Grounding Intelligence in Movement

Melanie Segado
University of Pennsylvania
melanie.segado@gmail.com

Felipe Parodi
University of Pennsylvania

Jordan K. Matelsky
University of Pennsylvania

Michael L. Platt
University of Pennsylvania

Eva B. Dyer
University of Pennsylvania

Konrad P. Kording
University of Pennsylvania
kording@seas.upenn.edu

Abstract

Recent advances in machine learning have dramatically improved our ability to model language, vision, and other high-dimensional data, yet they continue to struggle with one of the most fundamental aspects of biological systems: movement. Across neuroscience, medicine, robotics, and ethology, movement is essential for interpreting behavior, predicting intent, and enabling interaction. Despite its core significance in our intelligence, movement is often treated as an afterthought rather than as a rich and structured modality in its own right. This reflects a deeper fragmentation in how movement data is collected and modeled, often constrained by task-specific goals and domain-specific assumptions. But movement is not domain-bound. It reflects shared physical constraints, conserved morphological structures, and purposeful dynamics that cut across species and settings. **We argue that movement should be treated as a primary modeling target for AI.** It is inherently structured and grounded in embodiment and physics. This structure, often allowing for compact, lower-dimensional representations (e.g., pose), makes it more interpretable and computationally tractable to model than raw, high-dimensional sensory inputs. Developing models that can learn from and generalize across diverse movement data will not only advance core capabilities in generative modeling and control, but also create a shared foundation for understanding behavior across biological and artificial systems. Movement is not just an outcome, it is a window into how intelligent systems engage with the world.

1 Introduction

Embodied intelligence is fundamentally rooted in movement. Across species and contexts, animals – including humans – demonstrate remarkable adaptability in their interactions with objects and conspecifics, often achieving complex motor behaviors with zero or few-shot learning. Nearly all animals use subtle movements of the face and body to signal intent and convey internal state. All processing of information in the brain, from vision to language, has only one goal: emitting better movement primitives, or tokens, that improve evolutionary fitness.

Biological movements are highly context dependent, as are their interpretations, so a useful model clearly needs to accommodate context inputs. Biological organisms excel at making inferences based on the in-context movements of others – recognizing actions, predicting trajectories, detecting disorders. As such, it is clear that proper biological movement understanding is possible. There is therefore every reason to believe that AI systems should be able to excel at this task as well.

The fact that current AI does not do well with biological movement can be seen as a special case of Moravec's paradox. The paradox tells us that these 'easy' motor tasks for animals are precisely the ones AI struggles with most. Interestingly, it's not just producing good movements that AI finds challenging (e.g., for robotics), but all of movement modeling and analysis is challenging, including video generation, and especially in contexts like medicine1.

Movement, unlike language and vision (or more recent foundation model targets like time series[43, 77, 30], graphs[46, 92, 75], particles[9], and neural data[5, 21]), does not have any single prototypical data type to model. Instead, there are many types of data relating to biological movement. This includes visual observations (videos), information about embodiment (physical form, body structure, biomechanical context), and trajectory-based timeseries (e.g., accelerometer data, computer-vision-derived poses). There are even neural and physiological signals recorded during movement (EMG, electrophysiology, EEG). While the surface-level data (pixels, accelerometer readings, neural signals) have varying levels of abstraction and dimensionality, they are often just different views or consequences of the same underlying movements. A good model of movement should be able to effectively use all these modalities.

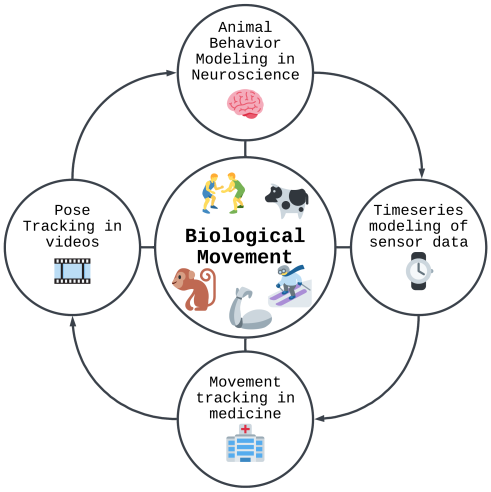
Animal Behavior Modeling in Neuroscience
Pose Tracking in videos
Biological Movement
Timeseries modeling of sensor data
Movement tracking in medicine

Figure 1: **Domains with biological movement at their core.** Biological movement is central to neuroscience, medicine, computer vision, and sensor modeling—each offering unique but interconnected perspectives on how movement is tracked, modeled, and understood.

Biological movement is a truly vexing problem: it lies at a unique intersection between highly structured data and high dimensional unstructured data. The body itself is extremely structured, with rough skeletal geometry conserved over hundreds of millions of years [19]. The skeleton and musculature within a species is highly, if not perfectly, conserved [24]. At the same time, the body around the skeleton is deformable as is the world that animals interact with. As such, biological movement lies in a uniquely interesting part of the world modeling space.

Meanwhile, existing multi-modal foundation models, such as those focusing on video, fail to model these causal and physical attributes despite having been trained on millions of videos of movement (e.g., producing videos with 180-degree rotations of the neck, or bodies floating up off the ground after a fall). The problem is so hard that extensive tests with current video generative models suggest that anything involving collisions between bodies leads to model failure. Current state of the art models fail catastrophically at generating simple high-fives or stumbling over a rock in our informal experiments.

Based on an analysis of the movement field, **our position is that the machine-learning community should intentionally build overarching models of movement, rather than hoping scaled video or task-specific systems will converge on the same capability**.

2 The landscape of existing movement models

There are many models through which movement is *partially* represented, such as video generators or time-series modeling[33, 27, 78]. Collectively, these models are promising as foundational elements for an ML-driven, overarching movement models, since each offers unique and complementary information. Consequently, a unifying framework must effectively integrate elements from these diverse approaches to achieve a more comprehensive understanding of movement.

