MedBLIP: Bootstrapping Language-Image Pretraining from 3D Medical Images and Texts

Qiuhui Chen[0000−0002−1210−1306] and Yi Hong✉[0000−0002−9065−3691]

Department of Computer Science and Engineering,
Shanghai Jiao Tong University, Shanghai 200240, China
✉yi.hong@sjtu.edu.cn

Abstract. Vision language pretraining (VLP) models have proven effective in numerous computer vision applications. In this paper, we focus on developing a VLP model for the medical domain to facilitate computer-aided diagnoses (CAD) based on image scans and text descriptions from electronic health records. To achieve this, we introduce MedBLIP, a lightweight CAD system that bootstraps VLP from off-the-shelf frozen pre-trained image encoders and large language models. We incorporate a MedQFormer module to bridge the gap between 3D medical images and 2D pre-trained image encoders and language models. To evaluate the effectiveness of our MedBLIP, we have collected over 30,000 image volumes from five public Alzheimer's disease (AD) datasets: ADNI, NACC, OASIS, AIBL, and MIRIAD. On this large-scale AD collection, our model demonstrates impressive performance in zero-shot classification of healthy, mild cognitive impairment (MCI), and AD subjects, and also shows its capability in medical visual question answering (VQA) on the M3D-VQA-AD dataset. The code and pre-trained models are available at https://github.com/Qybc/MedBLIP.

Keywords: Vision Language Pretraining · Computer-Aided Medical Diagnosis · Alzheimer's Disease

1 Introduction

Electronic health records (EHR), which include radiology images, lab and test results, and patient demographics, play a crucial role in clinical diagnosis. For instance, diagnosing Alzheimer's Disease (AD) involves not only brain imaging but also physical and neurological exams, as well as various diagnostic tests, with these results typically presented in textual form. Over the past decades, researchers have amassed extensive collections of EHRs for AD studies, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) [28], the National Alzheimer's Coordinating Center (NACC) [5], and the Open Access Series of Imaging Studies (OASIS) [25]. Despite this wealth of data, developing methods to diagnose AD based on these diverse EHRs remains a significant challenge in computer-aided diagnosis (CAD). This challenge is particularly pronounced in the integration of heterogeneous medical data, such as images and texts, from

various sources. Beyond merely estimating disease categories or risks, we also prefer a CAD system that can accurately interpret and respond to medical images and textual data by answering disease-related questions, thereby enhancing their utility in clinical diagnosis settings.

Recently, large vision language pretraining (VLP) models, e.g., CLIP [29], BLIP [20,19], Flamingo [1], LLaMA [34], have achieved significant success in various downstream computer vision applications, such as classification [4], segmentation [39], and generating responses based on visual and textual information. These VLP models learn multimodal representations by aligning features from extensive image and text datasets into a common space. In the medical domain, researchers have introduced Medical Bootstrapping Language-Image Pretraining (MedCLIP) [38], which learns generic representation from large-scale medical image-text pairs, demonstrating generalization across various medical tasks, particularly when medical data or labels are limited. The recent Med-Flamingo [26] model is a multimodal few-shot learner with generative medical visual question answering abilities. Another large language model (LLM), Med PaLM [30,31], encodes clinical knowledge to answer medical questions spanning medical exams, medical research, and consumer medical questions. However, these models primarily address general medical questions and focus on 2D medical image scans, and the potential of LLMs in medical diagnosis remains underexplored.

In this paper, we focus on integrating medical scans with textual information in EHRs, such as age, gender, and lab results, to enhance AD diagnosis. Our objective is to develop a VLM tailored for CAD scenarios, capable of fusing diverse types of medical data. We address three key challenges: (1) Extending a 2D image encoder to extract features from 3D medical images; (2) Aligning image and text features to learn multimodel representations; and (3) Developing a lightweight language model for CAD question-answering. Inspired by BLIP-2 [19], we propose MedBLIP, as shown in Fig. 1, a bootstrapping language-image pretraining model designed to fuse 3D medical images and texts using a query mechanism. We first employ a learnable patch embedding to bridge the gap between 3D medical images and a pre-trained 2D image encoder, significantly reducing the required amount of image data for learning. Next, we introduce MedQFormer, which uses learnable queries to align visual and textural features needed by a language model. Finally, we select BioMedLM [36] as our base language model and fine-tune it with the LoRA [15] technique. Our CAD model, MedBLIP, is lightweight and can be trained on a single NVIDIA RTX 3090 GPU.

To train and evaluate the effectiveness of our proposed MedBLIP model, we collect more than 30,000 medical image volumes from five public AD datasets, including ADNI [28], NACC [5], OASIS [25], AIBL [11], and MIRIAD [24]. After pretraining on most of the images from ADNI, NACC, and OASIS datasets, we evaluate our MedBLIP on three tasks: (1) zero-shot classification, which directly applies pre-trained MedBLIP to classify unseen subjects from AIBL and MIRIAD datasets into three classes, i.e., normal controls (NC), mild cognitive impairment (MCI), and AD; (2) zero-shot medical visual question answering (VQA), which generates an initial diagnosis for an unseen AIBL or MIRIAD

