Self-supervised Multi-modal Training from Uncurated Image and Reports Enables Zero-shot Oversight Artificial Intelligence in Radiology

Sangjoon Park*[1], Eun Sun Lee†[,2], Kyung Sook Shin[3], Jeong Eun Lee†[,3], and Jong Chul Ye†[,‡,4]

[1]Department of Radiation Oncology, Yonsei College of Medicine, Seoul, Korea
[2]Chung-Ang University Hospital, Seoul, Korea
[3]Department of Radiology, Chungnam National University Hospital, Chungnam National University College of Medicine, Daejeon, Korea
[4]Kim Jaechul Graduate School of AI, KAIST, Daejeon, Korea
*This work was mainly conducted when the first author was affliated with KAIST.
†Co-corresponding authors.
‡Correspondence should be addressed to J.C.Y. (jong.ye@kaist.ac.kr)

Oversight AI is an emerging concept in radiology where the AI forms a symbiosis with radiologists by continuously supporting radiologists in their decision-making. Recent advances in vision-language models sheds a light on the long-standing problems of the oversight AI by the understanding both visual and textual concepts and their semantic correspondences. However, there have been limited successes in the application of vision-language models in the medical domain, as the current vision-language models and learning strategies for photographic images and captions call for the web-scale data corpus of image and text pairs which was not often feasible in the medical domain. To address this, here we present a model dubbed Medical Cross-attention Vision-Language model (Medical X-VL), leveraging the key components to be tailored for the medical domain. Our medical X-VL model is based on the following components: self-supervised uni-modal models in medical domain and fusion encoder to bridge them, momentum distillation, sentence-wise contrastive learning for medical reports, and the sentence similarity-adjusted hard negative mining. We experimentally demonstrated that our model enables various zero-shot tasks for oversight AI, ranging from the zero-shot classification to zero-shot error correction. Our model outperformed the current state-of-the-art models in two different medical image database, suggesting the novel clinical usage of our oversight AI model for monitoring human errors. Our method was especially successful in the data-limited setting, which is frequently encountered in the clinics, suggesting the potential widespread applicability in medical domain.

In recent years, deep learning has made significant strides in the development of vision and language models, particularly in the medical field, bringing us closer to human-level intelligence. However, despite these successes, there has been limited progress in building models that can correlate visual and language concepts, unlike the human perception that can seamlessly integrate both modalities. This issue has been a long-standing topic of research in the field of artificial intelligence (AI)[1]. Fortunately, recent advances in vision-language models, a multi-modal model trained on a vast corpus of image-text pairs with the goal of learning shared concepts between images and texts, has led to remarkable results in downstream tasks such as image-text retrieval, vision question answering, visual grounding, and more, which require a deep understanding of both visual and language information. Consequently, the vision-language model has revolutionized the field of multi-modal vision-language research, leading to a significant number of studies in recent years[2–8].

The rapid advances of VLP have been indebted to the introduction of vision transformer (ViT)[9], which processes the images as a set of small patches similar to those of several words for a sentence with the transformer model for natural language processing[10]. Thanks to the intrinsic similarities between the ways of processing images and sentences through the self-attention mechanism of the transformer, direct attention between the image patches and words are possible, facilitating more straightforward cross-attention between modalities. The recent works have demonstrated that the transformer-based vision-language models trained with the web-scale image and text data pairs have the generic capability for multiple downstream vision-language tasks[2,11–14]. Compared with individual models specialized for each task and modality, the vision-language models trained with massive

data exhibit superior performances along with the amortization of training cost, enabling to push the limit of model capacity for both domains to reach human-level performances.

In medical fields, rather than completely replacing clinicians' decision-making tasks, there is also an increasing need for oversight AIs to alert clinically significant abnormalities or to detect and correct rare but critical errors in their clinical decision-making. In particular, the decision of radiologists usually comes in the form of medical reports so that visual language models that can understand both the medical images and the reports are an essential step toward the wide acceptance of an oversight AI.

Thanks to the recent surge of interest in introducing AI into the medical field, there has been a proliferation of self-supervised pre-trained models that are specialized in specific medical domain data and modalities[15–19] Furthermore, uncurated medical data such as radiograph image and report pairs are already abundant in hospitals, but the absence of manual annotation to discrete labels for traditional supervised learning impedes the utilization of those uncurated image and text pairs to build a robust model. Therefore, making the model directly learn from the uncurated image-report pairs will greatly increase the usability of data, thereby enabling to develop a robust model that can efficiently adapt to various downstream tasks. Nevertheless, there exist limited studies on the vision-language models in the medical domain[15–18,20–22] applied for narrow range of tasks, where the pairs of image and sentence are frequently used in the form of radiographs, pathology slides, and corresponding reports, and there is no study proposing the method to bride the uni-modal pre-trained vision and language models in medical domain to build the robust multi-modal model, enabling the zero-shot oversight AI in various applications.

Directly adapting the vision-language model in computer vision to the medical domain may result in suboptimal performance due to the different characteristics between the two domains. Compared with the photographic images and captions where billion-scale image-text pairs can be utilized with web crawling[2,11], the amount of image-text pairs for medical images is often not sufficient to enable learning a firm relation between visual semantics and textual concept. Furthermore, the diversities between the different images and reports are often subtle in medical domains than photographic images. For radiographs as an example, the standardized imaging protocols make them consistent in anatomical patterns, and the abnormal findings in radiographs are usually subtly different in appearance[23]. Likewise, the medical reports usually take the confined words and the sentence structures for a better workflow, producing the structured patterns of words in a sentence except for some keywords to describe the key findings. In addition, there also exist linguistic challenges in medical domain, ranging from the difference in sentence structure like the common usage of the negation to the frequent use of the domain-specific medical terms rarely used in the general domain. For instance, the *"There is no boy in the picture"* would be rather awkward and unlikely to be used in the caption for the photographic image, but the descriptions like *"There is no finding suggesting pneumonia"* are frequently used in radiology reports as it provides important information about the absence of abnormal findings. These discrepancies indicate the need for the text and image encoders specialized for the medical domain, as well as a novel model architecture that can bridge them.

To address this, here we present a model dubbed *Medical Cross-attention Vision-Language model (Medical X-VL)*, leveraging the key components to be tailored for the medical domain. Our medical X-VL model is based on the following components: self-supervised uni-modal models in medical domain and fusion encoder to bridge them, momentum distillation, sentence-wise contrastive learning for medical reports, and the sentence similarity-adjusted hard negative mining. We experimentally demonstrated that our model outperforms the current state-of-the-art medical vision-language and self-supervised models in tasks for the oversight AI, ranging from the zero-shot disease detection to detect not only seen disease classes during the pre-training but also unseen disease class like COVID-19 to zero-shot detection and correction of human errors, which can be catastrophic and even life-threatening (1). In addition to the quantitative measures, we performed qualitative analysis on the medical X-VL model by visualizing the cross-attention between images and words, providing another merit to visualize the word-region level attention, providing the transparent interpretation on the model's behavior. Finally, we extended our model to another medical domain data with limited number, to validate the adaptability of our model in wider clinical applications (Figure 1).

