Vision Language Model (VLM)-based Alzheimer's Disease Detection

Muhammad Ahmad Ashraf **Umer Raja**

Abstract

This paper presents a Vision-Language Model (VLM)-based approach for classifying Alzheimer's Disease (AD) in both binary and multi-class settings. We use BioMedCLIP as the baseline model and extend it by incorporating a dedicated classification head and domain-specific textual prompts, fine-tuned using the OASIS-I MRI dataset. Each MRI image is paired with a custom caption describing key neuroanatomical features and pathological indicators to enhance multimodal fusion and model robustness. To improve interpretability, we integrate Eigen-CAM, which generates activation heatmaps highlighting brain regions that influence the model's predictions, thereby increasing transparency and supporting clinical decision-making. In the multi-class setting, the fine-tuned BioMedCLIP model with a ViT/B-16 vision backbone achieved the highest accuracy of 85.27% and an AUC of 0.9654, while in the binary classification task, it achieved an average accuracy of 93.14% and AUC of 0.9751, outperforming the baseline models by a large margin. Our codebase can be found at: https://github.com/UmerSR/CS-437-Final-Project/

1. Introduction

1.1. Background and Motivation

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder characterized by cognitive decline and brain atrophy. Early diagnosis is crucial for patient care, and neuroimaging (especially MRI) has become a key tool for detecting AD-related changes. In recent years, deep learning models (e.g., convolutional neural networks and autoencoders) have achieved promising accuracy (often around 85–90% in distinguishing AD patients from healthy controls) using MRI data [1]. However, these vision-only models process images in isolation and act as "black boxes," offering little interpretability into their decisions. This lack of interpretability is a known barrier to clinical adoption of AI, as physicians are less likely to trust models without transparent reasoning [2]. There is a clear need for approaches that maintain high diagnostic accuracy and provide insights into why a prediction was made.

Vision-Language Models (VLMs) represent the state-of-the-art in deep learning by combining advances in computer vision and natural language processing. VLMs are fundamentally multimodal, meaning they simultaneously process and integrate visual and textual data. Models like CLIP align images with text in a shared feature space, using Transformer-based architectures to fuse information from both modalities [3]. By leveraging both image content and descriptive language, VLMs capture richer contextual information and produce more interpretable outputs than vision-only models. These models have demonstrated remarkable capabilities across domains—from captioning images to cross-modal retrieval—indicating their potential as general-purpose foundation models.

In healthcare, multimodal learning is emerging as an important paradigm. VLMs have been applied to tasks like generating radiology reports from X-ray images and medical visual question answering [3]. Despite these successes, there remains a significant gap in applying VLMs to neurodegenerative diseases like AD. Brain MRI datasets for AD typically lack rich annotated captions or reports required for training, and subtle anatomical changes in AD can be difficult to describe textually. Chiumento et al. highlighted the shortage of comprehensive image-text datasets for neurodegenerative conditions [4]. This limitation motivates our work, which aims to harness VLMs for AD diagnosis, hypothesizing that integrating visual and textual information improves both accuracy and explainability.

1.2. Recent Vision-Language Approaches

Several studies have begun exploring VLM-based methods for Alzheimer's diagnosis. Lee et al. introduced a graph-based VLM approach linking brain MRI regions with corresponding text descriptions, achieving 88.7% accuracy on the ADReSSo dataset and providing interpretability via linguistic markers [5]. Can et al. proposed VisTA, fine-tuning BioMedCLIP on MRI-text pairs using contrastive learning, achieving approximately 88% accuracy and generating clinician-aligned explanations [6].

Chiumento and Liu addressed textual annotation scarcity by generating synthetic radiology reports using GPT-4, fine-

tuning a VLM on these synthetic captions to generate diagnostic summaries directly from MRIs, achieving promising preliminary results [4]. Feng et al. integrated MRI images with textual and tabular patient data via a large language model (LLM), achieving state-of-the-art performance on the ADNI dataset and highlighting the potential of language models in multimodal diagnostic frameworks [7].

Overall, these studies underscore the potential of multimodal learning in improving AD detection, yet also highlight significant data-related challenges and the necessity of carefully curated training datasets.

1.3. Research Objectives

Significant challenges remain in adapting general-purpose VLMs to MRI-based AD detection due to limited textual annotations specific to brain imaging and domain-specific nuances. Our project aims to address these challenges by fine-tuning BioMedCLIP, a biomedical VLM trained on 15 million image–text pairs [8]. We extend BioMedCLIP by adding a dedicated classification head and incorporating domain-specific captions during fine-tuning. Each MRI image is paired with a custom caption highlighting key neuroanatomical features or pathological changes. This multimodal fusion aims to improve the robustness and explainability of the AD detection task.

Additionally, we prioritize model interpretability by integrating Eigen-CAM, a variant of class activation mapping. Eigen-CAM computes principal components of convolutional feature maps, producing robust activation heatmaps highlighting brain regions influential in the model's predictions [9]. This method provides visual explanations, supporting clinical decision-making by making the model's reasoning transparent.

