Foundation Models for Spatiotemporal Tasks in the Physical World

Zhe Jiang * Yu Wang † Zelin Xu *

Abstract

Foundation models such as ChatGPT are poised to transform society by providing general intelligence for problem-solving in healthcare, education, and law. They are also expected to make dramatic impacts in the way of AI solving spatiotemporal tasks in the physical world, such as smart manufacturing, intelligent transportation, and Earth system modeling. However, one major handicap is that existing foundation models do not understand the spatiotemporal knowledge of the physical world, leading to unexpected model behaviors and significant safety risks. This paper discusses emerging opportunities and unique challenges in integrating foundation models with physical components for solving spatiotemporal tasks. We also identify several new research directions to enhance the safety of such integrated models by spatiotemporal-knowledge-guided in-context-learning, verification, safety alignment, and the development of physics-informed geo-foundation models, as well as new benchmarking datasets and evaluation metrics.

1 Introduction

A foundation model is a large deep neural network model trained on a vast quantity of data at scale (often by self-supervised or semi-supervised learning) such that it can be easily adapted to a wide range of downstream tasks [5, 8, 24, 20]. Foundation models (e.g., ChatGPT) are poised to transform our society by providing general intelligence in healthcare, education, and law [5]. More importantly, they are also expected to make dramatic impacts in the way of AI solving spatiotemporal tasks in the physical world. For example, a foundation model, when integrated with a digital map engine as a plugin, can provide more intelligent location-based services, e.g., planning a week-long road trip from Orlando, Florida to Long Beach, California that passes through five national parks and a stop at a historical landmark. Similarly, a foundation model can translate a user's task in natural language into executable programs to control a robot [25] or call a physical simulator. Traditionally, such tasks are often solved by customized programs for a narrowly defined set of tasks. A program needs to be re-configured for a new task. Foundation models are transformative in that they provide a general and intelligent interface that can take complex tasks from multiple domains.

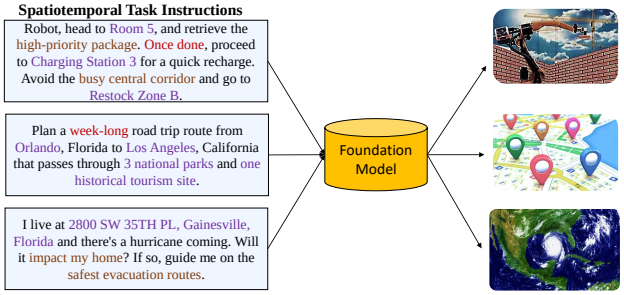

Figure 1: A vision of Emerging Opportunities.

However, one major handicap is that foundation models lack genuine understanding and reasoning of *spatiotemporal knowledge of the physical world*. Therefore, it can generate results that appear plausible but are spurious (hallucination [28]). For example, it has been found that a large language model can provide incorrect answers for physical reasoning questions such as if objects fall proportionately to their weight [16]. As another example in intelligent transportation, a foundation model may face difficulties in spatial reasoning between locations and recommend incorrect routes based on complex user requirements. The consequences are dire in high-stake decision-making applications such as evacuation planning.

A fundamental question is how to ground a foundation model with the physical knowledge of the world so that it can solve spatiotemporal tasks with safety assurance. Our vision is that addressing this question requires not only integrating a foundation model with physical engine plugins but also grounding physical knowledge into the foundation model pipeline (e.g., spatiotemporal-knowledge-guided prompt, alignment or fine-tuning, outcome verification).

2 Emerging Opportunities

Figure 1 illustrates emerging opportunities for foundation models in solving spatiotemporal tasks in the physical world. The key idea is to integrate a foundation model with a physical engine plugin, such as a robot,

*Department of Computer & Information Science & Engineering, University of Florida (zhe.jiang@ufl.edu, zelin.xu@ufl.edu).

†Department of Mechanical and Aerospace Engineering, University of Florida (yuwang1@ufl.edu).

a high-definition map engine, or physical simulators of the Earth system processes (or its AI surrogate) [18]. A spatiotemporal task description is first translated into physical commands, e.g., step-wise robotic commands, atomic map queries, or configuration files for physical simulators. The physical engine will execute the commands and provide solutions or take action.

