
AGA: AN ADAPTIVE GROUP ALIGNMENT FRAMEWORK FOR STRUCTURED MEDICAL CROSS-MODAL REPRESENTATION LEARNING

A PREPRINT

Li Wei

School of Computing and Artificial Intelligence
Southwest Jiaotong University
Chengdu, China 611756
liweii0521@163.com

Gong Xun*

School of Computing and Artificial Intelligence
Southwest Jiaotong University
Chengdu, China 611756
xgong@home.swjtu.edu.cn

Li Jiao

Department of Gastroenterology
The Third People's Hospital of Chengdu
Chengdu, China 610031
cyljiao@163.com

Sun Xiaobin

Department of Gastroenterology
The Third People's Hospital of Chengdu
Chengdu, China 610031
xbsun1197@163.com

August 1, 2025

ABSTRACT

Learning medical visual representations directly from paired medical images and reports has emerged as a promising direction in representation learning. However, existing vision-language pretraining (VLP) methods in the medical domain often oversimplify clinical reports into single entities or fragmented tokens, overlooking their inherent structured nature. Moreover, contrastive learning paradigms typically rely on large quantities of hard negative samples, which poses challenges when dealing with small scale medical datasets. To address these issues, we propose Adaptive Grouped Alignment (AGA), a novel framework for learning structured information from paired medical images and reports. Specifically, we design a bidirectional grouping mechanism based on a sparse similarity matrix. Given an image-report pair, we first compute a fine-grained similarity matrix between each text token and each image patch. For each token, we select the top-matching patches to form a visual group, and conversely, for each patch, we select the most semantically related tokens to form a language group. To enable adaptive grouping, we introduce two threshold gating modules, Language-grouped Threshold Gate and Vision-grouped Threshold Gate, which dynamically learn similarity thresholds for group construction. The group representation corresponding to each token or patch is computed as a weighted average over the elements in its group, where the weights are given by their similarity scores. To align each token representation with its corresponding group representations, we propose an Instance-aware Group Alignment (IGA) loss, which operates solely within individual image-text pairs, eliminating the need for external negative samples and thereby alleviating the reliance on large scale hard negatives. Finally, we employ a Bidirectional Cross-modal Grouped Alignment (BCGA) module to facilitate fine-grained alignment between visual and linguistic group representations. Extensive experiments on both public and private datasets across various downstream tasks, including image-text retrieval and classification (in both fine-tuning and zero-shot settings), demonstrate the effectiveness of our proposed framework.

Keywords Representation learning · Vision-language pretraining · Contrastive learning · Grouped alignment

*corresponding author

1 Introduction

Advances in medical imaging technologies have revolutionized healthcare practices and significantly improved patient outcomes. However, the rapidly increasing volume of imaging studies in recent years has posed substantial challenges, including the fact that annotating medical imaging datasets requires domain expertise, imposing a growing burden on radiologists and incurring prohibitive costs at scale. Consequently, the development of effective medical image models is hindered by the scarcity of large scale manually annotated datasets. To overcome this limitation, vision-language pretraining (VLP) methods that learn visual representations of medical images directly from radiology reports without any additional manual annotation have become mainstream Chauhan et al. [2020], Zhang et al. [2017, 2022a]. These methods aim to learn general medical visual representations from detailed clinical narratives authored by physicians, which can then be transferred to downstream tasks. Recent works leverage such medical reports as supervisory signals and optimize the multimodal representations by maximizing the mutual information between global representations of paired images and reports Zhang et al. [2022a], Bannur et al. [2023]. Nonetheless, considering that pathological findings often occupy only small regions within an entire image, several studies Huang et al. [2021], Wang et al. [2022] have explored learning local representations for medical language-image tasks. Furthermore, methods such as those in Gao et al. [2025], Cheng et al. [2023], Zhang et al. [2023a] improve semantic-driven contrastive learning by segmenting medical reports into sentences instead of individual tokens, facilitating fine-grained cross-modal interactions. These advances mark significant progress in medical VLP. However, these approaches commonly oversimplify medical reports into single entities or fragmented tokens and neglect their inherent structured nature, as illustrated in Fig. 1(a). Sentence-level representations typically compress entire sentences into global vectors, which tend to conflate multiple independent clinical entities, anatomical locations, and attribute information. This leads to semantic ambiguity and impedes precise alignment with corresponding regions in medical images.