2.1 Movement characterization & understanding

Pose estimation Precisely estimating pose is central to movement modeling. Historically this was only possible using marker-based motion capture, which was expensive and not feasible at scale or in natural settings[23, 73]. However, computer vision has completely revolutionized the field, making it possible to get precise pose estimates across species from hand-held video. Whereas earlier computer-vision approaches often required extensive, task-specific annotated datasets (e.g., for training tools like initial versions of DeepLabCut[54], or SLEAP[61]), foundation model approaches generalize to new species and contexts with minimal (or no) finetuning (e.g., SuperAnimal [97]). Human pose estimation specifically has experienced impressive gains with pre-trained models that work across different ages and complex poses, which was previously not possible (e.g., ViTPose [94, 95]). This includes the ability to detect and track entities across frames using object-detection models like RT-DETR [105]. For a long time better pose estimation was a blocker for movement modeling [73] but this has effectively been solved by computer vision.

Bodies are volumetric, and the volumetric properties have a big influence on how they move. Mesh recovery from video is improving and can provide information about body size and geometry (e.g., PromptHMR[87], SMPLer-X[12], SMAL[106], Penn Avian Mesh[7]), hand configurations (MANO[67]), facial expression (DiffusePoseTalk[76], clothing (ChatGarment[8]), and hair (DifFLocks[68]). Any model that aims to model movements of specific bodies needs to include all of these features.

The underlying skeletal structure and musculature of bodies defines *how* they can move. To capture this, the outputs of pose models (2D, 3D, and Mesh) are increasingly being integrated with biomechanical simulators (e.g., OpenSim[22], Myosuite[11], Mujoco[81]). Models that make explicit use of kinematic data result in estimates that are more aligned with the underlying skeletal structure (e.g., BioPose [41]). Biomechanical simulators provide complementary information and are important for training models that unify across representations.

Action recognition Actions are composed of sequences of poses, and correctly characterizing these sequences is a key challenge for movement models. Video-action models such as VideoMAE V2[86] already resolve at least 700 distinct human actions[13], yet their animal-centric peers still run an order of magnitude lower: MammalNet[15] tracks 12 canonical behaviors across 173 mammal taxa, while ChimpACT[51] pushes toward finer primate granularity but remains well under the 100-class mark—even when the footage is fully "in-the-wild." Video foundation models like VideoMAE V2 increasingly achieve robust multimodal understanding by learning shared or aligned representations across modalities (e.g., text, audio). Likewise, multimodal large language models like LLaVA[44] are increasingly being paired with video and used for action recognition (e.g., LLaVAction [96]).

Beyond video and text, sensor-based activity recognition aims to detect specific activities from wearable sensors such as Inertial Measurement Units (IMUs) or electromyography (EMG) signals. Doing so is critical for applications in human-computer interaction (e.g., for gesture-based wearable controllers [64] or adaptive prosthetics[59, 63]) and healthcare (e.g., rehabilitation [89]. Many models have been proposed for analyzing sensor data, either in isolation or in conjunction with video, including approaches that perform cross-modal prediction from video to sensor timeseries (e.g., [18], and benchmarks like emg2pose [71]). The importance of activity recognition extends beyond human applications. Animal activity recognition, leveraging both video and sensor data is increasingly vital for ecological studies [53, 58], conservation efforts, and applications such as livestock management in agriculture[40].

Ultimately, movement models will need to accommodate and integrate information from many different data streams, likely by combining merits of the growing number of multimodal frameworks (e.g., Act-ChatGPT[56], HumanOmni[104], HIS-GPT[103], Meta-transformer[102], Audiopalm[69], Imagebind[29], Omnibind[88], CoMP[16], M^{3}GPT[49]) to achieve increasingly nuanced understanding.

Limitations Pose estimators and action recognition models may be able to *describe* movements, but do not typically capture the underlying intent, the nuances of execution quality in a way meaningful for performance analysis or pathology, or the critical environmental and interaction context that gives movement its full meaning. They often struggle with the variability and unscripted nature of real-world actions, particularly those that are not easily categorized or segmented. Furthermore, current approaches often lack the layered understanding required to interpret the same pose sequence differently based on context, such as distinguishing a voluntary handshake from a tremor with diagnostic significance. This highlights the gap between merely describing movement kinematics or labeling actions and truly understanding the dynamic, contextual, and functional aspects of biological movement.

2.2 Generating movement

Generative Models The capacity of a model to generate novel, plausible movement serves as a potent indicator of its underlying representational quality. One influential line of research approaches movement generation analogously to language modeling, capitalizing on the compositional structure inherent in poses and actions. Models such as ChatPose [27] and MotionGPT[38] exemplify this by discretizing continuous movements into a vocabulary of motion "tokens." These are then synthesized using autoregressive strategies, akin to Generative Pre-trained Transformers (GPTs). A key advantage of this token-based paradigm is its natural alignment with textual descriptors, facilitating intuitive text-to-motion generation and enabling tasks like motion infilling (completing sequences between given tokens) and short-term forecasting (predicting subsequent tokens)[38].

A movement model should be able to predict what a human/animal's next movement will be, and where they will be at a future point in time. This is an exceptionally complex task, because movement is stochastic and depends on a variety of factors ranging from internal motivations to environmental drivers to physical constraints[62]. Models that predict solely based on past movements inevitably fail after very few time-steps [6, 57]. To accurately forecast movement, models need to account for goals and intent [25, 70].

Moreover, diffusion models[36] are increasingly applied to directly denoise text prompts into structured skeleton-based motion; the Human Motion Diffusion Model (MDM), for example, treats a 3D pose sequence as a trajectory to be denoised step-by-step into fluid, text-conditioned skeletons[80, 35]. Building on this, MotionDiffuse [100] incorporates foot-contact and diversity losses to produce physically plausible skeletal clips from free-form textual prompts. While these successes across both visual rendering and skeletal animation highlight the immense potential for high-fidelity movement generation, extending these capabilities to achieve nuanced, controllable, and holistic full-body articulated motion consistently presents an ongoing objective, suggesting a need to bridge current specialized methodologies.

Physics-based movement simulation Movements produce and respond to forces in the environment, so generated movements should do so as well. Progress on physics-informed movement generation has been enabled by biomechanical simulators, the curation of human motion datasets that include motion dynamics (torques and forces [90]), and also methods like force estimation from volumetric mesh. Generative models trained on motion dynamics data have been shown to replicate accurate human gait kinematics[78], and models that integrate a force loss generate more realistic meshes [99] showing that the integration of physics-grounded movement data improves real-world relevance.

