Foundation Models for Exploration Geophysics

Qi Liu, Zenggui Chen, Jianwei Ma*

School of Earth and Space Sciences, Institute for Artificial Intelligence, Peking University, Beijing, China

*Corresponding email: jwm@pku.edu.cn

Abstract Recently, large models, or foundation models, have exhibited remarkable performance, profoundly impacting research paradigms in diverse domains. Foundation models, trained on extensive and diverse datasets, provide exceptional generalization abilities, allowing for their straightforward application across various use cases and domains. Exploration geophysics is the study of the Earth's subsurface to find natural resources and help with environmental and engineering projects. It uses methods like analyzing seismic, magnetic, and electromagnetic data, which presents unique challenges and opportunities for the development of geophysical foundation models (GeoFMs). This perspective explores the potential applications and future research directions of GeoFMs in exploration geophysics. We also review the development of foundation models, including large language models, large vision models, and large multimodal models, as well as their advancement in the field of geophysics. Furthermore, we discuss the hierarchy of GeoFMs for exploration geophysics and the critical techniques employed, providing a foundational research workflow for their development. Lastly, we summarize the challenges faced in developing GeoFMs, along with future trends and their potential impact on the field. In conclusion, this perspective provides a comprehensive overview of the development, hierarchy, applications, development workflow, and challenges of foundation models, highlighting their transformative potential in advancing exploration geophysics.

1 Introduction

As artificial intelligence evolves from rule-based systems to the era of machine learning, data-driven deep learning methods have emerged as a significant advancement. These methods utilize neural networks to uncover complex patterns within data across a wide range of applications. With the rapid growth of data volumes and continuous advancements in computational resources, deep learning-based foundation models have gained significant attention. Foundation models are large-scale, pre-trained models developed on vast and diverse datasets, often leveraging self-supervised or unsupervised learning techniques. The aim of foundation models is to create a flexible and adaptable base that can support a variety of tasks and applications. In recent years, the emergence of foundation models has achieved great success in various fields, including natural language processing models represented by ChatGPT[1], image segmentation models such as Segment Anything Model[2] (SAM), and video processing models represented by Sora[3], among others. Trained on massive datasets and with a vast number of model parameters, these models push the boundaries of deep learning, demonstrating its remarkable potential. The exceptional performance of foundation models has profoundly influenced the research paradigms in various fields, such as geoscience[62,63], chemistry[4,5], biology[6], finance[7,8], medicine[9], and remote sensing[10-12], among others. However, the application and development of foundation models in exploration geophysics are still in an initial stage.

Geophysics is a scientific discipline that investigates and analyzes the Earth's structures and states using physical principles and multimodal geophysical data. Exploration geophysics utilizes geophysical techniques and data to study the Earth's subsurface structure, aiding in the discovery of natural resources such as oil, gas, and minerals. Sub-disciplines like remote sensing and seismology also contribute valuable insights. Remote sensing uses satellite images to study surface conditions

and monitor environmental changes, while seismology focuses on earthquakes and provides insights into the Earth's internal composition and dynamics. These sub-disciplines collectively offer a comprehensive understanding of the Earth's subsurface and surface states. The procedure of geophysics mainly includes three stages: data acquisition (multimodal geophysical data, such as seismic data, remote sensing images, gravity data, atmospheric data, etc.), data processing (multiple tasks, such as first-arrival picking, interpolation, and denoising), and data interpretation (seismic imaging, weather forecasting, earthquake detection, and so on). The data processing stage has the following features: (1) large amounts of data; (2) multimodal data; (3) multiple tasks. The interpretation stage significantly relies on human analysis and experience. The extensive multimodal datasets provide a foundation for training and fine-tuning GeoFMs, while the complex, multi-task nature of the field highlights the potential for GeoFMs to significantly enhance exploration geophysics.

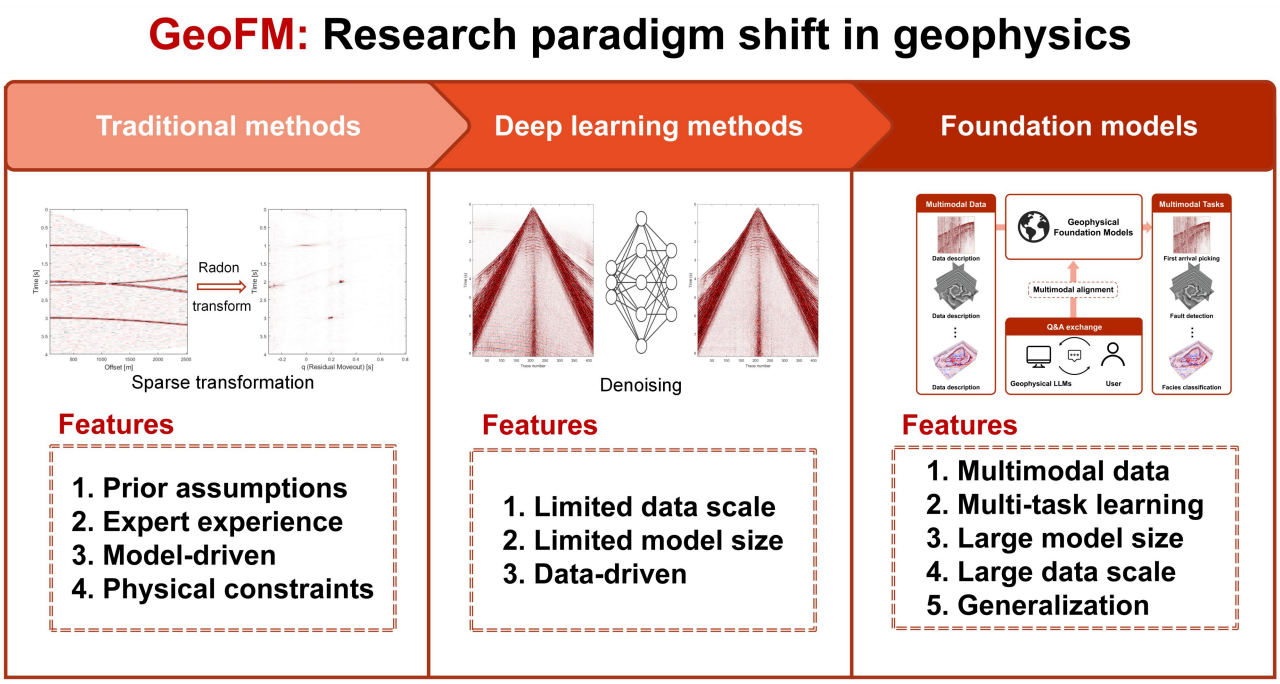


Figure 1 Research paradigm shift in exploration geophysics, the transition from traditional methods to deep learning methods and foundation models with corresponding features.

Over the past decade, the rapid development of deep learning methods has had a profound impact on the research paradigm in exploration geophysics, which has shifted from traditional methods to deep learning-based data-driven approaches[13], and further to foundation models, as shown in Figure 1. Traditional methods are usually model-driven and follow specific assumptions. Moreover, they are intricately dependent on the experience of experts. For example, the linearity assumption presumes that the seismic events exhibit linearity in small windows[14], while transform-domain-based methods assume that the data are sparse[15] or low-rank after transformations[16], like the Fourier[17], Curvelet[18], and Radon[19] transforms. These traditional methods are effective in the seismic data processing stage. However, when faced with complex field data, they may not be able to obtain reasonable denoising or interpolation results due to the invalid assumptions. With the exponential growth of data across various domains, data-driven deep learning methods have emerged as a focal point of research. Deep learning methods have been utilized in various fields of exploration geophysics, such as first-arrival picking[20], interpolation[21], inversion[22], etc. Review articles about deep learning in geophysics have recently been published[13,23], providing a detailed account of the application of deep learning methods in geophysics. Nonetheless, the current deep learning methods in geophysics are mostly based on the limited parameter volumes and training data, which limits the generalization of the model and thereby restricts the application of deep learning-based methods in the exploration industry. In light of the developments of foundation models, these limitations are gradually being alleviated. GeoFMs, trained on a large amount of multimodal geophysical data, demonstrate strong generalization capabilities across various geophysical downstream tasks, assisting researchers in conducting academic studies and exploration operations in the field of geophysics.

The success of foundation models across various fields has demonstrated the immense potential of deep learning methods, which are bound to bring profound changes to the paradigm of geophysical research. This paper aims to offer an overview of recent foundation models research in exploration geophysics, examine how the era of foundation models has impacted the paradigm of geophysical research, and explore potential research directions for GeoFMs. The main contributions of this paper are summarized as follows:

- **Introduction to Foundation Models and Their Applications in Geophysics:** This paper introduces the concept of foundation models, including large language models, large vision models, and large multimodal models. It reviews the development of these models and explores their applications within exploration geophysics, emphasizing their potential to transform the field. The paper also provides a brief overview of the use of foundation models in related geophysical sub-disciplines, illustrating their broader relevance to geophysics.
- **Hierarchy of GeoFMs:** The paper presents the hierarchy of GeoFMs into four distinct levels: task-specific model, modality-specific model, multimodal model, geophysical agent and copilot. We detail the specific characteristics and applications of each level, examine the interrelationships between these levels, and highlight their interaction and complementarity in the context of geophysical applications.
- **Proposed Generalized Workflow for GeoFM Development:** We propose a generalized workflow for the development of geophysical foundation models, which comprises four stages: data preparation, pretraining, multimodal alignment, and downstream task adaptation. We discuss the key techniques used at each stage and provide practical

examples to illustrate the application of this workflow in the development of GeoFMs.

- **Applications of GeoFMs in Exploration Geophysics:** The potential applications of GeoFMs in various areas of exploration geophysics, including data processing, geophysical imaging, and inversion, are discussed. The paper uses first-arrival picking as a case study to illustrate the specific application of GeoFMs in seismic data analysis. Additionally, it addresses the development strategies for different tasks and the challenges associated with applying GeoFMs to these tasks.
- **Challenges and Future Directions for GeoFM Development:** The paper outlines the key challenges that GeoFMs are likely to encounter during their development and deployment. It also offers an outlook on future trends and potential applications of GeoFMs, emphasizing the transformative potential of these models in advancing exploration geophysics.

The structure of the paper is organized as follows: Section 2 introduces the background of foundation models. Section 3 reviews the development and applications of foundation models within exploration geophysics and other related fields. In Section 4, GeoFMs are classified into four levels, with a discussion of the interrelationships between these levels. Section 4 also presents a generalized workflow for the development of GeoFMs, outlining the key stages and techniques involved. Section 5 examines the applications of GeoFMs across various areas of exploration geophysics, illustrated through a case study on first-arrival picking. Section 6 discusses the challenges faced by GeoFMs and provides an outlook on future trends. Finally, Section 7 concludes the paper by summarizing its key contributions.

2 The Background of Foundation Models

By offering highly generalizable capabilities that can adapt across a wide range of tasks, the development of foundation models has reshaped numerous fields. This section introduces the background of foundation models, focusing on three key categories: Large Language Models (LLMs), Large Vision Models (LVMs), and Large Multimodal Models (LMMs). Each of these categories has demonstrated exceptional potential in their respective domains and collectively contributes to advancements in artificial intelligence.

2.1 Large Language Models

LLMs are pre-trained on vast amounts of text data[24,25] and are capable of handling any text-formatted task without the necessity of task-specific fine-tuning. The remarkable zero-shot generalization capability of LLMs is attributed to the in-context learning paradigm[26], which allows LLMs to recognize various patterns in natural language and learn prompts through next-token prediction in autoregressive training.

The introduction of the Transformer framework[27] has revolutionized the development of LLMs. These models, built upon the self-attention mechanism, are capable of capturing long-distance dependence in text and enable parallel training with large parameters, thereby laying the methodological foundation for LLMs training. LLMs require extensive text data collected from various sources and are trained on thousands of GPUs or TPUs, with the scale of training reaching billions of parameters. Notable examples include the GPT[1] series developed by OpenAI, which utilizes a decoder-only Transformer architecture and incorporates reinforcement learning from human feedback[28] to fine-tune model responses. The GPT models have set new benchmarks for natural language understanding and generation tasks. Google's PaLM series[29] has achieved efficient

training of super-large parameters on TPUs with breakthrough performance in many downstream tasks. Meta AI released the LLaMA series[30] and open-sourced the model parameters, providing a foundation for the research of LLMs. Recently, the Claude 3 series[31] launched by Anthropic has made breakthrough progress in text tasks with ultra-long sequences. In addition, there are other high-performance LLMs such as LaMDA[32], GLM[33], OPT[34], and Chinchilla[35]. These LLMs have achieved inspiring results in natural language processing tasks, gradually revolutionizing the work patterns of professionals across various fields, including geoscience[36,121], remote sensing[36], chemistry[5,37], medicine[38-40], and others. In the field of geophysics, LLMs with geophysical knowledge can serve as the foundation for developing geophysical artificial intelligence, leveraging multimodal alignment to access various GeoFMs.

2.2 Large Vision Models

Most LVMs are primarily built upon Vision Transformer (ViT)[41] or convolutional neural network (CNN) architectures and are pretrained on large-scale image and video datasets. These models have achieved state-of-the-art results in various downstream tasks. In the segmentation task, SAM[2] has attracted widespread attention due to its outstanding segmentation performance and remarkable generalization ability. Furthermore, other fields can utilize it for data segmentation, such as identifying the initial arrivals in seismic datasets. Another notable development is the image restoration foundation model named SUPIR[42], which has achieved advanced results in various types of image restoration tasks. In addition, multimodal LVMs have made significant breakthroughs in text-to-image generation and text-controlled image editing, such as the contrastive language-image pre-training (CLIP) model[43], DALL-E[44] 2, Google Imagen[45], and Stable Diffusion[46].

Despite the impressive achievements, most current LVMs are either trained for single visual

tasks or are heavily dependent on LLMs for guidance. As a result, the development of pure LVMs has emerged as a recent area of interest[47]. The success of LLMs is largely due to the in-context learning[26] paradigm, which enables them to complete any text-formatted tasks through prompting. However, the absence of the in-context learning paradigm in LVMs hinders the flexible specification of language tasks in visual prompts as in LLMs[47]. Recently, researchers have started to explore new training paradigms for LVMs, with the objective of developing universal LVMs capable of tackling a variety of visual tasks. Bai et al. [48] (2023) defined a common format, "visual sentences", to enable the specification of visual tasks, and constructed a large-scale dataset to support their work. Specifically, "visual sentences" articulate visual tasks through sequences of images, with the model being capable of predicting the next image by learning patterns within the provided image sequence. Guo et al.[49] (2024) proposed a method of tokenizing the images first, followed by autoregressive training on the tokens. These efforts aim to advance the development of general LVMs, providing valuable insights for the evolution of foundation models in other disciplines[50,51]. Such advancements are crucial for the applicability of LVMs to more specialized domains, including geophysics, where complex seismic data interpretations are often required.

2.3 Large Multimodal Models

Large multimodal models (LMMs) represent a significant advancement in artificial intelligence by integrating information from multiple data modalities, such as text, images, audio, and others. This integration facilitates a more profound and nuanced comprehension of tasks across various areas. By comprehending the intricate interactions among many data modalities, LMMs provide enhanced adaptability relative to single-modality models.

The development of multimodal foundation models[51] has seen notable contributions that

significantly enhance the capabilities of artificial intelligence in understanding and generating across multiple data types. A prominent example is the CLIP[43] model developed by OpenAI, which learns visual representations from natural language descriptions, enabling zero-shot image classification and other downstream tasks with impressive accuracy. CLIP utilizes a contrastive learning framework to align image and text representations, effectively bridging the gap between visual and linguistic information. Flamingo[52] by DeepMind have further advanced the integration of vision and language, allowing AI to perform visual question answering, image captioning, and even interactive conversations involving visual content. The architecture of Flamingo effectively integrates both image and text streams, underscoring the growing trend of developing AI systems that are more dynamic and versatile. In addition, Meta's ImageBind[53] represents a groundbreaking step toward true multimodal integration by learning joint embeddings across six different modalities, including images, text, audio, depth, thermal, and inertial measurements. ImageBind's approach aims to create a shared semantic space that can simultaneously understand and generate from multiple data types. DALL-E[54], Google Imagen[55], GPT-4o[1], and Blip-2[57] are all notable multimodal models, demonstrating the power of integrating language and vision in creative and analytical applications. The ability of multimodal models to understand context and perform complex, cross-domain tasks makes them particularly useful in geophysics, where data often comes in multiple forms (e.g., seismic shot gathers, geological images, well logs, and textual reports).

