VPL: Visual Proxy Learning Framework for Zero-Shot Medical Image Diagnosis

Jiaxiang Liu[*,1,3] **Tianxiang Hu**[*,1] **Huimin Xiong**[*,1] **Jiawei Du**[3,4]
Yang Feng[2] **Jian Wu**[1] **Joey Tianyi Zhou**[3,4] **Zuozhu Liu**[†,1]

[1] ZJU-Angelalign R&D Center for Intelligence Healthcare, Zhejiang University, China
[2] Angelalign Research Institute, Angelalign Technology Inc., China
[3] Centre for Frontier AI Research (CFAR), Agency for Science, Technology and Research (A*STAR), Singapore
[4] Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore
{jiaxiang.21, zuozhuliu}@intl.zju.edu.cn

Abstract

Vision-language models like CLIP, utilizing class proxies derived from class name text features, have shown a notable capability in zero-shot medical image diagnosis which is vital in scenarios with limited disease databases or labeled samples. However, insufficient medical text precision and the modal disparity between text and vision spaces pose challenges for such paradigm. We show analytically and experimentally that enriching medical texts with detailed descriptions can markedly enhance the diagnosis performance, with the granularity and phrasing of these enhancements having a crucial impact on CLIP's understanding of medical images; and learning proxies within the vision domain can effectively circumvent the modal gap issue. Based on our analysis, we propose a medical visual proxy learning framework comprising two key components: a text refinement module that creates high-quality medical text descriptions, and a stable Sinkhorn algorithm for an efficient generation of pseudo labels which further guide the visual proxy learning. Our method elevates the Vanilla CLIP inference by supplying meticulously crafted clues to leverage CLIP's existing interpretive power and using the feature of refined texts to bridge the vision-text gap. The effectiveness and robustness of our method are clearly demonstrated through extensive experiments. Notably, our method outperforms the state-of-the-art zero-shot medical image diagnosis by a significant margin, ranging from 1.69% to 15.31% on five datasets covering various diseases, confirming its immense potential in zero-shot diagnosis across diverse medical applications.

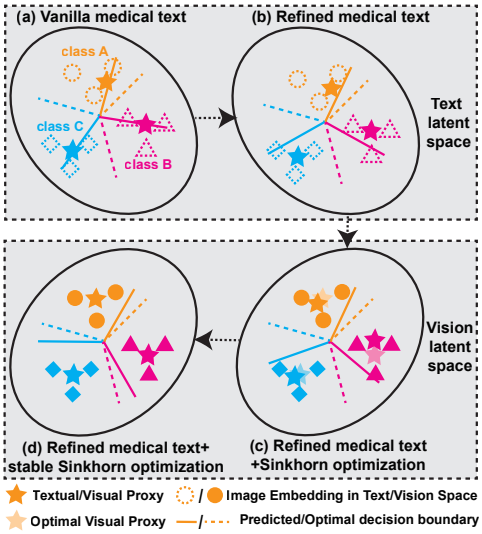

Figure 1: Schematic diagram of textual/visual proxies and decision boundaries learned by the vanilla method and our method. Refining medical texts brings predicted decision boundaries closer to the optimal. The Sinkhorn optimization facilitates a shift from text to visual proxies, mitigating modal gap and aligning predicted boundaries with optimal ones. While our stable Sinkhorn optimization, by addressing numerical instability, further enhances the algorithm's capacity for precise boundary approximation, thereby yielding better performance.

1 Introduction

In the realm of vision-language models (VLMs) (Do et al., 2021; Liu et al., 2021a; Wang et al., 2023a; Li et al., 2023; Gai et al., 2024; Lai et al., 2024a; Chen et al., 2023d; Xin et al., 2024; Jing et al., 2018; Liu et al., 2021c; Yang et al., 2021; Liu et al., 2021b,d), large-scale pretrained models such as Contrastive Language-Image Pre-Training (CLIP) have demonstrated exceptional capabilities across a spectrum of visual and linguistic tasks, particularly excelling in zero-shot recognition tasks (Radford et al., 2021; Liu et al., 2023c). Recent advancements have sought to adapt the principles underlying CLIP to the field of medical image analysis. For instance, initiatives like MedCLIP signify an effort to tailor large-scale VLMs for medical contexts, exploring applications in medical image classification (Wang et al., 2022). A pivotal area of exploration within this domain is zero-shot medical image classification, a task of paramount im-

*Contributed equally. †Corresponding author.

Findings of the Association for Computational Linguistics: EACL 2024, pages 9978–9992

portance in real-world healthcare settings. This is especially relevant where exhaustive datasets covering all possible pathologies are unattainable, and annotated medical images are scarce (Gai et al., 2024; Liu et al., 2023a; Chen et al., 2023c,b; Yang et al., 2024).

However, the application of existing VLMs like CLIP to zero-shot medical image classification encounters significant challenges (Radford et al., 2021). First, models such as CLIP, which are pre-trained on internet image-text pairs, may lack the specificity required for medical contexts, leading to suboptimal performance in medical applications (You et al., 2023). Moreover, the efficacy of zero-shot inference using CLIP hinges on the model's ability to interpret category texts. These texts in medical imaging are often highly specialized and abstract medical terminologies, posing an intrinsic challenge for existing VLMs (Chen et al., 2019, 2020; Li et al., 2020). Furthermore, recent studies (Liang et al., 2022; Qian et al., 2024) have consistently demonstrated that there is a sustained discrepancy between image and text features in zero-shot classification even in general contexts, and current contrastive multi-modal learning paradigms like CLIP training appear to be insufficient in mitigating this disparity (Radford et al., 2019; Chen et al., 2023a). These challenges, among others, highlight the imperative for devising specialized methodologies for the adaptation of VLMs towards achieving efficacious zero-shot classification in the realm of medical imaging (Zhao et al., 2023).

Efforts to counter the scarcity of medical image-text pairs in VLMs' pre-training datasets and the complexity of medical terminologies include enriching these models with detailed class descriptions (Liu et al., 2023c). Mirroring the diagnostic procedure of human experts, this involves leveraging human or generative model expertise to create thorough symptoms of medical conditions, unleashing the power of pre-trained VLMs. Preceding studies emphasize the importance of the text enrichment design, as the model performance is markedly influenced by the precision of descriptions or prompts for querying Large Language Models (Menon and Vondrick, 2023; Liu et al., 2023c; Ren et al., 2023). For our classification task, ensuring medical class descriptions are comprehensive, class-reflective, and distinct is crucial.

The vanilla CLIP inference essentially translates image features into the text feature space (Wang et al., 2023b), positioning these image features within the textual domain. However, recent discoveries have highlighted that the inherent discrepancies between image and text feature spaces may impede the efficacy of using text features as classification proxies. (Qian et al., 2024). A more effective approach involves finding proxies in the vision domain (Xin et al., 2023), bypassing the fundamental gap between the two spaces, a gap that is otherwise non-trivial to overcome. In CLIP inference, image labels are unknown, making the key to success the efficient use of CLIP-generated image and class text features for obtaining visual proxies. Despite its potential, visual proxy learning using these features, particularly in medical image classification, is still underexplored, thus merits further investigation.

