Residual-based Language Models are Free Boosters for Biomedical Imaging Tasks

Zhixin Lai [*†] Jing Wu [* ‡] Suiyao Chen [* §] Yucheng Zhou [¶] Naira Hovakimyan [‡]

Abstract

In this study, we uncover the unexpected efficacy of residual-based large language models (LLMs) as part of encoders for biomedical imaging tasks, a domain traditionally devoid of language or textual data. The approach diverges from established methodologies by utilizing a frozen transformer block, extracted from pre-trained LLMs, as an innovative encoder layer for the direct processing of visual tokens. This strategy represents a significant departure from the standard multi-modal vision-language frameworks, which typically hinge on language-driven prompts and inputs. We found that these LLMs could boost performance across a spectrum of biomedical imaging applications, including both 2D and 3D visual classification tasks, serving as plug-and-play boosters. More interestingly, as a byproduct, we found that the proposed framework achieved superior performance, setting new state-of-the-art results on extensive, standardized datasets in MedMNIST-2D and 3D. Through this work, we aim to open new avenues for employing LLMs in biomedical imaging and enriching the understanding of their potential in this specialized domain. The code is available at https://github.com/ZhixinLai/LLMBoostMedical

1. Introduction

Modern healthcare research is multifaceted, integrating various disciplines and technologies to improve patient outcomes [10], healthcare delivery [9], and disease prevention [24]. One of the most critical components is biomedical imaging. The ability to classify and segment medical images accurately and swiftly is essential for clinicians, reducing errors and improving patient care. Recent advancements in artificial intelligence (AI) for vision, such as Vision Transformers (ViTs), have significantly contributed to these areas. These AI models enhance the accuracy and efficiency of medical image analysis, aiding in the development of computer-aided diagnostic systems in clinical applications. By learning from large volumes of medical data, AI technologies can produce accurate diagnostic results across a range of medical applications. Their performance is often comparable to that of experienced clinicians, highlighting the transformative impact of AI in healthcare and its growing role in improving diagnostic processes.

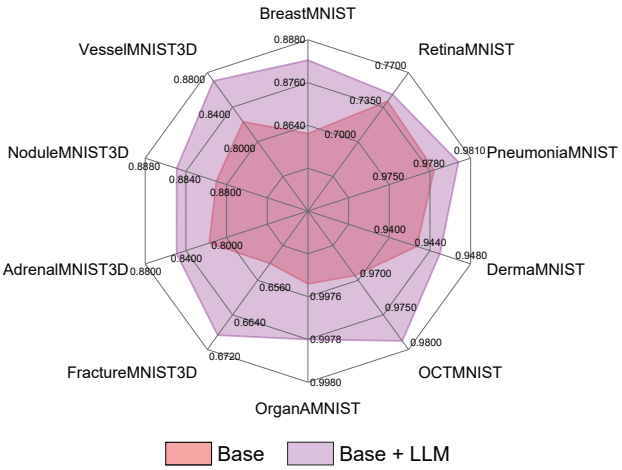

Figure 1. R-LLM benefits baseline models on a broad range of datasets in biomedical imaging tasks under the AUC metric.

Despite the promising capabilities of ViTs in biomedical imaging, we still face significant challenges that hinder further performance enhancements. First, the challenge lies in the data requirement for training these models. Effective training demands extensive, meticulously labeled datasets. However, in the realm of biomedical imaging, creating such datasets is particularly burdensome. The need for expert knowledge is paramount due to the fine-grained nature of medical images. This process is not only time-intensive but also incurs significant financial costs, making it a substantial barrier to progress. Second, the optimization of ViT presents a critical challenge similar to the broader computer vision domain. Achieving the best performance necessi-

*These authors contributed equally to this work.
†Cornel University
‡University of Illinois at Urbana-Champaign
§University of South Florida
¶University of Macau

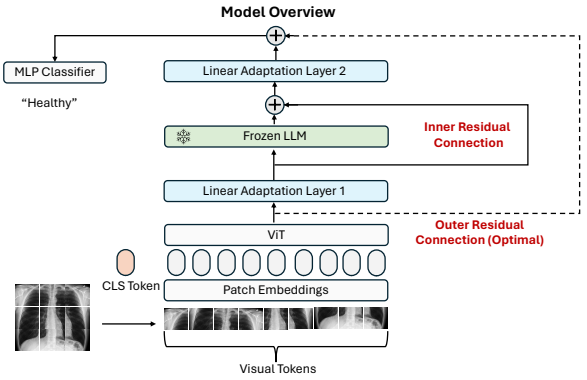

Figure 2. The proposed framework of applying language models as a booster for biomedical imaging classification task. We use Vision Transformer (ViT) from [17] for demonstration.

tates rigorous parameter tuning, a process that requires a deep understanding of the model architecture and consumes considerable computational resources. This level of optimization, while crucial for maximizing model efficacy, is a demanding task that often stretches beyond practical limits in terms of time and computational expense. Confronted with these two significant challenges, this research focuses on exploring strategies to enhance the performance of ViT in biomedical imaging without accumulating larger datasets or dramatically increasing computational demands.