Limitations Biological movement is extremely precise – in the case of some activities the difference between successful and unsuccessful movements on the order of millimeters and fractions of degrees in the spatial domain, and milliseconds in the temporal domain. Auto-regressive motion generators produce very smoothed, average representations of what specific movements *look like*, but fail to produce detailed actions, and will often produce unrealistic or incoherent results when filling in motions or forecasting. Mesh timeseries generation, as it currently stands, is not useful for applications that require precision.

Generative video remains imprecise, often misrepresenting bodily movements in ways that violate physical plausibility– even when models incorporate physics-based constraints. Recent systems such as Sora [10] or Google Veo 2 (now Veo 3) continue to struggle with producing physically realistic depictions of even simple actions, such as stumbling or coordinated interactions like playing the violin. Despite advances in visual fidelity, these models demonstrate a persistent gap in understanding the structured, biomechanical nature of movement.

2.3 Learning to move in world models

Reinforcement Learning (RL) and world models RL offers a powerful framework for training agents to perform complex movement tasks through interaction with an environment and the use of reward functions. In RL, agents learn to optimize their actions to achieve specific goals, such as navigating a terrain or manipulating an object.

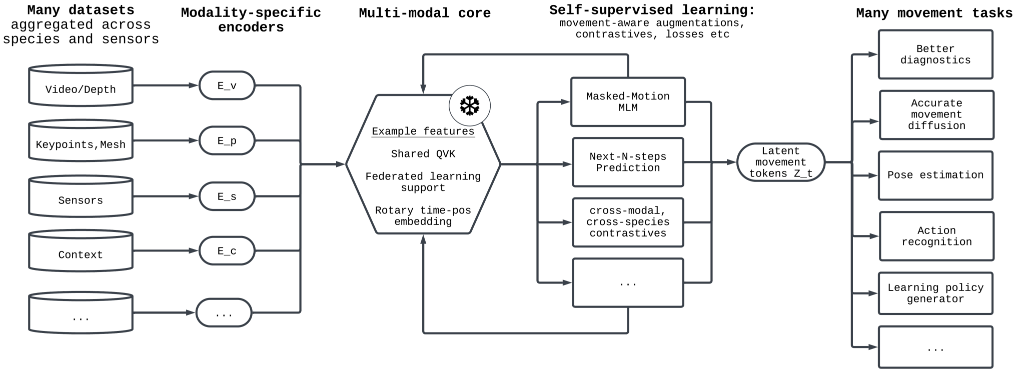

Figure 2: **Unifying framework for movement modeling across species and sensors.** The ML community has already developed all of these components. What's needed is coordinated effort to combine them into a purpose-built framework for learning movement features directly from aggregated movement data.

Complementary to RL, world models (e.g., NVIDIA Cosmos[2], Google Genie 2[60]) provide a way for agents to learn a model of their environment's dynamics, enabling them to plan and make decisions more effectively [33, 55, 34]. By learning to predict the consequences of their actions, agents can use world models to anticipate future states and select optimal movement strategies.

RL agents trained in world models exhibit a wide range of complex and adaptive behaviors [91], especially focused on how agents move in relation to their environments [74, 93]. For instance, world models have enabled the development of sophisticated controllers for legged robots (e.g., Boston Dynamics' Spot), allowing them to achieve agile and robust locomotion over challenging terrains[66, 42]. Similarly, RL and world models are being explored to create robots that can interact more naturally and effectively with humans, adapting their movements to social cues and user preferences[17].

Limitations Despite their potential, RL and world models face several limitations. One significant challenge lies in specifying precise goals, as RL fundamentally relies on well-defined goals and reward functions, which can be difficult to specify for the nuanced and multifaceted objectives inherent in complex, real-world movement tasks. Furthermore, the model accuracy of world models is crucial for effective planning, but learning a sufficient model of the environment is often intractable, and inaccuracies can lead to suboptimal or even dangerous behavior (see also: "sim-to-real gap", [84, 98, 55]). Learning realistic constraints on movement also remains a substantial hurdle despite the integration of RL with musculoskeletal models[72].

Furthermore, the resulting RL policies, even when learned within sophisticated world models, often fail to capture or transfer knowledge to inherently different movement contexts, such as distinct developmental stages or other species. For instance, an RL agent trained to emulate adult human locomotion will not inherently understand or replicate an infant's fidgeting, thereby limiting its use in developmental modeling even though the two motions are deeply biologically linked. Similarly, the movement strategies learned by such an agent are typically not informed by, nor easily adaptable to, the diverse locomotion patterns of other animal species without extensive, separate retraining. Traditionally, this retraining involves a complete redefinition of the agent's morphology and reward structures within the RL framework. [3] This task-specific nature means that the movement understanding derived from many RL and world model systems remains highly specialized, posing a hurdle for creating models with truly general or foundational insights into movement.

3 Towards a unifying framework for movement modeling

Despite impressive improvements in the past few years, existing approaches do not model movement as a unified concept, nor are they precise enough for real-world use. We propose that it is therefore necessary to model movement itself using a unifying framework, not just as a product of video or sensor data, but as a structured, embodied, and generalizable domain (**Fig. 2**).

The resulting models should satisfy the following:

1. **Cross-modal**, integrating signals such as video, IMU, EMG, GPS, neural data, and facial movements.
2. **Biomechanically and physically constrained** across bodies and species.
3. **Context-aware**, capable of grounding motion in tasks, environments, and goals
4. **Generalizable** across settings, supporting low-shot adaptation in clinical, ecological, and interactive applications.

These are not simply larger action recognizers or more detailed movement generators. They are overarching models that treat movement as a fundamental signal, like language or vision, capable of supporting a wide range of scientific and applied domains. Accomplishing this will require coordinated effort to combine existing datasets and fill in gaps, pre-train using identity/privacy preserving approaches that prioritize movement realism, and evaluate on benchmarks that are tailored to real-world impact.

Below, we outline the steps that are needed to model movement, and the contributions that the ML community can make to enable it.

3.1 Step 1: Aggregate a movement data pile

A first step towards a unifying framework is acknowledging the sheer volume and diversity of existing movement data, spanning multiple species, contexts, and sensor modalities. High-fidelity motion capture (MoCap) systems continue to provide gold-standard 3D kinematic data, with established datasets like AMASS ([52]) and Human3.6M[37] being complemented by newer collections capturing nuanced whole-body kinematics (e.g., [28]) and rich human-object interactions (e.g.,[48]). The ubiquity of video recording has led to an explosion of data from both controlled and in-the-wild settings. Efforts like Motion-X ([45]) are unifying legacy datasets and incorporating new web-scraped videos to create internet-scale resources of human motion meshes.

