Review

Not peer-reviewed version

Physic Grounded Vision Foundation Models for Human Computer Interaction in Embodied Environments

Gulnaz Rati [*] , Rafael Mendes , Aisha Noor

Posted Date: 10 October 2025

doi: 10.20944/preprints202510.0649.v1

Keywords: physic-ground vision models; foundation models; human-computer interaction; embodied intelligence; multimodal learning; differentiable physics; causal reasoning; assistive robotics; embodied cognition; interdisciplinary AI; sim-to-real transfer; ethical AI design; interactive systems

Review

Physic Grounded Vision Foundation Models for Human Computer Interaction in Embodied Environments

Gulnaz Rati [1,*], Rafael Mendes [2] and Aisha Noor [3]

[1] University of São Paulo, Brazil
[2] Federal University of Rio de Janeiro, Brazil
[3] University of Campinas, Brazil
* Correspondence: gulnaz.zati@usp.br

Abstract

Physic-ground vision foundation models for human-computer interaction represent a transformative paradigm in artificial intelligence, as they extend beyond conventional data-driven perception to incorporate explicit reasoning about the physical laws and causal structures that govern the real world. Unlike earlier generations of vision models that excelled at pattern recognition but faltered when faced with tasks demanding robust predictions of dynamics, affordances, or embodied interactions, these new approaches explicitly embed principles of physics into large-scale multimodal architectures. This integration allows systems to not only recognize objects and interpret scenes but also anticipate outcomes, model constraints, and interact in ways that are coherent with the embodied experience of humans. The result is a class of foundation models that hold profound implications for applications ranging from assistive robotics and healthcare to education, design, and collaborative work, where safety, interpretability, and physical plausibility are paramount. In this review, we explore the conceptual underpinnings, methodological innovations, and broad implications of physic-ground vision foundation models. We highlight the technical advances that have made large-scale physical reasoning feasible, including differentiable physics simulators, causal representation learning, and multimodal integration strategies that combine visual, tactile, and proprioceptive inputs into unified frameworks. We also examine the computational challenges inherent in simulating high-dimensional physical dynamics, the scarcity of richly embodied datasets, and the difficulties of bridging synthetic-to-real gaps in interactive environments. Beyond technical considerations, we emphasize the interdisciplinary nature of the field, drawing on insights from cognitive science, neuroscience, robotics, and the social sciences to show how these models can be both technically robust and socially meaningful. Crucially, we discuss the ethical, philosophical, and societal implications of deploying physic-ground systems in real-world contexts. By enabling machines to act in ways that are physically grounded, these models reshape the balance of autonomy and control in human-computer interaction, raising questions of trust, accountability, equity, and inclusivity. They also invite deeper reflection on the nature of intelligence itself, as machines begin to approximate forms of embodied reasoning once considered unique to humans. Looking forward, we argue that the future of physic-ground vision foundation models depends not only on technical breakthroughs but also on interdisciplinary collaboration and human-centered design, ensuring that these systems serve as partners in creativity, learning, and problem-solving rather than opaque or paternalistic arbiters of human activity. In this way, physic-ground models embody both the extraordinary potential and the immense responsibility that defines the next era of human-computer interaction.

Keywords: physic-ground vision models; foundation models; human-computer interaction; embodied intelligence; multimodal learning; differentiable physics; causal reasoning; assistive robotics; embodied cognition; interdisciplinary AI; sim-to-real transfer; ethical AI design; interactive systems

1. Introduction

In recent years, the rapid advancements in artificial intelligence and computer vision have converged to give rise to a new class of models that promise to fundamentally transform the way humans interact with machines: vision foundation models. These large-scale models, trained on diverse and massive datasets, have demonstrated remarkable generalization capabilities across a wide spectrum of vision tasks, ranging from image classification and object detection to semantic segmentation, pose estimation, and video understanding. However, the majority of early foundation models primarily focused on abstract visual recognition tasks without explicitly grounding their representations in the underlying physics of human perception, embodiment, and interaction with the physical world. This limitation has led to a growing recognition of the necessity for *physic-grounded vision foundation models*, where perceptual understanding is tightly linked with the physical dynamics of environments, the structural regularities of the real world, and the embodied context of human-computer interaction (HCI) [1]. The essence of grounding vision models in physics lies in bridging the gap between visual perception and the laws that govern the natural world. Humans, for example, interpret visual stimuli not merely as static pixel arrays but through an intuitive understanding of motion, forces, material properties, occlusion, gravity, and causal relationships. When reaching for an object, humans inherently anticipate its weight and frictional properties; when navigating an environment, they effortlessly account for geometric constraints, stability, and affordances. Traditional deep learning models, despite their unprecedented accuracy in benchmark tasks, have largely overlooked such physical priors, often resulting in brittle generalization when deployed in real-world interactive scenarios. Consequently, the vision community has witnessed a paradigm shift towards embedding physical reasoning into the core architectures and training objectives of foundation models, aiming to endow them with robust, interpretable, and transferable capabilities for dynamic HCI applications. Within the domain of human-computer interaction, the integration of physic-ground vision foundation models introduces a fundamentally richer paradigm for designing intelligent systems that can perceive, reason, and respond in a manner more aligned with human expectations. For instance, in virtual and augmented reality, physics-informed vision models enable more natural and immersive interactions by predicting realistic object dynamics, hand-object interactions, and environmental feedback. In assistive technologies, such models enhance the interpretability and reliability of systems designed for accessibility, such as gesture recognition for individuals with motor impairments or smart prosthetic devices that adapt based on learned physical constraints. Furthermore, in collaborative robotics, physic-grounded perception facilitates safe and seamless cooperation between humans and machines, as robots equipped with such models can better anticipate human intentions, understand shared workspaces, and predict the outcomes of joint actions. By coupling visual understanding with the implicit principles of physics, foundation models thus open pathways for more robust and human-centered HCI design [2]. The transition towards physic-ground foundation models also raises deeper theoretical and methodological questions regarding the sources of grounding, the mechanisms for integrating physics priors, and the evaluation benchmarks that truly reflect physical reasoning in interactive settings [3]. Existing approaches span a diverse range, including explicit physics engines integrated with neural networks, self-supervised learning based on real-world dynamics, differentiable simulators that guide representation learning, and multimodal frameworks that align visual perception with proprioceptive or tactile signals. Moreover, large-scale synthetic datasets have been curated to mimic physical laws in controlled settings, while real-world data collection strategies emphasize embodied interaction to teach models intuitive physics [4]. Despite these strides, the field is still grappling with open challenges such as scalability, efficiency, interpretability, and cross-domain generalization [5]. Addressing these challenges is critical to ensure that physic-grounded vision models not only excel in controlled laboratory tasks but also adapt seamlessly to the unpredictability of real-world human environments [6]. It is important to note that the incorporation of physics into vision models is not merely a matter of performance optimization but also a philosophical step towards aligning artificial systems with the embodied cognition of humans. Cognitive science and neuroscience suggest that human perception is inherently predictive and physics-driven, constantly simulating future states of the environment based on prior experiences and sensory cues. Replicating such embodied predictive capabilities in machines represents a profound milestone for artificial intelligence, one that reshapes HCI by fostering trust, transparency, and alignment with human expectations [7]. Unlike static recognition models, physic-grounded systems can account for uncertainty, reason about causality, and anticipate downstream consequences of actions, thus enabling interactions that are not only reactive but proactive and anticipatory. This alignment brings the promise of moving beyond shallow task-specific interaction toward rich, adaptive, and context-aware systems that continuously learn and evolve alongside their human counterparts [8]. In summary, the advent of physic-ground vision foundation models marks a pivotal moment in the evolution of human-computer interaction. By embedding physical reasoning and grounding visual perception in the fundamental principles that govern the real world, these models transcend the limitations of traditional vision systems and pave the way for interactive technologies that are more robust, natural, and human-centric [9]. The scope of this review is to systematically analyze the emerging research directions, methodological innovations, challenges, and application domains where physic-ground vision foundation models are reshaping HCI [10]. Through this exploration, we seek to highlight not only the technical contributions but also the broader implications for building intelligent systems that can understand, predict, and collaborate with humans in a manner that feels intuitively aligned with our shared physical reality [11].