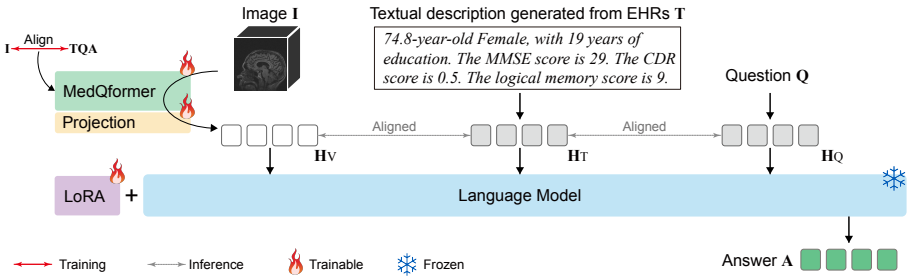

Fig. 1. Overview of MedBLIP architecture: A CAD system for medical diagnosis using multimodal representation learning in a language model with electronic health records.

subject based on input images and text descriptions and also provides some reasons for making such decision; and (3) zero-shot close-ended VQA on AD-related disease in the M3D-VQA dataset [2], which applies our MedBLIP to select answers from a set of possible choices when presenting with 3D medical scans and related questions.

Overall, our contributions of this paper are summarized below:

- We propose MedBLIP, a lightweight CAD system pre-trained on electronic health records comprising images and texts, capable of zero-shot classification and medical VQA. The architecture of our CAD system is versatile, with the potential to incorporate additional modalities and extend to diseases beyond AD.
- We propose the MedQFormer module, which extracts 3D medical image features and aligns them with textural features to be fused into a language model (LM). This module facilitates the alignment of diverse medical data types into the common space of the LM, offering a generic solution applicable to various medical applications.
- We have assembled a large-scale dataset containing over 30,000 3D image scans and accompanying texts to study AD, achieving promising results in classification and VQA tasks. This demonstrates the potential of leveraging LLMs for AD diagnosis with explanations to some extent. Additionally, our system operates directly on raw images without any preprocessing, making it practical and easy to use.

2 Related Works

Vision Language Pretraining. Data collected from different modalities typically provide different views about the data, which often complement each other and provide more complete information to facilitate a holistic understanding of the data. Vision-language pre-training (VLP) aims to learn multimodal foundation models, showing improved performance on various vision-and-language

tasks [29]. We can roughly divide current VLP models into two categories when fusing multi-modal inputs: light fusion and heavy fusion.

The approaches in the light fusion category focus on multimodal alignment, which facilitates text matching, retrieval, and other downstream tasks, with representative methods like CLIP [29] and ALIGN [17]. These methods directly align image representations with the corresponding text representations using a contrastive loss. DeCLIP [22] exploits inter/intra-modality supervision to train a CLIP-like model with fewer data. On the other hand, the heavy fusion category focuses on incorporating multimodal information with an attention mechanism to perform additional tasks. For instance, ALBEF [21] proposes a contrastive alignment, which is followed by deeper fusion with a multimodal encoder. Methods such as BLIP [20], MoCo [14], CoCa [40] incorporate a decoder and add image-to-text generation as an auxiliary task. Heavy fusion can interpret VQA, captions, and other downstream tasks that require more information for fusion and understanding.

Medical image-text representation learning has been investigated based on contrastive learning as well. CheXzero [33] directly applies the CLIP on large-scale chest X-ray datasets to enable zero-shot classification of unseen findings in images. MedCLIP [38] decouples paired images and texts and uses soft targets of semantic similarities to learn from unpaired medical images and texts. BioViL-T [3] proposes a novel multi-image encoder to augment the current image representation with information from previous images. Most existing medical VLP are designed based on 2D images, since compared to 3D image volumes, 2D slices are sufficient to form a large-scale dataset for learning. However, in this paper, we aim to develop a medical VLP based on 3D image volumes with relatively few parameters and limited data size, i.e., a lightweight medical VLP for learning a 3D medical image and text representation.

LLMs for Multimodal Understanding. Recently, using large language models (LLMs) as decoders in vision-language tasks has gained significant attention. This approach takes advantage of cross-modal transfer, which allows sharing knowledge between language and multimodal domains. VisualGPT [6] and Frozen [35] have demonstrated the advantage of employing a pre-trained language model as a vision-language model decoder. Flamingo [1] freezes a pre-trained vision encoder and language model and then fuses vision and language modalities with gated cross-attention. BLIP-2 [19] designs a Q-Former to align the visual features from the frozen visual encoder with large language models, like FLAN-T5 [8] and OPT [42]. FROMAGe [18] freezes large language models and visual encoders, and fine-tunes linear mapping layers to achieve cross-modality interactions. This method shows strong zero-shot performances on contextual image retrieval and multimodal dialogue tasks. Built upon PaLM [7], PaLM-E [10] employs features from sensor modalities and integrates real-world continuous sensor modalities into an LLM, thereby establishing a connection between real-world perceptions and human languages. GPT-4 [27] presents powerful visual understanding and reasoning abilities after pre-training on a vast collection of image-text data.