2. Methodology

2.1. Database Selection

This study employs the *processed OASIS-I dataset*, publicly available via Kaggle[1], for the task of Alzheimer's disease (AD) prediction. The dataset comprises approximately **80,000 T1-weighted MRI brain scans** collected from around **400 patients**, with each subject represented by **60 axial 2D slices** extracted from volumetric 3D MRI data. Each image is categorized into one of **five diagnostic classes** reflecting the severity of cognitive decline:

These class labels are based on clinical assessments and represent progressive stages of Alzheimer's disease.

A notable advantage of this dataset is its preprocessing: the original 3D *NiFTi (Neuroimaging Informatics Technology Initiative)* format has been converted into standard 2D image slices. This significantly reduces the preprocessing burden and enables direct integration with *Vision-Language Models (VLMs)* that operate on 2D image inputs.

[1] https://www.kaggle.com/datasets/ninadaithal/imagesoasis

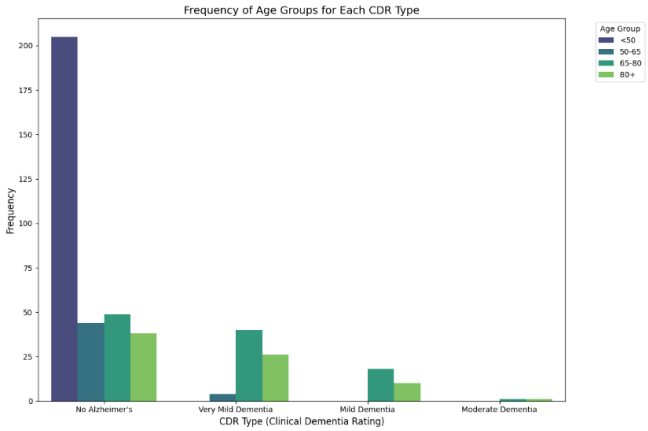

Figure 1. Graph showing frequency of age groups for each CDR type to assess class breakdown with respect to demography

To ensure consistency and reduce inter-slice variability, we selected only a single representative slice per patient for training and evaluation. Specifically, the **middle axial slice (slice index 130)** was used, as it provides a balanced anatomical view including key regions implicated in AD such as the *hippocampus*, *entorhinal cortex*, and *ventricular system*. These structures are known to exhibit early pathological changes in Alzheimer's disease, making this slice particularly informative.

Each MRI volume in the dataset is available in four variations: **MPR1**, **MPR2**, **MPR3**, and **MPR4**, which correspond to different acquisition protocols or preprocessing pipelines. To maintain consistency and reduce variability from heterogeneous imaging parameters, only the **MPR1** images were utilized across all subjects in this study.

A significant challenge with the dataset is the presence of **class imbalance**, especially in the more severe AD stages (as illustrated in Figure 1). To address this, we implemented a targeted augmentation strategy: for underrepresented classes, we included slices adjacent to the middle slice—namely, **slices 129 and 131**, and expanded further outward as needed. This approach preserved anatomical relevance while enhancing class balance, thereby improving model training stability and generalizability.

2.2. Caption Generation

Each image used in this study was paired with a corresponding textual caption using demographic and diagnostic information provided in the OASIS-I metadata file, avail-

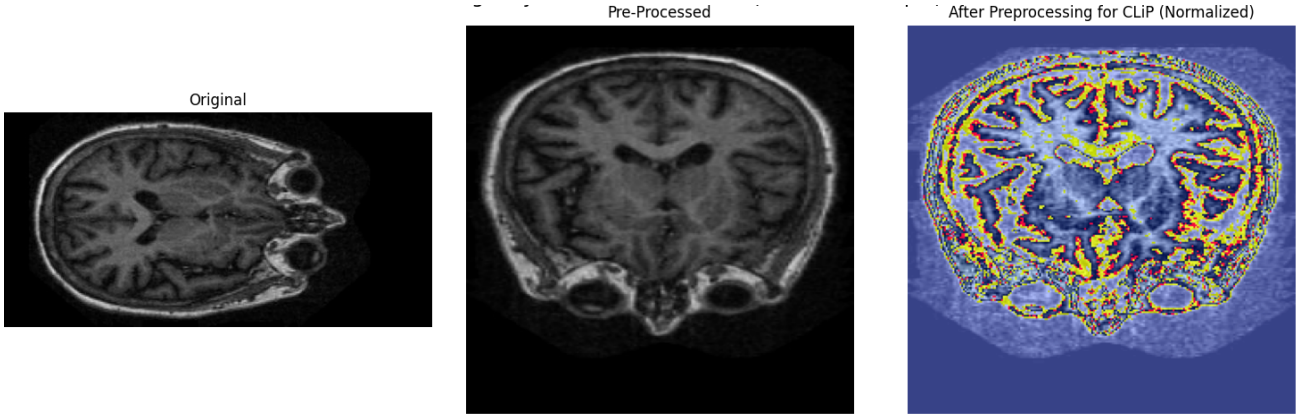

Figure 2. Stages of the preprocessing pipeline of MRI scans for BioMedCLIP input

able via Kaggle[2]. This metadata includes several clinically relevant variables such as **Age**, **Gender**, **Handedness**, along with cognitive and anatomical metrics including **MMSE**, **eTIV**, **nWBV**, and **ASF**. Each MRI slice was associated with a unique patient identifier, allowing precise alignment between the image and its corresponding metadata. This enabled the generation of structured, individualized captions containing both demographic and clinical data for use in training and evaluation of VLMs.

