Are Natural Domain Foundation Models Useful for Medical Image Classification?

Joana Palés Huix [1, 2, 3 * †]   Adithya Raju Ganeshan [1, 2, †]   Johan Fredin Haslum [1, 2, 3]

Magnus Söderberg [3]   Christos Matsoukas [1, 2, 3]   Kevin Smith [1, 2]

[1] KTH Royal Institute of Technology, Stockholm, Sweden [2] Science for Life Laboratory,
Stockholm, Sweden [3] AstraZeneca, Gothenburg, Sweden

Abstract

The deep learning field is converging towards the use of general foundation models that can be easily adapted for diverse tasks. While this paradigm shift has become common practice within the field of natural language processing, progress has been slower in computer vision. In this paper we attempt to address this issue by investigating the transferability of various state-of-the-art foundation models to medical image classification tasks. Specifically, we evaluate the performance of five foundation models, namely SAM, SEEM, DINOv2, BLIP, *and* OPENCLIP *across four well-established medical imaging datasets. We explore different training settings to fully harness the potential of these models. Our study shows mixed results.* DINOv2 *consistently outperforms the standard practice of* IMAGENET *pretraining. However, other foundation models failed to consistently beat this established baseline indicating limitations in their transferability to medical image classification tasks.*

1. Introduction

Recently there has been a surge of interest in the use of foundation models trained on large-scale datasets within the field of computer vision. This has resulted in a growing trend of adapting these models for a wide range of downstream applications with minimal effort.

Foundation models, characterized by their substantial size and self-supervised training on diverse datasets, possess remarkable capabilities for generating meaningful representations across multiple domains [1]. These models offer significant advantages as parameter initialization strategies for a wide range of downstream tasks. While the concept of large-scale pretraining originated in the field of Natural Language Processing (NLP) with BERT [9], subse-
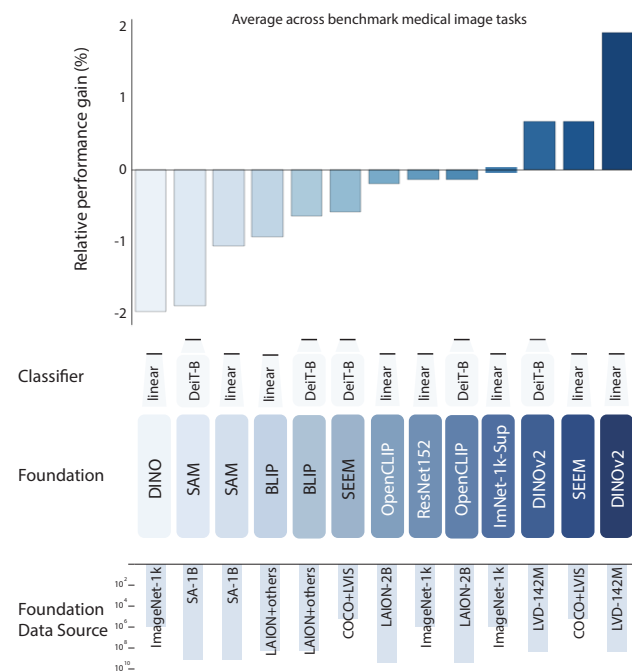

Figure 1. Foundation models applied to benchmark medical image classification tasks. We report different combinations of foundation models (linear heads or Vision Transformers (VIT) with linear heads) and classifiers for relative gain in performance for each setting. We also provide the size of the source data on which the foundation models were trained, ranging from LVD+COCO (2×10^5) to LAION-2B (2×10^9). The baseline consists of a VIT foundation pretrained on IMAGENET-1k in a supervised manner and fine-tuned for the medical task. While some combinations perform below the baseline, DINOv2 consistently outperforms standard IMAGENET pretraining.

quent advancements like BART [18], RoBERTa [21], GPT [2, 30, 31], have further propelled the utilization of task-agnostic models. This paradigm shift has not only driven research on adaptation methods but has also fueled a growing interest in comprehending the inner workings of foun-

---

*Corresponding author: Joana Palés Huix <joanaph@kth.se>

†These authors contributed equally to this work.

dation models [1, 43].

The field of computer vision is arguably currently undergoing a similar transformation. Until recently, computer vision relied heavily on models pretrained on IMAGENET [7] using supervised learning. However, recent advancements have led to the emergence of alternative computer vision foundation models using larger datasets and self-supervision in a move that diverges from IMAGENET as the primary source of pretraining data. These models can be broadly categorized into two types: feature encoding models trained on pretext tasks, such as DINO [3, 27], CLIP [29] and BLIP [19], and models designed specifically to tackle specific tasks such as segmentation in the case of SAM [16] and SEEM [45]. Notably, exceptional zero-shot capability and strong generalization across a wide range of tasks is reported for each of these models.

Due to constraints related to privacy and ethical considerations, tasks in medical imaging suffer from a shortage of data. As such, it is an area that can greatly benefit from transfer learning [24, 25]. It follows that the adoption of foundation models for medical tasks is expected to provide significant advantages as well. However, domain shifts between the data that foundation models are pretrained on and the target medical tasks poses a considerable challenge. Thus, there is a need to test the effectiveness of these models specifically in this domain. Although some attempts have been made to adapt foundation models to the medical domain (*e.g.* adapting SAM for medical segmentation) [8, 23, 28, 34, 37, 38, 40] to the best of our knowledge, a comprehensive evaluation of cutting-edge foundation models for medical image classification is lacking.