In this paper, we propose a novel **V**isual **P**roxy **L**earning (VPL) framework for zero-shot medical image diagnosis. Specifically, we create specialized prompts for the expert to produce refined text descriptions for each disease, comprehensive and sufficiently distinct to harness the potential of VLMs. Furthermore, we propose a stable iterative algorithm based on Sinkhorn algorithm (Kruithof, 1937; Sinkhorn, 1964) that efficiently generates pseudo labels utilizing image features and those of the refined text. Proxy learning in the vision space directed by these pseudo labels profoundly enhances the efficacy of text enrichment, taking its impact to a higher level. Figure 1 showcases our improvements. The introduction of refined medical text facilitates a shift in proxies and decision boundaries within the text space towards the optimal state. Subsequent to the recovery of visual proxies, Stable Sinkhorn, outperforming the standard, yields superior proxies and decision boundaries in the vision space. We verify the effectiveness of each design through extensive experiments. The main contributions are summarized as:

- To address the semantic and modality gaps between text proxies and visual spaces in medical text classification within the CLIP paradigm, we propose VPL. This approach consists of a strategy for refining medical texts and a stable method for optimizing visual pseudo-labels in the logarithmic domain. These techniques aim to mitigate performance losses caused by these gaps.
- Comprehensive experiment results show that VPL robustly surpasses all baselines across a diverse array of evaluated datasets that span

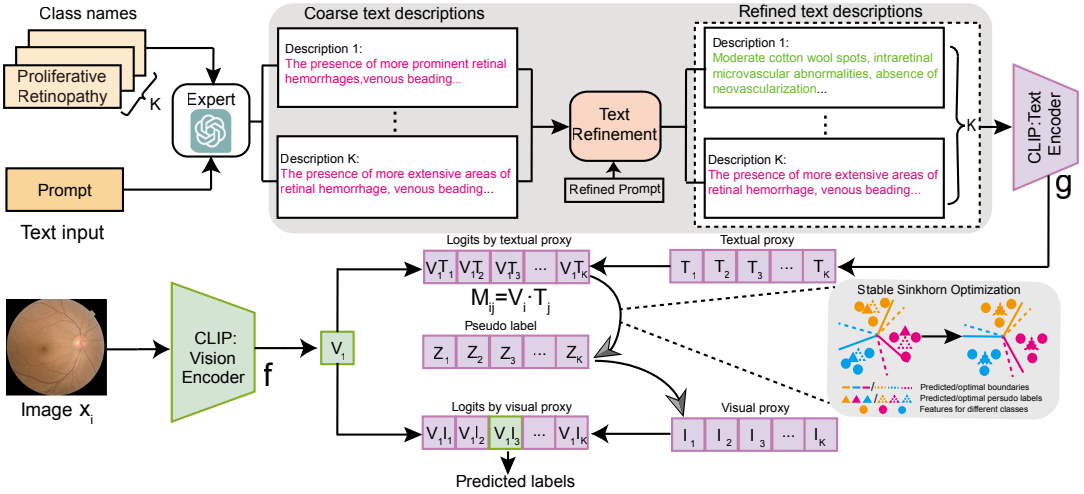

Figure 2: Model Pipeline. We initiate by inputting a meticulously crafted prompt into ChatGPT, which produces initial coarse text descriptions for each disease. These often contain overlapping information across different classes. Then, our TextRefinement module transforms these into distinct and clear text descriptions, with similar content highlighted in red and texts with reduced redundancy marked in green. Following this, a text encoder generates textual proxies, which merged with the visual features from the image encoder, yield the predicted logits. Finally, employing our newly developed stable Sinkhorn algorithm, we enhance the visual proxies, culminating in more accurate final predictions.

medical conditions including pneumonia, tuberculosis, diabetic retinopathy, and brain tumors. Notably, it establishes a substantial margin over the baselines in dataset IDRiD with a significant performance increase of 15.31%.

2 Related Work

2.1 Zero-shot Diagnosis

Zero-shot image classification has become essential for scenarios where extensive data labeling is impractical (Xian et al., 2018; Lampert et al., 2013). This approach, capable of categorizing images into unseen training classes, leverages semantic understanding and transfer learning. By learning from diverse images paired with descriptive texts, models like CLIP (Radford et al., 2021) grasp complex visual concepts, enabling them to classify images beyond their training data (Menon and Vondrick, 2023; Ren et al., 2023). Particularly transformative in medical image classification (Mahapatra et al., 2021, 2022; Dufumier et al., 2021), zero-shot learning addresses the rarity and diversity of medical conditions often absent in training datasets. Models like CLIP's ability to interpret and analyze medical images, even those depicting untrained conditions (Liu et al., 2023b), is invaluable, broadening the diagnostic scope and aiding in early detection and treatment of rare or emerging conditions, underscoring its potential in advancing healthcare diagnostics (Liu et al., 2023c).

2.2 Iterative Proportional Fitting Procedure

Iterative Proportional Fitting Procedure (IPFP), also known as biproportional fitting in various fields such as statistics, economics, and computer science (Stephan, 1942; Bacharach, 1965; Idel, 2016; Chang et al., 2023), has a rich history, being reexplored many times since its first appearing in (Yule, 1912). The essence of IPFP lies in transforming an initial matrix, denoted as A, into a fitted matrix P that is closest to A but conforms to certain row and column constraints. The attribution of the proof for both uniqueness and convergence within this process is credited to Sinkhorn's seminal work (Sinkhorn, 1964), which subsequently led to the naming of the algorithm in his honor.

3 Method

3.1 Proxy Learning Formulation

We focus on zero-shot image classification using CLIP (Radford et al., 2021). Let f, g denote the vision encoder and text encoder. Denote the image dataset for inference as $\{x_i\}_{i=1}^N$, and the image classes as $\{c_j\}_{j=1}^K$, the image features and text proxies can be represented as $\{\mathbf{x}_i\}_{i=1}^N, \{\mathbf{z}_j\}_{j=1}^K$ where $\mathbf{x}_i = f(x_i), \mathbf{z}_j = g(c_j)$. The classification process may be interpreted as the assignment of a given image feature $\mathbf{x}$ to a proxy $\mathbf{z}$, identified as the one demonstrating the greatest similarity to the image feature $\mathbf{x}$. Approaching from a foundational perspective, it becomes evident that the most suit-

able proxy for this task naturally emerges as $\mathbf{W}^* = \arg\min_{\mathbf{W}} \sum_i -log(\frac{\exp(\mathbf{x}_i^T \mathbf{w}_{y_i}/\tau)}{\sum_{j=1}^K \exp(\mathbf{x}_i^T \mathbf{w}_j/\tau)})$ where y_i are ground truth labels, which advocates for the training of proxies within the vision space. Recent work (Liang et al., 2022) have shown that the image and text feature spaces generated by the trained CLIP are distinct with a clear margin, and textual proxies inherently possess an inescapable deviation from the optimal proxies. Given the absence of y_i, it is reasonable to utilize image and class text features to create pseudo labels P_i' for supervising the proxy training in the vision space:

$$\hat{\mathbf{W}} = \arg\min_{\mathbf{W}} \sum_i KL(P_i' || P_i), \quad (1)$$
$$\text{where} \quad P_{i,j} = \frac{\exp(\mathbf{x}_i^T \mathbf{w}_j/\tau)}{\sum_{k=1}^K \exp(\mathbf{x}_i^T \mathbf{w}_k/\tau)}, \forall j = 1, ..., K.$$

Note that this problem is convex and thus the visual proxy can be obtained using the standard gradient descent.

3.2 Pseudo Label Generation

3.2.1 Optimization Problem Equivalence

Evidence from recent analyses (Qian et al., 2024) suggest that pseudo labels P_i' play a pivotal role in the performance of the learned visual proxy. A naive choice of P_i' is the distribution inferred by the textual proxy, which can be regarded as a solution of an optimization problem (Proposition 1).