The primary diagnostic label was derived from the **CDR (Clinical Dementia Rating)**, a widely used clinical scale to assess the severity of dementia:

- **0** – Non-Demented
- **0.5** – Very Mild Dementia
- **1** – Mild Dementia
- **2** – Moderate Dementia

The description of the cognitive and anatomical metrics used in the caption are defined as follows:

- **MMSE (Mini-Mental State Examination):** A standardized 30-point questionnaire used to measure cognitive impairment. Scores between 24–30 indicate normal cognition, while scores below 24 are associated with various levels of cognitive decline.
- **eTIV (Estimated Total Intracranial Volume):** This metric estimates the overall cranial volume (typically ranging from 1400–1600 cm^3 in healthy adults) and is useful for normalizing brain volume measures.
- **nWBV (Normalized Whole Brain Volume):** Represents the proportion of brain volume relative to eTIV. Healthy individuals usually exhibit values between 0.70 and 0.85; lower values may indicate brain atrophy, a hallmark of Alzheimer's disease.
- **ASF (Atlas Scaling Factor):** Indicates the degree of transformation required to align the subject's brain to a standard template. A value near 1.0 suggests minimal scaling; deviations may reflect anatomical differences or registration issues.

A standardized structure for the captions was devised, with an example shown below for clarity:

```
Patient ID: OAS1_0033_MR1,
Gender: F, Handedness: R,
Age: 80 years, Education: 4.0,
Socioeconomic Status: 2.0,
MMSE: 29.0 <MMSE Description>,
eTIV: 1323 <eTIV Description>,
nWBV: 0.735 <nWBV Description>,
ASF: 1.326 <ASF Description>
```

This caption structure provides a comprehensive representation of each subject, combining quantitative and categorical features that are potentially predictive of dementia severity. These captions were used as the textual modality input for training and evaluating the VLM. It is important to note that the same caption was assigned to all image slices corresponding to a given patient, ensuring consistency across multiple views of the same subject.

2.3. Baseline Model

The baseline model employed in this study is BioMedCLIP, a Vision-Language Model (VLM) fine-tuned on over 15

[2] https://www.kaggle.com/datasets/ninadaithal/oasis-1-shinohara/

million biomedical image-text pairs. This large-scale pretraining enables the model to learn robust multimodal representations that capture subtle semantic associations between medical imagery and textual descriptions, necessary for Alzheimer's disease detection.

Image inputs were prepared to align with the BioMedCLIP model's expected input using the following algorithm:

Algorithm 1 Image Preprocessing for BioMedCLIP

Require: Original MRI slice image I
Ensure: Preprocessed image suitable for BioMedCLIP
Stage 1: Custom Preprocessing
Resize image I to 224×224 pixels using bilinear interpolation
Apply rotation to I to ensure correct anatomical orientation
Store intermediate result as I'
Stage 2: CLIP Preprocessing
Normalize I' using BioMedCLIP's preprocessing function (normalization etc.)
Convert normalized image to tensor format compatible with model input
Assign output as $\hat{I}$
return $\hat{I}$

During training and inference, each input image was paired with its corresponding caption as previously curated, along with a label prompt that functioned as the classification anchor. Specifically, the model received concatenated inputs in the form:

```
<Caption>. This is a photo of:
<Label>.
```

The model's performance was evaluated by computing the cosine similarity between the image embeddings and the text embeddings for each class prompt. The predicted class was assigned based on the label with the highest similarity score, effectively framing classification as a retrieval task in the shared embedding space.

2.4. First Improvement

The first improvement implemented involved fine-tuning the BioMedCLIP model on the curated Alzheimer's dataset for binary Alzheimer's disease detection. Due to resource constraints, a lightweight fine-tuning strategy was adopted. Specifically, all parameters of the model were frozen except for the final two transformer blocks of the visual encoder and the projection head. This selective unfreezing allowed the model to adapt to the domain-specific features of our dataset while minimizing memory and compute overhead.

To further enhance the model's classification capability, a dedicated classification head was introduced. After encoding the input image and its corresponding caption using BioMedCLIP, the resulting image and text embeddings (both 512-dimensional vectors) were concatenated and passed through a lightweight feedforward neural network to produce label logits. The rationale behind this concatenation was to enable both modalities—visual and textual—to jointly inform the prediction. This multimodal fusion not only improves classification accuracy but also allows the text to act as a semantic regularizer, encouraging the model to learn more meaningful and robust representations from both inputs.

This network produced raw label logits, which were then passed through a softmax function to yield class probabilities. This architecture allowed the model to leverage both visual and textual context for robust classification. The network remained consistent throughout the following improvements and is shown below:

```
FFN(x) = Linear(1024, 256)
→ ReLU()
→ Dropout(0.3)
→ Linear(256, N),
```

where N denotes the number of target classes (e.g., 2 for binary classification or 5 for multiclass classification).