LLMs, trained on extensive textual data, have shown impressive versatility, applying their capabilities far beyond their initial linguistic applications. In computer vision, for instance, LLMs have demonstrated an intriguing capacity to engage with and interpret visual tokens, converting them into a structured, tokenized format. This integration often occurs within a multi-modal vision-language framework. Here, visual tokens are typically interfaced with LLMs through linear projection layers, or by employing cross-attention mechanisms that facilitate interaction between visual and linguistic tokens. As we delve deeper into the potential of LLMs in computer vision, a compelling question emerges: Can these models, originally designed for language processing, adeptly manage purely visual tasks, without any dependence on linguistic elements?

In pursuit of understanding the capability of LLMs in visual tasks, our research offers a novel and affirmative insight. We introduce an approach that has been largely unexplored until now: utilizing a residual-based LLM (R-LLM) block as an efficient encoder for visual data. This method is distinct in its simplicity and effectiveness, with a significant performance boost on biomedical imaging tasks, as shown in Figure 1. Specifically, it involves three integral steps, as depicted in Figure 2: Firstly, we integrate a frozen transformer block from an LLM into the visual encoder's architecture. Secondly, to ensure compatibility and effective information transfer, trainable linear layers are strategically positioned around the LLM block, enabling seamless feature dimension alignment. Third, a residual connection before and after the frozen LLM is introduced. Finally, while the transformer block remains frozen to retain its pre-trained characteristics, the other modules are unfrozen and undergo regular optimization during the training phase.

Remarkably, the proposed straightforward approach yields significant performance improvements across a broad range of tasks in biomedical imaging, including both 2D and 3D classification tasks. This enhancement is consistently observed with various publicly available large language models, such as LLaMA, and across different transformer blocks within these LLMs. As shown in Figure 2a, the methodology innovates by treating LLM transformers as a booster of biomedical encoders, deviating significantly from the traditional perspective in vision-language models. Three key features distinguish our application of LLM transformers: First, their operation is entirely independent of language components, such as prompts, inputs, or outputs, marking a significant departure from traditional usage. Second, our method is adaptable both with and without pre-training, providing flexibility and bypassing the reliance on pre-trained models. Third, we simplify using LLMs by treating transformer blocks as distinct, modular units. This innovative approach not only challenges but also reshapes the conventional application of LLMs, particularly in the complex field of biomedical imaging tasks. In summary, our paper makes the following primary contributions:

- We introduce a novel residual-based framework that incorporates a frozen transformer block from pre-trained LLMs as a visual encoder layer, enhancing the learning of various biomedical imaging tasks. This innovative approach is tailored to adapt to the diverse and complex nature of biomedical images.
- Extensive experiments have been conducted across multiple datasets and scales, including BreastMNIST, DermaMNIST, FractureMNIST3D, etc. Surprisingly, the approach achieves state-of-the-art (SoTA) results, surpassing the performance of previous models. This underscores the effectiveness of our method in a wide array of medical imaging contexts.
- We provide in-depth discussions and ablation studies to dissect and understand the components of our proposed framework. These studies offer insights into the functionality and efficacy of each module, providing a comprehensive understanding of why and how our approach achieves its superior performance.

2. Related Work

2.1. Large Language Model

In the realm of large language models, evolution began with the pretraining of transformers [17] using masked token

prediction. This approach significantly enhances the versatility of language models across various tasks and modalities [16, 33, 49]. Following these advancements, the focus shifted towards developing larger-scale models, as guided by the scaling law [27]. This direction led to the creation of groundbreaking models such as GPT [7], LLaMA [43], OPT [53], BLOOM [48], and PaLM [12]. These models, with their tens of billions of parameters, unveiled the potential for advanced in-context learning and exceptional zero-shot performance across various tasks. However, the increasing complexity and size of these models presented new challenges in adaptability and efficiency. Addressing this, several papers have introduced innovative tuning methods, such as LoRA [22] and Q-LoRA [15], which aim to enhance the flexibility of these large models without the need for extensive retraining. For our work, we build upon this foundation and unveil an interesting discovery: the transformer blocks in such LLMs possess the unique capability to interact with biomedical data.

2.2. Vision Transformer

The Vision Transformer introduced by [17] exemplifies how a purely transformer-based model can achieve notable success in image classification. In ViT, images are divided into patches (tokens), and transformer layers are utilized to model the global interrelations among these patches for effective classification. Building upon this, the T2T-ViT [52] refines the tokenization process by recursively aggregating neighboring tokens, thereby enriching the representation of local structures. Similarly, the Swin Transformer [34] introduces a local window-based self-attention mechanism, with a shifted window scheme for comprehensive in-window and cross-window interaction modeling. In biomedical imaging, these technologies have also led to more accurate and efficient medical image segmentation and classification [13, 18, 45], leveraging transformers to handle variable-length inputs and capture long-distance dependencies.

2.3. Language Models for Visual and Biomedical Imaging Tasks

In the general vision domain, the advent of large language models (LLMs) has catalyzed a wave of innovative applications due to their generative capabilities. Notably, LLMs are being utilized to merge vision algorithms with user queries, enabling more interactive and user-specific outcomes, as explored in recent studies [41]. Another area of advancement is in visual programming, where LLMs play a central role in visual reasoning and in-context learning [20]. Furthermore, the versatility of LLMs as decoders is increasingly recognized, with their ability to translate latent visual features into meaningful output tokens [46]. Common methodologies in this domain involve projecting visual features directly onto the input layers of LLMs [19, 31, 36], or leveraging latent bottleneck structures to encode visual tokens more effectively [3, 26, 30, 46].