Proposition 1. *Given M ($M_{i,j} = \mathbf{x}_i^T \mathbf{z}_j$), τ, let H denote the function to compute the entropy of matrices, consider the optimization problem*

$$\max_P (< M, P > + \tau H(P)), \quad (2)$$
$$s.t. \quad \forall i, \sum_j P_{i,j} = \frac{1}{n}; \forall i, j, P_{i,j} \geq 0.$$

The solution is

$$P_{i,j} = \frac{\exp(\mathbf{x}_i^T \mathbf{z}_j/\tau)}{\sum_{k=1}^K \exp(\mathbf{x}_i^T \mathbf{z}_k/\tau)}. \quad (3)$$

3.2.2 Hypothesis 1: Toward Text Refinement

In medical image classification, the categorical granularity provided by conventional medical taxonomies may be insufficient for the CLIP model to capture the full semantic intricacies in the data. Prior research suggests that a nuanced interpretation and subsequent refinement of these categorical definitions can significantly enhance the CLIP model's semantic comprehension, leading to notable improvements in classification performance (Menon and Vondrick, 2023). Consequently, we believe such text refinement can yield more accurate pseudo labels, which in turn, are instrumental for the efficacy of proxy learning methods. We consider modifying the logits matrix, transitioning from $M_{i,j} = f(x_i)^T g(c_j)$ to $M'_{i,j} = f(x_i)^T g(h(c_j))$, where h represents a text processing that enhances the semantic richness of each disease category so descriptions $h(c_j)$ are not only informative but also distinctly discernible across different disease categories.

Hypothesis 2: Toward Stable Sinkhorn Algorithm Building upon the original optimization problem, it becomes pertinent to explore the incorporation of a reference distribution across image classes $q \in R^K$ which can be defined as $q_j = \frac{\tilde{q}_j^\gamma}{\sum_{k=1}^K \tilde{q}_k^\gamma}$, where $\tilde{q}_j = \frac{1}{n} \sum_i \frac{\exp(\mathbf{x}_i^T \mathbf{z}_j/\tau)}{\sum_{k=1}^K \exp(\mathbf{x}_i^T \mathbf{z}_k/\tau)}$. This additional constraint on the matrix P leads to

$$\max_P (\langle M, P \rangle + \tau H(P)), \quad (4)$$
$$\text{s.t.} \quad \forall i, \sum_j P_{i,j} = \frac{1}{n}; \forall j, \sum_i P_{i,j} = q_j; \ \forall i, j, \ P_{i,j} \geq 0.$$

thereby refining the solution to better align with the intricacies of the image classification task we are addressing.

Proposition 2. *The new optimization problem has a unique solution of the form*

$$P = diag(u) A diag(v), \quad (5)$$

where $diag(u), diag(v)$ are two diagonal matrices with diagonals taken from vectors u, v, and $A = e^{\frac{M}{\tau}}$.

Proof. Introducing two dual variables $d^{(1)} \in \mathbb{R}^N$, $d^{(2)} \in \mathbb{R}^K$ for each marginal constraint, the Lagrangian reads

$$\mathcal{L}(P, d^{(1)}, d^{(2)}) = \langle P, M \rangle + \tau H(P) - \langle d^{(1)}, P\mathbf{1}_n - \frac{1}{n}\mathbf{1}_n \rangle - \langle d^{(2)}, P^T \mathbf{1}_K - q \rangle.$$

First order conditions yield

$$\frac{\partial \mathcal{L}(P, d^{(1)}, d^{(2)})}{\partial P_{i,j}} = M_{i,j} - \tau \log(P_{i,j}) - d_i^{(1)} - d_j^{(2)} = 0,$$

which results in the expression

$$P_{i,j} = e^{-d_i^{(1)}/\tau} e^{M_{i,j}/\tau} e^{-d_j^{(2)}/\tau},$$

which can be rewritten in the form $diag(u) A diag(v)$. □

Consider jointly the form of matrix P from proposition 2 and the constraint on its rows and columns as delineated in the optimization problem, the optimization task essentially equates to estimate matrix P given matrix $A = e^{\frac{M}{\tau}}$ such that

$$P_{i,j} = u_i A_{i,j} v_j, \sum_j P_{i,j} = \frac{1}{n}, \sum_i P_{i,j} = q_j. \quad (6)$$

An intuitive resolution is consecutively updating the matrix to fulfill the row and column conditions, known as the Sinkhorn algorithm (Kruithof, 1937; Sinkhorn, 1964). However, such an algorithm can suffer from computational instability due to underflow and overflow issues, as well as significant rounding errors, especially considering that the optimal matrix in our optimization problem has entries spanning a range that can include extremely small values. Thus, we propose to update the matirx within the logarithmic domain:

$$\text{Set } \log(P_{i,j}^{(0)}) = \frac{M_{i,j}}{\tau}, \text{ and for all } \eta \geq 1, \quad (7)$$

$$\log(P_{i,j}^{(2\eta-1)}) = \log(P_{i,j}^{(2\eta-2)}) + \log\left(\frac{1}{n}\right) - \log\left(\sum_{k=1}^{K} P_{i,k}^{(2\eta-2)}\right), \quad (8)$$

$$\log(P_{i,j}^{(2\eta)}) = \log(P_{i,j}^{(2\eta-1)}) + \log(q_j) - \log\left(\sum_{k=1}^{n} P_{k,j}^{(2\eta-1)}\right). \quad (9)$$

This technique can effectively compresses the dynamic range of the matrix entries. Furthermore, the logarithmic domain simplifies multiplicative operations into additive ones, enhancing computational stability and accuracy.

3.3 Pipeline

Under CLIP, we compute the similarity scores for image-text query pairs $(x, c), c \in \{c_1, ..., c_K\}$. The classification of image x is is determined by identifying the category $\hat{y} \in \{1, ...K\}$ that yields the highest similarity score between x and $c_{\hat{y}}$. Taking into account both hypotheses, we advocate for obtaining the pseudo labels for proxy learning via applying the stable sinkhorn algorithm to solve the optimization problem with the text refined logits M'. Then, we use the pseudo labels to supervise visual proxy learning, summarized in Algorithm 1. The pipeline is shown in Figure 2.

Step1: Text Enhancement and Feature Extraction The image x is processed through the CLIP visual encoder to obtain its visual representation: $V = \text{VisualEncoder}(x)$. In parallel, ChatGPT is queried with our designed prompt to generate major symptoms for each diagnostic category: $s_j = \text{ChatGPT}(\text{prompt1}, c_j)$, referencing the paradigm of (Liu et al., 2023c). In empirical investigations, we observed that ChatGPT can generate analogous descriptions for distinct classes.