Most recently, several domain-specific multimodal LLMs have been developed. ChatCAD [37] combines visual and linguistic information processed by various networks as inputs of large language models to develop a medical-image CAD model, which provides a condensed report and offers interactive explanations and medical recommendations. Open-ended MedVQA [32] employs a multi-layer perceptron (MLP) network that maps the extracted visual features from a frozen vision encoder to a set of learnable tokens, which develops an open-ended VQA for diagnoses and treatment decisions. Differently, our MedBLIP explores a lightweight framework that works on 3D medical scans and aligns different types of medical data for CAD.

3 MedBLIP

3.1 Problem Formulation

We design a CAD system in the form of a dialogue, specifically for automatic AD diagnosis. Given inputs of a brain image scan I collected from a subject and a textual description T generated from this subject's EHRs in natural language, our CAD aims to generate an answer $A = \{A_0, A_1, ..., A_N\}$, composed of N tokens, for a question asked in natural language Q. The answer generation is conditioned on all inputs $\{I, T, Q\}$. To achieve this goal, we build a CAD model based on a large language model and find its optimal parameters θ^* by maximizing the conditional log-likelihood below:

$$\theta^* = \arg\max_{\theta} \sum_{i=1}^{N} \log P\left(A_i \mid A_{<i}, I, T, Q; \theta\right). \tag{1}$$

3.2 Network Framework

Our medBLIP model is structured as an encoder-decoder architecture, featuring a dual-stream encoder and a language model (LM) serving as the decoder, as illustrated in Fig. 1. Specifically, the dual-stream encoder processes inputs from two modalities: a vision sub-encoder for the image I, and a text sub-encoder for the textual description T and the question Q. The language model, defined as a causal language transformer, generates the answer A in an auto-regressive manner. This approach ensures that the model can effectively integrate visual and textual information to produce contextually relevant responses.

Vision Encoding Stream. To encode a 3D brain image volume while fully leveraging a current large model to reduce data requirements, we utilize a pre-trained 2D vision encoder to extract 3D visual features. This approach necessitates addressing two key challenges: (1) bridging the domain and dimension gaps between a 2D vision encoder and 3D medical scans, and (2) aligning image features with textural ones to map all inputs into the latent space of the LM decoder, thereby facilitating multimodal representation learning.

Drawing inspiration from [19], we introduce a query network based on a transformer encoder. This network maps the visual features into a visual prefix $H_v = \{v_1, v_2, \cdots, v_{\ell_v}\} \in \mathbb{R}^{\ell_v \times e}$ for the language model, where ℓ_v represents the length of the vision embedding sequence and e denotes the embedding size. Additionally, we incorporate a lightweight, learnable projection mechanism that adapts 3D image volumes to the input requirements of the pre-trained 2D image encoder. Our proposed medical query transformer (MedQFormer) effectively addresses these challenges, ensuring seamless integration of 3D medical imaging with 2D vision encoders and alignment with textual data. A detailed discussion of the MedQFormer is provided in Sec. 3.3.

Language Encoding Stream. To process the textural descriptions within subjects' EHRs, excluding image scans, we initially employ a standard tokenization procedure as outlined in [16]. This method yields a sequence of tokens, specifically the textual description $\mathbf{T} = \{t_1, t_2, \cdots, t_{\ell_t}\} \in \mathbb{R}^{\ell_t \times e}$, the question $\mathbf{Q} = \{q_1, q_2, \cdots, q_{\ell_q}\} \in \mathbb{R}^{\ell_q \times e}$, and the answer $\mathbf{A} = \{a_1, a_2, \cdots, a_{\ell_a}\} \in \mathbb{R}^{\ell_a \times e}$, where ℓ_t, ℓ_q, ℓ_a represent the lengths of the embedding sequences for the text, question, and answer, respectively. Subsequently, these tokens are embedded using the embedding function of a pre-trained language model. This embedding process transforms the token sequences into high-dimensional vectors that the language model can process. The embedded tokens $\mathbf{T}$, $\mathbf{Q}$, and $\mathbf{A}$ are thus aligned in a shared latent space with visual encoders via MedQformer, enabling the model to integrate and leverage multimodal data for comprehensive analysis.

Prompt Structure. To create a structured prompt in line with current VQA methodologies employed in language models [19,32], we precede the question and answer tokens with tokenized descriptive strings, specifically formatted as **question:** and **answer:**. The embeddings of the image and text description are positioned prior to the question tokens. Consequently, our prompt template is structured as follows:

$$\begin{aligned} p = & [v_1, v_2, \cdots, v_{\ell_x}, t_1, t_2, \cdots, t_{\ell_t}, \textbf{Question}: \textit{What} \\ & \textit{will this subject be diagnosed with?} \textbf{Answer:}], \end{aligned} \tag{2}$$

which is subsequently fed as input to the language model below.

Language Model. Following established practices in language modeling systems [36], we treat VQA as a conditional text generation task, optimizing the standard maximum likelihood objective during training. The language model receives the prompt sequence as input and generates the answer $\mathbf{A}$ token by token. Specifically, at each time step i, the model outputs logits, which parameterize a categorical distribution $p_\theta(\mathbf{A})$ over the vocabulary tokens. This distribution is represented as follows:

$$\log p_\theta(\mathbf{A}) = \sum_{l_a} \log p_\theta \left(a_i \mid v_1, \ldots, v_{\ell_v}, t_1, \ldots, t_{\ell_t}, q_1, \ldots, q_{\ell_q}, a_1, \ldots, a_{i-1} \right). \tag{3}$$

The parameters of the language model are initialized from a pre-trained model, which has been previously pre-trained on large web-collected datasets [9,13].

