Leveraging Projected Residual Stream Decomposition for Mechanistic Probing and Parameter-Efficient Repair of Biomedical CLIPs

Omar Abdelwahed and Yiming Xiao

Department of Computer Science and Software Engineering,
Concordia University, Montreal, Canada
omar.abdelwahed@mail.concordia.ca, yiming.xiao@concordia.ca

Abstract. Despite excellent strengths in zero-shot generalization, biomedical vision–language models (e.g., CLIP) often require further adaptation for downstream tasks, potentially due to intrinsic representation deficiencies from multi-modal pretraining. While fine-grained mechanistic understanding of these representation deficiencies in biomedical CLIPs is highly beneficial to guide, assess, and explain new pretraining and finetuning techniques, related studies are still scarce. Therefore, by leveraging the recent Projected Residual Stream Decomposition, we reveal four representation defects in the popular BiomedCLIP and UniMed-CLIP models with identical underlying architectures, including model redundancy (component concept overlap & output embedding anisotropy), faulty attention heads, layer-wise concept erasure, and mean class prompt bias. In addition, we mechanistically compare the two models beyond their training data differences. Finally, with these insights, we explore several tailored, parameter-efficient/training-free model repair strategies to effectively enhance their performance in diagnostic tasks (max $\sim$ 44% accuracy gain, with 1.6% parameter tuning). By using six distinct public biomedical imaging datasets (6 modalities & tasks) for mechanistic probing and model repair, our study offers valuable insights for future exploration and adoption of biomedical CLIP models.

Keywords: Biomedical Vision–Language Models · Mechanistic Interpretability · Projected Residual Stream.

1 Introduction

Contrastive Language–Image Pretraining (CLIP) models that learn a shared image-text embedding space enable powerful zero-shot image classification [22], but they often require task-specific adaptation to ensure accuracy due to representation deficiencies from self-supervised pretraining. This is particularly important for biomedical CLIP models, such as BiomedCLIP [27] and UniMed-CLIP [11], whose applications are high-stakes. While recent parameter-efficient finetuning methods [13] for CLIP models have shown great success, in-depth mechanistic understanding of CLIP models' representation deficiencies could further guide, assess, and explain novel CLIP pretraining and finetuning techniques,

as well as in-depth model comparisons. However, unlike large language models (LLMs), there is a lack of such mechanistic probing for CLIP models.

The Projected Residual Stream (PRS) decomposition [5,7] is a mechanistic interpretability method that factors a Transformer model's hidden representation into additive contributions from residual updates of individual internal components (attention heads & MLPs) across different layers. Although it has been used in LLMs to enable component- and pathway-wise causal understanding of model behaviors, such as concept encoding, its adoption in CLIP models remains limited despite the potential benefits. Recently, TextSpan [7] extends this idea to semantically explain CLIP's vision encoder components, offering a new avenue to facilitate intuitive mechanistic probing of CLIP models.

To address the aforementioned knowledge gap, we use PRS decomposition for mechanistic diagnoses of two popular biomedical CLIP models (BiomedCLIP & UniMed-CLIP) with the same model architecture, across six distinct public datasets. Our main contributions include: **First**, we reveal four representation deficiencies that were studied in LLMs for the vision encoders in biomedical CLIP models, including *model redundancy (component concept overlap & output embedding anisotropy)*, *faulty attention heads*, *layer-wise concept erasure*, and *mean class prompt bias*. **Second**, for each deficiency type, we explore corresponding training-free or parameter-efficient model repair strategies to enhance the image classification performance, with a maximum accuracy gain of ∼44% and confirm the impact of the deficiencies. **Lastly**, we mechanistically compare BiomedCLIP and UniMed-CLIP beyond their training differences.