3 Foundation Models in Geophysics

In recent years, geophysicists have increasingly focused on the development of foundation models in geophysics, leading to significant advancements across various subfields by leveraging multimodal geophysical data. These foundation models have achieved substantial progress,

transforming multiple areas of geophysical research and applications. Figure 2 provides an overview of GeoFMs in geophysics, illustrating how these models excel at various downstream tasks and are applicable across multiple subfields. Below, we primarily focus on the advancements of GeoFMs in exploration geophysics while also summarizing key developments in other geophysical domains.

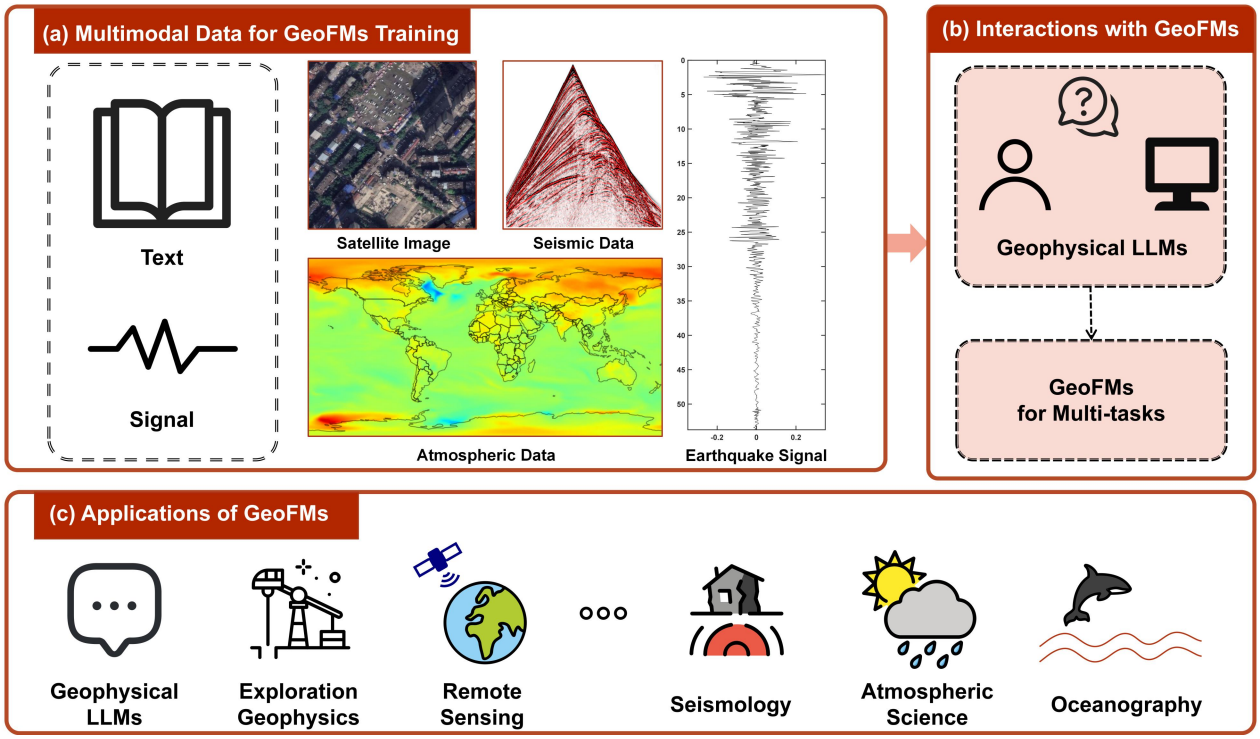


Figure 2 Overview of GeoFMs in geophysics. (a) GeoFMs are trained on multimodal geophysical data to acquire knowledge in the geophysical domain. (b) Various tasks in the field of geophysics can be effectively accomplished by interacting with geophysical LLMs and invoking different GeoFMs. (c) Applications of GeoFMs in the field of geophysics.

3.1 Exploration Geophysics

Exploration geophysics focuses on investigating subsurface structures to support natural resource exploration. Unlike areas such as medicine, remote sensing, and other disciplines, the application and development of GeoFMs in exploration geophysics are still in an initial stage. The relevant foundation model research in exploration geophysics is summarized in Table 1. Currently, interpretive tasks are the primary focus of foundation model development in exploration geophysics.

In contrast, the lack of large, high-quality datasets hinders tasks related to data processing, limiting progress in this area. The seismic imaging domain, on the other hand, remains heavily reliant on physical constraints, which complicates the integration of foundation models. As a result, related work in these two areas is still relatively sparse compared to interpretive tasks.

Sheng et al.[77] (2023) developed the first seismic foundation model (SFM) based on migrated seismic data using a Masked Autoencoder[83] (MAE) architecture for pre-training, achieving state-of-the-art results in various seismic downstream tasks, which marks a significant exploration in geophysical research in the era of foundation models. Han et al.[58] (2024) introduced a multi-attributes masking contrastive learning (MAMCL) approach using multimodal geophysical data, achieving excellent performance in explainable seismic facies analysis. In this paper, we present a novel approach by applying the existing foundation model SAM to the task of first-arrival picking in seismic data. Subsequently, Guo et al. [57] (2024) proposed fine-tuning large vision models using multimodal geophysical data to accomplish tasks such as geophysical data analysis. Kumar et al. (2024) developed a ViT-based[41] foundation model for numerical simulation studies in the oil and gas industry. Numerical simulations, which are computationally intensive with long turnaround times, can benefit significantly from this approach. Additionally, Liu et al.[67] (2024) provided a comprehensive review of the development and applications of large model technology in the oil and gas sector.

The rapid development of foundation models in exploration geophysics has opened up new avenues for advanced data analysis and subsurface exploration. From seismic analysis to multimodal geophysical data interpretation, GeoFMs are proving to be highly effective in addressing complex challenges. While much of the research is still in its early stages, the demonstrated success of these

models in a variety of tasks suggests a promising future.

Reference	Data Type	Methods	Key Features
SFM[86]	Migrated Seismic Data	MAE Pretraining	Foundation model, can be adapted on downstream tasks
GeoFMs(ours)	Shot Gather	SAM	First-arrival picking
CDFMA[57]	Multimodal	Fine-tuning	Geophysical data analysis
MAMCL[58]	Seismic Facies	Contrastive Learning	Seismic facies analysis
RockGPT[59]	Digital Rocks	Generative Pretraining	Random reconstruction on digital rocks data
Kumar et al. [60]	Reservoir Simulation Models	ViT Pretraining	Oil reservoir forecasting

3.2 Remote Sensing

Remote sensing, which employs sensors on satellites or aerial platforms, is a crucial tool for collecting geophysical data, facilitating the monitoring and analysis of the Earth's surface and environment. Remote sensing data primarily consists of various types of image data, including synthetic aperture radar (SAR) images, thermal infrared images, optical images, among others. Currently, remote sensing stands as the most developed domain for the application of foundation models within geophysics. Many tasks in remote sensing bear similarities to those in natural language processing (NLP) and computer vision (CV), contributing to this progress. Tasks such as image classification, object detection, and segmentation in remote sensing are conceptually analogous to those in CV, while spatiotemporal data analysis in remote sensing shares similarities with NLP tasks involving sequential data. Therefore, by fine-tuning, we can easily adapt existing large pre-trained models from NLP and CV to remote sensing applications.

In recent years, a number of comprehensive reviews have emerged, highlighting the growing interest in the development and application of foundation models in remote sensing. Data et al.[64] (2023) provided a detailed review of the evolution of Earth observation techniques, focusing on remote sensing as a key method. This work highlights the increasing challenges posed by the growing volume of multimodal remote sensing data and emphasizes the need for advanced AI techniques, such as foundation models, to effectively extract and utilize this wealth of information. Lu et al.[65] (2023) explored the role of self-supervised learning in enhancing model performance for tasks like scene classification and object detection in remote sensing. Their study underlines how self-supervised learning approaches have been instrumental in advancing foundational models by providing more effective pre-training strategies that leverage large amounts of unlabeled data. Jiao et al.[66] (2023) conducted a survey of foundation models in remote sensing, evaluating a range of models across multiple datasets. Their work confirmed the strong performance of foundation models in this field, particularly in tasks such as feature extraction, data classification, and highlighted important distinctions in model effectiveness for different types of remote sensing data.

3.3 Seismology

Seismology is the scientific study of earthquakes and the propagation of elastic waves through the Earth or other planet-like bodies. This field involves analyzing and interpreting seismic data collected from seismographs and other instruments, thereby enhancing the understanding of the Earth's internal structure. Recent advancements in deep learning have significantly enhanced the field of seismology, with these methods now being broadly applied across a diverse range of tasks within the discipline. These tasks include, but are not limited to phase picking[56-58], first-motion polarity classification[56,59], earthquake detection[58,60,61], event location[62,63], and earthquake prediction[64-

[66]. However, these methods are often trained on specific tasks using limited datasets, which restricts their wide application due to their limited generalization ability.

Inspired by the success of foundation models in various fields, researchers have begun to explore the use of vast amounts of seismic data to train GeoFMs in the field of seismology. SeisCLIP[67] is a foundation model in the field of seismology, utilizing contrast learning on multimodal data for pretraining. This foundation model can be applied to a variety of downstream tasks, such as event classification, localization, and focal mechanism analysis tasks through fine-tuning with small datasets. Li et al.[68] (2024) introduced Seismogram Transformer (SeisT), a foundation model designed for various earthquake monitoring tasks, including earthquake detection, seismic phase picking, first-motion polarity classification, and so on. This model was trained on the DiTing dataset[97] and evaluated for its generalization ability on the PNW dataset[98]. Both datasets encompass a significant number of seismic events and corresponding labels such as arrival times, magnitude, and first-motion polarity. Additionally, SeisLM[69] (2024) introduces a foundational model for analyzing seismic waveforms. SeisLM learns general waveform patterns, enabling strong performance in tasks such as event detection, phase-picking, onset time regression, and foreshock–aftershock classification. This model further demonstrates the potential of foundation models in seismology, showing promise in multiple fundamental tasks related to earthquake monitoring and analysis. These GeoFMs have demonstrated substantial potential to address different seismological tasks, significantly impacting the research paradigm within the field of seismology.

3.4 Atmospheric Science

Atmospheric science is a scientific discipline dedicated to the study and understanding of the Earth's atmosphere, including aspects such as weather and climate and their interactions with other

systems on Earth. This field is crucial for weather forecasting, climate prediction, and understanding changes in our environment due to natural and human-made factors. Over the past decade, weather forecasting has consistently relied on numerical weather prediction methods[69,70], which simulate transitions of atmospheric states based on partial differential equations.

Recently, deep learning-based weather prediction methods[71,72] have emerged as promising tools for accelerating weather forecasting. However, these methods, trained with limited data, do not achieve the accuracy of conventional numerical weather prediction techniques. To address these limitations, several GeoFMs have been developed in the field of atmospheric science. Trained with vast amounts of data, these GeoFMs have exhibited outstanding performance in both prediction speed and accuracy. Zhu et al.[70] (2023) explored the potential of foundation models in earth and climate sciences, identifying eleven key features needed for an optimal earth foundation model. Their work provides important guidelines for the future development of GeoFMs, highlighting aspects such as data diversity, model scalability, and physical interpretability. Similarly, Zhang et al.[71] (2023) evaluated the potential of foundation models in atmospheric science, examining their performance across tasks including climate data processing, physical diagnosis, forecasting, and adaptation. Their findings illustrate the diverse applications of foundation models in atmospheric research, emphasizing their ability to enhance accuracy and efficiency compared to traditional methods. Zhang et al.[73] (2023) developed NowcastNet using an end-to-end optimization architecture that integrates physical-evolution schemes to generate high-resolution, physically plausible nowcasts. Bi et al.[74] (2023) presented Pangu-Weather, a GeoFM designed for rapid and accurate weather prediction, which was trained on 39 years of global weather data. This model demonstrates superior performance compared to the leading numerical weather prediction model of that period, thereby

highlighting the transformative potential of GeoFMs in atmospheric science.

3.5 Oceanography

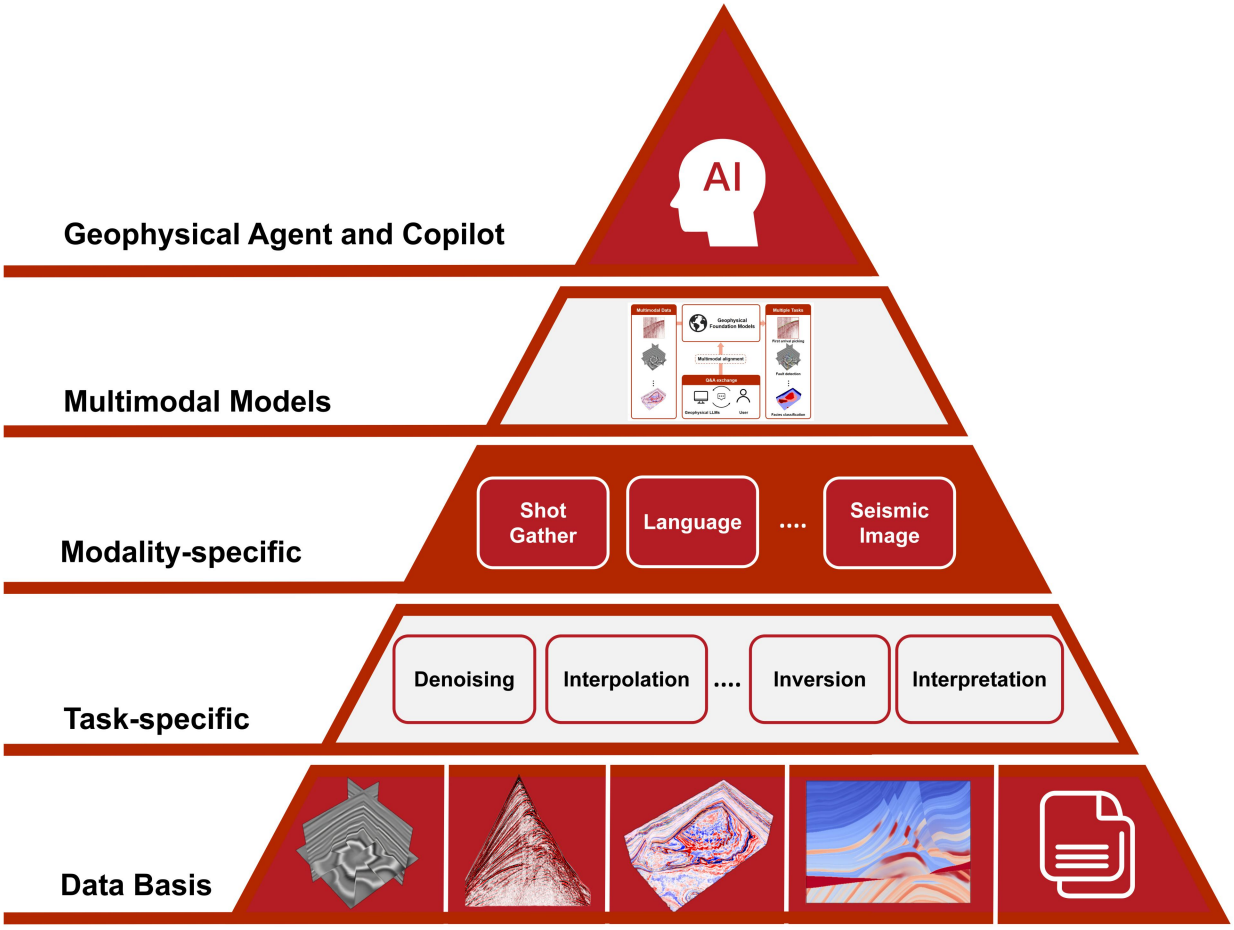
Oceanography, or ocean science, investigates the intricacies of the oceans that cover over 70% of the Earth's surface. This field is vital for understanding marine life and biodiversity, assessing the ocean's role in climate regulation, and examining their impact on global economies. Inspired by the remarkable success of foundation models in general domains, oceanographers have begun to explore the potential of GeoFMs in the field of oceanography.

