How Does Diverse Interpretability of Textual Prompts Impact Medical Vision-Language Zero-Shot Tasks?

Tony Wang
Department of Earth Science and Engineering
Imperial College London
Exhibition Rd, South Kensington, London SW7 2AZ
tony.wang23@imperial.ac.uk

Che Liu *
Department of Earth Science and Engineering
Imperial College London
Exhibition Rd, South Kensington, London SW7 2AZ
che.liu21@imperial.ac.uk

Arcucci Rossella
Department of Earth Science and Engineering
Imperial College London
Exhibition Rd, South Kensington, London SW7 2AZ
r.arcucci@imperial.ac.uk

Abstract

Recent advancements in medical vision-language pre-training (MedVLP) have significantly enhanced zero-shot medical vision tasks such as image classification by leveraging large-scale medical image-text pair pre-training. However, the performance of these tasks can be heavily influenced by the variability in textual prompts describing the categories, necessitating robustness in MedVLP models to diverse prompt styles. Yet, this sensitivity remains underexplored. In this work, we are the first to systematically assess the sensitivity of three widely-used MedVLP methods to a variety of prompts across 15 different diseases. To achieve this, we designed six unique prompt styles to mirror real clinical scenarios, which were subsequently ranked by interpretability. Our findings indicate that all MedVLP models evaluated show unstable performance across different prompt styles, suggesting a lack of robustness. Additionally, the models' performance varied with increasing prompt interpretability, revealing difficulties in comprehending complex medical concepts. This study underscores the need for further development in MedVLP methodologies to enhance their robustness to diverse zero-shot prompts.

1 Introduction

Medical Vision Language Pre-training (MedVLP) is a rapidly developing topic within the machine learning community[1–8]. For downstream tasks, models pre-trained with MedVLP is being applied to zero-shot diagnosis tasks, taking only image input and a textual prompt to describe the category name [9–15], thereby diminishing the data requirements and also enabling generalisation to open-set tasks. Recent Medical VLP(MedVLP) models, such as BioViL[9], MedKLIP[15], and KAD[13], have achieved superior performance in zero-shot diagnosis for chest X-ray (CXR) images with diverse and advanced pre-training techniques.

*Corresponding author: che.liu21@imperial.ac.uk

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

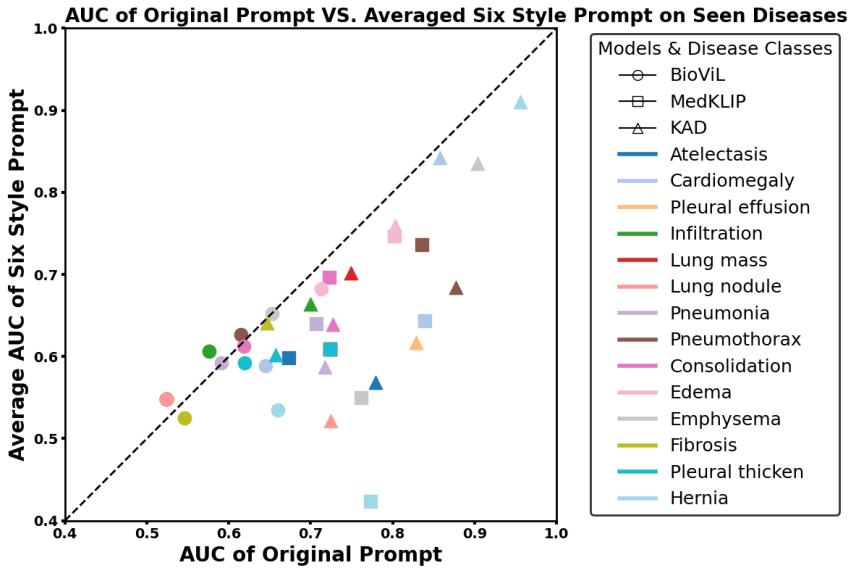

Figure 1: Comparison of original prompt and six style prompts' zero-shot image classification performance on seen disease classes of BioViL[9], MedKLIP[15] and KAD[13]. The X-axis shows the AUC performance of the original prompt, and the Y-axis shows the macro average of AUC performance of six style prompts. The dashed line shows the ideal scenario, where the model shows consistent performance on seen classes regardless of the prompt style.

Our research identifies a critical limitation in MedVLP models: their high sensitivity to variations in textual prompts within zero-shot classification tasks. These models experience significant performance degrading when using diverse styles of textual prompts, as depicted in Figure 1. Ideally, a MedVLP model should provide consistent conclusions across disease classes, regardless of the prompt style, whether it uses simplified disease names or detailed CXR descriptions. This is crucial since clinicians can recognize the disease through various descriptions. Additionally, for disease classes not seen during pre-training, the MedVLP model should benefit from detailed, highly interpretable text prompts that provide comprehensive descriptions.

In our study, we utilise the large language model (LLM) GPT-4o [16] to construct six different styles of text prompts for various diseases, each ranked by their interpretability. We evaluated three mainstream MedVLP models—BioViL[9], MedKLIP[15], and KAD[13]—using these six styles of prompts on three publicly available benchmark test datasets: ChestX-ray14 [17], CheXpert [18], and COVIDx CXR-4 [19]. Quantitative evaluation revealed that when using prompts with styles different from those used in the original pre-training, the models' performance decreased by an average of 10.17% in AUC score across all models, even when the diseases in the test set were already present in their pre-training datasets.

In addition to benchmarking MedVLP models with diverse textual prompts on zero-shot classification tasks, we analyse the varying degrees of performance degradation across different prompt styles. Based on these observations, we proposed a suggested retraining recipe for MedVLP models. This recipe is intended to help the community design robust MedVLP models that can effectively handle diverse textual prompts.

2 Related Work

2.1 General Vision Language Pre-training

Vision-Language Pre-training (VLP) learns cross-modal representations from large-scale paired image-text data for various downstream tasks. Recent studies, such as CLIP [20], ALIGN [21], ALBEF [22], and LiT [23], use contrastive learning on extensive multimodal datasets, scaling up both data and model sizes to enhance vision-language representation. Alternatively, approaches like BeiT3 [24], SLIP [25], and A-FLIP [26] focus on making VLP more cost-effective by reducing model size and data requirements while maintaining high performance. In the medical domain, VLP applies to

tasks like radiology report generation, disease diagnosis, and clinical decision-making. However, due to the fine-grained nature of these tasks and the need for clinical expertise, medical VLP (MedVLP) remains a significant challenge for ongoing research.

2.2 Medical Zero-shot Classification Task

By leveraging large, diverse datasets like radiology reports paired with CXR images [27], recent MedVLP models can identify diseases directly, without fine-tuning, by using their learned visual and textual representations to compute the similarity between input images and textual prompts. For instance, BioViL [9] redesigns vision-language models for better alignment with clinical texts, while BioViL-T [10] incorporates temporal data for enhanced zero-shot capabilities. ConVIRT [11] employs bidirectional contrastive learning to align medical images and text, and GLoRIA [12] uses an attention-based framework to learn global and local representations. MedKLIP [15] integrates external medical knowledge to improve zero-shot classification and grounding. MAVL [14] uses dual-head transformers in a multi-aspect description framework to enhance disease recognition. KAD [13] leverages Unified Medical Language System (UMLS) knowledge graphs within a query-based transformer architecture for superior zero-shot performance in CXR diagnosis. Our study aims to thoroughly investigate the sensitivity of mainstream MedVLP models to text prompts, providing insights into their robustness and generalisation capabilities.

2.3 Prompt Engineering for Zero-shot Task

Recent studies [28–31] aim to enhance general VLP models' zero-shot performance without costly retraining by using detailed, informative text prompts during inference. One such study, Xplainer [32], seeks to improve medical zero-shot diagnosis using a similar approach. However, Xplainer adopts a classification-by-description method and focuses on enhancing the explainability of zero-shot diagnosis, without fully exploring the models' adaptability to various prompt styles or thoroughly evaluating their sensitivity. Our research systematically investigates the performance of mainstream MedVLP models across diverse prompt styles with varying levels of interpretability.