For animal studies, datasets such as MammAlps ([83]) leverage camera traps for large-scale wildlife monitoring. Furthermore, egocentric and multimodal capture approaches are providing rich, first-person perspectives; EgoBody[101] pairs head-mounted RGB-D video with eye-gaze and full-body meshes in social scenes, while Nymeria[50]) utilizes advanced sensors like Project Aria[26] alongside third-person cameras for human activity capture in natural environments. Complementing these are extensive logs from low-cost wearable sensors, exemplified by CAPTURE-24 [14] with over 2,500 hours of free-living wrist-accelerometer data. Physiological data streams, such as the HD-sEMG "Hyser" bank featuring EMG data for hand gestures[39], add another layer of detail. Domain-specific databases are also being contributed, from industrial datasets monitoring worker movements to agricultural datasets like MmCows[85], which records synchronized UWB tags, inertial sensors, climate logs, and extensive video footage of dairy cattle. This vast and varied "data pile" underscores the opportunity and challenge for developing comprehensive movement models.

Open challenges the ML community can address:

Standardizing conventions: We have lots of data, but combining it is extremely complicated. Developing standardized dataset conventions (similar to the BIDS format for neuroimaging data[4]), and building data-loaders that can accept flexible input types and translate into a standard convention (e.g., building on efforts like OpenMMLab's keypoint converter, or HumanML3D[32]), would enable aggregation across datasets/sensors/species.

Curating datasets to fill in gaps: Movement datasets need to include more context. This includes pairing video/sensor data with audio, text-based action labels, medical scores/diagnoses, object-interactions, social relationships, time of day, ambient temperature, and physiological measures. While a growing number of multimodal datasets exist, there are still gaps that need to be filled. A priority here is datasets that include both movement data (from one or more sensor streams) and rich, annotated context.

3.2 Step 2: Pretrain a multimodal backbone

A flexible, general purpose backbone for movement tasks would be a game-changer for the field of movement science, reducing the time and effort needed to obtain latent features of movement that are useful for downstream tasks. However, for this to have impact, training objectives need to be aligned with the needs of end users. Context needs to be considered, fine-grained details must be preserved, and biomechanical/physical constraints must be strictly respected.

Open challenges the ML community can address:

Design architectures for multimodal contextual integration. Contextual data needs to be included, but the model should not attempt to learn everything about the world – in fact, world models already attempt this – but rather integrate contextual data streams (e.g., environmental sounds, object identities) as ’co-teachers’ (akin to approaches in models like CLIP). This will allow them to inform the learned movement latents without having the model attempt to learn everything about the ’world’, thereby duplicating efforts in building comprehensive world models.

Make federated learning easy. Much of the interesting, medical movement data that exists can only be analyzed on hospital servers (e.g., videos of newborn infants in the NICU). Federated learning is currently complicated and inaccessible. There is a huge opportunity for the ML community to build user-friendly, open source federated learning training pipelines so models can be trained on sensitive medical data.

Develop movement-aware augmentations Standard image augmentations degrade critical features in biological movement data. Noise corrupts subtle signs like tremor, and left–right flips erase lateralized deficits. Domain-informed augmentations are essential, and the ML community should work closely with experts to formalize and encode their expertise.

Design movement-aware loss functions: Realism should be prioritized. This should include adopting losses for joint-angle violations, non-physical accelerations, bone-length inconsistencies, implausible forces, and contact violations to improve realism of generated model outputs[99].

3.3 Step 3: Evaluate on high-impact use-cases

The true measure of movement models’ success should be their practical utility and alignment with societal benefit, especially given the significant resources required for their training. Evaluation must encompass a broader set of critical metrics. This includes assessing generalization and transferability through cross-species and cross-task benchmarks, as well as validating biomechanical and physical realism using metrics explicitly geared to diagnostic needs for movement disorders.

Model performance evaluation should also consider privacy robustness, ensuring reliable performance even when employing privacy-preserving methods.

Open challenges the ML community can address:

Develop outcome-aligned benchmarks. In many domains, the utility of a movement model is binary: either it captures the core features of movement — or it fails. Near-miss representations are diagnostically meaningless. Evaluation must reflect this and reward models that produce functionally valid latent representations. Specifically, benchmarks could test causal understanding through physical perturbation recovery or counterfactual movement generation under anatomical or pathological constraints.

Benchmark cross-domain generalization. Movement models should be tested across species, body plans, and tasks. Can a model trained on infant reaching also represent gait in Parkinson’s patients or agility in non-human animals? Cross-domain generalization is key to scalability. Models must demonstrate they can predict compensatory patterns, preserve diagnostic asymmetries, and correlate with clinical outcomes – not merely reproduce statistical regularities in training data.

Incentivize robustness to privacy constraints. Evaluation protocols should reward models that perform well under privacy preserving methods, ensuring that real-world deployment in healthcare and research is feasible and ethical.

4 Societal stakes & Ethical risks

Movement models must be built with explicit attention to risks to privacy, algorithmic bias, and personal safety. Movement data—whether drawn from medical settings, wearables, social behavior, or animal tracking—poses unique risks. It is inherently biometric, often longitudinal, and deeply entangled with questions of identity, ability, health, and physical safety. Modeling it at scale, with generative and predictive capability would be incredibly beneficial—but also opens the door to misuse. These concerns are echoed by many in the AI community who have voiced increasing unease about the centralization of power, the opacity of learned representations, and the potential for social harm when AI systems are developed without robust guardrails. Without appropriate safeguards, movement models could be co-opted for surveillance, behavioral scoring, or automated exclusion, thereby undermining the very communities they are intended to help.

Risks to privacy Movement data, by its nature, often contains identifiable biometric information such as gait patterns, posture, and even subtle facial expressions captured in video. Recent work has shown that diffusion and transformer-based generative models are prone to memorization of training data, including identifiable faces, gestures, and sequences [31]. In the context of movement modeling, this raises concerns of biometric re-identification, even when datasets are nominally anonymized.