In this work, we aim to bridge this gap by conducting a thorough evaluation and analysis of state-of-the-art foundation models in the context of medical image classification, shedding light on their applicability and performance within the medical domain.

The findings of our study are as follows:

- Not all foundations transfer well to the medical domain. Some fail to beat the IMAGENET-1k baseline. DINOv2 shows performance gains, SEEM matches while SAM, BLIP and OPENCLIP lag behind.

- Foundation models require adaptation to downstream tasks. Wherein low-level features demonstrate transferability to medical tasks, while the later layers undergo task-specific adaptation upon unfreezing of the foundation model.

- Appending a DEIT classifier to the foundation models results in marginal performance gains. The obtained marginal gains suggest that the high-level features may not effectively serve as inputs to the classifier.

The code to reproduce our experiments can be found at https://github.com/joanaapa/Foundation-Medical.

2. Related Work

Foundation models have sparked tremendous excitement in the community. In the domain of segmentation tasks, two foundation models have recently gained attention. The first model, SAM [16], incorporates a VIT-based encoder and a lightweight transformer-based decoder, showcasing remarkable zero-shot segmentation capabilities. SAM has gained considerable attention in the medical field, despite being a relatively recent publication. Several studies have conducted assessments of SAM's performance in medical imaging tasks *out-of the-box* [4, 8, 12, 26, 34, 44]. Furthermore, researchers have explored slight modifications of the SAM model to tailor it specifically for medical imaging applications [28, 38] or fine-tuned it for specific datasets [23].

SEEM [45], another foundation model for segmentation demonstrates excellent performance on numerous benchmarks. However, at present, there is a dearth of literature exploring SEEM's application in the medical domain. Despite conducting an exhaustive search, we found no relevant references regarding its usage within the medical domain, making it a compelling candidate for our study.

DINOv2, introduced by Oquab et al. [27], is a VIT-based model that leverages a robust unsupervised pretraining method. Building upon DINO [3], DINOv2 benefits from training on a large-scale curated dataset, resulting in representations that capture the semantic meaning of images remarkably well. This has allowed DINOv2 to demonstrate effective performance across various tasks, including pixel-level and image-level classification.

The CLIP family of models is pretrained on image-text pairs [29]. The combination of language and vision in its representations equips CLIP with versatile multimodal analysis capabilities, which prove to be powerful in various domains. The adaptability of CLIP to the medical domain was demonstrated by Wang et al. [37]. Their study showcased the superior adaptability of CLIP, surpassing state-of-the-art models specific to radiology. Noteworthy studies by Tiu et al. [35], Zhang et al. [42], and Huang [13] explore similar approaches as CLIP, showcasing its outstanding zero-shot capabilities. In comparison to its counterparts initialized with ImageNet weights, CLIP consistently outperforms strong baselines with remarkable data efficiency. BLIP [19] unifies vision-language understanding and generation by employing uni-modal encoders, image-grounded text encoders, and image-grounded text decoders, catering to various image-text tasks. Pre-training involves optimizing Image-Text Contrastive, Image-Text Matching, and Language Modeling objectives. The majority of the discussed studies utilize CNN-based vision backbones initialized with pre-trained weights from IMAGENET. Some also incorporate in-domain pretraining on top of IMAGENET initialization.

Table 1. **Frozen** foundation models applied to medical tasks.

Classifier (Unfrozen)	Foundation (Frozen)	APTOS2019, κ ↑ $n = 3{,}662$	DDSM, AUC ↑ $n = 10{,}239$	ISIC2019, Rec. ↑ $n = 25{,}333$	CheXpert, AUC ↑ $n = 224{,}316$
	DINOv1	0.878 ± 0.004	0.912 ± 0.002	0.588 ± 0.015	0.738 ± 0.001
	IMNET-1K-SUP	0.864 ± 0.007	0.908 ± 0.001	0.513 ± 0.015	0.719 ± 0.001
	RESNET152	0.824 ± 0.003	0.883 ± 0.001	0.461 ± 0.027	0.712 ± 0.000
Linear	SAM	0.873 ± 0.010	0.916 ± 0.004	0.402 ± 0.014	0.724 ± 0.001
	SEEM	0.852 ± 0.003	0.891 ± 0.001	0.527 ± 0.007	0.731 ± 0.001
	OPENCLIP	0.857 ± 0.003	0.897 ± 0.002	0.489 ± 0.010	0.702 ± 0.001
	DINOv2	0.881 ± 0.002	0.905 ± 0.001	0.569 ± 0.012	0.722 ± 0.000
	BLIP	0.818 ± 0.014	0.845 ± 0.005	0.416 ± 0.021	0.647 ± 0.000
DEIT-B	SAM	0.890 ± 0.005	0.940 ± 0.007	0.740 ± 0.015	0.788 ± 0.003
	SEEM	0.887 ± 0.002	0.925 ± 0.006	0.747 ± 0.028	0.777 ± 0.001
	OPENCLIP	0.903 ± 0.006	0.948 ± 0.004	0.748 ± 0.010	0.790 ± 0.001
	DINOv2	0.901 ± 0.005	0.945 ± 0.008	0.790 ± 0.010	0.798 ± 0.001
	BLIP	0.895 ± 0.009	0.945 ± 0.004	0.763 ± 0.011	0.785 ± 0.000

Table 2. **Unfrozen** foundation models applied to medical tasks.