Beyond traditional visual tasks such as image tasks objection detection [42] and image understanding [11], researchers in the biomedical imaging field have developed datasets that bridge the gap between vision and language [25, 32, 47]. Utilizing these specialized datasets, significant advancements have been made in applying general-domain vision-language models to biomedical imaging [6, 23, 54]. These models have shown promising results in enhancing the analysis and interpretation of medical images. However, they still require careful alignment between the visual and linguistic modalities or an additional mapping process to translate visual information into the language space.

Recent advancements in the vision domain have illuminated the potential of using transformer blocks from LLMs as general-purpose encoder layers for visual data [37]. This perspective marks a departure from their traditional roles, primarily confined to encoding textual data, decoding tokenized outputs, or facilitating alignment between modalities. Instead, the pre-trained blocks may discern informative visual tokens and amplify their impacts on feature representation. Inspired by this, we hypothesize that a similar idea could be effectively adapted to biomedical imaging tasks.

3. Method

In this section, we first introduce the overall framework of the proposed method in Section 3.1. Following this, we highlight the key design and differences between the framework and previous methods in Section 3.2.

3.1. The Overall Framework

We now formally introduce our comprehensive framework that harnesses the power of LLM as a free booster for biomedical imaging tasks. The entire workflow of this framework is delineated in Figure 2. Traditionally, the framework begins by taking a biomedical image as input, denoted as x. It then utilizes a vision transformer-based encoder, $\mathcal{F}_{\mathcal{V}}$, to transform x into a feature embedding z. This process is followed by an MLP-based classifier $\mathcal{F}_{\mathcal{C}}$ for the final classification task, correlating with the corresponding label y. For the supervised learning, we define it as

$$\begin{aligned} \mathcal{F}_{\mathcal{V}}(x) &= z, \\ \mathcal{F}_{\mathcal{C}}(z) &= y. \end{aligned} \tag{1}$$

Following the baseline framework, we incorporate a pre-trained block from LLM, specifically selecting a block from LLaMA [43] in this study. We denote this LLM block as $\mathcal{F}_{\mathcal{L}}$. To effectively integrate $\mathcal{F}_{\mathcal{L}}$ into the vision-based pipeline, we introduce two additional adaptation layers: $\mathcal{F}_{\mathcal{E}}$ and $\mathcal{F}_{\mathcal{D}}$. The layer $\mathcal{F}_{\mathcal{E}}$ is positioned before $\mathcal{F}_{\mathcal{L}}$, while $\mathcal{F}_{\mathcal{D}}$ follows it. These layers serve a critical function in aligning

the dimensions between the vision data and the language model, ensuring seamless interoperability and efficient processing within our hybrid framework. Very importantly, we strategically implement a residual connection [21], positioned both before and after the LLM block. This setup allows an efficient exchange of gradient information and the passage of visual embedding through a shortcut path. Such an architecture not only facilitates the learning process but also ensures that crucial information is effectively preserved and communicated across models with different modalities, i.e., vision and language. We formally formulate this as

$$\begin{aligned} \mathcal{F}_{\mathcal{E}} \cdot \mathcal{F}_{\mathcal{V}}(x) &= r, \\ \mathcal{F}_{\mathcal{D}} \cdot (\mathcal{F}_{\mathcal{L}}(r) + r) &= z, \\ \mathcal{F}_{\mathcal{C}}(z) &= y. \end{aligned} \tag{2}$$

During training, we freeze all the parameters of $\mathcal{F}_{\mathcal{L}}$, the LLM transformer block. Meanwhile, the rest of the modules, including two adaptation layers, $\mathcal{F}_{\mathcal{E}}$ and $\mathcal{F}_{\mathcal{D}}$, are trained simultaneously. Following the previous paradigm [37], the approach modifies the behavior of LLM transformers to accommodate the stark differences between visual and textual data formats. Specifically, there are two critical adaptations. First, in LLMs, auto-regressive masks are typically used to simulate the sequential nature of text generation. However, in visual data, such as image tokens, the information is presented simultaneously rather than sequentially. Recognizing this, we forgo using auto-regressive attention masks in our framework. Instead, we employ attention masks solely to denote the presence of padded tokens in the visual data. Second, the positional embeddings utilized in LLMs, like the rotary positional embedding in LLaMA [43], are not typically chosen for visual encoders. Hence, for the sake of simplicity and to maintain consistency with conventional visual backbones, we opted to remove the LLMs' positional embeddings from our system.

3.2. Comparison with Previous Methods

At first glance, the proposed methods may appear akin to those used in prior vision-language model research, such as in video language retrieval [31], FROMAGe [28], and LiMBeR [36], where bridging the gap between vision and language spaces is achieved through linear layers. However, a distinctive aspect of our approach is the absence of an alignment between these two modalities' spaces. In essence, $\mathcal{F}_{\mathcal{E}}$ is not constrained to map features directly from the vision to the language space, differing fundamentally from these previous methods. This conclusion and design are consistent with the previous results shown in [37]. To be more specific, the method we propose distinguishes itself in several critical ways. Unlike prevailing approaches, it does not depend on a pre-trained encoder such as CLIP [39], ALBEF [29] and Coca [51], enabling the model to be trained entirely from scratch. This independence from pre-existing models offers greater flexibility and adaptability. Additionally, the method functions and operates autonomously from language-based inputs or prompts, which are applicable to general biomedical imaging Tasks. Most notably, our approach represents a pioneering attempt to employ a residual connection to facilitate information exchange among different modalities, a design particularly novel in biomedical imaging. These three aspects - independence from pre-trained models, autonomy from language-based inputs, and the innovative use of residual connections across modalities - collectively underscore the distinctiveness and innovation of our method in advancing biomedical imaging technology.