2. Mathematical Formulation of Physic-Grounded Vision Models

To formally articulate the underpinnings of physic-ground vision foundation models for human-computer interaction, it is essential to construct a mathematical framework that integrates perception, physical reasoning, and interactive dynamics [12]. Unlike conventional computer vision approaches that primarily optimize for pixel-wise or feature-level objectives, physic-ground models embed physical constraints, dynamic equations, and causal structures directly into the representation space. Let us denote the raw sensory input (e.g., image or video frames) by $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$, where H and W are the height and width of the visual frame and C denotes the number of channels (typically $C = 3$ for RGB data) [13]. The objective of the vision model is to map $\mathbf{I}$ into a latent representation $\mathbf{z} \in \mathbb{R}^d$ such that it captures not only semantic abstractions but also the physical state variables underlying the observed scene. We define the set of latent physical states as $\mathcal{S} = \{\mathbf{x}_t\}_{t=1}^{T}$, where each $\mathbf{x}_t \in \mathbb{R}^n$ represents the state of the system at time t [14]. This state typically encodes position $\mathbf{p}_t \in \mathbb{R}^3$, velocity $\mathbf{v}_t \in \mathbb{R}^3$, orientation $\mathbf{R}_t \in SO(3)$, and possibly higher-order derivatives or material properties. The evolution of these states follows Newtonian mechanics or more general physical laws, represented as a differential equation:

$$\frac{d\mathbf{x}_t}{dt} = f_{\text{dyn}}(\mathbf{x}_t, \mathbf{u}_t, \mathbf{F}_t), \tag{1}$$

where $\mathbf{u}_t$ is a control input (e.g., an action from the human or computer agent), and $\mathbf{F}_t$ encapsulates external forces such as gravity, friction, or contact interactions [15]. The challenge in physic-ground vision models is to infer $\mathbf{x}_t$ directly from visual observations $\mathbf{I}_{1:T}$ without explicit annotations, a task commonly framed as solving an inverse problem:

$$\hat{\mathbf{x}}_t = \arg\min_{\mathbf{x}_t} \ \mathcal{L}\big(\Phi(\mathbf{I}_{1:T}), \mathbf{x}_{1:T}\big), \tag{2}$$

where Φ denotes the feature extraction mapping from images to latent representations, and $\mathcal{L}$ represents a loss function that enforces consistency between predicted states and physical dynamics. In addition to dynamics, human-computer interaction requires modeling the bidirectional coupling between human intent and machine response [16]. Let $\mathbf{a}_t^h$ denote human actions and $\mathbf{a}_t^m$ denote machine actions [17]. The interaction can be modeled as a joint policy:

$$\pi(\mathbf{a}_t^h, \mathbf{a}_t^m | \mathbf{x}_t, \mathbf{I}_t) = \pi_h(\mathbf{a}_t^h | \mathbf{x}_t, \mathbf{I}_t) \cdot \pi_m(\mathbf{a}_t^m | \mathbf{x}_t, \mathbf{a}_t^h), \tag{3}$$

where π_h represents the probabilistic distribution of human actions given perceived states, and π_m represents the machine's adaptive policy conditioned on both the state and the human's actions. The physic-grounded nature of the model ensures that $\mathbf{x}_t$ embodies physically meaningful attributes, thereby allowing π_m to align its behavior with human expectations in a physically plausible manner. To further formalize this grounding, consider a constraint-based optimization problem where the objective is to learn $\mathbf{z}$ such that it satisfies both semantic reconstruction and physical consistency:

$$\min_{\theta} \mathbb{E}_{\mathbf{I}_{1:T}} \Big[\mathcal{L}_{\text{vision}}(\mathbf{I}_{1:T}, \hat{\mathbf{I}}_{1:T}) + \lambda \, \mathcal{L}_{\text{physics}}(\mathbf{x}_{1:T}, f_{\text{dyn}}) \Big], \tag{4}$$

where θ are the parameters of the foundation model, $\mathcal{L}_{\text{vision}}$ enforces reconstruction or prediction of visual sequences, and $\mathcal{L}_{\text{physics}}$ enforces compliance with physical dynamics, with λ controlling the trade-off [18]. This formulation illustrates how physics acts as a regularizer, shaping the latent space toward physically grounded abstractions that improve both robustness and generalization [19]. A particularly important aspect in HCI contexts is the notion of *affordances*, which represent the actionable possibilities offered by objects or environments. We denote affordance functions as $\mathcal{A} : \mathbf{x}_t \mapsto \mathbb{R}^k$, where each dimension corresponds to the feasibility of a specific interaction (e.g., grasping, pushing, lifting). Affordance reasoning can be expressed as a probabilistic mapping:

$$P(\text{action} | \mathbf{I}_t, \mathbf{x}_t) = \sigma\big(W \cdot g(\mathbf{z}_t, \mathbf{x}_t) + b\big), \tag{5}$$

where $g(\cdot)$ is a nonlinear function combining latent visual features $\mathbf{z}_t$ and physical states $\mathbf{x}_t$, W and b are learnable parameters, and $\sigma(\cdot)$ denotes the softmax function. By embedding affordances into the predictive framework, physic-ground vision foundation models not only reconstruct scenes but also anticipate feasible human-machine interactions [20]. Finally, we emphasize that the full pipeline of a physic-ground foundation model can be represented as a multi-objective optimization:

$$\min_{\theta} \; \alpha \, \mathcal{L}_{\text{semantic}} + \beta \, \mathcal{L}_{\text{temporal}} + \gamma \, \mathcal{L}_{\text{physics}} + \delta \, \mathcal{L}_{\text{interaction}}, \tag{6}$$

where $\mathcal{L}_{\text{semantic}}$ ensures traditional visual understanding, $\mathcal{L}_{\text{temporal}}$ ensures temporal consistency, $\mathcal{L}_{\text{physics}}$ encodes adherence to physical constraints, and $\mathcal{L}_{\text{interaction}}$ optimizes human-computer synergy. The coefficients $\alpha, \beta, \gamma, \delta$ are hyperparameters balancing the contributions of each objective. This holistic formulation underscores the fact that physic-ground vision foundation models must balance perception, reasoning, and interaction within a unified mathematical structure to fully realize their potential in the next generation of HCI systems [21].

3. Embodied Context and Human-Centric Design Principles

The role of physic-ground vision foundation models in human-computer interaction extends far beyond the abstract realm of visual perception or the narrowly defined scope of computational benchmarks. At its core, the notion of grounding models in physical principles is tightly linked with the concept of embodiment, where intelligent agents are not isolated reasoning engines but rather entities that exist, perceive, and act within a physical world governed by natural laws and human expectations [22]. From the perspective of human-centric design, this shift introduces a profound transformation in the way interactive systems are conceived: rather than being programmed as rigid tools that respond only to pre-specified commands or static input-output mappings, they become adaptive collaborators capable of predicting, interpreting, and even anticipating human actions within physically coherent environments [23]. The embodiment of such models ensures that their learned representations inherently capture the affordances, constraints, and causal dynamics of real-world objects and spaces, thereby fostering a more intuitive alignment between artificial intelligence and human cognitive processes. For instance, when a person extends their hand toward a virtual object in augmented reality, a physic-ground foundation model can recognize not only the gesture but also infer the intention behind it, anticipate the potential outcome of the interaction based on object mass, shape, or stability, and generate feedback that reflects these dynamics. Such predictive and physically grounded interaction is vital for building trust, as users are more likely to accept and seamlessly integrate systems whose responses are consistent with their embodied experiences and expectations of the physical world [24]. An essential aspect of this integration is the recognition that human-computer interaction does not occur in isolation but in dynamic and often unpredictable contexts where multiple agents, objects, and environmental factors interact simultaneously [25]. Physic-ground vision foundation models address this complexity by embedding an implicit understanding of causal dependencies and physical regularities into their reasoning process. Instead of passively analyzing frames or segments of data, these models simulate possible trajectories, account for uncertainties, and continuously refine their predictions in real-time as new sensory inputs arrive [26]. Such simulation-driven perception is directly inspired by cognitive science, where humans are understood to construct forward models of the world that allow them to mentally test possible actions and outcomes before executing them. Translating this principle into artificial systems means that an HCI framework powered by physic-ground vision models can not only detect an object on a desk but also simulate whether it is stable if pushed, whether it can be safely grasped given its geometry and friction, or whether it might obstruct other interactions if moved. This continuous interplay between perception, simulation, and interaction unlocks a richer set of possibilities for designing interfaces that are no longer restricted to static interpretations but dynamically adapt to unfolding scenarios, making them more robust to the variability of real-world conditions. Equally important is the notion of interpretability and transparency, both of which become significantly more achievable when models are grounded in physics. Traditional deep learning architectures, while often accurate in specific benchmarks, are criticized for being black-box systems that provide little insight into their decision-making processes. By contrast, physic-ground models are inherently constrained by equations of motion, conservation laws, and structural relationships that can be explicitly examined and validated. This incorporation of physical principles serves as a natural form of regularization, guiding models toward plausible and human-understandable predictions [27]. For human-computer interaction, such transparency is not merely a technical advantage but a fundamental requirement, as users are more likely to trust and adopt systems that can justify their decisions in terms of physical reasoning rather than opaque statistical correlations [28]. Consider, for example, a collaborative robot operating alongside a human worker: when deciding whether to lift a shared object, a physic-ground model can communicate not only that the action is feasible but also that the expected load is within safe limits given mass estimates and frictional properties [29]. This ability to provide explanations in a language grounded in human physical intuition marks a critical step toward responsible and human-centered AI deployment. Beyond transparency, physic-ground vision foundation models also redefine the scalability and adaptability of interactive systems. By embedding generalizable physical principles into their core, such models are better equipped to transfer knowledge across tasks, environments, and modalities [30]. Whereas conventional machine learning systems often require extensive retraining when exposed to new conditions, physic-ground models leverage their internalized understanding of invariants like gravity, inertia, or object permanence to quickly adapt to unfamiliar scenarios. This adaptability is particularly significant in the realm of HCI, where the diversity of users, contexts, and goals cannot be exhaustively pre-programmed. For instance, in assistive technology for individuals with motor impairments, physic-ground models can generalize from one set of gestures or actions to entirely new forms of interaction by recognizing the underlying physical intent, such as pointing, grasping, or supporting. Similarly, in immersive environments such as virtual reality, they can maintain consistency in object dynamics even when rendering conditions, interaction devices, or user behaviors vary significantly. This universality of grounding offers a path toward systems that are not only scalable in terms of computational performance but also scalable in terms of their ability to interact meaningfully across a wide range of human contexts [31]. Finally, the implications of physic-ground foundation models for the future of human-computer interaction are deeply interdisciplinary, intersecting with fields such as cognitive psychology, robotics, human factors engineering, and design studies [32]. The process of embedding physical reasoning into vision systems demands collaboration between engineers who design efficient algorithms, cognitive scientists who provide insights into human perception and embodiment, and designers who envision how these systems can be integrated into everyday human practices [33]. The extreme richness of this intersection ensures that the adoption of physic-ground vision foundation models is not merely a technical trend but a holistic paradigm shift that reshapes the very principles of interaction design [34]. It highlights a trajectory where computational intelligence is not detached from human experience but fundamentally aligned with the embodied, predictive, and physically coherent ways in which humans engage with their environments. This convergence suggests that the future of HCI will increasingly move toward systems that are not only functional but also seamlessly woven into the texture of human life, anticipating needs, adapting to contexts, and communicating in a manner that reflects the shared physical reality between humans and machines.

