Adapting BLIP to Medical VQA with Domain Pretraining and Self-Distillation

Francesco Palma
Politecnico di Milano
francesco1.palma@mail.polimi.it

Sara Zappia
Politecnico di Milano
sara.zappia@mail.polimi.it

Abstract

Adapting general vision–language models (VLMs) to clinical domains is challenging due to the domain gap between web and medical imaging data. We propose a lightweight adaptation method for BLIP, a widely used web captions pretrained VLM, on the medical visual question answering (VQA) task. Our approach combines domain-adaptive pretraining on the ROCOv2 radiology image–caption corpus with progressive self-distillation, where an exponential moving average (EMA) teacher stabilizes learning and mitigates catastrophic forgetting. Without major architectural changes or reliance on large-scale medical corpora, this pipeline improves BLIP's overall exact-match accuracy by 8% (74.2% vs 66.2%) on VQA-RAD compared to standard fine-tuning. These findings suggest that carefully designed adaptation strategies can improve the performance of general-purpose VLMs in clinical VQA settings.

1. Introduction

Visual Question Answering combines both image understanding and natural language reasoning to produce reliable answers. In medical applications, for example radiology, VQA models can assist clinicians, reducing workloads and improving decision support. However, deploying general-purpose VLMs like BLIP in healthcare is difficult due to the domain shift between web-scale training data and specialized medical datasets. This often degrades performance when models are fine-tuned on small medical datasets like VQA-RAD. Previous research introduced specialized architectures or resource-intensive pretraining pipelines, which require substantial computational resources and annotated data. An efficient adaptation technique that uses strong general VLMs without architectural changes would therefore be valuable for practical clinical deployment. We show that BLIP—a general-purpose VLM—can be effectively adapted for clinical VQA through targeted domain adaptation and training stabilization. Instead of proposing new architectures, we focus on practical strategies for medical domain transfer. Our contributions are:

- **Baseline establishment:** We provide a reference performance benchmark by fine-tuning BLIP on VQA-RAD, quantifying its out-of-the-box capability on clinical VQA.
- **Targeted domain adaptation:** We perform domain-adaptive pretraining on ROCOv2 using BLIP's original objectives to bridge the domain gap between web and medical data.
- **Stabilization via self-distillation:** We introduce progressive online self-distillation with an exponential moving average teacher to prevent catastrophic forgetting during adaptation.

2. Related work

2.1. General Vision–Language Models

Vision–language models (VLMs) [1] learn aligned representations across visual and textual modalities. Contrastive approaches like CLIP [7] uses large-scale image–text pairs to enable robust zero-shot classification, while generative models like BLIP [5] combine image–text contrastive (ITC), image–text matching (ITM), and language modeling (LM) objectives for both understanding and generation tasks. These models achieve state-of-the-art performance on web-based benchmarks but struggle with domain shift when applied to specialized domains like radiology. The gap between natural images and clinical data motivates domain adaptation strategies to transfer these models to healthcare applications.

2.2. Medical Visual Question Answering

Medical VQA [6] extends vision–language reasoning to clinical tasks, with benchmarks like VQA-RAD [4] providing radiology question–answer pairs and medical corpus such as ROCOv2 [8] offering image–caption pairs for adaptation. While domain-adaptive pretraining [2] is common in medical VQA, recent work has focused enhancing pretraining objectives and modalities rather than simply adapting general models.

2.3. Model Adaptation and Self-Distillation

Domain-adaptive pretraining (DAPT) has proven effective for bridging distribution gaps in both NLP [2] and vision–language domains. By continuing pretraining on in-domain data, models can better align with specialized vocabularies and visual characteristics. Knowledge distillation [3] and its online variants, particularly exponential moving average (EMA) teacher methods [9], have demonstrated success in stabilizing training across various domains.

3. Proposed approach

3.1. Overview

We propose a two-stage pipeline for preparing BLIP to medical VQA. The first stage involves domain-adaptive pretraining (DAPT) on a large-scale radiology image-caption dataset (ROCOv2) to align the model with the medical domain. Our DAPT stage employs a self-distillation framework with a teacher-student setup, where the teacher model is an exponential moving average (EMA) of the student model. This framework leverages multi-task learning, combining image-text contrastive learning, image-text matching, and language modeling objectives. The self-distillation mechanism adaptively weights the contribution of each task based on the student's performance relative to the teacher, ensuring balanced and stable training. The second stage involves fine-tuning the domain-adapted model on the VQA-RAD dataset for the VQA task. This approach aims to mitigate the domain shift between natural images and radiology images and to leverage the rich visual and textual information in the radiology domain.