Bi et al.[75] (2024) introduced OceanGPT, the first oceanographic LLM pre-trained for various ocean science tasks. To get around the problems of getting ocean data, the Doinstruct domain construction framework was suggested. This framework lets multiple agents work together to build an ocean instruction dataset. Xiong et al.[76] (2023) presented AI-GOMS, a GeoFM employing the Fourier-based masked autoencoder architecture, which was designed for predicting ocean variables over a 30-day period. In addition to these language-based models, MarineInst (2024) was introduced as a foundational model specifically designed for marine visual analysis. MarineInst provides instance masks and captions for marine objects, addressing the unique challenges inherent in marine image analysis. It demonstrates strong generalization across various downstream visual tasks, making it a valuable tool for tasks such as identifying marine species, mapping ocean habitats, and monitoring environmental changes.

4 Hierarchy and Development Workflow of GeoFMs

This section provides a comprehensive exploration of GeoFMs, focusing on both their hierarchy and a generalized workflow for their development. The hierarchy of GeoFMs is divided into four distinct levels: task-specific models, modality-specific models, multimodal models, geophysical

agent and copilot. This division provides insights into the unique characteristics and applications of each type, as well as how they interact to advance geophysical research. Furthermore, a generalized workflow for developing GeoFMs is introduced, detailing the four key stages: data preparation, pretraining, multimodal alignment, and task-specific adaptation. Each stage is accompanied by a discussion of the core techniques employed, along with a practical example to demonstrate their implementation. This structured approach provides a roadmap for advancing the development and application of GeoFMs in the field of exploration geophysics.



(a) Hierarchy of GeoFMs

Figure 3 The hierarchy of GeoFMs, from geophysical agent and copilot to data basis.

4.1 Hierarchy of GeoFMs

Figure 3 illustrates the hierarchy of GeoFMs, organized from the highest to lowest level of intelligence: Geophysical agent and copilot, multimodal model, modality-specific model, task-specific model, and data basis. This subsection provides a detailed overview of each hierarchy level, discussing their unique characteristics and the ways in which they interact and complement each other in geophysical research.

4.1.1 Geophysical Agent and Copilot

Geophysical agent and copilot represent the highest level of intelligence in the hierarchy of GeoFMs. They are an integration of different levels of GeoFMs, combining various capabilities to achieve advanced intelligence and adaptability. Geophysicists can use these models as intelligent assistants or "co-pilots" who can not only understand and analyze data, but also provide support for a wide range of geophysical tasks.

The geophysical agent is an autonomous system[77] designed to understand user instructions and analyze diverse geophysical data, enabling automated workflow design and decision-making for exploration tasks. For instance, the geophysical agent can understand different types of seismic data through integrated multimodal GeoFMs, create automated workflows, and use task-specific GeoFMs or relevant software APIs to process and interpret seismic data, which helps guide exploration operations. This type of model aims to mimic expert-level understanding, including aspects such as pattern recognition, decision-making heuristics, and contextual analysis, thereby reducing the reliance on domain-specific knowledge from experts for routine seismic tasks, and enabling new insights into complex geophysical challenges.

In the field of exploration geophysics, traditional standardized workflows and processing

interpretation software are already quite mature, but their usage and operation often require specialized knowledge and extensive experience, which creates a need for an assisting tool. The geophysical copilot addresses this need by serving as an essential assistant[78] that integrates user input with advanced data analysis capabilities. The design of the geophysical copilot allows it to comprehend various exploration data modalities, including well logs, shot gathers, and seismic images, and aids geophysicists in creating workflows specific to their tasks. By incorporating task-specific GeoFMs or using relevant software APIs, the geophysical copilot helps validate hypotheses, generate insights, and optimize data processing and interpretation tasks. This interactive approach ensures that geophysicists retain control over decision-making while gaining the benefit of enhanced support and operational efficiency.

4.1.2 Geophysical Multimodal Models

Geophysical multimodal models represent the next level in the hierarchy of GeoFMs, providing an integrated approach to understanding geophysical data by combining different data modalities. These models enhance the ability to identify patterns and relationships that may not be evident when each type of data is analyzed independently, offering a cohesive understanding of complex geophysical tasks. Geophysical multimodal models are built upon modality-specific models, leveraging their outputs to align and integrate information from different data types.

The primary advantage of multimodal models lies in their ability to extract complementary information from diverse modalities, which is crucial for complex geophysical processing and interpretations. For example, in exploration geophysics, combining seismic data with electromagnetic measurements enhances the accuracy of subsurface imaging and resource estimation. Furthermore, these models serve as a bridge between modality-specific models and higher-level

GeoFMs by providing enriched, unified latent data representations that can be used for more comprehensive analysis. Multimodal model training always starts with unimodal pretraining, where each modality-specific model learns independently, followed by a multimodal alignment[79] phase that integrates these individual representations into a cohesive whole. By utilizing shared embedding spaces and contrastive learning, these models ensure that information from different sources is effectively harmonized, allowing for comprehensive analysis across data types. By doing so, they lay the groundwork for developing geophysical agent and copilot that can make more informed decisions based on a holistic understanding of different geophysical data.

4.1.3 Modality-specific Models

Modality-specific models represent a fundamental level in the hierarchy of GeoFMs, focusing on the representation learning of geophysical data from a single modality. These models are specifically designed to handle one type of geophysical data, such as seismic, electromagnetic, gravity, well logs, or text, and to optimize their performance for tasks that are unique to that particular data type. The primary role of modality-specific models is to learn rich representations that can serve as strong embeddings for subsequent multimodal model training, and to provide a solid foundation for fine-tuning on downstream geophysical tasks. Modality-specific models often employ advanced self-supervised learning techniques such as generative pretrained transformers[1] and MAE[83] to learn these powerful representations, enabling them to capture complex relationships within the data.

In the field of exploration geophysics, modality-specific models are increasingly being explored for various data processing and interpretation tasks that require a high level of expertise focused on a single modality. For instance, Sheng et al.[101] (2023) developed SFM based on migrated seismic data,

which serves as an effective pretrained model for a range of downstream tasks. By using task-specific adaptation, this seismic modality-specific model has shown great performance in tasks like interpolation, fault identification, and inversion. This shows how useful and flexible modality-specific models are in geophysical research.

Modality-specific models play a crucial role in connecting different levels of GeoFMs. They act as foundational building blocks that provide specialized embeddings for more complex GeoFMs. Multimodal models often use the embeddings from these models as their basis, enhancing the representation power by providing well-optimized, modality-specific features that can effectively integrate with other modalities. Through techniques such as fine-tuning, modality-specific models can be adapted to various downstream tasks, enhancing their performance in task-specific applications. The fact that modality-specific models act as a link between task-specific models and higher-level integrated systems shows how important they are for building a cohesive and multilayered geophysical modeling framework.

4.1.4 Task-specific Models

Task-Specific Models are at the foundational level of GeoFMs, specifically designed to address well-defined seismic tasks using targeted learning approaches. Unlike modality-specific models that focus on mastering a particular data type, task-specific models aim to solve particular problems such as denoising, interpolation, first-arrival picking, fault detection, or seismic imaging, utilizing the embeddings learned from modality-specific models to enhance their performance. These models are tailored to handle specific challenges in exploration geophysics, directly contributing to the efficiency and accuracy of operational workflows. In exploration geophysics, task-specific models have been utilized to handle highly specialized tasks that require precise problem-solving capabilities.

For example, a model designed for fault detection may be used to pinpoint fault locations in seismic datasets, while another model for seismic inversion may focus on estimating subsurface properties based on seismic reflection data.

Task-specific models are essential in the broader context of GeoFMs, serving as the endpoints for applying foundational knowledge to practical applications. By using the rich representations learned from modality-specific models, task-specific models can achieve a higher level of accuracy and generalization capability in their respective tasks. This approach ensures that each specific problem is addressed with optimal precision, while also allowing task-specific models to be seamlessly integrated into multimodal models or geophysical agent and copilot, which require specialized task solutions as part of their overall functionality.

4.1.5 Data Basis

The data basis serves as the fundamental building block for all GeoFMs. It involves the collection, preparation, and curation of diverse geophysical datasets used as input for training models within the GeoFMs hierarchy. This includes various geophysical data types, such as seismic records, electromagnetic measurements, gravity data, well logs, geophysical text data, and task-specific data pairs for different geophysical challenges. For example, if we want to train a foundation model for prestack seismic data denoising, we first need a large amount of shot gathers to train a foundation model for the prestack data modality. Next, we train a model specifically for seismic denoising by using numerous denoising sample pairs for different types of noise.

Ensuring quality, diversity, and comprehensiveness of these datasets is crucial for developing reliable models across all GeoFMs levels. The data basis forms the groundwork for all modeling efforts, supporting the training of modality-specific models as well as task-specific models, enabling

end-to-end training to solve specific geophysical challenges. For modality-specific models, the data basis provides the necessary diversity and richness to learn strong, generalizable representations for each modality. For task-specific models, the data basis supplies carefully curated, task-oriented data pairs that allow models to be trained effectively for precise applications. This interconnected structure ensures that each level of the GeoFM hierarchy benefits from a robust data foundation, enabling seamless progression from foundational data analysis to complex geophysical interpretations.

4.2 Development Workflow for GeoFMs

To effectively develop GeoFMs, a structured and systematic approach is essential. This section introduces a generalized workflow designed to guide the development of GeoFMs, which covers key stages including data preparation, pretraining, multimodal alignment, and adaptation for task-specific applications, as shown in Figure 4. We outline these stages to provide a comprehensive workflow that can adapt to various geophysical tasks, ensuring a consistent methodology for building powerful and versatile GeoFMs capable of addressing the complex challenges inherent in exploration geophysics.

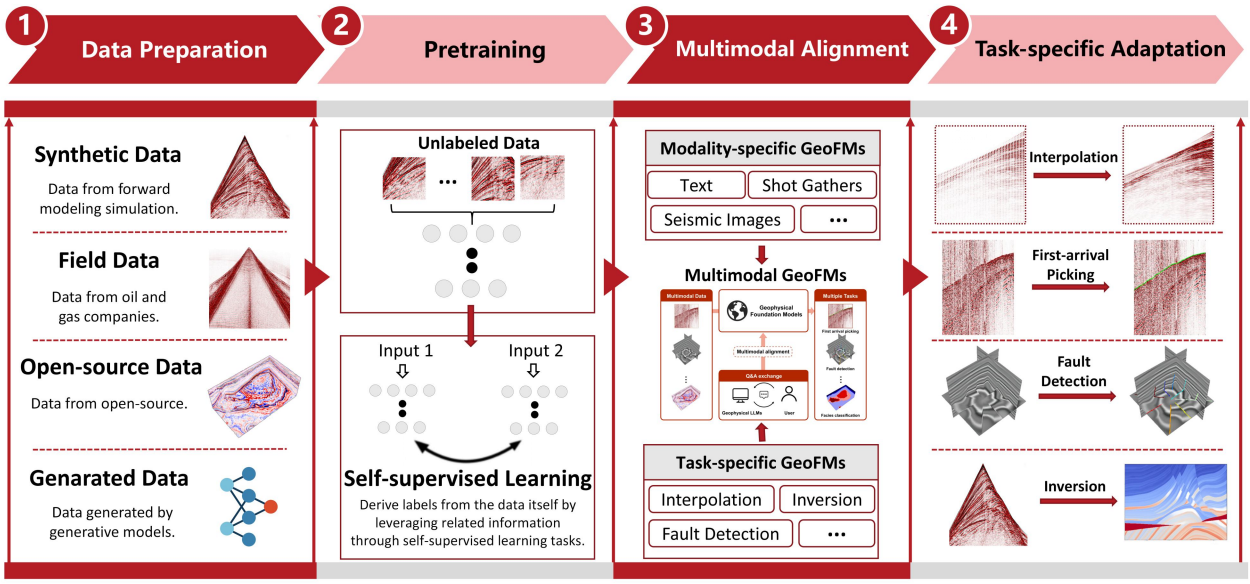


Figure 4 The development workflow of GeoFMs, including data preparation, pretraining, multimodal alignment, and task-specific adaptation. The illustration of self-supervised learning is repainted from the survey proposed by Liu et al.[169]

4.2.1 Data Preparation

Data serves as the cornerstone for developing robust GeoFMs, where the hierarchy of GeoFMs into different levels emphasizes the importance of comprehensive datasets. Data preparation is crucial not only for training task-specific models but also for enabling the development of higher-level models, such as modality-specific and multimodal GeoFMs. The quality and diversity of data directly impact the model's ability to generalize across different geophysical applications, making data preparation a foundational aspect of the entire workflow. In the context of GeoFM development, data preparation involves several key steps, including data collection, preprocessing, and labeling, as shown in Figure 5.

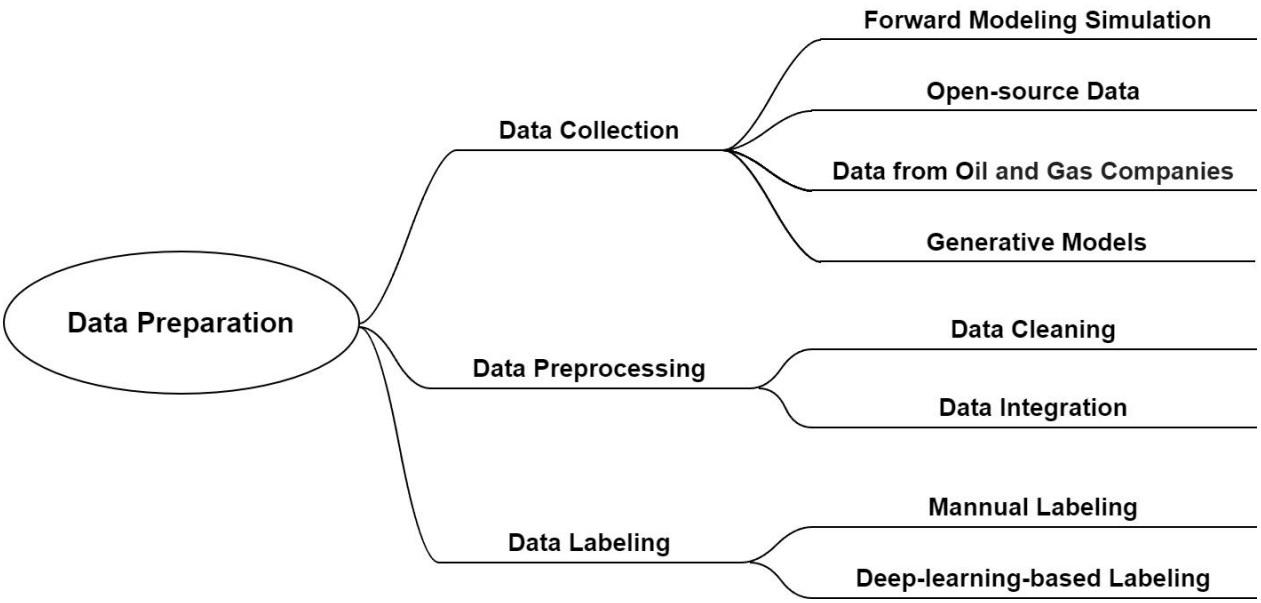


Figure 5 Key steps for data preparation in the development workflow of GeoFMs.

The primary sources of data for GeoFMs include forward modeling simulation, open-source data, proprietary data from oil and gas exploration companies, and data generated through generative models. Forward modeling simulations, utilizing techniques like finite-difference modeling and finite-element algorithms[138] based on the wave equation, provide synthetic datasets where subsurface properties and geological scenarios are precisely defined. While these synthetic datasets are invaluable for understanding theoretical responses and testing algorithms, they often lack the complexity, noise characteristics, and variability of real-world data, limiting their utility for accurate model performance in practical scenarios. Open-source data offers access to real geophysical data, enabling initial model training and validation. However, these datasets are typically limited in volume and may vary in quality, making them insufficient for training high-performance GeoFMs. The most critical data comes from proprietary sources held by oil and gas exploration companies. These large-scale, high-quality datasets, which encompass a wide range of geological settings and subsurface complexities, are crucial for developing GeoFMs capable of generalizing effectively to real-world exploration tasks. Accessing this data often requires collaboration between industry and

research institutions, highlighting the importance of multi-institutional partnerships. To address data scarcity, generative models such as GAN, VAE, and diffusion models[135] offer promising solutions by learning the statistical distributions of real data and generating new, realistic samples. These models can augment existing datasets, simulate rare events, and address underrepresented scenarios, improving the robustness and generalization of GeoFMs. Combining these diverse sources of data is key to building powerful models that can tackle the complexities of geophysical exploration.