4. Experinments and Results

In this section, we conduct extensive empirical evaluations and experiments to validate the effectiveness of our proposed method as a cost-free, plug-and-play booster for biomedical imaging tasks. We begin by detailing the datasets utilized in our study in Section 4.1. Subsequently, in Section 4.2, we delve into the experiments conducted on 2D classification tasks. Following this, Section 4.3 will cover the 3D classification tasks, providing insights into the implementation details, experiments conducted, and the results derived from these tasks. Lastly, we conduct a series of ablation studies to understand and explore variants of the proposed method in Section 4.4.

4.1. Datasets

We carefully selected datasets from MedMNIST V2 [50], supplemented with other public datasets. Specifically, the chosen datasets encompass a broad spectrum of imaging types featuring both 2D and 3D images. Additionally, these datasets provide a diverse range of classification challenges, including both binary and multi-class tasks.

We commence our testing with a foundational 2D dataset, comprising 780 images, to carry out binary classification tasks. This initial phase is for a preliminary evaluation of our proposed approach. Progressing from there, we expand the scale of the datasets under investigation, transitioning from hundreds to over 100,000 images. Given the limited availability of 3D datasets, our selection for 3D image analysis includes four datasets, each containing thousands of images under similar scales. We described the details of the datasets as follows:

BreastMNIST, drawing from a dataset of 780 breast ultrasound images [2], classifies these images into three categories: benign, malignant, and normal. Given that the dataset comprises low-resolution images, the task has been simplified into a binary classification framework.

RetinaMNIST is derived from the DeepDRiD (Deep Diabetic Retinopathy) dataset [8], featuring data from 628 patients and encompassing 1600 retina fundus images.

Dataset	BreastMNIST		RetinaMNIST		PneumoniaMNIST		DermaMNIST		OCTMNIST		OrganAMNIST	
Backbone	ViT-S		ViT-S		ViT-S		ViT-S		ViT-S		ViT-S	
R-LLM	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
ACC	**87.17**	**87.17**	54.25	**57.00**	**94.23**	93.91	78.95	**79.50**	83.60	**85.10**	95.19	**95.22**
AUC	86.17	**88.23**	74.09	**74.78**	97.83	**98.01**	94.27	**94.50**	96.93	**97.88**	99.75	**99.78**

Table 1. Performance comparison of 2D classification results of the proposed framework with and without the Residual-based LLM as a booster, evaluated using the AUC and ACC metrics. The highest-performing results are highlighted in **bold** for clarity.

PneumoniaMNIST, adapted from an existing dataset [38], is comprised of 5,856 pediatric chest X-ray images. This dataset is particularly focused on the classification of pneumonia and is structured into two binary classes: 'pneumonia' and 'normal.'

DermaMNIST is derived from the HAM10000 dataset [44], a substantial compilation of multi-source dermatoscopic images showcasing common pigmented skin lesions. This dataset encompasses 10,015 dermatoscopic images, each with dimensions of 450 × 600 pixels.

OCTMNIST is derived from a previously established dataset [14], consisting of 109,309 valid optical coherence tomography (OCT) images collected specifically for the study of retinal diseases. The dataset encompasses four distinct types of retinal conditions, which form the basis for a multi-class classification task.

OrganAMNIST originates from 3D computed tomography (CT) images utilized in the Liver Tumor Segmentation Benchmark (LiTS) [1] with 58,850 images. To obtain organ labels for these images, bounding-box annotations of 11 body organs from a separate study were employed [35].

FractureMNIST3D is derived from the RibFrac Dataset [4], featuring about 5,000 rib fractures from 660 CT scans. We adhere to the official dataset division for experiments.

AdrenalMNIST3D, derived from Zhongshan Hospital affiliated with Fudan University, encompasses shape masks from 1,584 adrenal glands (792 patients). It includes 3D shapes of adrenal glands for binary classification. This dataset is randomly divided into training, validation, and test sets, with 1,188, 98, and 298 cases, respectively, ensuring a patient-level split.

NoduleMNIST3D is developed from a substantial public lung nodule dataset derived from thoracic CT scans. The dataset is partitioned in a 7:1:2 ratio into training, validation, and test sets. The images, spatially normalized to a 1mm×1mm×1mm spacing, are center-cropped to a uniform size of 28×28×28 for analysis.

VesselMNIST3D comprises 103 3D brain vessel models derived from reconstructed MRA images. From these models, 1,694 healthy vessel segments and 215 aneurysm segments have been generated. The source dataset has been divided into training, validation, and test sets in a 7:1:2 ratio, facilitating a comprehensive evaluation of the models across various samples.

4.2. 2D Classification

We now dive into the experiments of 2D classification tasks for biomedical images. We will first introduce the detailed implementation and then move to the corresponding results.

4.2.1 Implementation Details

For 2D classification experiments, all images are initially resized to a resolution of 224 x 224 pixels. We train each model using a batch size of 128, employing an AdamW optimizer for 100 epochs. The initial learning rate is set at 0.0005, coupled with a weight decay of 0.05. We utilize the ViT small model as the encoder pre-trained on ImageNet along with the llama-7b while keeping all parameters unfrozen for end-to-end training, except those in the LLaMA model. All these experiments are carried out on NVIDIA A6000 GPUs.