3 Methods

3.1 Overview

In this study, we aim to evaluate the sensitivity of MedVLP methods to different textual prompts. Specifically, we focus on three mainstream MedVLP models: BioViL[9], MedKLIP[15], and KAD[13], which have demonstrated strong performance in zero-shot classification of CXR images in their original studies. In this section, we first provide an overview of the selected MedVLP methods, followed by the design of diverse prompt styles. These prompts are then ranked by interpretability using LLMs to further evaluate how the MedVLP models are affected by the level of prompt interpretability.

3.2 Preliminary

In this section, we introduce the three mainstream MedVLP methods utilised in our experiments to investigate the impact of diverse prompts on zero-shot CXR classification tasks. Additionally, we present the original prompt styles used in their respective studies Notably, we did not retrain or re-implement their methods, as our focus is solely on zero-shot inference with diverse prompts. Therefore, we adopted the official code and pre-trained weights from the GitHub repositories provided by the original authors. The three MedVLP methods frameworks are shown in Figure 2.

- **BioViL[9]**: BioViL is one of the first studies to introduce a CXR domain-specific VLP model. BioViL employs advanced text augmentation, regularisation techniques, and multiple pre-training strategies, resulting in significant improvements in both image and text model performance across various medical benchmarks. The text inputs used in pre-training primarily consist of sentences from the Impression and Findings sections of MIMIC-CXR radiology reports (e.g., 'Specifically, no evidence of edema.' and 'There is no focal consolidation, pleural effusion, or pneumothorax.').
- **MedKLIP[15]**: MedKLIP utilises a unique triplet extraction module to simplify radiology reports into structured triplets. It focuses on integrating domain-specific knowledge into vision-language pre-training and introduces an entity translation module that leverages a medical knowledge base to translate simple entities into informative descriptions. The text inputs used in pre-training are

descriptive sentences (e.g., 'Cardiomegaly, sometimes referred to as megacardia or megalocardia, is a medical condition in which the heart is enlarged.').

- **KAD[13]**: KAD leverages an external knowledge graph [13] to enhance auto-diagnosis for CXR. Additionally, KAD includes a transformer-based Disease Query Network (DQN) that uses disease names as queries for flexible zero-shot evaluations. The text inputs used in pre-training are simple entity names (e.g., 'pacemaker,' 'nodule,' 'pneumonia').

We also considered other notable studies in zero-shot diagnosis. For example, BioViL-T [10], an enhancement of the original BioViL [9] architecture, performs well with data containing temporal information. However, it shows minimal improvements on standard datasets, making it less relevant for our study. MAVL [14] achieves impressive results with multi-aspect disease descriptions but requires specific prompt formats during inference, reducing its flexibility and making it incompatible with our approach.

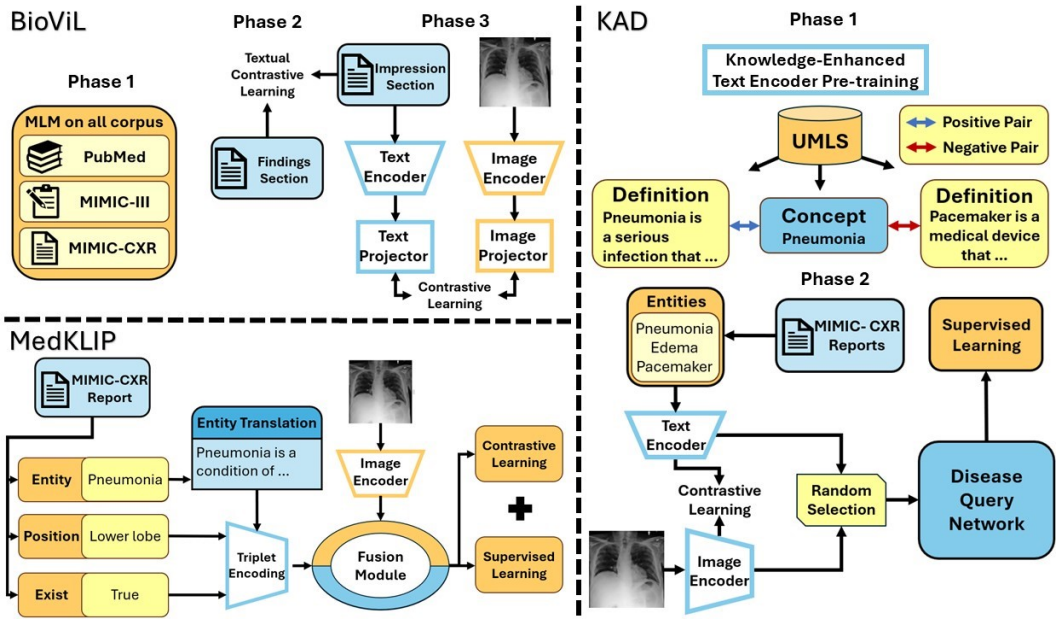

Figure 2: Framework of Three Mainstream MedVLP Models. **BioViL**: Phase 1 conducts a Masked Language Modelling (MLM) on a diverse corpus, including PubMed abstracts [33], MIMIC-III clinical notes [34], and MIMIC-CXR radiology reports [27]. Phase 2 involves textual contrastive learning between the Findings section and the Impression section of MIMIC-CXR reports. Phase 3 projects encoded image and text representations into a global space, then applies contrastive learning between them. **MedKLIP**: Pre-training involves extracting entity, position, and existence triplets from MIMIC-CXR reports. The model then translates simple entities into detailed descriptions and feeds these triplets into the fusion module together with encoded X-ray images. Lastly, it applies contrastive learning between image and text representations, and supervised learning based on the prediction results. **KAD**: Phase 1 pre-trains the knowledge-enhanced text encoder by applying contrastive learning between definition and concept pairs extracted from the Unified Medical Language System (UMLS) knowledge graph. Phase 2 applies combined contrastive learning between encoded entities and images and supervised learning on the disease query network by randomly selecting encoded entity and image pairs.

3.3 Design of Diverse Prompts

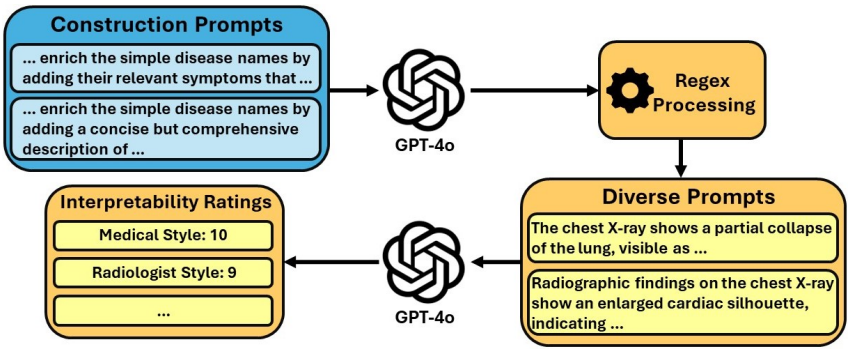

Figure 3: Pipeline of diverse prompt generation and interpretability score rating.

In this section, we describe the pipeline as depicted in Figure 3. We constructed text prompts with varying levels of interpretability and styles to test the generalisation capabilities and sensitivity of the models. Specifically, we created six distinct prompt styles for each disease class:

- **Disease Name**: This style provides a concise, specific name for the condition without additional details.
  Example: Atelectasis
  Interpretability Score: **3**
- **Disease Symptom**: This style combines the disease name with key symptoms or radiographic findings associated with the condition. It focuses on the most prominent clinical and imaging characteristics.
  Example: lung opacity, volume loss, mediastinal shift
  Interpretability Score: **7**
- **Disease Attribute**: This style gives a detailed breakdown of the disease's features, including the appearance of borders, presence of fluid, location, opacity, patterns, shape, and texture. It provides a comprehensive description that helps in visualising the condition on imaging.
  Example: border: clear but may show a shift of structures towards the collapsed area. fluid: no fluid accumulation typically present. location: localised to a segment, lobe, or entire lung. opacity: increased opacity in the affected area, appearing whiter than normal. other: displacement of structures like the diaphragm and trachea towards the affected side. patterns: linear or wedge-shaped opacity, with volume loss. shape: varies based on extent; can be a triangular or band-like appearance. texture: homogeneous texture within the collapsed region.
  Interpretability Score: **9**
- **Disease Description(Plain English)**: This style describes the disease in straightforward, easy-to-understand language, focusing on how it would appear on a CXR imaging, and what that implies. It's intended for a general audience or non-specialists.
  Example: The chest X-ray shows a partial collapse of the lung, visible as a white area where the lung tissue has lost its air content. This can appear as a dense, triangular shape pointing towards the hilum, and the affected area may cause a shift in nearby structures, such as the heart or trachea, towards the side of the collapse.
  Interpretability Score: **8**
- **Disease Description(Radiologist Style)**: This style is used by radiologists and focuses on the interpretation of imaging findings. It provides a brief summary of what is seen on the X-ray or scan, and is intended for use in radiology reports or discussions among specialists.
  Example: Radiographic findings on the chest X-ray demonstrate a partial collapse of the lung, often seen as increased density in the affected area.
  Interpretability Score: **9**
- **Disease Description(Medical Style)**: This style provides a summary of the disease with clinical terms, often used in medical reports or documentation. It describes the imaging findings and their implications in a precise, formal manner, intended for healthcare professionals.
  Example: Imaging reveals a collapse of lung tissue in the left lower lobe with volume loss and

mediastinal shift towards the affected side. Bronchial obstruction is evident.
Interpretability Score: **10**

The GPT-4o prompts used to construct the diverse prompts are included in full as supplementary tables 1, 2, 3, 4.

3.4 Prompt Interpretability Rating

We again leveraged GPT-4o to rate the six prompt styles based on interpretability, using a scale from 1 to 10. A score of 10 represents the most interpretable and informative prompt for diagnosing the disease, while a score of 1 represents the opposite. This rating system helped us systematically assess the trend of the models' performance with different input prompt styles and interpretability levels. The GPT-4o prompt used to rate the interpretability of diverse prompts are provided as supplementary table 5.

4 Experimental Setting

4.1 Datasets

The three mainstream MedVLP models tested in this study, BioViL [9], MedKLIP [15], and KAD [13], primarily utilised the MIMIC-CXR [27] dataset in their pre-training procedures. MIMIC-CXR is a publicly available dataset comprising 227,835 radiographic studies from 65,379 patients. Each study includes a corresponding CXR image and a free-text radiology report. Notably, since the MIMIC-CXR dataset was released prior to the COVID-19 pandemic, it does not contain any COVID-19 related cases.

To evaluate the generalisation capabilities of these models, we utilised three publicly available and widely-used datasets: ChestX-ray14[17], CheXpert[18], and COVIDx CXR-4[19].

- **ChestX-ray14[17]** consists of 112,120 CXR images across 14 disease classes: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening, and Hernia. All 14 diseases have corresponding samples appearing in the MIMIC-CXR dataset. Our tests strictly followed the official train-test split, using a test set that includes 25,597 chest X-ray samples.
- **CheXpert[18]** contains 224,316 CXR images. We used the official test set, which includes 500 CXR images annotated by radiologists. Following the original paper, our study focuses on the evaluation of 5 observations on the official test set: Atelectasis, Cardiomegaly, Consolidation, Edema and Pleural Effusion. All these classes have corresponding samples appearing in the MIMIC-CXR dataset.
- **COVIDx CXR-4[19]** is a major expansion of the dataset series COVIDx CXR-4. It includes 84,818 CXR images from 45,342 patients. The dataset has two classes: COVID-19 positive and COVID-19 negative. In this study, we used the official test set, which is perfectly class-balanced with 4,241 images in each category, totalling 8,482 images.

4.2 Implementation

To ensure fair comparison, all experiments were conducted on the same software environment and same device with RTX 3070 Mobile GPU. Before testing with our six prompt styles, we first evaluated each model on its baseline prompt style. The baseline style refers to the original prompt styles used in the respective studies: BioViL used "Findings suggesting + disease name", MedKLIP employed short disease descriptions, and KAD used the disease names alone. Performance on the baseline prompt serves as a benchmark, against which we compare the performance on other prompt styles to assess the models' true generalisation capabilities.

We then replaced the original prompts with the six different styles we constructed. All models used in the experiments were the original versions provided by the respective studies, without any further fine-tuning, to maintain a zero-shot setting. For KAD, which offers three different image encoder sizes (224px, 512px, and 1024px), we report results tested with the 512px-size encoder, as performance trends were similar across all sizes.

We used the image pre-processing methods described in the respective studies' original papers: For BioViL, we resized images to 512px and applied a 480px centre crop; For MedKLIP, we resized images to 224px and normalised them using global mean and standard deviation; For KAD, we

resized images to 512px and normalised them using global mean and standard deviation. Similarly, text pre-processing was also performed according to the methods outlined in the original studies.

5 Results and Analysis

We categorise the disease classes in our datasets into two groups: seen classes and unseen classes.

- Seen classes are those that appear in the MIMIC-CXR dataset, which are used during the pre-training process of all three mainstream MedVLP models we focus on in this study. This category includes 14 disease classes: Atelectasis, Cardiomegaly, Pleural effusion, Infiltration, Lung mass, Lung nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural thicken and Hernia. These disease classes come from two datasets, **ChestX-ray14** and **CheXpert**. Out of these 14 disease classes, 5 of them exist in both datasets: Atelectasis, Cardiomegaly, Consolidation, Edema and Pleural effusion. The results we present are the **macro average** of the performance from both datasets.
- Unseen classes consist of one disease class, COVID-19, which solely comes from the **COVIDx CXR-4** dataset.

The seen classes help us identify the models' sensitivity across varied prompt styles, while the unseen classes are used to test the models' zero-shot inference ability on diseases not directly learned during pre-training. An ideal model should be both robust to prompt style variations on known diseases and able to effectively use highly interpretable prompts to improve predictions on unseen diseases.

In this study, we use the Area Under the Curve (AUC) as the main metric for evaluating model performance due to its ability to provide a comprehensive measure of the models' ability to discriminate between classes across all threshold levels. In addition to AUC, we also present F1 scores and accuracy (ACC) metrics in tables included in supplementary materials to provide a more rounded assessment of the models' performance across different prompt styles. In this section, we used abbreviations for prompt styles in graphs, namely: Disease Name -> Name, Disease Symptom -> Symptom, Disease Attribute -> Attribute, Disease Description(Plain English) -> Plain ENG, Disease Description(Medical Style) -> MED Style, Disease Description(Radiologist Style) -> RAD Style.

For KAD, since the baseline style is identical to the Disease Name style, we only show the result for baseline style.

5.1 Performance on Seen Classes

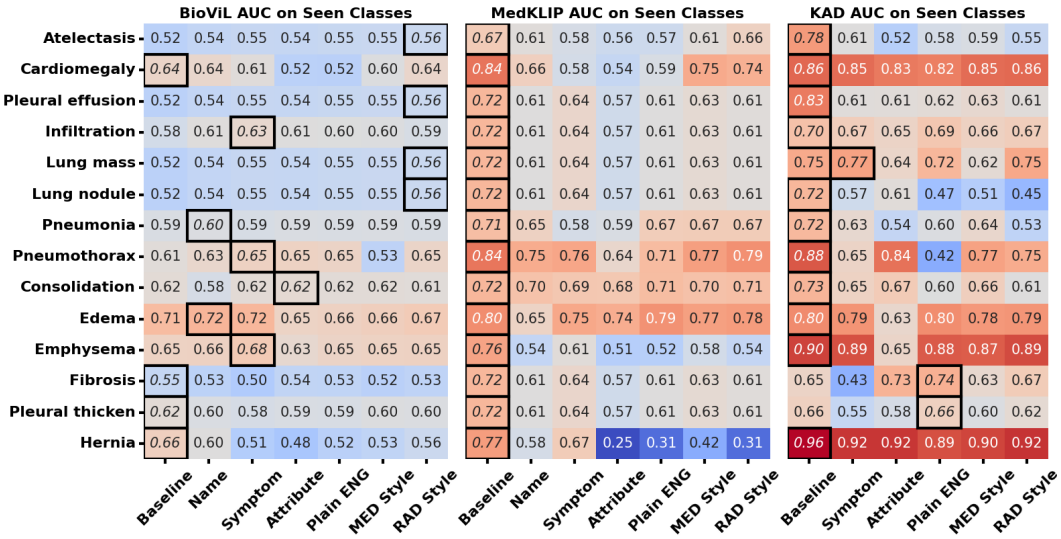

Figure 4: Heatmap demonstrating the performance of different models on **seen disease classes** with all prompt styles. The best performing prompt style of each disease class is highlighted with thick cell border and italic font.