3.2. Base Model Architecture

We build upon BLIP (Bootstrapping Language-Image Pre-training) [5], a unified vision-language framework that excels in both understanding and generation tasks through its multimodal mixture of encoder-decoder architecture. BLIP employs a Vision Transformer (ViT-B/16) visual encoder that processes 224×224 images into patch embeddings. The model dynamically operates in three modes through parameter sharing:

- **Unimodal Encoder**: ViT for image encoding, BERT for text encoding
- **Image-Grounded Text Encoder**: Adds cross-attention layers for multimodal fusion
- **Image-Grounded Text Decoder**: Causal self-attention for autoregressive generation

The architecture shares embeddings and feed-forward networks across components while maintaining separate self-attention mechanisms, achieving optimal parameter efficiency (224M parameters). This enables flexible switching between VQA modes: understanding-based classification and generation-based answering. We initialize from BLIP's checkpoint pre-trained on 129M web image-text pairs using captioning and filtering (CapFilt) methodology, providing robust vision-language grounding for medical domain adaptation.

3.3. DAPT with Progressive Self Distillation

3.3.1 Medical Dataset

We use the ROCOv2 (Radiology Objects in Context) corpus [8], comprising 80k radiology image-caption pairs. To ensure domain relevance, we filter for primary radiology modalities (CT, MRI, X-ray) using regular expression pattern matching, yielding 45k high-quality medical image-text pairs. This curated dataset provides the domain-specific supervision essential for effective medical adaptation while maintaining computational efficiency through token-length filtering (maximum 64 tokens per caption).

3.3.2 Pretraining Objectives

Our DAPT employs BLIP's three complementary objectives, optimized for medical imaging through targeted adaptations. The mathematical formulations are summarized in Figure 1.

$$\mathcal{L}_{ITC} = \frac{1}{2}\left(\mathcal{L}_{I\text{ß}T} + \mathcal{L}_{T\text{ß}I}\right)$$

$$\mathcal{L}_{ITM} = \mathbb{E}_{(I,T)}\left[y \log p + (1-y)\log(1-p)\right]$$

$$\mathcal{L}_{LM} = -\mathbb{E}_{(I,T)} \sum_{t=1}^{L} \log P(w_t | w_{<t}, I; \theta)$$

$$\mathcal{L}_{DAPT} = \mathcal{L}_{ITC} + \mathcal{L}_{ITM} + \mathcal{L}_{LM}$$

Figure 1. Mathematical formulation of domain-adaptive pretraining objectives.

Objective Specifications:

- **Image-Text Contrastive (ITC):** Aligns representations through bidirectional contrastive learning using cosine similarity between image and text [CLS] tokens with temperature scaling $\tau = 0.07$.
- **Image-Text Matching (ITM):** Performs fine-grained alignment via binary classification. Employs hard negative mining by selecting the most challenging negative samples based on ITC similarity scores.
- **Language Modeling (LM):** Enables conditional text generation through autoregressive prediction, training the model to generate medically accurate descriptions.

Our DAPT approach preserves the core BLIP methodology—architecture and three-task formulation—while introducing specific adaptations. These include removal of momentum encoders for ITC simplification, clinically-focused negative mining, and medical-optimized hyperparameters. This balanced approach maintains BLIP's architectural advantages while specializing for radiology applications through $\mathcal{L}_{DAPT}$.

3.4. Progressive Self-Distillation Framework

To enhance medical domain adaptation, we propose a progressive self-distillation (PSD) framework that dynamically balances learning across the three pre-training objectives—Image-Text Contrastive (ITC), Image-Text Matching (ITM), and Language Modeling (LM)—through adaptive teacher-student knowledge transfer.

3.4.1 Teacher-Student Architecture

We maintain an exponential moving average (EMA) teacher model that provides stable training targets while gradually incorporating student improvements. The teacher parameters θ_{teacher} are updated from the student parameters θ_{student} at each optimization step:

$$\theta_{\text{teacher}} \leftarrow \tau \cdot \theta_{\text{teacher}} + (1-\tau) \cdot \theta_{\text{student}} \tag{1}$$

where τ is a momentum parameter that progressively increases from 0.95 to 0.995 during training, ensuring stable initialization followed by gradual teacher refinement.