4. Architectural Abstractions and System-Level Integration

The design of physic-ground vision foundation models for human-computer interaction is not only a matter of constructing better perception pipelines but also one of orchestrating multiple architectural components into a cohesive system that spans sensing, reasoning, and interaction [35]. At a high level, such systems can be described as layered architectures, where raw sensory inputs such as visual frames, depth signals, or proprioceptive feedback are progressively transformed into abstract representations that encode both semantic meaning and physical consistency. The first layer of this architecture typically involves low-level feature extraction from high-dimensional sensory data, where convolutional or transformer-based encoders distill pixel intensities into compact embeddings. These embeddings, however, are insufficient if they remain detached from the physical world, so the second layer integrates dynamical reasoning mechanisms that map latent features onto physical state variables such as positions, velocities, orientations, or contact forces. A subsequent layer then incorporates predictive modeling, simulating forward trajectories under various hypothetical actions and evaluating possible outcomes using physical laws [36]. Finally, at the top of the hierarchy, an interaction layer aligns predicted outcomes with human intent and machine responses, thereby closing the loop between perception, reasoning, and action in a way that is deeply grounded in physical coherence. This hierarchical design reflects the fact that meaningful interaction cannot be achieved by perception alone; rather, it requires a systemic approach where each stage of computation respects and enforces the physical regularities of the environment in which the interaction occurs. One of the most significant implications of such an architectural design is that it enables systems to achieve robustness and generalization not through brute-force memorization of visual patterns but through principled reasoning about physical constraints. By embedding explicit or implicit models of physical dynamics into their internal representations, physic-ground vision foundation models can extrapolate beyond observed data and predict how novel situations will unfold, which is essential for interactive systems that must operate in diverse, unstructured, and unpredictable human environments. For example, a robot trained in one environment may encounter objects of unfamiliar shapes or materials in another, yet if its perceptual pipeline is grounded in principles of mass, friction, and force interaction, it can still infer feasible ways to grasp, move, or interact with these objects [37]. This ability to generalize across domains and tasks arises not from memorizing a dataset of possible interactions but from leveraging the invariants that underlie the physical world. For human-computer interaction, this translates into systems that are not brittle or fragile but adaptive, capable of extending their learned knowledge to situations never explicitly encountered during training [38]. Such adaptability is critical in real-world contexts where human actions are highly variable, environmental conditions are dynamic, and the success of interaction depends on rapid, reliable, and physically consistent responses [39]. To illustrate this hierarchical flow and integration of components, Figure **??** provides a simple schematic representation of a physic-ground vision foundation model architecture for HCI. The figure captures the multi-layer structure of the system, beginning with visual inputs at the left and culminating in interaction-level outputs at the right. Each stage is conceptualized as both a computational and physical abstraction, where information is progressively refined to capture semantic meaning, physical state, predictive dynamics, and interaction affordances. The representation, while simplified, conveys the principle that physic-ground vision models must operate as integrated systems rather than as isolated modules, ensuring that the final outputs are coherent with both human expectations and physical laws. The architectural abstraction illustrated above demonstrates that grounding vision models in physics is not a peripheral add-on but a central organizing principle that fundamentally shapes the flow of information across the entire system. By embedding physical reasoning at intermediate stages, the model ensures that the representations passed forward to higher levels are not arbitrary abstractions but physically meaningful constructs that can support predictive and interactive tasks [40]. Moreover, this design philosophy suggests a path toward unifying disparate components of artificial intelligence, where perception, reasoning, and action are no longer treated as separate silos but as deeply interconnected layers bound together by the shared language of physics. For HCI researchers and practitioners, this integrated view offers a framework for constructing systems that are simultaneously powerful, interpretable, and aligned with the embodied nature of human experience, thereby advancing the ultimate goal of creating interactive technologies that feel not artificial but seamlessly embedded within the dynamics of everyday human life.

5. Comparative Perspectives and Emerging Benchmarks

In order to fully appreciate the transformative role of physic-ground vision foundation models in human-computer interaction, it is instructive to compare them systematically against both traditional computer vision approaches and the more recent generation of large-scale vision-language or multi-modal foundation models that are not explicitly grounded in physics. Such comparisons shed light not only on the advantages that physical grounding brings in terms of robustness, interpretability, and adaptability, but also on the limitations and challenges that must be overcome before these models can be deployed at scale in real-world interactive systems [41]. Traditional computer vision models, for example, were largely designed as task-specific solutions: a convolutional neural network trained for image classification would rarely generalize to tasks such as object tracking or pose estimation without retraining or significant modification. By contrast, foundation models trained on massive datasets have demonstrated impressive generalization across multiple tasks, but they often fail when confronted with out-of-distribution scenarios or when required to reason about physical dynamics beyond static visual features [42]. This weakness underscores the critical insight that generalization in the absence of physical grounding remains shallow, as systems may misinterpret plausible-looking but physically impossible scenarios [43]. For human-computer interaction, this limitation is especially problematic because users demand systems that not only recognize and respond but do so in a way that is consistent with the underlying rules of the physical world [44]. Physic-ground vision foundation models address this gap by incorporating physical priors and constraints directly into their training and inference pipelines. This incorporation takes multiple forms, including the integration of differentiable physics engines, the use of self-supervised objectives based on temporal consistency, and the alignment of visual representations with state variables such as force vectors, material properties, or contact dynamics. The result is a class of models that, while computationally more complex, exhibit qualitatively different behaviors in interactive contexts. For example, when predicting human hand-object interactions in augmented reality, a purely vision-based model may recognize the hand and the object but fail to predict the feasibility of the grasp given the object's mass and friction properties [45]. A physic-ground model, by contrast, can anticipate the dynamics of the interaction, predicting not only whether the grasp will succeed but also whether the object might slip or deform. Similarly, in robotics, purely visual models may succeed at static object recognition but fail when required to plan safe trajectories around dynamic obstacles [46]. Physic-ground models, through their explicit modeling of causal relationships and physical dynamics, are far more capable of generating safe, reliable, and human-aligned interaction policies. These differences are not marginal improvements but fundamental shifts in the very capabilities of the models, marking a decisive step forward in the field of HCI [47]. To consolidate these insights, Table 1 provides a comparative overview of representative classes of models across multiple dimensions relevant to human-computer interaction. The table spans traditional vision models, vision-language foundation models, and physic-ground vision foundation models, highlighting their relative strengths and weaknesses across key criteria such as generalization, interpretability, robustness, adaptability, and alignment with human physical intuition [48]. The purpose of this table is not only to summarize technical differences but also to provide a benchmark for evaluating progress in the field [49]. As the table shows, physic-ground models consistently outperform their predecessors in domains requiring predictive reasoning, causal understanding, and interaction with the physical world, though challenges remain in computational scalability, data efficiency, and standardization of benchmarks. These challenges point toward fertile avenues for future research, where innovations in efficient training, multimodal fusion, and embodied simulation environments will likely play pivotal roles [50].

Table 1. Comparative overview of different classes of vision models for human-computer interaction. Physic-ground vision foundation models integrate physical reasoning with perception and interaction, providing distinct advantages in robustness, interpretability, and adaptability.

