

PhD thesis (3 year): Multimodal physically-consistent world modeling

Inria, Paris, France

Starting date: January - March 2027

Supervisor: Fabio Pizzati (Research Scientist, PEPR Chaire d’Attractivité, Inria)

Co-Supervisor: Raoul de Charette (HDR, Research Director, Inria)

Timeline and application

- There are **two ways to apply** to this PhD position:
 - **Option 1 (preferred): ELLIS PhD Program – Deadline: October 31st, 2026.** Candidates are strongly encouraged to apply through the [ELLIS PhD Program](#), specifying **Raoul de Charette as an ELLIS advisor**. Priority will be given to candidates applying through this route.
 - **Option 2: Direct application – Deadline: November 20th, 2026.** Candidates can alternatively apply directly by sending an email to raoul.de-charette@inria.fr and pizzati.fabio@gmail.com, including their CV, names and contact details of 2–3 referees, a short motivation letter (1 page max.), and copies of their latest transcripts and diplomas.
- Selection criteria include scientific excellence (prior publications in top-tier venues are a plus), knowledge of world models and generative AI, coding proficiency, and academic background. Applicants must hold a Master’s degree.
- Applications will be reviewed, and candidates interviewed, on a *rolling basis*. We are committed to non-discrimination, openness, and transparency, and actively support equality, diversity, and inclusion within our research community. We particularly encourage **women and candidates from underrepresented minorities** to apply. Applications from all backgrounds are welcome and will be reviewed through a fair, open, and transparent recruitment process.

The PhD thesis, funded by PEPR AI, will be conducted at the [Inria Paris](#) research lab in the [Astra-Vision group](#), within the 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 academia and industry.

1 Scientific context

World models aim to learn representations of the world that enable predicting its evolution from observations and actions [HS18, AAB⁺25]. Recent advances in video generation have made it possible to synthesize increasingly realistic future scenes, opening promising applications in robotics, autonomous driving, and interactive environments. However, despite their visual realism, these models still struggle to reproduce fundamental physical phenomena, including object collisions, contact dynamics, and conservation laws [TZC⁺26, YPP⁺26]. Recent physics-guided generative approaches have started addressing these limitations [RBI⁺26], but important questions remain about whether existing models learn the underlying physical mechanisms governing a scene or mainly reproduce visual correlations observed during training.

In the physical world, events are inherently multimodal. A collision between two objects, for example, produces not only visual motion but also acoustic signals, changes in acceleration, and potentially tactile or force measurements. Each modality provides complementary information about the same physical interaction, which may be difficult to infer from visual observations alone. Recent developments in multimodal representation learning have demonstrated the possibility of jointly encoding heterogeneous sensory signals, including images, audio, depth, and inertial measurements [GENL⁺23]. Similarly, emerging audiovisual generative models suggest that joint training across modalities can improve visual generation, particularly for object interactions [WLH⁺25]. Yet, it remains unclear whether these improvements reflect a stronger understanding of physical phenomena, or merely better statistical alignment between different observations. While recent research has demonstrated that physical plausibility can emerge from the geometry of pretrained visual representations [IPI⁺26], we still have limited understanding of how physical knowledge is internally represented and shared across modalities.

This raises a fundamental question: *Can complementary information across sensory modalities enable world models to learn more physically consistent representations and predictions?* In this thesis, we aim to investigate how physical

knowledge is learned and represented across different modalities, how this knowledge can be exploited to improve multimodal generative models, and how effective multimodal learning strategies can be developed at scale. Beyond perceptual realism and audiovisual synchronization, our ambition is to move towards world models that maintain a physically coherent representation of events and their consequences. Such models would have important implications for generative media and embodied AI, where reliable predictions of the physical world are essential, and could ultimately enable world models with verifiable physical consistency for well-defined physical phenomena.

2 Thesis outline

The aim of this PhD thesis is to develop physically consistent multimodal world models by exploiting complementary physical information across sensory modalities. Rather than assuming that jointly processing different signals automatically leads to better physical understanding, we propose to investigate which physical properties are learned from each modality, how these properties are internally represented and shared, and which learning strategies are most effective. We will initially focus on audiovisual data, which can be collected at scale from existing video datasets [CXVZ20], while progressively considering additional sensors, such as inertial, tactile, and force measurements, particularly relevant for embodied applications [GENL+23, YMZ+22].

