International Journal of Research and Review in Applied Science, Humanities, and Technology
Vol 1, Issue 2 December 2024 ISSN: 3048-975X
https://ijrasht.com/

Foundation Models Beyond Language: A Survey of Vision, Biology, Robotics, and Multimodal Systems

Amit, Department of Computer Science and Engineering (AI & ML), Panipat Institute of Engineering and Technology (PIET), Samalkha, Panipat, Haryana, India 2822429.cse@pietgroup.co.in

Anshu, Department of Computer Science and Engineering (AI & ML), Panipat Institute of Engineering and Technology (PIET), Samalkha, Panipat, Haryana, India 2822433.cse@pietgroup.co.in

Riya, Department of Computer Science and Engineering (AI & ML), Panipat Institute of Engineering and Technology (PIET), Samalkha, Panipat, Haryana, India 2822409.cse@pietgroup.co.in

Shruti, Department of Computer Science and Engineering (AI & ML), Panipat Institute of Engineering and Technology (PIET), Samalkha, Panipat, Haryana, India 2822346.cse@pietgroup.co.in

Taruna, Department of Computer Science and Engineering (AI & ML), Panipat Institute of Engineering and Technology (PIET), Samalkha, Panipat, Haryana, India taruna.cse-aiml@piet.co.in

Abstract— Foundation models — large neural networks pretrained on broad data and adapted to many downstream tasks — first rose to prominence in natural language processing, but the same underlying recipe of large-scale self-supervised pretraining followed by task-specific adaptation has since been extended well beyond text. This paper surveys foundation models developed for vision, biological sequence and structure data, robotics, and multimodal settings that jointly reason over combinations of these modalities. We introduce a taxonomy organized around modality and pretraining objective, describe representative architectures and self-supervised learning strategies used outside the language domain, and examine applications spanning medical imaging, autonomous robotics, genomics, drug discovery, and multimodal virtual assistants. We summarize benchmarks used to evaluate these systems, present a worked case study of a vision-language-action model for robotic manipulation, and discuss challenges including modality-specific data scarcity, the difficulty of defining meaningful scaling laws outside text, and the added complexity of evaluating embodied and multimodal competence. We conclude by outlining directions connecting foundation models across modalities with embodied learning, scientific discovery, and unified multimodal reasoning.

Index Terms—foundation models, self-supervised learning, multimodal learning, vision transformers, robotics foundation models, genomic language models, vision-language-action models

I. INTRODUCTION

The term "foundation model" was coined to describe a class of large neural networks trained on broad, often internet-scale data using self-supervised objectives, producing general-purpose representations that can be adapted, typically through fine-tuning or prompting, to a wide variety of downstream tasks without training a new model from scratch for each task. This paradigm was popularized first and most visibly in natural language processing, where large transformer-based language models pretrained on massive text corpora demonstrated strong performance across an enormous range of downstream language tasks using comparatively little task-specific labeled data.

The success of this recipe — large-scale self-supervised pretraining followed by lightweight downstream adaptation — has since been extended well beyond text. Vision foundation models pretrained on billions of images now provide general-purpose visual representations usable across classification, detection, and segmentation tasks. Biological foundation models pretrained on protein and DNA sequence databases capture structural and functional patterns useful for downstream tasks in structural biology and genomics. Robotics foundation models pretrained on large, diverse collections of robot interaction data aim to provide general-purpose control policies adaptable to new tasks and embodiments. Multimodal foundation models jointly pretrained on paired data spanning two or more modalities, such as images and text, aim to produce representations and generative capabilities that span modality boundaries entirely.

Extending the foundation model paradigm beyond language is not a simple matter of substituting a different data type into an unchanged recipe. Each modality presents distinct challenges: images lack the natural discrete token structure of text, requiring different tokenization and self-supervised objective design; biological sequences carry structural and evolutionary constraints not present in natural language; robotic interaction data is far scarcer and more expensive to collect than web text or images, since it typically requires physical hardware and, in many cases, human supervision or teleoperation; and multimodal settings require architectures and training objectives capable of aligning

representations across fundamentally different data types. This paper surveys how the field has adapted the foundation model paradigm to address these modality-specific challenges.

The remainder of this paper proceeds as follows. Section II reviews background on what constitutes a foundation model and the historical progression from language to other modalities. Section III presents a taxonomy organized by modality and pretraining strategy. Section IV describes core methods in technical detail. Section V surveys applications. Section VI summarizes benchmarks. Section VII offers a comparative discussion. Section VIII presents an illustrative case study of a vision-language-action model. Section IX discusses implementation tooling. Section X considers ethical implications. Section XI discusses open challenges, Section XII outlines future directions, and Section XIII concludes.