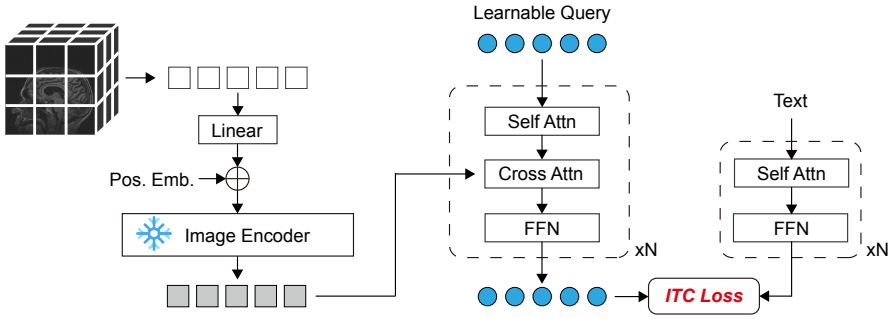

Fig. 2. Illustration of our proposed MedQformer, which aligns 3D visual and textural features for learning within the unified latent space of the language model.

3.3 MedQFormer

To bridge the gap between 3D medical images and 2D vision encoders pre-trained on natural images, and to align visual features to the text latent space, we draw inspiration from BLIP-2 [19] and utilize a query encoder to extract and align vision features accordingly.

Image Feature Extraction. To handle a 3D image volume I, we divide it into a set of 3D sub-volumes $\{Iv_i\}_{i=1}^{N_v}$. Each sub-volume is then projected into 1D image embeddings $\{E_i = f_{\varphi_1}(Iv_i)\}_{i=1}^{N_v}$ using a linear projection f_{φ_1}. To incorporate positional information, we add learnable position embeddings f_{φ_2}. These embeddings are then input into a standard pre-trained vision encoder to extract desired image features. Although the pre-trained vision encoder f_ϕ has fixed parameters ϕ, our approach utilizes learnable linear projection and position embeddings to adapt the 2D vision encoder for the 3D medical domain. Consequently, we develop a medical vision encoder with learnable parameters φ_1 and φ_2, mapping a volumetric image I into N_v visual features $f_1, \cdots, f_{N_v} = \{f_\phi(f_{\varphi_1}(Iv_i), f_{\varphi_2}(Iv_i))\}_{i=1}^{N_v}$. The final output is a set of image embeddings $IE = (f_1, \cdots, f_{N_v})$ for each input image volume I, where f_i represents the extracted visual features for each 3D sub-volume.

Query Encoder. To map the visual features $\{f_i\}_{i=1}^{N_v}$ into the common language space, we utilize a set of L learnable queries $qry_i \in \mathbb{R}^{d_e}$, where d_e is the dimension of the query embeddings. These queries are structured as transformers that interact with the image encoder to refine visual feature extraction and with a text transformer to extract texturally aligned visual features. As depicted in Fig. 2, these learnable queries interact with each other through self-attention layers and then interact with image features through cross-attention layers, followed by a Feed-Forward Network (FFN). This interaction process yields a visual prefix H_v that is aligned with textural features through image-text contrastive learning, making it suitable for processing by a language model.

3.4 Training MedBLIP

Learnable Parameters. Fine-tuning a language model with a small, domain-specific dataset can potentially diminish its generalization capabilities. To address this issue, we explore two parameter-efficient strategies that specifically adapt the attention blocks within the language models.

- **Frozen LM.** During training, the parameters of the language model remain completely frozen. In this configuration, only the 3D vision query network undergoes updates through backpropagation.
- **Low-Rank Adaptation (LoRA).** We incorporate learnable weight matrices into the query Q_w and value V_w components of the attention blocks at each layer of the frozen language model, represented as $W + \triangle W$, following the approach outlined by [15]. In this configuration, both the 3D vision query network and the learnable weight matrices are trained concurrently.

Objective Functions. Our MedBLIP model includes distinct loss functions for the MedQformer and LM modules. As detailed in Sec. 3.3, the MedQformer comprises a transformer image encoder E_I and a transformer text encoder E_T. During training, we utilize sets of image-text pairs (I, T) and image-diagnosis Q&A pairs (I, Q&A). To achieve multimodal representation alignment, we employ the image-text contrastive (ITC) learning loss as described in [29], leading to our feature alignment loss:

$$\mathcal{L}_{FA} = \mathrm{ITC}\left(E_I(I), E_T(T)\right) + \mathrm{ITC}\left(E_I(I), E_T(Q\&A)\right). \tag{4}$$

Similar to BLIP-2 [19], we select the output query embedding with the highest similarity to the text to compute the ITC Loss. To supervise the LM component, we compute the language generation loss $\mathcal{L}_{LG}$ using cross entropy. Consequently, the final loss function is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{FA} + \lambda_{LG}\mathcal{L}_{LG}, \tag{5}$$

where λ_{LG} is a hyperparameter to balance these two loss terms.