4.2.2 Results

In demonstrating the effectiveness of the R-LLM as a booster for 2D classification tasks, we primarily utilize Accuracy (ACC) and Area under the ROC Curve (AUC) as evaluation metrics. ACC, being a threshold-based metric, is particularly sensitive to class discrepancy as it evaluates discrete prediction labels. In contrast, AUC is a threshold-free metric suited for assessing continuous prediction scores. Given the diversity in dataset sizes and types in our experiments, employing both ACC and AUC provides a comprehensive assessment of our method's performance across varying conditions.

The results in Table 1 demonstrate that integrating the LM consistently enhances performance across various datasets and evaluation metrics. Notably, the most significant accuracy gains, approximately 1 to 3 percent, are observed in datasets such as RetinMNIST, OCTMNIST, and DermaMNIST. While improvements in other datasets are less pronounced, this could be attributed to our approach of applying a uniform set of hyperparameters across all experiments to showcase the LM's general applicability. The relatively modest enhancements in certain cases might result from this methodological choice, as it potentially limits the fine-tuning of hyperparameters tailored to

Method \Dataset	BreastMNIST	RetinaMNIST	PneumoniaMNIST	DermaMNIST	OCTMNIST	OrganAMNIST
ResNet-18	83.3	49.3	86.4	75.4	76.3	93.5
ResNet-50	84.2	51.1	88.4	73.1	77.6	94.7
Auto-sklearn	80.3	51.5	85.5	71.9	60.1	76.2
AutoKeras	83.1	50.3	87.8	74.9	76.3	90.5
Google AutoML	86.1	53.1	94.6	76.8	77.1	88.6
MedVIT-S	**89.7**	56.1	**96.1**	78.0	78.2	92.8
ViT-S + R-LLM	87.2	**57.0**	93.9	**79.5**	**85.1**	**95.2**

Table 2. Performance comparison of 2D classification results (ACC) with the previous SoTA methods. The best values are shown in **bold**.

Dataset	FractureMNIST3D		AdrenalMNIST3D		NoduleMNIST3D		VesselMNIST3D	
R-LLM	✗	✓	✗	✓	✗	✓	✗	✓
ACC (ViT-3D)	53.33	**54.58**	81.88	**82.89**	86.77	**89.68**	90.05	**91.10**
AUC (ViT-3D)	64.80	**65.15**	81.98	**83.86**	91.48	**92.39**	82.55	**83.71**
ACC (ViViT-S)	53.75	**55.00**	79.87	**81.21**	85.81	**86.45**	88.74	**90.31**
AUC (ViViT-S)	65.54	**66.20**	81.04	**82.12**	86.55	**88.76**	83.88	**84.56**
ACC (ViViT-M)	53.33	**56.25**	81.54	**83.22**	85.81	**87.42**	89.27	**90.58**
AUC (ViViT-M)	65.21	**66.87**	81.70	**84.91**	88.10	**88.49**	82.31	**87.03**

Table 3. Performance comparison of 3D classification results of the proposed framework with and without the Residual-based LLM as a booster, evaluated using the AUC and ACC metrics. The highest-performing results are highlighted in **bold** for clarity.

each specific dataset's characteristics. Interestingly, we noticed that R-LLM did not contribute to improving the ACC metric in the PneumoniaMNIST dataset. This observation can be attributed to the dataset's imbalanced nature, with a pneumonia-to-normal ratio of approximately 3:1. Consequently, accuracy can be misleading in such an imbalanced setting, as the baseline may achieve better accuracy simply by predicting most samples as the majority class. As we switch from ACC to AUC, we can see a more fair comparison and consistently observe that R-LLM continues to benefit the classification tasks.

More surprisingly, when the LLM booster is integrated into the basic ViT model, it not only matches but, in some cases, even surpasses existing SoTA results. As outlined in Table 2, this novel approach achieves unparalleled accuracy in datasets like BreastMNIST, RetinaMNIST, DermaMNIST, and OCTMNIST. Most notably, our method outperforms the SoTA on OCTMNIST by a remarkable margin of nearly 7 percent.

4.3. 3D Classification

We now move to the experiments of 3D classification tasks for biomedical images. Similarly, we will first introduce the detailed implementation and then the corresponding results.

4.3.1 Implementation Details

For the 3D classification experiments, each model is trained using a batch size of 128, employing an AdamW optimizer across 100 epochs. The initial learning rate is 1×10^{-5}. We adopt the ViViT [5] and ViT3D [17], both modified with

Method \Dataset	FractureMNIST3D	AdrenalMNIST3D	NoduleMNIST3D	VesselMNIST3D
ResNet-18 + 3D	50.8	72.1	84.4	87.7
ResNet-18 + ACS	49.7	75.4	84.7	**92.8**
ResNet-50 + 3D	49.4	74.5	84.7	91.8
ResNet-50 + ACS	49.4	78.5	84.1	85.8
Auto-sklearn	51.7	80.2	87.4	91.5
AutoKeras	45.8	70.5	83.4	89.4
ViT3D-M + R-LLM	54.6	82.9	**89.7**	91.1
ViViT-M + R-LLM	**56.3**	**83.2**	87.4	90.6

Table 4. Performance (ACC) comparison of 3D classification with the previous SoTA methods. The best values are shown in **bold**.

three channels to accommodate the 3D input, alongside the llama-7b model. The ViT3D model comprises 130.3M parameters. For ViVit, we utilize two encoder sizes: ViViT-Small (ViViT-S) and ViT-Medium (ViViT-M), containing 49.2M and 258.6M parameters, respectively. All parameters, except for those in LLaMA, are kept unfrozen for end-to-end training. These experiments are conducted on NVIDIA A6000 GPUs.