II. BACKGROUND AND MOTIVATION

A. What Makes a Model a Foundation Model

A foundation model is generally characterized by three properties: it is trained on a broad base of data rather than a narrowly curated dataset for a single task; it is trained using a self-supervised or weakly supervised objective that does not require expensive manual labeling at the scale of the full training corpus; and it is designed to be adapted to a wide range of downstream tasks, often with substantially less task-specific data than would be required to train a comparable model from scratch. This last property, broad downstream adaptability from a single pretrained model, is what most clearly distinguishes the foundation model paradigm from earlier practices of training separate, task-specific models for each application.

The economic and practical appeal of this paradigm is considerable: the substantial cost of large-scale pretraining is incurred once and amortized across many downstream applications, while the cost of adapting the pretrained model to a specific task, whether through fine-tuning, prompting, or lightweight adapter modules, is generally far lower than training a comparable task-specific model from randomly initialized weights, particularly in domains where labeled data for any single downstream task is limited.

B. From Language to Multimodal Foundation Models

The specific self-supervised objectives that proved effective for language, such as predicting a masked or subsequent word given surrounding context, exploit the fact that text is naturally discrete and sequential, with a next-token prediction objective providing a rich and scalable training signal from raw, unlabeled text at essentially unlimited scale. Extending this recipe to vision required developing analogous self-supervised objectives suited to continuous, spatially structured image data, such as predicting masked image patches or learning representations that remain invariant to different augmented views of the same image. Extending the recipe to biological sequences required accounting for the specific statistical structure of protein and DNA sequences, which, while also discrete and sequential like text, are governed by evolutionary and structural constraints quite different from the grammatical and semantic structure of natural language.

Multimodal foundation models introduce an additional requirement beyond single-modality self-supervision: a mechanism for aligning representations across modalities, so that, for example, an image of a dog and the text description "a photograph of a dog" are mapped to similar points in a shared representation space, enabling downstream tasks such as image captioning, visual question answering, and text-to-image generation that inherently require reasoning across modality boundaries rather than within a single modality alone.

C. Historical Context

Self-supervised representation learning in vision predates the foundation model era, with earlier work on autoencoders and contrastive representation learning establishing that useful visual representations could be learned without manual labels well before the term "foundation model" was coined. The transformer architecture, originally developed for language, was subsequently adapted to vision through architectures that treat an image as a sequence of patches processed analogously to word tokens, enabling much of the same large-scale pretraining infrastructure developed for language models to be repurposed for vision with comparatively modest architectural modification. This cross-

pollination of architectural ideas between modalities, particularly the widespread adoption of the transformer architecture across vision, biology, and robotics, has been an important practical enabler of the rapid extension of the foundation model paradigm beyond its original language setting, since it allowed researchers in other modalities to build directly on the substantial engineering and optimization infrastructure already developed for training large transformer-based language models.

III. A TAXONOMY OF FOUNDATION MODELS BEYOND LANGUAGE

Foundation models beyond language can be organized along two dimensions: the modality or modalities of data used for pretraining, and the nature of the self-supervised or weakly supervised objective employed. Single-modality foundation models are pretrained exclusively on data from one modality, such as images alone or protein sequences alone, and are typically adapted to downstream tasks within that same modality, though the learned representations sometimes transfer usefully to related modalities as well. Multimodal foundation models are pretrained jointly on paired or interleaved data spanning two or more modalities, most commonly images and text, and are designed to support downstream tasks that require reasoning across modality boundaries.

Along the objective dimension, contrastive pretraining objectives learn representations by encouraging matched pairs of inputs, such as an image and its paired caption, or two augmented views of the same image, to have similar representations while encouraging mismatched pairs to have dissimilar representations. Generative or reconstructive objectives instead train a model to reconstruct masked or corrupted portions of an input, or to generate an entire output conditioned on a partial or transformed input, and have been adapted across essentially every modality discussed in this paper, from masked image modeling in vision to masked residue prediction in protein language models. Table I summarizes representative foundation models organized along both dimensions.

TABLE I. TAXONOMY OF FOUNDATION MODELS BEYOND LANGUAGE

Modality	Objective Type	Representative Models	Primary Downstream Use
Vision (single-modality)	Contrastive / self-distillation	SimCLR, DINO, MAE	Classification, detection, segmentation
Biology (protein/DNA)	Masked sequence modeling	ESM, ProtBERT, DNABERT	Structure/function prediction, design
Robotics	Behavior cloning / multi-task pretraining	RT-1, RT-2, Octo	General-purpose robotic control
Multimodal (vision-language)	Contrastive + generative	CLIP, Flamingo, vision-language-action models	Captioning, VQA, embodied instruction following

IV. CORE METHODS

A. Vision Foundation Models

Contrastive vision foundation models learn representations by generating two randomly augmented views of the same image, such as different crops or color distortions, and training an encoder such that the two views' representations are pulled together while representations of different images are pushed apart, without requiring any manual labels. Self-distillation approaches extend this idea by training a student network to match the output of a slowly updated teacher network applied to a differently augmented view of the same image, and have been observed to produce representations in which semantically meaningful structure, such as object boundaries, emerges without any explicit segmentation supervision.