Algorithm 2 Classification Pipeline for BioMedCLIP

Require: Preprocessed image I, curated caption T
Ensure: Predicted label $\hat{y}$
Encode caption T using BioMedCLIP text encoder to obtain text embedding E_T
Encode image I using BioMedCLIP visual encoder to obtain image embedding E_I
Concatenate embeddings: $E = \texttt{torch.cat}([E_I, E_T])$ along dimension 1
Pass E through a feedforward neural network (FFN) to obtain raw logits z
Apply softmax (for multiclass) or sigmoid (for binary classification) to z to compute class probabilities p
Assign and return predicted label: $\hat{y} = \arg\max(p)$

2.5. Second Improvement

The second improvement extended the previous experiment by shifting the focus to multiclass classification and evaluating whether the vision component of the Vision-Language Model (VLM) could effectively capture class-specific features associated with different stages of Alzheimer's disease. This setup aimed to test the discriminative capability of the visual encoder in separating fine-grained categories beyond the binary setup.

To explore the effect of different visual backbones, we con-

ducted a comparative analysis by replacing the original ViT-B/16 visual encoder in BioMedCLIP with an alternative convolutional neural network architecture—VGG-16. The motivation behind this substitution lies in the interpretability and proven performance of convolutional models like VGG-16, which are known to capture localized and hierarchical visual features. While transformer-based encoders such as ViT leverage global self-attention, convolutional models often preserve spatial hierarchies more effectively, which could be beneficial in medical imaging tasks where local features (e.g., hippocampal atrophy) play a critical role.

Since the output of VGG-16 does not natively match the 512-dimensional embedding space of ViT-B/16, a projection head was introduced to bridge this gap. Specifically, the VGG feature extractor was followed by a small fully connected projection module that maps the high-dimensional convolutional output to a 512-dimensional vector. This ensured compatibility with the rest of the BioMedCLIP architecture, enabling downstream modules—including the text encoder and classification head—to process image embeddings in a consistent representation space.

The modified architecture followed the same fine-tuning strategy as before. For the VGG-16 variant, all layers were frozen except for the deeper convolutional layers starting from `vgg.features[24:]`, which correspond to the final convolutional block. This selective unfreezing allowed the model to adapt higher-level features specific to Alzheimer's classification while preserving the general representations learned from ImageNet pretraining.

2.6. Experimental Setup

The models were fine-tuned and trained using a single NVIDIA T4 GPU on Google Colab. The training procedure was designed to balance performance with computational feasibility given hardware constraints and dataset size. Table 1 summarizes the key hyperparameters and setup used for model training.

Component	Setting
GPU	NVIDIA T4 (Google Colab)
Learning Rate	1×10^{-5}
Epochs	10
Loss Function (Binary)	Binary Cross-Entropy
Loss Function (Multi-class)	Cross-Entropy Loss
Optimizer	Adam (default hyperparameters)
Train-Test Split	80/20
Random Seed	42

Table 1. Training configuration and hyperparameters used in the fine-tuning of BioMedCLIP.

For the binary classification task in the first improvement, the target labels were binarized such that label 0 represented non- demented individuals, while labels 0.5, 1, and 2 were grouped into a single class indicating the presence of Alzheimer's disease. This binarization enabled the model to perform a simplified disease detection task in addition to the full multiclass classification setting.

3. Results

3.1. Binary Classification (First Improvement)

To evaluate the effectiveness of our proposed fine-tuning strategy, we compared the performance of the vanilla BioMedCLIP model with our improved BioMedCLIP variant across three different prompts. Each model was assessed using two key metrics: Accuracy and AUC (Area Under the Receiver Operating Characteristic Curve). The results are presented in Table 2.

Prompt	Model	Accuracy	AUC
Prompt 1	Vanilla BioMedCLIP	0.4314	0.4509
	Fine-tuned Model	**0.9314**	**0.9754**
Prompt 2	Vanilla BioMedCLIP	0.5196	0.6631
	Fine-tuned Model	**0.9314**	**0.9750**
Prompt 3	Vanilla BioMedCLIP	0.4804	0.3558
	Fine-tuned Model	**0.9314**	**0.9750**
Average (Vanilla BioMedCLIP)		0.4771	0.4899
Average (Fine-tuned Model)		**0.9314**	**0.9751**

Table 2. Comparison of binary classification performance between vanilla and fine-tuned BioMedCLIP across three diagnostic prompts. Fine-tuned BioMedCLIP significantly outperforms the baseline on both Accuracy and AUC.

The results indicate a substantial improvement in both Accuracy and AUC after applying our fine-tuning strategy. The vanilla BioMedCLIP model demonstrated high variability in its performance across different prompts, with AUC scores ranging from as low as 0.35 to 0.66. This suggests that the model, without domain-specific adaptation, struggled to consistently interpret Alzheimer's-related imaging features based solely on the prompt.

In contrast, the fine-tuned model achieved remarkably stable and high performance across all prompts, achieving an average accuracy of 93.14% and an AUC of 0.9751. This improvement can be attributed to the model's ability to better encode Alzheimer's-specific visual features and textual cues after fine-tuning on our curated dataset.

It is important to note that during this phase of evaluation, automatically generated captions were *not* prepended to the prompts. The prompts consisted purely of diagnostic language (e.g., "This is a photo of `<Label>`") which are available in Appendix 6.1. Despite this, the fine-tuned model demonstrated robust performance, highlighting its capacity to generalize visual patterns of disease progression without

explicit caption supervision.

These results validate the effectiveness of lightweight domain-specific adaptation of VLMs for medical imaging tasks, even under limited computational resources.