2.1 Robotic Planning and Control Code Generation. Traditionally, robotic planning and control relied on experts to design specific rules and algorithms, such as instructing a robot to stack boxes. However, these algorithms often struggled with variations, such as different box sizes or stacking multiple boxes [13]. There are emerging methods that use foundation models to generate robotic planning and decision-making codes [1]. With users input details, the model will generate customized robotic programs. Recent research proposes foundation models with Planning Domain Definition Language (PDDL), a common abstract language for automated planning and scheduling, through chain-of-thoughts [26] and automated debugging [23].

2.2 Intelligent location-based services. In the current practice, location-based services are implemented by customized geospatial data management and mining algorithms for narrowly defined tasks. For example, one can find nearby restaurants or hotels with several predefined filters (e.g., price, cuisine, customer ratings) or find the fastest route from one location to another. In contrast, integrating foundation models, digital map engines, and GIS technologies provides new opportunities to provide broader and more complex location-based services. For example, one can ask a virtual assistant on an electrical vehicle to find a route from work to home with a stop at 6:30 p.m. for dinner at an Italian restaurant and a walk in a nearby mall.

2.3 AI-enabled Earth engine assistant. The current foundation models can only solve relatively straightforward problems about the Earth, such as the longest river or the highest mountain in the world. They are unable to answer more complex questions such as the flood inundation maps near my house in the next few hours, or the anticipated sea level rise in Florida coasts in the upcoming decades. Solving such problems requires numerical simulation models of the Earth system processes (or their AI surrogates) fused with observation data. However, existing models are often developed in silos for specific tasks. Foundation models can potentially provide a generic interface to interact with a suite of different Earth system models by translating a user's question into configuration files. For example, when a user asks for the forecasted flood inundation extent near his beach house in the next 24 hours, a foundation model can translate this question into configuration files to run a regional ocean circulation model (for storm surge) and an inundation model (for flood inundation extent).

3 Novel Technical Challenges

Several new technical challenges exist for foundation models to solve physical tasks.

First, existing foundation models are typically pretrained on large historical datasets comprising texts and images (videos). How to extend the models to geospatial and spatiotemporal data, which exhibit unique formats (e.g., geometric points, polylines, polygons, and Geo-Raster layers) and characteristics (e.g., spatiotemporal autocorrelation, heterogeneity) is non-trivial.

Second and more importantly, foundation models are trained without an understanding of the physical knowledge. Grounding foundation models with physical knowledge is critical to avoid fatal mistakes in high-stake spatiotemporal applications, such as disaster response. For instance, integrating physical knowledge (e.g., conservation laws) could improve a geo-foundation model's ability to simulate the dynamics of floods or storm surges. However, this is non-trivial due to the size, complexity, and black-box nature of foundation models, the difficulties in model retraining and fine-tuning, and the need for a good representation of physical knowledge.

Third, a foundation model may generate incorrect solutions in an over-confident manner (hallucination), which can be misleading. There is a need to develop approaches to quantify the uncertainty of the outputs from a foundation model as well as approaches for safety verification and alignment.

4 Future Research Directions

Figure 2 illustrates our vision of how to ground spatiotemporal knowledge into foundation models (e.g., Large Language Models and Geo-Foundation Models [17]). We identify several future research directions.

4.1 Spatiotemporal-Knowledge-Guided In-Context Learning, Model Verification, and Alignment. Research has found that proper prompting plays a crucial role in retrieving effective answers from a foundation model. For example, adding few-shot examples to break down a complex task into subtasks, also called chain-of-thoughts (COT), is beneficial [26]. For complex tasks in the physical world, task decomposition requires spatiotemporal logical reasoning beyond the simple COT. Traditionally, task decomposition in robotic planning relied heavily on human expertise.

