Physically Grounded Multimodal World Models for Embodied Robot Interaction via 3D Representation Alignment

Maurice Anderson
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
manderson@binghamton.edu

Brant Ryen
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
contactbrent@uab.edu

Abstract

The deployment of autonomous robots in unstructured, human-centric environments demands perception-action systems that are simultaneously multimodal, physically coherent, and spatially precise. While recent advances in foundation models have produced powerful language-vision representations and high-fidelity 3D scene reconstructions, a fundamental gap persists between passive recognition and dynamic physical interaction. This paper presents a system-level investigation into physically grounded multimodal world models that integrate 3D representation alignment as a central scaffold for embodied robot control. We examine the architectural trade-offs involved in constructing world models that merge visual, linguistic, proprioceptive, and physical simulation modalities within unified 3D coordinate frames. The discussion focuses on how explicit 3D geometric alignment across sensing streams enables more robust sim-to-real transfer, facilitates long-horizon planning, and supports physical reasoning under uncertainty. We analyze infrastructure and deployment challenges, including latency constraints, computational sustainability, and the governance of data that couples real-world physics with learned priors. Furthermore, we address issues of fairness, bias, and policy implications arising from physically grounded systems that learn behavioral policies from human demonstrations and synthetic experience. By synthesizing insights from neural rendering, multimodal learning, robotics, and AI governance, the paper offers a forward-looking perspective on building physically intelligent robots that are not only capable but also accountable in their interactions with the world.

Keywords

physically grounded world models; multimodal alignment; 3D representation learning; embodied AI; robot interaction; simulation-to-real transfer; governance.

1. Introduction

The ambition to field general-purpose robots that can assist in homes, hospitals, and industrial settings hinges on their ability to form rich internal models of the physical world. Unlike narrow automation that relies on structured environments and preprogrammed routines, physically interactive robots must continuously infer geometry, material properties, object affordances, and task dynamics from a stream of multimodal sensor data. This challenge has historically been fragmented across separate research communities: computer vision

producing increasingly accurate 3D reconstructions, natural language processing enabling instruction following, and control theory delivering stable reactive behaviors. Recent years, however, have witnessed a convergence driven by deep learning architectures that can consume and relate information from multiple modalities [1, 2]. World models, internal predictive simulations of environment dynamics, have emerged as a promising substrate for fusing these capabilities, yet their designs frequently overlook the physical grounding that makes predictions actionable for embodied agents [1]. The integration of explicit 3D geometric structures into such world models—and the alignment of representations across vision, language, and physics—stand as a critical frontier for robotics.

Foundational work on world models demonstrated that agents can learn compressed spatiotemporal representations of visual environments and use them to dream up plausible futures, dramatically improving sample efficiency on control tasks [1]. In parallel, contrastive language-image pretraining established that aligned multimodal embedding spaces could bridge semantic concepts with visual appearance, enabling zero-shot generalization [2]. Similarly, neural radiance fields and their successors revolutionized the fidelity of 3D scene representations by learning continuous volumetric functions from posed image collections [3]. Yet each of these lines of advancement addresses only a fragment of the embodied interaction problem. A robot must not merely look at a mug and describe it; it must grasp the mug while accounting for its mass distribution, anticipated friction, and the fluid dynamics of its contents. Achieving this level of competence requires a physically grounded world model that aligns 3D structure with multimodal sensory streams and physical simulation engines in a coherent, queryable form.

The system-level challenges are formidable. Building a world model that is simultaneously accurate enough for fine manipulation and efficient enough for real-time closed-loop control forces delicate architectural trade-offs. Representations that are rich in physical detail—such as signed distance functions or particle-based deformable models—are computationally expensive to update and propagate, whereas lightweight latent state vectors may discard critical geometric nuance needed for contact-rich tasks. Moreover, the alignment of representations across vision, language, and physics is not a monolithic process but a multifaceted integration problem spanning coordinate frame conventions, dimensionality mismatches, and differing semantic granularities. The incorporation of large language models for task planning introduces additional complexity, as textual descriptions of physical interactions often omit the precise low-level parameters that govern mechanical stability. A systems perspective is therefore essential, one that considers not only algorithmic accuracy but also latency budgets, hardware heterogeneity, sensor calibration drift, and the lifelong learning required in dynamic environments.