Masked image modeling, an alternative generative approach analogous to masked language modeling, trains a model to reconstruct randomly masked patches of an input image given the surrounding unmasked patches, using a vision transformer architecture that processes the image as a sequence of patch tokens. This objective has proven highly effective at scale and, notably, requires substantially less compute per training step than contrastive approaches that typically require processing multiple augmented views of each training image simultaneously, though the two families of objectives are not mutually exclusive and several state-of-the-art vision foundation models combine elements of both.

B. Biological and Genomic Foundation Models

Protein language models apply the masked sequence modeling objective directly to amino acid sequences, treating each amino acid as an analogue of a word token, and have been shown to learn representations that capture meaningful structural and evolutionary information, including secondary structure and residue-residue contact patterns, purely from patterns of co-occurrence and substitution across large databases of evolutionarily related protein sequences, without any explicit structural supervision during pretraining. These learned representations have subsequently been used as inputs to downstream structure prediction, function prediction, and protein design models, in some cases substantially reducing the computational cost of structure prediction relative to methods that rely more heavily on searching large databases of related sequences at inference time.

Genomic foundation models apply analogous self-supervised sequence modeling objectives to DNA sequences, learning representations useful for downstream tasks such as predicting the functional effect of a genetic variant, identifying regulatory elements, or predicting gene expression from sequence context. Genomic sequence modeling presents distinctive challenges relative to protein and natural language modeling, including much longer effective context lengths required to capture long-range regulatory interactions and a comparatively lower information density per nucleotide relative to amino acids or words, both of which have motivated specialized architectural adaptations for long-context genomic sequence modeling.

C. Robotics Foundation Models

Robotics foundation models aim to learn general-purpose control policies from large, diverse datasets of robot demonstrations spanning multiple tasks, environments, and, in more recent work, multiple distinct robot embodiments, in contrast to the traditional practice of training a separate control policy from scratch for each individual task and robot. Vision-language-action models extend multimodal foundation model architectures by adding action prediction as an additional output modality, typically by fine-tuning a pretrained vision-language model to output robot control commands conditioned on a visual observation and a natural language task instruction, allowing the resulting model to leverage broad visual and linguistic knowledge acquired during vision-language pretraining even though comparatively little of that original pretraining data involved robotic control directly.

A central challenge specific to robotics foundation models, relative to vision or language foundation models, is the comparative scarcity and cost of collecting large-scale, diverse robot interaction data, since unlike text or images, robot demonstration data generally requires either physical robot hardware operating in the real world or a sufficiently realistic simulation environment, both of which are far more expensive and slower to scale than collecting text or images from the web. This has motivated substantial collaborative efforts to aggregate robot demonstration data across many research institutions and robot platforms into shared, standardized datasets suitable for large-scale pretraining, an approach sometimes described as building an "internet-scale" dataset for robotics by pooling many smaller, individually collected datasets.

D. Multimodal Foundation Models

Contrastive vision-language models learn a shared representation space for images and text by training on large datasets of image-caption pairs collected from the web, encouraging matched image-caption pairs to have similar representations while unmatched pairs are pushed apart, a training objective that has proven effective at learning representations transferable to a wide range of downstream vision-language tasks, including zero-shot image classification performed simply by comparing an image's representation to the representations of candidate class-name text descriptions.

Generative multimodal models extend beyond representation alignment to support open-ended multimodal generation, such as producing a free-form natural language answer to a question about an input image, or generating an image conditioned on a natural language text description, typically by combining a pretrained vision encoder with a large language model, connected through a learned mapping that projects visual representations into the language model's input space, allowing the language model's existing linguistic and reasoning capabilities to be extended to visual inputs with comparatively modest additional multimodal-specific training.

E. Self-Supervised Pretraining Objectives Beyond Language

Across vision, biology, and robotics, self-supervised pretraining objectives can be broadly grouped into contrastive objectives, which learn representations through relative comparison between examples, and generative or predictive objectives, which learn representations by requiring a model to reconstruct or predict some portion of its input. An important practical consideration in selecting between these objective families is the relationship between the pretraining objective and the intended downstream use: contrastive objectives tend to produce representations well suited to discriminative downstream tasks such as classification, while generative objectives, particularly when the pretraining task itself resembles the ultimate downstream generative task, such as reconstructing masked image regions as a precursor to image editing or generation, tend to transfer more directly to downstream generative applications.

F. Scaling Laws Beyond Language

Empirical scaling laws relating model performance to model size, dataset size, and training compute, first characterized in detail for language models, have been investigated across other modalities with mixed results. Vision foundation models have generally exhibited scaling behavior qualitatively similar to language models, with performance improving predictably as model and dataset size increase, though the specific quantitative scaling exponents differ from those observed in language. Scaling behavior for robotics foundation models remains comparatively less well characterized, in significant part because the scarcity of large-scale, diverse robot interaction data has so far limited experiments to a smaller range of dataset sizes than has been explored in vision or language, leaving open the question of whether robotics foundation models will continue to improve predictably with further data scaling in the manner observed in other modalities, or whether the field will encounter distinct data-scarcity-driven limitations.