Bias, algorithmic phrenology The misuse of movement data to infer sensitive traits like health conditions or psychological states carries a significant risk of bias and discrimination in areas such as employment, insurance, and surveillance. Furthermore, models trained on narrowly defined datasets risk performing algorithmic phrenology, reviving discredited attempts to link physical traits to the nonphysical. There is also a critical concern that movement models could unintentionally marginalize individuals with disabilities by defining 'normal' or 'typical' movements in a way that excludes or misinterprets the movements of those with different physical abilities.

Risk mitigation To mitigate these risks, we advocate for a design approach rooted in responsible AI development principles. This means building privacy-respecting pipelines (e.g., federated or split learning), embedding morphological and demographic variation during training, and ensuring transparency around model intent, capabilities, and limitations. It also means engaging with domain experts and impacted communities throughout the model development process, not after deployment.

Concerns about potential harms underscore the importance of intentional design. Movement models have the potential to improve clinical care, conservation, assistive technology, and performance science—but only if they are aligned with values of equity, safety, and scientific rigor. By embedding the lessons of AI safety from the outset, we can shape this new modeling paradigm toward outcomes that benefit society as a whole.

5 Alternative views

Scaling existing models will solve this We are proposing that purpose-built movement models are necessary to advance the field, but one could argue that same progress in modeling movement could simply be achieved by continuously scaling existing, specialized models. This perspective suggests that the increasing realism and familiarity of current approaches—from generative video models that treat movement as pixels to simulators that require ground-up environment construction, or RL agents that demand extensive goal specifications—could eventually converge onto something like a overarching movement model if simply scaled further. The increasing visual fidelity and physical realism achieved by these scaled systems might make this an appealing proposition.

However, we argue that scaling existing approaches will not solve the fundamental challenges inherent in truly understanding and generating sophisticated biological movement. Scaling alone fails to address core issues such as physical realism, interpretability, and embodiment unless curated constraints are applied. Critically, these models, even at larger scales, will struggle to tackle complex and interesting questions unique to movement science, such as cross-species comparisons and translations, developmental changes, and the subtle manifestations of neurological disorders.

Furthermore, current modeling paradigms are largely species- and modality-specific, meaning they would not seamlessly scale to include diverse medical or animal data as a unified whole unless such integration is intentionally designed. Duplicating specialized pipelines across various domains is not only costly and inefficient but also precludes the emergence of generalizable movement intelligence. A dedicated, unified framework for movement modeling, which intentionally learns fundamental properties of movement across all data types and species, is essential to move beyond the limitations of scaled, fragmented approaches.

Task specific models are better The assertion that a movement model trained on all movement data across tasks and species will outperform task-specific models derives from the massive success of foundation models in language and vision. Admittedly, earlier attempts to create broad "generalist" models (e.g., GATO[65], Unified IO[47]) yielded mixed results, often performing on par with, or sometimes worse than, existing specialized models.

This will not be the case for an overarching movement model. Unlike previous generalist models which were trained on disparate tasks often lacking a cohesive underlying principle (evidenced by the lack of transfer between skills like image captioning and RL policy making), a dedicated movement model would focus on a coherent set of tasks all reliant on precise movement representations. Such a model is therefore well-positioned to benefit from increasingly diverse training data, leveraging strong indications that proficiency in one movement task frequently enhances performance in others. Conceptually, this approach aligns more closely with highly successful architectures like GPT-4[1], Gemini[79], and Llama[82], which excel by deeply modeling specific, rich domains, rather than with the earlier, less potent generalist AIs.

Beyond achieving better performance, a main advantage to training a flexible movement 'backbone' would be to streamline development efforts and increase adoption outside the field of ML. The fragmented nature of movement has resulted in widespread duplication of effort, particularly when applied to real-world problems. Each researcher, doctor, or engineer has to weave together a patchwork of different libraries, each requiring their own packages and data-formatting pipeline, many of which are never supported beyond the conference they were submitted to. Hundreds of new models with incremental advances are published each year, making it near-impossible for researchers outside the field to even know which models are available to them. While initiatives like OpenMMLab[20] have significantly lowered the barrier to entry, the workflow still stands in stark contrast to the ease with which non-ML scientists can now analyze and generate text.

6 Conclusion

Movement modeling matters. Developing a model that can recognize, understand, forecast, and generate movement would transform our understanding of movement across species, transform our ability to recognize early signs of disorder and develop precision rehabilitation strategies, and advance the development of robots that predict and synchronize with our movements.

Here, we argued that overarching models of movement should be developed as a distinct and essential focus for ML systems, and present a framework to unify across species and sensors. We argue that existing approaches have only approximated some elements of movement, and that none have truly modeled movement as a unified concept. The widespread success of task-general foundation models in various other domains, and the growing momentum to develop movement models across all domains, makes this both an ambitious goal but also a logical next step.

There are massive amounts of movement data that have yet to be aggregated, there exist flexible architectures that can handle diverse data inputs, and there exists scalable compute on which to train. Every component that is needed to model movement exists – all that is needed is coordinated effort across domains.

The machine learning community is uniquely positioned to take on this challenge, recognizing movement as a foundational domain with the potential for broad scientific impact, clinical breakthroughs, and real-world applications. By uniting efforts across disciplines and data sources, we can build models that not only reflect how organisms move, but also how they act, adapt, and learn. Embracing movement as a core modality opens the door to a deeper understanding of intelligence, understood not only as an abstract computational process but as something grounded in physical interaction, goal-directed behavior, and adaptation to a dynamic world.

References

[1] Josh Achiam et al. “Gpt-4 technical report”. In: *arXiv preprint arXiv:2303.08774* (2023).

[2] Niket Agarwal et al. “Cosmos world foundation model platform for physical ai”. In: *arXiv preprint arXiv:2501.03575* (2025).

[3] Shuang Ao et al. “Curriculum reinforcement learning via morphology-environment co-evolution”. In: *arXiv preprint arXiv:2309.12529* (2023).

[4] Stefan Appelhoff et al. “MNE-BIDS: Organizing electrophysiological data into the BIDS format and facilitating their analysis”. In: *Journal of Open Source Software* 4.44 (2019), p. 1896.

[5] Mehdi Azabou et al. “A unified, scalable framework for neural population decoding”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 44937–44956.