3.2. Multiclass Classification (Second Improvement)

To assess the impact of fine-tuning in a multiclass diagnostic setting, we evaluated three configurations: the vanilla BioMedCLIP model (ViT/B-16), the fine-tuned BioMedCLIP (ViT/B-16), and a variant where the vision encoder was swapped with a fine-tuned VGG-16 head. Table 3 summarizes the overall performance of all the models.

The fine-tuned BioMedCLIP model with the ViT/B-16 vision head achieved the highest overall accuracy (85.27%) and AUC (0.9654), significantly outperforming both the baseline and the VGG-16 variant.

The per-class accuracy graph in Figure 2 demonstrates how well the finetuning strategy allows the finetuned BioMedCLIP models to successfully differentiate between classes. However, the vanilla BioMedCLIP model still suffers from result variability due to prompt inconsistencies, as observed in the baseline performance when comparing per-class accuracy, which was 0%. In contrast, the fine-tuned BioMedCLIP with ViT/B-16 and VGG-16 variants attained 80.7% and 85.96%, respectively. These inconsistencies highlight the sensitivity of the model to prompt formulation, which is an important consideration when assessing its robustness.

Although the VGG-16 variant showed a slight edge in classifying the *Non-Demented* category, it struggled significantly in detecting *Very Mild Dementia*, suggesting a trade-off between feature granularity and generalization. The fine-tuned CLIP model, however, exhibited more balanced performance across all classes, demonstrating its overall robustness.

While the fine-tuned BioMedCLIP model leads to the best accuracy, the results raise important questions regarding the interpretability of these outcomes. As we explore deeper into the model's structures and layers, it becomes evident that while the model achieves high performance, understanding the mechanisms driving these results is crucial. This investigation into interpretability will guide further refinements and enhance the transparency of domain-adaptive fine-tuning methods. Ultimately, the superior performance of the fine-tuned CLIP model reinforces the value of modality alignment within the vision-language pretraining paradigm, particularly when optimized for Alzheimer's-related imaging features.

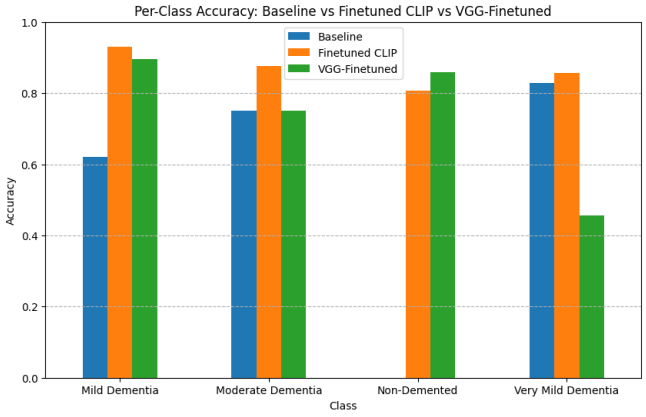

Figure 3. Per-class accuracy results for multiclass classification across model variants

4. Discussion

4.1. How Effective is finetuned BioMedCLIP without the Classification Head?

In our experiments, we initially fine-tuned the BioMedCLIP model with its classification head, which yielded promising results in terms of both accuracy and AUC. However, to better understand the role that the lightweight classification head plays in classification performance, we decided to evaluate the model without it. Specifically, we removed the classification head and used basic cosine similarity between image-text pairs from the fine-tuned BioMedCLIP model to perform classification.

The results, as shown in Table 4, reveal that without the classification head, the performance of the fine-tuned BioMedCLIP model drops significantly. The accuracy decreased from 85.27% to 15.69%, and the AUC also dropped from 0.9654 to 0.0863, indicating a significant decline in classification performance. This suggests that the model's learned representation is highly dependent on the classification head for achieving accurate results.

Model	Accuracy
Vanilla BioMedCLIP	0.4314
Fine-tuned BioMedCLIP (without classification head)	0.1569

Table 4. Accuracy comparison of BioMedCLIP with and without the classification head.

These results imply that the model's ability to make accurate predictions relies significantly on the classification head, and without it, the learned representation fails to generalize well for classification tasks. The poor performance could indicate that the model's representation is not robust enough to perform well without the additional guidance provided by the classification head.

Model	Accuracy	AUC
Vanilla BioMedCLIP (ViT/B-16)	0.4109	0.7032
Fine-tuned BioMedCLiP with Classification Head (ViT/B-16)	**0.8527**	**0.9654**
Fine-tuned BioMedCLiP with Classification Head (VGG-16)	0.7519	0.9517

Table 3. Overall multiclass classification performance across model variants on our custom prompts

To improve these results, several approaches could be explored. One possible improvement is unfreezing more parameters during fine-tuning, allowing the model to learn better representations that are less dependent on the classification head. Additionally, an alternative training strategy could involve first fine-tuning BioMedCLIP alone, followed by training the classification head in conjunction with the fine-tuned model, which may lead to better performance without sacrificing classification accuracy.

4.2. Impact of Captions on Model Classification

4.2.1. Vanilla BioMedCLIP

As discussed in Section 3, the vanilla BioMedCLIP model exhibits significant variability in classification accuracy when subjected to different, yet semantically similar, prompts. This inconsistency highlights the model's sensitivity to prompt formulation, where even minor linguistic changes can lead to drastic performance shifts. Such behavior suggests that the model's text encoder may not be robust enough to consistently align varied natural language inputs with the corresponding visual features.