2 Methods and Materials

Datasets. For the study, we use six public datasets, one per imaging modality, covering ultrasound (BUSI [1]), MRI (BTMRI [19]), chest X-ray (COVID-19 [4]), histopathology (LungColon [3]), endoscopy (Kvasir [21]), and abdominal CT (CTKidney [10]). Each dataset has C=3∼8 classes, a 70%/10%/20% stratified training/validation/test split, and modality-specific class prompts (e.g., "`An MRI of [class]`") [12]. These splits are performed at the image level, as patient identifiers are not provided with these public datasets. All metrics and results are averaged over the 6 datasets and three seeds, where each seed defines a distinct split shared across all methods and baselines.

CLIP and zero-shot image classification. In CLIP-based zero-shot image classification, each class $y = c$ is written with a predefined context as a class prompt (e.g., *glioma* → "`an MRI of a glioma`") and encoded as t_c. With an encoded image feature x, its class prediction is decided by $\hat{y} = \arg\max_c \cos(x, t_c)$[1].

Projected residual stream decomposition & semantic interpretation. For a CLIP model, PRS decomposition can represent an image embedding x as additions of residual stream contributions of individual components projected in the shared image–text space:

[1] $\cos(\cdot)$ represents the cosine similarity score.

$$x \approx z^{(0)} + \sum_{l,h} z^a_{l,h} + \sum_l z^m_l, \qquad z^{(0)}, z^a_{l,h}, z^m_l, x \in \mathbb{R}^{512} \tag{1}$$

Here, $z^a_{l,h}$ and z^m_l are the contributions of Attention Head h and MLP in Layer l, respectively; $z^{(0)}$ is the stem. BiomedCLIP and UniMed-CLIP (both ViT-B/16) have 12 layers and 12 heads/layer (indices start from 0 in our later description), with 156 component contributions in total.

To facilitate natural-language-based interpretation of individual components' contributions, we follow the approach of TextSpan and build a biomedical text corpus $\mathcal{T}$ of short phrases with GPT-4 [20] (109∼157 phrases per dataset/modality), covering general categories of visual appearance, diagnostic findings, and imaging modality. The semantic concept of each component in the image encoder is revealed by finding textual concepts from the corpus that maximize the variance of the head/MLP activation patterns.

CLIP model representation deficiencies. As CLIP encourages global semantic alignment under often noisy and biased data distributions, entangled, shortcut-prone, and unevenly organized representation can occur, resulting in poor performance. Here, we investigate four *representation deficiencies* that were studied in LLMs in biomedical CLIP models: **1)** ***Model redundancy*** refers to the presence of overlapping semantic concepts in multiple model internal components; we investigate such redundancy in CLIP models' attention heads and MLPs, and then characterize the resulting *image output embedding anisotropy*, where the final representations concentrate in a few dominant directions rather than using the full shared space [6,18,25]. **2)** ***Faulty attention heads*** possess faulty concepts in the image encoder rather than task-centric ones, contributing to misclassification. **3)** ***Concept erasure*** is a layer-wise removal of semantic concepts from the encoder's final representation; we focus on adjacent-layer concept erasure to identify local concept destruction. **4)** In image classification, ***Mean class prompt bias*** is manifested as the predictions being heavily influenced by a fixed contextual prompt (i.e., mean of template-based class prompts) than by class-discriminative concepts, thus degrading the classification accuracy. We measure these deficiencies and devise tailored model repairs in Sec. 3.

Evaluation metrics for representation deficiencies. We describe the metrics used to assess each deficiency in the same order: **1)** ***Model redundancy*** is quantified by the *effective rank*, r_{eff} [23,8]. For a matrix A whose rows are the semantic vectors to test, with its singular values $\{\sigma_1, \ldots, \sigma_Q\}$ and $p_i = \sigma_i / \sum_j \sigma_j$,

$$r_{\text{eff}}(A) = \exp\Big(-\sum_{i=1}^{Q} p_i \log p_i \Big), \qquad 1 \le r_{\text{eff}}(A) \le Q, \tag{2}$$