[6] Mehdi Azabou et al. “Relax, it doesn’t matter how you get there: A new self-supervised approach for multi-timescale behavior analysis”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 28491–28509.

[7] Marc Badger et al. “3D bird reconstruction: a dataset, model, and shape recovery from a single view”. In: *European conference on computer vision*. Springer. 2020, pp. 1–17.

[8] Siyuan Bian et al. “ChatGarment: Garment Estimation, Generation and Editing via Large Language Models”. In: *arXiv preprint arXiv:2412.17811* (2024).

[9] Joschka Birk, Anna Hallin, and Gregor Kasieczka. “OmniJet-α: the first cross-task foundation model for particle physics”. In: *Machine Learning: Science and Technology* 5.3 (2024), p. 035031.

[10] Tim Brooks et al. “Video generation models as world simulators”. In: *OpenAI Blog* 1 (2024), p. 8.

[11] Vittorio Caggiano et al. “MyoSuite–A contact-rich simulation suite for musculoskeletal motor control”. In: *arXiv preprint arXiv:2205.13600* (2022).

[12] Zhongang Cai et al. “Smpler-x: Scaling up expressive human pose and shape estimation”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 11454–11468.

[13] Joao Carreira et al. “A short note on the kinetics-700 human action dataset”. In: *arXiv preprint arXiv:1907.06987* (2019).

[14] Shing Chan et al. “CAPTURE-24: A large dataset of wrist-worn activity tracker data collected in the wild for human activity recognition”. In: *Scientific Data* 11.1 (2024), p. 1135.

[15] Jun Chen et al. “Mammalnet: A large-scale video benchmark for mammal recognition and behavior understanding”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 13052–13061.

[16] Yitong Chen et al. “CoMP: Continual Multimodal Pre-training for Vision Foundation Models”. In: *arXiv preprint arXiv:2503.18931* (2025).

[17] Marvin H Cheng, Po-Lin Huang, and Hao-Chuan Chu. “Bio-Inspired Motion Emulation for Social Robots: A Real-Time Trajectory Generation and Control Approach”. In: *Biomimetics* 9.9 (2024), p. 557.

[18] Mia Chiquier and Carl Vondrick. “Muscles in action”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 22091–22101.

[19] MI Coates. “The Devonian tetrapod Acanthostega gunnari Jarvik: postcranial anatomy, basal tetrapod interrelationships and patterns of skeletal evolution”. In: *Earth and Environmental Science Transactions of the Royal Society of Edinburgh* 87.3 (1996), pp. 363–421.

[20] MMHuman3D Contributors. *Openmmlab 3d human parametric model toolbox and benchmark*. 2021.

[21] Wenhui Cui et al. “Neuro-gpt: Towards a foundation model for eeg”. In: *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2024, pp. 1–5.

[22] Scott L Delp et al. “OpenSim: open-source software to create and analyze dynamic simulations of movement”. In: *IEEE transactions on biomedical engineering* 54.11 (2007), pp. 1940–1950.

[23] Yann Desmarais et al. “A review of 3D human pose estimation algorithms for markerless motion capture”. In: *Computer Vision and Image Understanding* 212 (2021), p. 103275.

[24] Rui Diogo and Virginia Abdala. *Muscles of vertebrates: comparative anatomy, evolution, homologies and development*. CRC Press, 2010.

[25] Markos Diomataris et al. “WANDR: Intention-guided human motion generation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 927–936.

[26] Jakob Engel et al. “Project aria: A new tool for egocentric multi-modal ai research”. In: *arXiv preprint arXiv:2308.13561* (2023).

[27] Yao Feng et al. “Chatpose: Chatting about 3d human pose”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, pp. 2093–2103.

[28] Saeed Ghorbani et al. “MoVi: A large multi-purpose human motion and video dataset”. In: *Plos one* 16.6 (2021), e0253157.

[29] Rohit Girdhar et al. “Imagebind: One embedding space to bind them all”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 15180–15190.

[30] Mononito Goswami et al. “Moment: A family of open time-series foundation models”. In: *arXiv preprint arXiv:2402.03885* (2024).

[31] Xiangming Gu et al. “On memorization in diffusion models”. In: *arXiv preprint arXiv:2310.02664* (2023).

[32] Chuan Guo et al. “Generating diverse and natural 3d human motions from text”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 5152–5161.

[33] David Ha and Jürgen Schmidhuber. “World models”. In: *arXiv preprint arXiv:1803.10122* (2018).

[34] Danijar Hafner et al. “Mastering diverse domains through world models”. In: *arXiv preprint arXiv:2301.04104* (2023).

[35] Xiaoxu Han, Hegen Xu, and Huilling Sun. “Motion Diffusion Model for Long Motion Generation”. In: *2024 5th International Conference on Machine Learning and Computer Application (ICMLCA)*. IEEE. 2024, pp. 603–608.

[36] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.

[37] Catalin Ionescu et al. “Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments”. In: *IEEE transactions on pattern analysis and machine intelligence* 36.7 (2013), pp. 1325–1339.

[38] Biao Jiang et al. “Motiongpt: Human motion as a foreign language”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 20067–20079.

[39] Xinyu Jiang et al. “Open access dataset and toolbox of high-density surface electromyogram recordings”. In: *PhysioNet* (2021).

[40] Natasa Kleanthous et al. “Deep transfer learning in sheep activity recognition using accelerometer data”. In: *Expert Systems with Applications* 207 (2022), p. 117925.

[41] Farnoosh Koleini et al. “BioPose: Biomechanically-accurate 3D Pose Estimation from Monocular Videos”. In: *arXiv preprint arXiv:2501.07800* (2025).

[42] Hang Lai et al. “World Model-based Perception for Visual Legged Locomotion”. In: *arXiv preprint arXiv:2409.16784* (2024).

[43] Yuxuan Liang et al. “Foundation models for time series analysis: A tutorial and survey”. In: *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*. 2024, pp. 6555–6565.

[44] Bin Lin et al. “Video-llava: Learning united visual representation by alignment before projection”. In: *arXiv preprint arXiv:2311.10122* (2023).

[45] Jing Lin et al. “Motion-x: A large-scale 3d expressive whole-body human motion dataset”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 25268–25280.

[46] Jiawei Liu et al. “Towards graph foundation models: A survey and beyond”. In: *arXiv preprint arXiv:2310.11829* (2023).