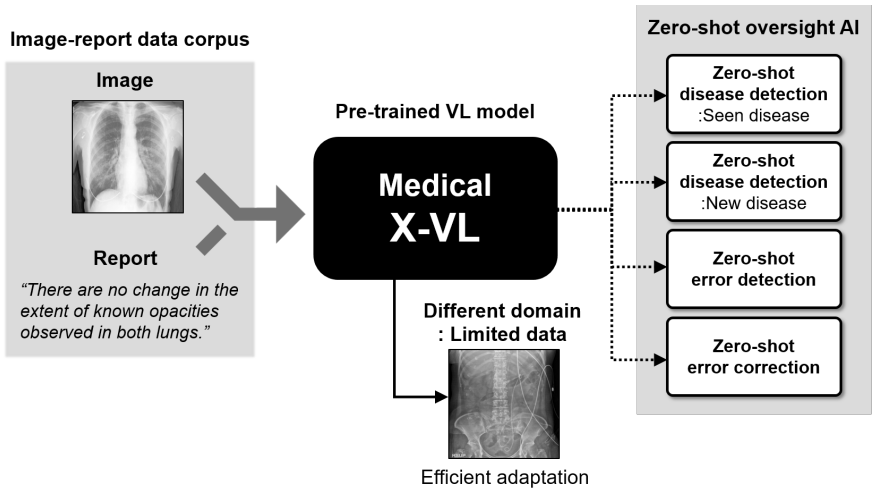


Figure 1: **The proposed vision-language model, medical X-VL, is self-supervised on an uncurated image-report corpus, and can be used for various zero-shot oversight AI tasks and facilitates efficient adaptation to medical data from different domains.**

Results

Overview of the proposed model Currently, most vision-language models can be classified into two distinct architectures: the parallel dual encoder architecture and the multi-modal fusion encoder architecture. The parallel dual encoder architecture refers to a model that uses cross-modal contrastive loss to allow uni-modal encoders to embed the encoded features in the same embedding space, as suggested in pioneering works in the general domain, such as CLIP[11] and ALIGN[2], as well as in the medical domain[15,17,18] (Supplementary Fig. S1a). While this approach is intuitive and capable of positioning vision and text representations in the same embedding space, it relies solely on contrastive loss, which requires a large number of image-text pairs to achieve proper alignment in the embedding space. This can be a significant limitation, particularly in the medical domain, where there may be insufficient image-text pairs compared to the general domain. Additionally, since this approach uses parallel uni-modal encoders without a bridging module between them, there may be limitations in utilizing downstream applications that require multi-modal representations, such as image-guided text generation or completion. In contrast, the design using a multi-modal fusion encoder performs direct fusion between image and text representations using self-attention (Supplementary Fig. S1b) or cross-attention (X-attention) (Supplementary Fig. S1c). This approach enables obtaining an image-text fused representation, which can be utilized to perform various downstream tasks requiring multi-modal understanding. Several medical vision-language models have demonstrated the ability to perform diverse tasks requiring multi-modal understanding in the medical domain using this structure[20,21]. However, the approach of jointly learning image and text inputs without cross-modal alignment often faces challenges in ensuring that the multi-modal encoder successfully learns image-text interaction, as visual and textual features are not aligned.

A recently proposed method[3] leverages the strengths of both designs by introducing CLIP-style pre-alignment before the multi-modal fusion. Furthermore, to overcome the limitation that the web-scale database used for VLP in the general domain contains many noisy pairs, a momentum distillation method utilizing knowledge distillation from a momentum teacher model was employed to obtain informative features that cannot be obtained through one-hot labels. These innovations make it possible for the model to achieve SOTA performance in various vision-language tasks with relatively few data and enable data-efficient learning of the vision-language model. For more detailed information on the concept of the momentum distillation, please refer to the Supplementary material. Building upon prior research and models, we introduce designs that are suitable for medical domain data, specifically medical image and report pairs. Figure 1 illustrates the proposed medical X-VL model and the learning approach devised for the medical domain.

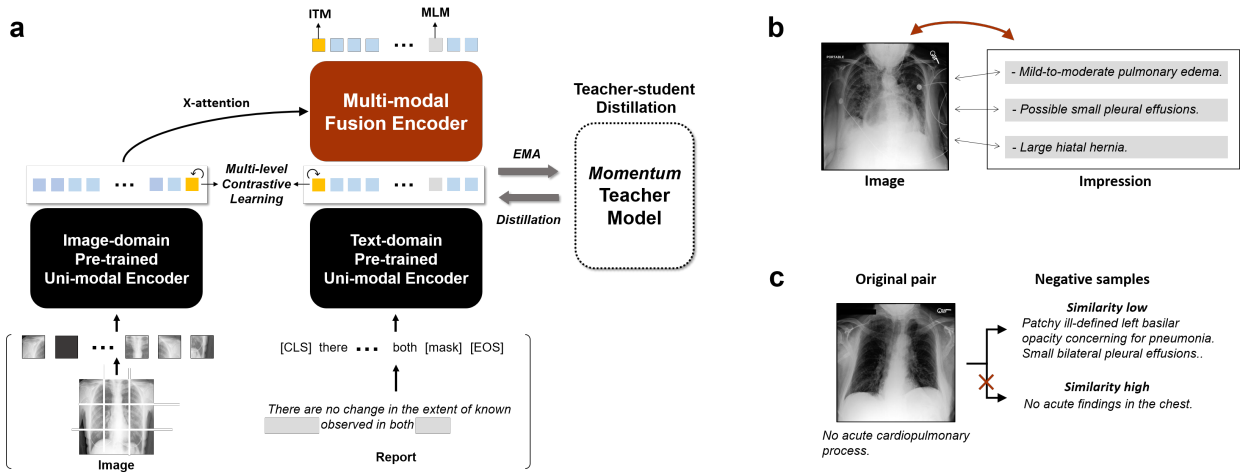


Figure 2: **Architecture and key components of the medical X-VL model. (A) The model is based on a cross-attention (X-attention) based multi-modal vision language architecture that utilizes contrastive learning and momentum distillation methods. The cross-attention-based fusion encoder connects the uni-modal encoders that are self-supervised in each uni-modal domain. (B) To account for medical reports being composed of multiple sentences, contrastive learning is introduced between each sentence and image. (C) Since similar reports belonging to the same class are likely to exist between medical reports, text similarity is calculated to exclude those with high similar, thereby improving negative mining performance.**

Based on the models proposed in previous work[3], we introduced contraptious methods suitable for medical domain data, specifically medical image and report pairs. Figure 2a illustrates the proposed learning method consisting of a fusion encoder that performs cross-attention between uni-modal encoders, which are self-supervised model using their respective domain data. As the domain-specific self-supervised model, we used DINO[24] pre-trained model on MIMIC-CXR training set as the visual encoder and the self-supervised CXR-BERT model[17] as the text encoder. This approach enables stable training by enabling the fusion module to work mainly to bridge the pre-trained uni-modal encoders during the learning process of vision-language model.

Additionally, we proposed an additional contrastive learning suitable for medical domain data as shown in Figure 2b. Unlike a short sentence that describes the overall content of an image in the general domain, a medical report usually consists of several sentences, each representing a specific observation. Inspired by previous studies that used the global-to-local as well as the global-to-global contrastive learnings to improve the performance[25], we introduced contrastive learning between individual sentences and images to ensure that the individual observations can align well with the visual representation. The masked language modeling (MLM) loss enables fine-grained correlation between the image and individual words by predicting the masked words with the help of image information. To compute the image-text matching (ITM) loss that directly determines whether a given image-text pair is correct, it is necessary to construct negative samples. Although hard negative mining proposed in previous works is an effective negative sampling method, it may select a positive sample as a negative sample in medical domain data where multiple images may exhibit the same observation and therefore semantically identical image or report can be sampled as the negative sample in the same batch. Therefore, we alleviate this issue by calculating the text similarity, as shown in Figure 2c, to avoid selecting those with high text similarities. Ablating any of these elements resulted in sub-optimal performance, as shown in the Supplementary Table 1, indicating that each component is essential for achieving optimal performance. For a detailed description of the model architecture and learning objectives, refer to the Method section.

To enable the detection of clinically significant abnormalities and oversight AI for detecting human errors without explicit labeling, we leveraged zero-shot learning. For this purpose, we utilized the MIMIC-CXR dataset, which includes 377,110 image pairs. We trained the model using 371,951 images excluding the test set.