4 Experiments

4.1 Datasets and Experimental Settings

We collect more than 30,000 image volumes from five public datasets for studying AD/Dementia and evaluate our CAD system MedBLIP on separating subjects with AD or mild cognitive impairment (MCI) from normal controls (NC). Table 1 reports the demographic statistics of these five datasets.

ADNI [28]. This dataset has 10,387 volumetric T1 MRI scans that went through a series of pre-processing steps, including denoising, bias field correction, skull stripping, and affine registration to the SRI24 atlas, with an image size of $138 \times 176 \times 138$. For testing, we subject-wisely sample a subset of 200 images in each class (i.e., NC, MCI, AD), which is named ADNI-3x200.

Datasets		ADNI [28]	NACC [5]	OASIS [25]	AIBL [11]	MIRIAD [24]
#Images		10387	15354	3020	1002	708
#Text	#F/#M	4710/5677	9058/6296	1798/1222	471/531	393/315
	Age (#)	45-95 (10386)	19-102 (15354)	18-98 (3020)	42-96 (1002)	55-87 (708)
	#Educ	9860	15329	2300	-	-
	#SES	-	-	2153	-	-
	#MMSE	9385	7867	2293	1002	268
	#CDR	9401	15354	2300	1002	46
	#LM	7189	7654	-	1002	-
	Diagnosis (#)	NC,MCI, AD E/L/S/PMCI (10387)	NC, IMCI, MCI, DEM (14277)	DEM, Non-DEM (336)	NC, MCI, AD (997)	NC, AD (708)

Table 1. Demographic statistics of the AD datasets we have collected for experiments. F: Female, M: Male, Educ: Education level, SES: Socio-Economic Status, MMSE: Mini-Mental State Examination, CDR: Clinical Dementia Rate, LM: Logical Memory, E/L/S/PMCI: early, late, stable, and progressive MCI, IMCI: Impaired not MCI, and DEM: demented. # indicates the number.

NACC [5]. This dataset has a large amount of raw volumetric T1 MRI scans with a variety of resolutions. We select those MRIs having 100∼256 slices in all three dimensions, resulting in 15,354 images. Unlike the ADNI dataset, we directly use the raw data; but similarly, we sample subject-wisely a NACC-3x200 subset for testing.

OASIS [25]. We collect 3020 volumetric T1 MRIs from OASIS 1&2. These scans went through pre-processing with denoising and skull stripping and have a size of $256 \times 256 \times 256$. Since OASIS 1 only releases some clinical reports but with no diagnoses (e.g. NC, MCI or dementia), we use all images from OASIS 1 for pre-training. For testing, we sample subject-wisely an OASIS-2x200 subset from OASIS 2 to separate demented and non-demented subjects.

AIBL [11]. This dataset has 1002 volumetric T1 MRI scans with sizes of $160 \times 240 \times 256$, which are collected from demented, MCI, or healthy subjects. We do not use this data for training; for testing, we sample a balanced subset with 200 images each for NC, MCI, and dementia classes.

MIRIAD [24]. We collect 708 raw volumetric T1 MRI scans, which have an image size of $124 \times 256 \times 256$. This is a binary classification dataset with two labels, i.e., demented and not-demented subjects. We sample a balanced subset with a 1:1 positive and negative ratio, resulting in 2×200 images for testing. No images are used for training to perform zero-shot experiments.

As a result, we have most images from ADNI, NACC, and OASIS datasets for pretraining and save images from AIBL and MIRIAD datasets for zero-shot testing. In total, we held 1000 subjects with 2600 samples out for evaluation. To simplify the preprocessing step, all images are first padded to a cube and then scaled to a unified size of $224 \times 224 \times 224$ as inputs.

Table 2. Quantitative comparison of our MedBLIP with baseline methods on five datasets. The classification performance is measured in the mean accuracy (ACC(%)) with five runs. The best scores are in bold. (†: zero-shot)

Methods		LM size	Learnable #params	ADNI -3x200	NACC -3x200	OASIS -2x200	AIBL†	MIRIAD†
FLAN-T5 [8]	Text only	3.4B	-	37.0%	39.5%	46.7%	33.3%	60.0%
Ours w/ T5	Frozen		151M	50.5%	69.2%	61.3%	54.7%	64.0%
	LoRA		156M	64.0%	77.3%	75.8%	59.2%	66.8%
BioGPT [23]	Text only	1.5B	-	25.7%	21.7%	28.3%	26.7%	50.0%
Ours w/ BioGPT	Frozen		151M	56.3%	66.5%	66.0%	60.7%	55.2%
	LoRA		156M	62.2%	72.3%	71.7%	62.4%	59.7%
BioMedLM [36]	Text only	2.7B	-	62.5%	63.5%	61.8%	65.7%	46.3%
Ours w/ BioMedLM	Frozen		151M	71.2%	82.0%	79.8%	77.8%	66.1%
	LoRA		154M	**78.7%**	**83.3%**	**85.3%**	**80.8%**	**71.0%**

Implementation Details. For the frozen image encoder, we choose pre-trained ViT-G/14 from EVA-CLIP [12], which is demonstrated to be effective in BLIP-2 [19]. For the input image with a size of $224 \times 224 \times 224$, the patch size and the stride are both set as 32, resulting in image features with the size of 344×1408. For the MedQformer, we use 32 learnable queries, where each query has a dimension of 768 and the number of hidden layers N is set to 12. Regarding language models, we have three options, i.e., FLAN-T5 [8], BioGPT [23], and BioMedLM [36]. FLAN-T5 is an instruction-trained model with 3B parameters trained on C4 WebText [9]. BioGPT and BioMedLM are both GPT models relying on GPT-2 architecture, pre-trained on PubMed and biomedical data from the Pile [13], with a size of 1.5B and 2.7B parameters, respectively. All our models can fine-tune on a single NVIDIA RTX 3090 GPU. We use the AdamW optimizer with a learning rate of 5e-3. The hyperparameter λ_{LG} is set to 1.