Classifier (Unfrozen)	Foundation (Unfrozen)	APTOS2019, κ ↑ $n = 3{,}662$	DDSM, AUC ↑ $n = 10{,}239$	ISIC2019, Rec. ↑ $n = 25{,}333$	CheXpert, AUC ↑ $n = 224{,}316$
	DINOv1	0.886 ± 0.004	0.958 ± 0.004	0.775 ± 0.007	0.794 ± 0.001
	IMNET-1K-SUP	0.904 ± 0.006	0.960 ± 0.003	0.823 ± 0.008	0.797 ± 0.002
	RESNET152	0.899 ± 0.002	0.960 ± 0.003	0.817 ± 0.007	0.807 ± 0.000
Linear	SAM	0.894 ± 0.007	0.950 ± 0.006	0.798 ± 0.010	0.801 ± 0.001
	SEEM	0.902 ± 0.003	0.958 ± 0.003	0.835 ± 0.009	0.808 ± 0.001
	OPENCLIP	0.903 ± 0.013	0.945 ± 0.012	0.818 ± 0.008	0.806 ± 0.001
	DINOv2	$\mathbf{0.909} \pm 0.009$	$\mathbf{0.966} \pm 0.003$	$\mathbf{0.859} \pm 0.007$	$\mathbf{0.812} \pm 0.001$
	BLIP	0.891 ± 0.006	0.961 ± 0.003	0.798 ± 0.016	0.797 ± 0.000
DEIT-B	SAM	0.890 ± 0.005	0.951 ± 0.004	0.778 ± 0.013	0.796 ± 0.002
	SEEM	0.901 ± 0.007	0.946 ± 0.002	0.808 ± 0.022	0.802 ± 0.001
	OPENCLIP	0.907 ± 0.011	0.955 ± 0.010	0.809 ± 0.013	0.805 ± 0.001
	DINOv2	0.904 ± 0.007	$\mathbf{0.966} \pm 0.004$	0.822 ± 0.011	$\mathbf{0.812} \pm 0.001$
	BLIP	0.897 ± 0.001	0.955 ± 0.002	0.806 ± 0.016	0.799 ± 0.000

3. Methods

The primary objective of this study is to evaluate the efficacy of vision foundation models in the context of medical image classification. Specifically, we aim to assess the transferability of features generated by these models to the medical domain and identify the best utilization strategies. The foundation models we consider were trained with different data and different objectives, so it stands to reason that some models will be better suited to transfer features for medical tasks than others. We conducted a series of experiments employing two distinct approaches: using the foundation model as a standalone model (with a linear prediction head) and stacking it with a model referred to as the "Classifier model". For consistency, DEIT [36], a member of the transformer family, was used as the classifier.

3.1. Datasets

To ensure a comprehensive assessment of feature transferability to the medical domain, we carefully chose four publicly available and well-established datasets. These, encompass a wide range of imaging modalities, color scales, and dataset sizes, providing diversity in the evaluation scenarios. For each dataset, we employed specific evaluation metrics tailored to the nature of the classification task, enabling a detailed analysis of model performance.

- **APTOS 2019** [15] - The dataset comprises a collection of 3,662 high-resolution images related to diabetic retinopathy, categorized into five severity classes. The evaluation of model performance on this dataset is measured using the quadratic Cohen kappa metric.

- **CBIS-DDSM** [17] - The dataset used in this study consists of 10,239 mammography images for the detection of anomaly masses. The model evaluation on this dataset was conducted using the ROC-AUC metric.

- **ISIC 2019** [5] - This dataset comprises a set of 25,331 dermoscopic images. These images were used for the classification of skin lesions into 9 distinct diagnostic categories. The evaluation of model performance on this dataset was assessed using the recall metric.

- **CHEXPERT** [14] - This dataset contains a series of 224,316 chest X-rays containing 14 different diagnostic observations, encompassing a range of conditions and abnormalities. The model performance was evaluated using the ROC-AUC metric.

3.2. Foundation models

The foundation models chosen for this study are all based on the transformer architecture. In order to maintain fairness in our comparisons, we specifically selected versions of these models that were as closely matched in size as feasible. This approach ensures that any performance differences observed can be attributed to the specific architectural variances and not to significant differences in model size or complexity. The foundation models we consider in this study are:

- **SAM** (88.8M parameters) [16] - SAM is an encoder-decoder architecture specifically designed for promptable segmentation tasks. It has been trained on a dataset consisting of 11 million high-resolution images along with 1 billion associated masks. We utilized the image encoder of SAM as the foundation model. When fine-tuning SAM, we use a resolution of 224 which deviates from the original training resolution

- **SEEM** (29.9M parameters) [45] - SEEM employs a generic encoder-decoder architecture comprised of separate encoders for vision and text, followed by a decoder for mask generation which works on the joint embedding space of vision and text. A combination of COCO2017 [20] and REF-COCO dataset [41] were utilized for training. Our feature extractor is the Focal transformer [39] vision backbone.