Table 1: Performance of zero-shot oversight AI to alert clinically significant abnormality

	Average	Atelectasis	Cardiomegaly	Consolidation	Pulmonary edema	Pleural effusion
AUC						
X-VL (simple)	0.879 (0.842-0.911)	0.803 (0.763-0.839)	0.874 (0.841-0.903)	0.880 (0.830-0.923)	0.903 (0.867-0.934)	0.934 (0.910-0.955)
X-VL (detailed)	0.881 (0.843-0.912)	0.812 (0.773-0.850)	0.869 (0.835-0.899)	0.897 (0.848-0.936)	0.901 (0.863-0.930)	0.924 (0.896-0.948)
CheXzero (simple)	0.878 (0.839-0.913)	0.787 (0.744-0.826)	0.898 (0.870-0.923)	0.904 (0.838-0.955)	0.888 (0.856-0.919)	0.917 (0.888-0.941)
CheXzero (detailed)	0.830 (0.782-0.873)	0.696 (0.647-0.743)	0.863 (0.830-0.894)	0.809 (0.723-0.886)	0.877 (0.840-0.913)	0.902 (0.872-0.930)
F1						
X-VL (simple)	0.630 (0.552-0.705)	0.666 (0.607-0.718)	0.692 (0.636-0.745)	0.413 (0.286-0.552)	0.638 (0.554-0.709)	0.740 (0.677-0.798)
X-VL (detailed)	0.629 (0.552-0.701)	0.653 (0.595-0.705)	0.693 (0.635-0.747)	0.433 (0.308-0.558)	0.629 (0.548-0.702)	0.739 (0.673-0.796)
CheXzero (simple)	0.645 (0.565-0.720)	0.641 (0.588-0.699)	0.748 (0.692-0.799)	0.515 (0.369-0.657)	0.612 (0.538-0.677)	0.708 (0.638-0.770)
CheXzero (detailed)	0.579 (0.504-0.652)	0.547 (0.489-0.601)	0.694 (0.638-0.745)	0.351 (0.235-0.473)	0.610 (0.530-0.687)	0.693 (0.628-0.754)
Radiologists*	0.615 (0.552-0.672)	0.692 (0.646-0.731)	0.678 (0.634-0.718)	0.385 (0.280-0.485)	0.583 (0.511-0.645)	0.737 (0.689-0.783)

* Results are from the previous work[18].

Zero-shot oversight AI to alert clinically significant abnormality Detecting clinically significant abnormalities using zero-shot learning can be considered a zero-shot classification problem for the abnormal class. The model's performance was evaluated on the CheXpert competition's test set data of 500 images[26], classifying five abnormality labels, atelectasis, cardiomegaly, consolidation, pulmonary edema, and pleural effusion. Evaluation was conducted in two ways: one was calculating the matching score between the simple prompts of *"pathology"* and *"no pathology"* as proposed in a previous work[18], and the other was calculating the score between the *class-specific detailed descriptions* selected by clinicians and *"no pathology"* as proposed in another work[15]. For example of the detailed description for a given abnormality class, refer to Supplementary Fig. S2. Through the latter method, the model's understanding of fine-grained details supporting the class estimation can be evaluated, unlike just simple class classification with simple prompts. We mainly compared our model with the SOTA zero-shot classification model, CheXzero, as well as other medical vision-language models and self-supervised models.

Table 1 shows the model performance compared to the current SOTA model and radiologists. Using the simple prompts, the model showed excellent performance that was not statistically significantly different from the previous SOTA model in terms of area under the receiver operating characteristic curve (AUC), based on the average of metrics for the five classes. Similarly, in terms of F1 score, the model showed better performance based on the average of five classes than the radiologists as well as the the current SOTA model, albeit not statistically significant. When using detailed description for each abnormality class, the improved performances of the proposed model over the current SOTA model were pronounced. Specifically, in terms of AUC, the proposed model showed statistically significantly better performances for detection of atelectasis, and also showed higher performances in all other labels without statistical significance. In terms of F1-score, the proposed model showed the trend toward better detection performances based on the average of metrics for all abnormality classes than both the current SOTA model and the radiologists. Compared to the current SOTA model, the proposed model demonstrated trends toward better results in the metric averaged over the five classes. In the zero-shot detection utilizing detailed descriptions for each class, the current SOTA model showed a significant degradation in performance, indicating that it is inadequate for detecting abnormalities based on the detailed descriptions for each class, rather than just the class name. On the other hand, the proposed model exhibited almost no compromise in performance, showing that the model more accurately understands the detailed descriptions that can explain each abnormality class.

Table 2: Comparison of zero-shot classification performance with medical vision-language models and self-supervised models

	Model	Mean AUC
Supervised	DenseNet-121*	0.901
Self-supervised	GLoRIA*	0.534
	ConVIRT-ResNet-50-1%*	0.870
	ConVIRT-ResNet-50-10%*	0.881
	ConVIRT-ResNet-50-100%*	0.881
	ConVIRT-ViT-1%*	0.725
	ConVIRT-ViT-10%*	0.809
	ConVIRT-ViT-100%*	0.856
	CheXzero	0.878
	X-VL (ours)	0.881

* Results are from the previous work[18].

Table 3: Zero-shot error detection performance for the critical radiology report errors

	Average	Mismatch	Location	Extent	False-negative	False-positive
AUC						
X-VL	0.855 (0.796-0.905)	0.956 (0.937-0.972)	0.724 (0.640-0.797)	0.703 (0.602-0.804)	0.926 (0.890-0.958)	0.966 (0.913-0.995)
CheXzero	0.744 (0.686-0.799)	0.826 (0.797-0.854)	0.589 (0.516-0.666)	0.638 (0.552-0.717)	0.829 (0.798-0.865)	0.839 (0.765-0.895)
F1						
X-VL	0.516 (0.392-0.631)	0.621 (0.526-0.691)	0.315 (0.206-0.431)	0.352 (0.215-0.500)	0.623 (0.522-0.710)	0.672 (0.493-0.821)
CheXzero	0.171 (0.126-0.227)	0.288 (0.237-0.339)	0.098 (0.063-0.146)	0.113 (0.073-0.156)	0.238 (0.188-0.292)	0.119 (0.068-0.200)

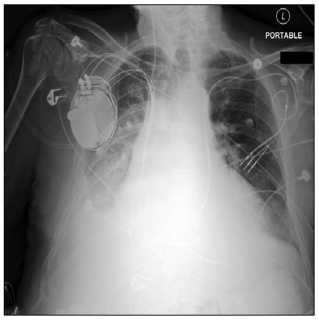
Table 2 presents the performance comparison of our proposed model with various medical vision-language models and self-supervised models other than the current SOTA model, CheXzero. The results indicate that our approach achieves better or comparable performance to self-supervised models that have been fine-tuned with some (1-50%) or all (100%) labeled data as well as medical vision-language models without the need for explicit label training.

Combined, the results demonstrate that our model can perform zero-shot oversight AI to alert clinically significant abnormalities with a more fine-grained understanding of various abnormalities.

Zero-shot oversight AI to detect critical radiology report error In radiology, errors can occur in many aspects through the reading process, which lead to fatal results though not frequently occur. Based on the result that our model showed the best performance in correlating the detailed components of image and text, we verified whether the trained model can detect critical human errors in a zero-shot manner. Inspired by the previous studies[27,28], we classified human errors into five classes for simulation: image-report mis-registration (mismatch error), error in description for location (location error), error in description for extent (extent error), false estimate of no finding (false-negative error), and false estimate of abnormal finding (false-positive rror). We designed an error generator to produce these errors with a probability of 5%, and evaluated whether the vision-language model could detect the existence of these errors without any fine-tuning process. For more details on the simulation of the radiology report errors, refer to the Method section.

Overall, our model significantly outperformed the current SOTA model in the oversight task of detecting human errors (Table 3). When examining each type of error, our proposed model provided significantly better performances for the mismatch, false-negative and false-positive errors in terms of AUC, although the perfor-

a **Location error**



True: In comparison with the study of ___ the monitoring and support devices are unchanged. there is again substantial enlargement of the cardiac silhouette with pulmonary vascular congestion and bilateral pleural effusions more prominent on the right.

Wrong: In comparison with the study of ___ the monitoring and support devices are unchanged. there is again substantial enlargement of the cardiac silhouette with pulmonary vascular congestion and bilateral pleural effusions more prominent on the left.

Corrected: In comparison with the study of ___ the monitoring and support devices are unchanged. there is again substantial widening of the cardiac silhouette with pulmonary vascular congestion and bilateral pleural effusions more prominent on the right.

b **Extent error**

True: New consolidation at the base the left lung could be either atelectasis or pneumonia, accompanied by stable small left pleural effusion. Chest is otherwise unchanged, including normal size heart, minimally dilated upper lobe pulmonary vessels, but no pulmonary edema.