Dimension	**Traditional Vision Models**	**Vision-Language Foundation Models**	**Physic-Ground Vision Foundation Models**
Generalization Across Tasks	Narrow, task-specific, requiring retraining	Broad generalization across perception tasks but weak on dynamics	Strong generalization, especially in tasks requiring physical reasoning and interaction
Interpretability	Limited interpretability, black-box features	Partially interpretable through multimodal alignment	Enhanced interpretability via physically meaningful latent variables
Robustness to Out-of-Distribution Scenarios	Fragile, easily fails under distribution shift	More robust than traditional, but fails in physically implausible cases	High robustness due to grounding in physical constraints and dynamics
Adaptability in HCI Contexts	Requires explicit reprogramming for new contexts	Can adapt linguistically but not physically	Adaptable to new contexts by leveraging invariants such as gravity, friction, and object permanence
Alignment with Human Physical Intuition	Weak, often counterintuitive outputs	Moderate, can describe scenarios linguistically but lacks causal reasoning	Strong alignment, anticipates outcomes in ways consistent with human embodied cognition
Computational Complexity	Moderate, efficient for specific tasks	High due to large-scale multimodal training	Very high due to integration of physical simulations and multimodal inputs
Applications in HCI	Limited to recognition, detection, tracking	Strong for multimodal interfaces (vision + language)	Broad spectrum: robotics, AR/VR, assistive technologies, collaborative systems

Taken together, the comparative evidence suggests that physic-ground vision foundation models represent not just an incremental improvement but a paradigm shift in the design of intelligent systems for human-computer interaction [51]. By explicitly modeling the causal, dynamic, and physical aspects of the world, these models are able to achieve a level of robustness and adaptability that was previously out of reach for purely data-driven or vision-language approaches [52]. At the same time, this shift opens new challenges, particularly regarding efficiency and scalability: training such models often requires access to large-scale simulation environments, extensive computational resources, and carefully curated multimodal datasets that capture both visual and physical information. Furthermore, standardizing evaluation benchmarks for physic-ground reasoning remains an open problem, as existing datasets rarely test for the physical plausibility of predicted interactions. Addressing these challenges will require not only technical innovation but also cross-disciplinary collaboration between AI researchers, cognitive scientists, human factors engineers, and designers [53]. Ultimately, the comparative framework provided here highlights both the extraordinary promise of physic-ground models and the urgent need for further research to realize their full potential in building interactive systems that are robust, interpretable, adaptive, and deeply aligned with the embodied experiences of human users.

6. Challenges, Limitations, and Open Research Directions

While physic-ground vision foundation models hold extraordinary promise for advancing human-computer interaction, their development and deployment remain fraught with technical, conceptual, and practical challenges that require careful examination [54,55]. One of the most pressing challenges is the issue of computational scalability [56]. Unlike conventional deep learning models that focus primarily on perceptual tasks, physic-ground models must simulate or approximate complex physical dynamics, often in real-time, to support interactive applications. This requirement dramatically increases computational costs, as physics engines or differentiable simulators embedded in learning pipelines can become prohibitively expensive when scaled to high-dimensional inputs and long temporal horizons [13]. For instance, simulating the fine-grained dynamics of soft-body deformation or multi-contact interactions in virtual environments requires solving large systems of equations that are computationally intensive, and incorporating these simulations into the training of large-scale foundation models multiplies the demand for resources. Even when approximations are employed, there exists a tension between fidelity to physical laws and computational efficiency, leading to a fundamental trade-off that must be addressed before such models can become widely deployable in real-world HCI systems [57]. Another significant limitation lies in the availability and quality of training data. While large-scale image and video datasets have fueled the rise of conventional foundation models, datasets that adequately capture physical interactions, causal dependencies, and embodied affordances are comparatively scarce [58]. Constructing such datasets often requires either expensive real-world data collection with specialized sensors, such as motion capture systems or haptic devices, or the use of synthetic environments where physical parameters can be explicitly controlled. However, synthetic data raises questions of transferability, as models trained in simulated environments may fail to generalize when exposed to the variability, noise, and unpredictability of the real world [59]. Bridging this so-called “sim-to-real gap” remains a formidable challenge, demanding advances in domain adaptation, transfer learning, and hybrid approaches that combine real-world and simulated data [60]. Moreover, the ethical and practical implications of large-scale data collection in embodied contexts must also be carefully considered, as the capture of human motion, gesture, and interaction data introduces privacy concerns that are particularly sensitive in HCI applications. A further set of challenges arises from the inherent complexity of integrating multiple modalities into a unified physic-ground framework. Human-computer interaction rarely relies on vision alone; instead, it involves a rich fusion of visual, auditory, tactile, and proprioceptive inputs. Designing models that can seamlessly combine these heterogeneous modalities while respecting physical constraints is a non-trivial problem. Vision may provide spatial and semantic information, but tactile data may be essential for inferring frictional properties, while proprioceptive feedback is critical for understanding joint angles or motor constraints. Current approaches often treat multimodal integration as an afterthought, but for physic-ground models to reach their full potential, they must be built on architectures that are inherently multimodal and capable of reasoning jointly across diverse sensory streams. Achieving this integration requires innovations in representation learning, alignment strategies, and training objectives that ensure different modalities contribute coherently to physical reasoning rather than introducing redundancy or conflict. Interpretability and evaluation represent yet another open frontier [61]. While grounding models in physical principles theoretically improves interpretability, in practice the internal representations of deep learning systems often remain opaque [62]. There is a pressing need for evaluation frameworks that go beyond accuracy on benchmark datasets and instead assess whether models respect fundamental physical constraints such as conservation of energy, stability under perturbations, or causal consistency across temporal sequences [63]. Without such benchmarks, claims of physical grounding risk becoming superficial or misleading [64]. At the same time, interpretability must be tailored to the specific needs of human-computer interaction: it is not enough for a model to be physically consistent if its reasoning cannot be communicated in a way that is understandable and actionable for end users. For example, in assistive robotics, users may require explicit explanations of why a model refuses to execute a requested action—such as the prediction that the object is too heavy or unstable to lift safely—rather than opaque denials based on internal error thresholds [65]. This need for user-facing interpretability raises important design questions about how physic-ground reasoning can be expressed in natural and intuitive terms [66]. Finally, there are deeper philosophical and conceptual questions about the very nature of physical grounding in artificial intelligence. While embedding physical constraints into vision models aligns them more closely with human embodied cognition, it remains an open question whether these models genuinely "understand" physics in the way humans do or whether they merely approximate physical laws within narrow domains. Humans, for instance, possess an intuitive physics that is not derived from exact equations but from embodied interaction with the world, where predictions are probabilistic, approximate, and deeply tied to lived experience [67]. Replicating such embodied intuition in machines requires more than embedding Newtonian dynamics; it demands models that can reason about counterfactuals, adapt their intuitions to novel materials or contexts, and learn from sparse, noisy, and incomplete data in the same way humans do. This deeper grounding challenges researchers to rethink not only technical architectures but also theoretical frameworks for what it means to align artificial perception with the embodied, causal, and predictive nature of human experience. Taken together, these challenges underscore that the path toward fully realizing physic-ground vision foundation models for human-computer interaction is both promising and complex. Overcoming the barriers of computational scalability, data scarcity, multimodal integration, interpretability, and conceptual grounding will require sustained innovation across disciplines. Progress will not come from incremental improvements in network architectures alone but from holistic rethinking of how artificial intelligence should perceive, reason, and interact in ways that resonate with the embodied, physical, and causal nature of human life. In this sense, the limitations of current systems are not obstacles to be avoided but opportunities to redefine the future of interaction design, pushing the boundaries of AI toward a paradigm that is not only computationally powerful but also fundamentally aligned with the principles that govern the physical and social worlds we inhabit [68].