Data preprocessing is a critical step in preparing geophysical datasets for training GeoFMs. It consists of two main stages: data cleaning and data integration. Data cleaning involves removing errors and inconsistencies that can adversely affect model performance. This includes eliminating "bad traces" in seismic records, which may arise from faulty sensors or poor signal quality, as well as filtering out strong noise interference from external sources. Additionally, redundant or duplicate data points are removed to ensure that the model trains on unique and relevant samples, further enhancing the quality of the data. Data integration focuses on transforming and structuring the data to fit the input-output format required by the model. This includes formatting the data to match the network's architecture, such as adjusting seismic data to fixed-length time windows or integrating geophysical measurements into compatible formats. For different tasks, end-to-end datasets are curated by combining various data sources, such as seismic, well logs, and gravity data, to create comprehensive training sets. Common operations like normalization and standardization are also applied to ensure consistent data scaling and prevent features with large magnitudes from skewing the model's learning process.

Data labeling is essential for preparing high-quality labeled datasets that support the training of GeoFMs. It involves two main strategies: manual labeling and deep learning-assisted labeling.

Manual labeling relies on specialists or professional software to annotate geophysical data, including tasks like data processing, seismic imaging, and generating detailed analysis reports. These datasets are crucial for ensuring accurate annotations that reflect complex geological conditions. On the other hand, deep learning-assisted labeling leverages advanced deep learning models to facilitate and enhance the labeling process. For instance, SAM models can assist in automatically labeling first-arrival picking, a task that will be further discussed in subsequent sections. The integration of deep learning techniques into the labeling process offers distinct advantages, particularly in terms of scalability and consistency. These methods not only expedite the annotation of large datasets but also ensure a level of uniformity in labeling that is often challenging to achieve through manual annotation alone. Moreover, deep learning-based approaches have the capacity to tackle complex, high-dimensional tasks that would be otherwise cumbersome and prone to human error. As such, the role of deep learning in data labeling is becoming increasingly pivotal in the context of GeoFM development, enabling the creation of comprehensive, high-quality labeled datasets that are essential for the effective training of robust models.

4.2.2 Pretraining

Pretraining is a crucial phase in the development of GeoFMs, as it allows models to learn generalized representations from large-scale geophysical datasets before being fine-tuned for specific tasks. During this stage, GeoFMs are exposed to vast amounts of data to capture underlying structures and features, which can then be transferred to downstream applications. Pretraining provides the foundational knowledge necessary for GeoFMs to excel across various geophysical tasks, significantly reducing the data and computational requirements for subsequent task-specific adaptations. We mainly focus on self-supervised pretraining, which has become a key paradigm in

foundation model development. In this approach, the model can leverage unlabeled data by creating auxiliary tasks to learn useful representations and is allowed to extract features without requiring explicit human-provided labels. Here, we explore two primary self-supervised pretraining paradigms: generative pretraining and contrastive pretraining, as shown in Figure 6.

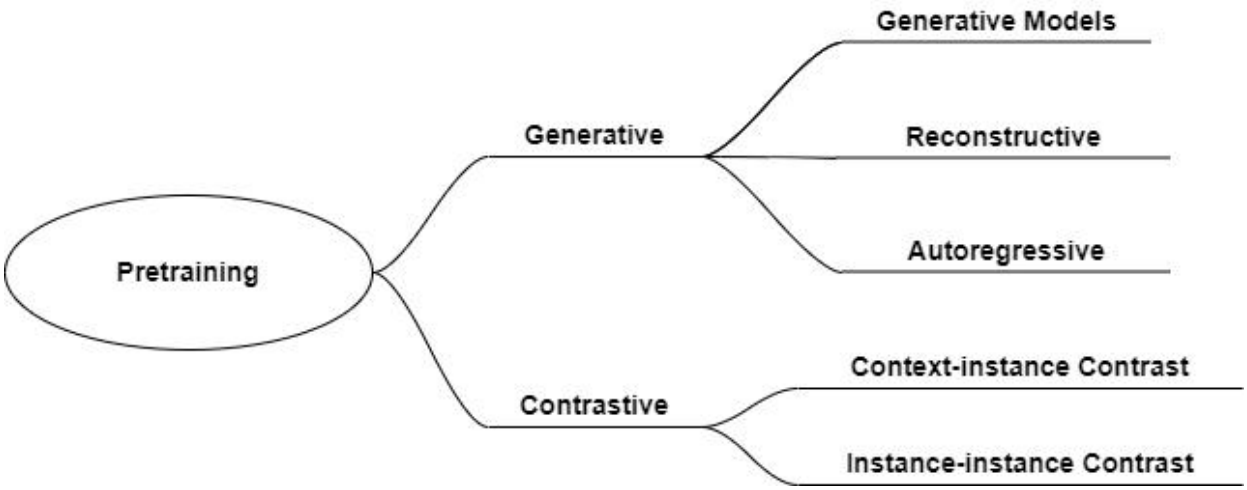


Figure 6 An overview of self-supervised pretraining methods.

Generative pre-training is a general method used to train models on large amounts of unlabeled data before fine-tuning them for specific downstream tasks. The goal is for the model to learn general patterns, distributions, and representations in data that can be applied across various applications. Generative pretraining typically involves generative models, reconstructive tasks, and autoregressive tasks.

One of the key strategies in generative pretraining is using generative models, such as generative adversarial networks[133] (GANs) and diffusion models[134,135], to learn representations from large, unlabeled geophysical datasets. These models are trained to capture data distribution, which is crucial for pretraining GeoFMs. GANs, composed of a generator and a discriminator, are commonly used in pretraining tasks where the goal is for the generator to learn to produce synthetic data that closely mirrors real data distributions. In the context of geophysical data, GANs can be employed to

learn representations of seismic waveforms or geological features by training the generator to produce data that is indistinguishable from actual seismic data. Although GANs are widely used for data generation, their primary role in pretraining GeoFMs is to help the model learn broad, generalized patterns and representations of geophysical processes that can be transferred to specific downstream tasks. Diffusion models, such as Stable Diffusion[136], are another powerful tool for pretraining, offering an alternative to GANs with more stable learning dynamics. Diffusion models learn to reverse the process of noise addition, gradually denoising data to reconstruct it. This denoising process enables the model to learn complex data distributions without requiring explicit labels, making it a strong candidate for pretraining on unlabeled geophysical datasets. Stable Diffusion, in particular, has demonstrated the ability to generate high-quality, detailed samples from noisy data, which can be invaluable in learning representations of seismic data, geological structures, and other geophysical phenomena. The model's ability to reconstruct data by progressively refining it aligns well with the process of extracting useful features and patterns from complex geophysical datasets.

Another good way to use self-supervised pretraining is with reconstructive tasks. In these tasks, the model learns to guess what parts of the input data are missing or broken, which helps it build a compact and useful representation of the data. This approach is particularly useful for learning generalizable features from unlabeled geophysical datasets. The two main strategies in this paradigm are mask prediction and autoencoder. Mask prediction, as exemplified by MAE[116], is inspired by BERT[137] which converts images into sequences similar to text and predicts the masked portions. In the context of seismic data, MAE can be used to mask portions of seismic input and train the model to reconstruct the missing sections, encouraging the model to capture essential geological features. In

the geophysical field, the seismic foundation model (SFM) proposed by Sheng et al.[110] employs this approach, using masked migrated seismic data to learn valuable representations. Additionally, autoencoder, which learn to encode data into a lower-dimensional latent space and then decode it back to the original input, are also effective for learning data representations without labeled samples. These approaches align well with data processing tasks in seismic analysis, such as noise attenuation and data enhancement. Besides, this approach is particularly beneficial in seismic interpretation tasks, where high-dimensional data is often incomplete or corrupted.

Autoregressive tasks represent a powerful pretraining strategy where the model generates data sequentially, predicting each token based on previously generated tokens. The GPT[1] models are based on autoregressive generation and have achieved remarkable success, especially for NLP tasks. During autoregressive pretraining, the model learns to capture the temporal and contextual dependencies within data sequences, enabling it to generate coherent and contextually relevant outputs. Beyond NLP, the autoregressive paradigm has been extended to other domains, including computer vision and multimodal tasks, with notable implementations such as AnyGPT[130]. In the context of geophysical applications, autoregressive models offer significant promise for modeling sequential or time-series data, such as seismic data, well logs, or multimodal geophysical measurements. These models can learn the dependencies between past and future data points, facilitating the generation of contextually relevant geophysical interpretations. While seismic data necessitates specialized pretraining to account for its unique characteristics, textual data—such as exploration reports, drilling logs, or geophysical analysis reports—can benefit from the rich pretrained knowledge embedded in LLMs. LLMs, which are pretrained on vast corpora of textual data, excel in understanding both structured and unstructured text. By fine-tuning these models on

domain-specific geophysical datasets, such as exploration reports or drilling logs, they can be adapted to generate highly relevant and contextually accurate outputs in the domain of exploration geophysics, thereby supporting tasks such as report generation, interpretation of geophysical findings, and decision-making.

Contrastive pretraining focuses on learning representations by distinguishing between positive and negative data samples. The model is trained to maximize the distance between dissimilar samples and minimize the distance between similar samples within a shared representation space. Two prominent paradigms of contrastive learning are context-instance contrast and instance-instance contrast, each focusing on different levels of representation and learning objectives. As contrastive learning technique is mainly used in multimodal alignment tasks, this subsection provides a brief overview, with further discussions on its application in multimodal geophysical data integration in the following subsection.

Context-instance contrast focuses on learning the relationship between a local feature and its broader context. By connecting a specific feature to its global context, the model better understands the dependencies between local and global components. In GeoFMs applications, this method is useful when studying seismic data or geological features in relation to larger regional structures. For instance, when interpreting a local seismic trace, the model learns to relate it to broader geological formations or regional trends. Instance-instance contrast emphasizes local feature-to-feature relationships, where the model focuses on distinguishing individual instances based on their inherent characteristics. In GeoFMs, this method helps the model to recognize and classify local features, which are crucial for tasks like fault detection, horizon identification, or reservoir characterization. Contrastive pretraining helps the model distinguish between similar and dissimilar seismic traces,

improving its understanding of geological patterns. Specifically, contrastive pretraining involves constructing positive and negative pairs of seismic traces, where positive pairs are similar (e.g., traces from the same geological layer), and negative pairs are dissimilar (e.g., traces from different regions or with distinct features). Techniques like SimCLR[131] or MoCo[132] can be adapted, where augmentations are applied to create different views of the same seismic trace. These views are then contrasted with traces from different areas to help the model learn distinguishing features.

4.2.3 Multimodal Alignment

Multimodal alignment is a critical technique that enables the integration of different types of data (or modalities) into a unified representation space. Transformer-based models, which have gained widespread attention in various multimodal tasks such as image-text pairing[44], text-to-speech alignment[146], and video analysis[147], offer a robust framework for achieving cross-modal alignment. Geophysical models, which require simultaneous processing of various data types for more accurate analysis and interpretation, greatly benefit from this alignment. In the context of GeoFMs, this can involve aligning geophysical data, like seismic traces, well logs, and geological images, with other modalities, such as natural language descriptions, technical reports, or other data types related to geological analysis. To achieve this, two main components are involved: the modality encoder and the modality interaction, as shown in Figure 7.

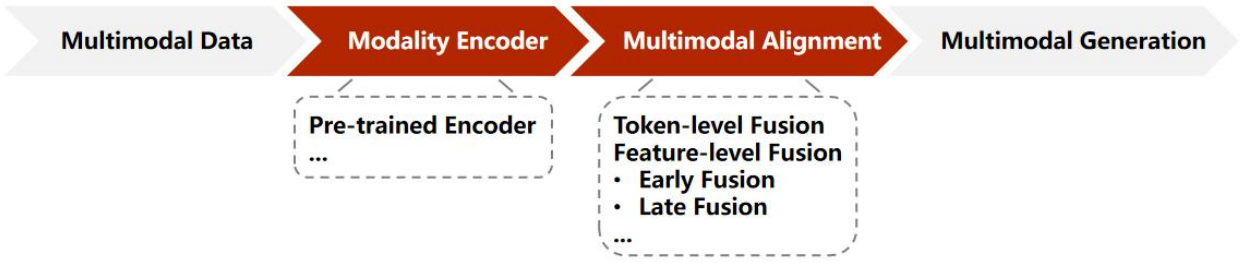


Figure 7 An overview of the multimodal generation process.

In multimodal alignment, the modality encoder plays a crucial role in transforming raw input data into compact, semantically dense feature representations. These representations capture the essential characteristics of the data, enabling the model to process and analyze complex, heterogeneous information more effectively. A common strategy is to leverage pre-trained encoders that have been trained on large-scale datasets to align features from different modalities, such as images and text. For instance, in models like CLIP[44], a visual encoder is trained to associate image features with textual descriptions. This allows the encoder to be fine-tuned or adapted to new tasks, making it easier to integrate and process multimodal data, such as combining seismic data with geological reports or images in the case of GeoFMs.

Once the data from different modalities has been encoded into feature spaces, the modality interaction plays a crucial role in aligning and fusing these representations so that a unified model can understand and reason over them effectively. There are two primary ways of integrating multimodal information into a shared framework: token-level fusion[148] and feature-level fusion[149,150].

Token-level fusion transforms features from different modalities into token representations, which can then be combined and processed by the model. This method is particularly effective for multimodal tasks where each modality can be represented as a sequence of tokens. The transformed tokens from each modality are then concatenated and sent into the model. For instance, BLIP-2[57] proposed a Querying Transformer (Q-Former) that uses learnable queries to extract relevant features from the characteristics of different modalities, and these selected features are then fed as prompts into the LLM. Advanced methods, X-Former[151], enhance visual representations through an innovative interaction mechanism, which combines high-frequency, detailed features from masked image modeling with low-frequency, semantically rich features from contrastive learning. In contrast,

some methods[152,153] use the MLP as a learnable connector to link features from different modalities.

Feature-level fusion, on the other hand, involves deep integration between the features of different modalities, where each modality's features interact at a deeper level. This approach allows for more complex interactions between modalities, enabling the model to learn richer representations from cross-modal relationships. The survey proposed by Xu et al.[168] summarizes six types of transformer-based cross-modal interactions. The common methods include: (1) Early Summation[139], where token embeddings from different modalities are weighted and summed before being processed; (2) Early Concatenation[140], which merges modality embeddings into a single sequence for unified processing; (3) Hierarchical Attention (Multi-stream to One-stream)[141], where separate streams for each modality are fused at a later stage; (4) Hierarchical Attention (One-stream to Multi-stream)[142], where a shared Transformer initially processes the data, followed by independent streams; (5) Cross-Attention[143], where each modality attends to the other by using query embeddings from one modality to attend to key and value embeddings from the other; and (6) Cross-Attention to Concatenation[144], where the cross-attended features from each modality are concatenated and processed together before passing through another layer. For GeoFMs, feature-level fusion could be applied to combine features from seismic data, geological images, and well logs at deeper levels. For instance, a deep cross-attention mechanism might be used to fuse seismic features with geological image features or textual descriptions of geological strata. Dual interaction layers can be introduced between seismic data and geological features, allowing the model to refine its understanding of how seismic signals correspond to specific geological structures.

The LanguageBind[145] approach proposes using language as the central modality to align other data types effectively. This strategy, which has been applied successfully to vision, infrared, depth,

and audio data, can be adapted to exploration geophysics to bring together seismic, well logs, electromagnetic, and other data. By freezing the pretrained encoder and using contrastive learning, models for each geophysical modality can be trained to align with the central modality, resulting in a unified feature space across all modalities.

Besides, a geoscientific corpus could be used as the central language modality, while seismic data and other geophysical measurements could be aligned to this space using encoders trained via contrastive learning. The geophysical data can be represented through embedding techniques that are similar to the approach of treating depth or infrared data as RGB images. Multimodal large language models can combine seismic waveforms, electromagnetic, and gravity data with geological textual information to automatically identify subsurface structures, such as reservoirs or faults, thereby improving interpretation efficiency.

4.2.4 Adaptation for Task-specific Applications

When adapting pre-trained GeoFMs for specific geophysical tasks, fine-tuning is crucial to customize the model effectively while minimizing computational requirements. Given the large size of pre-trained models, Parameter-Efficient Fine-Tuning (PEFT) methods offer an effective solution by allowing adaptation with minimal modifications to the core model. Here, we discuss four commonly used fine-tuning approaches: additive fine-tuning, selective fine-tuning, reparametrized fine-tuning, and hybrid fine-tuning, as shown in Figure 8.