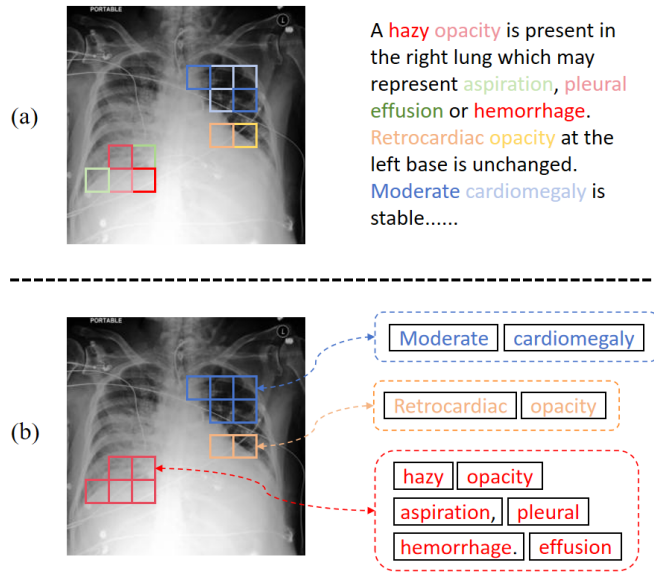


Figure 1: Fine-grained alignment. (a) illustrates the conventional word-to-patch alignment approach. (b) shows our proposed group-wise alignment strategy, where colors denote the corresponding alignment relationships.

Furthermore, due to the inherent difficulty in acquiring medical data, sample sizes are often limited. Traditional contrastive learning heavily relies on a large number of hard negative samples Kalantidis et al. [2020], Hjelm et al. [2018], Zhang et al. [2025, 2022b]. Although prototype-based contrastive methods He et al. [2025], Zhang et al. [2024] reduce the need for negative mining by aligning samples with class prototypes, these approaches typically depend on static class prototypes constructed across samples. Such static prototypes struggle to accommodate the presence of multiple heterogeneous semantic units within a single medical image-text pair (e.g., multiple lesions, anatomical sites, or attribute combinations). This issue is particularly pronounced under weakly supervised or unlabeled conditions, where prototypes fail to accurately capture the diversity and fine-grained semantics inherent within individual samples.

To address the aforementioned challenges, we propose a novel VLP framework with Adaptive Grouped Alignment, termed AGA. Unlike existing methods that simplify images and reports into single entities or fragmented tokens, our framework aligns local text token embeddings with their corresponding Token-Grouped Visual (TGV) embeddings using an Instance-aware Group Alignment (IGA) loss, as shown in Fig. 1(b). This is achieved by learning structured

information among patches through a grouping mechanism. For example, a single word in a report may correspond to multiple localized visual regions. Conversely, we generate Patch-Grouped Language (PGL) embeddings by aligning each visual patch with a group of semantically related textual tokens. The proposed group representations are constructed from semantically coherent subsets of tokens or patches, enabling more natural correspondence with localized visual patterns in medical images. By associating each text token (or image patch) with its corresponding semantic group representation, the model is able to better capture hierarchical and compositional structural information, thereby facilitating more precise cross-modal alignment. To realize this, we first compute a pairwise similarity matrix for each image-text pair. To form a visual group for each text token, we introduce a Language-grouped Threshold Gate, which progressively sparsifies the similarity matrix during training by adaptively lowering the grouping threshold via a momentum-based mechanism. The resulting similarity scores are row-normalized to assign a weight to each image patch embedding, and the weighted sum of these embeddings constitutes the corresponding TGV representation, enabling dynamic grouping. A similar procedure is applied to generate the PGL embedding for each image patch by aggregating its most relevant textual tokens.