Algorithm 1 VPL Framework

Input: Unlabeled image set $\{x_i\}_{i=1}^N$, class names $\{c_j\}_{j=1}^K$, expert (chatgpt) prompt $prompt1, prompt2$, CLIP vision and text encoders $\{f, g\}$, expert model (chatgpt), text similarity threshold δ

Output: Predicted labels $\hat{y}_i$

1: Obtain text descriptions $s_j = \text{Expert}(\text{prompt1}, c_j)$.
2: Evaluate class descriptions with δ. Obtain refined text descriptions $s_j = \text{TextRefinement}(\text{prompt2}, c_j)$.
3: Extract features $V_i = f(x_i)$ and $T_j = g(s_j)$.
4: Obtain logits matrix $M \in \mathbb{R}^{N \times K}$: $M_{i,j} = V_i \cdot T_j$.
5: Obtain reference distribution q, and then pseudo labels P' by Eq. 7-9. Refine P' with one-hot conversion.
6: Optimize visual proxies $\{I_j\}_{j=1}^K$ by Eq. 1
7: **return** $\hat{y}_i = argmax_j \; V_i \cdot I_j$

For example, when employing this procedure, class 'moderate nonproliferative retinopathy' and 'severe nonproliferative retinopathy' both include descriptions like retinal hemorrhages and venous beading. This homogeneity obscured the nuanced differences essential for accurate classification, thus led us to use a *TextRefinement* module as shown in Figure 3. This module examines each class pair, activating a specialized prompt for one class if their descriptions' cosine similarity exceeds a threshold δ (See Appendix B.3): $s_j = \text{TextRefinement}(\text{prompt2}, c_j)$, where prompt2 is designed to regenerate a differentiating description for that class. To illustrate with the previously mentioned retinopathy classes, prompt2 applying to class "moderate nonproliferative retinopathy" reads "Compared to severe nonproliferative retinopathy, according to published literature, what are useful medical visual features for distinguishing moderate nonproliferative retinopathy in a photo?" which differentiates the two classes based on distinct visual characteristics. The refined symptoms are then sent into the CLIP text encoder to obtain representations: $T_j = \text{TextEncoder}(s_j)$.

Step2: Obtaining Pseudo Labels Using Stable Sinkhorn Extracted features $\{V_i\}_{i=1}^{N}$ and $\{T_j\}_{j=1}^{K}$ serve as the foundational inputs for the computation of the logits matrix $M : M_{i,j} = V_i \cdot T_j$ which is the input of our stable sinkhorn algorithm (Eq. 7 - 9). The algorithm involves an iterative process that alternatively normalizes columns and rows of the matrix within the log domain for a set number of iterations. Following the final iteration, we exponentiate the log matrix to obtain a preliminary result, which is then transformed into the pseudo label matrix P' using one-hot conversion, as described in (Sohn et al., 2020; Qian et al., 2024).

Step3: VPL and Inference With pseudo labels, VPL is governed by Eq. 1. The learning process is executed through gradient descent using an adaptive learning rate, with an initial state of $\{I_j\}_{j=1}^{K}$ set to be $\{T_j\}_{j=1}^{K}$ (Qian et al., 2024). After the completion of weight updates, the learning process produces the better visual proxies $\{I_j\}_{j=1}^{K}$. A score function S is formulated to assess the similarity of an image-text pair (x, c) which entails computing the similarity between the feature representation of the image x and the visual proxy of class c: $\mathrm{S}(x, c) = V \cdot I$. Going over all categories, the one with the maximum score is taken as the predicted diagnosis of x:

$$\hat{y} = \underset{j \in \{1,\ldots,K\}}{\operatorname{argmax}} \mathrm{S}(x, c_j) = \underset{j \in \{1,\ldots,K\}}{\operatorname{argmax}} V \cdot I_j. \quad (10)$$

4 Experiments

4.1 Experimental Setup

Dataset & Evaluation Metric. For a comprehensive assessment, we follow the five datasets as CMD (Liu et al., 2023c) to conduct extensive performance validation, including Pneumonia (Kermany et al., 2018), Montgomery (Jaeger et al., 2014), Shenzhen (Jaeger et al., 2014), IDRiD (Porwal et al., 2018) and BrainTumor (Liu et al., 2023c) Dataset. They cover data from different countries and different disease types, as shown in Table 5. The classification accuray is used as the evaluation metric in our experiments.

Implementation Detail. We perform our experiments in PyTorch framework on a NVIDIA GEFORCE RTX 3090 GPU. We employ two versions of pre-trained CLIP vision encoders, i.e., ViT-L/14 and ViT-L/14@336px, for inference. We added hyperparameter search experiments for proxy learning, selecting a temperature τ of 0.01 and 20 iterations (see appendix Figure 7).

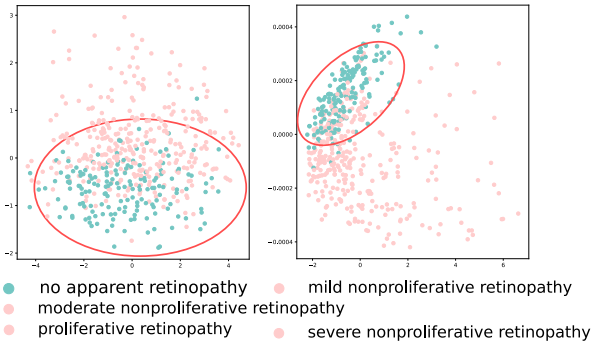

Figure 3: Results of PCA visualization. (a) with vanilla medical text input and textual proxies for inference. (b) with refined medical text input and stable Sinkhorn algorithm to generate visual proxies for inference.

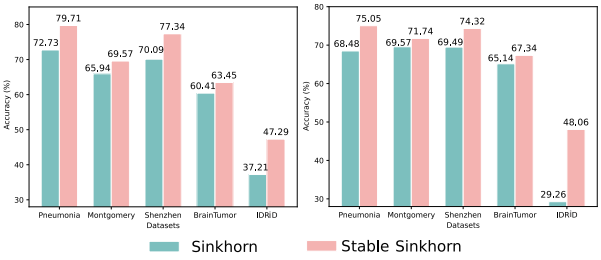

Figure 4: Performance comparison of Stable Sinkhorn and Sinkhorn under Refined Texts. The left and right subfigures denote the results under ViT-L/14 and ViT-L/14@336px.

4.2 Main Results

We evaluate the proposed method and compare it to the state-of-the-art zero-shot image classification methods, including CLIP (Radford et al., 2021), VCD (Menon and Vondrick, 2023), the medical diagnosis method CMD (Liu et al., 2023c), and InMaP (Qian et al., 2024), with two backbones on five datasets as depicted in Section 4.1. Results are shown in Table 1. Note that since the GPT-3 model as used in the VCD has limitations in interpreting medical terms, we replace it with the upgraded GPT-3.5 model. We can observe that on all datasets, regardless of the backbone used, our method consistently and conspicuously outperforms the baselines, which indicates its highly effective and robust capabilities in medical diagnosis. Notably, our method surpasses all baselines by a large margin of 15.31% on dataset IDRiD which uniquely demands the evaluation of severity levels rather than the binary detection tasks in other datasets, underscoring our method's superior capability in processing complex medical data and discerning subtle differences across severity levels.

To demonstrate the improvements in representation learning brought by VPL, we perform PCA

Table 1: Comparison to baselines with different backbones on Various Datasets

	Pneumonia	Montgomery	Shenzhen	BrainTumor	IDRiD
Model	ViT-L/14	ViT-L/14	ViT-L/14	ViT-L/14	ViT-L/14
CLIP (Radford et al., 2021)	64.55	60.14	50.76	57.36	06.80
VCD (Menon and Vondrick, 2023)	72.97	63.77	64.65	58.04	18.45
CMD (Liu et al., 2023c)	73.36	59.42	68.13	62.61	20.38
InMaP (Qian et al., 2024)	66.75	68.12	69.79	63.96	32.75
VPL (Ours)	**75.05**(+1.69)	**71.74**(+3.62)	**74.32**(+4.53)	**67.34**(+3.38)	**48.06**(+15.31)
Model	ViT-L/14@336px	ViT-L/14@336px	ViT-L/14@336px	ViT-L/14@336px	ViT-L/14@336px
CLIP (Radford et al., 2021)	71.14	57.97	50.76	57.19	11.65
VCD (Menon and Vondrick, 2023)	72.97	62.32	69.94	57.19	18.45
CMD (Liu et al., 2023c)	73.17	57.97	68.88	58.38	13.59
InMaP (Qian et al., 2024)	70.27	60.87	73.72	57.53	38.37
VPL (Ours)	**79.71**(+6.54)	**69.57**(+7.25)	**77.34**(+3.62)	**63.45**(+5.07)	**47.29**(+8.92)

visualization (Abid et al., 2018) according to the grading predictions of the model with backbone ViT-L/14 on dataset IDRiD and compare the effects in Figure 3. Subfigures (a), (b) denote VPL with different learning scheme versions, as described in the caption. We can observe from the visualization that our complete model (b) has distinct improvements over the model with naive learning schemes (a). Using our model, the individual characteristics of the same type of disease are closer, and the feature distinction of different clusters of disease categories is more obvious, which shows the effectiveness of our designs for feature learning.