Wrong: New consolidation at the base the left lung could be either atelectasis or pneumonia, accompanied by stable large left pleural effusion. Chest is otherwise unchanged, including normal size heart, minimally dilated upper lobe pulmonary vessels, but no pulmonary edema.

Corrected: New consolidation at the base the left lung could be either atelectasis or pneumonia, accompanied by stable small left pleural effusion. Chest is otherwise unchanged, including normal size heart, minimally dilated upper lobe pulmonary vessels, but no pulmonary edema.

Figure 3: **(A) Example for the correction of the location error. Notably, other than the wrong words ("left" → "right"), it also changes another word ("enlargement") to another ("widening") without significantly changing the meaning. (B) Example for the correction of the extent error ("large" → "small").**

mances for the other error classes were also better without statistical significance. In terms of F1-score, the proposed model showed better detection performances for all error types with statistical significances. Considering that detecting errors in sentences requires fine-grained understanding of the components of the sentence than simply classifying image with the corresponding class, the result again demonstrates the superiority of our model in terms of detailed understanding of sentence components.

The structure of the medical X-VL model with a multi-modal fusion module provides another merit of enabling zero-shot "correction" as well as zero-shot detection of errors. Thanks to the image-guided MLM in pre-training objectives, the model can correct erroneous expression by substituting wrong words (red text) with correct ones (blue text) referring to the image when masking each word within a sentence one by one and predicting it (Figure 3). The model was able to replace not only the words with "location error (Figure 3a)" but also words with "extent error (Figure 3a)" with those of similar meanings. Interestingly, it was observed that even another word (e.g. enlargement) without error was replaced with different words with similar meanings (e.g. widening). This behavior of the model was understandable, considering the nature of one-by-one substitution approach using MLM by the model.

Model applicability to unseen disease and scalability of uni-modal pre-trained model to multi-modal understanding The MIMIC-CXR dataset consists of patient data collected from 2011 to 2016 and does not include cases of coronavirus disease (COVID-19) that first emerged in late 2019. Therefore, for models trained with vision-language pre-training using MIMIC-CXR, COVID-19 can be considered an unseen disease. The detailed findings (bilateral peripheral and basal multifocal airspace ground glass opacity or consolidation) indicating COVID-19 can also be observed in other infectious diseases, suggesting that accurate diagnosis of new diseases can be achieved with high accuracy by detecting the presence of these detailed findings. Table 4 shows zero-shot

Table 4: Zero-shot detection performance for unseen abnormality of COVID-19

	Direct class name	Detailed impression
AUC		
X-VL	0.799	0.800
	(0.781-0.816)	(0.785-0.816)
CheXzero	0.684	0.778
	(0.664-0.702)	(0.760-0.795)
F1		
X-VL	0.823	0.802
	(0.809-0.836)	(0.788-0.815)
CheXzero	0.722	0.798
	(0.707-0.737)	(0.784-0.811)

detection performances of unseen abnormality of COVID-19. When using the detailed description for zero-shot detection, there was no statistically significant difference between the proposed model and the current SOTA model, although the proposed model showed trend toward better performances for both AUC and F1-score.

The term *"COVID-19"* is not included in the training of vision-language models, and therefore, direct detection using a simple prompt with *"COVID-19"* and *"no COVID-19"* is expected to result in significant performance degradation. Interestingly, while the current state-of-the-art model shows a marked deterioration, our proposed model does not experience a decrease in performance, presenting significantly better performance in terms of both AUC and F1-score. The success of our model can be attributed to its design, which bridges domain-specific pre-training models. The text encoder of our model employs CXR-BERT[17], a self-supervised model trained not only on MIMIC-CXR and MIMIC-III data, but also on PubMed abstracts, which contain COVID-19 information from recent publications. As a result, the in-domain knowledge acquired during uni-modal self-supervised learning on a text corpus was extended to the multi-modal domain, facilitating the effective detection of previously unseen diseases. These results suggest the potential to expand the knowledge of a uni-modal model pre-trained in a specific domain to other domains through multi-modal combination, highlighting the scalability of currently available large uni-modal pre-trained models.

Verification of image-text correlation via qualitative analysis of cross-attention In contrast to CLIP-based models that visualize task-agnostic self-attention or only the class-level attention, our cross-attention-based model offers the additional advantage of visualizing word-patch level cross-attention between images and sentences. This provides a more transparent interpretation of the model's behavior by considering the meaning of each word component. To achieve this, we performed qualitative analysis using Grad-CAM[29] visualization for the fusion module's cross-attention, as suggested in ALBEF[3], as shown in the illustrated cases of Fig. 4. Without any supervision for the region-word correlations, the medical X-VL model correctly focuses on the regions related to each word, demonstrating its ability to understand the relationship between the semantics of the image and the textual concept. Notably, the model not only identifies important clinical findings such as "Congestion" and "Pacemaker," but also understands the location ("Mediastinal," "Cardiac") and relationships ("Left ... Lower," "Left ... Upper").

Application to clinical data in different domain In order to further demonstrate the broad applicability of our model, we extended our research to the field of abdominal radiography. We obtained a total of 5,772 image-text pairs of abdominal radiographs from Chung Ang University Hospital (CAUH) for training, and an additional 734 abdominal radiographs from the Chungnam National University Hospital (CNUH) registry for external validation. The model was initialized with pre-trained weights from chest radiographs, as there are similarities between chest and abdominal radiographs, despite differences in the field of view. The main abnormal findings in abdominal radiographs, ileus and pneumoperitoneum, were evaluated. Unlike chest radiographs, the medical reports for abdominal radiographs typically do not include descriptions of the location or extent of the disease, but rather simplified reports. Therefore, the zero-shot error detection performance was evaluated only for three types of

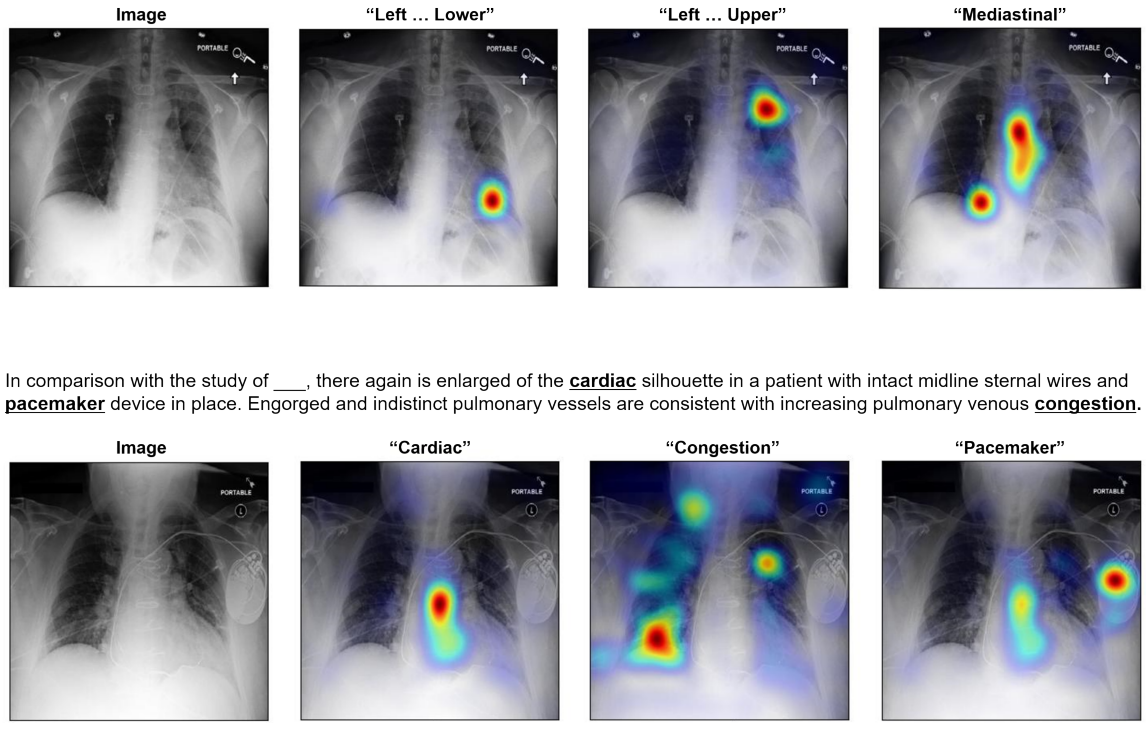


Figure 4: **(A-B) Exemplified cases of the Grad-CAM visualization of the cross-attention maps corresponding to each word. The medical X-VL model not only grounds the important clinical findings (“Congestion", “Pacemaker") but also understands the location (“Mediastinal", “Cardiac") and the relationships (“Left ... Lower", “Left ... Upper").**

errors: mismatch, false-negative, and false-positive.