To align local token embeddings with their corresponding group representations while mitigating the reliance on a large number of hard negatives, we introduce an IGA loss. Unlike conventional prototype-based contrastive learning approaches that operate at the class level, IGA constructs multiple semantically coherent group representations within each instance, and guides each image patch or text token to align with its associated semantic group. This method not only preserves the semantic aggregation advantages of prototype-based learning, but also provides enhanced instance-level modeling capacity and structural expressiveness.

In summary, the main contributions of this work are as follows:

1. We propose a novel group alignment framework, termed AGA, which constructs group representations for both textual tokens and visual patches by computing a sparse similarity matrix for each image-report pair. By learning from these group-level representations, the model captures structured information and enhances cross-modal feature expressiveness.
2. To enable adaptive grouping, we introduce learnable Threshold Gates that dynamically select grouping thresholds. Additionally, we design an Instance-aware Group Alignment (IGA) loss that aligns token embeddings with their corresponding group representations. This alignment is performed within each individual image-report pair, eliminating the need for hard negatives and allowing for fine-grained representation learning in a more efficient manner.
3. We employ a Bidirectional Cross-modal Grouped Alignment (BCGA) module to align group representations across modalities, and conduct extensive experiments on both public and private medical datasets. The results on downstream tasks such as image-text retrieval and classification, under both finetuning and zero-shot settings, demonstrate the effectiveness and generalizability of the proposed framework.

The remainder of this paper is organized as follows. Section 2 briefly reviews related work. Section 3 presents the overall architecture of our model, including the training process and alignment process. Section 4 reports the experimental datasets and results. In Section 5, we provide a detailed discussion. Finally, Section 6 concludes the paper.

2 Related work

2.1 Medical vision-language pretraining

Vision-language pretraining (VLP) on large scale medical image-text datasets has emerged as a widely adopted paradigm for learning generic visual representations, supporting various downstream tasks and serving as a foundation for visual encoders in multimodal foundation models Zhang et al. [2017], Wang et al. [2025a], Zhang et al. [2022a], Bica et al. [2024]. By aligning global image and text representations in a shared latent space using matched and mismatched image-text pairs, these models have demonstrated strong performance in image-level vision tasks such as classification, coarse-grained retrieval, and visual question answering Bannur et al. [2023], Pan et al. [2022], Li et al. [2025], Fan et al. [2025]. However, they often suffer from the drawback of discarding fine-grained information. To address this, Huang et al. [2021] made significant contributions by employing attention-based mechanisms to contrast image regions with words in paired reports, enabling the learning of localized visual representations that better capture the fine-grained patterns present in medical images. Building on this, several studies Gao et al. [2025], Cheng et al. [2023], Zhang et al. [2023a,b] have advanced semantics-guided contrastive methods by segmenting medical reports into sentences rather than individual words, thus facilitating localized cross-modal interactions. Other works have focused on multi-scale contrastive learning. For instance, Liao et al. [2021] optimized the estimation of mutual information between local image features and sentence-level text representations, improving fine-grained

alignment. Seibold et al. [2022] assumed that each sentence conveys distinct diagnostic information and proposed aligning images with corresponding sentences. Palepu and Beam [2023] further introduced entropy-based regularization on token representations to penalize image patch similarities, encouraging more informative alignments. Nevertheless, sentence-level embeddings compress the entire sentence into a single global vector, leading to information mixing. This makes it difficult to identify which specific words correspond to which image regions, thereby limiting the flexibility of structural modeling.