Figure 5: Bar charts demonstrating the performance difference of non-baseline prompt styles with baseline prompt style of different models on **seen disease classes**.

In this section, we discuss the performance of the three mainstream MedVLP models on seen disease classes. To visualise the results, we provide three heatmaps in Figure4 displaying the AUC values, and three bar charts in Figure5 showing the percentage difference of non-baseline prompt styles with baseline prompt style.

BioViL demonstrates the most stable performance across various prompt styles compared to the other two models. However, it also exhibits an unremarkable overall performance, achieving a mean of 0.588 across all disease classes and prompt styles. The percentage differences in AUC between the baseline style and average of other prompt styles for BioViL range from a marginal improvement of +0.16% for the Disease Description (Radiologist Style) to a slight decrease of -2.5% for the Disease Attribute style.

MedKLIP, despite being designed to reduce dependency on specific prompt styles during training, shows a massive decrease in performance in all prompt styles that differed from the baseline style. The AUC for the Disease Description (Plain English) style, which is closest to the baseline style used in MedKLIP's training, still shows a decrease of -18.48%. The performance drops in non-baseline styles can be attributed to the model's final pre-training step, where the baseline style prompts are directly encoded and used in contrastive learning between entity descriptions and CXR images.

KAD's performance, while generally the most outstanding across the models, shows significant performance drops when tested with prompt styles different from the baseline style. The baseline style for KAD yields excellent results, reaching an average of 0.780 over all disease classes. The other prompt styles result in AUC decreases ranging from -11.2% on Disease Description (Medical Style) to -13.38% on Disease Attribute, highlighting the model's sensitivity to prompt style variation.

5.2 Performance on Unseen Classes

In this section, we discuss the performance of the three mainstream MedVLP models on unseen disease classes. We provide a bar/line chart in Figure6 to visualise the models' performance trend with varying prompt interpretability score.

For unseen classes, BioViL does not demonstrate the ability to leverage more detailed, informative prompts to learn new knowledge about these diseases. The performance is similar across different prompt styles, with no clear trend indicating improvement as prompt interpretability increases. For

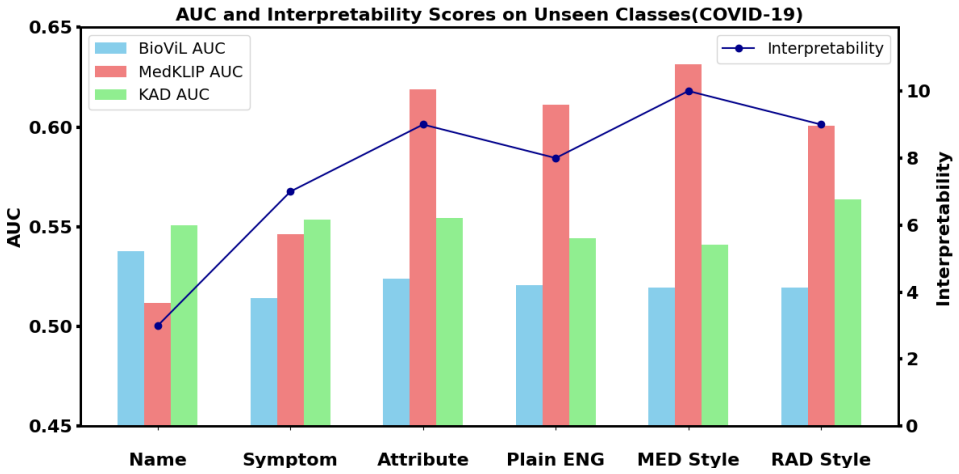

Figure 6: Graph showing both the interpretability rating of different prompt styles and the performance of different models on **unseen disease class(COVID-19)**. The right Y-axis shows the scale of interpretability scores of each prompt style. The left Y-axis shows the scale of the AUC scores.

instance, the highest AUC achieved on COVID-19 class is 0.538 with Disease Name style while the other styles achieve only 0.524 or lower.

MedKLIP shows a significant ability to learn from information within prompts during inference on unseen classes. For COVID-19, the model's AUC increases from 0.511 (Disease Name style) to 0.631 (Disease Description Medical Style). It is also notable that the model's performance scales with the prompt styles' interpretability score, reflecting the model's ability to utilise highly interpretable prompts effectively.

KAD shows some initial understanding of the disease class COVID-19, with AUC score using the simple disease names reaching 0.550. This observation can likely be explained with the fact that KAD's knowledge encoder was pre-trained with knowledge from UMLS database, which includes few pieces of introductions of this disease. However, KAD fails to demonstrate the ability to utilise high interpretability prompts, with all prompt styles yielding AUCs close to the baseline style (0.550 for Disease Name/baseline style and up to 0.564 for the Disease Description Radiologist style).

5.3 Overall Analysis

The results clearly highlight the main limitations of the current mainstream MedVLP models in handling variations in prompt styles during CXR zero-shot diagnosis.

- **BioViL:** While less sensitive to diverse prompt styles, BioViL delivers mediocre overall performance. It also shows minimal improvement when using high-interpretability prompts on unseen classes.
- **MedKLIP:** MedKLIP's performance on seen classes is heavily influenced by the prompt styles used during pre-training, with severe performance drops observed when using non-baseline styles. However, MedKLIP does exhibit noticeable performance improvement on unseen diseases when provided with highly interpretable prompts.
- **KAD:** Despite its strong overall performance, KAD is highly sensitive to variations in prompt styles. It struggles with learning from more detailed prompts on unseen diseases and shows only marginal performance gains with increasing interpretability ratings.

5.4 Proposed Vision-Language Pre-training Recipe

Based on our findings in Sections 5.1-5.3, we propose the following ideal recipe for future MedVLP model development:

- **Incorporate Domain Knowledge-Enhanced Approaches:** Both KAD [13] and MedKLIP [15] utilise external knowledge databases, incorporating informative medical domain knowledge, which significantly enhances their zero-shot diagnosis performance. This is evidenced by their superior performance compared to BioViL, in both seen and unseen classes.

- **Pre-train with Informative Text:** MedKLIP's scaling performance with prompt interpretability in unseen classes suggests that pre-training should incorporate prompts containing more descriptive and interpretable information. Including such informative text in the pre-training phase enables the model to better utilise the information in prompts during inference, thereby improving its performance.
- **Ensure Diverse Style of Text in Pre-training Dataset:** The performance inconsistency observed in MedKLIP and KAD when using baseline style prompts versus other styles underscores the importance of incorporating diverse text styles in the pre-training dataset. To enhance the model's adaptability and robustness during inference, it is crucial to include a broader range of prompt styles, ranging from simple disease names to detailed descriptions.

6 Conclusion

In this study, we conducted the first systematic evaluation of the sensitivity of three mainstream MedVLP methods to varying textual prompts across 15 distinct diseases in zero-shot classification tasks. We developed 6 unique prompt styles to replicate real-world clinical scenarios, ranked them by interpretability, and used these to assess the models' performance. Our analysis reveals that existing MedVLP models exhibit significant fluctuations in performance depending on the prompt styles, exposing a considerable gap in their robustness. Furthermore, the inconsistencies observed with more interpretable prompts suggest challenges in the models' ability to understand complex medical concepts. These findings underscore the need for further advancements in MedVLP techniques to improve their handling of diverse zero-shot prompts. We hope this work will inspire further research and innovation in the field of medical vision-language pre-training.

References

[1] Che Liu, Cheng Ouyang, Yinda Chen, Cesar César Quilodrán-Casas, Lei Ma, Jie Fu, Yike Guo, Anand Shah, Wenjia Bai, and Rossella Arcucci. T3d: Towards 3d medical image understanding through vision-language pre-training. *arXiv preprint arXiv:2312.01529*, 2023.

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

[3] Che Liu, Sibo Cheng, Chen Chen, Mengyun Qiao, Weitong Zhang, Anand Shah, Wenjia Bai, and Rossella Arcucci. M-flag: Medical vision-language pre-training with frozen language models and latent space geometry optimization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 637–647. Springer, 2023.