where a low r_{eff} indicates that the rows concentrate on a few semantic directions. We first apply it inside the CLIP image encoder, stacking the 156 component contributions of Eq. 1 into $A \in \mathbb{R}^{156\times 512}$ to measure concept overlap across attention heads and MLPs. We then apply the r_{eff} to the normalized image embeddings X

to characterize the resulting anisotropic geometry of the output space. **2) *Faulty attention heads*** are identified on the misclassified images in the validation set by the *wrong-class margin* $M_{l,h} = cos(z^a_{l,h}, t_p) - cos(z^a_{l,h}, t_y)$, which measures how much a head supports the falsely predicted class t_p over the true class t_y. We utilize the top 5 heads with the highest $M_{l,h}$ for the repair by finetuning them. **3) *Concept erasure*** is quantified by the cosine similarity between adjacent-layer MLP contributions, $\cos(z^m_l, z^m_{l+1})$, computed on every adjacent pair in the CLIP image encoder, where a negative value indicates a local concept erasure. **4) *Mean class prompt bias*** is measured with respect to the mean class prompt per task $\bar{u} = \text{norm}(\frac{1}{C}\sum_c t_c)$. Specifically, we compute the cosine similarity between $\bar{u}$ and layer-wise attention head contribution $\cos(\sum_h z^a_{l,h}, \bar{u})$. A value near one indicates that the layer lacks class-discriminative concepts.

Model repair strategy training and baselines. For related parameter-efficient model repairs, finetuning uses AdamW [16] with learning rate 10^{-4}, batch size 32, cosine decay, and early stopping at patience of 5. All data-dependent choices, the edit directions, the strength $\alpha \in [0, 2]$, the rank r, and the faulty heads were decided based on the validation set while results are reported on the held-out test set. For baseline references, we include LoRA-based [9] and different variants of layer-wise CLIP model finetuning.

3 Mechanistic Diagnoses and Targeted Model Repairs

The evaluations of representation deficiencies are summarized in Fig. 1 for both BiomedCLIP and UniMed-CLIP while Table 1 presents the performance of the proposed training-free and parameter-efficient model repair methods, along with qualitative saliency map comparisons of model repairs in Fig. 2.

Model component redundancy. In the first step, we evaluate the effective ranks, r_{eff} for the image encoders of BiomedCLIP and UniMed-CLIP to assess the level of concept overlaps in their internal components. Among the 156 internal components (attention heads & MLPs), the mean effective ranks across the datasets for UniMed-CLIP and BiomedCLIP are $\mathbf{65.8 \pm 3.1}$ (from 60.2 on COVID-19 to 69.5 on CTKidney) and $\mathbf{43.1 \pm 0.7}$ (from 42.1 on BTMRI to 44.2 on Kvasir), respectively. This suggests a high level of redundancy in concept directions for both models, particularly for BiomedCLIP.

Output embedding anisotropy. Concept redundancy in model internal components can affect the final image embeddings, which are important for how well CLIP separates and aligns visual concepts. Therefore, we next use the effective ranks of the output image embeddings in both CLIP models to evaluate their embedding anisotropy level. Out of the 512 available semantic directions, the resulting average effective ranks (across datasets) are $\mathbf{57.9 \pm 9.1}$ (from 43.3 on BUSI to 70.9 on Kvasir) and $\mathbf{48.3 \pm 6.4}$ (from 41.1 on BUSI to 60.9 on Kvasir) for UniMed-CLIP and BiomedCLIP, respectively. BiomedCLIP exhibits higher embedding anisotropy while both models have relatively high redundancy in the embedding space, consistent with comparisons in the model internal components.