4.2 Experimental Results

Medical Classification. Table 2 reports the evaluation of our MedBLIP using different language models and settings. The language models, i.e., FLAN-T5, BioGPT, and BioMedLM, show their capability of performing monomodal medical CAD, i.e., using text descriptions only, to some extent. Among these three language models, BioMedLM performs the best, showing that it captures some dependencies between prompts and inherent knowledge when generating answers. By adding the visual modality even without fine-tuning, the performance of our model on all datasets has improved significantly. The accuracy improvement varies within [4.0%, 44.8%], and the BioGPT benefits the most from the visual input. This result indicates the necessity of using image scans in diagnosis. Using the fine-tuning technique LoRA, our performance is further improved, with at least 1.3% and at most 13.5% improvement in accuracy. Over-

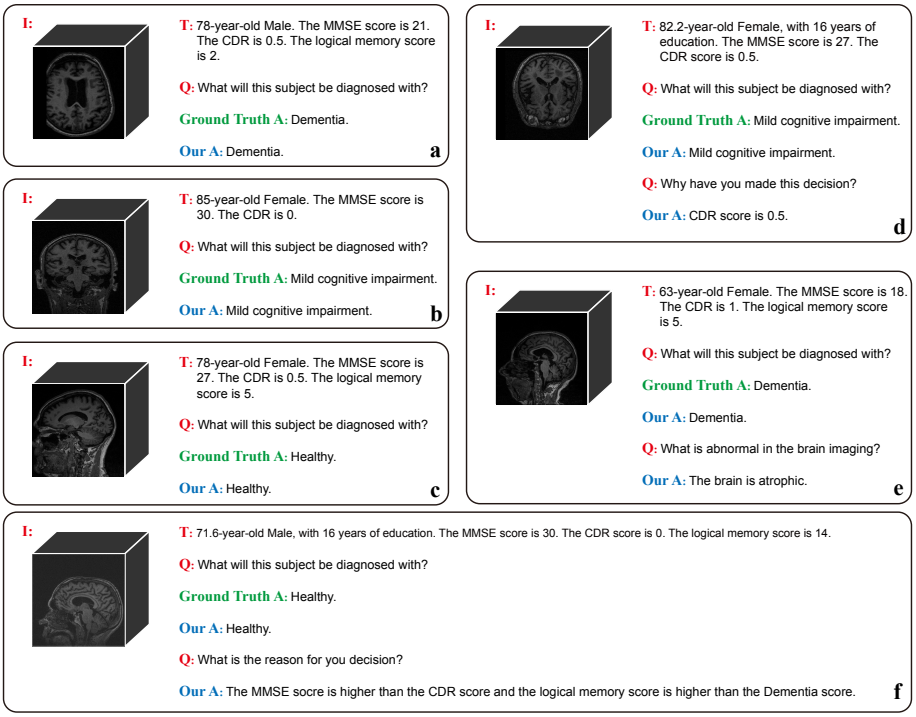

Fig. 3. Samples of zero-shot results on the AIBL dataset, which are generated by our MedBLIP built on BioMedLM with LoRA fine-tuning.

all, our MedBLIP built upon BioMedLM and LoRA fine-tuning shows the best performance on all datasets.

Qualitative Analysis. Figure 3 (a-c) visualizes the zero-shot CAD process on unseen subjects sampled from the AIBL dataset. Take Fig. 3(b) for example, although the text description of this subject shows no significant difference from those of healthy subjects, e.g., the CDR score is 0; while in brain scans the hippocampus and ventricle show the presence of abnormal atrophy. Our MedBLIP provides the correct diagnosis of MCI for this subject.

Figure 3 (d-f) qualitatively shows the zero-shot medical VQA ability of our MedBLIP on the AIBL dataset. Since our approach is generative, after a simple classification-based question, MedBLIP provides a natural way of performing VAQ and presents the chain of thoughts. While MedBLIP may generate unsatisfactory answers to users' questions due to various reasons, including inaccurate medical knowledge from the LLM, activating the incorrect reasoning path, or lacking up-to-date information on new image content.