[4] Che Liu, Sibo Cheng, Miaojing Shi, Anand Shah, Wenjia Bai, and Rossella Arcucci. Imitate: Clinical prior guided hierarchical vision-language pre-training. *IEEE Transactions on Medical Imaging*, 2024.

[5] Zhongwei Wan, Che Liu, Mi Zhang, Jie Fu, Benyou Wang, Sibo Cheng, Lei Ma, César Quilodrán-Casas, and Rossella Arcucci. Med-unic: Unifying cross-lingual medical vision-language pre-training by diminishing bias. *Advances in Neural Information Processing Systems*, 36, 2024.

[6] Che Liu, Cheng Ouyang, Sibo Cheng, Anand Shah, Wenjia Bai, and Rossella Arcucci. G2d: From global to dense radiography representation learning via vision-language pre-training. *arXiv preprint arXiv:2312.01522*, 2023.

[7] Che Liu, Anand Shah, Wenjia Bai, and Rossella Arcucci. Utilizing synthetic data for medical vision-language pre-training: Bypassing the need for real images. *arXiv preprint arXiv:2310.07027*, 2023.

[8] Jiuming Qin, Che Liu, Sibo Cheng, Yike Guo, and Rossella Arcucci. Freeze the backbones: a parameter-efficient contrastive approach to robust medical vision-language pre-training. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1686–1690. IEEE, 2024.

[9] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision-language processing. In *European Conference on Computer Vision*, pages 1–21, 2022.

[10] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15016–15027, 2023.

[11] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Proceedings of the 38th International Conference on Machine Learning (PMLR)*, volume 182, pages 2–25, 2022. Highest Starred Implementation: `https://github.com/edreisMD/ConVIRT-pytorch`.

[12] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3942–3951, 2021. `https://github.com/marshuang80/gloria`.

[13] Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. Knowledge-enhanced visual-language pretraining on chest radiology images. *Nature Communications*, 14(1):4542, 2023.

[14] Vu Minh Hieu Phan, Yutong Xie, Yuankai Qi, Lingqiao Liu, Liyang Liu, Bowen Zhang, Zhibin Liao, Qi Wu, Minh-Son To, and Johan W. Verjans. Decomposing disease descriptions for enhanced pathology detection: A multi-aspect vision-language pre-training framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11492–11501, June 2024.

[15] Qi Wu, Xiang Jiang, Hao Luo, Yingyao Gao, Fei Feng, Hao Zhu, Yulei Huang, Yanwu Xu, Xiangyang Xue, Xiaohong Xu, and Yi Zhang. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21372–21383, 2023.

[16] OpenAI. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`, 2024.

[17] X. Wang et al. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2097–2106, 2017.

[18] J. Irvin et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019.

[19] Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports*, 10(1):19549, Nov 2020.

[20] A. Radford et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (PMLR)*, volume 139, pages 8748–8763, 2021.

[21] C. Jia et al. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning (PMLR)*, pages 4904–4916, 2021.

[22] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34:9694–9705, 2021.

[23] X. Zhai et al. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18123–18133, 2022.

[24] W. Wang et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

[25] N. Mu, A. Kirillov, D. Wagner, and S. Xie. Slip: Self-supervision meets language-image pre-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 529–544. Springer, 2022.

[26] H. Xue et al. Stare at what you see: Masked image modeling without reconstruction. *arXiv preprint arXiv:2211.08887*, 2022.

[27] Alistair E. Johnson et al. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6:1–8, 2019.

[28] Y. Ge, J. Ren, A. Gallagher, Y. Wang, M. Yang, H. Adam, L. Itti, B. Lakshminarayanan, and J. Zhao. Improving zero-shot generalization and robustness of multi-modal models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11093–11101, Los Alamitos, CA, USA, June 2023. IEEE Computer Society.

[29] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 14274–14289. Curran Associates, Inc., 2022.

[30] Muhammad Uzair Khattak, Muhammad Ferjad, Naseer Muzzamal, Luc Van Gool, and Federico Tombari. Learning to prompt with text only supervision for vision-language models. *arXiv preprint arXiv:2401.02418*, 2024.

[31] Anshul Gupta, Pierre Vuillecard, Arya Farkhondeh, and Jean-Marc Odobez. Exploring the zero-shot capabilities of vision-language models for improving gaze following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 615–624, June 2024.

[32] Chantal Pellegrini, Matthias Keicher, Ege Özsoy, Petra Jiraskova, Rickmer Braren, and Nassir Navab. Xplainer: From x-ray observations to explainable zero-shot diagnosis. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 420–429, Cham, 2023. Springer Nature Switzerland. `https://github.com/ChantalMP/Xplainer`.

[33] Pubmed. `https://pubmed.ncbi.nlm.nih.gov/`.

[34] Alistair E. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Marzyeh Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):1–9, 2016.

A Appendix / supplemental material

Supplementary Table 1: Construction prompts of diverse prompts Part A

Prompt Style	Construction Prompt
Disease Symptom	You're a helpful AI radiologist. Help me enrich the simple disease names by adding their relevant symptoms that will help the diagnosis when only looking at the patient's chest x-ray image. In other words, add the symptoms of the input disease that are commonly shown on the chest X-ray images of the patient if the patient is confirmed to have the input disease. There are no limits on how many symptoms to add, but only include the ones that are most common. Use [SEP] as the separator. Let's think step by step. Start by listing all the possible symptoms in their most common names that can be shown on the chest x-ray image of the patient if the patient is confirmed with the input disease. Then pick the highly possible/common symptoms from them. Last, compile the highly possible symptoms, and put them into . The text should be concise and follow the format in the examples below. Here is one example. In your reply, only include content after "your output:". """ My input: Lung cancer Your output: Possible symptoms: Nodule, Mass, Atelectasis, Pleural Effusion, Lymphadenopathy, Cavitation, Infiltrates, Rib Erosion. Most common symptoms: Nodule, Mass, Atelectasis. Final Output:Lung cancer [SEP] Nodule [SEP] Mass [SEP] Atelectasis """ Now process the following inputs: "atelectasis", "cardiomegaly", "pleural effusion", "infiltration", "lung mass", "lung nodule", "pneumonia", "pneumothorax", "consolidation", "edema", "emphysema", "fibrosis", "pleural thicken", "hernia", "COVID-19"

Supplementary Table 2: Construction prompts of diverse prompts Part B

Prompt Style	Construction Prompt
Disease Attribute	You're a helpful AI radiologist. Help me enrich the simple disease names by adding their relevant attributes that will help the diagnosis when only looking at the patient's chest x-ray image. Describe the disease from 8 visual attributes that describes the patient's chest X-ray image. The 8 visual attributes are: border, fluid, location, opacity, other, patterns, shape, texture. The text should be concise and follow the format in the examples below. Here are two examples. In your reply, only include content after "Your output:". """ My input: normal Your output: "border: clear and smooth, with the edge of the lung tissue appearing as a thin, curved line against the ribs.", "fluid: no fluid or effusion accumulation.", "location: fills the chest cavity, from just below the collarbones to just above the diaphragm.", "opacity: balanced, neither too opaque (white) nor too transparent (dark).", "other: symmetric appearance between two chest's sides; clear visibility of the heart, ribs, spine, and diaphragm; bronchial tubes and blood vessels are visible as white lines or tree-branch patterns against the darker lung tissue.", "patterns: no cloudy or patchy areas, no concentrated white or black spots.", "shape: lungs appear as two large, oval or triangular areas on either side of the heart.", "texture: uniform with small, branching white lines representing the bronchi and blood vessels." My input: effusion Your output: "border: clear, sharp border along the top of the effusion.", "fluid: fluid accumulation is the main feature, which causes a cloudy appearance.", "location: typically located at the base of the lungs, between the lung and chest wall.", "opacity: more opaque, appearing whiter or cloudier than the surrounding lung tissue.", "other: possible displacement of other structures such as the heart or trachea; reduction in lung volume; and increased density at the base of the lung.", "patterns: no specific patterns, effusion spreads out in the pleural space.", "shape: typically appears as a meniscus, or curved shape, at the lung base.", "texture: smooth texture, without any grainy or mottled appearance." """ Now process the following inputs: "atelectasis", "cardiomegaly", "pleural effusion", "infiltration", "lung mass", "lung nodule", "pneumonia", "pneumothorax", "consolidation", "edema", "emphysema", "fibrosis", "pleural thicken", "hernia", "COVID-19"