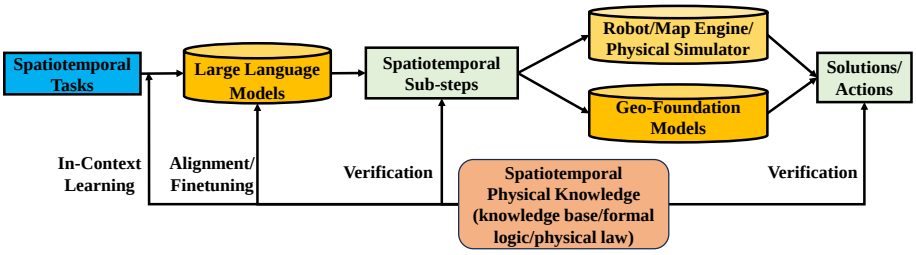

Figure 2: Our envisioned framework of physics-grounded foundation models for spatiotemporal tasks.

We envision a more rigorous prompting framework through formal logical reasoning. Linear temporal logic (LTL) [3] uses a set of rules to encode our understanding of how to fulfill the task and systematically use these rules to break down the tasks, enabling automated task reasoning and algorithm development. For applications where well-defined logic expression and decomposition are unavailable, we may resort to data mining (e.g., spatial decision trees [12]) to learn logic from human demonstrations.

Another direction for the safety assurance of foundation models in physical tasks is formal verification and safety alignment [21]. It is well-known that the behavior of foundation models may not always align with the intended user intent. The primary solution is fine-tuning the models through reinforcement learning with human feedback (RLHF) [2]. However, collecting human feedback is much more difficult for physical tasks compared with common question-answering. One potential direction is to replace human feedback with physical laws and constraints (e.g., conservation law, formal logic). Recent advancements in reinforcement learning methods allow for direct integration with logical expressions [7]. When combined with existing RLHF techniques, these methods can improve the alignment of pre-trained foundation models across a range of applications. Similarly, formal verification can inform the fine-tuning process if misalignments are detected [11].

4.2 Physics-Informed Geo-Foundation Model

There are several recent works on extending foundation models from computer vision to geo-domains with Earth imagery as well as climate and weather simulations, such as IBM-NASA Prithvi [10] and Microsoft ClimateX [19]. However, the model architecture itself is still based on the common vision transformer (ViT), which does not incorporate physical knowledge and constraints in model training. Although there exists extensive research in physics-informed machine learning in the literature [14, 27], direct application of these methods to foundation models is non-trivial. For instance, due to the size, complexity, and black-box nature of foundation models, once they are pre-trained, it is difficult to re-train and fine-tune. One potential strategy could be exploring parameter-efficient fine-tuning (PEFT) [9] with physical knowledge integrated into the loss function. Another strategy is to apply physics-guided post-processing of model outputs.

4.3 Benchmarking datasets and evaluation metrics. Another important direction is to develop new benchmark datasets for physics-grounded foundation models. For example, the Sen1Floods11 [6] dataset is used to evaluate the IBM-NASA geo-foundation model in flood inundation mapping, but the dataset only labels flood pixels with exposed water surface without considering the complete flood extent, especially those pixels obscured by obstacles like tree canopies. It is important to design a benchmarking dataset that labels the complete flood pixels based on the physics of water flows on terrain. For intelligent location-based services, existing benchmarking datasets such as the Geographical Question Answer set [22] only pose straightforward queries about spatial object relationships. It is necessary to develop a benchmark dataset with more comprehensive and complex queries. In robotics, there are existing open challenges in robotic planning and control in open-world scenarios, e.g., DARPA Robotics Challenge [15] and the Robo World Cup [4]. Evaluation metrics should reflect the consistency between a model's outputs with physical constraints.

Acknowledgement

This work is supported by the NSF under Grant No. IIS-2147908, IIS-2207072, CNS-1951974, OAC-2152085.

References

[1] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

[2] Y. Bai, A. Jones, and et al. Training a helpful and harmless assistant with reinforcement learning from human feedback, Apr. 2022.

[3] C. Baier and J.-P. Katoen. *Principles of Model Checking*. The MIT Press, 2008.

[4] T. Balch, P. Stone, and G. Kraetzschmar. *RoboCup 2000: Robot Soccer World Cup IV*. Springer, 2001.

