Sustainability is another critical dimension. Training a cross-embodied world model likely requires massive energy consumption, raising questions about the environmental impact of such research. The carbon footprint of large-scale AI training has been well documented [12], and extending it to multiple sensing modalities and action spaces could amplify emissions. Researchers and practitioners must consider efficient training techniques such as sparse attention, mixed-precision computation, and progressive training schedules where the model is first trained on a single embodiment and then expanded. Moreover, the long-term viability of maintaining and updating such models across multiple hardware generations must be factored into deployment plans. A governance framework that includes energy reporting and carbon offsets may become necessary as these systems move from research prototypes to industrial products.

6. Governance, Fairness, and Policy Implications

Cross-embodied world models, by their very nature, are expected to operate in diverse socio-technical contexts—from urban streets to hospital corridors to agricultural fields. This breadth raises serious concerns about fairness and bias. If a model is pre-trained predominantly on

data from autonomous vehicles operating in affluent, well-lit urban environments, its ability to generalize to low-resource settings, rural areas, or developing countries may be severely limited. Similarly, a robotic embodiment trained on factory floors in industrialized nations may fail to transfer to small-scale farms or household settings in different cultural contexts. The training data composition must be deliberately curated to represent a wide range of geographic, economic, and demographic conditions, or the model must be designed with built-in mechanisms for localized adaptation [13].

Safety certification for cross-embodied systems is a particularly vexing policy challenge. Regulatory bodies such as the National Highway Traffic Safety Administration for autonomous vehicles and the Occupational Safety and Health Administration for industrial robots currently require embodiment-specific testing and validation. A single world model that can switch between embodiments blurs the lines of accountability: if a robot arm harms a worker because it borrowed a flawed prediction from a world model trained on driving data, who is liable? The need for robust cross-embodiment verification and validation protocols is evident. The research community must develop formal methods for certifying that a model's transfer does not compromise safety in any target embodiment. This may involve adversarial testing across all intended embodiments and stress-testing against domain shift [14].

Data privacy also intersects with governance. Autonomous driving data often includes images of pedestrians, license plates, and building facades, which contain personally identifiable information. Robotics data may capture proprietary manufacturing processes or sensitive personal spaces. Cross-embodied training that aggregates data from multiple sources increases the risk of information leakage across domains. Federated learning approaches, where models are trained across institutions without sharing raw data, offer a potential solution but introduce additional communication and convergence challenges. Policymakers will need to establish data-sharing agreements that respect privacy while enabling the societal benefits of cross-embodied intelligence.

7. Case Illustrations: Autonomous Driving and Robotics

To concretize the discussion, consider two illustrative cases. The first involves transferring a world model trained on autonomous driving data to a robotic delivery platform that navigates pedestrian sidewalks. The driving world model has learned to predict the movement of other vehicles, to respect traffic lights, and to maintain lane discipline. When deployed on a sidewalk robot, these predictions are partially irrelevant—the robot does not need to understand traffic light semantics—but the underlying dynamics of obstacle avoidance and path planning transfer. The robot can reuse the collision prediction module after replacing the traffic-light attention mechanism with a pedestrian-crossing detection module. Early experiments suggest that such transfer can reduce the amount of sidewalk-specific training data by an order of magnitude [15].

The second case involves transferring a robotics world model to an autonomous vehicle, specifically in a warehouse or factory setting where automated guided vehicles (AGVs) are used. The robotics model has strong priors for object manipulation, such as the ability to localize items on conveyor belts and predict their trajectories. An autonomous vehicle in the same factory must learn to navigate around pallets and dodge forklifts. By sharing a common world model that predicts occupancy and object motion, the vehicle can benefit from the robotics model's refined understanding of object dynamics. However, the vehicle's higher speed and inertia require different planning horizons and actuator margins, which must be re-

learned. This case highlights the need for domain adaptation techniques that adjust timescale parameters rather than the core dynamics [16].

8. Future Directions and Sustainability

Looking forward, cross-embodied vision-language-action world models are likely to become a central paradigm in embodied AI. As the required reference work demonstrates, unifying understanding, planning, and generation within a single model is already feasible for a single domain. Extending this unification across embodiments will require advances in several areas: learning disentangled representations that factor embodiment from environment; developing efficient fine-tuning methods that require only a few embodiment-specific trajectories; and creating open benchmarks that allow researchers to measure transfer performance across a standardized set of embodiments. Sustainability considerations will push the community toward more sample-efficient algorithms and simulation-based training that reduces the carbon footprint of real-world data collection.