This paper adopts such a systems-focused lens to examine how 3D representation alignment can be leveraged as the organizational principle for building physically grounded multimodal world models for embodied robot interaction. We survey the architectural building blocks, analyze structural trade-offs, and discuss deployment realities that extend well beyond laboratory benchmarks. The inquiry is organized as follows. Section 2 reviews foundational concepts and related work, situating our discussion within the broader landscape of world models, multimodal learning, and neural 3D representations. Section 3 outlines a conceptual system architecture and dissects the key design decisions. Section 4 deepens the analysis of 3D representation alignment as a core enabling mechanism. Section 5 confronts the practical challenges of embodied interaction and robot deployment. Section 6 expands the scope to

governance, fairness, and sustainability implications. Finally, Section 7 concludes with reflections on the path toward accountable physical intelligence in robotic systems.

2. Foundational Concepts and Related Work

World models as formalized in recent deep learning literature provide an agent with an internal simulator of the environment, learned from experience and capable of generating imagined trajectories. The canonical approach involves a variational autoencoder that compresses high-dimensional observations into a compact latent state, a recurrent dynamics model that predicts state transitions, and a controller trained entirely within the model's imagination [1]. This architecture achieved remarkable results on visually rich control tasks, yet it operated on 2D image observations without explicit 3D structure, limiting its ability to reason about occlusion, viewpoint change, and physical consistency. Concurrently, the field of computer vision made dramatic leaps in 3D scene understanding. Neural Radiance Fields enabled novel view synthesis with unprecedented quality by encoding a scene as a continuous volumetric radiance field parameterized by a multilayer perceptron [3]. These representations, however, were originally designed for static scenes and view synthesis, not for dynamic interaction or integration with physics simulation.

On the multimodal front, the alignment of language and vision has been transformed by models that map images and text into a shared embedding space using contrastive objectives on internet-scale data [2]. Such models enable a form of semantic grounding where linguistic descriptions can direct visual attention and object recognition, but they lack any explicit notion of 3D spatial relationships or physical dynamics. The extension of alignment to additional modalities—such as depth, audio, thermal, and inertial measurement units—has further demonstrated that a unified embedding can serve as a powerful interface for downstream tasks [7]. What remains underexplored is the tight coupling of these semantically rich embeddings with geometrically precise 3D representations that are physically actionable. A bridge between the two worlds has begun to emerge in work that enforces physics coherence during generative modeling through feature-level and 3D representation alignment, showing that constraining latent representations with physically grounded 3D cues can significantly improve the temporal consistency of synthesized dynamics [8]. These efforts underscore that alignment is not only a semantic convenience but a structural precondition for reliable physical reasoning.

The robotics community has long grappled with the physical grounding challenge, though often through the lens of simulation-to-reality transfer rather than representation alignment. Domain randomization, wherein visual and physical parameters of a simulator are varied during training to promote robustness to real-world variations, has been a cornerstone technique [6]. Physics engines such as MuJoCo provide efficient differentiable or quasi-differentiable simulation of rigid-body dynamics, contacts, and actuation [13]. However, these simulators operate on idealized geometric descriptions and material models that can diverge sharply from real-world physics, particularly for deformable objects, granular media, and fluid interactions. The gap between simulation and reality therefore persists not only in appearance but in the fundamental dynamics. This gap motivates the pursuit of hybrid models that learn residual physics from real data while retaining the stability and interpretability of analytic simulators. Relational inductive biases that structure neural networks as graph-based computations over entities and relations have proven useful for learning physical systems, because they naturally capture the pairwise interactions that characterize Newtonian mechanics [5].