The central hypothesis of this thesis is that different modalities may acquire different levels of physical understanding, even when jointly trained on the same observations. For example, acoustic signals may provide stronger evidence about contact events, while visual observations provide more information about object trajectories. Understanding these differences, and the internal mechanisms through which physical representations emerge and interact, could enable the design of more effective multimodal training objectives, allowing physical knowledge acquired from one modality to improve the representations and predictions of others. Importantly, we will seek computationally efficient strategies that exploit existing pretrained models, rather than requiring large-scale training from scratch.

The thesis will be articulated into three research directions. In Year 1 (Sec. 2.1), we will investigate how physical knowledge emerges across modalities in existing multimodal models. We will first evaluate physical consistency within and across generated modalities, potentially introducing a dedicated benchmark. We will then investigate how physical concepts are internally represented and shared across modalities using interpretability methods. These findings will establish the experimental foundations and identify the physical relationships to exploit throughout the thesis.

In Year 2 (Sec. 2.2), we will leverage these insights to investigate new training and inference-time strategies for improving physical consistency across modalities. We envision exploring cross-modal regularization, representation alignment, and modality-dependent learning mechanisms, including mixture-of-experts architectures and iterative refinement. We will also investigate their applicability to action-conditioned world models, where complementary signals such as vision and touch may improve the prediction of physical interactions.

Lastly, in Year 3 (Sec. 2.3), our longer-term ambition will be to investigate how multimodal physical learning can be generalized to arbitrary sensor modalities and scaled efficiently. We will explore automatic strategies to identify useful modality combinations, discover appropriate learning objectives, and optimize multimodal training. This may include LLM-based research agents for hypothesis generation and experimental exploration [LLL+24], together with compute-efficient approaches for transferring insights from small-scale experiments to larger models [MTB+25]. Ultimately, the thesis aims to establish principles for developing scalable multimodal world models with increasingly reliable physical representations, contributing towards the longer-term objective of physically consistent world modeling with verifiable constraints.

2.1 Understanding multimodality in world models (Year 1)

Multimodal models are increasingly trained to jointly process or generate visual, acoustic and other sensor observations of the same scene [GENL+23, WLH+25]. Yet, it remains unclear whether different modalities acquire the same understanding of underlying physical events. For example, in a billiards sequence, a model may generate an impact sound at the expected instant while depicting the second ball moving before contact, or conversely produce plausible trajectories with a mistimed acoustic response. Such discrepancies raise a fundamental question: when trained on the same multimodal data, which modalities learn physical plausibility and causal relationships more effectively, and why?

The first project will investigate the physical consistency of multimodal outputs, initially focusing on pretrained audiovisual generative models. We aim to understand whether different modalities exhibit different levels of physical plausibility, and whether their joint generation produces physically consistent outcomes. Building on existing evaluations [YPP+26, LGS+26, XLH+25], we envision a targeted benchmark of controlled audiovisual interactions, considering collisions, contact dynamics and material-dependent phenomena. For example, a visible collision should produce a compatible acoustic event, while an audible impact should correspond to a plausible visual interaction when its source is observable. Through controlled variations of collision dynamics, materials or event ordering, we will investigate which physical relationships are captured by each modality and when inconsistencies emerge. This will provide a systematic understanding of the strengths and limitations of existing multimodal generators.

Building on these findings, the second project will investigate how physically meaningful representations emerge within multimodal networks, and how physical knowledge is shared across modalities. Recent studies have identified physical

quantities within intermediate representations of visual world models [JGB⁺26, KSTD26], while other findings suggest that physical plausibility can be characterized through geometric properties of pretrained visual features [IPI⁺26]. We aim to extend these investigations to multimodal architectures, studying whether concepts such as acceleration, restitution or material properties can be identified within modality-specific representations, and whether the same physical properties are encoded differently across modalities. We will employ representation probing, layer-wise analyses and controlled interventions, exploiting the phenomena isolated in the first project. We will further study whether multimodal networks learn independent physical representations or develop shared concepts through cross-modal interactions. For instance, can acoustic features encoding an impact influence the visual representation of the resulting motion, and which network components mediate this interaction? These investigations will reveal how complementary physical knowledge emerges and is exchanged across modalities, providing foundations for the training strategies explored in Year 2.