[47] Jiasen Lu et al. “Unified-io: A unified model for vision, language, and multi-modal tasks”. In: *arXiv preprint arXiv:2206.08916* (2022).

[48] Jiaxin Lu et al. “HUMOTO: A 4D Dataset of Mocap Human Object Interactions”. In: *arXiv preprint arXiv:2504.10414* (2025).

[49] Mingshuang Luo et al. “M ^{3} GPT: An Advanced Multimodal, Multitask Framework for Motion Comprehension and Generation”. In: *arXiv preprint arXiv:2405.16273* (2024).

[50] Lingni Ma et al. “Nymeria: A massive collection of multimodal egocentric daily motion in the wild”. In: *European Conference on Computer Vision*. Springer. 2024, pp. 445–465.

[51] Xiaoxuan Ma et al. “Chimpact: A longitudinal dataset for understanding chimpanzee behaviors”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 27501–27531.

[52] Naureen Mahmood et al. “AMASS: Archive of motion capture as surface shapes”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 5442–5451.

[53] Axiu Mao et al. “FedAAR: A novel federated learning framework for animal activity recognition with wearable sensors”. In: *Animals* 12.16 (2022), p. 2142.

[54] Alexander Mathis et al. “DeepLabCut: markerless pose estimation of user-defined body parts with deep learning”. In: *Nature neuroscience* 21.9 (2018), pp. 1281–1289.

[55] Yutaka Matsuo et al. “Deep learning, reinforcement learning, and world models”. In: *Neural Networks* 152 (2022), pp. 267–275.

[56] Yuto Nakamizo and Keiji Yanai. “Act-ChatGPT: Introducing Action Features into Multimodal Large Language Models for Video Understanding”. In: *International Conference on Pattern Recognition*. Springer. 2024, pp. 250–265.

[57] Aymeric Orhan et al. “Combining Model-based and Data-based approaches for online predictions of human trajectories”. In: *2024 10th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob)*. IEEE. 2024, pp. 1764–1771.

[58] Ryoma Otsuka et al. “Exploring deep learning techniques for wild animal behaviour classification using animal-borne accelerometers”. In: *Methods in Ecology and Evolution* 15.4 (2024), pp. 716–731.

[59] Nawadita Parajuli et al. “Real-time EMG based pattern recognition control for hand prostheses: A review on existing methods, challenges and future implementation”. In: *Sensors* 19.20 (2019), p. 4596.

[60] J Parker-Holder et al. “Genie 2: A large-scale foundation world model”. In: *URL: https://deepmind. google/discover/blog/genie-2-a-large-scale-foundation-world-model* (2024).

[61] Talmo D Pereira et al. “SLEAP: A deep learning system for multi-animal pose tracking”. In: *Nature methods* 19.4 (2022), pp. 486–495.

[62] Giovanni Pezzulo et al. “Generating meaning: active inference and the scope and limits of passive AI”. In: *Trends in Cognitive Sciences* 28.2 (2024), pp. 97–112.

[63] Hari Prasanth et al. “Wearable sensor-based real-time gait detection: A systematic review”. In: *Sensors* 21.8 (2021), p. 2727.

[64] Ctrl-labs at Reality Labs et al. “A generic noninvasive neuromotor interface for human-computer interaction”. In: *Biorxiv* (2024), pp. 2024–02.

[65] Scott Reed et al. “A generalist agent”. In: *arXiv preprint arXiv:2205.06175* (2022).

[66] Robert Riener, Luca Rabezzana, and Yves Zimmermann. “Do robots outperform humans in human-centered domains?” In: *Frontiers in Robotics and AI* 10 (2023), p. 1223946.

[67] Javier Romero, Dimitrios Tzionas, and Michael J Black. “Embodied hands: Modeling and capturing hands and bodies together”. In: *arXiv preprint arXiv:2201.02610* (2022).

[68] Radu Alexandru Rosu et al. “DiffLocks: Generating 3D Hair from a Single Image using Diffusion Models”. In: *arXiv preprint arXiv:2505.06166* (2025).

[69] Paul K Rubenstein et al. “Audiopalm: A large language model that can speak and listen”. In: *arXiv preprint arXiv:2306.12925* (2023).

[70] Andrey Rudenko et al. “Human motion trajectory prediction: A survey”. In: *The International Journal of Robotics Research* 39.8 (2020), pp. 895–935.

[71] Sasha Salter et al. “emg2pose: A large and diverse benchmark for surface electromyographic hand pose estimation”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 55703–55728.

[72] Pierre Schumacher et al. “Dep-rl: Embodied exploration for reinforcement learning in overactuated and musculoskeletal systems”. In: *arXiv preprint arXiv:2206.00484* (2022).

[73] Nidhi Seethapathi et al. “Movement science needs different pose tracking algorithms”. In: *arXiv preprint arXiv:1907.10226* (2019).

[74] Merkourios Simos, Alberto Silvio Chiappa, and Alexander Mathis. “Reinforcement learning-based motion imitation for physiologically plausible musculoskeletal motor control”. In: *arXiv preprint arXiv:2503.14637* (2025).

[75] Li Sun et al. “Riemanngfm: Learning a graph foundation model from riemannian geometry”. In: *Proceedings of the ACM on Web Conference 2025*. 2025, pp. 1154–1165.

[76] Zhiyao Sun et al. “Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models”. In: *ACM Transactions on Graphics (TOG)* 43.4 (2024), pp. 1–9.

[77] Sabera Talukder, Yisong Yue, and Georgia Gkioxari. “Totem: Tokenized time series embeddings for general time series analysis”. In: *arXiv preprint arXiv:2402.16412* (2024).

[78] Tian Tan et al. “GaitDynamics: A Generative Foundation Model for Analyzing Human Walking and Running”. In: *Research Square* (2025), rs–3.

[79] Gemini Team et al. “Gemini: a family of highly capable multimodal models”. In: *arXiv preprint arXiv:2312.11805* (2023).

[80] Guy Tevet et al. “Human motion diffusion model”. In: *arXiv preprint arXiv:2209.14916* (2022).

[81] Emanuel Todorov, Tom Erez, and Yuval Tassa. “Mujoco: A physics engine for model-based control”. In: *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE. 2012, pp. 5026–5033.