An additional complication in characterizing scaling behavior for biological foundation models is that the effective diversity of available sequence data is fundamentally limited by evolutionary history itself, rather than by data collection effort alone: unlike web text, which continues to grow as new content is created, or images, which can be collected essentially without bound, the number of naturally occurring, evolutionarily distinct protein families is finite, meaning that simply collecting more sequence data eventually yields diminishing returns in novel information content, a qualitatively different scaling regime than the comparatively open-ended data availability that has driven continued scaling in vision and language.

V. APPLICATIONS

A. Medical Imaging

Vision foundation models pretrained on large collections of natural images or, increasingly, on large collections of medical images specifically, have been adapted to downstream medical imaging tasks including tumour detection, organ segmentation, and disease classification from radiological, pathological, and ophthalmological imagery, often achieving strong performance even when fine-tuned on comparatively small, institution-specific labelled datasets, addressing the chronic scarcity of large labelled medical imaging datasets relative to natural image domains.

Multimodal foundation models combining medical imaging with associated clinical text, such as radiology reports, have also been developed to support tasks such as automated report generation from an imaging study, or retrieval of similar historical cases given a new patient's imaging and clinical presentation, extending the vision-language

foundation model paradigm into a specialized clinical setting with correspondingly higher requirements for accuracy and interpretability given the consequential nature of clinical decisions.

B. Autonomous Robotics and Embodied AI

Robotics foundation models are increasingly used as a general-purpose starting point for a wide range of downstream manipulation and navigation tasks, substantially reducing the amount of task-specific demonstration data required to train a functional control policy for a new task relative to training entirely from scratch, and, in the case of models trained across multiple robot embodiments, sometimes enabling a degree of direct transfer of learned skills to robot hardware not represented at all in the original pretraining data.

Beyond individual manipulation tasks, embodied AI research more broadly explores agents that must perceive, plan, and act within simulated or physical three-dimensional environments over extended time horizons, a setting in which foundation models pretrained on vision and language provide useful prior knowledge about object semantics and common-sense physical relationships, even though such models were not originally trained with embodied action in mind, illustrating a degree of beneficial knowledge transfer across modalities that has been an important motivation for continued investment in multimodal and robotics foundation model research.

C. Genomics and Drug Discovery

Protein and genomic foundation models are used throughout drug discovery pipelines, supporting tasks such as predicting how a candidate small molecule will interact with a target protein, predicting the functional consequences of a genetic mutation relevant to disease risk, and, in generative applications, proposing novel protein sequences expected to fold into a desired structure or perform a desired biochemical function, extending the discussion of AI-driven scientific discovery to specifically leverage the foundation model paradigm of broad pretraining followed by targeted downstream adaptation.

In agricultural and environmental biotechnology, genomic foundation models pretrained across diverse plant and microbial genomes are being explored to accelerate crop improvement and the engineering of microorganisms for applications such as biodegradation of pollutants or industrial enzyme production, extending the foundation model paradigm's benefits, namely reduced downstream labeled-data requirements, to biological domains that have historically received less machine learning research attention than human health applications.

D. Multimodal Virtual Assistants

Multimodal foundation models increasingly underpin consumer-facing virtual assistants capable of accepting images, and in some systems audio or video, alongside text as input, supporting use cases such as answering questions about a photographed document, describing the contents of an image for users with visual impairments, or assisting with visually grounded tasks such as identifying a plant species or troubleshooting a physical device from a photograph.

These assistants increasingly support multi-turn, mixed-modality conversations in which a user might share an image, ask a follow-up question referring back to an earlier part of the conversation, and receive a response that integrates information across multiple previously shared images and turns of dialogue, a capability that places substantially greater demands on a model's ability to maintain and reason over an extended, mixed-modality context relative to single-turn image-captioning or visual question answering tasks that dominated earlier multimodal benchmark evaluation.

E. Remote Sensing and Earth Observation

Vision foundation models pretrained specifically on satellite and aerial imagery support downstream applications including land-use classification, deforestation monitoring, crop yield estimation, and disaster damage assessment, often adapted from models originally pretrained on natural ground-level photography, though performance has generally improved further when pretraining is instead conducted directly on remote sensing imagery, which differs systematically from natural images in viewing angle, spectral characteristics, and spatial resolution.

Remote sensing foundation models also increasingly incorporate multiple satellite data modalities beyond standard optical imagery, such as synthetic aperture radar, which can penetrate cloud cover and operate independently of daylight, and multispectral or hyperspectral sensors that capture wavelength information beyond the visible spectrum relevant to tasks such as vegetation health assessment, with multimodal pretraining across these complementary data sources showing improved robustness relative to models trained on optical imagery alone, particularly for monitoring applications in regions with persistent cloud cover where optical imagery is frequently unavailable.