This also indicates that the vanilla model, in its zero-shot form, may not generalize well without task-specific adaptation or prompt engineering. Moreover, it points to an underexplored area in captioning strategies: the need for more controlled or learned prompt formulations that reduce ambiguity and improve alignment between modalities. Without these, the model remains heavily dependent on the phrasing and structure of input prompts, limiting its utility in practical clinical or diagnostic applications.

Future work could investigate prompt tuning or the integration of learned textual adapters to mitigate this variability and enable more consistent zero-shot performance.

4.2.2. Fine-tuned BioMedCLIP

In this experiment, we aimed to evaluate the impact of captioning strategies on classification performance. Specifically, we compared the finetuned model's performance when provided with a generic caption versus our custom caption along with the prompt used earlier in the binary classification task.

As shown in Table 5, the results are almost identical between the two captioning strategies. This suggests that the text embeddings may not be playing a significant role in improving the classification task. One potential reason for this could be that the tokenizer used had a context length of 256 tokens, which might not adequately support the use of detailed or lengthy captions. This limitation may prevent the model from fully utilizing richer textual information when dealing with more elaborate captions. The results imply that, at least within the given context length, longer or more descriptive captions may not offer substantial performance gains.

Further empirical testing is required to explore whether more detailed captions would improve performance if the context length limitation was addressed or if a different tokenization strategy was employed.

Additionally, the fact that both captioning strategies yielded similar results could indicate that the majority of the decision-making power lies with the visual head of the BioMedCLIP model. The classification head may be relying more heavily on image features for classification, rather than leveraging textual information. This could be particularly true in tasks where the visual data provides more discriminative power than the text, suggesting that the visual modality plays a dominant role in driving the classification performance in this specific setup.

Caption Type	Accuracy	AUC Score
Generic Caption	93.14%	0.9761
Custom Caption + Prompt	93.14%	0.9754

Table 5. Comparison of model performance with generic and custom captions.

4.3. Eigen-CAM Visualizations

To better understand the spatial attention patterns of our models during prediction, we utilized Eigen-CAM, a gradient-based class activation mapping technique. Eigen-CAM is particularly effective in scenarios where the majority of layers are frozen, as it computes class-discriminative heatmaps using the dominant eigenvectors of the gradient-weighted activation maps without requiring architectural changes. This property made it a suitable choice for our experiments, given the frozen nature of the vision encoders during fine-tuning.

The visualizations allow us to assess whether the ViT/B-16 and VGG-16 vision heads are attending to clinically relevant

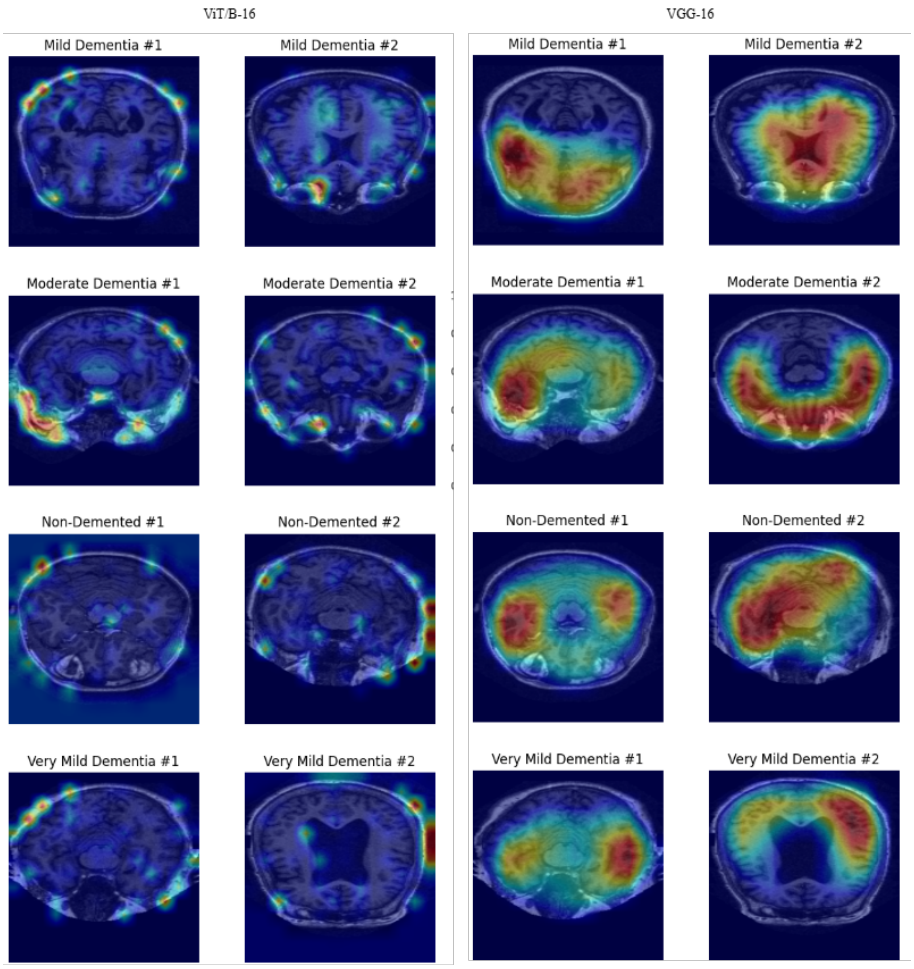

Figure 4. Eigen-CAM visualizations for different visual encoders (ViT/B-16: left, VGG-16: right) across different dementia classes.