[82] Hugo Touvron et al. “Llama: Open and efficient foundation language models”. In: *arXiv preprint arXiv:2302.13971* (2023).

[83] Devis Tuia. “MammAlps: A multi-view video behavior monitoring dataset of wild mammals in the Swiss Alps”. In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2025*. 2025.

[84] Keyon Vafa et al. “Evaluating the world model implicit in a generative model”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 26941–26975.

[85] Hien Vu et al. “MmCows: A Multimodal Dataset for Dairy Cattle Monitoring”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 59451–59467.

[86] Limin Wang et al. “Videomae v2: Scaling video masked autoencoders with dual masking”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 14549–14560.

[87] Yufu Wang et al. “PromptHMR: Promptable Human Mesh Recovery”. In: *arXiv preprint arXiv:2504.06397* (2025).

[88] Zehan Wang et al. “Omnibind: Large-scale omni multimodal representation via binding spaces”. In: *arXiv preprint arXiv:2407.11895* (2024).

[89] Suyao Wei and Zhihui Wu. “The application of wearable sensors and machine learning algorithms in rehabilitation training: A systematic review”. In: *Sensors* 23.18 (2023), p. 7667.

[90] Keenon Werling et al. “Addbiomechanics dataset: Capturing the physics of human motion at scale”. In: *European Conference on Computer Vision*. Springer. 2024, pp. 490–508.

[91] Jialong Wu et al. “ivideogpt: Interactive videogpts are scalable world models”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 68082–68119.

[92] Lianghao Xia and Chao Huang. “Anygraph: Graph foundation model in the wild”. In: (2024).

[93] Xuan Xiao et al. “Robot learning in the era of foundation models: A survey”. In: *Neurocomputing* (2025), p. 129963.

[94] Yufei Xu et al. “Vitpose: Simple vision transformer baselines for human pose estimation”. In: *Advances in neural information processing systems* 35 (2022), pp. 38571–38584.

[95] Yufei Xu et al. “Vitpose++: Vision transformer for generic body pose estimation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.2 (2023), pp. 1212–1230.

[96] Shaokai Ye et al. “LLaVAction: evaluating and training multi-modal large language models for action recognition”. In: *arXiv preprint arXiv:2503.18712* (2025).

[97] Shaokai Ye et al. “SuperAnimal pretrained pose estimation models for behavioral analysis”. In: *Nature communications* 15.1 (2024), p. 5165.

[98] Albert Yu et al. “Natural language can help bridge the sim2real gap”. In: *arXiv preprint arXiv:2405.10020* (2024).

[99] Ye Yuan et al. “Physdiff: Physics-guided human motion diffusion model”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, pp. 16010–16021.

[100] Mingyuan Zhang et al. “Motiondiffuse: Text-driven human motion generation with diffusion model”. In: *IEEE transactions on pattern analysis and machine intelligence* 46.6 (2024), pp. 4115–4128.

[101] Siwei Zhang et al. “Egobody: Human body shape and motion of interacting people from head-mounted devices”. In: *European conference on computer vision*. Springer. 2022, pp. 180–200.

[102] Yiyuan Zhang et al. “Meta-transformer: A unified framework for multimodal learning”. In: *arXiv preprint arXiv:2307.10802* (2023).

[103] Jiahe Zhao et al. “HIS-GPT: Towards 3D Human-In-Scene Multimodal Understanding”. In: *arXiv preprint arXiv:2503.12955* (2025).

[104] Jiaxing Zhao et al. “HumanOmni: A Large Vision-Speech Language Model for Human-Centric Video Understanding”. In: *arXiv preprint arXiv:2501.15111* (2025).

[105] Yian Zhao et al. “Detrs beat yolos on real-time object detection”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, pp. 16965–16974.

[106] Silvia Zuffi et al. “3D menagerie: Modeling the 3D shape and pose of animals”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 6365–6373.

Technical Appendices and Supplementary Material

Table 1: Examples of challenging, high-impact goals in movement modeling and analysis

Domain	Example goal	Challenging task
Neuroscience		
Neuromotor interfaces	High-dimensional control	Intent prediction and action recognition
Neuroethology	Understand social dynamics	Relating movements with neural activity during natural behavior
Health & Medicine		
Rehabilitation	Personalize rehabilitation plans and outcome tracking using movement data	Low-shot, context-aware adaptation across patient populations
Biomarker Discovery	Detect early-stage motor degradation in Parkinson's	Fine-grained temporal modeling that preserves diagnostic signals
Child Development	Diagnose atypical motor development in infants within first few months of age	Learn from rare, unlabeled or unstructured motor trajectories
Climate & Environmental Health	Track exertion, instability, or gait adaptation in extreme heat or smoke	Movement models conditioned on environmental covariates
Human–Machine Interaction		
Assistive Devices	Optimize prosthetics or exoskeletons in daily use	Dynamic adaptation to user-specific control patterns
Human–Robot Collaboration	Anticipate and respond to human movement intent	Shared control spaces for fluid joint action
Science		
Evolutionary Biology	Compare locomotion across phylogenetic lineages	Latent spaces that reflect morphological and functional divergence
Animal Behavior & Conservation	Model behavioral adaptation across habitats	Transfer across species, climates, and movement vocabularies
Performance, Adaptation, and Resilience		
Peak Performance	Optimize motion in high-stakes sports and occupational tasks	Detection of micro-asymmetries and fatigue under real-world stress
Space Medicine	Track neuromotor adaptation in microgravity	Few-shot generalization to off-Earth conditions, low-signal environments
Extreme Environments	Model movement under cold, hypoxia, or heavy gear	Environment-aware prediction of failure points and adaptation strategies
Creative Domains		
Generative video	Generate precise movements that accurately reflect text prompts	Physically grounded movement across species
Choreography	Generate novel, physically valid movement sequences	Constrained generative movement models for creativity and rehearsal
Industry & Public Policy		
Occupational Safety	Forecast injury risk from asymmetrical or repetitive motion	Real-time risk estimation in safety-critical domains
Public Infrastructure	Design inclusive environments for mobility and access	Simulation of population-level movement from diverse bodies and devices
Robotics		
Social Robots	Generate human-like movements that allow for social interaction	Perform real-time interpretation of human movement and generate appropriate adaptive responses
Legged Robots	Match the ease with which animals adapt to different terrains	Implementation of biomimetic strategies