Recent advances in generative modeling have further expanded the toolkit for world model construction. Diffusion models, originally developed for high-fidelity image synthesis, have been adapted for trajectory planning by learning to denoise sequences of states and actions, enabling flexible behavior synthesis that can incorporate constraints and goals [11]. Meanwhile, frameworks for embodied AI research such as Habitat provide standardized platforms where agents interact with photorealistic 3D environments, offering training and evaluation sandboxes that sit between pure simulation and the physical world [12]. These platforms facilitate large-scale experimentation but still rely on rigid-body physics and predetermined action spaces, leaving open the question of how to integrate fine-grained contact mechanics, tactile sensing, and deformable body dynamics into a unified world model. The survey of reinforcement learning in robotics remains a foundational reference for understanding the breadth of control strategies available, from model-free policy gradient methods to model-based planning [4].

3. System Architecture for Physically Grounded Multimodal World Models

Designing a physically grounded multimodal world model for embodied interaction demands an architecture that reconciles several inherently conflicting requirements: geometric fidelity, physical accuracy, multimodal semantic richness, and computational tractability for real-time control loops. At the highest level, such a system may be conceptualized as a modular yet tightly coupled assembly of four subsystems: a multimodal encoder responsible for producing aligned embeddings from heterogeneous sensor streams; a 3D geometric backbone that constructs and maintains a persistent spatial representation of the environment; a physics-infused dynamics predictor that evolves the representation forward in time under plausible physical laws; and a policy head that translates task specifications and world state into motor commands. The interfaces between these subsystems, and the degree to which they share parameters or latent variables, define the system's capacity for generalization, its vulnerability to compounding errors, and its computational footprint.

A central architectural choice concerns whether the 3D representation serves as an explicit intermediate layer that all modalities are forced to address, or whether it remains implicit—emerging only within the latent dynamics of a black-box neural predictor. Explicit 3D alignment has the advantage that it naturally handles viewpoint changes, supports collision reasoning, and can be directly queried by classical planning algorithms. Representations such as 3D Gaussian splats, which model scenes as collections of anisotropic Gaussian primitives with view-dependent appearance, offer a compelling compromise between rendering quality and real-time performance [10]. These structures can be incrementally updated as the robot moves and can be aligned with point clouds from depth sensors, making them attractive as a shared geometric substrate. In contrast, implicit approaches that rely entirely on learned latent states, as in the original world models work [1] and its dreamer-style descendants [9], trade geometric transparency for the flexibility to learn compressed dynamics that may capture subtle physical relationships not easily expressed in explicit geometry. The trade-off is not merely expressive; it directly impacts the auditability and safety of the system, since explicit 3D representations permit human operators to inspect and validate the robot's internal understanding of physical space.

The multimodal encoder must fuse data that differ dramatically in sampling rate, dimensionality, and noise characteristics. Camera images arrive at tens of hertz with millions of pixels, while joint encoders produce kilohertz streams of low-dimensional vectors, and language instructions arrive asynchronously as tokens. Aligning these streams within a

unified 3D frame requires spatial registration that is itself a learned capability, susceptible to calibration drift and domain shift. Vision transformers and their hierarchical variants offer a scalable way to process image and point cloud data by treating patches or local neighborhoods as tokens within an attention mechanism [18, 19]. Language inputs are typically processed by pretrained transformer decoders and projected into a shared embedding space [2, 17]. The critical design decision lies in how these embeddings are fused with the 3D backbone: some architectures inject language features into the 3D representation as additional per-point attributes, enabling language-conditioned segmentation and affordance prediction directly in the geometric domain; others keep language separate and route it into a policy network that interprets the 3D state from outside. The former approach strengthens the physical grounding of linguistic references but increases the dimensionality of the geometric representation, potentially straining memory and compute budgets in long-duration missions.