VI. BENCHMARKS AND EVALUATION

Evaluating foundation models beyond language requires benchmarks tailored to each modality's characteristic downstream tasks. Vision foundation models are commonly evaluated through linear probing or fine-tuning on standard image classification, detection, and segmentation benchmarks, as well as through zero-shot transfer performance for multimodal vision-language models. Biological foundation models are evaluated on structure prediction accuracy, variant effect prediction accuracy, and downstream function prediction tasks drawn from curated biological databases. Robotics foundation models are evaluated through success rate on held-out manipulation or navigation tasks, often in both simulated and physical robot environments, given the comparative difficulty and cost of large-scale physical robot evaluation relative to simulation.

Table II summarizes representative benchmarks across these modalities. A common evaluation theme across all of these modalities, echoing similar concerns raised regarding causal machine learning and AI for scientific discovery earlier in this survey series, is the distinction between in-distribution benchmark performance and genuine out-of-distribution generalization, which is particularly emphasized in robotics evaluation, where a model's ability to generalize to novel objects, environments, or robot embodiments not represented in its training data is often considered a more meaningful measure of genuine progress than performance on tasks closely resembling the training distribution.

TABLE II. REPRESENTATIVE BENCHMARKS FOR FOUNDATION MODELS BEYOND LANGUAGE

Benchmark	Modality	Task
ImageNet / linear probe suite	Vision	Image classification transfer
COCO	Vision	Object detection and segmentation
VQA / visual reasoning benchmarks	Multimodal	Visual question answering
CASP	Biology	Protein structure prediction
Open X-Embodiment benchmarks	Robotics	Cross-embodiment manipulation success rate

VII. COMPARATIVE ANALYSIS AND DISCUSSION

Across the modalities surveyed in this paper, vision foundation models are the most methodologically mature, benefiting from architectural and optimization techniques directly inherited from language modeling and from an abundance of large-scale unlabelled image data comparable in scale to text corpora used for language model pretraining. Biological foundation models occupy an intermediate position: pretraining data, while smaller in scale than web text or images, is nonetheless substantial for well-studied biological sequence databases, and the resulting models have already demonstrated clear practical impact, particularly in structure prediction. Robotics foundation models remain the least mature of the modalities surveyed, constrained fundamentally by the comparative scarcity of large-scale, diverse physical interaction data, a constraint that is unlikely to be resolved simply by further algorithmic innovation absent continued substantial investment in physical data collection infrastructure.

Multimodal foundation models introduce a distinct set of trade-offs related to alignment quality between modalities: representations that align well for one downstream task, such as zero-shot classification, do not necessarily align equally well for another, such as fine-grained visual question answering requiring detailed spatial reasoning, and much of the recent methodological innovation in multimodal foundation models has focused specifically on improving the granularity and reliability of cross-modal alignment, for example through architectures that preserve more detailed spatial visual information when projecting into a language model's input space, rather than relying solely on a single, coarse image-level representation.

A further point of comparison across modalities concerns the relative maturity of open ecosystems supporting independent research and reproduction. Vision and multimodal foundation models benefit from a comparatively rich ecosystem of openly released pretrained checkpoints, training code, and standardized evaluation harnesses, allowing academic and smaller industrial research groups to build directly on state-of-the-art pretrained models without needing to reproduce costly pretraining themselves. Biological foundation models have followed a broadly similar open-release pattern for many prominent models, particularly protein language and structure prediction models, reflecting norms of open data sharing already well established in structural biology prior to the foundation model era. Robotics foundation models, by contrast, have generally been released less openly, whether due to the specialized physical hardware required for independent reproduction, competitive commercial considerations, or the comparative immaturity of shared community norms around open robotics research relative to the other modalities discussed in this survey.

TABLE III. COMPARATIVE MATURITY OF FOUNDATION MODEL MODALITIES

Modality	Pretraining Data Scale	Methodological Maturity	Primary Bottleneck
Vision	Very large (billions of images)	High	Fine-grained spatial/compositional reasoning
Biology	Large for well-studied sequence families	Moderate to high	Coverage of understudied organisms/families
Robotics	Comparatively small	Low to moderate	Physical data collection cost and scarcity
Multimodal	Large (web-scale paired data)	Moderate to high	Cross-modal alignment granularity

VIII. ILLUSTRATIVE CASE STUDY

Consider a representative vision-language-action model designed to control a robotic arm performing tabletop manipulation tasks specified in natural language, such as "pick up the red block and place it in the bowl." The model is built by starting from a pretrained vision-language foundation model, which has already learned rich, general-purpose visual and linguistic representations from large-scale web image-text data, and extending its output space to include a sequence of low-level robot action tokens, such as discretized end-effector position and gripper commands, in addition to its original text output capability.