[5] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[6] D. Bonafilia, B. Tellman, T. Anderson, and E. Issenberg. Sen1floods11: A georeferenced dataset to train and test deep learning flood algorithms for sentinel-1. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 210–211, 2020.

[7] A. K. Bozkurt, Y. Wang, M. Zavlanos, and M. Pajic. Control synthesis from linear temporal logic specifications using model-free reinforcement learning. In *IEEE International Conference on Robotics and Automation*, pages 10349–10355, Paris, France, 2020.

[8] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[9] N. Ding, Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, S. Hu, Y. Chen, C.-M. Chan, W. Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023.

[10] J. Jakubik, S. Roy, C. Phillips, P. Fraccaro, D. Godwin, B. Zadrozny, D. Szwarcman, C. Gomes, G. Nyirjesy, B. Edwards, et al. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*, 2023.

[11] S. Jha, S. K. Jha, P. Lincoln, N. D. Bastian, A. Velasquez, and S. Neema. Dehallucinating large language models using formal methods guided iterative prompting. In *2023 IEEE International Conference on Assured Autonomy (ICAA)*, pages 149–152. IEEE, 2023.

[12] Z. Jiang, S. Shekhar, P. Mohan, J. Knight, and J. Corcoran. Learning spatial decision tree for geographical classification: a summary of results. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, pages 390–393, 2012.

[13] S. Jiménez, J. Segovia-Aguas, and A. Jonsson. A review of generalized planning. *The Knowledge Engineering Review*, 34:e5, 2019.

[14] A. Karpatne, R. Kannan, and V. Kumar. *Knowledge Guided Machine Learning: Accelerating Discovery Using Scientific Knowledge and Data*. CRC Press, 2022.

[15] E. Krotkov, D. Hackett, L. Jackel, M. Perschbacher, J. Pippine, J. Strauss, G. Pratt, and C. Orlowski. The darpa robotics challenge finals: Results and perspectives. *The DARPA robotics challenge finals: Humanoid robots to the rescue*, pages 1–26, 2018.

[16] R. Liu, J. Wei, S. S. Gu, T.-Y. Wu, S. Vosoughi, C. Cui, D. Zhou, and A. M. Dai. Mind's eye: Grounded language model reasoning through simulation. In *The Eleventh International Conference on Learning Representations*, 2022.

[17] G. Mai, W. Huang, J. Sun, S. Song, D. Mishra, N. Liu, S. Gao, T. Liu, G. Cong, Y. Hu, C. Cundy, Z. Li, R. Zhu, and N. Lao. On the opportunities and challenges of foundation models for geospatial artificial intelligence, Apr. 2023.

[18] G. Mialon, R. Dessì, M. Lomeli, C. Nalmpantis, R. Pasunuru, R. Raileanu, B. Rozière, T. Schick, J. Dwivedi-Yu, A. Celikyilmaz, et al. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*, 2023.

[19] T. Nguyen, J. Brandstetter, A. Kapoor, J. K. Gupta, and A. Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

[20] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

[21] L. Pan, A. Albalak, X. Wang, and W. Y. Wang. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. *arXiv preprint arXiv:2305.12295*, 2023.

[22] D. Punjani, K. Singh, A. Both, M. Koubarakis, I. Angelidis, K. Bereta, T. Beris, D. Bilidas, T. Ioannidis, N. Karalis, et al. Template-based question answering over linked geospatial data. In *Proceedings of the 12th workshop on geographic information retrieval*, pages 1–10, 2018.

[23] T. Silver, S. Dan, K. Srinivas, J. B. Tenenbaum, L. P. Kaelbling, and M. Katz. Generalized planning in PDDL domains with pretrained large language models, May 2023.

[24] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[25] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J.-R. Wen. A survey on large language model based autonomous agents, Sept. 2023.

[26] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[27] Z. Xu, T. Xiao, W. He, Y. Wang, and Z. Jiang. Spatial knowledge-infused hierarchical learning: An application in flood mapping on earth imagery. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pages 1–10, 2023.

[28] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, et al. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.