2.2 Aligning physics across modalities (Year 2)

Building on the first-year analysis, we will investigate how complementary physical knowledge across modalities can be exploited to improve the physical plausibility of multimodal generators. If one modality learns a particular physical relationship more reliably or with fewer examples, can this knowledge provide a training signal for the other modalities? Rather than encouraging generic feature similarity, we will explore physics-aware cross-modal objectives based on the phenomena and internal representations identified in Year 1, such as contact timing, event ordering, motion consequences and consistency between observed interactions. Possible strategies include cross-modal representation regularization, event-level distillation and adaptive weighting of modality-specific training signals, inspired by representation alignment techniques [YKJ⁺25].

We will also study how such signals can improve the training of multimodal mixture-of-experts transformers [LJH⁺25]. For example, can expert specialization or routing be driven by the physical information carried by each modality, so that acoustic evidence is emphasized for impacts while visual evidence remains central for trajectories? Beyond training strategies, we will investigate whether the physical representations and cross-modal mechanisms identified in Year 1 can be exploited at inference time. In particular, iterative refinement strategies, potentially based on looped transformers [FYC⁺26], could allow multimodal models to progressively exchange physical information across modalities and improve the consistency of their predictions without extensive retraining.

Beyond audiovisual generation, we envision extending this investigation to action-conditioned world models, where the physical consequences of an action are particularly important [YYZ⁺26]. In robotic manipulation, for instance, a gripper closing around an object may look successful in video while tactile or force measurements reveal insufficient contact or incipient slip; such observations could supervise physically meaningful predictions of object motion and action outcomes [YMZ⁺22, HSB⁺25]. A central objective will be to assess whether these cross-modal learning strategies yield gains under comparable data and compute budgets. We will primarily build on pretrained backbones and parameter-efficient adaptation, avoiding the need to train large multimodal models from scratch.

2.3 Multimodal learning with arbitrary modalities (Year 3)

Building on the understanding of multimodal physics acquired in Year 1 and the training methods developed in Year 2, our longer-term ambition will be to move from hand-designed modality combinations toward physically consistent world models capable of exploiting a broad range of sensors. Existing approaches already connect heterogeneous signals such as video, audio, depth and inertial measurements [GENL⁺23], but offer limited guidance on which modalities actually improve learned physical dynamics or how they should be integrated efficiently. A central question is therefore: can we automatically discover which modalities are useful for a given physical phenomenon, how they should interact, and how these choices should evolve with model scale?

To address this ambitious objective, we envision exploring automated modality selection, fusion mechanisms, training objectives and curricula. In particular, LLM-based research agents could exploit natural-language descriptions of available sensors and the physical diagnostics developed earlier in the thesis to propose hypotheses, conduct controlled experiments and refine multimodal training strategies [LLL⁺24]. Since exhaustive foundation-model training would be computationally prohibitive, we will investigate whether experiments with frozen pretrained encoders, compact proxy models and parameter-efficient adaptation can identify strategies that transfer to larger world models [MTB⁺25]. Beyond a particular architecture, the goal will be to establish principles and tools for scalable multimodal learning, contributing toward the long-term vision of world models that produce physically consistent predictions and, for well-defined phenomena, eventually satisfy verifiable physical constraints.