4.3 Ablation Studies

Effects of Text Refinement. We compare the performance with and without text refinement both in conjunction with the Stable Sinkhorn algorithm and in its absence, as exhibited in Table 2. It can be observed that on the basis of using stable Sinkhorn algorithm, replacing Vanilla text input with our refined texts, i.e., from 1(c) to 1(d) and from 2(c) to 2(d) in Table 2, consistently brings performance improvements. In the case without stable Sinkhorn algorithm, i.e., from 1(a) to 1(b) and from 2(a) to 2(b), the text refinement also demonstrates positive influence on all datasets except Montgomery. Besides, we compare our text refinement to the CMD's text enrichment strategy (Liu et al., 2023c) which essentially does not perform text similarity reduction. Since the text descriptions between different disease categories are quite different generated from dataset Pneumonia, Montgomery and Shenzhen, meaning there is no need to reduce text repeatability, we only perform the comparison on dataset

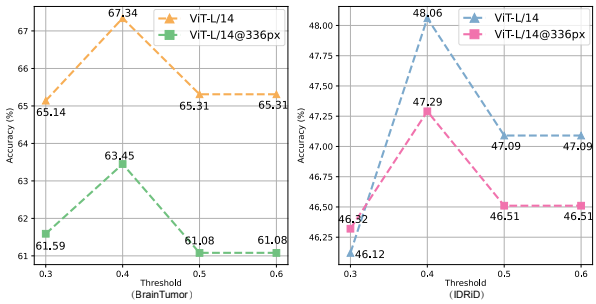

Figure 5: Ablation study on the text cosine similarity threshold. It is observed that 0.4 represents the optimal threshold.

BrainTumor and IDRiD. For fair comparison, all comparison experiments are employed the stable Sinkhorn to obtain visual proxies. The results are shown in Table 3, revealing that our similarity reduction between different categories is necessary and our text refinement is superior to CMD.

Effects of Stable Sinkhorn Algorithm. Similarly, we check the effect of stable Sinkhorn algorithm, as shown in Table 2. Implementing the stable Sinkhorn algorithm, i.e., from 1(a) to 1(c), 1(b) to 1(d), from 2(a) to 2(c), 2(b) to 2(d), yields accuracy enhancements across almost all scenarios, showing the effectiveness of the proposed algorithm. In addition, we compare the performance of the visual proxies in (Qian et al., 2024) learned with pseudo labels from the standard Sinkhorn and our visual proxies learned with pseudo labels from the proposed stable Sinkhorn. Results are illustrated in Figure 4, demonstrating that our algorithm design can indeed boosts performance, which aligns with our analysis in Hypothesis 2.

Interesting interaction between Text Refinement and Stable Sinkhorn. We find an interesting phe-

Table 2: Ablation study on different components with two types of backbone on five datasets.

Backbone		Refined Texts	Stable Sinkhorn	Pneumonia	Montgomery	Shenzhen	BrainTumor	IDRiD
ViT-L/14	1(a)			64.55	60.14	50.76	57.36	06.80
	1(b)	✓		72.97	42.03	59.82	58.38	24.61
	1(c)		✓	62.96	68.84	72.51	67.34	46.12
	1(d)	✓	✓	**75.05**	**71.74**	**74.32**	**67.34**	**48.06**
ViT-L/14@336px	2(a)			71.14	57.97	50.76	57.19	11.65
	2(b)	✓		72.97	42.03	56.04	57.19	37.02
	2(c)		✓	73.87	64.49	77.19	62.27	46.71
	2(d)	✓	✓	**79.71**	**69.57**	**77.34**	**63.45**	**47.29**

Table 3: Performance comparison of our text refinement approach with advanced text design in CMD (Liu et al., 2023c).

Dataset	Method	ViT-L/14	ViT-L/14@336px
BrainTumor	w/o TextRefinement	65.31	61.08
	VPL	**67.34**	**63.45**
IDRiD	w/o TextRefinement	47.09	46.51
	VPL	**48.06**	**47.29**

Table 4: Accuracy (%) comparison of ours and Flamingo. The * signifies a special case: a perfect score demonstrated on Shenzhen dataset, likely due to its inclusion in OpenFlamingo's training set.

	Ours	OpenFlamingo
Pneumonia (Kermany et al., 2018)	**79.71**	72.97
Montgomery (Jaeger et al., 2014)	**69.57**	57.97
Shenzhen (Jaeger et al., 2014)	77.34	**100***
BrainTumor (Liu et al., 2023c)	**63.45**	57.02
IDRiD (Porwal et al., 2018)	**47.29**	32.36

nomenon from Table 2: while implementing either text refinement or the stable Sinkhorn individually in the zero-shot diagnosis process occasionally yields failures, their combined application, i.e., from 1(b) to 1(d), 1(c) to 1(d), from 2(b) to 2(d), 2(c) to 2(d) in Table 2, consistently succeeds. This suggests a minor instability when these designs are used separately, yet also hints a synergistic interaction that enhances performance robustness and effectiveness when used together. We attributes this synergy to the effectiveness of text refinement in identifying class proxies within the vision domain. The precise nature of this interaction warrants deeper investigation in future studies.

Effects of text similarity threshold δ selection. As we conduct text similarity reduction in BrainTumor and IDRiD dataset, we record the results under different text similarity threshold δ choices on these two datasets in Figure 5. It can be seen that on both datasets, regardless of the backbone choice, our model consistently achieve the best performance when the threshold is 0.4. Therefore, we unanimously set $\delta = 0.4$ as the text similarity threshold in all our experiments.

4.4 Discussion

Currently, pretrained multimodal large models show great application potential in the field of medical diagnosis due to their powerful zero-shot inference capabilities (Awadalla et al., 2023; Lai et al., 2024b). We compare VPL to advanced multimodal large models, OpenFlamingo (Awadalla et al., 2023; Alayrac et al., 2022). For medical image diagnosis, OpenFlamingo adopts a medical visual question answering approach, inquiring the question of "Is this an image of {Diagnostic Category}?" Our experiments employ the OpenFlamingo 9B model (Zhu et al., 2023; Awadalla et al., 2023). The results in Table 4 display that VPL exceeds OpenFlamingo on most datasets except Shenzhen dataset. Although OpenFlamingo achieves 100% accuracy on the Shenzhen dataset, its performance on other datasets is consistently lower. This perfect score likely results from the Shenzhen dataset being included in OpenFlamingo's training data, making its generalization ability on this dataset unreliable. While for the rest datasets, VPL achieves an accuracy improvement from 6.43% to 14.93% compared with OpenFlamingo. Despite OpenFlamingo being trained on more data, its generalization ability lags behind ours, further demonstrating the superiority of VPL.

We applied VPL to the BioMedCLIP (Zhang et al., 2023), observing substantial improvements across four public datasets, highlighting its potential as a promising, training-free CLIP application (see Apendix Table 7).