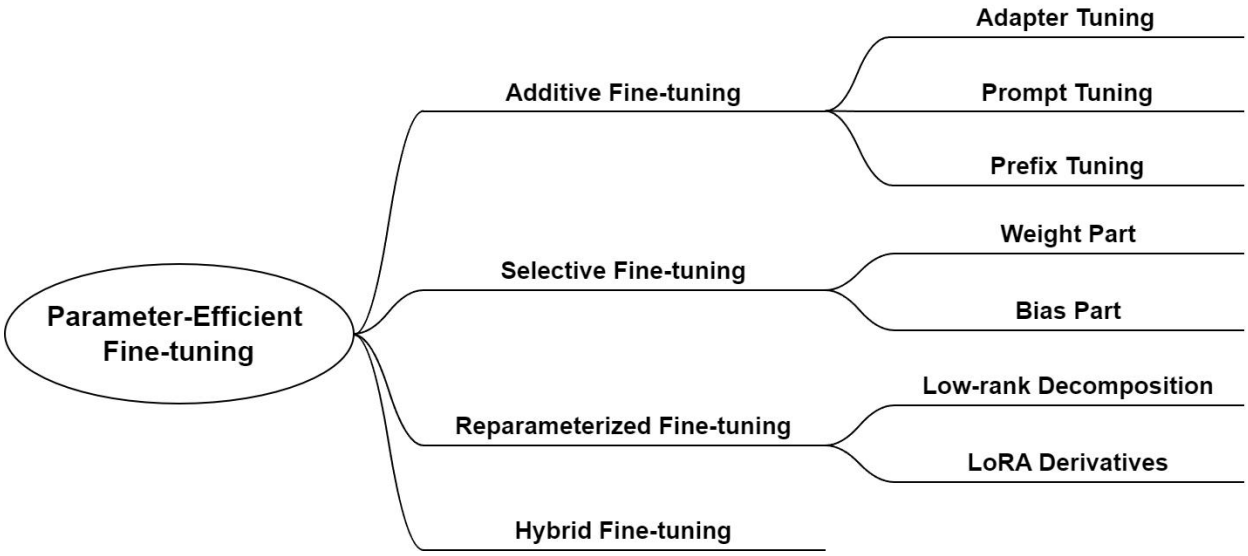


Figure 8 An overview of Parameter-Efficient Fine-Tuning.

Additive fine-tuning methods focus on introducing extra parameters to the pre-trained model without modifying its core structure. These methods add task-specific components, such as adapters or additional prompts, which can be fine-tuned while leaving the rest of the model frozen. This allows for task adaptation with minimal computational overhead. One popular method is the use of adapter[154] layers, which are small neural network modules added between the layers of the pre-trained model. Adapters are trained during fine-tuning while the pre-trained weights remain fixed. This technique is highly parameter-efficient, as it introduces only a small number of additional parameters relative to the entire model. Adapters have been used successfully in NLP tasks like question answering and text classification, as well as in vision tasks. A more recent approach, prompt tuning[155], involves adding learnable embeddings (prompts) to the input sequence. These prompts help the model attend to the most relevant features for the task at hand without altering the underlying weight of the model. Soft prompts[156] can be seen as a form of "task-specific attention" that guides the model's focus. Further, prefix-tuning[157] optimizes a continuous set of prompt vectors (prefix) added to the input of a pre-trained model to guide its generation task. This approach allows

the model to adapt to the new task by training only the prefix tokens while keeping the rest of the model fixed.

Selective fine-tuning involves updating only a small subset of the model's parameters, typically targeting specific layers or components that are most relevant for the task. This reduces the computational burden and helps mitigate overfitting by avoiding unnecessary adjustments to the entire model. For instance, bias adjustment[158,159] involves updating only the bias terms in certain layers of the model, leaving the majority of weights untouched. Another common approach is layer-wise fine-tuning, where only certain layers of the model are updated. For example, Gheini et al.[160] propose to fine-tune only the cross-attention layers. Such a method is particularly useful in cases where the lower layers capture general knowledge that does not need to be altered. Furthermore, LT-SFT[161] learns sparse, real-valued, task-specific and language-specific masks based on the Lottery Ticket Hypothesis, which are then composable with a pretrained model to enable efficient fine-tuning.

Reparameterized fine-tuning techniques aim to decompose model parameters into more efficient representations, making it easier to adapt the model to new tasks without significantly increasing computational costs. Low-Rank Adaptation (LoRA[113]) is a popular reparameterization method where weight matrices are decomposed into low-rank matrices, making it possible to update only the low-rank components during fine-tuning. This drastically reduces the number of trainable parameters, making it particularly useful for large models. The LoRA+[162] method improves upon the original LoRA by setting different learning rates for the two low-rank adapter matrices, correcting the suboptimal feature learning that arises when both matrices are updated with the same learning rate. The advanced approach, Laplace-LoRA[163], applies Laplace approximation to the low-rank adaptation

parameters of fine-tuned large language models, improving their calibration by estimating uncertainty. The LoRA Dropout[164] method introduces random noise and increases sparsity in the learnable low-rank matrices of LoRA-based models to control overfitting during fine-tuning, improving generalization and model calibration, and is further enhanced by a test-time ensemble strategy.

Hybrid fine-tuning[165,166] combines different strategies to take advantage of their individual strengths, leading to highly efficient fine-tuning methods. For instance, hybrid fine-tuning adapts the most critical parts of the model while introducing minimal overhead by combining additive and selective techniques.

These four types of fine-tuning offer a variety of tools for efficiently adapting GeoFMs to specific tasks in exploration geophysics. These fine-tuning methods make sure that GeoFMs can adapt to new geophysical environments, give more accurate geological interpretations, and lower computational costs. They do this by selectively updating model parameters or adding new parts like adapters and prefix tokens. For example, to improve the model's ability to understand specific geophysical tasks, such as lithology classification or seismic interpretation, instruction tuning[83] is performed using instruction-based data. Geoscientific questions (e.g., "Identify the main lithology in this log section") are paired with corresponding geophysical data and expected responses. This process helps adapt the pretrained model to domain-specific tasks and ensures it can effectively follow instructions in various geoscientific contexts. Besides, alignment tuning[167] focuses on improving the alignment across different modalities to ensure the model can interpret combined inputs. For instance, the model is trained to jointly process logs and seismic data while understanding corresponding geological descriptions. The model uses projection-based or query-based interfaces

and other connectors to make sure that multimodal geophysical data is interpreted in a way that makes sense. This makes it easier to think about features and structures below the surface.

4.2.5 Examples

In this subsection, we use the representative work in the field of exploration geophysics, the Seismic Foundation Model (SFM) proposed by Sheng et al.[110] (2023), as an example to illustrate how this workflow functions. SFM is a modality-specific GeoFM, and its main workflow includes data preparation, modality-specific pretraining, and task-specific adaptation, as illustrated in Figure 4. In the data preparation phase, Sheng et al. collected 192 open-source 3D migrated seismic datasets and converted them into 2D seismic data slices along the crossline and inline for subsequent training. In the pre-training phase, SFM used MAE-based self-supervised learning techniques to extract data features. Finally, in the downstream task adaptation phase, SFM adapts to different downstream tasks by training a decoder with a small number of parameters. The development workflow we proposed mentions these processes and methods, highlighting their applicability and importance in the development of GeoFMs.

5 Applications and Perspectives of Foundation Models in Exploration Geophysics

In this section, we explore the potential applications and future prospects of GeoFMs across different stages of exploration geophysics, including seismic data processing, imaging, and interpretation. As the development of GeoFMs in exploration geophysics is still in its early stages, this section aims to present a forward-looking perspective on how these models could transform the field. By examining possible applications, we illustrate the immense potential of GeoFMs in

advancing geophysical workflows and improving the efficiency and accuracy of exploration activities.

Figure 9 presents an overview of the potential applications of GeoFMs in exploration geophysics. For different seismic tasks, a variety of seismic foundation models can be utilized to handle data from different modalities and generate the desired outputs. Additionally, LLMs with exploration geophysics knowledge can be leveraged to facilitate user interaction and orchestrate different GeoFMs, completing complex seismic tasks through multimodal alignment and decision-making. This section aims to highlight not only the current applications but also the future possibilities and advancements that GeoFMs could bring to the field of exploration geophysics.

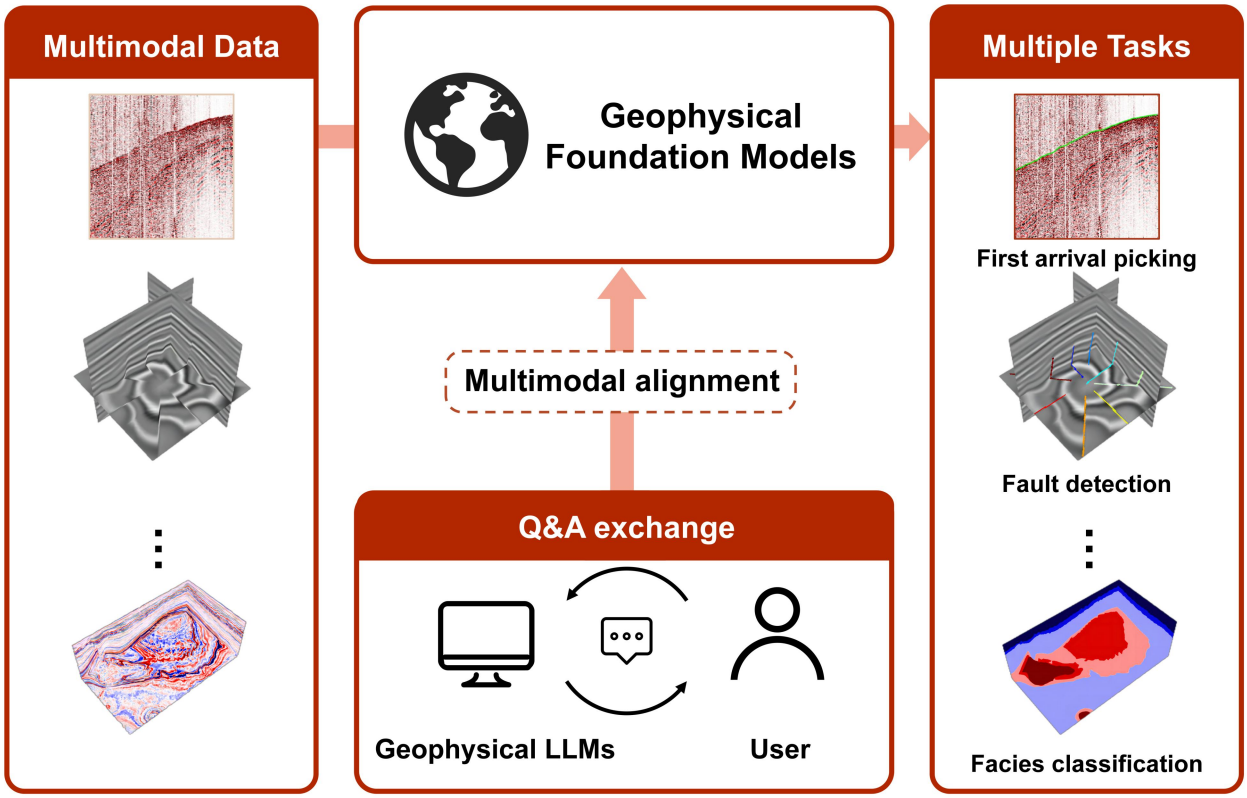


Figure 9 Overview of the GeoFMs application in the field of exploration geophysics. Users issue commands by interacting with seismic LLMs to process different modalities of input data through multimodal alignment using suitable foundation model, and outputting the results of target tasks, such as first-arrival picking, fault detection, facies classification, etc.

5.1 Seismic Data Processing

Seismic data processing aims to transform the collected raw seismic records into a format that facilitates subsequent seismic imaging and interpretation. Currently, GeoFMs in exploration geophysics are primarily applied to seismic data interpretation, while their development in data processing is still at an early stage. Therefore, this section will discuss the prospects of applying GeoFMs to data processing, including interpolation and denoising. Additionally, as GeoFMs in exploration geophysics have mostly focused on interpretation tasks so far, we propose a first-arrival picking method based on SAM to demonstrate the application of foundation models in geophysics and illustrate their potential in data processing tasks.

5.1.1 First-arrival Picking Based on SAM

In seismic data processing, first-arrival picking plays a pivotal role in estimating the subsurface velocity structure, as it pinpoints the initiation of the first signal arrivals. In recent years, deep learning methods have been able to achieve excellent first-arrival picking results on similar datasets after training. However, the generalization performance of deep learning methods is significantly affected by the noise and differences in data distribution.

SAM[2], developed by Meta AI, is considered the first foundation model for computer vision. With its training conducted on a massive corpus of data, which encompasses millions of images and billions of masks, SAM demonstrates a robust capability of delivering effective segmentation results across a wide range of image segmentation tasks. The first-arrival picking in seismic data is also a segmentation task. Thus, employing SAM for this task presents promising prospects. Therefore, we propose the application of SAM directly to the task of first-arrival picking in seismic data, to demonstrate the impact of large models on the paradigm of exploration geophysics research.

SAM is composed of three parts: a prompt encoder, an image encoder, and a mask decoder. The prompt encoder serves to encode the prompt (mask, points, box, and text) into embedding, while the image encoder is a pre-trained model based on the ViT[41] architecture. The mask decoder is a lightweight module that updates both image and prompt embeddings through cross-attention, which is ultimately used for dynamic mask outputs[2]. The process of the first-arrival picking based on SAM is shown in Figure 10. There are two ways to use SAM. The first is to directly utilize the automatic segmentation feature of SAM and then select the largest mask as the first-arrival picking result, since the seismic data usually occupies the largest part of the image. The second method involves manually setting prompts, which can achieve better results when a more refined segmentation is needed, and it has a faster running speed. However, the drawback is that it requires manual settings of the prompts.