This extended model is then fine-tuned on a dataset combining a comparatively large amount of general vision-language data with a smaller but still substantial amount of robot demonstration data collected across many distinct manipulation tasks and, in more advanced systems, across multiple distinct robot platforms, using a training objective that predicts the correct action tokens given a visual observation and a natural language instruction. Because the underlying vision-language backbone was pretrained on broad, general-purpose data, the resulting model exhibits a notable degree of generalization to novel objects and instructions not explicitly present in the robot demonstration data, for example correctly identifying and manipulating an object described using a natural language phrase that appeared during general vision-language pretraining but never in the robot-specific fine-tuning data.

This case study illustrates the central practical advantage of the foundation model paradigm for robotics: rather than requiring an impractically large amount of task-specific, embodiment-specific robot demonstration data to achieve broad generalization, the approach leverages far more abundant general-purpose vision and language data to bootstrap most of the model's semantic and visual understanding, reserving the comparatively scarce and expensive robot-specific data for teaching the model how to translate that general understanding into physical action, a division of labour directly analogous to the pipeline architectures for combining perception and reasoning discussed in earlier neuro-symbolic AI research and applied here across the boundary between general-purpose perception and embodied physical control.

IX. IMPLEMENTATION CONSIDERATIONS AND TOOLING

A. Pretrained Model Availability and Adaptation

A substantial number of vision, biological, and multimodal foundation models have been released as open-weight pretrained checkpoints, allowing practitioners to adapt these models to downstream tasks using comparatively modest computational resources relative to the cost of pretraining from scratch, typically through parameter-efficient fine-tuning techniques that update only a small fraction of the model's parameters, or through simple linear probing that trains only a lightweight task-specific output layer on top of frozen pretrained representations. Robotics foundation models have generally been released less openly than their vision and language counterparts, in part reflecting the comparatively higher cost and specialized hardware requirements involved in independently reproducing or extending robot-specific pretraining.

B. Computational and Infrastructure Requirements

Pretraining a large foundation model in any modality typically requires substantial specialized computational infrastructure, including large clusters of hardware accelerators and high-throughput data pipelines capable of streaming very large training datasets without becoming a bottleneck to accelerator utilization. This infrastructure requirement has historically concentrated foundation model pretraining capability among a comparatively small number of well-resourced research organizations, though the growing availability of openly released pretrained checkpoints has substantially broadened practical access to foundation model capabilities for downstream adaptation, even among organizations lacking the resources to conduct large-scale pretraining themselves.

C. Deployment and Serving Considerations

Deploying a large foundation model for real-time downstream use, particularly in latency-sensitive settings such as robotics or interactive multimodal assistants, introduces engineering considerations distinct from the pretraining process itself, including model compression, quantization, and distillation techniques used to reduce a large pretrained model's computational footprint sufficiently for deployment on constrained hardware, such as an onboard robot computer or a mobile device, without unacceptably degrading downstream task performance. These deployment-oriented techniques have become an increasingly important complementary research area alongside core pretraining methodology, since a foundation model's practical value ultimately depends on its ability to be served reliably and efficiently within the computational constraints of its intended downstream application.

X. ETHICAL AND SOCIETAL CONSIDERATIONS

Foundation models trained on large, web-scraped or otherwise broadly collected datasets can inherit and amplify biases present in that underlying data, a concern well documented in language models that extends directly to vision and multimodal foundation models, for example through skewed representation of certain demographic groups in web-scraped image datasets, or through biased associations learned between visual attributes and text descriptions. Because foundation models are subsequently adapted to a wide range of downstream applications, biases present in the underlying pretrained representation can propagate into many otherwise unrelated downstream systems, making bias auditing and mitigation at the foundation model level a particularly high-leverage intervention point relative to attempting to separately address bias in each individual downstream application.

In robotics and embodied AI specifically, safety considerations take on additional physical dimensions beyond those typically discussed for purely digital foundation models, since a robotics foundation model's errors can result in direct physical harm to people, property, or the robot itself, motivating substantial ongoing research into safe exploration, robust failure detection, and conservative deployment practices, such as extensive simulated and constrained physical testing prior to broader real-world deployment, that are specific to the embodied setting and are less directly applicable to purely text- or image-based foundation model applications.

XI. CHALLENGES AND OPEN PROBLEMS

A central challenge shared across biological and robotics foundation models is data scarcity relative to the scale of data available for vision and language pretraining, constraining model capacity and increasing the risk of the resulting model overfitting to idiosyncrasies of a comparatively narrow training distribution rather than learning genuinely broad, transferable representations of the kind achieved by the largest vision and language foundation models.

A second challenge concerns evaluation methodology, particularly for robotics and embodied AI, where physical evaluation is expensive, slow, and difficult to standardize across different research groups' hardware and environmental setups, complicating direct comparison between different robotics foundation models in the way that a shared, static benchmark dataset more straightforwardly enables for vision or language models.

A third challenge involves cross-modal alignment quality in multimodal foundation models, where representations that appear well aligned according to coarse benchmark metrics can nonetheless exhibit subtle failures in fine-grained cross-modal reasoning, such as correctly counting or spatially localizing objects described in a complex natural language query.