Zero-shot Medical VQA. We further assess the performance of our MedBLIP model on close-ended VQA utilizing the M3D-VQA-AD dataset, a subset of AD-related diseases filtered from the open 3D VQA dataset M3D-VQA [2]. In

Table 3. Zero-shot medical VQA results of our MedBLIP on M3D-VQA-AD close-ended dataset.

methods	Accuracy
M3D-LaMed [2]	72.88%
Ours	**77.96%**

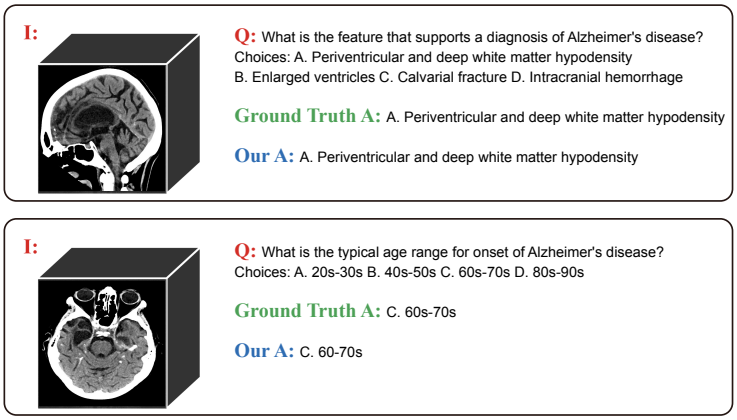

Fig. 4. Samples of zero-shot medical VQA results on the M3D-VQA-AD dataset.

closed-ended VQA, questions and choices serve as prompt inputs, while answers are used as supervision signals. The accuracy is calculated by directly comparing the answers with the choices. As shown in Table 3, our model significantly outperforms M3D-LaMed on AD-related closed-ended VQA tasks. Figure 4 visualizes samples of zero-shot medical VQA results on the M3D-VQA-AD dataset.

Ablation Study. We perform ablation studies from three aspects to answer the following three questions: (1) Why use a 2D pre-trained vision encoder instead of a trainable large vision encoder? (2) Will a prompt structure make a difference in the final CAD result? and (3) Why need the ITC loss between the image and diagnosis Q&A?

(1) Benefit of using a frozen 2D pre-trained vision encoder. To demonstrate the effectiveness of our lightweight image encoder based on the 2D pre-trained model, we take the query output embedding from MedQformer and compare it with features extracted from trainable ViT-G [41] from the initial weight of EVA-CLIP [12] on ADNI. We add a linear classification head with the cross-entropy loss. Table 4 reports that MedQFormer achieves slightly reduced performances, i.e., 0.6% lower than ViT-G in accuracy, but with much fewer parameters (only 15.1% of ViT-G's). This lightweight module benefits downstream tasks and allows building our model on language models and training it on one GPU. We also see that benefiting from this lightweight visual encoder, our MedBLIP outputs ViT-G by an improvement of 6.5% in the classification accuracy on ADNI.

Table 4. Comparison among a large vision encoder, MedQFormer, and MedBLIP on the ADNI dataset.

Visual features	#Params	Accuracy
ViT-G [12]	1B	72.2%
Our MedQFormer	**151M**	71.6%
Our MedBLIP w/ BioMedLM and LoRA	154M	**78.7%**

Table 5. Comparison between different prompt structures. (†: zero-shot)

Setting	ADNI -3x200	NACC -3x200	OASIS -2x200	AIBL†	MIRIAD†
Regular (I&T, Q, A)	78.7%	**83.3%**	**85.3%**	80.8%	**71.0%**
Alternative (Q, I&T, A)	**79.3%**(+0.6)	82.8%(-0.5)	82.5%(-2.8)	**82.8%**(+2.0)	70.8%(-0.2)

(2) Effect of using different prompt structures. To answer the second question above, we investigate the order of three prompting components, i.e., image and text features, the question, and the answer, and its effect on our model's performance. We treat the one with the question in the middle as the regular prompt structure and compare it to the one starting with the question. Table 5 shows that on some datasets our MedBLIP prefers the regular prompt, but this is not always the case. We conclude that the prompt strategy will not make a huge difference in the final performance of our model.

(3) Necessity of using two ITC loss functions. Besides the regular ITC loss between image and text pairs, we have another one between image and diagnosis Q&A, as presented in Eq. 4. Table 6 demonstrates that by adding the second ITC loss function, the classification accuracy improves on all datasets. This result is consistent with our motivation of adding the ITC loss between image and diagnosis Q&A, since it enforces the learnable queries to extract image features related to CAD.

5 Conclusion and Discussion

In this paper, we propose a novel CAD system MedBLIP that fuses medical multimodal data, i.e., image and text, from EHRs and shows its capability of

Table 6. Ablation study on loss functions. (†: zero-shot)

Loss Function	ADNI -3x200	NACC -3x200	OASIS -2x200	AIBL†	MIRIAD†
ITC(I,T)	71.7%	80.5%	82.5%	74.7%	66.8%
ITC(I,T) + ITC$(I,Q\&A)$	**78.7%** (+7.0)	**83.3%** (+2.8)	**85.3%** (+2.8)	**80.8%** (+6.1)	**71.0%** (+4.2)

performing zero-shot classification and medical VQA. Our MedBLIP introduces MedQFormer, a lightweight trainable 3D vision encoder that acts as a bridge between 3D medical images and a large frozen 2D vision encoder and as a bridge between 3D medical images and language models. Moreover, MedBLIP operates with low computational costs, as it smartly combines large pre-trained vision and language models with no need to train them from scratch or a large dataset in a specific medical domain. Our experiments demonstrate the effectiveness of our approach by outperforming several baselines, which sheds new light on further exploring medical multimodal CAD.