The dynamics predictor is where physical grounding is most stringently tested. Existing world model implementations typically learn a recurrent state-space model that predicts future latent states without explicit enforcement of physical conservation laws [9]. While this yields impressive short-horizon predictions, the models can produce physically implausible trajectories when rolled out over many steps, due to accumulation of errors that violate momentum conservation, contact constraints, or object permanence. Incorporating a differentiable physics engine as a prior—either by training the predictor to output residual corrections to an analytic simulator or by using the simulator's gradients to regularize the learned dynamics—can dramatically improve long-horizon coherence. Such hybrid approaches require careful synchronization of the coordinate frames used by the learned components and the simulator, which brings 3D representation alignment to the fore. For instance, if the simulator expects object meshes defined in a canonical frame but the perception system tracks objects as deformable point clouds, a mapping must be learned that respects both the geometric identity of objects and their simulated physical properties. The computational cost of running a differentiable physics engine inside the training loop is a serious constraint that system designers must weigh against the benefits of physical consistency, and recent work on efficient parallel simulation and hardware acceleration is beginning to shift this balance.

The policy head must translate the aligned, physically grounded world state into real-time actions. Many current systems separate the world model training from policy learning, using the former as a generative environment for model-based reinforcement learning [9] or as a prior for planning with trajectory optimization. An alternative is to train the policy end-to-end with the world model, allowing gradients from the policy loss to shape the representation itself. However, this tight coupling can lead to representations that are fragile and overfit to the specific task distribution, undermining the generality that is a primary motivation for building world models in the first place. Meta-learning strategies that optimize for fast adaptation across a distribution of tasks provide one route to preserving generality while still enabling end-to-end training [14]. The decision about coupling also reverberates into the hardware deployment: tightly coupled systems may require on-robot fine-tuning that demands substantial compute, whereas modular systems can be trained offline but risk representation mismatch when deployed in novel environments. Careful system profiling and latency budgeting are therefore not afterthoughts but integral to architectural selection.

4. 3D Representation Alignment as a Core Enabling Mechanism

The term alignment in the context of multimodal world models extends beyond the matching of embedding spaces; it encompasses the coordination of geometric reference frames, physical scale, temporal synchronization, and semantic correspondence. 3D representation alignment refers specifically to the process of ensuring that all sensory and learned representations of the environment are coherently referenced to a shared three-dimensional coordinate system that is consistent with physical laws. This alignment operates at multiple levels: low-level sensor registration that corrects for calibration errors, mid-level object-centric alignment that associates visual features with persistent object identities and poses, and high-level conceptual alignment that maps language-grounded task descriptions onto geometric affordances. When executed effectively, 3D representation alignment acts as a skeleton upon which the muscles of multimodal perception and physical prediction can attach, giving the entire system the structural integrity needed for reliable embodied interaction.

At the sensor registration level, modern robotic platforms carry an array of cameras, lidar units, and inertial sensors whose extrinsic calibrations can drift due to thermal expansion, mechanical shock, or wear. Traditional calibration procedures are manual and static, ill-suited to lifelong autonomy. Self-supervised alignment methods that use mutual information or contrastive losses between modalities can continuously refine these extrinsics online, relying on the geometric consistency of the fused 3D representation as the supervisory signal. For example, a neural radiance field optimized from camera images can be compared with lidar depth maps after projecting both into a common coordinate frame; discrepancies can then backpropagate to update the estimated sensor poses. This form of alignment is deeply integrated with the world model itself, blurring the boundary between perception and state estimation. The Mip-NeRF extension, which incorporates anti-aliasing by modeling conical frustums rather than infinitesimal rays, is particularly relevant for aligning representations across sensors with different spatial resolutions and viewing geometries, as it naturally handles scale ambiguity [20].

At the object level, 3D representation alignment must solve a dynamic correspondence problem: matching segmented object proposals across time and across modalities while respecting physical continuity. Cognitive mapping and planning approaches that build topological graphs over metric spatial representations have demonstrated robust long-range navigation by anchoring perception in durable landmarks [15]. World models aiming for fine manipulation require an even denser alignment, where each point on an object's surface is associated with a material property and a predicted contact response. Large-scale pretraining on synthetic data generated by physics simulations has enabled the learning of such dense correspondences for rigid and articulated objects, but extending this to deformable and fluid materials remains an open challenge. The PhysAlign framework, by enforcing physics coherence during image-to-video generation through feature and 3D representation alignment, demonstrates that generative objectives can serve as a powerful inductive bias for learning these correspondences without requiring exhaustive manual annotation [8]. The crucial insight is that forcing a generative model to produce video frames that obey physical laws, as measured by a 3D-aware discriminator or a physics loss, creates a representation that implicitly aligns appearance, motion, and geometry in a physically meaningful way.

