PhD Thesis:
Understanding and Improving Physics Capabilities of Vision Foundation Models

Inria, Paris, France

Expected starting date: September - November 2025

Advisor: Raoul de Charette (HDR, Research Director, Inria)
Co-supervisor: Tuan-Hung Vu (Senior Researcher, Inria / Valeo.ai)

Timeline and application

- Candidates are encouraged to **apply asap** and no later than **Tuesday May 20th, 2025**.
  - Apply via email (raoul.de-charette@inria.fr) with you resume, two contacts for reference, a motivation letter (1 page max.), and the copy of your latest transcripts and diplomas.
  - Criteria to select candidates are: scientific excellence (good prior publications is a plus), knowledge of foundation models, coding proficiency, academic profile. Applicants must hold a Master degree.
- Selected candidates will be interviewed (remote or onsite) May 21-27, 2025.
- Results will be communicated in two-stages: **a)** Conditional-acceptance by May 30th, **b)** Acceptance by mid-June, subject to funding from PRAIRIE-PSAI and validation from the doctoral school. The thesis start is Sept-Nov. 2025.

The PhD thesis, funded by PR[AI]RIE-PSAI, is an academic PhD thesis conducted at the Inria research lab, in Paris. We encourage diverse profiles to apply and commit to an open, transparent and merit-based application process.
Legal mention: Non-discrimination, ouverture et transparence. L'ensemble des partenaires de PR[AI]RIE-PSAI s'engagent à soutenir et promouvoir l'égalité, la diversité et l'inclusion au sein de ses communautés. Nous encourageons les candidatures issues de profils variés, que nous veillerons à sélectionner via un processus de recrutement ouvert et transparent.

1 Scientific context

With the emergence of foundation models trained on internet-scale datasets [BHA+21], scene understanding with computer vision has drastically progressed in the last years [ANK+25]. However, interaction with the physical world requires more than just object localization, semantic segmentation or geometric understanding. It demands knowledge about the physical properties of the objects and understanding of the fundamental laws of physics to better apprehend how to interact with our world. Psychophysics studies have shown that human perception relies on a physical model of the world which develops in the first months of the life [Bai94], at a stage where infants have access only to limited visual scenarios. With the rise of large language models, there has been a growing number of studies showing that foundation models exhibit some emergence of visual common sense which can be measured through spatial awareness [ZZXZ24], 3D reasoning [EBRM+24], video prediction [KYL+24] or Visual Question Answering (VQA) [CML+25]. However, such ability is not yet well understood and seems to arise as a by-product at the cost of training on billions of data samples. Despite massive amount of data, foundation models were shown to drastically fail at understanding even some of the simplest physics concepts such as gravity [LMP+25] or mechanical forces [MCS+25] not to mention complex deformation. Nevertheless, grasping these physical concepts is crucial to accurately predict the next state of the world (e.g., *Is the car moving or stationary?*, *How fast this object is falling?*, *Will the objects collide?*).

In today's real-world applications, computer vision practitioners rely on auto-regressive world models [HRY+23], forecasting [KKGK24], or video prediction [Ope24] — all of which may be seen as surrogate tasks for the underlying physics understanding. Explicitly learning physics knowledge has evident applications in growing fields like generative AI, to improve visual realism and learning efficiency per data, or robotics, where comprehensive understanding of physical laws enables better prediction of the consequences resulting from an action.

2 Thesis outline

The objective of the PhD is to improve the *explicit* understanding of physics in Vision Foundation Models (VFM). While the latter are typically trained on reconstruction objectives (*e.g.*, future or masked patch reconstruction), they have shown some emergence of physics [MCS+25, GBA+25] but still fail at accurately modeling basic physics. In this context, the goal is to enable *physics-grounded VFM* capable of understanding Newtonian dynamics from real-world videos, for example being able to predict pixel-wise rigid body dynamics (*e.g.*, gravity, forces, motion, *etc.*) that model the macro interactions between

objects in the scene. The resulting physics-grounded VFM are expected to showcase better capability at downstream applications relying on implicit physics modeling such as forecasting, motion estimation, scene parsing, *etc.*

To address the challenge of physics-grounded VFM, the PhD thesis will be articulated in three chronological phases. In year 1 (Sec. 2.1), we will start by evaluating the ability of existing VFMs to capture physics knowledge and propose a physics benchmark building on prior work of our team that produces real-world videos with pixel-level physics annotation. In year 2 (Sec. 2.2), we will seek the incorporation of physics knowledge in VFM, designing physics-grounded VFM, for example by finetuning existing VFMs with explicit frugal physics-learning objectives. Finally, year 3 (Sec. 2.3) will explore other forms of trainings, and the ability to extend our findings to generative AI or dynamics of non-rigid bodies.

Scientific outcomes are expected to be published in top-tier computer vision (CVPR, ICCV, ECCV) and machine learning venues (NeurIPS, ICLR, ICML), with *all results open sourced to the community*. Further, we will encourage international visits and collaborations outside the lab, for example with an ELLIS Lab.

2.1 Evaluating physics in VFM (1st year)

Measuring the physics ability of VFMs on real-world videos would require physically annotated datasets – a virtually impossible endeavor. Instead, we will build on our recent work [PVBC25] that animates real-world scenes, captured with 3D Gaussian Splatting [KKLD23], from the emulation of the particles with controllable Newtonian forces [XZQ+24]. Different from VFM benchmarking datasets, which are either confined to 2D simulations [KYL+24] or providing only sequence level annotation [MCS+25], our method enables *pixel-level annotations of forces* (gravity, friction, normal, *etc.*). As a result, this will extend the literature, currently limited to generative evaluation, enabling predictive evaluation of physics reasoning in VFMs.