3.4.2 Adaptive Multi-Task Distillation

Distillation losses are dinamically weighted based on relative task difficulty. For each objective $k \in \{\text{ITC}, \text{ITM}, \text{LM}\}$, we compute the performance ratio:

$$r_k = \frac{\mathcal{L}_k^{\text{student}}}{\mathcal{L}_k^{\text{teacher}} + \epsilon} \tag{2}$$

where $\epsilon = 10^{-8}$ prevents numerical instability. Higher ratios indicate tasks where the student struggles relative to the teacher. These ratios are smoothed using exponential moving average:

$$\bar{r}_k \leftarrow \alpha \cdot \bar{r}_k + (1-\alpha) \cdot r_k \tag{3}$$

with $\alpha = 0.9$. The smoothed ratios are clamped to $[0.1, 5.0]$ to prevent extreme values and normalized via softmax to obtain final weights:

$$w_k = \frac{\exp(\bar{r}_k)}{\sum_{j \in \{\text{ITC,ITM,LM}\}} \exp(\bar{r}_j)} \tag{4}$$

3.4.3 Distillation Loss Formulation

We employ mean squared error between student and teacher task losses, weighted by the adaptive coefficients:

$$\mathcal{L}_{\text{distill}}^{\text{ITC}} = \text{MSE}(\mathcal{L}_{\text{ITC}}^{\text{student}}, \mathcal{L}_{\text{ITC}}^{\text{teacher}}) \cdot w_{\text{ITC}} \tag{5}$$

$$\mathcal{L}_{\text{distill}}^{\text{ITM}} = \text{MSE}(\mathcal{L}_{\text{ITM}}^{\text{student}}, \mathcal{L}_{\text{ITM}}^{\text{teacher}}) \cdot w_{\text{ITM}} \tag{6}$$

$$\mathcal{L}_{\text{distill}}^{\text{LM}} = \text{MSE}(\mathcal{L}_{\text{LM}}^{\text{student}}, \mathcal{L}_{\text{LM}}^{\text{teacher}}) \cdot w_{\text{LM}} \tag{7}$$

3.4.4 Progressive Training Strategy

The complete training objective combines the pre-training losses with adaptive distillation:

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{\text{ITC}} + \mathcal{L}_{\text{ITM}} + \mathcal{L}_{\text{LM}}}_{\text{pre-training objectives}} + \underbrace{\mathcal{L}_{\text{distill}}^{\text{ITC}} + \mathcal{L}_{\text{distill}}^{\text{ITM}} + \mathcal{L}_{\text{distill}}^{\text{LM}}}_{\text{adaptive distillation}} \tag{8}$$

We employ gradient accumulation (8 steps) and a progressive learning rate schedule with linear warmup from 5×10^{-9} to 4×10^{-7} followed by cosine decay to 1×10^{-9}.

This framework enables focused training on challenging objectives where the student model requires additional guidance, providing more effective medical domain adaptation than fixed-weight multi-task learning approaches. The progressive nature of both the teacher momentum and learning rate ensures stable optimization throughout the domain adaptation process.

3.5. Task-Specific Fine-Tuning

Following domain-adaptive pre-training, we specialize the model for medical visual question answering through targeted fine-tuning on the VQA-RAD benchmark. This stage transforms the domain-adapted vision-language model into a clinical decision support system capable of answering diverse medical questions about radiology images.

3.5.1 Dataset and Experimental Setup

We utilize the VQA-RAD dataset [4] comprising 3,515 question-answer pairs across 315 radiology images. The dataset exhibits balanced clinical characteristics with 57.8% closed-ended and 42.2% open-ended questions, covering abnormality detection, presence queries, counting tasks, and anatomical localization. We employ stratified sampling to preserve question type and answer category distributions across 70%/10%/20% training, validation, and test splits.

3.5.2 Fine-Tuning Methodology

The fine-tuning process optimizes the language modeling objective for medical answer generation:

$$\mathcal{L}_{\text{VQA}} = -\sum_{t=1}^{L} \log P(w_t | w_{<t}, I, Q) \tag{9}$$

where I represents the input medical image, Q is the clinical question, and w_t denotes the t-th token in the answer sequence of length L.

To prevent overfitting on limited medical data while leveraging domain-adapted representations, we employ selective parameter freezing:

$$\theta_{\text{visual}}^{(t)} = \theta_{\text{visual}}^{(0)} \quad \forall t \tag{10}$$

This strategy preserves the medical domain visual features learned during DAPT while specializing the textual components for clinical question answering.

3.5.3 Optimization and Regularization

We fine-tune using AdamW optimizer with learning rate 1×10^{-5}, weight decay 0.01, and gradient clipping at norm 1.0. The training employs cosine annealing with minimum learning rate 1×10^{-6} over 15 epochs. Early stopping with patience of 4 epochs prevents overfitting, and batch size of 8 ensures stable optimization given computational constraints.

3.5.4 Clinical Adaptation Benefits

This fine-tuning approach provides several clinical advantages:

- **Preserved Domain Knowledge**: Frozen visual encoder maintains radiology-specific feature representations
- **Specialized Language Generation**: Textual components adapt to medical terminology and answer patterns
- **Computational Efficiency**: Reduced parameter updates enable faster convergence on limited medical data
- **Robust Evaluation**: Comprehensive assessment across closed and open-ended clinical questions

The fine-tuning stage represents the final specialization step, transforming a general vision-language model into a clinically relevant tool for radiology question answering while maintaining the domain-adapted benefits from pre-training.