Supplementary Table 3: Construction prompts of diverse prompts Part C

Prompt Style	Construction Prompt
Disease Description (Plain English)	You're a helpful AI radiologist. Help me enrich the simple disease names by adding a concise but comprehensive description of the patient's chest X-ray image in plain English style. Focus on what will be shown on the patient's Chest X-ray image. The text should be concise and follow the format in the examples below. Here are two examples. In your reply, only include content after "Your output:". """ My input: lung cancer Your output: "The chest X-ray shows a suspicious mass in the left lung. This mass appears as a solid white area, indicating a potential tumor. The mass is approximately 4 centimeters in size and is located in the upper part of the left lung. There are also signs that the cancer may have spread to nearby lymph nodes, but no signs that it has spread to other parts of the body." My input: effusion Your output: "The chest X-ray shows an abnormal buildup of fluid in the space between the lungs and the chest wall, known as pleural effusion. This appears as a white area at the bottom of the lung. The fluid can cause the lung to partially collapse, making it harder to breathe." Now process the following inputs: "atelectasis", "cardiomegaly", "pleural effusion", "infiltration", "lung mass", "lung nodule", "pneumonia", "pneumothorax", "consolidation", "edema", "emphysema", "fibrosis", "pleural thicken", "hernia", "COVID-19"
Disease Description (Medical Style)	You're a helpful AI radiologist. Help me enrich the simple disease names by adding a concise but comprehensive description of the patient's chest X-ray image in plain English style. Focus on what will be shown on the patient's Chest X-ray image. The text should be concise and follow the format in the examples below. Here are two examples. In your reply, only include content after "Your output:". """ My input: lung cancer Your output: "The chest X-ray shows a suspicious mass in the left lung. This mass appears as a solid white area, indicating a potential tumor. The mass is approximately 4 centimeters in size and is located in the upper part of the left lung. There are also signs that the cancer may have spread to nearby lymph nodes, but no signs that it has spread to other parts of the body." My input: effusion Your output: "The chest X-ray shows an abnormal buildup of fluid in the space between the lungs and the chest wall, known as pleural effusion. This appears as a white area at the bottom of the lung. The fluid can cause the lung to partially collapse, making it harder to breathe." Now process the following inputs: "atelectasis", "cardiomegaly", "pleural effusion", "infiltration", "lung mass", "lung nodule", "pneumonia", "pneumothorax", "consolidation", "edema", "emphysema", "fibrosis", "pleural thicken", "hernia", "COVID-19"

Supplementary Table 4: Construction prompts of diverse prompts Part D

Prompt Style	Construction Prompt
Disease Description (Radiologist Style)	You're a helpful AI radiologist. Help me enrich the simple disease names by adding a concise but comprehensive description of the patient's chest X-ray image in professional radiologist diagnosis style. Focus on what will be shown on the patient's Chest X-ray image. The text should be concise and follow the format in the examples below. Here are two examples. In your reply, only include content after "Your output:". """ My input: lung cancer Your output: "Radiographic findings on the chest X-ray demonstrate a 4 cm irregular, well-defined mass in the left upper lobe, presenting as an opaque region with spiculated margins. The lesion shows increased density compared to surrounding tissues. There is evidence of ipsilateral mediastinal lymphadenopathy. No pleural effusion or distant metastatic lesions are identified. Findings are consistent with stage IIIA non-small cell lung carcinoma (NSCLC). Further evaluation with contrast-enhanced CT and PET-CT is recommended for comprehensive staging and assessment." My input: effusion Your output: "Radiographic findings on the chest X-ray demonstrate a large pleural effusion in the right hemithorax, characterized by a homogenous opacity with a meniscus sign. There is a noticeable blunting of the costophrenic angle and partial atelectasis of the adjacent lung parenchyma. No evidence of pneumothorax or significant mediastinal shift is observed. Further evaluation with ultrasound or contrast-enhanced CT is recommended to assess the nature and extent of the effusion." Now process the following inputs: "atelectasis", "cardiomegaly", "pleural effusion", "infiltration", "lung mass", "lung nodule", "pneumonia", "pneumothorax", "consolidation", "edema", "emphysema", "fibrosis", "pleural thicken", "hernia", "COVID-19"

Supplementary Table 5: Interpretability rating prompt

Interpretability Rating Prompt
Below are 6 different styles to describe a disease. Rate the 6 styles in terms of interpretability with scale of 1 to 10. 10 being the most interpretable and the most informative for diagnosing the disease, and 1 being the least interpretable and the least informative. *Following are the diverse prompts for all disease classes and prompt styles. Due to length issue, we omit them from this table. For all diverse prompts, see the supplementary materials of this paper.*

Supplementary Table 6: Results of zero-shot image classification on the **ChestX-ray14** dataset with the **BioViL** model. The best performing prompt style for each disease class is highlighted in bold.

Metric	Disease Class	Baseline	Name	Symptom	Attribute	Plain ENG	MED Style	RAD Style
AUC	atelectasis	0.524	0.539	0.549	0.545	0.550	0.553	**0.555**
	cardiomegaly	**0.645**	0.637	0.610	0.525	0.521	0.600	0.639
	pleural effusion	0.524	0.539	0.549	0.545	0.550	0.553	**0.555**
	infiltration	0.576	0.607	**0.634**	0.606	0.603	0.601	0.586
	lung mass	0.524	0.539	0.549	0.545	0.550	0.553	**0.555**
	lung nodule	0.524	0.539	0.549	0.545	0.550	0.553	**0.555**
	pneumonia	0.591	**0.597**	0.586	0.594	0.592	0.593	0.590
	pneumothorax	0.615	0.635	**0.654**	0.653	0.646	0.528	0.645
	consolidation	0.618	0.584	0.619	**0.624**	0.616	0.616	0.614
	edema	0.713	**0.725**	0.724	0.651	0.656	0.663	0.675
	emphysema	0.653	0.656	**0.678**	0.629	0.647	0.653	0.654
	fibrosis	**0.546**	0.527	0.500	0.536	0.534	0.521	0.532
	pleural thicken	**0.619**	0.595	0.584	0.591	0.589	0.599	0.597
	hernia	**0.659**	0.605	0.506	0.484	0.519	0.535	0.559
F1	atelectasis	0.224	0.231	0.242	0.240	0.243	0.244	**0.246**
	cardiomegaly	0.125	0.122	0.112	0.084	0.083	0.106	**0.129**
	pleural effusion	0.224	0.231	0.242	0.240	0.243	0.244	**0.246**
	infiltration	0.406	0.432	**0.458**	0.438	0.436	0.432	0.421
	lung mass	0.224	0.231	0.242	0.240	0.243	0.244	**0.246**
	lung nodule	0.224	0.231	0.242	0.240	0.243	0.244	**0.246**
	pneumonia	0.052	0.053	0.052	0.053	0.053	**0.053**	0.052
	pneumothorax	0.234	0.243	0.254	**0.254**	0.253	0.189	0.252
	consolidation	0.168	0.157	0.169	**0.171**	0.168	0.169	0.167
	edema	0.123	0.129	**0.129**	0.100	0.102	0.104	0.108
	emphysema	0.119	0.118	**0.126**	0.109	0.114	0.116	0.116
	fibrosis	**0.038**	0.036	0.033	0.037	0.036	0.035	0.036
	pleural thicken	**0.116**	0.106	0.103	0.105	0.104	0.108	0.107
	hernia	**0.012**	0.009	0.007	0.006	0.007	0.007	0.008
ACC	atelectasis	0.463	**0.536**	0.409	0.398	0.398	0.435	0.405
	cardiomegaly	0.581	0.579	0.561	0.339	0.334	0.495	**0.640**
	pleural effusion	0.463	**0.536**	0.409	0.398	0.398	0.435	0.405
	infiltration	0.511	**0.546**	0.536	0.476	0.477	0.495	0.473
	lung mass	0.463	**0.536**	0.409	0.398	0.398	0.435	0.405
	lung nodule	0.463	**0.536**	0.409	0.398	0.398	0.435	0.405
	pneumonia	0.283	0.292	0.276	0.299	0.286	**0.323**	0.282
	pneumothorax	0.382	0.383	0.414	0.423	0.446	**0.512**	0.448
	consolidation	0.322	0.323	0.344	**0.364**	0.338	0.360	0.323
	edema	0.530	0.556	**0.561**	0.412	0.420	0.439	0.451
	emphysema	0.443	0.418	**0.454**	0.343	0.366	0.378	0.379
	fibrosis	0.332	0.359	**0.367**	0.300	0.300	0.350	0.340
	pleural thicken	**0.455**	0.346	0.314	0.333	0.329	0.366	0.352
	hernia	**0.587**	0.489	0.338	0.283	0.308	0.338	0.363

Supplementary Table 7: Results of zero-shot image classification on the **ChestX-ray14** dataset with the **MedKLIP** model. The best performing prompt style for each disease class is highlighted in bold.