In evaluating the trained model for zero-shot detection of major abnormal findings in abdominal radiograph, the proposed model showed trend toward better performance than the current SOTA model both with the simple prompt and the detailed prompt for all abnormality classes, although not statistically significant, in terms of both AUC and F1-score (Table 5). The trend toward better performance of the proposed model was more pronounced in zero-shot detection of the error showing the larger differences in performance in terms of AUC for all error classes (Table 6), demonstrating once again the superiority of our proposed method for zero-shot oversight AI in different domain.

Discussion

Despite the remarkable achievements of deep learning-based AI models in various tasks, their successes have been limited to narrow domains, suggesting the presence of fundamental challenges yet to be addressed. While contemporary AI-based CAD models exhibit outstanding performance in detecting abnormalities within an image, they are primarily designed as standalone diagnostic tools that often disrupt common radiological workflows. Moreover, they lack the complementary ability to assist human readers in identifying errors in image descriptions due to their inability to interpret both image and text jointly.

Unlike the model trained with traditional supervised learning, where image-label pairs are manually anno-

Table 5: Performance of zero-shot oversight AI to alert clinically significant abnormality for abdominal radiograph

	Average	Ileus	Pneumoperitoneum
AUC			
X-VL (simple)	0.837 (0.765-0.898)	0.851 (0.810-0.887)	0.824 (0.720-0.910)
X-VL (detailed)	0.827 (0.755-0.890)	0.837 (0.796-0.875)	0.817 (0.714-0.905)
CheXzero (simple)	0.767 (0.696-0.833)	0.770 (0.716-0.823)	0.765 (0.676-0.843)
CheXzero (detailed)	0.815 (0.749-0.875)	0.827 (0.781-0.872)	0.803 (0.717-0.879)
F1			
X-VL (simple)	0.441 (0.321-0.556)	0.449 (0.376-0.520)	0.432 (0.267-0.592)
X-VL (detailed)	0.447 (0.316-0.571)	0.420 (0.346-0.493)	0.474 (0.286-0.650)
CheXzero (simple)	0.317 (0.221-0.428)	0.372 (0.300-0.457)	0.262 (0.143-0.400)
CheXzero (detailed)	0.366 (0.276-0.463)	0.454 (0.372-0.540)	0.279 (0.180-0.386)

Table 6: Zero-shot error detection performance for the critical radiology report errors in abdominal radiograph

	Average	Mismatch	False-negative	False-positive
AUC				
X-VL	0.816 (0.717-0.900)	0.837 (0.769-0.892)	0.785 (0.627-0.929)	0.826 (0.756-0.881)
CheXzero	0.616 (0.414-0.811)	0.695 (0.558-0.807)	0.426 (0.105-0.778)	0.728 (0.580-0.849)
F1				
X-VL	0.263 (0.151-0.408)	0.335 (0.209-0.451)	0.102 (0.008-0.303)	0.353 (0.236-0.469)
CheXzero	0.230 (0.129-0.331)	0.327 (0.188-0.443)	0.030 (0.003-0.083)	0.334 (0.196-0.467)

tated to train visual recognition, vision-language model uses uncurated image-text pairs. This approach has the advantage of allowing the model to learn rich semantics and broad coverage of visual concepts from free-form text, rather than being confined to discrete labels that offer dense but limited visual concepts[30]. By learning broad visual semantics and corresponding textual concepts together, the model can achieve good performance across a range of downstream tasks. However, a limitation of this approach is that the image-text pair lacks the powerful discriminative ability of dense labels used in traditional supervised learning. As a result, training a robust vision-language model often requires billion-scale image-text data, which is difficult to obtain in the medical domain, as demonstrated in the studies of CLIP[11] and ALIGN[2].

To overcome these issues, we utilized several key components in our study. Specifically, we adopted a multi-modal fusion encoder that bridges domain-specific pre-trained uni-modal models, leveraging domain-specific self-superivsed learning with single domain data and thereby allowing for stable learning even with fewer image-text pairs. We also employed momentum distillation, which was originally developed and used in a previous work[3], to effectively learn information that cannot be learned from one-hot labels and to alleviate the problems of noisy data pairs. This method was also found to be effective in medical domain data, which tends to have relatively small difference even between negative pairs, both visually and linguistically, than general domain data. There may exist weak positive relations with partially overlapping impressions even between negative pairs, and thus, a penalization method different to the one-hot approach is demanded. For this purpose, momentum distillation can be utilized, as the momentum teacher can provide different pseudolabels for pairs with different similarities, allowing to learn additional information between the labels. Additionally, we designed a sentence-wise

contrastive learning method between individual sentences that comprise a medical report and images, taking into account that medical reports are composed of multiple sentences, unlike general domain text captions with single sentence. This method is an extension of the global-to-local contrastive learning method[25] and allows images to be more closely correlated with the semantic components of individual sentences.

As a result, this structure enabled the model to better understand fine-grained image-sentence relations in the medical domain, allowing it to successfully perform not only zero-shot abnormality detection, but also error detection. The performance difference between our model and the current SOTA models was more pronounced in error detection, where the model should also detect partial errors of the given sentence, as well as when using a detailed prompt rather than a simple prompt. This implies that the model has a better understanding of more detailed meaning of words that constituting a sentence, beside the simple class names to produce more accurate results. Moreover, our proposed method offers several additional benefits. In contrast to previous models that employ separate image and text encoders for direct alignment[15,17,18], our approach utilizes a fusion encoder with multi-modal cross attention to facilitate image-guided text prediction through the MLM objective in pre-training. This allows for zero-shot error correction of incorrect words. The fusion of uni-modal self-supervised models through a fusion encoder may further enhance the model's scalability by leveraging large amounts of single-domain data with self-supervised learning. Our experiments with the unseen COVID-19 disease class demonstrated that extending the concept of "COVID-19" from the uni-modal self-supervised text encoder to the multi-modal domain was achievable. Given that well-aligned data in the medical domain is relatively scarce while large-scale domain-specific data is more abundant, this property suggests the potential to improve model scalability with single-domain data, reducing reliance on well-aligned image-text pairs.

Our study is not free of limitations. Firstly, despite our efforts to reduce reliance on labeled image-text data through domain-specific self-supervised learning, our approach still exhibits a certain degree of dependency on image-text paired data. Secondly, our zero-shot correction method using single word masking is effective only for correcting single-word errors and may not be suitable for correcting errors involving sentences of varying lengths, such as mis-registrations. Lastly, although we have developed a cross-attention-based model that demonstrates improved fine-grained understanding of semantics compared to previous approaches, the detection performance for location and extent errors, which involve subtle changes in a few words, was lower compared to other types of errors.

Despite the limitations, our proposed medical X-VL model, specifically designed for the medical domain, has demonstrated remarkable performance in various oversight AI tasks using a zero-shot approach, surpassing the current SOTA models. As AI models are increasingly used to assist medical professionals rather than replace them, our model trained on uncurated data without explicit labeling, but with a flexible understanding of vision-language concepts, would be highly beneficial in various clinical applications. Furthermore, the image-report pair structure and the multiple-sentence report structure are commonly found in medical imaging modalities beyond radiographs, suggesting that our approach has broad applicability in the field of medical imaging.

Methods

Details of model architecture Our medical X-VL model is based on the commonly used cross-attention-based multi-modal vision-language models such as those introduced in recent studies[3,12,25,31]. It connects the pre-trained uni-modal models in the medical domain by utilizing the CXR-BERT model[17] as the text encoder, which is a transformer model consisting of 12 layers and 12 multi-heads, and the vision transformer pre-trained on MIMIC training data using DINO self-supervised learning[24] for the vision encoder.