4. Experiments

4.1. Experimental Setup

We conduct three primary experiments to evaluate our domain adaptation with progressive self-distillation framework:

- **Baseline**: Standard BLIP model fine-tuned directly on VQA-RAD
- **DAPT**: Domain-adaptive pre-training on ROCO followed by VQA-RAD fine-tuning
- **DAPT + Self-Distillation**: Our full approach with adaptive multi-task distillation followed by VQA-RAD fine-tuning

All models use ViT-B/16 architecture with identical fine-tuning hyperparameters: learning rate 1×10^{-5}, batch size 8, weight decay 0.01, and cosine annealing over 15 epochs. The visual encoder remains frozen during fine-tuning to preserve domain-adapted representations. We evaluate using exact match accuracy and normalized edit similarity, reporting results separately for closed and open-ended questions to capture different clinical reasoning capabilities.

4.1.1 Main Results

Our progressive self-distillation framework demonstrates significant improvements across all metrics (Table 1). Domain-adaptive pre-training alone provides a 6.1% relative improvement in overall accuracy (0.662 → 0.723), with the full approach yielding 8.0% improvement over baseline (0.662 → 0.742). The gains are particularly pronounced for open-ended questions, where our method achieves 10.5% relative improvement in accuracy (0.492 → 0.597), indicating enhanced capability for complex clinical reasoning.

5. Conclusion

We presented a progressive self-distillation framework for adapting vision-language models to medical visual question answering. Our method combines domain-adaptive pre-training on medical image-caption pairs with adaptive multi-task distillation, followed by targeted fine-tuning on the VQA-RAD benchmark.

While direct numerical comparison with published VQA-RAD results is challenging due to variations in evaluation protocols and dataset splits, our primary contribution lies in demonstrating consistent relative improvements through our adaptation pipeline.

Key results demonstrate improvements over baseline approaches: 8.0% overall accuracy gain (74.2% vs 66.2%) and 10.5% improvement on open-ended questions (59.7%

Table 1. Performance Comparison on VQA-RAD Test Set

Method	**Overall Acc**	**Closed Acc**	**Open Acc**	**Overall Edit**	**Closed Edit**	**Open Edit**
Baseline	0.662	0.782	0.492	0.754	0.792	0.700
DAPT	0.723	0.815	0.592	0.806	0.825	0.778
DAPT + Self-Distill	**0.742**	**0.845**	**0.597**	**0.828**	**0.854**	**0.792**

vs 49.2%), highlighting enhanced clinical reasoning capabilities. The progressive distillation strategy effectively balances learning across vision-language objectives while preserving domain-specific representations.

Future work will address remaining challenges in complex clinical reasoning and rare pathology recognition, with potential extensions to multi-modal clinical contexts and uncertainty-aware predictions for safe clinical deployment.

References

[1] F. Bordes, R. Y. Pang, A. Ajay, A. C. Li, A. Bardes, S. Petryk, O. Mañas, Z. Lin, A. Mahmoud, B. Jayaraman, M. Ibrahim, M. Hall, Y. Xiong, J. Lebensold, C. Ross, S. Jayakumar, C. Guo, D. Bouchacourt, H. Al-Tahan, K. Padthe, V. Sharma, H. Xu, X. E. Tan, M. Richards, S. Lavoie, P. Astolfi, R. A. Hemmat, J. Chen, K. Tirumala, R. Assouel, M. Moayeri, A. Talattof, K. Chaudhuri, Z. Liu, X. Chen, Q. Garrido, K. Ullrich, A. Agrawal, K. Saenko, A. Celikyilmaz, and V. Chandra. An introduction to vision-language modeling, 2024. 1

[2] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith. Don't stop pretraining: Adapt language models to domains and tasks, 2020. 1, 2

[3] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network, 2015. 2

[4] J. Lau, S. Gayen, A. Ben Abacha, and D. Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5:180251, 2018. 1, 3

[5] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 1, 2

[6] Z. Lin, D. Zhang, Q. Tao, D. Shi, G. Haffari, Q. Wu, M. He, and Z. Ge. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, 143:102611, Sept. 2023. 1

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

[8] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. B. Abacha, A. G. Seco de Herrera, H. Müller, P. A. Horn, F. Nensa, and C. M. Friedrich. Rocov2: Radiology objects in context version 2, an updated multimodal image dataset. *Scientific Data*, 11(1), June 2024. 1, 2

[9] A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, 2018. 2