- **DINOv2** (86.5M parameters) [27] - This all-purpose VIT model is pretrained using a discriminative self-supervised method. The pretraining dataset (142 mil-
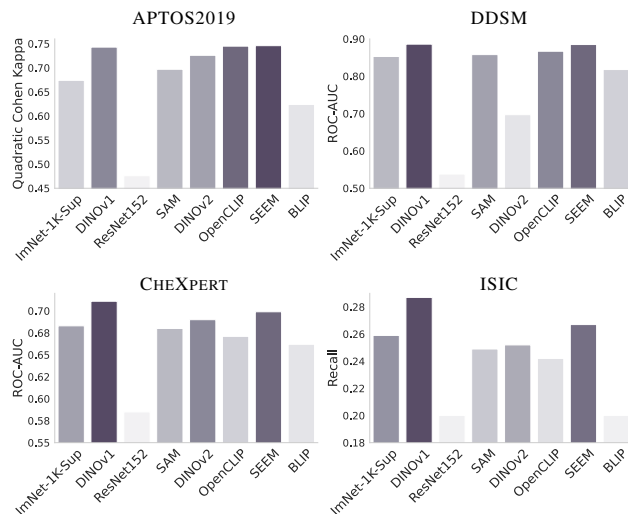
Figure 2. k-NN evaluation performance of different foundation and baseline models.

Figure 3. Training and inference times.

lion images) is a curation of 24 publicly available natural image datasets for classification, retrieval and segmentation tasks. We use the VIT-B model, which is a distilled version of their VIT-H architecture.

- **OPENCLIP** (86.2M parameters) [29] - A versatile VIT model, employing customized text-guided pre-training on a subset of the LAION-5B dataset [33]. The subset LAION-2B-EN contains 2 billion image-caption(in english) pairs obtained through web scraping. For our experiments, we utilize the base version of this model.

- **BLIP** (85.7M parameters) [19] - Utilizing vision transformers and BERT-based encoders, BLIP integrates visual and textual information through cross-attention and specialized tokens. Trained on 142 million images curated from datasets ranging from LAION-400M to COCO2017. Our feature extractor is the VIT-B vision backbone.

In addition to these foundation models, we consider as a baselines models trained on IMAGENET-1k (ILSVRC-2012) [32]

- **IMNET-1K-SUP** (86.6M parameters) - The standard practice for transfer learning to medical classification tasks is to use IMAGENET-1k pretrained networks [25] for initialization. In this work we use DEIT-B [36] pretrained in a supervised manner on IMAGENET-1k as our baseline.

- **DINOV1** (86.6M parameters) [3] - We also include in our comparisons this general-purpose VIT model, which is pretrained using an older self-supervised method compared to DINOV2, but on IMAGENET-1k
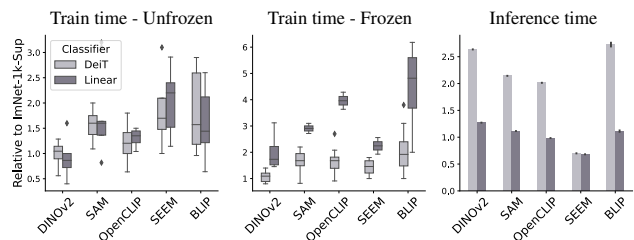
[32]. In this work we use the base model, which is based on the DEIT-B architecture.

- **RESNET** (60.4M parameters) [11] - We have also included this general purpose CNN backbone, which is pretrained on IMAGENET-1k. We use RESNET152 to match the complexity of other baselines.

3.3. Implementation Details

Prior to training, all images were downsampled to a resolution of 256×256, followed by a center crop to obtain a final size of 224×224. A set of standard augmentations was applied to the training images. These augmentations included normalization, vertical and horizontal flips, color jitter (adjusting brightness, contrast, saturation, and hue), and random resize crops. Each training set was divided into train (80%), validation (10%) and test (10%). In the case of APTOS, which is of a smaller size, the split was adjusted to 70%, 15% and 15%, respectively. Some foundation models like SAM were trained with different resolutions. To handle this, we use bicubic interpolation to align the position embeddings with the image scales and ensure compatibility [10].

In addition to the training resolution, the foundation model's architectures exhibit significant differences that necessitate adaptation for the medical tasks at hand. Specifically, when considering the SEEM and SAM adaptations of vision transformer architectures, the absence of a *[cls]* token requires us to employ global average-pooling of the patch representations obtained from the final layer in order to attach a linear head or perform a k-NN evaluation. In case of OPENCLIP, BLIP and DINOV2, we can make use of the original *[cls]* token without modification. In all cases, integrating a classifier model on top of the foundation model involves extracting patch representations from the foundation models, adding a projection layer, bypassing the patch embedding layer of the stacked classifier, and transmitting the patch representations to the classifier for the fine-tuning process. The appended Classifier's *[cls]* token is used for the classification task.

The models were fine-tuned using supervised learning on each medical dataset with the AdamW optimizer [22], employing a linear learning rate warmup strategy. To determine the optimal base learning rate, a search was conducted
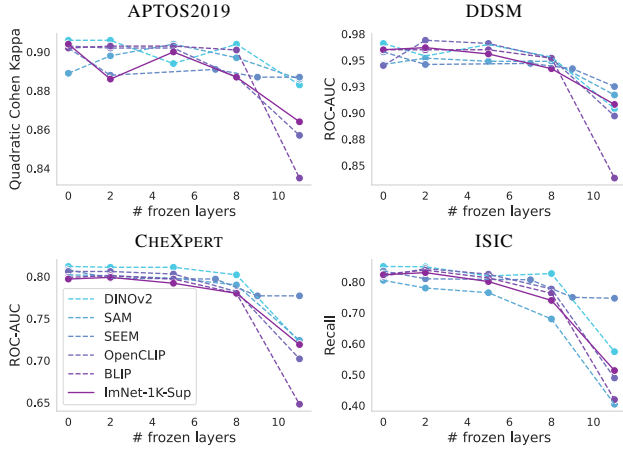
Figure 4. Impact of gradually freezing layers of the foundation model on performance for medical tasks (using a linear head).