2.2 Contrastive learning

Contrastive learning [Chen et al. [2020], He et al. [2020], Feng et al. [2020]] aims to learn an embedding space in which positive instances are mapped close to each other while negative instances are pushed far apart. A key challenge in contrastive learning is the effective identification of positive and negative pairs. To improve the efficiency of contrastive learning, some studies have proposed predicting the features of one view from another view [Chen et al. [2020], Grill et al. [2020]]. Furthermore, works such as [Chaitanya et al. [2020], Han et al. [2021], Taleb et al. [2022], Feng et al. [2020]] have introduced the power of contrastive learning into the medical imaging domain, achieving substantial progress. Recently, several prototype-based contrastive learning methods have been proposed to leverage semantic information at the prototype level within datasets. For example, [Wang et al. [2025b, 2021], Guo et al. [2022]] contrast instance features with their paired prototype features, while [Li et al. [2021], Caron et al. [2020], Wang et al. [2022]] employ clustering-based methods to perform prototype-to-prototype contrast. However, prototype alignment relies on aggregated prototypes at the class level, which is suitable for tasks with known labels and limited categories, but struggles to model fine-grained semantic structures and instance-level variations within samples.

3 Method

The goal of this work is to jointly learn global and local multimodal representations of medical images by leveraging medical reports, aiming to support downstream tasks with limited manual annotations. Unlike other approaches, local representations are not constructed from fragmented words or image patches but rather from novel grouped representations. Here, we first describe the image and text encoders used to extract features from each modality in Section 3.1. In Section 3.2, we formalize the computation and sparsification of the similarity matrix and introduce the threshold gates used for grouping. Finally, in Section 3.3, we present the alignment process and the associated alignment loss. The overall framework is illustrated in Fig. 2.

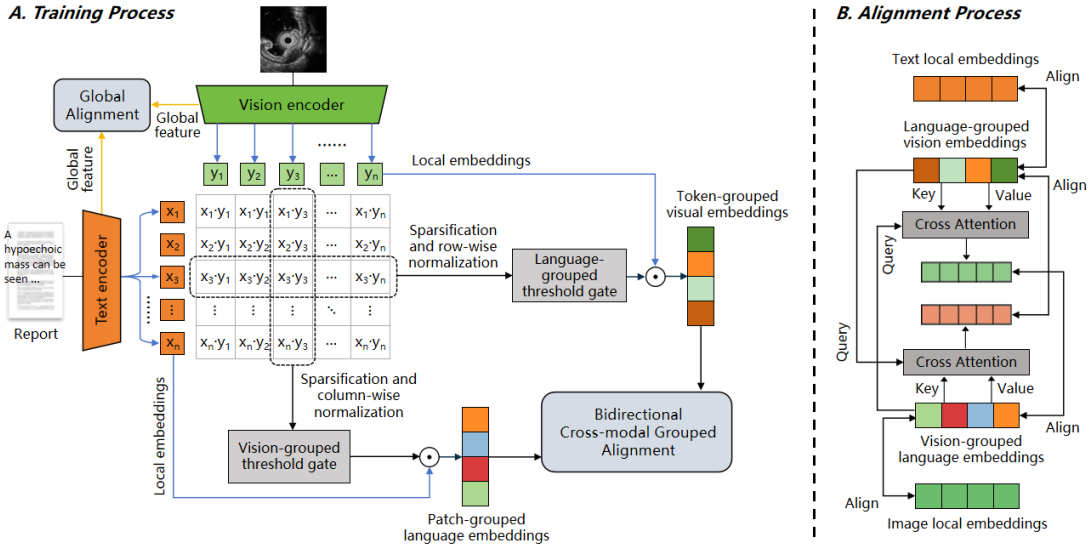


Figure 2: . Overview of the proposed adaptive group alignment (AGA) framework. Dynamic grouping is achieved via matrix sparsification and the grouping threshold gates, while the Bidirectional Cross-modal Grouped Alignment (BCGA) module facilitates effective cross-modal grouped alignment.