7. Ethical, Social, and Philosophical Considerations

As physic-ground vision foundation models for human-computer interaction progress toward real-world deployment, the ethical, social, and philosophical considerations surrounding their development become as central as the technical ones [59]. These considerations stem not only from the potential capabilities of such systems but also from the contexts in which they will be embedded, the populations they will serve, and the values they will inevitably encode. A first and pressing ethical concern lies in the domain of safety and responsibility. Because physic-ground models explicitly simulate or approximate the dynamics of the physical world, the consequences of their actions are not confined to digital abstractions but manifest in the material realm, where errors can cause tangible harm [69]. Unlike purely symbolic or linguistic systems, whose mistakes may remain in the domain of misinterpretation or misinformation, mistakes by physically grounded systems deployed in human-computer interaction contexts—such as assistive robotics, autonomous navigation, or medical interfaces—may lead to injury, property damage, or even loss of life [70]. Thus, the ethical bar for such systems must be considerably higher, demanding rigorous standards of reliability, transparency, and accountability before they are allowed to operate in environments where human safety is at stake. This shift raises complex questions about liability: if a physic-ground model predicts an unstable grasp but a user overrides it, who is responsible for the outcome [71]? Such questions demand not only technical safeguards but also regulatory frameworks that balance innovation with accountability. Another layer of complexity emerges in the social implications of widespread adoption. Human-computer interaction is not only about efficiency or utility; it is fundamentally about shaping how humans relate to the technologies that mediate their lives [72]. The introduction of models that can reason physically, predict outcomes, and proactively intervene in embodied tasks may transform expectations of autonomy, agency, and trust. For instance, an intelligent household assistant grounded in physics could prevent a child from performing a dangerous action, such as handling boiling water or climbing onto unstable furniture. While such interventions may appear unquestionably beneficial, they also redefine the balance of control between humans and machines, creating systems that do not merely execute commands but actively shape human behavior by constraining options based on physical predictions [73]. This raises the risk of paternalism, where machines decide what is "safe" or "appropriate" for humans without sufficient regard for individual autonomy [74]. Striking the balance between empowering users and protecting them from harm will be one of the most delicate design challenges for physic-ground vision foundation models, requiring not only technical precision but also deep engagement with social values and cultural norms [75]. Beyond safety and autonomy, there are philosophical questions concerning the very nature of grounding artificial intelligence in physical principles. Human cognition, as extensively studied in philosophy of mind and cognitive science, is deeply tied to the body and its interactions with the environment [76]. Our intuitions about weight, balance, causality, and affordances are not derived from abstract theories but from lived embodied experiences in a physical world [77]. By attempting to replicate or approximate these principles in machines, physic-ground models take a step closer to aligning artificial intelligence with the embodied nature of human cognition [78]. However, this alignment raises questions about whether such systems are genuinely "understanding" the world or simply reproducing patterns consistent with physical laws. Philosophical debates about strong versus weak AI resurface in this context: is a model that enforces conservation of energy or simulates collision dynamics truly reasoning in the way humans do, or is it merely enforcing constraints without awareness or comprehension? The distinction may seem abstract, but it has practical consequences for how humans interpret and relate to these systems [79]. If users anthropomorphize physic-ground models, attributing to them forms of understanding or intentionality they do not possess, they may overestimate their reliability or moral agency, leading to dangerous over-reliance. Conversely, dismissing them as mere statistical approximators may obscure the fact that their predictive capacity can, in many contexts, surpass human intuition. Moreover, embedding physics into AI systems has subtle implications for social equity and inclusion. If datasets for training such models are biased toward certain cultural practices, physical environments, or interaction styles, the resulting systems may perform poorly when deployed in other contexts, exacerbating global inequities. For example, a physic-ground model trained primarily in Western household environments may fail to generalize to different architectural structures, material properties, or interaction patterns found in other parts of the world [80]. Similarly, accessibility must remain a core priority: for individuals with disabilities, the interaction between embodied constraints and physical predictions may differ significantly from the normative assumptions baked into training data. Without careful design, such systems risk marginalizing the very populations they could most powerfully assist. This calls for a deliberate commitment to diversity and inclusivity in dataset creation, evaluation protocols, and deployment strategies, ensuring that physic-ground models serve a wide range of users in equitable ways. Finally, the development of these systems compels reflection on broader societal transformations [81]. As machines increasingly embody physical reasoning, the boundary between human and machine cognition may appear to narrow, inviting both fascination and anxiety. On one hand, the prospect of artificial systems that share with humans a grounding in the physical fabric of the world promises unprecedented collaboration and understanding. On the other hand, it raises existential questions about the uniqueness of human experience and the future role of humans in technologically saturated environments. Will the ability of machines to "understand" and act upon physical laws diminish the distinctiveness of human embodied intelligence, or will it instead amplify human creativity by offloading physical prediction to machines, freeing humans to focus on higher-order goals? These philosophical questions will shape not only academic debates but also public perceptions of artificial intelligence, influencing how societies choose to adopt, regulate, and integrate such systems. In this sense, the ethical, social, and philosophical considerations surrounding physic-ground vision foundation models are not peripheral concerns but central dimensions of their development, deployment, and long-term impact on the trajectory of human-computer interaction.

8. Interdisciplinary Integration and Cross-Domain Synergies

The advancement of physic-ground vision foundation models for human-computer interaction is unlikely to be achieved through progress in computer science alone. Instead, their full potential will be realized only when insights from a wide spectrum of disciplines converge, creating a collaborative ecosystem where engineering innovation, cognitive theory, social sciences, and even the arts intersect [82]. At the heart of this integration lies the recognition that physical grounding is not merely a technical constraint but a deeply human phenomenon, encompassing the way perception, action, and meaning coalesce in embodied experience. For instance, cognitive science has long studied how humans form internal representations of the world that are not purely visual but richly informed by tactile, auditory, and proprioceptive cues. These studies provide invaluable guidance for designing multimodal architectures that go beyond vision to incorporate a holistic sense of the environment. Neuroscience contributes further, offering insights into how the human brain efficiently encodes physical interactions, such as predicting trajectories or anticipating the consequences of actions, often under uncertainty. Translating such mechanisms into machine learning systems may inspire new algorithms that mirror human-like robustness and adaptability, allowing artificial models to function in dynamic and unpredictable environments with greater reliability. In parallel, the field of physics itself contributes foundational principles that constrain and guide the learning process. While traditional machine learning often operates in a data-driven, unconstrained manner, embedding physical principles ensures that learned representations respect causal and deterministic relationships in the real world. Yet, the interplay between exact physical laws and data-driven approximations opens fertile ground for interdisciplinary inquiry [83]. Applied mathematics, for example, can provide techniques for developing differentiable simulators, enabling efficient integration of physical reasoning into end-to-end training pipelines [84]. Robotics, on the other hand, contributes practical know-how about actuation, sensing, and control, grounding theoretical advances in real-world applications. The engineering challenges faced in robotics—such as dealing with sensor noise, actuator latency, or energy constraints—serve as testbeds for refining physic-ground models, ensuring that theoretical advances do not remain abstract but translate into tangible improvements in embodied performance [85]. Moreover, the synergy between robotics and computer vision is particularly promising: while vision provides high-dimensional perceptual input, robotics ensures that predictions and simulations are continually validated against real-world feedback, closing the loop between perception and action in a physically coherent manner [86]. The social sciences also play a pivotal role in shaping the development of physic-ground foundation models for HCI [87]. Human-computer interaction, at its core, is not only about enabling machines to perceive and predict physical dynamics but also about aligning those capabilities with human needs, values, and social practices [88]. Anthropology and sociology, for instance, provide nuanced perspectives on how different cultures interact with technology, revealing the contextual factors that influence the acceptability and usability of intelligent systems [26]. Psychology offers insights into user trust, cognitive load, and interaction patterns, which are essential for designing interfaces that feel intuitive and supportive rather than opaque or intrusive. Education research can guide the deployment of these systems in learning environments, ensuring that physically grounded simulations support pedagogy by aligning with how students conceptualize physical principles and problem-solving strategies. Even the arts, often overlooked in discussions of AI, can inspire novel approaches to interaction design, offering creative paradigms for how embodied systems might communicate physical reasoning to users in expressive, aesthetic, and emotionally resonant ways. Together, these diverse fields ensure that physic-ground models are not only technically sound but also socially embedded and culturally sensitive [89]. The implications of such cross-domain synergies extend far beyond individual applications, pointing toward a broader reconfiguration of how knowledge is produced and applied in the age of embodied artificial intelligence. By combining engineering precision, cognitive insight, social sensitivity, and creative imagination, physic-ground vision foundation models can become more than tools for efficiency; they can evolve into platforms for human-machine co-creation [90]. Consider, for example, the potential in collaborative scientific discovery: physic-ground systems could assist researchers in simulating experiments, predicting outcomes, and suggesting novel hypotheses based on physically plausible scenarios, effectively augmenting human intuition with machine-driven exploration. In healthcare, such systems might integrate medical imaging with biomechanical simulations to provide surgeons with predictive insights before complex procedures, blending vision, physics, and human expertise into a unified framework. In urban design, they could simulate the interactions of humans, vehicles, and infrastructure under different environmental constraints, supporting planners in creating safer and more livable spaces. These scenarios illustrate that the interdisciplinary integration of physic-ground models does not merely enhance technical capability but reshapes the very processes of innovation, decision-making, and creativity across multiple domains. Ultimately, the path forward for physic-ground vision foundation models will depend on the cultivation of collaborative ecosystems that encourage genuine dialogue between disciplines [91]. This requires not only technical frameworks that facilitate interoperability—such as standardized datasets, shared benchmarks, and open-source tools—but also institutional structures that support interdisciplinary research, including funding mechanisms, collaborative centers, and cross-training programs for students and researchers [8]. It also requires a cultural shift in how success is measured, moving beyond narrow benchmarks of accuracy or efficiency to encompass broader criteria such as inclusivity, interpretability, cultural relevance, and long-term societal impact [92]. By fostering such integrative practices, the field can move toward models that are not only computationally powerful and physically coherent but also deeply aligned with the diverse realities of human life. In this sense, interdisciplinary integration is not merely an auxiliary concern but the very foundation upon which the promise of physic-ground vision models for human-computer interaction will be realized, ensuring that these systems contribute meaningfully to both technological progress and human flourishing.