Repair strategy: To mitigate embedding anisotropy, we devise a *training-free edit* method (called "All-but-the-top" in Table 1) that removes a set of semantic directions B from the image embedding x with a strength α that is tuned on the validation set, leaving the class prompts and all weights untouched. Thus, the edited image embedding becomes:

$$x' = \frac{x - \alpha\, BB^\top x}{\|x - \alpha\, BB^\top x\|}, \qquad \hat{y} = \arg\max_c \ \cos(x', t_c). \tag{3}$$

Here, we set $B = [\bar{x}, v_1, \ldots, v_r]$, where $\bar{x}$ is the mean training embedding and $\{v_1, \ldots, v_r\}$ are the top r principal directions of the training embeddings. This removes the mean direction and the dominant directions that create the anisotropy, with r and α chosen on validation. The edit raises classification accuracy for BiomedCLIP from 51.1% to **56.8**% and for UniMed-CLIP from 49.3% to **55.4**%.

Faulty attention heads. The results of component-wise *wrong-class margin* $M_{l,h}$ are demonstrated in Fig. 1a for both BiomedCLIP and UniMed-CLIP. We find that the high values of the metric, which signifies strong faulty concepts, are primarily concentrated in the later layers (Layer 9∼11 for UniMed-CLIP and Layer 10 & 11 for BiomedCLIP). In addition, the sum of *wrong-class margins* over 144 attention heads in BiomedCLIP is **4.3**× that of UniMed-CLIP (2.97 vs. 0.70), and its worst single head is **3.8**× stronger (0.19 vs. 0.05).

Repair strategy: We propose a parameter-efficient model finetuning method ("Faulty heads" in Table 1) that only finetunes the top five attention heads with the highest *wrong-class margins* (highlighted in Fig. 1a), which is 1.6% of the vision encoder. This raises classification accuracy of BiomedCLIP from 51.1% to **94.6**% and UniMed-CLIP to **94.5**%. These are on par with different variants of LoRA-based and layer-wise model finetuning approaches, which achieve the highest accuracy of 95.5% and F1 of 95.3%, but require much more parameter adaptation. Notably, full finetuning trails all parameter-efficient approaches (92.8% on BiomedCLIP) because updating all weights on limited data distorts useful pretrained features [14,15]. This highlights the benefit of targeted intervention and helps confirm faulty attention heads' role in model performance. Interestingly, with TextSpan, we also find that the faulty heads exhibit the least concept overlap with the other components, and finetuning transforms their semantics towards more task-specific directions, with 51 of the 60 finetuned faulty heads (5 heads × 2 models × 6 datasets) landing on a diagnostic finding after finetuning. Take the detected faulty head L_{11}.H10 across both CLIP models on Kvasir as an example. With finetuning, its semantic concepts changed from "*a retroflexed view*" to "*esophagitis*" in BiomedCLIP and from "*a submucosal dissection*" to "*a lifted polyp*" in UniMed-CLIP. To assess whether these gains come from the head selection rather than the parameter budget alone, we finetune five randomly sampled heads with no overlap with the identified faulty heads, reaching $\mathbf{93.1 \pm 0.3}$% for BiomedCLIP and $\mathbf{94.1 \pm 0.2}$% for UniMed-CLIP. While the accuracy gap is modest, TextSpan shows mixed types of semantic changes for these random heads, unlike the consistent updates of the faulty ones. Specifically, some retain the same concept (e.g., L_7.H9 in UniMed-CLIP keeps "*nuclear*

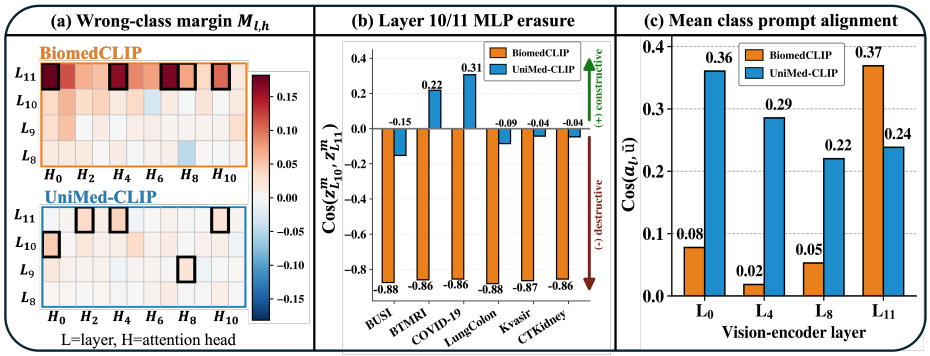

Fig. 1. Summary of biomedical CLIP model representation deficiency metrics: (a) Wrong-class margin for *Faulty attention heads* (finetuned heads in model repairs are highlighted); (b) cosine similarity scores of adjacent-layer MLPs' residual contributions for layer-wise concept erasure; (c) Mean prompt alignment for *Mean class prompt bias*.