Figure 5. Model performance when using different subsets of foundation layers in conjunction with a DEIT classifier.

across four values ranging from 10^{-5} to 10^{-3}. As the saturation point was reached, the learning rate was decreased by a factor of 10. A weight decay of 10^{-5} was applied, and this entire process was repeated five times to account for performance variance. The evaluation of the models was based on selecting the best-performing checkpoint, identified through analysis on the validation set.

3.4. Evaluation protocol

We employed a two-step approach to assess the performance of the foundation models. Initially, we kept the foundation models frozen and performed transfer learning by fine-tuning a linear classifier head on top. This enabled us to evaluate the adapted model's performance in comparison to the baseline (IMNET-1K-SUP). We conducted this evaluation across multiple medical datasets to gauge the model's efficacy. Subsequently, we appended a transformer classifier (DEIT-B) on top of the frozen foundation. We fine-tuned this combined architecture on the medical tasks and assessed its performance. To gain further insights, we repeated the aforementioned procedure while unfreezing the weights of the foundation model. This allowed the foundation model to adapt its features during fine-tuning, both with the linear head and with the classifier. To summarize, we considered four scenarios:

1. **Frozen with linear head.** Transfer learning to the medical tasks is done while keeping the foundation model frozen, only a linear head is fine-tuned to the task.

2. **Frozen with appended classifier.** Transfer learning to the medical tasks is done while keeping the foundation model frozen, a DEIT-B classifier is appended and fine-tuned for the task.

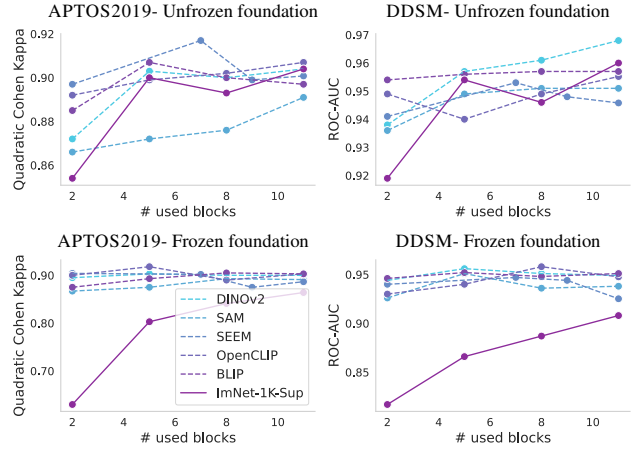3. **Unfrozen with linear head.** Both the foundation

model and a linear head are fine-tuned to the task.

4. **Unfrozen with appended classifier.** Both the foundation model and an appended DEIT-B classifier are fine-tuned for the task.

In the above setups we evaluated the final performance of the fine-tuned models on the medical tasks. We also consider the generalization capability and robustness to class separability of the features, which can be assessed using a k-Nearest Neighbour evaluation on the features extracted by the foundation model.

5. k-**NN evaluation.** We examine the applicability of the foundation models features for the medical tasks using k-Nearest Neighbors.

Furthermore we evaluate the impact of model complexity and resolution on the performance as well as training and inference times

6. **Model size.** We select the top-performing model and evaluation protocol, then examine various available models of differing sizes.

7. **Image resolution.** We upscale training resolution to 1024×1024 and evaluate using the best-identified scenario.

8. **Train and inference times.** We define training time as the number of iterations until the validation metrics reaches maximum performance (normalized across datasets). Inference time is calculated as the average time to classify a single instance. Both are reported relative to the baseline IMNET-1K-SUP.

Finally, to investigate how internal representations within the foundation models adapt to the medical tasks, we employ Centered Kernal Alignment (CKA) to compare models before and after fine-tuning.
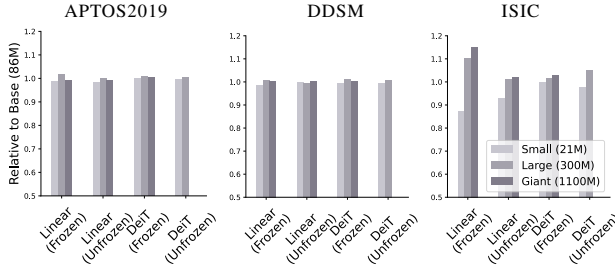
Figure 6. Performance of DINOv2 with different model complexities relative to the Base model.

9. **Centered Kernel Alignment (CKA).** We employ Centered Kernel Alignment (CKA) [6] to analyze the adaptation of internal representations within the foundation models to the medical tasks. CKA is a metric that measures the similarity between layers in different neural networks. This allows us to gain insights into the specific layers that undergo changes during the fine-tuning process. To visualize and quantify these changes, we compute a CKA heat map comparing the pre- and post-fine-tuned versions of the foundation models.

4. Experiments

In this section, we present the findings from the experiments outlined in Sections 3.2, 3.3, and 3.4.

4.1. Foundational Features for Medical Tasks

To measure the direct transferability of the features produced by the foundation models for medical tasks, we froze the foundation model and appended a linear head (1) or a classifier (2) as described in Section 3.4. The results of this experiment are provided in Table 1. When compared to the unfrozen fine-tuning, this approach was generally inferior.