High-level conceptual alignment between language and 3D geometry poses a different set of challenges. A phrase such as "grasp the mug by its handle" presupposes that the system can parse the functional anatomy of the mug in three dimensions. This requires mapping linguistic concepts to geometric regions defined not by low-level shape features alone but by their

functional roles in anticipated physical interactions. Recent approaches embed language and 3D point clouds into shared spaces and train on paired description-geometry data, yet such data are scarce and biased toward common household objects from specific cultural contexts. The resulting alignment can break down when robots encounter objects or interaction styles outside the training distribution, raising serious fairness and robustness concerns that will be addressed in Section 6. Nevertheless, the architectural principle stands: by anchoring all modalities in a single, physically referenced 3D frame, the system can leverage geometric reasoning as a universal medium for cross-modal translation, reducing the need for brittle direct mappings between language and low-level motor commands.

5. Embodied Interaction and Robot Deployment Trade-offs

Translating a physically grounded multimodal world model from a research prototype into a reliable deployed robotic system exposes a constellation of trade-offs that span hardware, software engineering, and operational policy. Latency is perhaps the most unforgiving constraint. In contact-rich tasks such as peg insertion or deformable object manipulation, control loops must operate at hundreds of hertz, yet a full update of a high-fidelity 3D world model may require tens to hundreds of milliseconds even on modern GPU accelerators. System architects must therefore deploy hierarchical architectures in which a fast, lower-fidelity reactive controller stabilizes the immediate interaction while a slower, model-based planner updates higher-level goals. The division of labor between these tiers directly depends on the granularity of the 3D representation: if the world model maintains a dense reconstruction of the scene, the reactive controller can be relatively simple, querying local geometry for collision avoidance; if the world model is coarser, the reactive layer must shoulder more responsibility and may itself require proprioceptive learned models that operate at high frequency. This interplay creates a durability challenge, as rapid control often relies on assumptions that higher-level updates eventually reconcile, and any violation can cascade into catastrophic failure.

Simulation-to-real transfer remains a central deployment hurdle despite extensive research on domain randomization [6] and system identification. World models that align 3D representations across simulation and reality offer a principled path forward: by ensuring that simulated and real observations project into compatible 3D spaces, the learned dynamics and policies can be expected to transfer with less degradation. However, the alignment itself can introduce brittleness if the 3D reconstruction pipeline fails under real-world conditions such as motion blur, specular reflections, or transparent objects—phenomena that are notoriously difficult to simulate faithfully. This suggests a design principle of “graceful degradation,” wherein the system explicitly models its own uncertainty over the 3D alignment and falls back to safer, more conservative behavioral modes when confidence is low. Probabilistic world models that maintain distributions over object poses and shapes, rather than point estimates, are a step in this direction but dramatically increase the computational burden.

The choice of platform and sensor suite also imposes structural constraints on the world model. A mobile manipulator equipped with wrist-mounted RGB-D cameras, a head-mounted panoramic camera, and joint torque sensors must fuse data over multiple coordinate frames, each with its own latency and error profile. The world model’s internal 3D representation must accommodate these heterogeneous sensors and gracefully handle failures of any single modality. Redundancy provides robustness but escalates cost and energy consumption. The sustainability implications of such sensor-heavy designs cannot be ignored, particularly when deployment is envisioned at scale across fleets of robots operating continuously. Energy-

efficient neural architectures, quantized inference engines, and event-driven sensors that only trigger computation when significant changes occur are all components of a sustainable deployment strategy. Moreover, the embodied nature of interaction means that data collection for continual world model improvement is inherently expensive and must be prioritized; training runs that require weeks of GPU time on static datasets represent a significant carbon footprint that must be justified by commensurate gains in safety and capability [22].