As the fusion encoder, we opted for the BERT$base$ model[32]. The image encoder transforms the input image I into a sequence of patch embeddings $p_{cls}, p_1, ..., p_N$, where p_{cls} denotes the `[CLS]` token embedding. Similarly, the input text T is converted into a sequence of word embeddings $w_{cls}, w_1, ..., w_M$, where w_{cls} rep-

resents the [CLS] token indicating the start of the sequence, and w_M signifies the end of the sentence as the [SEP] token. By utilizing cross-attention between the modalities, the word embeddings and patch embeddings are fused in the multi-modal fusion encoder, producing the fused word embeddings $v_{cls}, v_1, ..., v_M$.

Details of pre-training objectives The medical X-VL model is trained using three distinct learning objectives, namely contrastive learning for achieving cross- and intra-modal alignment, masked language modeling (MLM) to facilitate image-guided text completion, and image-text matching (ITM).

Contrastive Learning for Cross- and Intra-modal Alignment The objective of cross-modal contrastive learning is to align image and text features in a shared embedding space prior to fusion, through the application of uni-modal encoders. It operates by attracting positive image-text pairs towards each other, while pushing unmatched pairs apart. In the context of the encoded embeddings p_{cls} and w_{cls} of the [CLS] tokens of image I and text T, respectively, the similarity functions $sim(I,T)$ and $sim(T,I)$ can be defined as follows:

$$sim(I,T) = h_I(p_{cls})^\top h_T(w_{cls}), \quad sim(T,I) = h_t(w_{cls})^\top h_I(p_{cls}) \tag{1}$$

where h_I and h_T denote linear projectors with normalization layers for image and text features. Then, the normalized image-to-text and text-to-image similarities of each image-text pair are calculated as:

$$s_{i2t} = \frac{\exp\left(sim(I,T_m)/\tau\right)}{\Sigma_{m=1}^{M} \exp\left(sim(I,T_m)/\tau\right)}, \quad s_{t2i} = \frac{\exp\left(sim(T,I_n)/\tau\right)}{\Sigma_{n=1}^{N} \exp\left(sim(T,I_n)/\tau\right)} \tag{2}$$

where τ denotes the temperature parameter.

In contrast, intra-modal contrastive learning focuses on training the model to distinguish positive and negative samples within each modality based on their semantic differences. The normalized similarities between images and between texts can be defined similarly to those in cross-modal contrastive learning.

$$s_{i2i} = \frac{\exp\left(sim(I,I_n)/\tau\right)}{\Sigma_{n=1}^{N} \exp\left(sim(I,I_n)/\tau\right)}, \quad s_{t2t} = \frac{\exp\left(sim(T,T_m)/\tau\right)}{\Sigma_{m=1}^{M} \exp\left(sim(T,T_m)/\tau\right)} \tag{3}$$

where τ is the same temperature parameter used in the cross-modal contrastive learning.

Consequently, given the one-hot label similarity y, the cross-modal contrastive loss L_{CMC}, intra-modal contrastive loss L_{IMC} and overall contrastive loss $L_{contrastive}$ can be defined as the cross-entropy loss H:

$$L_{CMC} = \frac{1}{2}[H(y_{i2t}, s_{i2t}) + H(y_{t2i}, s_{t2i})] \tag{4}$$

$$L_{IMC} = \frac{1}{2}[H(y_{i2t}, s_{i2t}) + H(y_{t2i}, s_{t2i})] \tag{5}$$

$$L_{contrastive} = L_{CMC} + L_{IMC} \tag{6}$$

In our study, we employed the recent contrastive learning techniques[33] and utilized image and text queues to store the most recent Q samples from the momentum encoder for each modality. The size of the queue was set to 40,920 in our experiments. By calculating feature similarities between image and text with the aforementioned objectives, we applied hard negative mining for ITM, where negative pairs with high similarity were sampled more frequently.

Masked Language Modeling for Image-guided Text Completion The Masked Language Modeling (MLM) learning objective is commonly used to obtain language comprehension in which the model predicts the correct ground truth for masked word tokens w_{mask} while using the corresponding image as a guide. The masking process randomly replaces 15% of the word tokens with the `[MASK]` token, with 80% probability, a random word token with 10% probability, and the original token with 10% probability[32]. The masked text is represented as T^{mask}, the fusion encoder's prediction for `[MASK]` tokens is represented as $p^{mask}(I, T^{mask})$, and the ground truth for each word token is represented as y_w^{mask}. The MLM loss is defined using the cross-entropy loss H in our model.

$$L_{MLM} = H(y_w^{mask}, p^{mask}(I, T^{mask})) \tag{7}$$

The MLM task in our model, where the predictions for the `[MASK]` token are generated by the multi-modal fusion encoder that incorporates image representation into text tokens, can be seen as an image-assisted prediction of masked text tokens. This allows the model to learn the joint representation of images and text and their interdependence.

Image-Text Matching The image-text matching task aims to determine whether a given image-text pair is matched or unmatched. We used the fusion embeddings of two `[CLS]` tokens, which are obtained from the outputs of the image-to-text and text-to-image paths of the fusion encoder. These embeddings reflect the joint representation of the image-text pair. Binary classifiers were added after the fusion embeddings to predict whether an image-text pair is matched or unmatched, resulting in the prediction $c^{itm}(I, T)$. The ITM loss was defined using the cross-entropy loss H, where y^{itm} represents the ground truth label for image-text matching, as shown below:

$$L_{ITM} = H(y^{itm}, c^{itm}(I, T)) \tag{8}$$

We also implemented hard negative mining in the ITM task by prioritizing samples with high similarity scores s_{i2t} and s_{t2i} when selecting negative pairs from the batch for a given image or text[3]. This approach enables the model to better distinguish between semantically similar but fine-grained different images, which is particularly important in medical imaging, where images tend to have small differences due to standardized acquisition protocols. By using this strategy, we achieved performance improvements without any additional computational cost.

Combined, the overall pre-training objective L of the medical X-VL is:

$$L = L_{contrastive} + L_{MLM} + L_{ITM} \tag{9}$$

However, medical images can belong to multiple classes, and different images may represent the same

clinical findings. For instance, a pair of image-report with "No specific finding" and another pair with "No acute cardiopulmonary process" are different but describe the same clinical findings. Therefore, if we use the sample of "No acute cardiopulmonary process" as a negative sample for "No specific finding," which actually corresponds to a positive sample, an incorrect supervisory signal would be given, resulting in label noise that can negatively affect the learning process. Hence, we obtained the text feature of a sample obtained by hard negative mining and computed the cosine similarity between this text feature and that of the positive sample. If the cosine similarity exceeded the pre-defined threshold, the negative sample was resampled to ensure the cosine similarity was below this threshold.

Momentum Distillation In contrastive learning, negative samples may have similar context to positive ones and should be treated differently from entirely different negative samples. For example, when a radiograph shows bibasilar opacifications suggesting pneumonia, a report like "There exist lung opacities in both lower lobe suggesting the severe pleural effusion," which describes a different etiology for opacification, may be regarded as a negative sample. However, it should be penalized differently from an entirely unmatching description like "Both lung fields are clear and there is no remarkable finding." This issue is more critical in the medical domain, where the differences between images and reports are smaller than those between photographic images and captions, and the overlapping between the images or reports can be substantial. Similarly, for MLM, there may be other candidates with semantically identical meaning to the ground truth, like "No remarkable findings" and "No abnormality." However, the binary and one-hot coded labels for contrastive learning and MLM penalize all negative samples without considering their correctness.