Metric	Disease Class	Baseline	Name	Symptom	Attribute	Plain ENG	MED Style	RAD Style
AUC	atelectasis	**0.673**	0.611	0.584	0.562	0.565	0.613	0.655
	cardiomegaly	**0.839**	0.660	0.579	0.545	0.589	0.749	0.737
	pleural effusion	**0.723**	0.606	0.639	0.569	0.606	0.625	0.611
	infiltration	**0.723**	0.606	0.639	0.569	0.606	0.625	0.611
	lung mass	**0.723**	0.606	0.639	0.569	0.606	0.625	0.611
	lung nodule	**0.723**	0.606	0.639	0.569	0.606	0.625	0.611
	pneumonia	**0.707**	0.654	0.582	0.592	0.669	0.672	0.672
	pneumothorax	**0.836**	0.745	0.757	0.636	0.713	0.770	0.795
	consolidation	**0.723**	0.696	0.686	0.682	0.710	0.700	0.705
	edema	**0.802**	0.649	0.754	0.741	0.793	0.768	0.776
	emphysema	**0.761**	0.541	0.615	0.513	0.515	0.575	0.542
	fibrosis	**0.723**	0.606	0.639	0.569	0.606	0.625	0.611
	pleural thicken	**0.723**	0.606	0.639	0.569	0.606	0.625	0.611
	hernia	**0.772**	0.581	0.672	0.253	0.307	0.423	0.306
F1	atelectasis	**0.299**	0.265	0.255	0.259	0.265	0.272	0.286
	cardiomegaly	**0.295**	0.128	0.097	0.089	0.096	0.205	0.185
	pleural effusion	**0.254**	0.173	0.187	0.168	0.180	0.194	0.191
	infiltration	**0.254**	0.173	0.187	0.168	0.180	0.194	0.191
	lung mass	**0.254**	0.173	0.187	0.168	0.180	0.194	0.191
	lung nodule	**0.254**	0.173	0.187	0.168	0.180	0.194	0.191
	pneumonia	**0.100**	0.081	0.054	0.056	0.076	0.078	0.078
	pneumothorax	**0.453**	0.339	0.359	0.237	0.284	0.378	0.400
	consolidation	**0.225**	0.207	0.204	0.199	0.208	0.210	0.211
	edema	**0.193**	0.107	0.146	0.144	0.178	0.172	0.175
	emphysema	**0.296**	0.095	0.112	0.082	0.091	0.099	0.094
	fibrosis	**0.254**	0.173	0.187	0.168	0.180	0.194	0.191
	pleural thicken	**0.254**	0.173	0.187	0.168	0.180	0.194	0.191
	hernia	**0.100**	0.011	0.016	0.007	0.007	0.007	0.007
ACC	atelectasis	0.596	0.508	0.427	0.393	0.417	0.449	**0.621**
	cardiomegaly	0.914	0.804	0.446	0.322	0.356	**0.921**	0.897
	pleural effusion	**0.808**	0.581	0.603	0.441	0.508	0.536	0.553
	infiltration	**0.808**	0.581	0.603	0.441	0.508	0.536	0.553
	lung mass	**0.808**	0.581	0.603	0.441	0.508	0.536	0.553
	lung nodule	**0.808**	0.581	0.603	0.441	0.508	0.536	0.553
	pneumonia	0.883	**0.925**	0.509	0.646	0.836	0.802	0.852
	pneumothorax	**0.851**	0.800	0.816	0.475	0.776	0.851	0.850
	consolidation	0.659	0.674	**0.743**	0.611	0.639	0.619	0.705
	edema	0.878	0.790	0.742	0.811	0.860	**0.880**	0.858
	emphysema	**0.941**	0.332	0.646	0.086	0.253	0.389	0.311
	fibrosis	**0.808**	0.581	0.603	0.441	0.508	0.536	0.553
	pleural thicken	**0.808**	0.581	0.603	0.441	0.508	0.536	0.553
	hernia	**0.991**	0.809	0.842	0.047	0.068	0.120	0.009

Supplementary Table 8: Results of zero-shot image classification on the **ChestX-ray14** dataset with the **KAD** model. The best performing prompt style for each disease class is highlighted in bold.

Metric	Disease Class	Baseline	Symptom	Attribute	Plain ENG	MED Style	RAD Style
AUC	atelectasis	**0.779**	0.606	0.516	0.581	0.592	0.549
	cardiomegaly	**0.858**	0.850	0.831	0.822	0.851	0.855
	pleural effusion	**0.828**	0.611	0.613	0.620	0.632	0.610
	infiltration	**0.700**	0.666	0.648	0.687	0.655	0.665
	lung mass	0.749	**0.765**	0.642	0.724	0.623	0.754
	lung nodule	**0.725**	0.567	0.613	0.470	0.510	0.450
	pneumonia	**0.717**	0.631	0.543	0.596	0.636	0.531
	pneumothorax	**0.877**	0.646	0.841	0.421	0.766	0.746
	consolidation	**0.727**	0.651	0.675	0.602	0.664	0.606
	edema	**0.802**	0.789	0.635	0.801	0.782	0.788
	emphysema	**0.904**	0.889	0.645	0.875	0.871	0.893
	fibrosis	0.647	0.433	0.731	**0.737**	0.633	0.671
	pleural thicken	0.657	0.546	0.583	**0.659**	0.597	0.624
	hernia	**0.955**	0.918	0.919	0.888	0.903	0.922
F1	atelectasis	**0.402**	0.266	0.237	0.258	0.259	0.241
	cardiomegaly	0.362	**0.366**	0.346	0.305	0.366	0.362
	pleural effusion	**0.537**	0.350	0.349	0.359	0.365	0.346
	infiltration	**0.481**	0.458	0.444	0.467	0.447	0.453
	lung mass	0.297	**0.314**	0.183	0.281	0.165	0.286
	lung nodule	**0.262**	0.134	0.153	0.122	0.121	0.122
	pneumonia	**0.082**	0.057	0.048	0.049	0.058	0.047
	pneumothorax	**0.505**	0.245	0.434	0.201	0.350	0.321
	consolidation	**0.215**	0.177	0.195	0.151	0.187	0.159
	edema	0.160	0.154	0.092	**0.165**	0.150	0.163
	emphysema	**0.479**	0.445	0.117	0.384	0.382	0.458
	fibrosis	0.064	0.035	0.070	**0.084**	0.050	0.058
	pleural thicken	**0.136**	0.102	0.094	0.134	0.105	0.128
	hernia	**0.537**	0.504	0.510	0.403	0.531	0.530
ACC	atelectasis	**0.815**	0.393	0.236	0.364	0.391	0.210
	cardiomegaly	0.938	0.932	**0.939**	0.929	0.938	0.928
	pleural effusion	**0.812**	0.435	0.385	0.420	0.443	0.369
	infiltration	**0.670**	0.589	0.514	0.622	0.548	0.546
	lung mass	0.904	**0.905**	0.766	0.893	0.568	0.896
	lung nodule	**0.909**	0.437	0.588	0.106	0.148	0.134
	pneumonia	**0.681**	0.467	0.317	0.224	0.452	0.148
	pneumothorax	0.858	0.564	0.815	0.190	**0.865**	0.778
	consolidation	**0.603**	0.474	0.566	0.249	0.547	0.314
	edema	0.704	0.672	0.366	0.723	0.689	**0.739**
	emphysema	**0.957**	0.955	0.460	0.946	0.947	0.952
	fibrosis	**0.819**	0.063	0.700	0.815	0.612	0.667
	pleural thicken	0.663	**0.851**	0.200	0.650	0.461	0.789
	hernia	0.997	**0.997**	0.997	0.997	0.997	0.997

Supplementary Table 9: Results of zero-shot image classification on the **CheXpert** dataset with the **BioViL** model. The best performing prompt style for each disease class is highlighted in bold.