Model updating in the field introduces further governance and engineering complexity. A robot that learns from its own physical interactions will inevitably collect data that reflect the environment it inhabits, including human behaviors, private spaces, and potentially hazardous scenarios. The world model, by encoding this data into its 3D representations and learned dynamics, becomes a repository of sensitive information that may be susceptible to reconstruction attacks. Privacy-preserving techniques such as federated learning over robot fleets, differential privacy during training, and on-device learning with data minimization are crucial but must be implemented without undermining the real-time performance and physical grounding that define the system's core purpose. This tension between data utility and privacy is particularly acute for physically grounded models, because physical interaction data is inherently identifying—gait patterns, household layouts, and manipulation styles can all serve as biometric or contextual signatures.

6. Governance, Fairness, and Sustainability Implications

The integration of physically grounded world models into robots that operate in human environments is not solely a technical challenge; it is a socio-technical endeavor that requires robust governance frameworks, proactive fairness engineering, and a commitment to environmental sustainability. The behaviors of embodied robots are shaped by the data used to train their world models and the physical assumptions baked into their alignment procedures. If the 3D representations that underpin object understanding are learned predominantly from data collected in affluent, Western kitchens, for example, the robot may fail to recognize or safely manipulate utensils and food items common in other cultural contexts. This is a form of representational bias that extends classical algorithmic fairness concerns into the physical domain, where failures are not merely inconvenient but can cause property damage or physical injury. Such biases can be compounded by the spectral biases of neural networks, which tend to learn low-frequency features first and may overlook the fine geometric details that distinguish functionally critical object parts across cultures [25].

Governance mechanisms for embodied AI must address the entire pipeline from data collection to physical deployment. Standards for benchmarking world model performance should include diversity of environments, object sets, and interaction styles, and should report performance disaggregated by demographic and geographic variables. Regulatory frameworks that currently focus on autonomous vehicles and industrial robots will need to evolve toward more general service robots operating in private spaces, where expectations of safety and privacy are elevated. The fact that world models can be used to generate plausible simulations of real environments raises questions about consent: a hotel deploying a robot that builds a detailed 3D map of its corridors and guest movements must navigate the legal and ethical boundaries of surveillance, even when the data remains on-device. Transparency mechanisms, such as model cards adapted for physically grounded representations that disclose the provenance of training data, the physical assumptions encoded in the simulator, and the known failure modes, are essential for building public trust and enabling meaningful oversight.

Fairness in physically grounded world models also manifests through the dynamics of power and control. Robots endowed with rich internal models of the physical world can anticipate human actions and preemptively adapt, potentially creating asymmetric relationships where the robot's capacity to model a person exceeds the person's capacity to model the robot. This asymmetry calls for design practices that make the robot's internal state interpretable to non-experts, perhaps through augmented reality overlays that reveal the robot's 3D understanding in situ. The decision-making policies that translate world model predictions into actions should be auditable and contestable, and the very criteria by which a world model is judged as "accurate" must be examined critically. An accuracy metric that prioritizes global geometric consistency over the safe handling of fragile, culturally significant objects embeds a value judgment that may not align with the priorities of all stakeholders. Multistakeholder co-design processes, involving not only engineers but also social scientists, legal scholars, and community representatives, are necessary to surface and reconcile these value tensions.

Sustainability considerations further complicate the governance picture. The carbon footprint of training large-scale multimodal world models is substantial, and the incremental energy cost of running such models on deployed robots, multiplied across fleets, can rival the emissions of other industrial sectors if not carefully managed. System designers must therefore balance the benefits of ever-larger models and denser 3D representations against their environmental costs, exploring efficiency innovations such as sparse 3D architectures, neuromorphic computing substrates, and lifelong learning strategies that amortize initial training over long operational lifetimes. The governance of physically grounded models should include environmental impact assessments akin to those required for physical infrastructure projects. A model that achieves a marginal improvement in manipulation success rate at the cost of a tenfold increase in energy consumption may not be a net positive for society, particularly if the tasks it automates were previously performed by humans with minimal carbon footprint. These are policy choices that must be deliberated openly, not left solely to engineering optimization loops.