9. Concluding Reflections and Long-Term Outlook

As we approach the culmination of our exploration into physic-ground vision foundation models for human-computer interaction, it becomes evident that the significance of these systems transcends the boundaries of narrow technological domains and enters into a broader dialogue about the trajectory of artificial intelligence and its entanglement with human life. The development of models that do not merely perceive the visual world in terms of pixels and features but instead anchor their reasoning within the immutable constraints of physical laws represents a profound paradigm shift. This shift signals the maturation of AI from a largely descriptive science of patterns into a generative science of causal reasoning, one that not only recognizes what is but also anticipates what will be. In this sense, the long-term outlook for physic-ground models is both exhilarating and sobering: exhilarating because of their transformative potential across domains as varied as healthcare, education, design, and assistive technologies; sobering because of the immense responsibility that accompanies systems capable of shaping human behavior, influencing physical outcomes, and potentially redefining what it means to interact with intelligent machines in embodied contexts [93]. The most striking feature of physic-ground vision foundation models is their potential to reconfigure the relationship between humans and computers into one of mutual adaptation rather than unilateral instruction [94]. In traditional paradigms, human-computer interaction has often been framed as a sequence of commands and responses, with humans issuing directives and machines executing them within rigid constraints [95].

Physically grounded models, however, have the capacity to anticipate, negotiate, and adapt in ways that resonate with the natural rhythm of human embodied interaction. For example, an intelligent agent equipped with physical reasoning could anticipate the trajectory of a falling object, align its actions with human intentions, and intervene in ways that feel almost collaborative. Such interactions create the possibility of machines that do not merely follow instructions but instead engage in a form of co-creation, where both human and machine adjust dynamically to achieve shared goals. The philosophical implications of this are profound, as they point toward a future where computers are no longer passive instruments of human will but active participants in the unfolding fabric of human action and intention. Nevertheless, this long-term outlook also compels us to reflect critically on the risks and limitations that must be navigated with care [96]. The embedding of physical laws into AI systems, while promising greater robustness and interpretability, does not guarantee immunity from error, bias, or misuse. In fact, the very complexity of these systems may introduce new forms of fragility, such as overfitting to specific physical contexts, underestimating uncertainty in novel environments, or failing to account for the sociocultural dimensions of physical interaction [97]. Moreover, the growing autonomy that physic-ground systems afford to machines risks shifting power away from human agents if not carefully regulated and designed with transparency. For instance, in contexts like healthcare or autonomous driving, the line between human oversight and machine autonomy becomes blurred, raising questions about accountability and trust. It is therefore crucial that the long-term development of these models not only prioritize computational performance but also embed principles of ethical design, inclusive evaluation, and human-centered oversight at their core. Without such guardrails, the promise of physic-ground AI could be undermined by unintended consequences that erode public trust and hinder societal acceptance [98].

Another dimension of the long-term outlook concerns the co-evolution of human capabilities and artificial intelligence. As physic-ground vision foundation models become more pervasive, they will inevitably reshape human cognition, skill acquisition, and interaction patterns. Just as calculators reshaped mathematical practice and the internet transformed access to information, embodied AI systems grounded in physical reasoning may alter how humans engage with physical tasks, problem-solving, and even creative expression. In educational contexts, for example, students may increasingly rely on such systems to simulate experiments, predict physical outcomes, and explore counterfactual scenarios, shifting the focus of education from rote memorization of physical laws toward higher-level reasoning about their implications. In professional contexts, experts in fields such as architecture, engineering, or medicine may find their practices augmented by predictive systems that reveal physical consequences of designs or interventions that would have been invisible to human intuition alone. This co-evolution holds extraordinary promise for amplifying human creativity and productivity, yet it also raises the specter of dependency, where humans lose essential embodied skills or critical reasoning abilities through over-reliance on machine predictions. Striking a balance between augmentation and autonomy will therefore be one of the central challenges in managing the long-term societal impact of physic-ground AI.

Finally, when viewed against the backdrop of humanity's broader technological trajectory, physic-ground vision foundation models can be seen as a key milestone in the pursuit of artificial systems that are not alien to the human condition but instead deeply resonant with it. By aligning machine perception with the same physical realities that structure human life, these systems narrow the gap between artificial and human intelligence, fostering the conditions for more natural, intuitive, and meaningful interactions. Yet, in doing so, they also raise fundamental questions about the uniqueness of human cognition and the future role of humanity in a world increasingly populated by intelligent artifacts. Will the ability of machines to embody physical reasoning lead to deeper forms of partnership that extend human capabilities in unprecedented directions, or will it accelerate a displacement of human roles in ways that undermine social cohesion and individual purpose? The answer to this question is not predetermined but will depend on the choices researchers, designers, policymakers, and societies make in guiding the development of these technologies. In this sense, the long-term outlook for physic-ground vision foundation models is not merely a matter of technical feasibility but a matter of collective vision and responsibility, one that requires us to imagine and shape futures where embodied intelligence—both human and artificial—thrives in mutual flourishing.

References

1. Sharkey, N.E.; Sharkey, A.J. Adaptive generalisation. *Artificial Intelligence Review* **1993**, *7*, 313–328.
2. Xu, C.; Cao, B.T.; Yuan, Y.; Meschke, G. Transfer learning based physics-informed neural networks for solving inverse problems in tunneling. *arXiv e-prints* **2022**, pp. arXiv–2205.
3. Hao, Z.; Liu, S.; Zhang, Y.; Ying, C.; Feng, Y.; Su, H.; Zhu, J. Physics-informed machine learning: A survey on problems, methods and applications. *arXiv preprint arXiv:2211.08064* **2022**.
4. Kingma, D.P. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**.
5. de Wolff, T.; Lincopi, H.C.; Martí, L.; Sanchez-Pi, N. Mopinns: an evolutionary multi-objective approach to physics-informed neural networks. In Proceedings of the Proceedings of the Genetic and Evolutionary Computation Conference Companion, 2022, pp. 228–231.
6. Mowlavi, S.; Nabi, S. Optimal control of PDEs using physics-informed neural networks. *Journal of Computational Physics* **2023**, *473*, 111731.
7. Omidvar, M.N.; Li, X.; Mei, Y.; Yao, X. Cooperative co-evolution with differential grouping for large scale optimization. *IEEE Transactions on evolutionary computation* **2013**, *18*, 378–393.
8. Iwata, T.; Tanaka, Y.; Ueda, N. Meta-learning of Physics-informed Neural Networks for Efficiently Solving Newly Given PDEs. *arXiv preprint arXiv:2310.13270* **2023**.
9. Tang, Y.; Tian, Y.; Ha, D. Evojax: Hardware-accelerated neuroevolution. In Proceedings of the Proceedings of the Genetic and Evolutionary Computation Conference Companion, 2022, pp. 308–311.
10. Yu, G.; Ma, L.; Jin, Y.; Du, W.; Liu, Q.; Zhang, H. A survey on knee-oriented multiobjective evolutionary optimization. *IEEE transactions on evolutionary computation* **2022**, *26*, 1452–1472.
11. Eason, J.; Cremaschi, S. Adaptive sequential sampling for surrogate model generation with artificial neural networks. *Computers & Chemical Engineering* **2014**, *68*, 220–232.
12. Mazé, F.; Ahmed, F. Diffusion models beat gans on topology optimization. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2023, Vol. 37, pp. 9108–9116.
13. Goswami, S.; Anitescu, C.; Chakraborty, S.; Rabczuk, T. Transfer learning enhanced physics informed neural network for phase-field modeling of fracture. *Theoretical and Applied Fracture Mechanics* **2020**, *106*, 102447.
14. Wang, S.; Sankaran, S.; Perdikaris, P. Respecting causality for training physics-informed neural networks. *Computer Methods in Applied Mechanics and Engineering* **2024**, *421*, 116813.
15. Pellegrin, R.; Bullwinkel, B.; Mattheakis, M.; Protopapas, P. Transfer learning with physics-informed neural networks for efficient simulation of branched flows. *arXiv preprint arXiv:2211.00214* **2022**.
16. Garbet, X.; Idomura, Y.; Villard, L.; Watanabe, T. Gyrokinetic simulations of turbulent transport. *Nuclear Fusion* **2010**, *50*, 043002.
17. Chakraborty, S. Transfer learning based multi-fidelity physics informed deep neural network. *Journal of Computational Physics* **2021**, *426*, 109942.
18. Cao, L.; Hong, H.; Jiang, M. Fast Solving Partial Differential Equations via Imitative Fourier Neural Operator. In Proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN). IEEE, 2024, pp. 1–8.
19. Cheng, S.; Alkhalifah, T. Meta-PINN: Meta learning for improved neural network wavefield solutions. *arXiv preprint arXiv:2401.11502* **2024**.
20. Krishnapriyan, A.; Gholami, A.; Zhe, S.; Kirby, R.; Mahoney, M.W. Characterizing possible failure modes in physics-informed neural networks. *Advances in Neural Information Processing Systems* **2021**, *34*, 26548–26560.
21. Hospedales, T.; Antoniou, A.; Micaelli, P.; Storkey, A. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence* **2021**, *44*, 5149–5169.
22. Wandel, N.; Weinmann, M.; Neidlin, M.; Klein, R. Spline-pinn: Approaching pdes without data using fast, physics-informed hermite-spline cnns. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2022, Vol. 36, pp. 8529–8538.
23. Kharazmi, E.; Cai, M.; Zheng, X.; Zhang, Z.; Lin, G.; Karniadakis, G.E. Identifiability and predictability of integer-and fractional-order epidemiological models using physics-informed neural networks. *Nature Computational Science* **2021**, *1*, 744–753.
24. Gupta, A.; Zhou, L.; Ong, Y.S.; Chen, Z.; Hou, Y. Half a dozen real-world applications of evolutionary multitasking, and more. *IEEE Computational Intelligence Magazine* **2022**, *17*, 49–66.