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
- [CXVZ20] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020. 2
- [FYC⁺26] Tianyu Fu, Yichen You, Zekai Chen, Guohao Dai, Huazhong Yang, and Yu Wang. Think-at-hard: Dynamic looped transformers for improved reasoning. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306, pages 32129–32154. PMLR, 2026. 3
- [GENL⁺23] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 1, 2, 3
- [HS18] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, 2018. 1
- [HSB⁺25] Carolina Higuera, Akash Sharma, Chaithanya Krishna Bodduluri, Taosha Fan, Patrick Lancaster, Mrinal Kalakrishnan, Michael Kaess, Byron Boots, Mike Lambeta, Tingfan Wu, and Mustafa Mukadam. Sparsh: Self-supervised touch representations for vision-based tactile sensing. In *Proceedings of the 8th Conference on Robot Learning*, volume 270, pages 885–915. PMLR, 2025. 3
- [IPI⁺26] Christian Internò, Alexander Pondaven, Habon Issa, Fabio Pizzati, Francesco Pinto, Markus Olhofer, Ivan Laptev, Philip Torr, Eero P. Simoncelli, Barbara Hammer, and David Klindt. GeoPhys: The geometry of physical plausibility. *arXiv preprint arXiv:2606.20707*, 2026. 1, 3
- [JGB⁺26] Sonia Joseph, Quentin Garrido, Randall Balestriero, Matthew Kowal, Thomas Fel, Shahab Bakhtiari, Blake Aaron Richards, and Michael Rabbat. Interpreting physics in video world models. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306, pages 54687–54724. PMLR, 2026. 3
- [KSTD26] Jonas Kneifl, Jakub Skalski, Bartłomiej Twardowski, and Kamil Deja. Does physics live in the activations? localizing physical quantities in video diffusion models. *arXiv preprint arXiv:2610.03154*, 2026. 3
- [LGS⁺26] Tingle Li, Siddharth Gururani, Kevin J. Shih, Gantavya Bhatt, Sang-gil Lee, Zhifeng Kong, Arushi Goel, Gopala Anumanchipalli, and Ming-Yu Liu. Benchmarking single-factor physical video-to-audio generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1939–1949, 2026. 2
- [LJH⁺25] Yunxin Li, Shenyan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Uni-MoE: Scaling unified multimodal LLMs with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3424–3439, 2025. 3
- [LLL⁺24] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024. 2, 3
- [MTB⁺25] Ian Magnusson, Nguyen Tai, Ben Bogin, David Heineman, Jena D. Hwang, Luca Soldaini, Akshita Bhagia, Jiacheng Liu, Dirk Groeneveld, Oyvind Tafjord, Noah A. Smith, Pang Wei Koh, and Jesse Dodge. DataDecide: How to predict best pretraining data with small experiments. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 42487–42502. PMLR, 2025. 2, 3
- [RBI⁺26] David Romero, Ariana Bermudez, Viacheslav Iablochnikov, Hao Li, Fabio Pizzati, and Ivan Laptev. Learning to generate rigid body interactions with video diffusion models. In *European Conference on Computer Vision*, 2026. 1
- [TZC⁺26] Antonios Tragoudaras, Chenyu Zhang, Daniil Cherniavskii, Antonios Vozikis, Thijmen Nijdam, Derck W. E. Prinzhorn, Mark Bodracska, Nicu Sebe, Andrii Zadaianchuk, and Stratis Gavves. Evaluating newtonian mechanics in video generative models with real physical systems. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306, pages 121912–121951. PMLR, 2026. 1
- [WLH⁺25] Jianzong Wu, Hao Lian, Dachao Hao, Ye Tian, Qingyu Shi, Bialong Chen, Hao Jiang, and Yunhai Tong. Does hearing help seeing? investigating audio-video joint denoising for video generation. *arXiv preprint arXiv:2512.02457*, 2025. 1, 2
- [XLH⁺25] Tianxin Xie, Wentao Lei, Guanjie Huang, Pengfei Zhang, Kai Jiang, Chunhui Zhang, Fengji Ma, Haoyu He, et al. PhyAVBench: A challenging audio physics-sensitivity benchmark for physically grounded text-to-audio-video generation. *arXiv preprint arXiv:2512.23994*, 2025. 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 *International Conference on Learning Representations*, 2025. 3
- [YMZ⁺22] Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. In *Advances in Neural Information Processing Systems*, 2022. 2, 3
- [YPP⁺26] Jianhao Yuan, Fabio Pizzati, Francesco Pinto, Lars Kunze, Ivan Laptev, Paul Newman, Philip Torr, and Daniele De Martini. LikePhys: Evaluating intuitive physics understanding in video diffusion models via likelihood preference. In *International Conference on Learning Representations*, 2026. 1, 2
- [YYZ⁺26] Heng Yu, David D. Yuan, Juze Zhang, Changan Chen, Yao Feng, Michelle Baldonado, Steve Cousins, Li Fei-Fei, Jiajun Wu, and Ehsan Adeli. OpenWAM: An open framework for composable world-action models. *arXiv preprint arXiv:2610.07922*, 2026. 3