5 Conclusion

In this work, we focus on enhancing zero-shot classification of CLIP in medical image diagnosis. Our analysis reveals that imprecise texts, along with the vision-text modal gap, are key impediments to current classification performance. To address these, We propose VPL, which consists of a strategy to refine medical texts and a method to recover

class proxies in the visual space with the aid of medical textual proxies. Specifically, medical category texts are refined via LLMs to alleviate semantic ambiguities between medical texts and images. Then, by improving textual proxy predictions with unlabeled image data, visual proxies are learned through pseudo-labels, circumventing modal discrepancies. Experiments across various datasets consistently demonstrate zero-shot accuracy improvement in CLIP through our approach.

Limitation

The current limitation of this study is that it relies solely on the expert model ChatGPT (Ouyang et al., 2022) for detailed disease text descriptions. Engaging experts who are more intimately familiar with the nuanced classification of these diseases to craft more refined texts could enhance the representation of disease characteristics, thereby yielding a more optimized text proxy. Presently, our method is closely approaching the performance of supervised approaches. We aspire that in the future, through the ultimate refinement of text representations of diseases, we will surpass the performance of supervised methods.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 62106222), the Natural Science Foundation of Zhejiang Province, China (Grant No. LZ23F020008), and the Zhejiang University-Angelalign Inc. R&D Center for Intelligent Healthcare. This work is also supported by Jiawei Du's A*STAR Career Development Fund (CDF) C233312004.

References

Abubakar Abid, Martin J Zhang, Vivek K Bagaria, and James Zou. 2018. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature communications*, 9(1):2134.

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

Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, et al. 2023. Openflamingo.

Michael Bacharach. 1965. Estimating nonnegative matrices from marginal data. *International Economic Review*, 6(3):294–310.

Serina Chang, Zhaonan Qu, Jure Leskovec, and Johan Ugander. 2023. Inferring networks from marginals using iterative proportional fitting. In *The Second Learning on Graphs Conference*.

Man Chen, Wenquan Dong, Hao Yu, Iain Woodhouse, Casey M Ryan, Haoyu Liu, Selena Georgiou, and Edward TA Mitchard. 2023a. Multimodal deep learning for mapping forest dominant height by fusing gedi with earth observation data. *arXiv preprint arXiv:2311.11777*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Learning universal image-text representations.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Yinda Chen, Wei Huang, Xiaoyu Liu, Qi Chen, and Zhiwei Xiong. 2023b. Learning multiscale consistency for self-supervised electron microscopy instance segmentation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Yinda Chen, Wei Huang, Shenglong Zhou, Qi Chen, and Zhiwei Xiong. 2023c. Self-supervised neuron segmentation with multi-agent reinforcement learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 609–617.

Yinda Chen, Che Liu, Wei Huang, Sibo Cheng, Rossella Arcucci, and Zhiwei Xiong. 2023d. Generative text-guided 3d vision-language pretraining for unified medical image segmentation. *arXiv preprint arXiv:2306.04811*.

Tuong Do, Binh X Nguyen, Erman Tjiputra, Minh Tran, Quang D Tran, and Anh Nguyen. 2021. Multiple meta-model quantifying for medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 64–74. Springer.

Benoit Dufumier, Pietro Gori, Julie Victor, Antoine Grigis, Michele Wessa, Paolo Brambilla, Pauline Favre, Mircea Polosan, Colm Mcdonald, Camille Marie Piguet, et al. 2021. Contrastive learning with continuous proxy meta-data for 3d mri classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 58–68. Springer.

Xiaotang Gai, Chenyi Zhou, Jiaxiang Liu, Yang Feng, Jian Wu, and Zuozhu Liu. 2024. Medthink: Explaining medical visual question answering via multimodal decision-making rationale. *arXiv preprint arXiv:2404.12372*.

Martin Idel. 2016. A review of matrix scaling and sinkhorn's normal form for matrices and positive maps. *arXiv preprint arXiv:1609.06349*.

Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. 2014. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Baoyu Jing, Pengtao Xie, and Eric Xing. 2018. On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2577–2586.

Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131.

J Kruithof. 1937. Telefoonverkeersrekening. *De Ingenieur*, 52:15–25.

Zhixin Lai, Jing Wu, Suiyao Chen, Yucheng Zhou, Anna Hovakimyan, and Naira Hovakimyan. 2024a. Language models are free boosters for biomedical imaging tasks. *arXiv preprint arXiv:2403.17343*.

Zhixin Lai, Xuesheng Zhang, and Suiyao Chen. 2024b. Adaptive ensembles of fine-tuned transformers for llm-generated text detection. *arXiv preprint arXiv:2403.13335*.

Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. 2013. Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(3):453–465.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. 2023. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625.

Bo Liu, Li-Ming Zhan, and Xiao-Ming Wu. 2021a. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 210–220. Springer.

Fenglin Liu, Shen Ge, and Xian Wu. 2021b. Competence-based multimodal curriculum learning for medical report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3001–3012.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021c. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13753–13762.

Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, Xu Sun, et al. 2021d. Auto-encoding knowledge graph for unsupervised medical report generation. *Advances in Neural Information Processing Systems*, 34:16266–16279.

Jiaxiang Liu, Jin Hao, Hangzheng Lin, Wei Pan, Jianfei Yang, Yang Feng, Gaoang Wang, Jin Li, Zuolin Jin, Zhihe Zhao, et al. 2023a. Deep learning-enabled 3d multimodal fusion of cone-beam ct and intraoral mesh scans for clinically applicable tooth-bone reconstruction. *Patterns*, 4(9).

Jiaxiang Liu, Tianxiang Hu, Yan Zhang, Yang Feng, Jin Hao, Junhui Lv, and Zuozhu Liu. 2023b. Parameter-efficient transfer learning for medical visual question answering. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

Jiaxiang Liu, Tianxiang Hu, Yan Zhang, Xiaotang Gai, YANG FENG, and Zuozhu Liu. 2023c. A chatgpt aided explainable framework for zero-shot medical image diagnosis. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*.

Dwarikanath Mahapatra, Behzad Bozorgtabar, and Zongyuan Ge. 2021. Medical image classification using generalized zero shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3344–3353.

Dwarikanath Mahapatra, Zongyuan Ge, and Mauricio Reyes. 2022. Self-supervised generalized zero shot learning for medical image classification using novel interpretable saliency maps. *IEEE Transactions on Medical Imaging*, 41(9):2443–2456.

Sachit Menon and Carl Vondrick. 2023. Visual classification via description from large language models. In *The Eleventh International Conference on Learning Representations*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. 2018. Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data*, 3(3):25.

Qi Qian, Yuanhong Xu, and Juhua Hu. 2024. Intramodal proxy learning for zero-shot visual categorization with clip. *Advances in Neural Information Processing Systems*, 36.

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

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

Zhiyuan Ren, Yiyang Su, and Xiaoming Liu. 2023. Chatgpt-powered hierarchical comparisons for image classification. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Richard Sinkhorn. 1964. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879.

Mehreen Sirshar, Taimur Hassan, Muhammad Usman Akram, and Shoab Ahmed Khan. 2021. An incremental learning approach to automatically recognize pulmonary diseases from the multi-vendor chest radiographs. *Computers in Biology and Medicine*, 134:104435.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608.

Frederick F Stephan. 1942. An iterative method of adjusting sample frequency tables when expected marginal totals are known. *The Annals of Mathematical Statistics*, 13(2):166–178.