To address this problem, we utilized momentum teachers that generate pseudo labels for contrastive learning and MLM. These teachers are gradually updated with exponential moving averaging of the current models. During training, the model aims to match the pseudo labels generated from the momentum teacher by minimizing the distillation loss L_{dist} in addition to the overall loss L, with a weight parameter λ to balance the contributions of momentum distillation. This approach enables the model to better handle negative samples with similar context and achieve improved performance.

$$L_{total} = (1 - \lambda) \cdot L + \lambda \cdot L_{dist} \tag{10}$$

Details of dataset for vision-language model training To traing the vision-language model using chest radiographs, we employed the MIMIC-CXR dataset[34], which is an open-source database containing 377,110 pairs of images and their corresponding free-text radiology reports from 227,835 radiographic studies. The full radiology report is made up of various sections, including examination, indication, impression, technique, and comparison. As the impression section contains important summaries of the findings for the image, we used it for the training of vision-language model. We excluded 5,159 images without impressions among the images, used 371,951 images for training based on the official split, and used 3,651 images for zero-shot error detection evaluation on the test set.

To apply our method to other domains with limited data, we utilized the inpatient abdominal radiograph databases of two hospitals for training and evaluation. Specifically, we employed 5,772 images from CAUH for training and collected 734 images from the database of CNUH for external validation.

Details of the simulation for human error in radiology Radiology reports generated by clinicians can be subject to errors in various ways. To categorize clinically significant errors that occur, our study identified five types: mis-registration, incorrect location, incorrect extent, false-positive estimate, and false-negative estimate. To implement these errors, we designed an error generator to automatically produce them with a given probability.

When reports in the MIMIC-CXR dataset are classified using the CheXpert labeler, they can be divided into "no finding" and positive and negative estimates for each "abnormality" class. Mis-registration errors occur when a report from a different label class is erroneously registered (Supplementary Fig. S3a). This can occur when a previous patient's report is attached to the next patient, leading to potential fatal scenarios in clinical practice. To simulate this, we probabilistically replace the original report with a report from another label class. Incorrect location errors refer to errors where the description of a lesion's location (left-right, upper-lower, apical-basal, central-peripheral) is incorrect (Supplementary Fig. S3b). This can lead to confusion in making clinical diagnoses and treatment decisions. We implement this error type by replacing location-descriptive terms with their opposites (e.g., $right \rightarrow left$). Incorrect extent (severity) errors occur when the extent or severity of the findings (mild-severe, small-large, or minimal-extensive) is inaccurately described (Supplementary Fig. S3c). This can result in erroneous treatment plans due to the inaccurate information about disease severity. This error type is implemented by substituting words describing extent or severity with their antonyms (e.g., $mild \rightarrow severe$). False-positive estimate errors occur when there is actually no abnormal finding, but it is incorrectly interpreted as having some abnormality (Supplementary Fig. S3d). This can result in serious problems such as unnecessary treatment. We implement this error type by first determining if the report corresponds to "no finding," and then replacing it with a report corresponding to a certain abnormality. False-negative estimate errors occur when there is an abnormality present but not described in the report and is regarded as no finding (Supplementary Fig. S3e). This can lead to clinically significant problems by missing the opportunity to provide necessary treatment in time. This error type is implemented by detecting if the report describes a positive class for abnormalities and replacing it with a report corresponding to "no finding."

Implementation details We pre-processed the image data by removing margin space following the approach proposed in a previous work[20]. The pre-processing also involved Gaussian blurring, normalization, and resizing the images to 224 $\times$ 224. We used the pre-trained tokenizer of CXR-BERT with a vocabulary size of 30,522[17], and set the maximum word length of the report to 120. For the vision-language pre-training, we employed an AdamW optimizer with an initial and maximum learning rate of 0.00001 and 0.0001, respectively, for five epochs, including two epochs of warm-up period, and the batch size was 10. All experiments were conducted using Python version 3.8 and PyTorch library version 1.10 on two NVIDIA GeForce RTX 3090.

Details of evaluation. To evaluate the model performances for the zero-shot oversight tasks, we assessed the model in two aspects: zero-shot abnormality detection and zero-shot error detection. As zero-shot abnormality detection can be considered as a zero-shot classification problem, we assessed the zero-shot classification performance of the trained model using a subset of the CheXpert test dataset (500 samples). Following previous works[18], we calculated the area under the receiver operating curve (AUC) and F1-score for five major abnormalities (atelectasis, cardiomegaly, consoldiation, pulmonary edema, and pleural effusion) and compared the model performance with the current SOTA and other self-supervised and vision-language models. To calculate the probability of each abnormality class, we performed softmax evaluation for both the simple prompt method (abnormality and no abnormality) following previous work and the detailed prompt method for each abnormality. The descriptions for each abnormality used in the detailed prompt were five sentences for each class confirmed by board-certified radiologists in the previous paper[15], and new detailed prompts consisting of five sentences each were created by clinicians for the consolidation class for which no corresponding sentences were available. For example, the detailed prompt for "Pleural effusion" includes the following sentences: "A pleural effusion is present.", "Blunting of the costophrenic angles represents pleural effusions.", "Trace pleural fluid is present.", "The pleural space is partially filled with fluid.", and "Layering pleural effusions are present." By performing zero-shot detection using these detailed prompts, we aimed to evaluate how well the model understands detailed expressions beyond simple class names.

We utilized the COVIDx[35] dataset to evaluate the detection performance of the model on unseen diseases during the learning process of vision-language model. We randomly split 2,998 images from a total of 29,986 images and used them as a test set for evaluation. The images from the COVIDx dataset that were not included in the test set were not utilized during the training process.

We evaluated zero-shot error detection performance on the test set of MIMIC-CXR data that was not used for training, by calculating the AUC and F1-score for detection performance using simulated human errors of five types with a probability of 5%, as mentioned above.

Details of statistical analysis For statistical analysis, we used non-parametric bootstrapping. Random samples of the same size as the original dataset were repeatedly sampled with replacement 1,000 times. We estimated the difference in AUC and F1 metrics using the bootstrap samples. Confidence intervals were derived from the relative frequency distribution of the estimates over the re-samples, using the interval between the $100 \times (\alpha/2)$ and $100 \times (1 - \alpha/2)$ percentiles. We set α to 0.05 and considered differences beyond the confidence interval as statistically significant difference.

Ethic committee approval. The abdominal radiograph data collected for this study were ethically approved by the Institutional Review Boards of Chung-Ang University Hospital, Chungnam University Hospital, and the requirement for informed consent was waived.

Correspondence Correspondence and requests for materials should be addressed to Jong Chul Ye. (email: jong.ye@kaist.ac.kr).

Acknowledgements This research was supported by the KAIST Key Research Institute (Interdisciplinary Research Group) Project, the National Research Foundation of Korea under Grant NRF-2020R1A2B5B03001980, and Chungnam National University Hospital Research Fund, 2022.

Author Contributions S.P. performed all experiments, wrote the extended code, and prepared the manuscript. E.S.L and J.E.L collected data and provided clinical evaluation. J.C.Y. supervised the project in conception and discussion, and prepared the manuscript.

Competing Interests The authors declare that they have no competing financial interests.

Data Availability Part of the data is collected from open-sourced data repositories that are publicly available. The MIMIC-CXR database is available at https://physionet.org/content/mimic-cxr/2.0.0/. Subset of the CheXpert test data and corresponidng labels used for the evaluation of the model in zero-shot abnormality detection can be found at https://github.com/rajpurkarlab/cheXpert-test-set-labels. COVIDx dataset used for the evauation of the model in unseen disease is available tat https://www.kaggle.com/datasets/andyczhao/covidx-cxr2. Other parts of data used for the experiments for abdominal radiographs are not publicly available due to the patient privacy obligation. Interested users can request access to these data for research purposes, by contacting the corresponding author J.C.Y (jong.ye@kaist.ac.kr). The data can be shared after the IRB approval and de-identification along with the signed agreement on data transfer and usage. Replies to the initial request will be made within 10 working days. The use of data is limited only to the research purpose, and the redistribution is prohibited.

Code Availability Th code is available at following GitHub repository. $https://github.com/sangjoon-park/Medical_X-VL$

References

1. Boden, M. A. *Mind as machine: A history of cognitive science* (Oxford University Press, 2008).

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

3. Li, J. *et al.* Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* **34**, 9694–9705 (2021).

4. Cho, J., Lei, J., Tan, H. & Bansal, M. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, 1931–1942 (PMLR, 2021).

5. Chen, Y.-C. *et al.* Uniter: Universal image-text representation learning. In *European conference on computer vision*, 104–120 (Springer, 2020).