Acknowledgements

This research work was supported by the National Natural Science Foundation of China (NSFC) 62203303 and Shanghai Municipal Science and Technology Major Project 2021SHZDZX0102.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022)
2. Bai, F., Du, Y., Huang, T., Meng, M.Q.H., Zhao, B.: M3d: Advancing 3d medical image analysis with multi-modal large language models. arXiv preprint arXiv:2404.00578 (2024)
3. Bannur, S., Hyland, S., Liu, Q., Perez-Garcia, F., Ilse, M., Castro, D.C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., et al.: Learning to exploit temporal structure for biomedical vision-language processing. arXiv preprint arXiv:2301.04558 (2023)
4. Bao, H., Wang, W., Dong, L., Liu, Q., Mohammed, O.K., Aggarwal, K., Som, S., Piao, S., Wei, F.: Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. Advances in Neural Information Processing Systems **35**, 32897–32912 (2022)
5. Beekly, D.L., Ramos, E.M., Lee, W.W., Deitrich, W.D., Jacka, M.E., Wu, J., Hubbard, J.L., Koepsell, T.D., Morris, J.C., Kukull, W.A., et al.: The national alzheimer's coordinating center (nacc) database: the uniform data set. Alzheimer Disease & Associated Disorders **21**(3), 249–258 (2007)
6. Chen, J., Guo, H., Yi, K., Li, B., Elhoseiny, M.: Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18030–18040 (2022)
7. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311 (2022)
8. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022)

9. Dodge, J., Sap, M., Marasovic, A., Agnew, W., Ilharco, G., Groeneveld, D., Gardner, M.: Documenting the english colossal clean crawled corpus. arXiv preprint arXiv:2104.08758 (2021)
10. Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al.: Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378 (2023)
11. Ellis, K.A., Bush, A.I., Darby, D., De Fazio, D., Foster, J., Hudson, P., Lautenschlager, N.T., Lenzo, N., Martins, R.N., Maruff, P., et al.: The australian imaging, biomarkers and lifestyle (aibl) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of alzheimer's disease. International psychogeriatrics **21**(4), 672–687 (2009)
12. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva: Exploring the limits of masked visual representation learning at scale. arXiv preprint arXiv:2211.07636 (2022)
13. Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al.: The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027 (2020)
14. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
15. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
16. Jain, S.M.: Hugging face. In: Introduction to Transformers for NLP: With the Hugging Face Library and Models to Solve Problems, pp. 51–67. Springer (2022)
17. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning. pp. 4904–4916. PMLR (2021)
18. Koh, J.Y., Salakhutdinov, R., Fried, D.: Grounding language models to images for multimodal generation. arXiv preprint arXiv:2301.13823 (2023)
19. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
20. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
21. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021)
22. Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., Yan, J.: Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. arXiv preprint arXiv:2110.05208 (2021)
23. Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., Liu, T.Y.: Biogpt: generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics **23**(6) (2022)
24. Malone, I.B., Cash, D., Ridgway, G.R., MacManus, D.G., Ourselin, S., Fox, N.C., Schott, J.M.: Miriad—public release of a multiple time point alzheimer's mr imaging dataset. NeuroImage **70**, 33–36 (2013)

25. Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. Journal of cognitive neuroscience **19**(9), 1498–1507 (2007)
26. Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E.P., Rajpurkar, P.: Med-flamingo: a multimodal medical few-shot learner. In: Machine Learning for Health (ML4H). pp. 353–367. PMLR (2023)
27. OpenAI: Gpt-4 technical report (2023)
28. Petersen, R.C., Aisen, P.S., Beckett, L.A., Donohue, M.C., Gamst, A.C., Harvey, D.J., Jack, C.R., Jagust, W.J., Shaw, L.M., Toga, A.W., et al.: Alzheimer's disease neuroimaging initiative (adni): clinical characterization. Neurology **74**(3), 201–209 (2010)
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
30. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al.: Large language models encode clinical knowledge. Nature **620**(7972), 172–180 (2023)
31. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al.: Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617 (2023)
32. van Sonsbeek, T., Derakhshani, M.M., Najdenkoska, I., Snoek, C.G., Worring, M.: Open-ended medical visual question answering through prefix tuning of language models. arXiv preprint arXiv:2303.05977 (2023)
33. Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P.: Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. Nature Biomedical Engineering pp. 1–8 (2022)
34. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
35. Tsimpoukelli, M., Menick, J.L., Cabi, S., Eslami, S., Vinyals, O., Hill, F.: Multimodal few-shot learning with frozen language models. Advances in Neural Information Processing Systems **34**, 200–212 (2021)
36. Venigalla, A., Frankle, J., Carbin, M.: Biomedlm: a domain-specific large language model for biomedical text. MosaicML. Accessed: Dec **23** (2022)
37. Wang, S., Zhao, Z., Ouyang, X., Wang, Q., Shen, D.: Chatcad: Interactive computer-aided diagnosis on medical image using large language models. arXiv preprint arXiv:2302.07257 (2023)
38. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 (2022)
39. Xu, M., Zhang, Z., Wei, F., Lin, Y., Cao, Y., Hu, H., Bai, X.: A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. arXiv preprint arXiv:2112.14757 (2021)
40. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022)
41. Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L.: Scaling vision transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12104–12113 (2022)

42. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022)