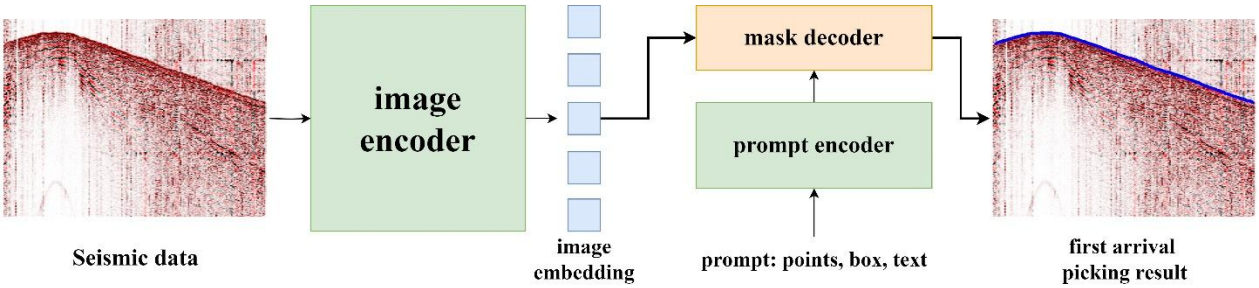
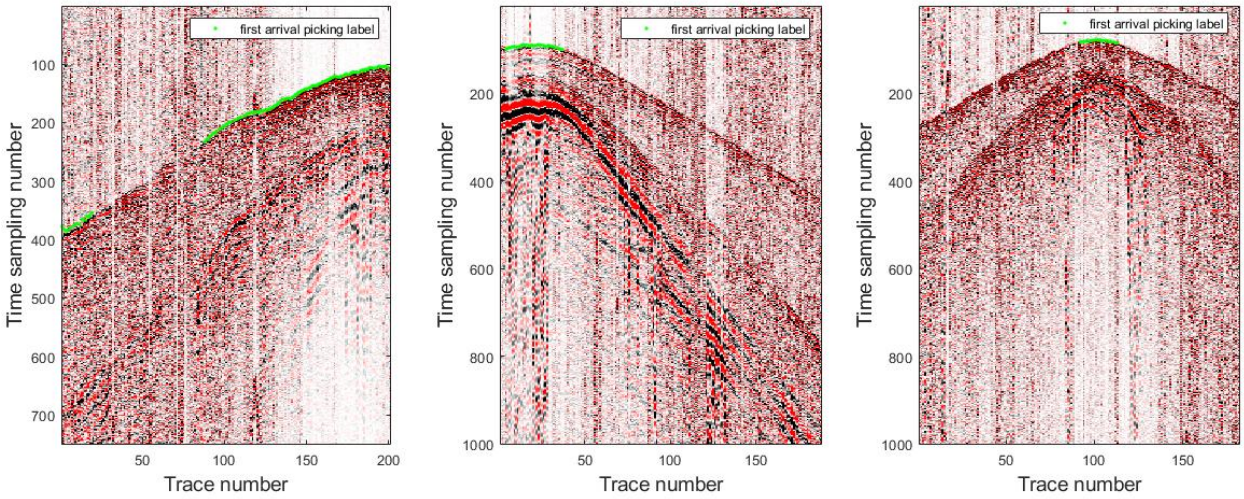


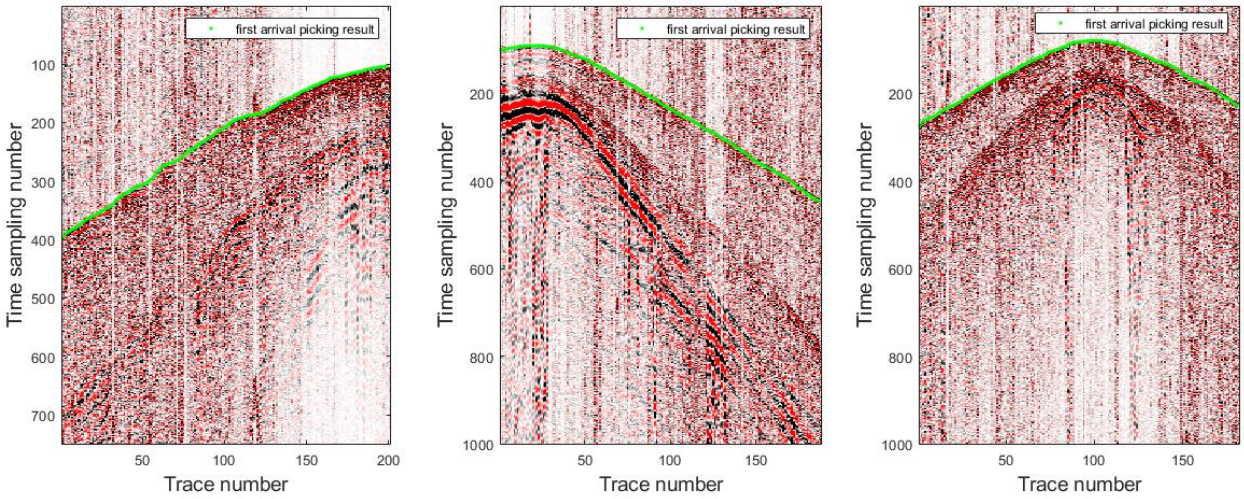
Figure 10 Overview of SAM-based first-arrival picking method. The seismic data and the prompt are encoded separately into embeddings by the image encoder and the prompt encoder, and subsequently fed into the mask decoder to generate the first-arrival picking results.

To better showcase the high generalization ability of SAM and its capability to fully automate the first-arrival picking process, we present the results obtained by the first method. The seismic data used for testing are derived from the public Halfmile and Sudbury datasets, published by St-Charles et al.[78] (2021). These datasets are very noisy, making it challenging to obtain accurate first-arrival

picking results. The first-arrival picking labels for 84.45% of the traces were provided by experts in these two datasets. Figure 11a-c present the line gathers and corresponding labels from the Halfmile and Sudbury datasets, which are challenging for manual picking and only provide a limited number of labels across the three line gathers. To demonstrate the generalization ability of the SAM-based method on seismic data and its robustness to noise, we directly apply SAM to the data without any special treatment. And the results, represented by green points, are shown in Figure 11d-f. Even in areas of strong noise interference where no labels are provided, the SAM-based first-arrival picking method can still produce effective results.



(a) test data (Halfmile) (b) test data (Sudbury) (c) test data (Sudbury)



(d) SAM-based result (Halfmile) (e) SAM-based result (Sudbury) (f) SAM-based result (Sudbury)

Figure 11 SAM-based first arrival picking results. (a-c) Test data from the Halfmile, Sudbury datasets, and the provided first-arrival picking labels, which were deemed valid by experts. (d-f) First-arrival picking results based on SAM.

Despite the excellent performance of the SAM-based first-arrival picking method on seismic data, there are still many issues worthy of further study. First, the efficiency of the current SAM-based method is relatively low. Taking the Halfmile dataset as an example, it takes about 5 seconds to process one line gather without parallelization, which is far from industrial-level applications. Hence, accelerating the SAM-based first-arrival picking method is an important research direction in the near future. Secondly, the existing SAM has been trained on image and text data and has not been adapted to seismic data. Consequently, it is unable to execute semantic segmentation tasks on seismic data, such as fault detection, identifying areas containing specific noise based on provided prompts, and so on. Potential future directions involve utilizing technologies such as Adapter[79] and LoRA[80] to fine-tune SAM on seismic data. These methods have already made breakthroughs in fields such as medicine[81] and remote sensing[82]. The first-arrival picking method based on SAM demonstrates considerable potential and superior performance, which could potentially influence the research paradigm in this field.

5.1.2 Interpolation and Denoising

Due to the limitations of the data collection conditions, raw seismic records always have problems such as missing data or containing various types of noise. The purpose of seismic data interpolation is to recover data from sparse samples, while the aim of denoising is to separate valid signals from noise. In the practice of seismic data processing, seismic data are typically transformed into two-dimensional line gathers, similar to the format of images. The resemblance has led to the

adoption of image processing techniques in most of the contemporary deep learning-based seismic data processing methods. Consequently, we can also adopt the image processing foundation models to develop seismic data processing foundation models.

In the raw seismic data, the missing data and noise are very complex, and the distribution of data varies wildly among different seismic data, impeding the generalization abilities of models and thus affecting the wide application of deep learning methods. To ensure that the trained model demonstrates robust generalization across various seismic datasets and processing tasks, there are two prevailing strategies for developing foundation models. The first strategy is based on pre-training and fine-tuning. This strategy initially utilizes a large-scale dataset to pretrain a model with a large number of parameters based on the MAE[83] framework, and then is fine-tuned in downstream tasks to achieve superior performance. Recently, Sheng et al.[77] (2023) developed the first seismic foundation model based on the pre-training and fine-tuning strategy. This model has demonstrated remarkable performance in seismic downstream tasks.

The second strategy is based on the seismic data priors of generative models, such as diffusion model[84,85], and GAN[86]. This strategy consists of two stages. The first stage is usually a corruption encoder, which encodes different types of missing data and noise into latent space. Subsequently, the second stage uses the embedding obtained from the first stage as prompts and employs a pre-trained diffusion model as the data prior for data restoration. Lin et al.[87] (2023) proposed diffusion models for the blind image restoration problem (DiffBIR), which pre-trained a restoration module to improve generalization capability and leveraged fixed stable diffusion through LAControNet for reconstruction. Wang et al.[88] (2023) presented StableSR, which only requires the fine-tuning of a lightweight, time-aware encoder to capture the degradation features.

In the realm of seismic data, the development of foundation models for interpolation and denoising presents substantial challenges. First, there is a notable lack of pre-trained models based on generative models serving as data priors in geophysics. Secondly, in contrast to image processing scenarios, seismic data embodies a wider variety and complexity of noise types. Consequently, the task of training an encoder to grasp the features of noise becomes a significant problem. Moreover, the most critical issue is the lack of large quantities of high-quality seismic data processing samples for training. These issues could be potential directions for future research in the advancement of GeoFMs.

5.2 Seismic Imaging

Seismic imaging is a technique that creates images of subsurface structures by analyzing seismic waves. It is a challenging problem because traditional methods such as the full waveform inversion (FWI) or seismic tomography are highly dependent on the constraints of physics equations. Currently, deep learning-based seismic imaging methods can be primarily classified into two categories. The first is end-to-end learning, which directly learns the mapping from seismic data to the imaging domain[22,89]. This learning strategy can yield promising results on synthetic datasets that are similar to the training data. However, once there are changes in data distribution or model parameters, the application of the trained model becomes limited by its generalization capability. The second category employs deep learning as an auxiliary tool to assist in completing certain steps in traditional seismic imaging methods. Sun et al.[90] (2023) utilized learned regularization as a constraint to optimize the FWI process. Ovcharenko et al.[91] (2019) extrapolated low-frequency from the respective high-frequency components of the seismic data to provide low-frequency information for FWI based on deep learning.

At present, the success of foundation models is mainly on applications that rely on experience and generative tasks. However, the comprehension and learning of physical laws for the solutions of partial differential equations (PDEs) are still in an initial stage. Seismic imaging, such as FWI, relies heavily on the forward modeling of wave equations, which is also the most time-consuming part of FWI. Therefore, it is extremely challenging to achieve a direct mapping from seismic data to images based on GeoFMs in the short term. Here, we propose two potential directions of development for foundation models in the field of seismic imaging. First, the acceleration of FWI is intrinsically linked to the efficient forward modeling of wave equations. Therefore, the initial research direction is to develop foundation models for solving PDEs. In recent years, the emergence of Fourier Neural Operator (FNO)[89] technology has provided a new perspective for seismic imaging[90,91], especially in handling the physics-informed constraints inherent in seismic data. FNOs learn the forward modeling process, mapping from the velocity model to seismic data, and can accelerate seismic imaging by speeding up the forward modeling stage. While FNOs offer a promising approach by directly learning mappings in the Fourier space, which can help improve the generalization capability of seismic imaging models to some extent, their limited generalization capability on complex real-world data restricts their practical application. This limitation arises because most FNOs are trained on synthetic datasets and random velocity models, which do not provide the data volume and parameter scale required for a foundation model. Ye et al.[92] (2024) introduced PDEformer, a foundation model for solving PDEs based on graph transformer architecture. However, there remains a gap before its application in seismic imaging. The second possible research direction is to integrate foundation models with traditional methods to overcome the inherent bottlenecks in conventional methods, such as the lack of low-frequency information. Most of these data preprocessing modules

can be integrated within the foundation models for the interpolation and denoising section discussed in Section 4.1.2. Inspired by the GeoFM of weather forecasting[74], predicting the gradient changes during the FWI process could be a strategy to accelerate FWI. Changing the gradient during the FWI process, similar to weather forecasting tasks, involves predicting a trend over a certain number of time steps at specific grid points, and can be corrected based on the result at the current moment.

5.3 Interpretation

Seismic interpretation involves the identification of geological features, including subsurface structures and properties, to locate potential target areas underground. The process includes tasks such as fault detection, geobody identification and segmentation, facies classification, and attribute analysis. Seismic interpretation is inherently complex due to the extensive data volume, the intricate nature of subsurface features, the uncertainties associated with expert interpretations, and the need to understand multimodal information. This multimodal nature aligns with the earlier discussion on multimodal levels, emphasizing the necessity of combining different data sources to effectively interpret geological features.

In the field of exploration geophysics, seismic interpretation is one of the fastest-growing areas for the application of GeoFMs. This rapid development can be largely attributed to the similarity between many seismic interpretation tasks and those found in computer vision. For instance, fault detection in seismic interpretation closely parallels edge detection in computer vision, while geobody segmentation is similar to object segmentation. These similarities have enabled foundational models from computer vision to be fine-tuned and adapted for seismic data interpretation, significantly accelerating progress in this area. For instance, Gao et al. (2023) introduced a foundation model empowered by a multi-modal prompt engine, integrating pre-trained vision foundation models fine-

tuned on seismic data, which demonstrated promising results for universal seismic geobody interpretation across surveys. Guo et al. (2023) introduced a cross-domain foundation model adaptation approach by adapting computer vision foundation models to geoscience. They developed a workflow that leverages existing LVMs and fine-tunes them for geoscientific tasks, demonstrating effectiveness in processing and interpreting data such as lunar images, seismic data, and DAS arrays. Another promising strategy is to train multimodal geophysical foundation models from scratch. For example, Han et al. (2024) introduced multi-attributes masking contrastive learning (MAMCL) technique for explainable seismic facies analysis. Traditional methods often require manual attribute selection and lack interpretability, while MAMCL uses an interpretable framework with a depthwise CNN for feature extraction and an iTransformer for feature aggregation. By employing an unsupervised contrastive learning strategy, MAMCL improves both efficiency and explainability in seismic facies interpretation.

6 Challenges and Future Trends of Geophysics in the Era of Foundation Models

6.1 Challenges

Despite the promising prospects of GeoFMs, there are still challenges to their development and application in the field of geophysics.

1) Data Scarcity and Quality: The limitations of available geophysical data pose a primary challenge in the development of GeoFMs. Although the volume of open-source geophysical data is massive, reaching TB scale, there are several limitations that hinder the development of GeoFMs: (1) Insufficient Data Diversity: While a single exploration area may have a large volume of seismic data,

there is a significant difference in data distribution among different exploration regions. To enable GeoFMs to have better generalization ability, diverse data from different exploration regions are required, rather than large amounts of data from a few target areas. (2) Lack of Labels: For field seismic data, there is often a shortage of corresponding high-quality labeled data, which presents a significant bottleneck in training models that require well-curated datasets for supervised learning. (3) Access Restrictions: Despite the large volume of geophysical data, there is not enough high-quality open-source data available for GeoFMs development. Much of the high-quality data remains inaccessible due to confidentiality, legal, and privacy issues, limiting its use for training and publication. This scarcity of high-quality open-source data significantly hampers the ability to train robust models. These challenges collectively impact the ability to train models that generalize effectively across different geophysical environments.

2) Benchmark: Deep learning methods have achieved great success and development in the field of exploration geophysics. However, unlike fields such as NLP or CV, there is currently a lack of unified benchmarks for comparison in the geophysics domain. In NLP and CV, widely accepted benchmark datasets and metrics provide a standardized way to evaluate and compare model performance. In exploration geophysics, different researchers often employ different preprocessing procedures, data selection, and evaluation metrics, which can lead to significant discrepancies in the results obtained, even when using the same neural network architecture for the same task. This lack of standardization makes it difficult to objectively evaluate and compare different models under a unified framework, hindering the establishment of best practices and slowing down progress in the field. Developing standardized benchmark datasets and evaluation criteria would be an important step toward fostering more consistent progress and enabling fair comparisons across different

approaches.

3) Computational Resources and Costs: Training foundation models, particularly those with multimodal capabilities, requires vast computational resources. The training process is a complex engineering challenge that demands significant time and financial investment. Foundation models often need large-scale data, multiple training iterations, and powerful hardware setups, which can lead to months of training time and substantial costs. The computational cost of training such models is often prohibitive, especially for smaller research institutions or companies without access to state-of-the-art hardware. Despite the fact that many oil and gas companies possess substantial computational resources, these resources are often not integrated in a way that meets the requirements for training foundation models. In exploration geophysics, the training of GeoFMs is particularly resource-intensive due to the need for large-scale simulations, multimodal data integration, and domain-specific adaptations. Although some companies have considerable computational capabilities, they are often fragmented and lack the cohesion necessary to meet the demands of large-scale model training. Additionally, the energy consumption associated with training large models raises environmental and sustainability concerns, necessitating more efficient training techniques and hardware optimizations. These factors collectively make the training of foundation models a resource-intensive endeavor that requires careful planning and efficient resource management.

4) Interpretability and Reliability: In exploration geophysics, the need for interpretability is paramount. Incorrect predictions can lead to misguided exploration efforts, resulting in significant financial losses. Therefore, models used in this field must not only be accurate but also interpretable and reliable. Foundation models, however, are inherently complex, often consisting of billions of

parameters, which makes them difficult to interpret and understand. The "black box" nature of large foundation models poses a challenge in gaining the trust of domain experts who need to understand how the model arrives at its conclusions, especially when the stakes are high. The lack of transparency in model decision-making can be a significant barrier to their adoption in geophysics, where experts require clear insights into the reasoning behind model outputs to make informed decisions. Developing methods to improve model interpretability—such as incorporating attention mechanisms, visualization tools, or explanatory modules—will be crucial in making GeoFMs more acceptable and trustworthy for real-world applications. Additionally, ensuring that models are reliable under different operational conditions and geological settings is essential, as the consequences of incorrect interpretations can have far-reaching implications, both economically and environmentally.