6. Lu, J., Batra, D., Parikh, D. & Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* **32** (2019).

7. Li, X. *et al.* Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, 121–137 (Springer, 2020).

8. Huang, Z., Zeng, Z., Liu, B., Fu, D. & Fu, J. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849* (2020).

9. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

10. Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017).

11. Radford, A. *et al.* Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763 (PMLR, 2021).

12. Yu, J. *et al.* Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917* (2022).

13. Wang, W., Bao, H., Dong, L. & Wei, F. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358* (2021).

14. Alayrac, J.-B. *et al.* Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198* (2022).

15. Zhang, Y., Jiang, H., Miura, Y., Manning, C. D. & Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, 2–25 (PMLR, 2022).

16. Huang, S.-C., Shen, L., Lungren, M. P. & Yeung, S. 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*, 3942–3951 (2021).

17. Boecking, B. *et al.* Making the most of text semantics to improve biomedical vision–language processing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, 1–21 (Springer, 2022).

18. Tiu, E. *et al.* Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering* 1–8 (2022).

19. Naseem, U., Khushi, M. & Kim, J. Vision-language transformer for interpretable pathology visual question answering. *IEEE Journal of Biomedical and Health Informatics* (2022).

20. Moon, J. H., Lee, H., Shin, W. & Choi, E. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *arXiv preprint arXiv:2105.11333* (2021).

21. Yan, B. & Pei, M. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation (2022).

22. Wu, C., Zhang, X., Zhang, Y., Wang, Y. & Xie, W. Medklip: Medical knowledge enhanced language-image pre-training. *medRxiv* 2023–01 (2023).

23. Xiang, T. *et al.* In-painting radiography images for unsupervised anomaly detection. *arXiv preprint arXiv:2111.13495* (2021).

24. Caron, M. *et al.* Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9650–9660 (2021).

25. Yang, J. *et al.* Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15671–15680 (2022).

26. Irvin, J. *et al.* Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 590–597 (2019).

27. Min, D., Kim, K., Lee, J. H., Kim, Y. & Park, C. M. Rred: A radiology report error detector based on deep learning framework. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, 41–52 (2022).

28. Yu, F. *et al.* Evaluating progress in automatic chest x-ray radiology report generation. *medRxiv* 2022–08 (2022).

29. Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626 (2017).

30. Yang, J. *et al.* Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19163–19173 (2022).

31. Dou, Z.-Y. *et al.* Coarse-to-fine vision-language pre-training with fusion in the backbone. *arXiv preprint arXiv:2206.07643* (2022).

32. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

33. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738 (2020).

34. Johnson, A. E. *et al.* Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data* **6**, 1–8 (2019).

35. Andy, Z. *et al.* Covidx cxr-2 dataset (2023). Https://www.kaggle.com/datasets/andyczhao/covidx-cxr2, Last accessed on 2023-03-28.

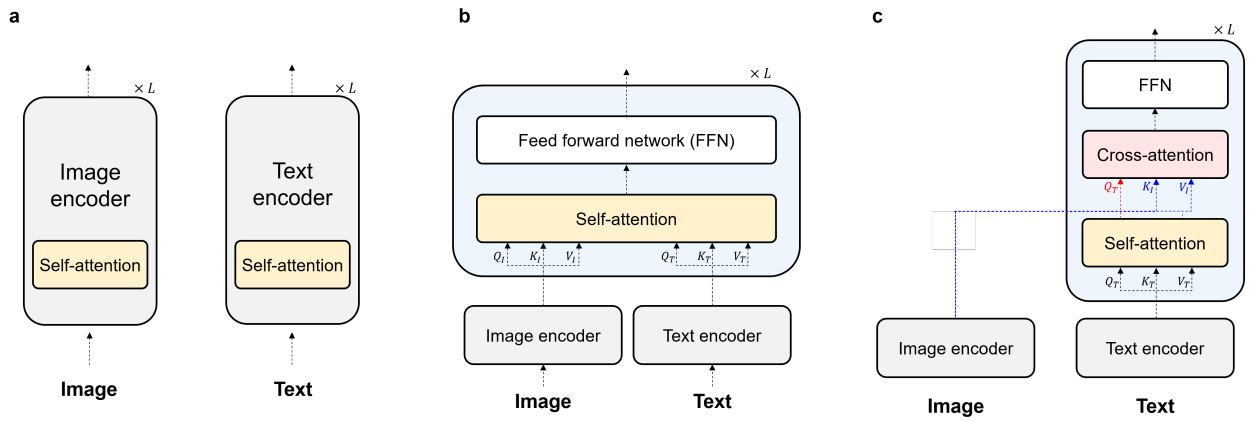
Supplementary Information

Definition of Momentum Distillation The learning paradigm of knowledge distillation involves transferring knowledge from a teacher model to a student model, which was previously explained in the context of self-training. Initially developed for compressing models, knowledge distillation is used to efficiently build a simpler student model by distilling the knowledge from a more complex teacher model. This approach is particularly useful for practical implementation of AI models in devices with limited computational resources. Moreover, knowledge distillation can be employed in a siamese design, where one model learns from the other model's predictions instead of relying on labels. Thus, several works on semi- and self-supervised learning have utilized knowledge distillation in self-training. Recent studies suggest that semi- or self-supervised learning methods based on the knowledge distillation framework can result in a model with performance similar to, or even better than, fully supervised models. Momentum distillation refers to a type of knowledge distillation where the teacher model, which has the same architecture as the student model, is updated gradually through exponential moving averages from the student model. This approach enables momentum-based updates to be transferred from the teacher to the student model during knowledge distillation.

Supplementary Table 1: Ablation study of the key component of medical X-VL model

	Average
AUC	
Proposed (X-VL)	0.881 (0.843-0.912)
No MLM	0.872 (0.834-0.906)
No momentum distillation	0.873 (0.835-0.906)
No sentencewise contrastive learning	0.868 (0.827-0.904)
No similarity constraint for negative sampling	0.869 (0.829-0.903)

Results are obtained using detailed prompt for five abnormality classes.



Supplementary Fig. S1: Illustration of the different vision-language model architectures. (A) Parallel dual encoder model. Multi-modal fusion encoder model with (B) self-attention and (B) cross-attention.

a

Simple prompt		**Detailed prompt**
Cardiomegaly.	*Mean*	The heart is mildly enlarged. Cardiomegaly is present. The heart shadow is enlarged. The cardiac silhouette is enlarged. Cardiac enlargement is seen.

b

Simple prompt		**Detailed prompt**
Pleural effusion.	*Mean*	A pleural effusion is present. Blunting of the costophrenic angles represents pleural effusions. Trace pleural fluid is present. The pleural space is partially filled with fluid. Layering pleural effusions are present.

Supplementary Fig. S2: Examples of the simple prompt (left) and the detailed prompt (right) for abnormality class (A) cardiomegaly and (B) pleural effusion. For the detailed prompt, evaluation is performed by calculating the mean of the logits of each description.

a **Mismatch**

Original Probable atelectasis at the right lung base. No definite consolidation.

Error Bibasilar opacities consistent with clinical diagnosis of pneumonia.

b **Location**

Original Left sided pleural effusion with underlying airspace disease likely representing atelectasis versus left lower lobe/lingular pneumonia.

Error Right sided pleural effusion with underlying airspace disease likely representing atelectasis versus right lower lobe/lingular pneumonia.

c **Extent**

Original Bibasilar opacities, most likely atelectasis, less likely pneumonia. Mild cardiomegaly.

Error Bibasilar opacities, most likely atelectasis, less likely pneumonia. Severe cardiomegaly.

d **False-positive**

Original No acute cardiopulmonary pathology.

Error Mild cardiomegaly with mild pulmonary vascular congestion. No evidence of pneumonia.

e **False-negative**

Original Mild pulmonary vascular engorgement and mild interstitial edema.

Error No acute cardiopulmonary abnormality.

Supplementary Fig. S3: Examples of (A) the mismatch, (B) the location, (C) the extent, (D) the false-positive, and (E) the false-negative errors.