atypia" on LungColon). Some switch into different concepts of the same category, such as L_5.H7 in BiomedCLIP, which changes from "*the retrosternal region*" to "*a bright spot*" on COVID-19, while others gain diagnostic concepts (e.g., L_4.H9 in UniMed-CLIP moves from "*a vertical line*" to "*a pituitary mass*" on BTMRI). Overall, only 33 of the 60 random heads across the two models land on a diagnostic finding after finetuning.

Layer-wise concept erasure. We compute the cosine similarity scores between the residual contributions of every adjacent-layer MLP pair for both CLIP models. In BiomedCLIP, the score of each pair is positive ($+0.17 \sim +0.87$), except that between Layer 10 and 11, resulting in a mean score of $\mathbf{-0.87}$ over all six datasets (Fig. 1b), signifying erasure of residual contributions towards the final embedding. On the other hand, UniMed-CLIP shows no concept erasure in adjacent-layer MLPs on average (very small negative scores exist in 4 datasets). A late component that writes contributions against an earlier one is well documented for LLMs [17,26,24], but not for CLIP. TextSpan further reveals that in BiomedCLIP, the top-1 semantic concepts in Layer 10 and 11 are related to diagnostic findings and imaging modality on BUSI, Kvasir and COVID-19 while on BTMRI, CTKidney and LungColon, the order is reversed. In comparison, UniMed-CLIP retains diagnostic finding concepts at both layers.

Repair strategy: We propose a training-free repair ("Erasure-gated MLP removal" in Table 1), which performs dynamically gated removal of residual contributions from the final embeddings x to mitigate the adverse erasure effect: $x' = \text{norm}(x - \alpha g\, z_l^m)$, with l chosen between Layer 10 and Layer 11 based on their coherence to the final embedding, along with α tuned on the validation set. Here, the gating factor $g = \max(0, -\cos(z_{L_{10}}^m, z_{L_{11}}^m))$ ensures that the contribution removal only occurs with adjacent-layer erasure in a dataset. For BiomedCLIP, where $g \approx 0.87$ on every dataset, this raises mean accuracy from 51.1% to **54.1**%