3.1 Image and text encoding

Given a batch of paired inputs $\{(x_1^v, x_1^t), \dots, (x_b^v, x_b^t)\}$, where (x_i^v, x_i^t) denotes the i -th image-text pair within the batch and b represents the batch size, we employ an image encoder and a text encoder to extract both global and local features from each modality. These global and local features are subsequently utilized for further processing within our framework. The encoders are trained jointly with our representation learning objectives.

Image encoding: To construct the image encoder E_v , we adopt the ResNet-50 He et al. [2016] architecture as the backbone, initialized with pretrained weights from ImageNet Russakovsky et al. [2015]. We extract local image features from the feature maps of the third bottleneck building block, while global image features are obtained from the final adaptive average pooling layer. Specifically, for the i -th image x_i^v , the corresponding patch embeddings are represented as $V_{i,l} = (v_{i,1}, v_{i,2}, \dots, v_{i,N})$ with $V_{i,N} \in \mathbb{R}^d$, where d denotes the feature dimension and N is the number of patch embeddings. The global image embedding is denoted as $\bar{v}_i \in \mathbb{R}^d$.

Text encoding: We employ a 12-layer BioClinicalBERT Devlin et al. [2019] model, pretrained on medical texts from the MIMIC-III dataset, as our text encoder E_t to obtain clinically relevant text embeddings. By aggregating the embeddings from the last four layers, we derive the local (word-level) embeddings $T_{i,l} = (t_{i,1}, t_{i,2}, \dots, t_{i,M_i})$, where each token embedding $t_{i,M_i} \in \mathbb{R}^d$ and M_i denotes the number of tokens in sample i . The global embedding of the report is represented as $\bar{t}_i \in \mathbb{R}^d$.

3.2 Sparsification of the similarity matrix

To learn structured information, we construct a grouped alignment by grouping words based on image patches and grouping image patches based on words. Specifically, for the i -th image-text pair, its similarity matrix is denoted as $S_i = [S_{i,mn}]_{M_i \times N}$, where $s_{i,mn} = v_{i,n} \cdot t_{i,m}$ represents the inner product between the n -th image patch embedding and the m -th word token embedding, with $n = 1, \dots, N$ and $m = 1, \dots, M_i$. For simplicity, we omit the index i in the following discussion. To obtain alignment weights, we first apply min-max normalization along each row (i.e., for each patch) of the similarity matrix to scale values into the range $[0, 1]$:

$$\hat{s}_{mn} = \frac{s_{mn} - \min_k s_{mk}}{\max_k s_{mk} - \min_k s_{mk}} \quad (1)$$

We sparsify the similarity matrix $S = (\hat{s}_{jk})_{1 \leq j \leq M, 1 \leq k \leq N}$ to facilitate learning and encourage each token to align with only a few patches, i.e.,

$$\tilde{s}_{jk} = \begin{cases} \hat{s}_{jk}, & \text{if } \hat{s}_{jk} \geq \sigma \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where σ denotes the sparsity threshold. The alignment weights are computed as:

$$\alpha_{jk} = \frac{\tilde{s}_{jr}}{\sum_{r=1}^R \tilde{s}_{jr}} \quad (3)$$

Here, α_{jk} represents the weight assigned to the k -th visual patch embedding in the group associated with token j . This approach enables a flexible mapping between each token and an arbitrary number of patch embeddings in the visual domain. For each token t_m , we compute the corresponding TGV embedding $p_m \in \mathbb{R}^d$ as:

$$p_m = \sum_{r=1}^R \alpha_{mr} v_r \quad (4)$$

as a weighted combination of the aligned patch embeddings, where R denotes the number of patches with non-zero alignment weights.

For each image patch v_n , we obtain its corresponding PGL embedding $q_n \in \mathbb{R}^d$ in the same manner.