5) Lack of Physical Laws: Many tasks in geophysics are highly dependent on physical principles. For example, geophysical imaging largely relies on constraints imposed by partial differential equations (PDEs) that govern wave propagation and subsurface mechanics. Foundation models must be able to incorporate these physical principles to ensure that their predictions are physically plausible and consistent with established scientific theories. However, integrating these physics-based constraints into data-driven models remains a challenging problem, as most existing foundation models are designed primarily for pattern recognition rather than incorporating domain-specific physical knowledge. Most of these models are trained using autoregressive and other similar approaches, which lack explicit integration of physical mechanisms. This gap hinders their ability to accurately represent geophysical processes that are fundamentally governed by physical laws. Achieving effective integration of physical laws in GeoFMs is crucial to enhance model reliability

and ensure that outputs are not only accurate but also scientifically valid.

6) Unclear Role and Applications: Currently, the traditional workflows and software for geophysical data processing and interpretation are well-established and have matured over decades of development. This maturity raises questions about the specific value that GeoFMs can bring to the field. While foundation models have shown promise in transforming industries such as natural language processing and computer vision, their exact role in exploration geophysics remains unclear. Unlike other domains where foundation models can directly replace or augment existing workflows, the adoption of GeoFMs in geophysics is less straightforward. The traditional workflows already provide reliable and well-tested methods for data processing, imaging, and interpretation. Therefore, it is essential to define the unique problems that GeoFMs can solve more effectively compared to existing solutions. For example, geophysical copilots could serve as valuable assistants, helping users operate complex software programs or manage downstream task-specific models more efficiently, thereby enhancing the overall productivity of geophysical workflows. However, without clearly identifying these specific applications, the integration of GeoFMs into the geophysical workflow may face skepticism and resistance. Thus, articulating a clear and compelling case for GeoFMs in exploration geophysics is a significant challenge that must be addressed to foster their widespread adoption.

6.2 Future Trends

To address the challenges highlighted in Section 6.1, this section outlines future trends that can guide the development and adoption of GeoFMs in exploration geophysics.

1) Data Integration and Benchmark Standardization: To overcome the challenge of data scarcity and quality, the integration of diverse geophysical datasets from multiple exploration regions

is essential. Developing unified benchmarks for GeoFMs will help establish standardized datasets and evaluation metrics, similar to those available in NLP and computer vision. These benchmarks will enable fair comparisons across different models and approaches, fostering more consistent progress in geophysics. Establishing well-curated open datasets and common benchmarks can provide a solid foundation for GeoFM training and evaluation, accelerating development in this domain.

2) Collaboration Among Institutions: Addressing the resource-intensive nature of GeoFM training requires collaboration among different stakeholders—oil and gas companies, AI companies, and academic institutions. Oil and gas companies can provide the high-quality geophysical data needed for training, while AI companies and universities can offer the computational resources and technical expertise required to develop large-scale foundation models. A collaborative approach, leveraging the strengths of different institutions, can help overcome individual limitations and facilitate the training of GeoFMs at scale.

3) Building Certainty and Interpretability in Complex Systems: Interpretability and reliability are crucial for the adoption of GeoFMs in exploration geophysics. Future research should focus on improving the transparency of model decision-making by incorporating explainable AI techniques. Attention mechanisms, visualization tools, and explanatory modules can help make GeoFMs more interpretable for geophysicists, allowing them to understand and trust the model's outputs. Moreover, ensuring the reliability of GeoFMs under different operational conditions is essential to avoid costly errors, making it a priority for future research to develop models that can provide consistent and reliable predictions.

4) Model Integration with Physical Mechanisms: Future trends in GeoFM development include

integrating physical principles into data-driven models. By embedding physical constraints, such as those imposed by partial differential equations, into the training process, GeoFMs can produce predictions that are consistent with established geophysical theories. Hybrid approaches that combine data-driven learning with physical modeling are expected to emerge as a powerful method for enhancing the reliability and scientific validity of GeoFMs. These methods will enable foundation models to handle the complexities of geophysical processes while adhering to the underlying physics.

5) Defining Clear Applications for GeoFMs in Exploration Geophysics: For GeoFMs to gain widespread acceptance, their role in exploration geophysics must be clearly defined. While traditional workflows and software are already mature, GeoFMs could provide unique value in automating complex tasks, integrating multimodal data, and enhancing data interpretation through better uncertainty quantification. Geophysical copilots, for instance, could assist in operating complex software, managing task-specific models, or integrating various data sources. Identifying specific use cases where GeoFMs outperform existing methods will help foster confidence in their adoption and create opportunities for more efficient geophysical workflows. Addressing these challenges will be crucial for realizing the full potential of foundation models in transforming geophysical research and exploration.

7 Conclusions

The emergence of foundation models has radically transformed the research paradigm in exploration geophysics, shifting from conventional, rule-based methodologies to a data-driven, foundation model-based framework adept at tackling intricate, multimodal geophysical challenges. This study offers an extensive overview of GeoFMs, including their present advancements, hierarchy,

development workflow, applications in exploration geophysics, as well as the challenges and future trends. We commenced with a review of foundation models, emphasizing their emergence and progress in the domain. We subsequently examined their hierarchy, highlighting the varied capacities of distinct types of GeoFMs and the ways in which these models can synergistically enhance one another. We introduced a generalized development workflow that outlines key stages from data preparation to downstream adaptation. Subsequently, we examined prospective uses of GeoFMs in exploration geophysics, concentrating on data processing, imaging, and interpretation while also discussing problems encountered and future trends in their advancement. The advancement of GeoFMs will create boundless opportunities for technological progress in geophysics and will also revolutionize the research paradigm in this discipline. The upcoming journey may present many challenges, but with persistent research and innovation, we stand poised to witness a revolution in the field of exploration geophysics.

8 Reference

1. Achiam, J., et al. Gpt-4 technical report. Preprint at https://doi.org/10.48550/arXiv.2303.08774 (2023).

2. Kirillov, A., et al. 2023. Segment anything. Proceedings of the IEEE/CVF International Conference on Computer Vision, 4015-4026 (2023).

3. OpenAI. Video generation models as world simulators. OpenAI: https://openai.com/research/video-generation-models-as-world-simulators (2015).

4. Castro Nascimento, C. M., & Pimentel, A. S. Do large language models understand chemistry? a conversation with ChatGPT. Journal of Chemical Information and Modeling. **63** (6), 1649-1655

(2023).

5. Guo, T., et al. What can large language models do in chemistry? A comprehensive benchmark on eight tasks. Advances in Neural Information Processing Systems, **36** (2023).

6. Li, Y., Xu, H., Zhao, H., Guo, H., & Liu, S. Chatpathway: Conversational large language models for biology pathway detection. NeurIPS 2023 AI for Science Workshop (2023).

7. Li, Y., Wang, S., Ding, H., & Chen, H. Large language models in finance: A survey. Proceedings of the Fourth ACM International Conference on AI in Finance (2023).

8. Wu, S., et al. BloombergGPT: A large language model for finance. Preprint at https://doi.org/10.48550/arXiv.2303.17564 (2023).

9. Moor, M., et al. Foundation models for generalist medical artificial intelligence. Nature, **616** (7956), 259-265 (2023).

10. Guo, X., et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. Preprint at https://doi.org/10.48550/arXiv.2312.10115 (2023).

11. Liu, F., et al. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. IEEE Transactions on Geoscience and Remote Sensing (Early Access) (2024).

12. Mall, U., et al., Remote sensing vision-language foundation models without annotations via ground remote alignment. International Conference on Learning Representations (2024).

13. Yu, S., & Ma, J. Deep learning for geophysics: Current and future trends. Reviews of Geophysics, **59** (3), e2021RG000742 (2021).

14. Spitz, S. Seismic trace interpolation in the F-X domain. Geophysics, **56** (6), 785-794 (1991).

15. Donoho, D. L., & Johnstone, I. M. Adapting to unknown smoothness via wavelet

shrinkage. Journal of the American Statistical Association, **90** (432), 1200-1224 (1995).

16. Oropeza, V., & Sacchi, M. Simultaneous seismic data denoising and reconstruction via multichannel singular spectrum analysis. Geophysics, **76** (3), V25-V32 (2011).

17. Sacchi, M. D., Ulrych, T. J., & Walker, C. J. Interpolation and extrapolation using a high-resolution discrete Fourier transform. IEEE Transactions on Signal Processing, **46** (1), 31-38 (1998).

18. Herrmann, F. J., & Hennenfent, G. Non-parametric seismic data recovery with curvelet frames. Geophysical Journal International, **173** (1), 233-248 (2008).

19. Trad, D., Ulrych, T., & Sacchi, M. Latest views of the sparse Radon transform. Geophysics, **68** (1), 386-399 (2003).

20. Hu, L., et al. First-arrival picking with a U-net convolutional network. Geophysics, **84** (6), U45-U57 (2019).

21. Mandelli, S., et al. Seismic data interpolation through convolutional autoencoder. SEG International Exposition and Annual Meeting (2018).

22. Yang, F., & Ma, J. Deep-learning inversion: A next-generation seismic velocity model building method. Geophysics, **84** (4), R583-R599 (2019).

23. Bergen, K. J., Johnson, P. A., de Hoop, M. V., & Beroza, G. C. Machine learning for data-driven discovery in solid Earth geoscience. Science, **363** (6433), eaau0323 (2019).

24. Naveed, H., et al. A comprehensive overview of large language models. Preprint at https://doi.org/10.48550/arXiv.2307.06435 (2023).

25. Zhao, W. X., et al. A survey of large language models. Preprint at https://doi.org/10.48550/arXiv.2303.18223 (2023).

26. Brown, T., et al. Language models are few-shot learners. Advances in Neural Information

Processing Systems, **33**, 1877-1901 (2020).

27. Vaswani, A., et al. Attention is all you need. Advances in Neural Information Processing Systems, **30** (2017).

28. Christiano, P. F., et al. Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, **30** (2017).

29. Chowdhery, A., et al. Palm: Scaling language modeling with pathways. Journal of Machine Learning Research, **24** (240), 1-113 (2023).

30. Touvron, H., et al. Llama 2: Open foundation and fine-tuned chat models. Preprint at https://doi.org/10.48550/arXiv.2307.09288 (2023)

31. Anthropic. Introducing Claude. Anthropic Blog https://www.anthropic.com/news/introducing-claude (2024).

32. Thoppilan, R., et al. LaMDA: Language models for dialog applications. Preprint at https://doi.org/10.48550/10.48550/arXiv.2201.08239 (2022).

33. Zeng, A., et al. GLM-130b: An open bilingual pre-trained model. International Conference on Learning Representations (2023).

34. Zhang, S., et al., OPT: Open pre-trained transformer language models." Preprint at https://doi.org/10.48550/arXiv.2205.01068 (2022).

35. Hoffmann, J., et al. An empirical analysis of compute-optimal large language models. Advances in Neural Information Processing Systems, **35**, 30016-30030 (2022).

36. Zhang, Y., Wei, C., Wu, S., He, Z., & Yu, W. (2023). GeoGPT: understanding and processing geospatial tasks through an autonomous GPT. arXiv preprint arXiv:2307.07930.

37. Kuckreja, K., et al. Geochat: Grounded large vision-language model for remote

sensing. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024).

38. Bran, A. M., et al. ChemCrow: Augmenting large-language models with chemistry tools. NeurIPS 2023 AI for Science Workshop (2023).

39. Li, Y., Wang, H., & Luo, Y. A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports. 2020 IEEE International Conference on Bioinformatics and Biomedicine (2020).

40. Wang, Z., et al. MedCLIP: Contrastive learning from unpaired medical images and text. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 3876-3887 (2022).

41. Qiu, J., et al. Large AI models in health informatics: Applications, challenges, and the future. IEEE Journal of Biomedical and Health Informatics, **27** (12), 6074-6087 (2023).

42. Dosovitskiy, A., et al. An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (2021).

43. Yu, F., et al. Scaling up to excellence: practicing model scaling for photo-realistic image restoration in the Wild. Preprint at https://doi.org/10.48550/2401.13627 (2024).

44. Radford, A., et al. Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning, **139**, 8748-8763 (2021).

45. Ramesh, A., et al. Hierarchical text-conditional image generation with clip latents. Preprint as https://doi.org/10.48550/arXiv.2204.06125 (2022).

46. Saharia, C., et al. Photorealistic text-to-image diffusion models with deep language

understanding. Advances in Neural Information Processing Systems, **35**, 36479-36494 (2022).

47. Rombach, R., et al. High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021).

48. Wang, J., et al. Review of large vision models and visual prompt engineering. Meta-Radiology, 100047 (2023).

49. Bai, Y., et al. Sequential modeling enables scalable learning for large vision models. Preprint at https://doi.org/10.48550/arXiv.2312.00785 (2023).

50. Guo, J., et al. Data-efficient large vision models through sequential autoregression. Preprint at https://doi.org/10.48550/arXiv.2402.04841 (2024).

51. Tiu, E., et al. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nature Biomedical Engineering, **6** (12), 1399-1406 (2022).

52. Liang, P. P., Zadeh, A., & Morency, L. P. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. ACM Computing Surveys, **56**(10), 1-42 (2024).

53. Alayrac, J. B., et al. Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems, **35**, 23716-23736 (2022).

54. Girdhar, R., et al. Imagebind: One embedding space to bind them all. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 15180-15190 (2023).

55. Ramesh, A., et al. Zero-shot text-to-image generation. International Conference on Machine Learning, 8821-8831 (2021).

56. Saharia, C., et al. Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems, **35**, 36479-36494 (2022).

57. Li, J., et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and

large language models. In International Conference on Machine Learning, 19730-19742 (2023).

58. Guo, Z., Wu, X., Liang, L., Sheng, H., Chen, N., & Bi, Z. (2024). Cross-Domain Foundation Model Adaptation: Pioneering Computer Vision Models for Geophysical Data Analysis. arXiv preprint arXiv:2408.12396.

59. Han, L., Wu, X., Hu, Z., Li, J., & Fang, H. (2024). MAMCL: Multi-attributes Masking Contrastive Learning for explainable seismic facies analysis. Computers & Geosciences, 193, 105731.

60. Zheng, Q., & Zhang, D. (2022). RockGPT: reconstructing three-dimensional digital rocks from single two-dimensional slice with deep learning. Computational Geosciences, 26(3), 677-696.

61. Kumar A. Vision Transformer Based Foundation Model for Oil Reservoir Forecasting[C]//85th EAGE Annual Conference & Exhibition (including the Workshop Programme). European Association of Geoscientists & Engineers, 2024, 2024(1): 1-5.

62. Zhang, H., Xu, J. J., Cui, H. W., Li, L., Yang, Y., Tang, C. S., & Boers, N. (2023). When geoscience meets foundation models: Towards general geoscience artificial intelligence system. arXiv preprint arXiv:2309.06799.

63. Hadid, A., Chakraborty, T., & Busby, D. (2024). When geoscience meets generative AI and large language models: Foundations, trends, and future challenges. Expert Systems, e13654.

64. DATA, M. (2024). Multimodal artificial intelligence foundation models: Unleashing the power of remote sensing big data in earth observation. Innovation, 2(1), 100055.

65. Lu, S., Guo, J., Zimmer-Dauphinee, J. R., Nieusma, J. M., Wang, X., VanValkenburgh, P., ... & Huo, Y. (2024). Ai foundation models in remote sensing: A survey. arXiv preprint arXiv:2408.03464.

66. Jiao, L., Huang, Z., Lu, X., Liu, X., Yang, Y., Zhao, J., ... & Feng, J. (2023). Brain-inspired

remote sensing foundation models and open problems: A comprehensive survey. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

67. LIU He , REN Yili , LI Xin , DENG Yue , WANG Yongtao , CAO Qianwen , DU Jinyang , LIN Zhiwei , WANG Wenjie. Research status and application of artificial intelligence large models in the oil and gas industry[J]. Petroleum Exploration and Development, 2024, 51(4): 910-923. https://doi.org/10.11698/PED.20240254

68. Si, X., Wu, X., Sheng, H., Zhu, J., & Li, Z. (2024). SeisCLIP: A seismology foundation model pre-trained by multi-modal data for multi-purpose seismic feature extraction. IEEE Transactions on Geoscience and Remote Sensing.

69. Liu, T., Münchmeyer, J., Laurenti, L., Marone, C., de Hoop, M. V., & Dokmanić, I. (2024). SeisLM: a Foundation Model for Seismic Waveforms. arXiv preprint arXiv:2410.15765.