A key aspect of this study will be the definition of physics-aware metrics beyond VQA [CML+25] and motion prediction [MCS+25] which only partially capture physics understanding. To address this, we will propose a combination of low level metrics such as pixel-wise error of the prediction, made possible by our pixel annotations, and high level metric measuring 'surprise' of the predicted latent [GBA+25]. Following this evaluation protocol will allow precise evaluation of the physics understanding of VFMs. A question that arise from such study is whether VFM trained only on vision (*i.e.*, Large Vision Model [BGM+24]) exhibit better physical understanding than those trained on modality-alignment objectives (*e.g.*, Vision-Language Model or Multimodal Large Language Model).

2.2 Physics-grounded VFM (2nd year)

The second phase of the PhD will focus on developing VFMs endowed with physics-aware representations using physics-driven learning objectives. Rather than full training of VFMs, to benefit from the large-scale priors learned, as well as for computational efficiency, we will prioritize leveraging existing VFMs – pretrained for discriminating semantic and geometric concepts [ZHH+23] – while seeking to incorporate physical laws and modeling the underlying physical interactions among semantic entities [BPE25].

We will explore different training strategies including (i) supervised learning on our generated dataset with detailed physics annotations (see Sec. 2.1) and (ii) self-supervised learning via proxy objectives designed to implicitly encourage physics modeling, such as the minimization of the surprise [BPE25] or the prediction of falling speed of synthetic objects [LMP+25]. Different from the literature, here the physics ability will no longer be *emerging* as a by-product from pattern correlation or memorization by large models from vasts amount of data, but be enabled by the *explicit* physics-driven learning objectives. We thus expect this grounding stage to be considerably more frugal than the original pre-training, thanks to the more explicit supervision.

A key practical challenge will be assessing the benefit of physics-aware representations over standard VFM representations in down-stream tasks, which are dominantly centered on semantic, geometric and dynamic reasoning. However, these tasks often lack mechanisms for capturing deeper physical reasoning, such as force interactions, stability, or causality.

2.3 Extension to other Trainings, Models and Forces (3rd year)

The last phase of the PhD will be explanatory, building on the findings of the previous years. We envisage several possible directions to extend to other forms of trainings, models and forces. **(i) Trainings** – Building on the insights that large models still benefit from representational alignment [YKJ+25], we will explore the effect of alignment with physics-grounded representations on the performance and ability to produce physically plausible outputs of downstream tasks. **(ii) Models** – While the PhD thesis focuses on discriminative VFMs, we will study the transposition of our findings to generative training, with the expectation to generate more physically realistic/plausible videos [LMP+25]. **(iii) Forces** – In addition to rigid body dynamics, a possible direction will be to investigate physics-grounded VFM in the case of non-rigid dynamics such as fluids, deformable dynamics or optics which have been studied only from the lens of generative AI [MCS+25].

3 Institution

The candidate will join the Astra project-team in Inria Paris. The team develops technologies linked to achieve sustainable mobility, improving safety and ensuring efficient road transport. Astra is structured into several research axes.

The PhD candidate will work in the Astra-Vision group (https://astra-vision.github.io) addressing robust visual scene understanding. Its research focuses on relaxing data and supervision while providing more interpretable models outputs. The group publishes in all major venues of computer vision and machine learning, and produce almost exclusively open source research. In the team, the PhD candidate is expected to actively contribute to group readings, seminars, discussions, team spirit, etc.

References

[ANK+25] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation models defining a new era in vision: a survey and outlook. *IEEE TPAMI*, 2025. 1

[Bai94] Renée Baillargeon. Physical reasoning in young infants: Seeking explanations for impossible events. *British Journal of Developmental Psychology*, 1994. 1

[BGM+24] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. In *CVPR*, 2024. 2

[BHA+21] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv*, 2021. 1

[BPE25] Anand Bhattad, Konpat Preechakul, and Alexei A Efros. Visual jenga: Discovering object dependencies via counterfactual inpainting. *arXiv*, 2025. 2

[CML+25] Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *ICLR*, 2025. 1, 2

[EBRM+24] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *CVPR*, 2024. 1

[GBA+25] Quentin Garrido, Nicolas Ballas, Mahmoud Assran, Adrien Bardes, Laurent Najman, Michael Rabbat, Emmanuel Dupoux, and Yann LeCun. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *arXiv*, 2025. 1, 2

[HRY+23] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv*, 2023. 1

[KKGK24] Efstathios Karypidis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. Dino-foresight: Looking into the future with dino. *CoRR*, 2024. 1

[KKLD23] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 2023. 2

[KYL+24] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model? – a physical law perspective. *arXiv*, 2024. 1, 2

[LMP+25] Chenyu Li, Oscar Michel, Xichen Pan, Sainan Liu, Mike Roberts, and Saining Xie. Pisa experiments: Exploring physics post-training for video diffusion models by watching stuff drop. *arXiv*, 2025. 1, 2

[MCS+25] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models learn physical principles from watching videos? *arXiv*, 2025. 1, 2

[Ope24] OpenAI. Sora: Generating videos with realistic motion and consistent appearance. https://openai.com/sora, 2024. 1

[PVBC25] Soumava Paul, Tuan-Hung Vu, Andrei Bursuc, and Raoul de Charette. Can Vision Foundation Models understand physics? *(in preparation)*, 2025. 2

[XZQ+24] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *CVPR*, 2024. 2

[YKJ+25] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *ICLR*, 2025. 2

[ZHH+23] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. In *NeurIPS*, 2023. 2

[ZZXZ24] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. A general protocol to probe large vision models for 3d physical understanding. *NeurIPS*, 2024. 1