Metric	Disease Class	Baseline	Name	Symptom	Attribute	Plain ENG	MED Style	RAD Style
AUC	atelectasis	0.646	0.644	0.750	0.753	0.751	0.750	**0.757**
	cardiomegaly	**0.720**	0.711	0.687	0.684	0.718	0.694	0.690
	consolidation	0.709	0.636	0.733	**0.749**	0.720	0.647	0.677
	edema	0.687	0.698	0.701	**0.703**	0.667	0.682	0.681
	pleural effusion	0.777	0.782	0.765	0.787	**0.799**	0.755	0.764
F1	atelectasis	0.496	0.495	0.611	0.615	0.612	0.610	**0.619**
	cardiomegaly	**0.573**	0.561	0.538	0.532	0.568	0.545	0.537
	consolidation	0.171	0.141	0.192	**0.205**	0.182	0.148	0.157
	edema	0.442	0.457	**0.464**	0.427	0.382	0.396	0.412
	pleural effusion	0.522	0.524	0.500	0.535	**0.557**	0.541	0.502
ACC	atelectasis	0.623	0.603	0.735	**0.738**	0.732	0.731	0.738
	cardiomegaly	0.714	0.701	**0.763**	0.689	0.690	0.720	0.654
	consolidation	0.551	0.490	0.623	**0.653**	0.597	0.536	0.516
	edema	0.849	0.850	**0.855**	0.807	0.796	0.795	0.820
	pleural effusion	0.710	0.707	0.674	0.725	0.750	**0.784**	0.683

Supplementary Table 10: Results of zero-shot image classification on the **CheXpert** dataset with the **MedKLIP** model. The best performing prompt style for each disease class is highlighted in bold.

Metric	Disease Class	Baseline	Name	Symptom	Attribute	Plain ENG	MED Style	RAD Style
AUC	atelectasis	**0.870**	0.844	0.825	0.811	0.839	0.838	0.787
	cardiomegaly	**0.899**	0.802	0.771	0.417	0.761	0.816	0.806
	consolidation	0.897	0.896	0.816	0.852	**0.905**	0.805	0.856
	edema	**0.924**	0.685	0.775	0.770	0.902	0.845	0.895
	pleural effusion	**0.909**	0.801	0.821	0.739	0.865	0.848	0.847
F1	atelectasis	**0.682**	0.663	0.639	0.649	0.652	0.649	0.584
	cardiomegaly	**0.706**	0.573	0.577	0.427	0.563	0.610	0.603
	consolidation	**0.456**	0.434	0.323	0.425	0.411	0.333	0.360
	edema	**0.621**	0.313	0.413	0.404	0.575	0.503	0.544
	pleural effusion	**0.648**	0.493	0.529	0.499	0.574	0.565	0.549
ACC	atelectasis	**0.811**	0.790	0.789	0.796	0.768	0.765	0.678
	cardiomegaly	**0.819**	0.744	0.722	0.337	0.713	0.759	0.714
	consolidation	**0.934**	0.909	0.867	0.930	0.934	0.915	0.913
	edema	0.882	0.723	0.807	0.831	0.855	0.868	**0.898**
	pleural effusion	**0.873**	0.786	0.811	0.742	0.831	0.842	0.811

Supplementary Table 11: Results of zero-shot image classification on the **CheXpert** dataset with the **KAD** model. The best performing prompt style for each disease class is highlighted in bold.

Metric	Disease Class	Baseline	Symptom	Attribute	Plain ENG	MED Style	RAD Style
AUC	atelectasis	**0.847**	0.728	0.626	0.765	0.750	0.730
	cardiomegaly	0.860	0.867	0.853	0.776	0.844	**0.871**
	consolidation	**0.867**	0.702	0.697	0.768	0.728	0.713
	edema	**0.932**	0.904	0.714	0.911	0.877	0.884
	pleural effusion	**0.964**	0.701	0.783	0.755	0.766	0.765
F1	atelectasis	**0.637**	0.524	0.466	0.557	0.541	0.550
	cardiomegaly	0.658	0.677	0.655	0.568	0.632	**0.680**
	consolidation	0.250	0.275	0.148	0.277	0.234	**0.314**
	edema	**0.667**	0.593	0.298	0.601	0.516	0.544
	pleural effusion	**0.783**	0.416	0.474	0.449	0.451	0.453
ACC	atelectasis	**0.794**	0.628	0.527	0.650	0.741	0.622
	cardiomegaly	0.764	**0.805**	0.752	0.695	0.723	0.803
	consolidation	**0.953**	0.942	0.425	0.845	0.867	0.945
	edema	**0.900**	0.867	0.433	0.873	0.806	0.838
	pleural effusion	**0.922**	0.600	0.730	0.770	0.647	0.766

Supplementary Table 12: Results of zero-shot image classification on the **COVIDx CXR-4** dataset with the **BioViL** model. The best performing prompt style for each disease class is highlighted in bold.

Metric	Disease Class	Baseline	Name	Symptom	Attribute	Plain ENG	MED Style	RAD Style
AUC	COVID-19	0.499	**0.538**	0.514	0.524	0.521	0.520	0.520
F1	COVID-19	0.467	0.507	0.480	0.472	0.513	0.495	**0.531**
ACC	COVID-19	0.499	**0.538**	0.514	0.524	0.521	0.520	0.520

Supplementary Table 13: Results of zero-shot image classification on the **COVIDx CXR-4** dataset with the **MedKLIP** model. The best performing prompt style for each disease class is highlighted in bold.

Metric	Disease Class	Baseline	Name	Symptom	Attribute	Plain ENG	MED Style	RAD Style
AUC	COVID-19	0.594	0.512	0.546	0.619	0.611	**0.631**	0.600
F1	COVID-19	**0.677**	0.667	0.668	0.674	0.668	0.675	0.667
ACC	COVID-19	**0.561**	0.501	0.507	0.548	0.516	0.536	0.502

Supplementary Table 14: Results of zero-shot image classification on the **COVIDx CXR-4** dataset with the **KAD** model. The best performing prompt style for each disease class is highlighted in bold.

Metric	Disease Class	Baseline	Symptom	Attribute	Plain ENG	MED Style	RAD Style
AUC	COVID-19	0.551	0.553	0.555	0.544	0.541	**0.564**
F1	COVID-19	0.663	0.540	0.663	0.660	**0.665**	0.662
ACC	COVID-19	0.524	**0.540**	0.529	0.531	0.527	0.530

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in both the Abstract and the Introduction1 clearly states the paper's contributions and scope, matching experimental results displayed in the Results and Analysis section5.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [NA]

   Justification: This study has no major limitations. We successfully carried out all the research we wished to conduct for this study.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions/hypothesis we made in the Introduction and Abstract are clearly proved in the Results and Analysis section5.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide the hardware environment used to conduct the experiments in Experimental Setting section4. We also provide the full code used to reproduce the experiment results as supplementary materials in the submission.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the full code used to reproduce the experiment results as supplementary materials in the submission. We also provide links to the official websites hosting the datasets used to conduct the experiments in supplementary materials.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide the full code used to reproduce the experiment results as supplementary materials in the submission. We also provide the test details in the Experimental Setting section4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Error bars or other statistical significance tests are not necessary for the purpose of this study.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the full code used to reproduce the experiment results as supplementary materials in the submission and testing parameters are defined within. We also provide the hardware resource used to conduct the experiments in the Experimental Setting section4..

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our study conform the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We clearly state the potential positive societal impacts in Introduction section1 and Conclusion section6. We do not think this study has any potential negative societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This study does not provide any trained models or release of data that has any risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We provide citations and provide URLs for all datasets used. We also provide citations to the original work of models we used to conduct the experiments.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This study does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This study does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This study does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.