Frozen and unfrozen foundations. We also measure the ability of foundation models to adapt to medical tasks by unfreezing their weights during fine-tuning, both with a linear head (3) and an appended classifier (4) as described in Section 3.4. The results of this experiment are provided in Table 2. Here, it can be seen that DINOv2 consistently provides significant boosts in performance when transferred to medical tasks. The results with linear classifiers in Table 2 demonstrate that DINOv2 outperforms the baseline across all datasets, yielding performance gains of up to 3.2%. The performance achieved on the APTOS2019 and ISIC highlights the potential for significant improvements even when dealing with small datasets. In fact, DINOv2 seems to outperform even self-supervised RESNETs, pretrained in-domain, as it can be seen in Appendix B.

k-**NN evaluation.** In Figure 2 we examine the adaptability of the features from the foundation models using k-Nearest Neighbors. Without any fine-tuning or supervision, this

Table 3. **Foundation models** trained at 1024×1024 resolution. Percentage gain relative to low resolution indicated in brackets.

Foundation (UnFrozen)	APTOS2019, $\kappa \uparrow$ $n = 3{,}662$	DDSM, AUC $\uparrow$ $n = 10{,}239$	ISIC2019, Rec. $\uparrow$ $n = 25{,}333$	CheXpert, AUC $\uparrow$ $n = 224{,}316$
SAM	0.908 (1.4)	0.984 (3.4)	0.804 (0.6)	0.820 (1.9)
SEEM	0.911 (0.9)	0.967 (0.2)	0.868 (3.3)	0.821 (1.3)
OPENCLIP	0.904 (0.1)	0.980 (3.5)	0.852 (3.4)	0.813 (0.7)
DINOv2	0.917 (0.8)	0.983 (1.7)	0.895 (3.6)	0.825 (1.3)
BLIP	0.909 (1.2)	0.984 (2.9)	0.839 (3.3)	0.813 (1.4)

tests how well the foundation model can separate the classes in the task at hand. The results are largely in line with the findings of Tables 1 and 2, but surprisingly DINOv1 performs the best overall, despite its poor performance when fine-tuned for the task. Another surprising observation is the consistently poor performance of IMNET-1K-SUP and RESNET. Their raw features perform poorly measured by k-NN but they adapt to the task through fine-tuning more readily than SAM and OPENCLIP.

Progressive freezing. It is known that different layers capture different image characteristics, starting from low level features in the first layers to high level at the last ones. Above, we only considered two scenarios, which were to use the features that the foundation provided out-of-the box or completely adapting the foundation during training. Here, we investigate the suitability of the features produced by each layer of the foundation models. Figure 4 illustrates the performance of foundation models when trained with a linear classifier by progressively freezing blocks. A consistent trend can be observed across almost all cases, including the baseline. Freezing up to the eighth block leads to a marginal decrease in performance, but beyond that point, there is a significant decline, which is particularly notable for ISIC. This may indicate a point where high-level foundational features become too specialized to the pretraining task, and unsuited for medical tasks. The only exception is SEEM not exhibiting this significant drop in performance.

Decapitating the foundation. Expanding on the concept that deeper layers tend to capture high-level features which may be less relevant for transfer learning, we explored what happens if we use only a part of the foundation as a feature extractor and propagated the tokens to a DEIT classifier. We decapitate the foundation model and append a DeiT-B classifier, then fine-tune. We consider two cases: where the foundation is either kept frozen, or fine-tuned along with the classifier. Better performance is observed when the foundation is fine-tuned along with the classifier. In this case, a clear trend is observed, indicating that incorporating more layers leads to improved performance (Figure 5 top). On the other hand, when the foundation model is used solely as a feature extractor without adaptation (Figure 5 bottom), it appears that relying on features from the early layers yields benefits, as they tend to capture more low-level (and more general) image information. The optimal point varies ac-
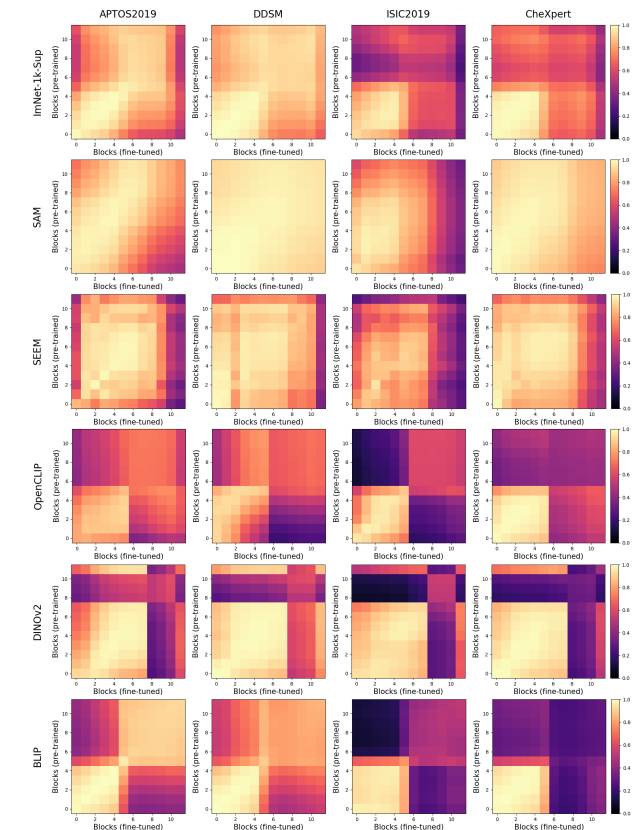
Figure 7. Layerwise CKA heatmaps pre-trained and fine-tuned foundation models using a linear classifier. For reference, the baseline IMNET-1K-SUP is included.