A fourth and increasingly discussed challenge concerns whether scaling laws observed for language and vision will continue to hold, or hold in modified form, as foundation model research extends further into modalities such as robotics where data scarcity may impose fundamental limits on achievable model scale that are not primarily a matter of insufficient computational investment, but rather a more fundamental constraint imposed by the physical cost of data collection itself.

A fifth challenge, cutting across all modalities discussed in this paper, concerns catastrophic forgetting and knowledge conflict during downstream adaptation: fine-tuning a pretrained foundation model on a narrow downstream task or dataset can degrade the broad general-purpose capabilities acquired during pretraining, sometimes substantially, unless adaptation is performed carefully using techniques such as parameter-efficient fine-tuning or careful learning rate scheduling specifically intended to preserve pretrained knowledge. This tension between specialization and generality is a recurring practical concern for practitioners deploying foundation models in production settings, where a model may need to perform well on a narrow target task while also retaining enough of its original broad competence to handle the inevitable edge cases and out-of-scope inputs that arise in real-world deployment.

XII. FUTURE RESEARCH DIRECTIONS

One prominent direction involves further unification of foundation models across modalities into single architectures capable of jointly processing and generating across text, images, audio, video, and robotic action within one coherent model, extending current multimodal and vision-language-action systems toward increasingly general-purpose multimodal foundation models capable of flexible reasoning and action across an even broader range of input and output modalities than currently supported by most systems.

A second direction focuses specifically on addressing the robotics data bottleneck through large-scale simulation, using increasingly realistic physics simulators combined with domain randomization techniques to generate synthetic robot interaction data at far greater scale and lower cost than physical data collection allows, while ongoing research on sim-to-real transfer aims to ensure that policies trained substantially or primarily in simulation continue to perform reliably when deployed on physical robot hardware.

A further direction concerns developing more rigorous, standardized evaluation methodology for foundation models across all modalities discussed in this paper, extending the kind of standardized benchmark practices well established

in vision and language to biological and robotics foundation models specifically, including explicit measurement of out-of-distribution generalization rather than in-distribution benchmark accuracy alone. Finally, continued research into parameter-efficient adaptation techniques is likely to remain important for broadening practical access to foundation model capabilities, allowing organizations without the resources to conduct large-scale pretraining to nonetheless adapt existing pretrained foundation models to specialized downstream applications using comparatively modest additional computational investment.

An additional promising direction involves foundation models that natively incorporate temporal and causal structure, moving beyond largely static, single-frame vision and language pretraining toward models trained on video and interactive data that captures how scenes and objects evolve and respond to actions over time. Such temporally aware foundation models could provide a more natural foundation for robotics and embodied AI applications specifically, since physical interaction is inherently temporal and consequence-driven in a way that a foundation model pretrained purely on static images or text is not naturally equipped to represent, potentially narrowing the gap in maturity between robotics foundation models and their more established vision and language counterparts discussed throughout this survey.

XIII. CONCLUSION

Foundation models have extended well beyond their original setting in natural language processing to encompass vision, biological sequence and structure data, robotics, and multimodal systems that jointly reason across combinations of these modalities. This survey reviewed the taxonomy of pretraining objectives and architectures used across these modalities, examined applications spanning medical imaging, autonomous robotics, genomics, multimodal assistants, and remote sensing, and presented a worked case study illustrating how a vision-language-action model combines general-purpose multimodal pretraining with comparatively scarce robot-specific data to achieve broad task generalization. While each modality faces distinct challenges, ranging from the severe data scarcity constraining robotics foundation models to the evaluation difficulties inherent in embodied and multimodal settings, the continued cross-pollination of architectural and training techniques across modalities suggests that foundation models will remain a central and unifying paradigm across an increasingly broad range of AI applications, extending well beyond the language domain in which the paradigm was first established.

As this field continues to mature, the modalities surveyed in this paper are likely to converge further rather than developing in isolation, with lessons learned in one modality, such as efficient long-context architectures developed for genomic sequence modeling, informing architectural choices in another, such as long-horizon video understanding, and with shared infrastructure and evaluation practices increasingly spanning what have historically been largely separate research communities organized around vision, language, biology, and robotics. This convergence, more than any single algorithmic breakthrough, may ultimately prove to be the most lasting contribution of the foundation model paradigm to the broader trajectory of artificial intelligence research.

REFERENCES

[1] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al., "On the opportunities and risks of foundation models," arXiv preprint, 2021.
[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. ICLR, 2021.
[3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in Proc. ICML, 2020.
[4] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in Proc. ICCV, 2021.
[5] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in Proc. CVPR, 2022.
[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in Proc. ICML, 2021.
[7] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al., "Flamingo: A visual language model for few-shot learning," in Proc. NeurIPS, 2022.
[8] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," Proceedings of the National Academy of Sciences, vol. 118, 2021.
[9] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, et al., "Evolutionary-scale prediction of atomic-level protein structure with a language model," Science, vol. 379, 2023.
[10] Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri, "DNABERT: Pre-trained bidirectional encoder representations from transformers model for DNA-language in genome," Bioinformatics, vol. 37, 2021.

