Informing Vision AI about Human Discoveries

Ayush Tewari
University Assistant Professor, Department of Engineering, University of Cambridge

Abstract: Physical perception, i.e., the ability to reason about the physical properties of a scene from visual observations, is a cornerstone of visual intelligence, enabling agents to interact meaningfully with their environment. It has broad applications across robotics, AR/VR, and computer graphics.

Recent advances have focused on reconstructing static 3D geometry by training on large datasets [4]. However, such models often fall short in understanding dynamic phenomena and complex material-light interactions. Attempts to model these effects directly from video often violate basic physical laws, revealing a critical gap in physical perception [3].

At their core, current approaches attempt to rediscover physical laws from raw data, making them data-inefficient, prone to shortcut learning [2], and fragile in generalization. This project instead asks: *Rather than forcing AI models to rediscover physics, how can we directly equip them with this knowledge?*

We propose to develop vision models that leverage centuries of physical understanding, including Newtonian mechanics, light transport, and material behavior. By embedding explicit physical knowledge into perception systems, we aim to create models that are more robust, controllable, and data-efficient.

We will pursue four primary research directions:

1. **Physical inductive biases:** How can we embed structured physical priors, such as forces, dynamics, and material behavior, into visual models to enable generalization to real-world video data?
2. **Soft and approximate constraints:** How can we balance physical plausibility with computational tractability, developing hybrid models that accommodate approximation error without sacrificing performance?
3. **Leveraging large language models (LLMs):** How can we use LLMs as knowledge distillers to inject symbolic and textbook physical knowledge into visual perception models?
4. **Continual learning and agentic AI:** How can physical priors be used to self-check and improve predictions through verifiers, enabling continual self-improvement and adaptive learning?

Together, these directions lay the foundation for a new class of physically informed visual AI—models that build upon, rather than rediscover, human knowledge. This is a project with a five year timescale, and my goal is to hire several PhD students and postdocs who will jointly work towards the larger goal.

1 Physical Inductive Biases

Current work in physical perception focuses primarily on static 3D reconstruction using synthetic datasets such as Objaverse XL [1]. Dynamic reconstruction from real-world videos remains underexplored due to the lack of structured supervision.

While recent video generative models produce compelling predictions, they lack physical controllability and often violate simple physical laws [3]. For example, objects may float in defiance of gravity, or respond unrealistically to collisions. These shortcomings arise from the absence of physical inductive biases—learned or hardcoded structures that encode how the world works.

Scope and Milestones

- Develop neural networks with inductive biases for representing 3D motion and geometry from monocular video (duration: 6–8 months).
- Extend these to include latent physical parameters (mass, friction) by building differentiable simulators or neural approximations that enable simulation under user-defined interactions (duration: 6–12 months). Here, we will focus on scenes with rigid objects.
- Scale up to complex dynamics (e.g., cloth, hair, articulated motion) with structured priors to enable generalization across object categories (duration: 12–18 months).

2 Soft Inductive Biases

The real world exhibits an enormous range of physical phenomena, many of which are too complex to simulate exactly. For instance, tracking every grain of sand flowing through a sieve is intractable, yet humans effortlessly understand the process. Our goal is to build visual systems that rely on *approximate physical models*, rather than exact simulation, while remaining robust to error.

Current learning frameworks typically assume perfect priors and supervision, e.g., known camera poses or Lambertian surfaces. In practice, physical assumptions are violated. We propose to build hybrid learning systems that combine approximate 3D models with flexible 2D reasoning layers that compensate for these violations.

Scope and Milestones

- Relax dependence on exact camera poses by developing robust hybrid models that operate under uncertain viewpoints (duration: 6–8 months).
- Investigate whether static inductive biases, combined with 2D correction mechanisms, can model dynamic scenes (duration: 6–8 months).
- Extend the models to handle complex appearance models, e.g., non-Lambertian reflectance and transparency, through hybrid 2D-3D architectures (duration: 6–8 months).

3 Leveraging LLMs

Instead of hand-designing physical priors, can we extract them from large language models? LLMs have effectively read the world's physics textbooks, and can be queried to extract structured knowledge, from equations of motion to material properties and chemical reaction dynamics.

In this direction, we aim to develop a pipeline where visual systems consult LLMs, e.g., Gemini from Google DeepMind, or GPT4 from OpenAI, to infer symbolic models or physical laws relevant to the current scene (e.g., elastic vs. rigid materials), generate constraints, architectures, or objectives for the vision system to follow, and interpret or verify outputs from a reasoning perspective, enabling physical reasoning without any individual human defining the inductive biases.

Scope and Milestones

- Build a toolkit for querying LLMs about physical properties and phenomena and integrate responses with differentiable models (duration: 4-6 months).
- Develop end-to-end prediction models where LLMs are uses to guide the training objective, but the test-time prediction is performed in a feedforward manner (duration: 6–8 months).
- Evaluate these hybrid systems on datasets with uncommon or rare physical effects, where classical training-based models fail (duration: 8–12 months).

4 Continual Learning and Agentic AI

Physical knowledge enables more than better perception—it enables self-evaluation. If a visual model violates conservation of momentum or predicts implausible motion, a physically-aware verifier can flag this. These verifiers can then be used to drive self-correction.

We will build vision systems with built-in critics that check predictions for physical consistency and trigger adaptation. Over time, these agentic systems can refine themselves, interactively query external tools (e.g., LLMs or simulators), and bootstrap their own training data.

Scope and Milestones

- Build differentiable verifiers to assess outputs from perception modules against known physical principles (duration: 6–8 months).
- Train agentic systems that iteratively refine their outputs by incorporating feedback from verifiers (duration: 8–12 months).
- Demonstrate continual learning by evaluating models on long-term or evolving video sequences (duration: 12–18 months).

5 Conclusion

This project aims to move beyond brute-force data-driven perception by infusing AI systems with human knowledge of physics and the visual world. Through structured inductive biases, soft constraints, LLM integration, and agentic learning loops, we envision a new generation of robust, controllable, and efficient visual intelligence. Over five years, the project will support multiple PhD students and postdocs, with milestones designed to advance both theory and practical capability in physical scene understanding. Over the course of these five years, we will slowly expand the scope of the project and be able to reason about more and more diverse scenes.

References

[1] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36:35799–35813, 2023.

[2] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[3] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? *arXiv preprint arXiv:2501.09038*, 2025.

[4] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025.