Exploring Low-level Physics Understanding through Data-Centric Computer Vision

Brian Nlong Zhao
briannlongzhao@gmail.com

1 Introduction

Computer vision has become pivotal in AI research, demonstrating remarkable abilities to analyze visual data and create new instances that mirror existing distributions. These vision models are widely applied across fields like image and video recognition, 3D spatial understanding, and especially visual content creation, offering automation opportunities in content creation, data augmentation, and simulation. State-of-the-art vision models are predominantly data-driven and trained on massive datasets. For example, text-to-image diffusion models such as DALL · E and Stable Diffusion, are trained on extensive text-image pairs, enabling them to generate images based on textual descriptions. Other visual foundational models, for example Segment Anything Model (SAM), benefits from a well-designed data engine that enhances the the quality of the data at a massive scale.

While state-of-the-art vision models have shown promising results in tasks such as static image segmentation and generation, **it remains unclear whether these models possess the ability to understand low-level dynamics and physics and generate physically consistent dynamic visuals, where scenes and objects are expected to exhibit motion.** This means that the visual models that incorporate a time dimension may not adhere to physics laws or reflect the natural dynamics of our daily observations. For instance, non-physics-aware object tracking models may not understand fundamental laws of inertia and will easily mixing up tracks when encountering two similar objects moving across each other. Humans or animals may have inconsistent limb proportions, impossible postures, or distorted features in generated visual contents, if not using a predefined and constrained 3D model. Visual generation models may also fail to represent realistic dynamics, such as object motion and deformation when encountering gravity or external forces. The lack of adherence to natural laws results in physically implausible outputs, undermining the utility of vision models in applications involving high degrees of realism. Thus, **my research aims to delve deeply into how well current visual models understand low-level physics and dynamics for visual analysis and generation, and how to enhance these abilities in vision models, particularly through a data-centric approach.**

2 Proposal

The questions that I therefore wanted to ask are, can these limitations in physics understanding be addressed by carefully curating training datasets to emphasize dynamic movements, thereby enabling visual models to learn more accurate statistical representations of physical dynamics? And based on this, does scaling-up the data allow models to better capture and generalize the underlying laws of physics in visual scenarios? To explore these questions, I propose a research plan that is structured around the following key steps.

2.1 Evaluation of Low-level Physics

The first step is to identify systematic tasks and evaluation methods that reveal the shortcomings of current vision models in understanding low-level physics. This involves, first, clearly defining how to determine if a vision model is capable of understanding low-level physics. There exist some datasets focusing on the ability of vision models to learn high-level physics properties. For example [7] collects a video dataset of the motions of various objects in different classical physics scenarios, such as objects sliding off a ramp or bouncing off from a surface, and also proposes some basic physical properties for machine learning model to learn and evaluate, such as material, volume, coefficient of friction, etc. It would therefore be interesting to extend similar ideas to evaluate how state-of-the-art vision models grasp low-level physical properties from images and videos. By low-level physics, it means beyond object level of motion, for example, pixel or voxel level, such as lighting, object deformation after interaction, fluid flow, etc. [12] demonstrates a vision model's ability to understand lighting and shadow by using inpainting to reconstruct a masked area, accurately predicting the shadow of an object, and further evaluating the model's ability by probing with different 3D properties. This concept could be extended by testing additional physical properties in video models, for example, using future frame prediction models to anticipate the movement and deformation of objects, given initial frames of objects with different states and materials. Additionally, systematic evaluation methods are required to assess the capability of models in understanding low-level physics in generated visual content. While benchmarks like [3] offer metrics for video generation, they focus on broader objectives such as motion classification accuracy, optical flow, and semantic consistency. Therefore, it is necessary to develop evaluation benchmarks specifically targeting low-level physical properties in generated videos and dynamic 3D models, focusing on aspects like object persistence and deformation consistency, and their correlation with forces.