4.3.2 Results

Similar to the 2D datasets, we present the results for 3D datasets, reinforcing the core assertion of this paper: that LMs serve as a free booster for general bioimaging tasks, including 3D analysis. As illustrated in Table 3, the results are reported for various datasets with and without the R-LLM incorporated. These results are spread across different types and scales of encoders, specifically including ViT3D, ViViT-S, and ViViT-M. Crucially, in all scenarios and across both ACC and AUC evaluation metrics, we observe marked improvements in model performance. This consistent enhancement underscores the versatility and effectiveness of the LLM as a booster in the realm of 3D biomedical imaging tasks.

For the comprehensive experiments, we follow the 2D experiment settings to compare the proposed method with previous SoTA approaches. Remarkably, in Table 4, our framework notched three SoTA results across four datasets without any additional hyperparameter tuning. Meanwhile, even more favorable outcomes might be attainable with further optimization and customization of training parameters.

Method	Dataset	Num. of Parameters	ACC	AUC
ViViT-M	FractureMNIST3D	258.6M	53.33	65.21
ViViT-M + MLP	FractureMNIST3D	294.6M	54.17	65.11
ViViT-M + R-LLM	FractureMNIST3D	294.6M	**56.25**	**66.87**
ViViT-M + R-LLM(FT)	FractureMNIST3D	1066.62M	53.57	64.70
ViViT-M	AdrenalMNIST3D	258.4M	81.54	81.70
ViViT-M + MLP	AdrenalMNIST3D	294.6M	81.88	82.89
ViViT-M + R-LLM	AdrenalMNIST3D	294.6M	**83.22**	**84.91**
ViViT-M + R-LLM(FT)	AdrenalMNIST3D	1066.62M	79.87	81.10

Table 5. Ablation study on model capacity and fine-tuning. The best values are shown in **bold**.

Method	Dataset	ACC	AUC
ViViT-M	FractureMNIST3D	53.33	65.21
ViViT-M + R-LLM	FractureMNIST3D	**56.25**	**66.87**
ViViT-M + Out R-LLM	FractureMNIST3D	55.83	65.60
ViViT-M + Hybrid R-LLM	FractureMNIST3D	55.00	65.50
ViViT-M	AdrenalMNIST3D	81.54	81.70
ViViT-M + R-LLM	AdrenalMNIST3D	**83.22**	**84.91**
ViViT-M + Out R-LLM	AdrenalMNIST3D	82.55	82.96
ViViT-M + Hybrid R-LLM	AdrenalMNIST3D	82.55	82.68

Table 6. Ablation study on the importance of residual structure.

4.4. Ablation and Visualization

To further prove the effectiveness of the proposed idea and the importance of the introduced LLM block, we conduct comprehensive experiments with models of varying capacities, detailed in Section 4.4.1. In these experiments, we assess how the models perform with different levels of complexity. Subsequently, in Section 4.4.2, we explore the potential benefits of unfreezing the LLM block. This step is aimed at fully leveraging the adaptability and fitting power of the LLM. Then, we highlight the importance of residual structure in Section 4.4.3. Lastly, Crad-CAM visualization is given in Section 4.4.4.

4.4.1 Model with Different Capacities

In evaluating the broad effectiveness of frozen LLM transformers, we considered whether the improvements could be attributed more to the expanded capacity of the linear adaptation layers, namely $\mathcal{F}_{\mathcal{E}}$ and $\mathcal{F}_{\mathcal{D}}$, rather than the pre-trained weights of the LLM block, $\mathcal{F}_{\mathcal{L}}$. To investigate this, we created a variant model, ViViT-M+MLP, which has a parameter count equivalent to that of ViViT+R-LLM. This variant omits the LLM block $\mathcal{F}_{\mathcal{L}}$, and keeps $\mathcal{F}_{\mathcal{E}}$ and $\mathcal{F}_{\mathcal{D}}$.

We adhered to the same training procedure outlined in Section 4.3 to ensure a fair comparison, focusing our experiments on the FractureMNIST3D and AdrenalMNIST3D datasets. The results, summarized in Table 5, show that ViViT-M+MLP, with its increased number of parameters, does outperform the baseline ViViT-M model. However, the improvement is relatively marginal. In contrast, the enhancement observed with ViViT-M+R-LLM is both robust and substantial across both metrics. These findings lead to a significant conclusion: the pre-trained weights of the LLM transformer are instrumental to the observed improvements, and the enhancements in our biomedical imaging tasks are not merely the result of increased model capacity.