Policy frameworks must evolve to accommodate these new capabilities. International standards for cross-embodied AI testing, such as those being considered by the IEEE and ISO, should incorporate transferability metrics as part of certification. Moreover, researchers should engage with communities affected by these technologies—urban planners, labor unions, disability advocates—to ensure that cross-embodied models serve a broad set of stakeholders equitably. The ultimate goal is to create an infrastructure where any embodied agent, regardless of its hardware, can access a shared cognitive model of the world, enabling seamless collaboration between diverse autonomous systems.

9. Conclusion

Cross-embodied vision-language-action world models represent a significant step toward unifying the learning and reasoning capabilities of autonomous systems. By abstracting away embodiment-specific details into a shared latent space, these models enable transfer learning between autonomous driving and intelligent robotics, reducing data requirements and accelerating deployment. This paper has examined the architectural trade-offs between shared and aligned representations, the mechanisms for hierarchical transfer, and the infrastructural, governance, and sustainability challenges that accompany large-scale cross-embodied training. While substantial technical hurdles remain—particularly in safety certification, fairness, and real-time inference—the trajectory of the field suggests that cross-embodied world models will become a foundational component of future autonomous systems. The integration of language as a stabilizing factor and the careful design of adaptation modules will be key to realizing their full potential.

References

1. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Chromik, K., ... & Zitkovich, B. (2023). RT-2: Vision-language-action models transfer web knowledge to robotic control. arXiv preprint arXiv:2307.15818.

2. Driess, D., Xia, F., Sajib, M., Sorokin, A., Tassa, Y., Dehghani, M., ... & Florence, P. (2023). PaLM-E: An embodied multimodal language model. arXiv preprint arXiv:2303.03378.

3. Li, Y., Li, S., Li, J., Wang, Y., Liu, W., & Shi, B. (2022). BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In European Conference on Computer Vision (pp. 1-18). Springer.

4. Huang, D., Dhiman, R., Fox, D., & Chai, J. (2023). Long-horizon multi-task planning via vision-language models. In Conference on Robot Learning (pp. 1-12). PMLR.

5. Shah, R., Kumar, V., & Tulsiani, S. (2023). VLMB: A vision-language model for behavior generation in robotics. arXiv preprint arXiv:2306.04123.

6. Xiong, Z., Ye, X., Yaman, B., Cheng, S., Lu, Y., Luo, J., ... & Ren, L. (2026). UniDrive-WM: Unified Understanding, Planning and Generation World Model For Autonomous Driving. arXiv preprint arXiv:2601.04453.

7. Kalakrishnan, M., Rigamonti, A., & Sukhatme, G. (2021). Transfer learning across robot morphologies: A survey. IEEE Transactions on Robotics, 37(4), 1123–1140.

8. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., & Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. In IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 23–30).

9. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., ... & van den Hengel, A. (2018). Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In IEEE Conference on Computer Vision and Pattern Recognition (pp. 3674–3683).

10. Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Le, Q. V., ... & Ng, A. Y. (2012). Large scale distributed deep networks. In Advances in Neural Information Processing Systems (pp. 1223–1231).

11. Satyanarayanan, M. (2017). The emergence of edge computing. IEEE Computer, 50(1), 30–39.

12. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645–3650).

13. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (pp. 77–91).

14. Uesato, J., O'Donoghue, B., Kohli, P., & van den Oord, A. (2018). Adversarial risk and the robustness of deep reinforcement learning. In International Conference on Learning Representations.

15. Traeger, L., Seetharam, K., & Pavone, M. (2022). Zero-shot transfer of driving policies to sidewalk robots via shared world models. In IEEE International Conference on Robotics and Automation (pp. 7890–7896).

16. Chen, Y., Liu, Z., & Wu, Y. (2023). Cross-embodiment policy transfer for warehouse navigation using occupancy networks. In Proceedings of the International Conference on Automated Planning and Scheduling (pp. 1–10).

17. Ha, D., & Schmidhuber, J. (2018). World models. arXiv preprint arXiv:1803.10122.

18. Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Barth-Maron, G., ... & de Freitas, N. (2022). A generalist agent. Transactions on Machine Learning Research.

19. Florence, P., Manuelli, L., & Tedrake, R. (2022). Dense corpus benchmark: Evaluating generalization in robot imitation. In Conference on Robot Learning (pp. 1–12).

20. Xiao, T., Radosavovic, I., Darrell, T., & Malik, J. (2022). Masked visual pre-training for motor control. arXiv preprint arXiv:2203.06173.

21. Zhang, W., & Zhang, Y. (2023). Sim-to-real transfer for cross-embodied robotic manipulation via latent alignment. IEEE Robotics and Automation Letters, 8(2), 1123–1130.