PhD thesis (3 year):
Simulatable Physics-aware World Models

Inria, Paris, France

Starting date: September – November 2026

Advisor: Raoul de Charette, PR[AI]RIE-PSAI Fellow (HDR, Research Director, Inria)
Co-supervisor: Fabio Pizzati (Research Scientist, MBZUAI)

Timeline and application

- Candidates are encouraged to **apply asap** and no later than **Sunday May 17th, 2026**.
  - Apply via email (raoul.de-charette@inria.fr) with: your resume, names and contact details of 2-3 referees, a short motivation letter (1 page max.), and the copy of your latest transcripts and diplomas.
  - Criteria to select candidates are: scientific excellence (prior publications in good venues is a plus), knowledge of world models, coding proficiency, academic profile. Applicants must hold a Master degree.
- Applications will be reviewed, and candidates interviewed, on a *rolling basis*. Besides, we are committed to non-discrimination, openness, and transparency. All PR[AI]RIE-PSAI partners are committed to supporting and promoting equality, diversity, and inclusion within their communities. We encourage applications from a wide range of backgrounds, which we will review through an open and transparent recruitment process.
- Results will be communicated in two-stages: **a)** Conditional-acceptance by May 22nd, **b)** Acceptance by mid-June, subject to funding from PR[AI]RIE-PSAI and validation from the doctoral school. The thesis start is Sept-Nov. 2026.

The PhD thesis, funded by PR[AI]RIE-PSAI will be conducted at the Inria Paris research lab in the Astra-Vision group group, within Astra research team.

Similar to all our previous works, the scientific outcomes are expected to be published in top-tier computer vision (CVPR, ICCV, ECCV) and machine learning venues (NeurIPS, ICLR, ICML), and results, models and code open sourced to the community. In the PhD, we will encourage international visits and collaborations outside the lab, whether with other labs of the PRAIRIE cluster or the ELLIS network.

1 Scientific context

Real-world computer vision systems rely on visual data to estimate the state of the world and to infer its evolution. For safety-aware applications, like robotics, the system's reliability directly depends on its ability to accurately estimate future states as they may influence the actions taken.

Recently, **world models** [AAB+25, HS18] have emerged as powerful implicit state models capable of predicting proxy of future states (*e.g.*, next frame prediction) from visual observations (*e.g.*, single or multi frame conditioning). However, for many applications like robotics [AAB+25], they still lack *controllability* (*i.e.*, precise control of the model output), *interpretability* (*i.e.*, ability to understand the models' reasoning), and *physics grounding* [CPV+26, ZCT+25, EBRM+24] (*i.e.*, adherence to physical laws). They also lack data-efficiency and require massive amounts of data sequences, arguably because they learn correlations at scale rather than discovering the causal representations of the world. Indeed, current world models are trained on self-supervised reconstruction tasks, like pixel-wise prediction of future frame(s) or prediction of next action(s) [AAB+25, LeC22], which has unlocked training on internet-scale datasets [ANK+25]. However, the nature of the representation learned remains unclear, and recent works have shown that they lack physical expressivity [CPV+26]. Contrary, cognitive science have long demonstrated that in the early stage of development [GUT11] humans acquire a causal model of the world which explain *why and how objects move*. This raises a central question for AI: *Which data is needed to learn a causal representation of the world?*.

In this thesis, **we aim at learning physics-aware world models, relying on coarse synthetic representations** which are expected to expose better interpretability and generalization capacity. The thesis topic also positions itself in a global landscape where major companies (AMI Labs, WorldLabs, *etc.*.) seek the development of more interpretable world models that also bypass the limitation of existing generative models highly sensitive to hallucinations and inevitable sensor malfunctions. Besides realism, grounding world models to physics-aware representations is crucial for computer vision agents (robots, cars, *etc.*) which evolve in a physical world.

2 Thesis outline

The aim of the PhD thesis is to develop **physics-aware world models**. To do so, instead of only optimizing a latent state of the world, we propose to develop models which jointly optimize a *simulatable* synthetic representation of the world.

Instead of seeking exhaustive digital twins, we will build on recent findings showing that coarse synthetic scenes (*e.g.*, made of primitives) can serve as powerful support basis to learn the dynamics of the world even transferable to complex tasks like video generation [RBL+25] or planning [FGZ+26]. Similarly, we expect that simple synthetic dynamics (*e.g.*, collision, gravity, *etc.*) can be used to steer pretrained models towards physical awareness. Preliminary literature [LGL+25] also suggests that using synthetic data not only enables a more nuanced representation of the physical world, but also allows for explicit modeling of physically relevant properties within images and videos. This is crucial as having access to such explicit physical representations of a scene enables direct interventions on the world representations, depending on explicit physics.