Figure 8. CKA of the last layer among the different foundation models.

cording to the dataset and model, with block 5 performing well for DDSM and different optimal points for AP-TOS2019 depending on the model.

Training and inference time. We also evaluate the training and inference times of the different models, as shown in Figure 3. Training a linear classifier on a frozen foundation took considerably longer compared to other configurations. Among the different foundation models, DINOv2 showed similar training times to the baseline IMNET-1K-SUP, while the others experienced increased training times, with SEEM and BLIP being the slowest when unfrozen. As for the inference times, adding a full DEIT classifier on the foundation results in a notable increase in inference time. However, SEEM is the fastest model in terms of inference as expected due to it's progressive reduction in number of spatial tokens throughout the network.

Model Complexity and Image resolution. Finally we analyze the top-performing model's performance by adjusting its complexity in Figure 6. For ISIC, using larger models as feature extractors improves performance, suggesting that features are more well-adapted with increased size. These benefits are not so clear for APTOS2019 and DDSM, pos-
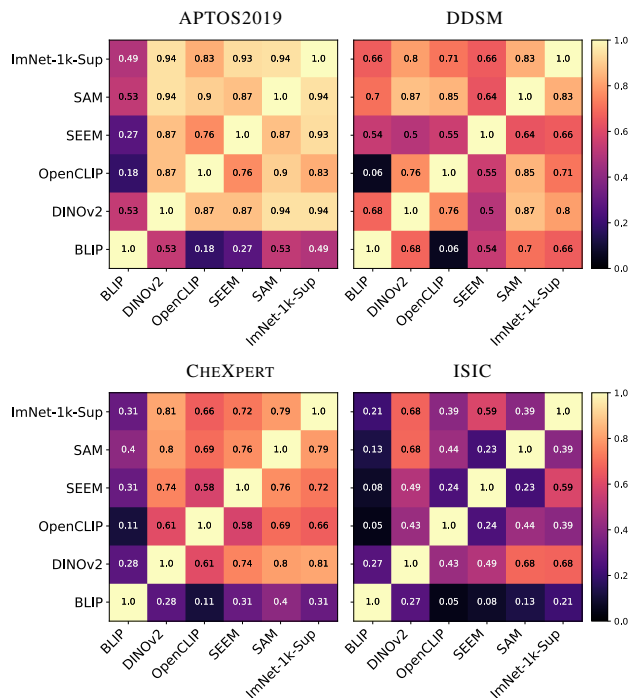
sibly due to the smaller size and/or larger domain shift. Resolution experiments in Table 3 show a universal improvement in performance when high-resolution images are used, with DINOv2 benefiting the most overall.

4.2. Internal Representation Adaptation

After considering the results in Section 4.1, an intriguing question arises: *why does stacking a complete transformer-based classifier on top of the foundation yield only marginal improvements compared to a linear head?* To address this question, we investigate how the learned representations are shared and adapted when propagated from the foundation model to the classifier network, especially considering the variations in dataset sizes and domains. To do this, we use the CKA similarity metric (9) as described in Section 3.4. This similarity analysis was performed under two distinct scenarios:

- **Figure 7** – Between the pre-trained foundation model and the corresponding model after fine-tuning with the addition of a linear layer.

- **Figure 8** – Between the last layer representation among different foundation models.

In Figure 7, we observed a trend in the adaptation of foundation models that is similar to IMNET-1K-SUP. Specifically, a significant transformation was observed in a considerable number of upper layers of the fine-tuned model. This transformation is dataset-specific, with the

ISIC dataset requiring a higher degree of adaptation. Interestingly, the low degree of change observed in the early layers suggests that foundation models possess exceptionally rich representations, requiring minimal adaptation. Similar analysis (Appendix Figure A.1) indicates that a high degree of adaptation is required from the stacked classifier in this unfrozen-foundation setting.

In Figure 8, we can observe how the features available in the last layer differ between foundation models for different datasets. Overall, ISIC leads to the highest difference among models, in line with our previous observation. Interestingly, BLIP remains to be the only model to learn very dissimilar features in comparison to the rest. Moreover, BLIP and OPENCLIP trained for similar tasks do not exhibit any feature similarity.

5. Discussion

Having considered both all-purpose foundation models and segmentation foundation models, our investigation indicates that DINOv2 serves as a solid base for transfer learning to medical tasks, requiring only a standard fine-tuning for optimal performance. Other models fail to outperform the baselines consistently – with the strategy of freezing the foundation model clearly failing. Interestingly, the compact SEEM model competes effectively with its larger counterparts.

We speculate on why foundation models excel in natural domain images but under perform in our tests. The architectural differences between SAM and SEEM, which are both pre-trained for segmentation, might lead to varied feature hierarchies. While both OPENCLIP and BLIP use text-guided pretraining, their objectives differ, potentially affecting representation quality with image-only input. Notably, DINOv2, without extra decoders or text conditioning, outperforms. This suggests that pretraining methods and architecture significantly influence transfer learning success.

We obtained the best performance when the foundation is fine-tuned, and show that employing a linear head generally yields equal or improved performance compared to using an appended classifier. When the foundation is frozen, stacking a classifier is necessary to get the best performance, but it is often inferior to the current standard practice of IMAGENET pretraining (IMNET-1K-SUP). Despite poor k-NN performance, pre-trained CNN's match their counterparts in a supervised setting.