brain regions associated with Alzheimer's disease. These include:

- Extreme shrinkage of the hippocampus [10]
- Enlarged ventricles [11]
- Cortical thinning or shrinkage [12]

The target layers for activation extraction were selected based on the structural characteristics of each model. For the ViT/B-16 model from CLIP, we used the patch embedding projection layer located in the visual trunk. This layer processes the raw image patches before they are passed through the transformer blocks and serves as an appropriate site for capturing spatial attention. For the VGG-16 model, we chose the last convolutional layer, specifically the layer at index 28 in the feature extractor. This is the final convolutional layer before the ReLU activation and max-pooling operations, making it a suitable candidate for generating class-discriminative activation maps.

As seen in Figure 4, the ViT/B-16 model frequently focuses on peripheral regions of the scan, including areas near the eyes or outer boundaries of the skull, rather than internal neuroanatomical regions critical to Alzheimer's diagnosis. This misalignment raises concerns about the interpretability of ViT-based predictions, as the model may be leveraging spurious correlations rather than medically relevant features.

In contrast, the VGG-16 model consistently highlights key brain regions across all dementia classes. For instance:

- In **Moderate Dementia**, activation is strong near the temporal lobe, aligning with hippocampal shrinkage.
- In **Non-Demented** scans, focus is more evenly distributed, suggesting the model recognizes the absence of abnormalities.
- In **Very Mild Dementia**, localized attention is visible near the ventricles and cortex, hinting at early structural changes.

Overall, although the ViT/B-16-based BioMedCLIP model achieves superior classification accuracy, its attention maps lack alignment with known medical indicators. On the other hand, the VGG-16 model offers greater explainability and clinical relevance in its focus regions, despite performing slightly worse in raw metrics. This highlights a critical trade-off between interpretability and accuracy—especially important in high-stakes domains like medical diagnostics.

In future work, a more robust comparison could be performed amongst various vision heads through full finetuning of the models by unfreezing all of the parameters of the vision encoders to fully explore interpretability. With the current efforts, however, it feels like combining the interpretability of CNN-based models with the strong multimodal alignment of transformer-based models could offer a promising direction for trustworthy and performant diagnostic systems based on VLMs.

5. Conclusion

While our study demonstrates promising results, it is important to acknowledge several limitations. First, limited computational resources restricted the scope of our comparative analysis against a broader range of baseline models. Additionally, our approach used a single caption per class rather than generating custom captions for each individual MRI slice, which may have constrained the full potential of the vision-language alignment. Finally, the relatively small size of the dataset may have affected the generalizability of the results.

Despite these constraints, this work highlights the effectiveness of fine-tuning a Vision-Language Model like BioMedCLIP with domain-specific captions and a dedicated classification head for Alzheimer's Disease classification. The model achieved strong performance in both binary and multiclass tasks, outperforming conventional baselines. By integrating Eigen-CAM, we gained valuable insight into the explainability of the model's predictions, making it more transparent and interpretable for clinical applications. Moreover, the use of Eigen-CAM also shed light on the architectural trade-offs between Vision Transformers and CNN-based models. Given that CNNs typically have fewer parameters and lower computational demands compared to ViT encoders, training well-optimized CNN-based models may offer a more sustainable and diagnostically efficient alternative—reducing both inference complexity and environmental impact.

Overall, our findings underscore the potential of multimodal, interpretable models in medical imaging, and pave the way for future work in scalable, sustainable, and explainable diagnostic AI systems.

References

[1] Aderghal, K. et al. 2018. Classification of Alzheimer's disease on imaging modalities with deep CNNs using cross-modal transfer learning. Proc. IEEE Int. Symp. Computer-Based Medical Systems (CBMS), 271–276.

[2] Marey, A. et al. 2024. Explainability, transparency, and black box challenges of AI in radiology. Egypt. J. Radiol. Nucl. Med., 55, 183.

[3] Hartsock, I., Rasool, G. 2024. Vision-language models for medical report generation and visual question answering. Front. Artif. Intell., 7, 1430984.

[4] Chiumento, F., Liu, M. 2024. Leveraging multimodal models for enhanced neuroimaging diagnostics in Alzheimer's disease. IEEE Int. Conf. on Big Data, arXiv:2411.07871.

[5] Lee, B. et al. 2025. Alzheimer's disease recognition using graph neural network by leveraging image-text similarity. Sci. Rep., 15, 997.

[6] Can, D.-C. et al. 2025. VisTA: Vision-Text Alignment Model for Alzheimer's diagnosis. arXiv:2502.01535.

[7] Feng, Y. et al. 2023. Large language models improve Alzheimer's diagnosis using multi-modality data. IEEE Int. Conf. on Medical Artificial Intelligence, 61–66.

[8] Zhang, S. et al. 2023. BioMedCLIP: multimodal biomedical foundation model pretrained from fifteen million pairs. arXiv:2303.00915.