[11] E. Nguyen, M. Poli, M. Faizi, A. Thomas, C. Birch-Sykes, M. Wornow, A. Patel, C. Rabideau, S. Massaroli, Y. Bengio, S. Ermon, S. A. Baccus, and C. Re, "HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution," in Proc. NeurIPS, 2023.
[12] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, et al., "RT-1: Robotics transformer for real-world control at scale," in Proc. RSS, 2023.
[13] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, et al., "RT-2: Vision-language-action models transfer web knowledge to robotic control," in Proc. CoRL, 2023.
[14] A. Radford et al., "Learning transferable visual models from natural language supervision (CLIP)," in Proc. 38th Int. Conf. Mach. Learn. (ICML), 2021, pp. 8748–8763.
[15] C. Zhou et al., "A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT," arXiv preprint arXiv:2302.09419, 2023.
[16] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li, et al., "Florence: A new foundation model for computer vision," arXiv preprint, 2021.
[17] X. Chen, X. Wang, S. Changpinyo, A. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer, et al., "PaLI: A jointly-scaled multilingual language-image model," in Proc. ICLR, 2023.
[18] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in Proc. NeurIPS, 2023.
[19] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., "DINOv2: Learning robust visual features without supervision," arXiv preprint, 2023.
[20] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," arXiv preprint, 2020.
[21] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," in Proc. CVPR, 2022.
[22] N. Tsimpoukelli, J. Menick, S. Cabi, S. M. A. Eslami, O. Vinyals, and F. Hill, "Multimodal few-shot learning with frozen language models," in Proc. NeurIPS, 2021.
[23] M. Y. Guan, T. Zhu, J. Ba, J. Wu, and D. Fried, "Foundation models for decision making: Problems, methods, and opportunities," arXiv preprint, 2023.
[24] S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. Springenberg, et al., "A generalist agent," Transactions on Machine Learning Research, 2022.
[25] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "RLBench: The robot learning benchmark and learning environment," IEEE Robotics and Automation Letters, vol. 5, 2020.
[26] B. Ichter, A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, E. Jang, et al., "Do as I can, not as I say: Grounding language in robotic affordances," in Proc. CoRL, 2022.
[27] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar, "Foundation models for generalist medical artificial intelligence," Nature, vol. 616, 2023.
[28] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine," Nature Medicine, vol. 28, 2022.
[29] S. Zhang, Y. Xu, N. Usuyama, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, et al., "BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs," arXiv preprint, 2023.
[30] A. Kirillov et al., "Segment anything," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2023, pp. 4015–4026.
[31] M. Reid, Y. Yamada, and S. S. Gu, "Can Wikipedia help offline reinforcement learning?," arXiv preprint, 2022.
[32] K. Fang, D. Xu, K. Hausman, T. Darrell, and V. Mnih, "Adaptive procedural task generation for hard-exploration problems," in Proc. ICLR, 2021.
[33] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in Proc. ICCV, 2023.
[34] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "ImageBind: One embedding space to bind them all," in Proc. CVPR, 2023.
[35] T. Xiao, H. Chan, P. Sermanet, A. Wahid, A. Brohan, K. Hausman, S. Levine, and J. Tompson, "Robotic skill acquisition via instruction augmentation with vision-language models," in Proc. RSS, 2023.
[36] C. Lu, H. Lu, K. Cobbe, J. Choi, J. Whiteson, and S. Levine, "Structured world models from human videos," in Proc. RSS, 2023.
[37] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, et al., "PaLM-E: An embodied multimodal language model," in Proc. ICML, 2023.
[38] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "BC-Z: Zero-shot task generalization with robotic imitation learning," in Proc. CoRL, 2021.
[39] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, et al., "E go4D: Around the world in 3,000 hours of egocentric video," in Proc. CVPR, 2022.
[40] T. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith, and L. Schmidt, "Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time," in Proc. ICML, 2022.
[41] L. Beyer, X. Zhai, A. Royer, L. Markeeva, R. Anil, and A. Kolesnikov, "Knowledge distillation: A good teacher is patient and consistent," in Proc. CVPR, 2022.
[42] R. Bommasani, K. Klyman, S. Longpre, S. Kapoor, N. Maslej, B. Xiong, D. Zhang, and P. Liang, "The foundation model transparency index," arXiv preprint, 2023.
[43] I. D. Raji, E. Denton, E. M. Bender, A. Hanna, and A. Paullada, "AI and the everything in the whole wide world benchmark," in Proc. NeurIPS Datasets and Benchmarks Track, 2021.
[44] A. Birhane, V. U. Prabhu, and E. Kahembwe, "Multimodal datasets: Misogyny, pornography, and malignant stereotypes," arXiv preprint, 2021.
[45] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," Journal of Machine Learning Research, vol. 17, 2016.