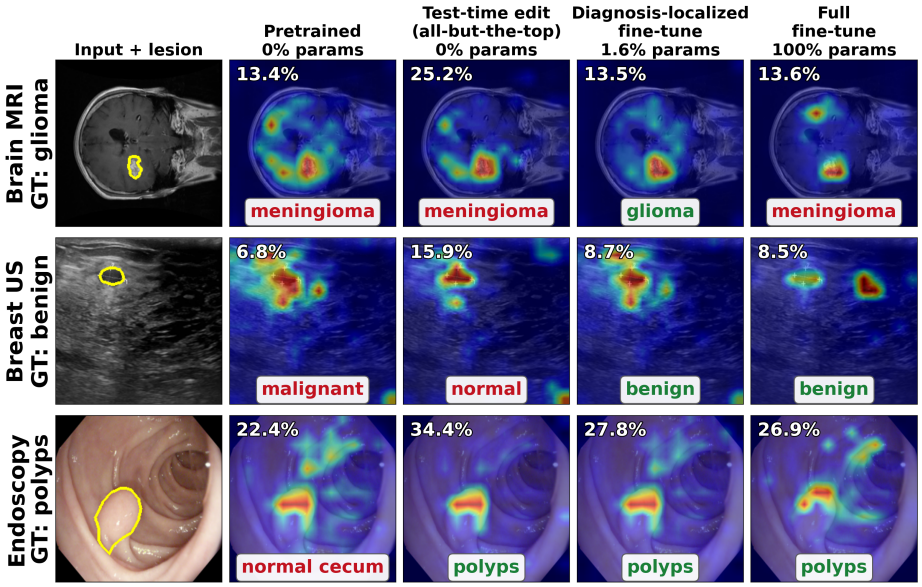

Fig. 2. Saliency map comparison across model repair strategies. The percentages in white fonts give the share of positive saliency inside the outlined lesion, and the predicted classes for each example are shown in each corresponding saliency map.

and macro-F1 from 50.0% to **52.6**%. For UniMed-CLIP, where g never exceeds 0.17, the same procedure moves neither measure, gaining 0.1% on each.

Mean class prompt bias. We compute the *mean class prompt alignment* $\cos(\sum_h z^a_{l,h}, \bar{u})$ at every Layer l. We showcase Layer 0,4,8 and 11 in Fig. 1c. The average alignment across six datasets is the strongest in Layer 11 for BiomedCLIP while for UniMed-CLIP, the strongest alignment occurs in Layer 0 and becomes weaker in later layers. As our class template contains the image modality concept, high alignment suggests that the residual contribution favors more generic concepts than disease-discriminative ones. Take the COVID-19 dataset as an example. With TextSpan, we find that in BiomedCLIP's Layer 11, the concept "*a medical chest X-ray*" explains 14.3% of the layer's activation variance, and no disease-related concepts reach the top ten; UniMed-CLIP's layer 11 instead ranks a disease-related finding first ("*bilateral perihilar opacities*").
Repair strategy: We employ a simple training-free model repair by reusing Eq. 3 for mitigating output embedding anisotropy, but with $B = [\bar{u}]$, to remove the mean class prompt concept [2] from the final embedding x. This raises the accuracy of BiomedCLIP to **55.1**% and UniMed-CLIP to **52.0**% (Table 1).

Comparison of different model repairs. As shown in Table 1, the training-free edits offer consistent but limited gains (maximum mean accuracy gain of **5.7**% and **6.1**% for BiomedCLIP and UniMed-CLIP), while all parameter-efficient finetuning strategies converge within $94.5 \sim 95.5$%. Notably, finetuning

Table 1. Accuracy and macro-F1 (%) over six datasets and three seeds, shown as mean±std across the seeds. Best per block results are in **bold**.

		BiomedCLIP		UniMed-CLIP	
Method	% params	Acc	F1	Acc	F1
Zero-shot (reference)	0	51.1±0.2	50.0±0.6	49.3±0.3	47.3±0.4
Training-free edits					
+ Prompt-mean removal	0	55.1±0.3	51.4±0.8	52.0±0.4	49.2±0.5
+ All-but-the-top	0	**56.8**±0.8	**53.8**±0.4	**55.4**±0.1	**53.4**±0.3
+ Erasure-gated MLP removal	0	54.1±0.4	52.6±0.5	49.4±0.3	47.4±0.4
Parameter-efficient finetuning					
LoRA (r=16)	3.2	95.2±0.6	95.0±0.6	95.3±0.2	94.9±0.4
LoRA (r=64)	11.4	**95.5**±0.7	**95.3**±0.8	**95.5**±0.2	95.1±0.3
Last 4 layers	33	95.1±0.6	94.8±0.7	95.1±0.4	94.8±0.4
Last 8 layers	66	95.1±0.8	94.6±1.0	95.4±0.9	**95.2**±0.9
Middle 4 layers	33	95.2±0.5	95.0±0.4	**95.5**±0.4	**95.2**±0.4
Faulty heads	1.6	94.6±0.8	94.2±0.9	94.5±1.0	94.2±1.1
Full encoder	100	92.8±0.8	92.2±1.0	94.8±0.5	94.6±0.5

Table 2. Per-dataset test accuracy (%), shown as mean±std over three seeds, for the zero-shot baseline and faulty-head finetuning.