The core of the PhD will consist of learning world models that not only predicts future outcomes given a sequence of actions, but also maintain an explicit physical representation of those outcomes, grounded in the elements present in the scene. For example, when processing a video of a bouncing ball, the model should associate the motion with physically meaningful features such as the coefficient of restitution or the material properties of the ball. By making these physical representations explicit, we can guide the generation of future states by controlling the evolution of these features, similar to how physical parameters are manipulated in simulation software such as Blender. This ultimately enables the learning of simulatable world representations, where outcomes can be controlled and grounded to real physics.

The PhD thesis will leverage world models for **video generation** as it is a well-established technique, requiring no annotations for training. We will benefit from existing models trained at scale on real-world data which we will finetune on our datasets created ad hoc or existing ones, at reasonable computation cost, to learn more nuanced physics with improved controlability. The PhD timeline will be articulated as follows: In year 1 (Sec. 2.1), we will explore ad-hoc **models for encoding and transferring synthetic knowledge** from environments of varying level of realism (texture, mesh, lighting, dynamics) to world models. In Year 2 (Sec. 2.2), we will address how to **jointly learning simulatable representation of the world**, by finetuning existing world models. Lastly, in Year 3 (Sec. 2.2) we will explore how to effectively integrate synthetic data into **larger-scale training of video world models**, and will study the properties of simulatable representations when heterogeneous synthetic datasets, capturing different types of physical phenomena, are used jointly during training.

2.1 Studying trade-offs in simulation realism (Year 1)

Synthetic simulated data can enable better world understanding. However, graphics-based simulations vary in complexity and realism depending on the effects modeled by the rendering engine (*e.g.*, ray tracing, specularity, rigid body dynamics, among others). Nowadays, relatively simple synthetic data are used for fine-tuning large video models to gain physical understanding [GHF+25, RBL+25], but there is still limited understanding of how the different factors that contribute to visual realism affect learning physical phenomena from such data. This question is fundamental and will lay the foundation for training data in the context of all the thesis. On one hand, highly realistic simulations can make data generation computationally expensive and difficult to scale. On the other, overly simplified data may limit the effectiveness of training. We therefore aim to answer the following question: "***which characteristics must be modeled to enable effective training on synthetic data?***".

To address this, we will construct a dataset starting from highly realistic simulations for example, leveraging the 3D generative models [SBH+26], which include high-quality textures, shading, ray-tracing effects, and dynamics, and progressively simplify them by systematically removing these elements to isolate their effects. Each dataset variant will focus on a specific physical phenomenon of interest, such as **rigid body dynamics, fluid dynamics, or optics**. We will then finetune open-source video generation models on these datasets variants, and analyze how realism affect the models' ability to synthesize the corresponding physical phenomena in generated videos. This will allow us to shed light on the trade-off between simulation realism and training effectiveness, providing insights that are broadly relevant to the thesis.

We will extend the insights of the first project as a stepping stone for a second project related to physical control enabled by training on synthetic data. More specifically, we aim to introduce explicit control over physical parameters of the scene, such as lighting direction and object material properties, by encoding them as latent conditioning vectors for the generation process, exploiting synthetic annotations [RBL+25]. First, this will require specific insights on data realism for *control*, rather than synthesis, drawing on the insights of the first project and expanding them. Moreover, it will enable seeking the optimal architecture and strategy to encode physical information as control signal, with new methodological contributions.

2.2 Joint simulatable world model generation (Year 2)

In the second year we will focus our efforts on the architectural and training contribution to learn a simulatable synthetic representation of the world directly derived from the world models latents. While existing world models encodes a latent representation at a given state, it is hardly interpretable and controllable.

Here, we will **jointly optimize an explicit coarse representation of the scene** following insights from the first year that will drive the granularity of the required synthetic details. The complexity here lies in the ability to decode such synthetic representation, as well as how to supervise it. We envision to explore two substantially different strategies. (a) We will leverage a pretrained synthetic autoencoder (trained on datasets of Year 1) to **align the generated representation** [FGZ+26] to

those of similar scenes, either using a real-to-sim model [TSL+24] or distribution alignments objectives similar to generative models [YKJ+25]. (b) We will explore **post-hoc reward functions** that encourages plausible synthetic representations, extending preliminary research that demonstrated the transferability of physics awareness via specific rewards [LZKS25]. Beyond learning synthetic representations, it will be used as a supervision signal in the world model, therefore leading to more controllable and physics-aware representations. Besides, we highlight that having an explicit (synthetic) world model allows greater diversity of generation, as states (such as position) of individual objects can be altered to produce varying futures.