4.4.2 End-to-end Fine-tuning

In examining whether fine-tuning the language transformer in the ViViT-M+R-LLM(FT) model is advantageous compared to maintaining it in a frozen state, we found an unexpected outcome. The results, as shown in Table 5, indicate a decline in performance with fine-tuning, in contrast to the consistent training of the ViViT-M+R-LLM. This suggests the difficulties in training large transformer models: there is a tendency to overfit with standard-scale datasets, and fine-tuning LLMs end-to-end is often time-intensive and complex. This observation reinforces our decision to keep the LLM transformers frozen within our framework. By doing so, we simplify the training process while also ensuring effectiveness, thereby avoiding the challenges associated with fine-tuning in complex transformer architectures.

4.4.3 Importance of Residual Structure

In this ablation study, the significance of the residual structure within our framework is meticulously examined. We found that incorporating such a structure in tandem with a Large Language Model (LLM) substantially enhances model performance. To elucidate this further, we introduce two variants of our Residual-based R-LLM: the 'Out R-LLM' and the Hybrid R-LLM. Out R-LLM is designed to incorporate the residual connection before the encoder $\mathcal{F}_{\mathcal{E}}$ and externally to the decoder $\mathcal{F}_{\mathcal{D}}$. This can be summarized as follows:

$$\begin{aligned} \mathcal{F}_{\mathcal{V}}(x) &= r, \\ \mathcal{F}_{\mathcal{D}} \cdot \mathcal{F}_{\mathcal{L}} \cdot \mathcal{F}_{\mathcal{E}}(r) + r &= z, \\ \mathcal{F}_{\mathcal{C}}(z) &= y. \end{aligned} \tag{3}$$

Hybrid R-LLM, blending the features of R-LLM and Out R-LLM, combines both internal and external residual structures. This approach offers an alternative method of integration. In line with our previous experiments, the performance of Hybrid R-LLM is evaluated on FractureMNIST3D and AdrenalMNIST3D datasets using the ACC and AUC metrics. The findings, presented in Table 6, indicate that R-LLM delivers the best results. However, any form of the residual structure consistently benefits the overall performance.

4.4.4 Visual Inspection

To validate the efficiency of LLM, we utilize Grad-CAM [40] to qualitatively analyze the performance of ViT-S

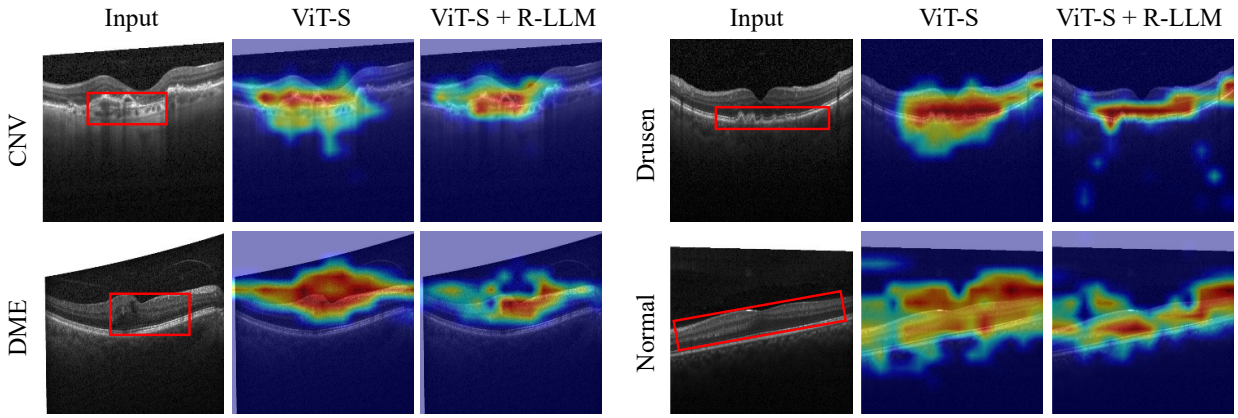

Figure 3. Visual inspection of ViT-S and ViT-S with R-LLM using Grad-CAM on original OCTMNIST dataset.

with R-LLM. We conduct training on the original OCTMNIST dataset [14], encompassing diverse retinal conditions: Choroidal Neovascularization (CNV), Diabetic Macular Edema (DME), Drusen, and Normal cases.

In Figure 3, significant regions are delineated by red rectangles, indicating areas crucial for medical diagnosis and analysis. Compared to the baseline, ViT-S enhanced with R-LLM demonstrates superior performance by closely aligning with these annotated red rectangles. This alignment enhances its ability to suppress attention toward extraneous background details effectively and to identify pivotal features essential for accurate diagnosis and analysis. This observation underscores the efficacy of our approach in medical image analysis tasks.

5. Discussion and Conclusion

5.1. Discussion

This study was primarily focused on methodically exploring a relatively under-investigated domain: the utility of pre-trained, frozen, and residual-based language transformers in biomedical imaging tasks. We have successfully demonstrated that these transformers can indeed serve as a 'free lunch', significantly boosting performance across various tasks. The experiments were carefully structured to cover a broad range of datasets and learning tasks, ensuring fair and meaningful comparisons with established baselines. Our focus was not exclusively on achieving state-of-the-art performance for every task, although this emerged as an unintended but welcome byproduct of our work.

This research not only confirms the value of LLMs in enhancing biomedical visual tasks but also opens the door for further exploration in this field. We urge fellow researchers to expand upon our work, potentially by enlarging the scope of experiments with more diverse datasets, which could lead to more universally applicable models in the industry. Moreover, we also recognize that our approach has not yet fully harnessed the specific traits of biomedical images, such as their fine-grained structures. Delving into these aspects could yield more nuanced insights and improvements, representing a vital and promising direction for future studies.