Patrik Szepesi and László Szilágyi. 2022. Detection of pneumonia using convolutional neural networks and deep learning. *Biocybernetics and Biomedical Engineering*, 42(3):1012–1022.

Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. 2023a. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1802–1812.

Zhengbo Wang, Jian Liang, Ran He, Nan Xu, Zilei Wang, and Tieniu Tan. 2023b. Improving zero-shot generalization for clip with synthesized prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3032–3042.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. Medclip: Contrastive learning from unpaired medical images and text. In *2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*.

Zhan Wu, Gonglei Shi, Yang Chen, Fei Shi, Xinjian Chen, Gouenou Coatrieux, Jian Yang, Limin Luo, and Shuo Li. 2020. Coarse-to-fine classification for diabetic retinopathy grading using convolutional neural network. *Artificial Intelligence in Medicine*, 108:101936.

Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. 2018. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5542–5551.

Yi Xin, Junlong Du, Qiang Wang, Ke Yan, and Shouhong Ding. 2024. Mmap: Multi-modal alignment prompt for cross-domain multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16076–16084.

Yi Xin, Siqi Luo, Pengsheng Jin, Yuntao Du, and Chongjun Wang. 2023. Self-training with label-feature-consistency for domain adaptation. In *International Conference on Database Systems for Advanced Applications*, pages 84–99. Springer.

Jin Yang, Daniel S Marcus, and Aristeidis Sotiras. 2024. Dynamic u-net: Adaptively calibrate features for abdominal multi-organ segmentation. *arXiv preprint arXiv:2403.07303*.

Xingyi Yang, Muchao Ye, Quanzeng You, and Fenglong Ma. 2021. Writing by memorizing: Hierarchical retrieval-based medical report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5000–5009.

Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K Hong, Woonhyuk Baek, and Byungseok Roh. 2023. Cxr-clip: Toward large scale chest x-ray language-image pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 101–111. Springer.

G. Udny Yule. 1912. On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, 75(6):579–652.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. 2023. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2(3):6.

Zihao Zhao, Yuxiao Liu, Han Wu, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Xiang Li, Zhiming Cui, Qian Wang, et al. 2023. Clip in medical imaging: A comprehensive survey. *arXiv preprint arXiv:2312.07353*.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2023. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*.

Appendix

In this section, we present additional implementation details, experiment results, and supplements. The content structure is outlined as follows:

A Vision-language Pre-training

Visual-language (VL) pre-training involves training multi-modal models on extensive datasets featuring both visual and textual elements (Do et al., 2021; Liu et al., 2021a; Wang et al., 2023a; Li et al., 2023), eg. images and captions, to learn joint representations that capture the complex interactions between the two modalities. Practically, due to the high cost of acquiring manually annotated datasets, most visual-language models (Chen et al., 2019, 2020; Li et al., 2020; Radford et al., 2021; Liu et al., 2023b) are trained with image-text pairs captured from the Internet (Jia et al., 2021; Sharma et al., 2018). As an important example, with pre-training on 400 million pairs of image and text from the Internet, the model CLIP (Radford et al., 2021) gained rich cross-modal representations and achieved amazing results on a wide range of visual tasks without any fine-tuning. Leveraging CLIP's extensive learning, we can create a framework for training-free medical image diagnosis.

B Appendix Method

B.1 Stable Sinkhorn

In this section, we propose the Stable Sinkhorn Algorithm, which is designed to compute a doubly stochastic matrix from an arbitrary non-negative matrix M, as shown in Algorithm 2. The algorithm

Algorithm 2 Stable Sinkhorn Algorithm

Require: M, τ, iter
Ensure: Stable matrix $\mathbf{P}$

$row, col \leftarrow \text{shape}(M)$
$\log\_P \leftarrow M/\tau$
$\log\_P \leftarrow \log\_P - \text{logsumexp}(\log\_P, 1)$ {Normalize columns first for stability}
for $i = 1$ to iter **do**
  $\log\_P \leftarrow \log\_P - \text{logsumexp}(\log\_P, 0)$ {Normalize rows}
  $\log\_P \leftarrow \log\_P - \text{logsumexp}(\log\_P, 1)$ {Normalize columns}
end for
return $\exp(\log\_P)$ {Convert log probabilities back to probabilities}

operates in the logarithmic domain to enhance numerical stability, especially when dealing with matrices with large dynamic ranges that may cause underflow or overflow in a conventional computing environment.

The input matrix M is first scaled by a temperature parameter τ, which controls the entropy of the resulting matrix. The scaled matrix is then subject to the log-sum-exp operation, which is a smooth approximation of the maximum function that helps in computing the normalization factor in the logarithmic domain. This operation is defined as $\text{logsumexp}(\mathbf{A}, \text{dim}) = \log(\sum(\exp(\mathbf{A}), \text{dim}))$, where $\mathbf{A}$ is a matrix and dim is the dimension over which to perform the summation.

The output of the algorithm is the exponentiated version of the logarithmic matrix, which yields a doubly stochastic matrix $\mathbf{P}$ with all rows and columns summing to one.

The log-sum-exp operation is a critical component of the Stable Sinkhorn Algorithm. To facilitate the computation of this operation, we define a macro in our implementation, which is robust against potential numerical issues. The `logsumexp` function, used in various numerical algorithms for stability, is defined mathematically as:

$$\text{logsumexp}(\mathbf{a}) = \log\left(\sum_i \exp(a_i)\right) \tag{11}$$

where $\mathbf{a} = [a_1, a_2, \ldots, a_n]$ is a vector of real numbers. This function takes a matrix and a dimension as inputs and performs the log-sum-exp operation along the specified dimension. This is particularly useful in avoiding numerical underflow or overflow when dealing with very small or large exponentials.

Algorithm 3 Image Classifier Optimization

Require: $feat, text\_classifier, plabel$
Ensure: $classifier$

$lr \leftarrow 10$, $iter \leftarrow 2000$, $\tau_i \leftarrow 0.04$, $\alpha \leftarrow 0.6$
$ins, dim \leftarrow \text{shape}(feat)$
$val, idx \leftarrow \text{torch.max}(plabel, \text{dim} = 1)$
$mask \leftarrow val > \alpha$
$plabel[mask, :] \leftarrow 0$
$plabel[mask, idx[mask]] \leftarrow 1$
$base \leftarrow feat^\top @plabel$
$classifier \leftarrow text\_classifier.clone()$
$pre\_norm \leftarrow \infty$
for $i \leftarrow 0$ **to** $iter - 1$ **do**
  $prob \leftarrow \text{F.softmax}(feat@classifier/\tau_i, \text{dim} = 1)$
  $grad \leftarrow feat^\top @prob - base$
  $temp \leftarrow \text{torch.norm}(grad)$
  if $temp > pre\_norm$ **then**
    $lr \leftarrow lr/2$
  end if
  $pre\_norm \leftarrow temp$
  $classifier \leftarrow classifier - (lr/(ins \times \tau_i)) \times grad$
  $classifier \leftarrow \text{F.normalize}(classifier, \text{dim} = 0)$
end for

B.2 Optimize Image Classifier

The proposed algorithm optimizes an image classifier by leveraging the feature space represented by the matrix $feat$ and the probability labels $plabel$. The optimization is guided by the text classifier parameters $text\_classifier$ and is conducted over $iter$ iterations with an adaptive learning rate lr. At each step, labels with confidence above a threshold α are considered reliable and used to construct a target matrix $base$ through a masking operation. The optimization process employs a softmax function with a temperature parameter τ_i to compute probabilities, which are then used to calculate the gradient $grad$ of the loss function with respect to the classifier parameters. 'F' represents the nn.functional function in PyTorch. The learning rate is halved whenever the norm of the current gradient exceeds the previous iteration's gradient norm, preventing overshooting in the parameter space. The classifier parameters are updated by subtracting the gradient, scaled by the learning rate and normalized by the feature space dimensions

and temperature parameter. Finally, the classifier parameters are normalized to ensure stability and convergence, as shown in Algorithm 3. The hyper-parameters for Optimize Image Classifier follow reference (Qian et al., 2024), ensuring alignment with established research.