2.2 Data-driven Perspective

With proper evaluation methods in place, the next step is to test the hypothesis that data scale and quality are critical for a vision model's ability to understand low-level physics through extensive experimentation. Significant efforts have been made to incorporate physical priors or constraints into vision models, such as SMPL [4] and SMAL [18] which propose parameterized, predefined 3D models for humans and animals to model posture and movement with physical constraints. Other methods, like [17], introduce additional constraint terms to mitigate unnatural motions, such as foot skating or slippage artifacts in 3D models. [11] uses a projection module to transform unnatural generated motions into physically plausible space in diffusion steps. While these approaches focus on integrating physics constraints, exploring these challenges from a data-centric perspective could be compelling. One initial approach would be to use existing datasets and frameworks to design experiments that control both the size and quality of the data. By varying these factors, we can analyze how they affect the model's performance using the proposed benchmarks, thereby gaining insights into the role data plays in enabling models to grasp low-level physics. If it can be demonstrated that data scale and quality are indeed crucial, the next step would involve collecting and curating datasets specifically tailored to improve vision models' understanding of low-level physics and dynamics. One potential approach to gathering large-scale, dynamic data is to leverage the vast amount of publicly available, unlabeled video content from the web. The focus would then shift towards developing specialized data scraping and processing tools for specific tasks, such as assembling a dataset of videos in-the-wild for reconstructing and generating motion of humans and objects. Such kind of tools would need to incorporate, for example, off-the-shelf detection, segmentation, and tracking models, along with carefully designed filtering and cleaning mechanisms, to ensure that the data is object-centric with rich and clean information of motion and dynamics.

2.3 Data-efficient Learning of Low-level Physics

After collecting and curating datasets for the evaluation tasks, the next step may involve designing specific representation and machine learning algorithms that can more effectively capture the underlying data distribution, thereby enhancing the model's ability to recognize and understand low-level physics. A typical scheme for a vision model to learn physical information from video is through an inverse rendering process, where models are designed to extract physical information of an object, for example, implicit latent parameter [6], or explicit shape and motion representations [5], and reconstruct the video through differentiable physics engine or rasterization methods for direct supervision and optimization. An interesting direction that I would like to explore is to insert low-level physics information into the explicit representation of the learned shape and motion. Popular explicit 3D representations such as mesh and 3D Gaussian splatting [2] allow for efficient rendering of 3D scenes and objects that consist of small shape units onto a 2D image plane. Taking 3D Gaussian splatting as example, each Gaussian has parameters such as position, variance, color, and opacity, allowing explicit representation of small portions of the object in 3D space. By introducing additional physical parameters to 3D Gaussian splatting or similar representations, it may become possible to learn object dynamics through inverse rendering from video data. For instance, embedding each Gaussian with a deformation-related coefficient would enable easy inference of the stiffness of objects from their 3D representation, such as Young's modulus used in [9, 13] and spring-mass model used in [16]. Additionally, introducing coefficients related to shear and elasticity would possibly clarify which parts of the object can move and how they would respond to external interactions or forces. This type of representation can be further enhanced by integrating self-supervisory dense features, such as those from DINO [1], to establish low-level associations between videos of different instances, as demonstrated in [8] and [10]. This approach could enable learning the dynamics of deformable or movable objects, facilitating easier editing and generation with the learned representation, if sufficient high-quality video data is available. Ultimately, the goal of these steps would be to enable computer vision models to understand low-level dynamics and physics of how different objects deform, move, behave, in the real-world.