25. Raissi, M.; Yazdani, A.; Karniadakis, G.E. Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. *Science* **2020**, *367*, 1026–1030.
26. Cho, W.; Jo, M.; Lim, H.; Lee, K.; Lee, D.; Hong, S.; Park, N. Parameterized physics-informed neural networks for parameterized PDEs. *arXiv preprint arXiv:2408.09446* **2024**.
27. Shi, Z.; Hu, Z.; Lin, M.; Kawaguchi, K. Stochastic Taylor Derivative Estimator: Efficient amortization for arbitrary differential operators. *arXiv preprint arXiv:2412.00088* **2024**.
28. Wu, H.; Luo, H.; Ma, Y.; Wang, J.; Long, M. RoPINN: Region Optimized Physics-Informed Neural Networks. *arXiv preprint arXiv:2405.14369* **2024**.
29. Markidis, S. The old and the new: Can physics-informed deep-learning replace traditional linear solvers? *Frontiers in big Data* **2021**, p. 92.
30. Peiró, J.; Sherwin, S. Finite difference, finite element and finite volume methods for partial differential equations. *Handbook of Materials Modeling: Methods* **2005**, pp. 2415–2446.
31. Zhang, D.; Lu, L.; Guo, L.; Karniadakis, G.E. Quantifying total uncertainty in physics-informed neural networks for solving forward and inverse stochastic problems. *Journal of Computational Physics* **2019**, *397*, 108850. https://doi.org/10.1016/j.jcp.2019.07.048.
32. Liu, Y.; Sun, Y.; Xue, B.; Zhang, M.; Yen, G.G.; Tan, K.C. A survey on evolutionary neural architecture search. *IEEE transactions on neural networks and learning systems* **2021**, *34*, 550–570.
33. Jiang, Z.; Jiang, J.; Yao, Q.; Yang, G. A neural network-based PDE solving algorithm with high precision. *Scientific Reports* **2023**, *13*, 4479.
34. Tian, Y.; Zhang, X.; Wang, C.; Jin, Y. An evolutionary algorithm for large-scale sparse multiobjective optimization problems. *IEEE Transactions on Evolutionary Computation* **2019**, *24*, 380–393.
35. Raissi, M.; Perdikaris, P.; Karniadakis, G.E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics* **2019**, *378*, 686–707.
36. Wong, J.C.; Chiu, P.H.; Ooi, C.; Dao, M.H.; Ong, Y.S. LSA-PINN: Linear boundary connectivity loss for solving PDEs on complex geometry. In Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN). IEEE, 2023, pp. 1–10.
37. Jin, Y.; Wang, H.; Chugh, T.; Guo, D.; Miettinen, K. Data-driven evolutionary optimization: An overview and case studies. *IEEE Transactions on Evolutionary Computation* **2018**, *23*, 442–458.
38. Yuan, G.; Zhuojia, F.; Jian, M.; Xiaoting, L.; Haitao, Z. CURRICULUM-TRANSFER-LEARNING BASED PHYSICS-INFORMED NEURAL NETWORKS FOR LONG-TIME SIMULATION OF NONLINEAR WAVE. *力学学报* **2023**, *56*, 1–11.
39. Wang, Y.; Zhong, L. NAS-PINN: neural architecture search-guided physics-informed neural network for solving PDEs. *Journal of Computational Physics* **2024**, *496*, 112603.
40. Battaglia, P.W.; Hamrick, J.B.; Bapst, V.; Sanchez-Gonzalez, A.; Zambaldi, V.; Malinowski, M.; Tacchetti, A.; Raposo, D.; Santoro, A.; Faulkner, R.; et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261* **2018**.
41. Gupta, A.; Mishra, B. Neuroevolving monotonic PINNs for particle breakage analysis. In Proceedings of the 2024 IEEE Conference on Artificial Intelligence (CAI). IEEE, 2024, pp. 993–996.
42. Gokhale, G.; Claessens, B.; Develder, C. Physics informed neural networks for control oriented thermal modeling of buildings. *Applied Energy* **2022**, *314*, 118852.
43. Deb, K.; Ehrgott, M. On Generalized Dominance Structures for Multi-Objective Optimization. *Mathematical and Computational Applications* **2023**, *28*, 100.
44. Banerjee, C.; Nguyen, K.; Fookes, C.; George, K. Physics-informed computer vision: A review and perspectives. *ACM Computing Surveys* **2024**, *57*, 1–38.
45. Cao, L.; Zheng, Z.; Ding, C.; Cai, J.; Jiang, M. Genetic programming symbolic regression with simplification-pruning operator for solving differential equations. In Proceedings of the International Conference on Neural Information Processing. Springer, 2023, pp. 287–298.
46. Lu, L.; Jin, P.; Karniadakis, G.E. Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. *arXiv preprint arXiv:1910.03193* **2019**.
47. Sung, N.; Wong, J.C.; Ooi, C.C.; Gupta, A.; Chiu, P.H.; Ong, Y.S. Neuroevolution of physics-informed neural nets: benchmark problems and comparative results. In Proceedings of the Proceedings of the Companion Conference on Genetic and Evolutionary Computation, 2023, pp. 2144–2151.
48. Wong, J.C.; Ooi, C.C.; Gupta, A.; Ong, Y.S. Learning in sinusoidal spaces with physics-informed neural networks. *IEEE Transactions on Artificial Intelligence* **2022**, *5*, 985–1000.