2.3 Synthetic data at scale (Year 3)

Now, after establishing (Year 1) trainings best practices and control methodologies and (Year 2) joint latent and synthetic world modeling, our goal will be to analyze the scaling of the proposed system, investigating how far synthetic data can be pushed for training. In particular, existing world models are particularly greedy and require massive amounts of real-world data. Here we will benefit from our joint representation to produce large scale synthetic data which exhibit greater diversity and controllability, while seeking to reduce the amount of real-world data. The main project will focus on training a medium-scale foundation video model on a joint dataset composed of both synthetic and real data, potentially with different relative sizes. This will allow to draw insights on the scaling laws of synthetic data. In particular, we will analyze the optimal balance between synthetic and real data, as well as the benefits of synthetic data for learning physical phenomena. We also aim to quantify the generalization of the trained system, potentially enabling processing real data in very different scenarios, or transferring the representation to other tasks (planning, 3D reconstruction, *etc.*). To make this feasible, we will reduce the computational cost by building on a pretrained model and limiting the number of trainable parameters, while still targeting a realistic mid-scale training setup.

3 Institution

The candidate will join Inria Paris, a dynamic and internationally-renowned centre which is established as a scientific and technological leader. They will be part of the Astra project-team which develops technologies linked to achieve sustainable mobility, improving safety and ensuring efficient road transport.

The PhD candidate will work in the Astra-Vision group (https://astra-vision.github.io), which addresses robust visual scene understanding by relaxing data and supervision while providing more interpretable models outputs. The group publishes regularly in all major venues of computer vision and machine learning, primarily producing open source research, and has received notable awards and grants. The candidate is expected to actively contribute to group readings, seminars, discussions, team spirit, etc.

References

[AAB+25] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 1

[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

[CPV+26] Sebastian Cavada, Soumava Paul, Tuan-Hung Vu, Andrei Bursuc, and Raoul de Charette. Newtphys: Do foundation models understand newtonian physics? *to appear on arXiv*, 2026. 1

[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

[FGZ+26] Lan Feng, Yang Gao, Eloi Zablocki, Quanyi Li, Wuyang Li, Sichao Liu, Matthieu Cord, and Alexandre Alahi. Rap: 3d rasterization augmented end-to-end planning. *ICLR*, 2026. 2

[GHF+25] Nate Gillman, Charles Herrmann, Michael Freeman, Daksh Aggarwal, Evan Luo, Deqing Sun, and Chen Sun. Force prompting: Video generation models can learn and generalize physics-based control signals. *NeurIPS*, 2025. 2

[GUT11] Noah D Goodman, Tomer D Ullman, and Joshua B Tenenbaum. Learning a theory of causality. *Psychological review*, 2011. 1

[HS18] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *NeurIPS*, 2018. 1

[LeC22] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *preprint*, 2022. 1

[LGL+25] Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Chih-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, et al. Diffusion renderer: Neural inverse and forward rendering with video diffusion models. In *CVPR*, 2025. 2

[LZKS25] Minh-Quan Le, Yuanzhi Zhu, Vicky Kalogeiton, and Dimitris Samaras. What about gravity in video generation? post-training newton's laws with verifiable rewards. *arXiv*, 2025. 3

[RBL+25] David Romero, Ariana Bermudez, Hao Li, Fabio Pizzati, and Ivan Laptev. Learning to generate rigid body interactions with video diffusion models. *arXiv preprint arXiv:2510.02284*, 2025. 2

[SBH+26] Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, Sanja Fidler, Jiahui Huang, Huan Ling, Jun Gao, and Xuanchi Ren. Lyra 2.0: Explorable generative 3d worlds. *arXiv*, 2026. 2

[TSL+24] Marcel Torne, Anthony Simeonov, Zechu Li, April Chan, Tao Chen, Abhishek Gupta, and Pulkit Agrawal. Reconciling reality through simulation: A real-to-sim-to-real approach for robust manipulation. *RSS*, 2024. 3

[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. 3

[ZCT+25] Chenyu Zhang, Daniil Cherniavskii, Antonios Tragoudaras, Antonios Vozikis, Thijmen Nijdam, Derck WE Prinzhorn, Mark Bodracska, Nicu Sebe, Andrii Zadaianchuk, and Efstratios Gavves. Morpheus: Benchmarking physical reasoning of video generative models with real physical experiments. *arXiv*, 2025. 1