B.3 Descriptions' Cosine Similarity

Cosine Similarity $= \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \|\vec{B}\|}$ is used to measure the similarity between two diseases descriptions embedding $\vec{A}$, $\vec{B}$ in terms of content or semantic information. For example, moderate and severe retinopathy are characterized by "more retinal hemorrhages and venous beading". To distinguish between them, we utilize cosine similarity to assess their semantic similarity. If similarity surpasses threshold δ, distinct descriptions are generated by refined prompt: moderate retinopathy is marked by "Presence of retinal hemorrhages," "absence of neovascularization"(Distinguished from severe retinopathy).

C Experiments

C.1 VPL, compared with supervised methods

This work compiles and presents data from five distinct datasets, including their distribution as shown in Table 5. The dataset is sourced from multiple countries, covering various diseases and including datasets commonly used in previous studies, ensuring diversity, fairness and consistency. It's evident that the datasets encompass a variety of disease types from multiple countries, as illustrated in Figure 6. The method proposed in this study achieved consistent improvement across all datasets, demonstrating its strong generalizability. Furthermore, this work conducts a comparative analysis with supervised algorithms recently applied to these five medical datasets, as detailed in Table 6. It's important to note that the performance metrics for the supervised methods in Table 6 were obtained from experiments on training, validation, and test sets designated by their respective authors. In contrast, our study's metrics are derived from inference tests on all images in the datasets, allowing only for an approximate comparison. Nevertheless, our method outperforms the supervised baseline methods and approaches the top performance seen in the referenced supervised methods. This signifies a narrowing gap between training-free and supervised learning, highlighting the superiority of our zero-shot medical image diagnosis approach. The effectiveness of the improvements made in study is further substantiated by Figure 3.

Notably, in the IDRiD dataset, which includes five types of retinal lesions with relatively subtle distinctions, the best performance achieved by supervised methods on the test set is only 56.19%. In comparison, our method falls short by a mere eight percentage points without any of the supervision or data required by supervised methods. This represents a significant advancement. If future work could further refine the description or representation of disease characteristics, it's plausible that this gap could be further reduced.

C.2 Effects of VPL in BioMedCLIP

We applied VPL to BioMedCLIP (Note that only the BioMedCLIP (ViT-B/16) (Radford et al., 2021; Zhang et al., 2023) model is publicly available), trained specifically for the medical domain, with results shown in Table 7. Across four public datasets and regardless of the domain knowledge used to train the CLIP (Radford et al., 2021) models, applying VPL with a training-free approach resulted in substantial improvements. This underscores the potential of VPL's inference paradigm to become a promising application of CLIP.

C.3 Expert Prompt1

Expert Prompt1

According to published literature, what are useful medical visual features for distinguishing *Category B* in a photo?

C.4 Expert prompt2

Expert Prompt2

Compared to *Category A*, according to published literature, what are useful medical visual features for distinguishing *Category B* in a photo?

Table 5: Data statistics of the used five datasets. GM and PCNSL denote glioblastoma multiforme and primary central nervous system lymphoma, respectively.

Dataset	Data Source	Disease	Number	Distribution
Pneumonia	China	Pneumonia	5232	Bacterial: Viral: Normal=2538: 1345: 1349
Montgomery	USA	Tuberculosis	138	Normal: Abnormal=80: 58
Shenzhen	China	Tuberculosis	662	Normal: Abnormal=326: 336
IDRiD	India	Diabetic Retinopathy	516	Normal: Mild: Moderate: Severe: Proliferative= 168: 25: 168: 93: 62
BrainTumor	China	Brain tumor	593	GM: PCNSL=338: 255

Table 6: Accuracy (%) of our method and supervised learning methods on the four datasets we use in our work. Note that the accuracy of our method is evaluated on the whole dataset, while results of supervised approaches are calculated on test sets split by the authors. ~ denotes approximate values, indicating that the accuracies are slightly different from those reported above.

	Montgomery	Shenzhen	Pneumonia	IDRiD
Supervised (baseline) Method	62.58 (Sirshar et al., 2021)	65.70 (Sirshar et al., 2021)	61.19 (Szepesi and Szilágyi, 2022)	52.28 (Wu et al., 2020)
Supervised (best citation) Method	88.40 (Sirshar et al., 2021)	81.11 (Sirshar et al., 2021)	97.21 (Szepesi and Szilágyi, 2022)	56.19 (Wu et al., 2020)
Best Acc (Ours)	~71.74	~77.34	~79.71	~48.06

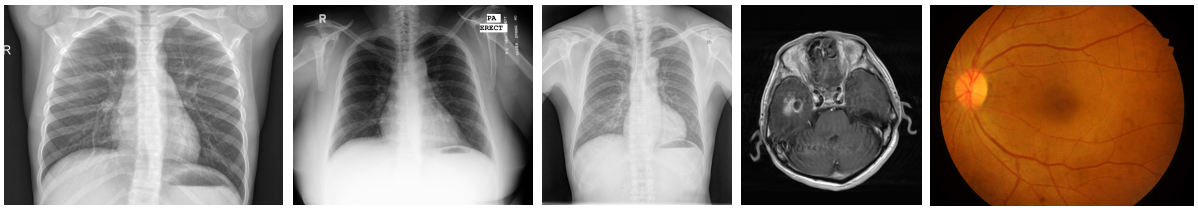

Figure 6: Examples from five different datasets. (a) is from dataset Pneumonia; (b) is from dataset Montgomery; (c) is from dataset Shenzhen; (d) is from dataset BrainTumor; (e) is from dataset IDRiD.

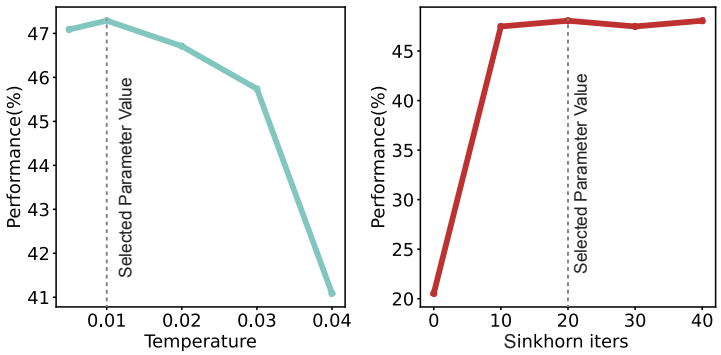

Figure 7: Stable sinkhorn parameter analysis.

Table 7: VPL with CLIP (ViT-L/14) and BioMedCLIP (ViT-B/16) on Various Datasets

	Pneumonia	Montgomery	Shenzhen	IDRiD
CLIP	64.55	60.14	50.76	06.80
VPL (CLIP)	**75.05**(+10.5)	**71.74**(+11.6)	**74.32**(+23.56)	**48.06**(+41.26)
BioMedCLIP	53.81	84.06	67.98	43.02
VPL (BioMedCLIP)	**71.72**(+17.91)	**90.58**(+6.52)	**83.38**(+15.4)	**46.71**(+3.69)