[9] Muhammad, M. B., Yeasin, M. 2020. Eigen-CAM: Class Activation Map using principal components. Proc. IJCNN 2020, 1–7.

[10] Kesslak, J. P., Nalcioglu, O., and Cotman, C. W. 1991. Quantification of magnetic resonance scans for hippocampal and parahippocampal atrophy in Alzheimer's disease. Neurology 41(1): 51–54.

[11] Thompson, P. M., Hayashi, K. M., de Zubicaray, G. I., et al. 2004. Mapping hippocampal and ventricular change in Alzheimer's disease. NeuroImage 22(4): 1754–1766.

[12] Dickerson, B. C., Bakkour, A., Salat, D. H., et al. 2009. The cortical signature of Alzheimer's disease: regionally specific cortical thinning relates to symptom severity. Cereb. Cortex 19(3): 497–510.

[13] Scahill, R. I., Schott, J. M., Stevens, J. M., Rossor, M. N., and Fox, N. C. 2002. Mapping the evolution of regional atrophy in Alzheimer's disease: unbiased analysis of fluid-registered serial MRI. Proc. Natl. Acad. Sci. U.S.A. 99(7): 4703–4707.

[14] Apostolova, L. G., Dutton, R. A., Dinov, I. D., et al. 2012. Hippocampal atrophy and ventricular enlargement in

normal aging, mild cognitive impairment, and Alzheimer's disease. Alzheimer Dis. Assoc. Disord. 26(1): 17–27.

[15] Jack, C. R. Jr., Petersen, R. C., Xu, Y. C., Waring, S. C., O'Brien, P. C., and Tangalos, E. G. 1997. Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease. Neurology 49(3): 786–794.

6. Appendix

6.1. Prompts used for Evaluation

- **Prompt 1**: The brain MRI scan of this patient is likely to be classified as `<Label>`.
- **Prompt 2**: A medical expert looking to assess for a neuro-degenerative disease would likely classify this as `<Label>`.
- **Prompt 3**: Based on the brain scan and given indicators, the patient most likely has the following diagnosis or stage of Alzheimer's disease: `<Label>`.

6.2. Biomedical Background

To substantiate the clinical observations in Section 4 of the paper, we provide authoritative references (from neurology, radiology, and neuroimaging journals) for each claim. The in-text citations [10]–[15] correspond to the new references in ACM format listed at the end.

Extreme shrinkage of the hippocampus in AD patients: Advanced Alzheimer's disease is associated with dramatic atrophy of the hippocampus. MRI volumetric studies show significant hippocampal volume loss (on the order of 40% smaller) in AD patients compared to age-matched healthy controls [10]. This profound hippocampal shrinkage is a well-documented hallmark of AD neuropathology and correlates strongly with cognitive impairment severity [10].

Enlargement of the ventricles: Widespread brain atrophy in AD leads to ex vacuo dilation of the cerebral ventricles. Significant enlargement of the lateral (and third) ventricles is consistently observed in Alzheimer's patients relative to normal elderly individuals [11]. This ventriculomegaly is a recognized imaging feature of AD, with ventricular expansion rates correlating with disease progression and severity [11].

Cortical thinning or shrinkage: Alzheimer's disease causes diffuse cortical neurodegeneration. MRI studies have identified a characteristic "cortical signature" of AD, involving marked thinning of the cerebral cortex in vulnerable brain regions (e.g., temporoparietal and frontal association cortices) [12]. This cortical atrophy is quantifiable even in mild AD, and the degree of regional cortical thinning correlates with symptom severity in early-stage dementia [12].

Early involvement of the temporal lobe in moderate dementia: AD pathology begins in the medial temporal lobe and then spreads to neocortical areas. By the time patients reach moderate dementia, atrophy extends throughout the temporal lobes. Longitudinal MRI analyses demonstrate that as AD progresses from mild to moderate, there is a shift of atrophy into the lateral temporal cortex, with inferolateral temporal regions showing the most accelerated volume loss [13]. In other words, the temporal lobe (beyond just the hippocampus) is heavily affected by the moderate stage of AD [13].

Absence of abnormalities in non-demented brains: In cognitively normal (non-demented) older adults, such pronounced neurodegenerative changes are not present. Healthy aging can lead to mild brain volume loss, but elderly controls do not exhibit the severe hippocampal atrophy or ventricular enlargement seen in AD patients [14]. Studies consistently find that Alzheimer's patients have significantly more hippocampal shrinkage and ventricular expansion than age-matched healthy controls, while normals show little or no such abnormalities beyond typical age-related change [14].

Early changes near ventricles and cortex in very mild dementia: Even at the very mild stage of AD (e.g., CDR 0.5), subtle atrophic changes can be detected in the medial temporal lobe (near the ventricles) and in cortical regions. For example, patients with very mild AD already have significantly reduced hippocampal volumes (1.75 SD below normal controls) [15], reflecting early atrophy of the hippocampus (adjacent to the ventricular temporal horn). Additionally, the entorhinal/parahippocampal cortex (a medial temporal cortical area) is among the first sites of AD-related atrophy, showing volume loss even in prodromal stages [15]. These findings confirm that incipient AD causes measurable atrophy near the ventricles and in cortical regions even at the very mild dementia stage [15].