Within the foundation model, our analysis shows that early layers see a substantial degree of feature reuse while adaptation of the deeper layers to the target task is critical – if the later layers are not adapted to the task they may impede performance.

In addition, we note the following key findings:

- Foundation models exhibit rich features that can be reused or adapted for medical image classification tasks, as demonstrated by the k-NN evaluation protocol. This is likely influenced by the large collection of natural images foundation models are trained on.

- Freezing the foundation models often leads to a drop in performance, indicating that the high-level features generated by the foundation models are not well-suited for medical tasks. The feature similarity analysis supports this observation, showing the emergence of novel high-level features in the last layers of unfrozen foundation models during fine-tuning.

- Compared to the other foundation models, fine-tuning DINOv2 demonstrated a notably faster convergence to achieve high levels of performance. This phenomenon can be plausibly attributed to fewer layers requiring adaptations which can be observed from the feature similarity maps.

- The correlation between the foundation model's training data and its transfer ability is unclear. However, a trend indicates improved performance with larger fine tuning datasets like Chexpert.

- The role of the *[cls]* token is unclear. While DINOv2 did perform best overall, the lack of a *[cls]* token did not prevent SEEM from performing nearly as well.

- High resolution fine tuning leads to improved performance in both, models pretrained at high and low resolution. Additionally, the base model tends to strikes the optimal balance between performance and model complexity.

6. Conclusions

In summary, the main finding of this study is that modern foundation models, notably DINOv2, serve as a solid base for transfer learning to medical tasks with minimal fine-tuning. The specific pretraining scheme and architecture are crucial factors affecting performance, but already today we have foundation models that dethrone IMAGENET pretraining for medical classification tasks. Looking to the future, as newer and larger foundation models continue to emerge, and with increasing demand for AI in medical applications, further research and exploration are needed to enhance the adaptability and effectiveness of foundation models in the medical domain. These advancements hold the potential to significantly improve medical image analysis and contribute to improved healthcare outcomes.

References

[1] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 1, 2

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 4, 12

[4] Dongjie Cheng, Ziyuan Qin, Zekun Jiang, Shaoting Zhang, Qicheng Lao, and Kang Li. Sam on medical images: A comprehensive study on three prompt modes. *arXiv preprint arXiv:2305.00035*, 2023. 2

[5] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018. 3

[6] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment, 2014. 6

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[8] Ruining Deng, Can Cui, Quan Liu, Tianyuan Yao, Lucas W Remedios, Shunxing Bao, Bennett A Landman, Lee E Wheless, Lori A Coburn, Keith T Wilson, et al. Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging. *arXiv preprint arXiv:2304.04155*, 2023. 2

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 4

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 4

[12] Sheng He, Rina Bao, Jingpeng Li, Jeffrey Stout, Atle Bjornerud, P. Ellen Grant, and Yangming Ou. Computer-vision benchmark segment-anything model (sam) in medical images: Accuracy in 12 datasets, 2023. 2

[13] Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3922–3931, 2021. 2

[14] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019. 3

[15] Sohier Dane Karthik, Maggie. Aptos 2019 blindness detection, 2019. 3

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 2, 3

[17] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4(1):1–9, 2017. 3

[18] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *ArXiv*, 2019. 1

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 2, 4

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 3

[21] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 1

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 4

[23] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023. 2

[24] Christos Matsoukas, Johan Fredin Haslum, Magnus Söderberg, and Kevin Smith. Pretrained vits yield versatile representations for medical images. *arXiv preprint arXiv:2303.07034*, 2023. 2, 12

[25] Christos Matsoukas, Johan Fredin Haslum, Moein Sorkhei, Magnus Söderberg, and Kevin Smith. What makes transfer learning work for medical images: feature reuse & other factors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9225–9234, 2022. 2, 4

[26] Sovesh Mohapatra, Advait Gosai, and Gottfried Schlaug. Brain extraction comparing segment anything model (sam) and fsl brain extraction tool. *arXiv preprint arXiv:2304.04738*, 2023. 2

[27] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 3

[28] Zhongxi Qiu, Yan Hu, Heng Li, and Jiang Liu. Learnable ophthalmology sam. *arXiv preprint arXiv:2304.13425*, 2023. 2

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 4

[30] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *ArXiv*, 2018. 1

[31] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 4

[33] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. 4

[34] Peilun Shi, Jianing Qiu, Sai Mu Dalike Abaxi, Hao Wei, Frank P-W Lo, and Wu Yuan. Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation. *Diagnostics*, 13(11):1947, 2023. 2

[35] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, pages 1–8, 2022. 2

[36] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 3, 4

[37] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022. 2

[38] Junde Wu, Rao Fu, Huihui Fang, Yuanpei Liu, Zhaowei Wang, Yanwu Xu, Yueming Jin, and Tal Arbel. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023. 2

[39] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. *Advances in Neural Information Processing Systems*, 35:4203–4217, 2022. 3

[40] Huahui Yi, Ziyuan Qin, Qicheng Lao, Wei Xu, Zekun Jiang, Dequan Wang, Shaoting Zhang, and Kang Li. Towards general purpose medical ai: Continual learning medical foundation model. *arXiv preprint arXiv:2303.06580*, 2023. 2

[41] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions, 2016. 3

[42] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text, 2022. 2

[43] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*, 2023. 2

[44] Tao Zhou, Yizhe Zhang, Yi Zhou, Ye Wu, and Chen Gong. Can sam segment polyps?, 2023. 2

[45] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023. 2, 3