49. Viana, F.A.; Subramaniyan, A.K. A survey of Bayesian calibration and physics-informed neural networks in scientific modeling. *Archives of Computational Methods in Engineering* **2021**, *28*, 3801–3830.
50. Jin, G.; Wong, J.C.; Gupta, A.; Li, S.; Ong, Y.S. Fourier warm start for physics-informed neural networks. *Engineering Applications of Artificial Intelligence* **2024**, *132*, 107887.
51. Lu, L.; Meng, X.; Mao, Z.; Karniadakis, G.E. DeepXDE: A deep learning library for solving differential equations. *SIAM review* **2021**, *63*, 208–228.
52. Khoo, Y.; Lu, J.; Ying, L. Solving parametric partial differential equations using the neural convolution. *SIAM Journal on Scientific Computing* **2021**, *43*, A1697–A1719.
53. Musekamp, D.; Kalimuthu, M.; Holzmüller, D.; Takamoto, M.; Niepert, M. Active Learning for Neural PDE Solvers. In Proceedings of the Proceedings of the 13th International Conference on Learning Representations (ICLR), 2025.
54. Fang, Z.; Zhan, J. Deep physical informed neural networks for metamaterial design. *IEEE Access* **2019**, *8*, 24506–24513.
55. Pham, V.T.; Le, T.L.; Tran, T.H.; Nguyen, T.P. Hand detection and segmentation using multimodal information from Kinect. In Proceedings of the 2020 International Conference on Multimedia Analysis and Pattern Recognition (MAPR), 2020, pp. 1–6. https://doi.org/10.1109/MAPR49794.2020.9237785.
56. Chen, Y.; Koohy, S. Gpt-pinn: Generative pre-trained physics-informed neural networks toward non-intrusive meta-learning of parametric pdes. *Finite Elements in Analysis and Design* **2024**, *228*, 104047.
57. Stanley, K.O.; Clune, J.; Lehman, J.; Miikkulainen, R. Designing neural networks through neuroevolution. *Nature Machine Intelligence* **2019**, *1*, 24–35.
58. Psaros, A.F.; Kawaguchi, K.; Karniadakis, G.E. Meta-learning PINN loss functions. *Journal of computational physics* **2022**, *458*, 111121.
59. Penwarden, M.; Zhe, S.; Narayan, A.; Kirby, R.M. A metalearning approach for physics-informed neural networks (PINNs): Application to parameterized PDEs. *Journal of Computational Physics* **2023**, *477*, 111912.
60. Négiar, G.; Mahoney, M.W.; Krishnapriyan, A.S. Learning differentiable solvers for systems with hard constraints. *arXiv preprint arXiv:2207.08675* **2022**.
61. Han, J.; Jentzen, A.; E, W. Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences* **2018**, *115*, 8505–8510.
62. Yang, S.; Tian, Y.; He, C.; Zhang, X.; Tan, K.C.; Jin, Y. A gradient-guided evolutionary approach to training deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems* **2021**, *33*, 4861–4875.
63. Cao, L.; Lin, Z.; Tan, K.C.; Jiang, M. Interpretable Solutions for Multi-Physics PDEs Using T-NNGP **2025**.
64. Cai, S.; Mao, Z.; Wang, Z.; Yin, M.; Karniadakis, G.E. Physics-informed neural networks (PINNs) for fluid mechanics: A review. *Acta Mechanica Sinica* **2021**, *37*, 1727–1738.
65. Cai, S.; Wang, Z.; Fuest, F.; Jeon, Y.J.; Gray, C.; Karniadakis, G.E. Flow over an espresso cup: inferring 3-D velocity and pressure fields from tomographic background oriented Schlieren via physics-informed neural networks. *Journal of Fluid Mechanics* **2021**, *915*, A102.
66. Heldmann, F.; Berkhahn, S.; Ehrhardt, M.; Klamroth, K. PINN training using biobjective optimization: The trade-off between data loss and residual loss. *Journal of Computational Physics* **2023**, *488*, 112211.
67. Zhou, M.; Mei, G. Transfer Learning-Based Coupling of Smoothed Finite Element Method and Physics-Informed Neural Network for Solving Elastoplastic Inverse Problems. *Mathematics* **2023**, *11*, 2529.
68. Cohen, T.; Welling, M. Group equivariant convolutional networks. In Proceedings of the International conference on machine learning. PMLR, 2016, pp. 2990–2999.
69. Nichol, A.; Schulman, J. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999* **2018**, *2*, 4.
70. Molnar, J.P.; Grauer, S.J. Flow field tomography with uncertainty quantification using a Bayesian physics-informed neural network. *Measurement Science and Technology* **2022**, *33*, 065305.
71. Wang, Y.; Yin, D.Q.; Yang, S.; Sun, G. Global and local surrogate-assisted differential evolution for expensive constrained optimization problems with inequality constraints. *IEEE transactions on cybernetics* **2018**, *49*, 1642–1656.
72. Miikkulainen, R.; Forrest, S. A biological perspective on evolutionary computation. *Nature Machine Intelligence* **2021**, *3*, 9–15.
73. Lai, X.; Wang, S.; Guo, Z.; Zhang, C.; Sun, W.; Song, X. Designing a shape–performance integrated digital twin based on multiple models and dynamic data: a boom crane example. *Journal of Mechanical Design* **2021**, *143*, 071703.

74. Tang, K.; Wan, X.; Yang, C. DAS-PINNs: A deep adaptive sampling method for solving high-dimensional partial differential equations. *Journal of Computational Physics* **2023**, *476*, 111868.
75. Qin, T.; Beatson, A.; Oktay, D.; McGreivy, N.; Adams, R.P. Meta-pde: Learning to solve pdes quickly without a mesh. *arXiv preprint arXiv:2211.01604* **2022**.
76. Karpatne, A.; Atluri, G.; Faghmous, J.H.; Steinbach, M.; Banerjee, A.; Ganguly, A.; Shekhar, S.; Samatova, N.; Kumar, V. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on knowledge and data engineering* **2017**, *29*, 2318–2331.
77. Lau, G.K.R.; Hemachandra, A.; Ng, S.K.; Low, B.K.H. PINNACLE: PINN Adaptive ColLocation and Experimental points selection. *arXiv preprint arXiv:2404.07662* **2024**.
78. Huhn, Q.A.; Tano, M.E.; Ragusa, J.C. Physics-informed neural network with fourier features for radiation transport in heterogeneous media. *Nuclear Science and Engineering* **2023**, *197*, 2484–2497.
79. Xiong, Y.; Duong, P.L.T.; Wang, D.; Park, S.I.; Ge, Q.; Raghavan, N.; Rosen, D.W. Data-driven design space exploration and exploitation for design for additive manufacturing. *Journal of Mechanical Design* **2019**, *141*, 101101.
80. Baxter, J.; Caruana, R.; Mitchell, T.; Pratt, L.Y.; Silver, D.L.; Thrun, S. Learning to learn: Knowledge consolidation and transfer in inductive systems. In Proceedings of the NIPS Workshop, http://plato. acadiau. ca/courses/comp/dsilver/NIPS95_LTL/transfer. workshop, 1995.
81. Xu, C.; Cao, B.T.; Yuan, Y.; Meschke, G. Transfer learning based physics-informed neural networks for solving inverse problems in engineering structures under different loading scenarios. *Computer Methods in Applied Mechanics and Engineering* **2023**, *405*, 115852.
82. Lu, B.; Moya, C.; Lin, G. NSGA-PINN: a multi-objective optimization method for physics-informed neural network training. *Algorithms* **2023**, *16*, 194.
83. Mahmoudabadbozchelou, M.; Jamali, S. Rheology-informed neural networks (RhINNs) for forward and inverse metamodelling of complex fluids. *Scientific reports* **2021**, *11*, 1–13.
84. Cai, S.; Li, H.; Zheng, F.; Kong, F.; Dao, M.; Karniadakis, G.E.; Suresh, S. Artificial intelligence velocimetry and microaneurysm-on-a-chip for three-dimensional analysis of blood flow in physiology and disease. *Proceedings of the National Academy of Sciences* **2021**, *118*, e2100697118.
85. Mattey, R.; Ghosh, S. A novel sequential method to train physics informed neural networks for Allen Cahn and Cahn Hilliard equations. *Computer Methods in Applied Mechanics and Engineering* **2022**, *390*, 114474.
86. Wang, Q.; Song, L.; Guo, Z.; Li, J.; Feng, Z. A Novel Multi-Fidelity Surrogate for Efficient Turbine Design Optimization. *Journal of Turbomachinery* **2024**, *146*.
87. Yang, L.; Meng, X.; Karniadakis, G.E. B-PINNs: Bayesian physics-informed neural networks for forward and inverse PDE problems with noisy data. *Journal of Computational Physics* **2021**, *425*, 109913.
88. Kapoor, T.; Wang, H.; Núñez, A.; Dollevoet, R. Transfer learning for improved generalizability in causal physics-informed neural networks for beam simulations. *Engineering Applications of Artificial Intelligence* **2024**, *133*, 108085.
89. Gao, Y.; Cheung, K.C.; Ng, M.K. Svd-pinns: Transfer learning of physics-informed neural networks via singular value decomposition. In Proceedings of the 2022 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2022, pp. 1443–1450.
90. Willard, J.; Jia, X.; Xu, S.; Steinbach, M.; Kumar, V. Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Computing Surveys* **2022**, *55*, 1–37.
91. Al Noman, A.; Tasneem, Z.; Sahed, M.F.; Muyeen, S.; Das, S.K.; Alam, F. Towards next generation Savonius wind turbine: Artificial intelligence in blade design trends and framework. *Renewable and Sustainable Energy Reviews* **2022**, *168*, 112531.
92. Chen, Y.; Lu, L.; Karniadakis, G.E.; Dal Negro, L. Physics-informed neural networks for inverse problems in nano-optics and metamaterials. *Optics express* **2020**, *28*, 11618–11633.
93. Li, Z.; Kovachki, N.; Azizzadenesheli, K.; Liu, B.; Bhattacharya, K.; Stuart, A.; Anandkumar, A. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895* **2020**.
94. Huang, H.M.; Raponi, E.; Duddeck, F.; Menzel, S.; Bujny, M. Topology optimization of periodic structures for crash and static load cases using the evolutionary level set method. *Optimization and Engineering* **2024**, *25*, 1597–1630.
95. Ollivier, Y.; Arnold, L.; Auger, A.; Hansen, N. Information-geometric optimization algorithms: A unifying picture via invariance principles. *Journal of Machine Learning Research* **2017**, *18*, 1–65.

96. Hennigh, O.; Narasimhan, S.; Nabian, M.A.; Subramaniam, A.; Tangsali, K.; Fang, Z.; Rietmann, M.; Byeon, W.; Choudhry, S. NVIDIA SimNet™: An AI-accelerated multi-physics simulation framework. In Proceedings of the Computational Science–ICCS 2021: 21st International Conference, Krakow, Poland, June 16–18, 2021, Proceedings, Part V. Springer, 2021, pp. 447–461.
97. Davi, C.; Braga-Neto, U. Multi-Objective PSO-PINN. In Proceedings of the 1st Workshop on the Synergy of Scientific and Machine Learning Modeling@ ICML2023, 2023.
98. Zhang, T.; Yan, R.; Zhang, S.; Yang, D.; Chen, A. Application of Fourier feature physics-information neural network in model of pipeline conveying fluid. *Thin-Walled Structures* **2024**, *198*, 111693.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.