Method	BUSI	BTMRI	COVID	LungC.	Kvasir	CTKid.	Mean
BiomedCLIP							
Zero-shot	41.0±2.3	63.3±1.2	32.0±0.3	67.1±0.4	60.7±1.5	42.6±0.8	51.1±0.2
Faulty heads	86.2±2.9	97.7±0.4	94.2±0.2	99.6±0.3	90.2±1.8	99.5±0.1	**94.6**±0.8
UniMed-CLIP							
Zero-shot	26.9±1.3	61.0±0.9	41.1±0.2	78.8±0.3	53.3±2.0	34.8±0.9	49.3±0.3
Faulty heads	83.2±6.2	98.5±0.3	96.8±0.9	100.0±0.0	88.8±0.5	99.8±0.1	**94.5**±1.0

only the faulty heads reaches this band with the least parameter adaptation (**1.6**%), while full-model finetuning performs worst among the finetuning strategies. Fig. 2 reflects the same trend at the image level, where the training-free edits mainly relocate the saliency onto the lesion regions (19.2% → **27.5**%), while finetuning the faulty heads converts the evidence into a correct classification.

4 Discussion

Although UniMed-CLIP was shown to outperform BiomedCLIP in zero-shot classification [11], their performances are similar in our study. This is likely due to the choice and splits of the datasets, considering UniMed-CLIP's training focuses more on radiological data. However, using PRS decomposition, we discover mechanistic differences between the two models, with more pronounced model redundancy, wrong-class margin in attention heads, local layer-wise concept erasure (Layer 10 & 11), and mean class prompt bias in BiomedCLIP. While we evaluate the four general representation deficiencies individually, model redundancy may also influence the others. For example, accumulations of redundant

concept representations in early layers can trigger concept erasures in later layers as a self-correction mechanism. In this study, we only focus on adjacent-layer concept erasure. In the future, more nuanced understanding of these mechanisms, their interplay, and further confirmation of the discovered trends with a broader range of datasets will be beneficial.

There are a few limitations for our exploratory study. *First*, as we only investigate BiomedCLIP and UniMed-CLIP, inclusion of more similar models with different architectures will enhance the values in mechanistic understanding of biomedical CLIP models, which face unique domain-specific challenges than those in natural visions. *Second*, we construct a set of biomedical text corpora with an LLM to help with intuitive understanding of the mechanistic probing by using TextSpan. Although the corpora have offered great values for this study, more comprehensive and clinician-validated corpora will further enhance the precision for semantic understanding of different model components and information flow in biomedical CLIP models. *Third*, we only focus on the CLIP image encoders for this study as class prompts have less variation than the input images, but we will investigate the text encoders jointly in the future. *Lastly*, the model repair strategies we proposed are preliminary. More sophisticated and efficient techniques are still required to better adapt the models, and this study showcases that fine-grained mechanistic understanding can guide such explorations in the future.

5 Conclusion

We use PRS decomposition to reveal four representation deficiencies in two popular biomedical CLIP models, and explore diagnosis-guided model repair strategies with both training-free and parameter-efficient approaches. We hope our mechanistic probing will offer valuable knowledge for the future exploration and adoption of biomedical vision-language models.

Disclosure of Interests. The authors have no competing interests.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in Brief **28**, 104863 (2020). https://doi.org/10.1016/j.dib.2019.104863
2. Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., Nanda, N.: Refusal in language models is mediated by a single direction. In: Advances in Neural Information Processing Systems (NeurIPS) (2024), arXiv:2406.11717
3. Borkowski, A.A., Bui, M.M., Thomas, L.B., Wilson, C.P., DeLand, L.A., Mastorides, S.M.: Lung and colon cancer histopathological image dataset (LC25000). arXiv preprint arXiv:1912.12142 (2019)
4. Chowdhury, M.E.H., Rahman, T., Khandakar, A., et al.: Can AI help in screening viral and COVID-19 pneumonia? IEEE Access **8**, 132665–132676 (2020). https://doi.org/10.1109/ACCESS.2020.3010287