7. Conclusion

Physically grounded multimodal world models, with 3D representation alignment as their organizing principle, represent a transformative yet demanding direction for embodied robot interaction. By unifying visual perception, natural language understanding, proprioceptive sensing, and physical simulation within coherent three-dimensional reference frames, these systems promise to endow robots with a deeper, more actionable understanding of the world. The system-level analysis presented in this paper reveals that the path toward this vision is strewn with intricate trade-offs: between geometric explicitness and learned flexibility, between computational tractability and long-horizon physical coherence, and between centralized cloud processing and distributed on-robot intelligence. Deployment realities tether these technical trade-offs to concrete engineering constraints, while the governance, fairness, and sustainability dimensions elevate them to matters of public concern. The advancement of physically grounded world models must therefore proceed as a multidisciplinary effort that integrates architectural innovation with ethical deliberation and lifecycle assessment. As the alignment of 3D representations across modalities matures from an algorithmic curiosity into a backbone of autonomous physical intelligence, it will be the care with which the broader socio-technical system is constructed that ultimately determines whether these machines become trustworthy partners or unpredictable agents in shared human spaces.

References

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

2. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (ICML).

3. Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2020). NeRF: Representing scenes as neural radiance fields for view synthesis. In Computer Vision – ECCV 2020 (pp. 405–421). Springer.

4. Kober, J., Bagnell, J. A., & Peters, J. (2013). Reinforcement learning in robotics: A survey. International Journal of Robotics Research, 32(11), 1238–1274.

5. Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., ... & Pascanu, R. (2018). Relational inductive biases, deep learning, and graph networks. arXiv preprint arXiv:1806.01261.

6. 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 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 23–30).

7. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., & Misra, I. (2023). ImageBind: One embedding space to bind them all. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 15180–15190).

8. Xiong, Z., Song, Y., He, L., Xiong, W., Yuan, Y., Qiao, F., & Jacobs, N. (2026). PhysAlign: Physics-Coherent Image-to-Video Generation through Feature and 3D Representation Alignment. arXiv preprint arXiv:2603.13770.

9. Hafner, D., Lillicrap, T., Norouzi, M., & Ba, J. (2020). Dream to control: Learning behaviors by latent imagination. In International Conference on Learning Representations (ICLR).

10. Kerbl, B., Kopanas, G., Leimkühler, T., & Drettakis, G. (2023). 3D Gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics, 42(4), 1–14.

11. Janner, M., Du, Y., Tenenbaum, J. B., & Levine, S. (2022). Planning with diffusion for flexible behavior synthesis. In Proceedings of the 39th International Conference on Machine Learning (ICML).

12. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., ... & Batra, D. (2019). Habitat: A platform for embodied AI research. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 9339–9347).

13. Todorov, E., Erez, T., & Tassa, Y. (2012). MuJoCo: A physics engine for model-based control. In 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 5026–5033).

14. Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the 34th International Conference on Machine Learning (ICML) (pp. 1126–1135).

15. Gupta, S., Davidson, J., Levine, S., Sukthankar, R., & Malik, J. (2019). Cognitive mapping and planning for visual navigation. International Journal of Computer Vision, 128(5), 1311–1330.

16. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 10684–10695).

17. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems, 33, 1877–1901.

18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, 30.

19. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (ICLR).

20. Barron, J. T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., & Srinivasan, P. P. (2021). Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 5855–5864).

21. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.

22. 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 (ACL) (pp. 3645–3650).

23. 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).

24. Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., ... & Mordatch, I. (2021). Decision transformer: Reinforcement learning via sequence modeling. In Advances in Neural Information Processing Systems, 34, 15084–15097.

25. Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F. A., ... & Courville, A. (2019). On the spectral bias of neural networks. In Proceedings of the 36th International Conference on Machine Learning (ICML) (pp. 5301–5310).