3 Past Experiences

My past research experience aligns closely with my proposal to tackle the problem of physics understanding of computer vision from a data-driven perspective. During my junior year, I was lucky to work under Professor Lin Zhang's guidance at Tongji University, where I started to learn basics of computer vision by joining some projects related to image feature matching and super-resolution. During my senior year and master's study at the University of Southern California, I had the opportunity of working with Dr. Yunhao Ge in Professor Laurent Itti's group. I initially joined a project focused on synthesizing detection and segmentation image data using image generation models—an efficient data augmentation technique that improved detection and segmentation model performance beyond the state of the art. In addition, I led the project DreamDistribution [14], a novel and efficient prompt-tuning method for visual generation models that better captures custom image data distributions for diverse personalized image and 3D generation. During my master's program, I also had the opportunity to intern at Microsoft Research Asia, where I led a project [15] to create a large-scale dataset for novel report generation tasks, reflecting real-world clinical scenarios. Both [14, 15] are have led to research paper submissions to ICLR 2025. Currently, I am a research assistant in Professor Jiajun Wu's group at Stanford University, working under the guidance of Dr. Shangzhe Wu. My current project involves collecting video from the web for animal motion reconstruction and generation, which closely aligns with my proposal of leveraging large-scale publicly available data for physics-aware visual understanding and generation. These experiences have provided me with a strong foundation in computer vision research, positioning me well to tackle the challenges of investigating and improving physics understanding in computer vision models from data-centric approaches.

4 Conclusion

This research line aims to investigate and improve how computer vision models can understand natural dynamics and physics, especially at a low and beyond object level. By addressing the current shortcomings in physical plausibility, the proposed work has the potential to significantly enhance the realism of dynamics of AI understanding and AI-generated visual content, making it more applicable and trustworthy in real-world scenarios.

References

[1] Mathilde Caron et al. “Emerging properties in self-supervised vision transformers”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 9650–9660.

[2] Bernhard Kerbl et al. “3D Gaussian Splatting for Real-Time Radiance Field Rendering”. In: *ACM Transactions on Graphics* 42.4 (July 2023).

[3] Yaofang Liu et al. “Evalcrafter: Benchmarking and evaluating large video generation models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 22139–22149.

[4] Matthew Loper et al. “SMPL: A Skinned Multi-Person Linear Model”. In: *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34.6 (Oct. 2015), 248:1–248:16.

[5] Keqiang Sun et al. “Ponymation: Learning 3D Animal Motions from Unlabeled Online Videos”. In: *arXiv preprint arXiv:2312.13604* (2023).

[6] Jiajun Wu et al. “Learning to See Physics via Visual De-animation”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017.

[7] Jiajun Wu et al. “Physics 101: Learning physical object properties from unlabeled videos”. In: *British Machine Vision Conference*. 2016.

[8] Shangzhe Wu et al. “MagicPony: Learning Articulated 3D Animals in the Wild”. In: 2023.

[9] Tianyi Xie et al. “PhysGaussian: Physics-Integrated 3D Gaussians for Generative Dynamics”. In: *arXiv preprint arXiv:2311.12198* (2023).

[10] Chun-Han Yao et al. “LASSIE: Learning Articulated Shape from Sparse Image Ensemble via 3D Part Discovery”. In: *NeurIPS*. 2022.

[11] Ye Yuan et al. “PhysDiff: Physics-Guided Human Motion Diffusion Model”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023.

[12] Guanqi Zhan et al. “What Does Stable Diffusion Know about the 3D Scene?” In: *arXiv preprint arXiv:2310.06836* (2023).

[13] Tianyuan Zhang et al. “PhysDreamer: Physics-Based Interaction with 3D Objects via Video Generation”. In: *European Conference on Computer Vision*. Springer. 2024.

[14] Brian Nlong Zhao et al. “Dreamdistribution: Prompt distribution learning for text-to-image diffusion models”. In: (2023).

[15] Brian Nlong Zhao et al. “Large Multimodal Model for Real-World Radiology Report Generation”. Dec. 2023. URL: https://www.microsoft.com/en-us/research/publication/large-multimodal-model-for-real-world-radiology-report-generation/.

[16] Licheng Zhong et al. “Reconstruction and Simulation of Elastic Objects with Spring-Mass 3D Gaussians”. In: *European Conference on Computer Vision (ECCV)* (2024).

[17] Yuliang Zou et al. "Reducing Footskate in Human Motion Reconstruction with Ground Contact Constraints". In: *Winter Conference on Applications of Computer Vision*. 2020.

[18] Silvia Zuffi et al. "3D Menagerie: Modeling the 3D Shape and Pose of Animals". In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. July 2017.