5. Elhage, N., Nanda, N., Olsson, C., et al.: A mathematical framework for transformer circuits. Transformer Circuits Thread (2021), https://transformer-circuits.pub/2021/framework/index.html
6. Ethayarajh, K.: How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP). pp. 55–65 (2019)
7. Gandelsman, Y., Efros, A.A., Steinhardt, J.: Interpreting CLIP's image representation via text-based decomposition. In: International Conference on Learning Representations (ICLR) (2024), arXiv:2310.05916
8. Garrido, Q., Balestriero, R., Najman, L., LeCun, Y.: RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In: Proceedings of the 40th International Conference on Machine Learning (ICML). PMLR, vol. 202, pp. 10929–10974 (2023)
9. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022), arXiv:2106.09685
10. Islam, M.N., Hasan, M., Hossain, M.K., Alam, M.G.R., Uddin, M.Z., Soylu, A.: Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography. Scientific Reports **12**(1), 11440 (2022). https://doi.org/10.1038/s41598-022-15634-4
11. Khattak, M.U., Kunhimon, S., Naseer, M., Khan, S., Khan, F.S.: UniMed-CLIP: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities. arXiv preprint arXiv:2412.10372 (2024)
12. Koleilat, T., Asgariandehkordi, H., Rivaz, H., Xiao, Y.: Biomedcoop: Learning to prompt for biomedical vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
13. Koleilat, T., Rivaz, H., Xiao, Y.: CLIP-SVD: Efficient and interpretable vision-language adaptation via singular values. Transactions on Machine Learning Research (TMLR) (2026), arXiv:2509.03740
14. Kumar, A., Raghunathan, A., Jones, R., Ma, T., Liang, P.: Fine-tuning can distort pretrained features and underperform out-of-distribution. In: International Conference on Learning Representations (ICLR) (2022), arXiv:2202.10054
15. Lee, Y., Chen, A.S., Tajwar, F., Kumar, A., Yao, H., Liang, P., Finn, C.: Surgical fine-tuning improves adaptation to distribution shifts. In: International Conference on Learning Representations (ICLR) (2023), arXiv:2210.11466
16. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2019)
17. McGrath, T., Rahtz, M., Kramár, J., Mikulik, V., Legg, S.: The hydra effect: Emergent self-repair in language model computations. arXiv preprint arXiv:2307.15771 (2023)
18. Mu, J., Bhat, S., Viswanath, P.: All-but-the-top: Simple and effective postprocessing for word representations. In: International Conference on Learning Representations (ICLR) (2018), arXiv:1702.01417
19. Nickparvar, M.: Brain tumor MRI dataset. https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset (2021)
20. OpenAI: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
21. Pogorelov, K., Randel, K.R., Griwodz, C., et al.: KVASIR: A multi-class image dataset for computer aided gastrointestinal disease detection. In: Proceedings of the 8th ACM Multimedia Systems Conference (MMSys). pp. 164–169 (2017). https://doi.org/10.1145/3083187.3083212

22. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML). pp. 8748–8763 (2021)
23. Roy, O., Vetterli, M.: The effective rank: A measure of effective dimensionality. In: 2007 15th European Signal Processing Conference (EUSIPCO). pp. 606–610 (2007)
24. Rushing, C., Nanda, N.: Explorations of self-repair in language models. In: International Conference on Machine Learning (ICML) (2024)
25. Timkey, W., van Schijndel, M.: All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2021), arXiv:2109.04404
26. Wang, K., Variengien, A., Conmy, A., Shlegeris, B., Steinhardt, J.: Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In: International Conference on Learning Representations (ICLR) (2023)
27. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., et al.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI **2**(1), AIoa2400640 (2025). https://doi.org/10.1056/AIoa2400640