5.2. Conclusion

In this research, we explored the unique potential of residual-based large language models, traditionally associated with text processing, as encoders for biomedical imaging tasks. This innovative application marks a significant shift from their usual text-centric roles. By integrating a frozen transformer block from pre-trained LLMs into visual encoders as a free booster, we discovered consistent enhancements in performance across a variety of 2D and 3D biomedical imaging tasks. These findings broaden the scope of LLM applications, suggesting their utility extends well beyond language processing. Our study aims to inspire further exploration in this nascent field, particularly in bridging the modality gap between vision and language and harnessing the full potential of LLMs within the biomedical imaging domain.

References

[1] Andrea Acevedo, Anna Merino, Santiago Alférez, Ángel Molina, Laura Boldú, and José Rodellar. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in brief*, 30, 2020. 5

[2] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020. 4

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 3

[4] Samuel G Armato, RY Roberts, and MF Mcnitt-Gray. A completed reference database of lung nodules on ct scans. *Academic Radiology*, 14(12):1455–1463, 2007. 5

[5] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 6

[6] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer, 2022. 3

[7] 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. 3

[8] Kanghao Chen, Yifan Mao, Huijuan Lu, Chenghua Zeng, Ruixuan Wang, and Wei-Shi Zheng. Alleviating data imbalance issue with perturbed input during inference. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 407–417. Springer, 2021. 4

[9] Suiyao Chen, Nan Kong, Xuxue Sun, Hongdao Meng, and Mingyang Li. Claims data-driven modeling of hospital time-to-readmission risk with latent heterogeneity. *Health care management science*, 22:156–179, 2019. 1

[10] Suiyao Chen, Xinyi Liu, Yulei Li, Jing Wu, and Handong Yao. Deep representation learning for multi-functional degradation modeling of community-dwelling aging population. *arXiv preprint arXiv:2404.05613*, 2024. 1

[11] Zhimin Chen, Longlong Jing, Yingwei Li, and Bing Li. Bridging the domain gap: Self-supervised 3d scene understanding with foundation models. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[12] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113, 2023. 3

[13] Yin Dai, Yifan Gao, and Fayu Liu. Transmed: Transformers advance multi-modal medical image classification. *Diagnostics*, 11(8):1384, 2021. 3

[14] DC Dataset. The 2nd diabetic retinopathy–grading and image quality estimation challenge, 2020. 5, 8

[15] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[16] 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. 3

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3, 6

[18] Yunhe Gao, Mu Zhou, and Dimitris N Metaxas. Utnet: a hybrid transformer architecture for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 61–71. Springer, 2021. 3

[19] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10867–10877, 2023. 3

[20] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962, 2023. 3

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[23] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951, 2021. 3

[24] Mohamed Said Ibrahim and Sameh Saber. Machine learning and predictive analytics: Advancing disease prevention in healthcare. *Journal of Contemporary Healthcare Analytics*, 7(1):53–71, 2023. 1

[25] 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*, pages 590–597, 2019. 3

[26] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 3

[27] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 3

[28] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal inputs and outputs. 2023. 4

[29] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi.

Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 4

[30] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 3

[31] Xudong Lin, Simran Tiwari, Shiyuan Huang, Manling Li, Mike Zheng Shou, Heng Ji, and Shih-Fu Chang. Towards fast adaptation of pretrained contrastive models for multi-channel video-language retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14846–14855, 2023. 3, 4

[32] Zudi Lin, Donglai Wei, Won-Dong Jang, Siyan Zhou, Xupeng Chen, Xueying Wang, Richard Schalek, Daniel Berger, Brian Matejek, Lee Kamentsky, et al. Two stream active query suggestion for active learning in connectomics. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 103–120. Springer, 2020. 3

[33] Shengjie Liu, Jing Wu, Jingyuan Bao, Wenyi Wang, Naira Hovakimyan, and Christopher G Healey. Towards a robust retrieval-based summarization system. *arXiv preprint arXiv:2403.19889*, 2024. 3

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3

[35] Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature methods*, 9(7):637–637, 2012. 5

[36] Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*, 2022. 3, 4

[37] Ziqi Pang, Ziyang Xie, Yunze Man, and Yu-Xiong Wang. Frozen transformers in language models are effective visual encoder layers. *arXiv preprint arXiv:2310.12973*, 2023. 3, 4

[38] Kai Qi and Hu Yang. Elastic net nonparallel hyperplane support vector machine and its geometrical rationality. *IEEE Transactions on Neural Networks and Learning Systems*, 33 (12):7199–7209, 2021. 5

[39] 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. 4

[40] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 7

[41] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[42] Shangquan Sun, Wenqi Ren, Tao Wang, and Xiaochun Cao. Rethinking image restoration for object detection. *Advances in Neural Information Processing Systems*, 35:4461–4474, 2022. 3

[43] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3, 4

[44] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 5

[45] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 36–46. Springer, 2021. 3

[46] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[47] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 3

[48] BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022. 3

[49] Jing Wu, Zhixin Lai, Suiyao Chen, Ran Tao, Pan Zhao, and Naira Hovakimyan. The new agronomists: Language models are experts in crop management. *arXiv preprint arXiv:2403.19839*, 2024. 3

[50] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023. 4

[51] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 4

[52] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 558–567, 2021. 3

[53] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 3

[54] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022. 3