70. Zhu, X. X., Xiong, Z., Wang, Y., Stewart, A. J., Heidler, K., Wang, Y., ... & Shi, Y. (2024). On the Foundations of Earth and Climate Foundation Models. arXiv preprint arXiv:2405.04285.

71. Zhang, L., Cui, H., Song, Y., Li, C., Yuan, B., & Lu, M. (2024). On the Opportunities of (Re)-Exploring Atmospheric Science by Foundation Models: A Case Study. arXiv preprint arXiv:2407.17842.

72. Zheng, Z., Chen, Y., Zeng, H., Vu, T. A., Hua, B. S., & Yeung, S. K. (2025). MarineInst: A Foundation Model for Marine Image Analysis with Instance Visual Description. In European Conference on Computer Vision (pp. 239-257). Springer, Cham.

73. Bi, Z., Zhang, N., Xue, Y., Ou, Y., Ji, D., Zheng, G., & Chen, H. (2023). Oceangpt: A large language model for ocean science tasks. arXiv preprint arXiv:2310.02031.

74. Chen, K., Liu, C., Chen, H., Zhang, H., Li, W., Zou, Z., & Shi, Z. (2024). RSPrompter: Learning

to prompt for remote sensing instance segmentation based on visual foundation model. IEEE Transactions on Geoscience and Remote Sensing.

75. Sultan, R. I., Li, C., Zhu, H., Khanduri, P., Brocanelli, M., & Zhu, D. (2023). GeoSAM: Fine-tuning SAM with sparse and dense visual prompting for automated segmentation of mobility infrastructure. arXiv preprint arXiv:2311.11319.

76. Huang, Z., et al. STU-Net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. Preprint at https://doi.org/10.48550/2304.06716 (2023).

77. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.

78. Lu, M. Y., Chen, B., Williamson, D. F., Chen, R. J., Zhao, M., Chow, A. K., ... & Mahmood, F. (2024). A multimodal generative AI copilot for human pathology. Nature, 634(8033), 466-473.

79. Liang, P. P., Zadeh, A., & Morency, L. P. (2024). Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. ACM Computing Surveys, 56(10), 1-42.

80. Christie, G., et al. Functional map of the world. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6172-6180 (2018).

81. Bastani, F., et al. SatlasPretrain: A large-scale dataset for remote sensing image understanding. Proceedings of the IEEE/CVF International Conference on Computer Vision, 16772-16782 (2023).

82. Zhan, Y., Xiong, Z., & Yuan, Y. RSVG: Exploring data and models for visual grounding on remote sensing data. IEEE Transactions on Geoscience and Remote Sensing, **61**, 1-13 (2023).

83. Liu, H., et al. Visual instruction tuning. Advances in Neural Information Processing Systems, **36**

(2024).

84. Ross, Z. E., Meier, M. A., & Hauksson, E. P wave arrival picking and first-motion polarity determination with deep learning. Journal of Geophysical Research: Solid Earth, **123** (6): 5120-5129 (2018).

85. Pardo, E., Garfias, C., & Malpica, N. Seismic phase picking using convolutional networks IEEE Transactions on Geoscience and Remote Sensing, **57** (9), 7086-7092 (2019).

86. Mousavi, S. M. et al. Earthquake transformer – an attentive deep learning model for simultaneous earthquake detection and phase picking. Nature Communications, **11** (1), 3952 (2020).

87. Zhao, M., et al. DiTingmotion: A deep learning first motion polarity classifier and its application to focal mechanism inversion. Frontier in Earth Science, **11**,1103914 (2023).

88. Yano, K., et al. Graph-partitioning based convolutional neural network for earthquake detection using a seismic array. Journal of Geophysical Research: Solid Earth, **126** (5), e2020JB020269 (2021).

89. Yang, S., Hu, J., Zhang, H., & Liu, G. Simultaneous earthquake detection on multiple stations via a convolutional neural network. Seismological Society of America, **92** (1), 246-260 (2021).

90. DeVries, P. M., et al. Deep learning of aftershock patterns following large earthquakes. Nature, **560** (7720), 632-634 (2018).

91. Zhang, X., et al. Locating induced earthquakes with a network of seismic stations in Oklahoma via a deep learning method. Scientific reports, **10** (1), 1941 (2020).

92. Rouet-Leduc, B., et al. Machine learning predicts laboratory earthquakes. Geophysical Research Letters, **44** (18), 9276-9282 (2017).

93. Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., & Anandkumar, A. (2020). Fourier neural operator for parametric partial differential equations. arXiv preprint

arXiv:2010.08895.

94. Lehmann, F., Gatti, F., Bertin, M., & Clouteau, D. (2023). Fourier neural operator surrogate model to predict 3D seismic waves propagation. arXiv preprint arXiv:2304.10242.

95. Yang, Y., Gao, A. F., Azizzadenesheli, K., Clayton, R. W., & Ross, Z. E. (2023). Rapid seismic waveform modeling and inversion with neural operators. IEEE Transactions on Geoscience and Remote Sensing, 61, 1-12.

96. Johnson, P. A., et al. Laboratory earthquake forecasting: A machine learning competition. Proceedings of the National Academy of Sciences, **118** (5), e2011362118 (2021).

97. Borate, P., et al. Using a physics-informed neural network and fault zone acoustic monitoring to predict lab earthquakes. Nature Communications, **14** (1), 3693 (2023).

98. Si, X., Wu, X., Sheng, H., Zhu, J., & Li, Z. SeisCLIP: A seismology foundation model pre-trained by multimodal data for multi-purpose seismic feature extraction. IEEE Transactions on Geoscience and Remote Sensing, **62** (2024).

99. Li, S., et al. SeisT: A foundation deep learning model for earthquake monitoring tasks. IEEE Transactions on Geoscience and Remote Sensing, **62** (2024).

100. Zhao, M., Xiao, Z., Chen, S., & Fang, L. (2023). DiTing: A large-scale Chinese seismic benchmark dataset for artificial intelligence in seismology. Earthquake Science, 36(2), 84-94.

101. Ni, Y., Hutko, A., Skene, F., Denolle, M., Malone, S., Bodin, P., ... & Wright, A. (2023). Curated Pacific Northwest AI-ready seismic dataset.

102. Ritchie, H., et al. Implementation of the semi-Lagrangian method in a high-resolution version of the ECMWF forecast model. Monthly Weather Review, **123** (2), 489-514 (1995).

103. Molteni, F., Buizza, R., Palmer, T. N., & Petroliagis, T. The ECMWF ensemble prediction

system: Methodology and validation. Quarterly Journal of the Royal Meteorological Society, **122** (529), 73-119 (1996).

104. Weyn, J. A., Durran, D. R., & Caruana, R. Can machines learn to predict weather? Using deep learning to predict gridded 500-hPa geopotential height from historical weather data. Journal of Advances in Modeling Earth Systems, **11** (8), 2680-2693 (2019).

105. Rasp, S., et al. WeatherBench: a benchmark dataset for data-driven weather forecasting. Journal of Advances in Modeling Earth Systems, **12** (11), e2020MS002203 (2020).

106. Zhang, Y., et al. Skilful nowcasting of extreme precipitation with NowcastNet. Nature, **619**, 526-532 (2023).

107. Bi, K., et al. Accurate medium-range global weather forecasting with 3D neural networks. Nature, **619**, 533-538 (2023).

108. Bi, Z., et al. OceanGPT: A large language model for ocean science tasks. Preprint at https://doi.org/10.48550/arXiv.2310.02031 (2024).

109. Xiong, W., et al. AI-GOMS: Large-AI driven global ocean modeling system. Preprint at https://doi.org/10.48550/2308.03152 (2023).

110. Sheng, H., et al. Seismic Foundation Model (SFM): a new generation deep learning model in geophysics. Preprint at https://doi.org/10.48550/arXiv.2309.02791 (2023).

111. St-Charles, P. L., et al. A multi-survey dataset and benchmark for first break picking in hard rock seismic exploration. Proceedings of the 2021 NeurIPS Workshop on Machine Learning for the Physical Sciences (2021).

112. Houlsby, N., et al. Parameter-efficient transfer learning for NLP. Proceedings of the 36th International Conference on Machine Learning, **97**, 2790-2799 (2019).

113. Hu, E. J., et al. Lora: Low-rank adaptation of large language models. International Conference on Learning Representations (2021).

114. Chen, T., et al. SAM-Adapter: Adapting segment anything in underperformed scenes. Proceedings of the IEEE/CVF International Conference on Computer Vision, 3367-3375 (2023).

115. Wang, D., et al. Samrs: Scaling-up remote sensing segmentation dataset with segment anything model. Advances in Neural Information Processing Systems, **36** (2024).

116. He, K., et al. Masked autoencoders are scalable vision learners. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022).

117. Song, Y., & Ermon, S. Generative modeling by estimating gradients of the data distribution. Advances in Neural Information Processing Systems, **32** (2019).

118. Ho, J., Jain, A., & Abbeel, P. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, **33**, 6840-6851 (2020).

119. Goodfellow, I., et al. Generative adversarial networks. Communications of the ACM, **63** (11), 139-144 (2020).

120. Lin, X., et al. DiffBIR: Towards blind image restoration with generative diffusion prior. Preprint at https://doi.org/10.48550/arXiv.2308.15070 (2023).

121. Wang, J., Yue, Z., Zhou, S., Chan, K. C., & Loy, C. C. Exploiting diffusion prior for real-world image super-resolution. Preprint at https://doi.org/10.48550/arXiv.2305.07015 (2023).

122. Araya-Polo, M., Jennings, J., Adler, A., & Dahlke, T. Deep-learning tomography. The Leading Edge, **37** (1), 58-66 (2018).

123. Sun, P., Yang, F., Liang, H., & Ma, J. Full-waveform inversion using a learned

regularization. IEEE Transactions on Geoscience and Remote Sensing, **61** (2023).

124. Ovcharenko, O., Kazei, V., Kalita, M., Peter, D., & Alkhalifah, T. Deep learning for low-frequency extrapolation from multioffset seismic data. Geophysics, **84** (6), R989-R1001 (2019).

125. Ye, Z., et al. PDEformer: Towards a foundation model for one-dimensional partial differential equations." Preprint at https://doi.org/10.48550/arXiv.2402.12652 (2024).

126. Deng, C., et al. K2: A foundation language model for geoscience knowledge understanding and utilization. Proceedings of the 17th ACM International Conference on Web Search and Data Mining (2024).

127. Lin, Z., et al. GeoGalactica: A scientific large language model in geoscience. Preprint at https://doi.org/10.18550/arXiv.2401.00434 (2023).

128. Taylor, R., et al. Galactica: A large language model for science. Preprint at https://doi.org/10.48550/arXiv.2211.09085 (2022).

129. Li, C., et al. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems (2024).

130. Zhan, J., Dai, J., Ye, J., Zhou, Y., Zhang, D., Liu, Z., ... & Qiu, X. (2024). Anygpt: Unified multimodal llm with discrete sequence modeling. arXiv preprint arXiv:2402.12226.

131. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597-1607). PMLR.

132. He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9729-9738).

133. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.

134. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in neural information processing systems, 33, 6840-6851.

135. Liu, Q., & Ma, J. (2024). Generative interpolation via a diffusion probabilistic model. Geophysics, 89(1), V65-V85.

136. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695).

137. Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

138. Marfurt, K. J. (1984). Accuracy of finite-difference and finite-element modeling of the scalar and elastic wave equations. Geophysics, 49(5), 533-549.

139. Gavrilyuk, K., Sanford, R., Javan, M., & Snoek, C. G. (2020). Actor-transformers for group activity recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 839-848).

140. Shi, B., Hsu, W. N., Lakhotia, K., & Mohamed, A. (2022). Learning audio-visual speech representation by masked multimodal cluster prediction. arXiv preprint arXiv:2201.02184.

141. Li, R., Yang, S., Ross, D. A., & Kanazawa, A. (2021). Ai choreographer: Music conditioned 3d dance generation with aist++. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 13401-13412).

142. Lin, J., Yang, A., Zhang, Y., Liu, J., Zhou, J., & Yang, H. (2020). Interbert: Vision-and-

language interaction for multi-modal pretraining. arXiv preprint arXiv:2003.13198.

143. Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems, 32.

144. Zhan, X., Wu, Y., Dong, X., Wei, Y., Lu, M., Zhang, Y., ... & Liang, X. (2021). Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 11782-11791).

145. Zhu, B., Lin, B., Ning, M., Yan, Y., Cui, J., Wang, H., ... & Yuan, L. (2023). Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. arXiv preprint arXiv:2310.01852.

146. Zhang, D., Li, S., Zhang, X., Zhan, J., Wang, P., Zhou, Y., & Qiu, X. (2023). Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. arXiv preprint arXiv:2305.11000.

147. Zhang, H., Li, X., & Bing, L. (2023). Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858.

148. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020, August). End-to-end object detection with transformers. In European conference on computer vision (pp. 213-229). Cham: Springer International Publishing.

149. Tsai, Y. H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L. P., & Salakhutdinov, R. (2019, July). Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the conference. Association for computational linguistics. Meeting (Vol. 2019, p. 6558). NIH Public Access.

150. Murahari, V., Batra, D., Parikh, D., & Das, A. (2020, August). Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In European Conference on Computer Vision (pp. 336-352). Cham: Springer International Publishing.

151. Swetha, S., Yang, J., Neiman, T., Rizve, M. N., Tran, S., Yao, B., ... & Shah, M. (2024). X-Former: Unifying Contrastive and Reconstruction Learning for MLLMs. arXiv preprint arXiv:2407.13851.

152. Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., & Cai, D. (2023). Pandagpt: One model to instruction-follow them all. arXiv preprint arXiv:2305.16355.

153. Pi, R., Gao, J., Diao, S., Pan, R., Dong, H., Zhang, J., ... & Zhang, T. (2023). Detgpt: Detect what you need via reasoning. arXiv preprint arXiv:2305.14167.

154. Chen, H., Tao, R., Zhang, H., Wang, Y., Li, X., Ye, W., ... & Savvides, M. (2024). Conv-adapter: Exploring parameter efficient transfer learning for convnets. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1551-1561).

155. Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691.

156. Petrov, A., Torr, P. H., & Bibi, A. (2023). When do prompting and prefix-tuning work? a theory of capabilities and limitations. arXiv preprint arXiv:2310.19698.

157. Li, X. L., & Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190.

158. Zaken, E. B., Ravfogel, S., & Goldberg, Y. (2021). Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. arXiv preprint arXiv:2106.10199.

159. Fu, C. L., Chen, Z. C., Lee, Y. R., & Lee, H. Y. (2022). Adapterbias: Parameter-efficient token-

dependent representation shift for adapters in nlp tasks. arXiv preprint arXiv:2205.00305.

160. Gheini, M., Ren, X., & May, J. (2021). Cross-attention is all you need: Adapting pretrained transformers for machine translation. arXiv preprint arXiv:2104.08771.

161. Ansell, A., Ponti, E. M., Korhonen, A., & Vulić, I. (2021). Composable sparse fine-tuning for cross-lingual transfer. arXiv preprint arXiv:2110.07560.

162. Hayou, S., Ghosh, N., & Yu, B. (2024). Lora+: Efficient low rank adaptation of large models. arXiv preprint arXiv:2402.12354.

163. Yang, A. X., Robeyns, M., Wang, X., & Aitchison, L. (2023). Bayesian low-rank adaptation for large language models. arXiv preprint arXiv:2308.13111.

164. Lin, Y., Ma, X., Chu, X., Jin, Y., Yang, Z., Wang, Y., & Mei, H. (2024). Lora dropout as a sparsity regularizer for overfitting control. arXiv preprint arXiv:2404.09610.

165. He, J., Zhou, C., Ma, X., Berg-Kirkpatrick, T., & Neubig, G. (2021). Towards a unified view of parameter-efficient transfer learning. arXiv preprint arXiv:2110.04366.

166. Hu, Z., Wang, L., Lan, Y., Xu, W., Lim, E. P., Bing, L., ... & Lee, R. K. W. (2023). Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. arXiv preprint arXiv:2304.01933.

167. Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3128-3137).

168. Xu, P., Zhu, X., & Clifton, D. A. (2023). Multimodal learning with transformers: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(10), 12113-12132.

169. Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., & Tang, J. (2021). Self-supervised

learning: Generative or contrastive. IEEE transactions on knowledge and data engineering, 35(1), 857-876.