3.3 Grouping threshold gate

To sparsify the similarity matrix, we introduce a sparsity threshold σ . Unlike using a fixed threshold, we design two dynamic mechanisms: the Language-grouped Threshold Gate and the Vision-grouped Threshold Gate, which adaptively adjust the threshold during training. The dynamic threshold selection process in the Language-grouped Threshold Gate is as follows:

$$\sigma_{tg} = \gamma_{tg} \cdot \sigma_{tg} + (1 - \gamma_{tg}) \cdot \bar{S} \quad (5)$$

where γ_{tg} denotes the momentum, which controls the update rate of the threshold, and \bar{S} represents the running average of the similarity matrix for the image-text pairs. Both image patches and text tokens originate from the same

image-text pair. Patches with similarity scores higher than the historical average similarity threshold are assigned to the corresponding token groups. In contrast, patches with lower similarity scores are considered irrelevant to the group. Accordingly, Eq. 2 can be reformulated as:

$$\tilde{s}_{jk} = \begin{cases} \hat{s}_{jk}, & \hat{s}_{jk} \geq \sigma_{tg} \\ 0, & \hat{s}_{jk} < \sigma_{tg} \end{cases} \quad (6)$$

Thus, enabling dynamic threshold selection and enhancing the flexibility of grouping.

Similarly, for the Vision-grouped Threshold Gate, a dynamically updated threshold is set by operating on the transposed similarity matrix S :

$$\sigma_{vg} = \gamma_{vg} \cdot \sigma_{vg} + (1 - \gamma_{vg}) \cdot \overline{S^T} \quad (7)$$

where γ_{vg} represents the momentum.

3.4 Alignment process

Global alignment: To capture global information, AGA employs a global contrastive loss Jia et al. [2021], Radford et al. [2021] at the level of global image embeddings \bar{v}_i and global text embeddings \bar{t}_i . Specifically, we optimize the similarity between each image and its corresponding text embedding, while minimizing the similarity with non-matching image-text pairs within the batch. The objective can be formulated as:

$$\mathcal{L}_g = -\frac{1}{2b} \sum_{i=1}^b \left(\log \frac{\exp(\phi(\bar{v}_i, \bar{t}_i)/\tau_1)}{\sum_{j=1}^b \exp(\phi(\bar{v}_i, \bar{t}_j)/\tau_1)} + \log \frac{\exp(\phi(\bar{t}_i, \bar{v}_i)/\tau_1)}{\sum_{j=1}^b \exp(\phi(\bar{t}_i, \bar{v}_j)/\tau_1)} \right) \quad (8)$$

where $\phi(\bar{v}_i, \bar{t}_j) = \frac{\bar{v}_i \cdot \bar{t}_j}{\|\bar{v}_i\|_2 \cdot \|\bar{t}_j\|_2}$, and τ denotes the temperature parameter.

Grouping alignment: The group alignment process, as illustrated in Fig. 2b, consists of two components. For the i -th image-text pair (x_i^v, x_i^t) , we first perform fine-grained alignment between text token embeddings and their corresponding TGV embeddings, as well as between image patch embeddings and their corresponding PGL embeddings. We introduce the IGA loss, which operates at the level of individual image-text pairs over sequences of tokens and patches, without requiring other pairs as negative samples. This design significantly reduces both computational and memory overhead. The IGA losses for text tokens and image patches are denoted as L_{tf} and L_{vf} , respectively:

$$\mathcal{L}_{tf} = -\frac{1}{2b} \sum_{i=1}^b \left[\frac{1}{M_i} \sum_{j=1}^{M_i} \left(\log \frac{\exp(\phi(p_i^j, t_i^j)/\tau_2)}{\sum_{k=1}^{M_i} \exp(\phi(p_i^j, t_i^k)/\tau_2)} + \log \frac{\exp(\phi(t_i^j, p_i^j)/\tau_2)}{\sum_{k=1}^{M_i} \exp(\phi(t_i^j, p_i^k)/\tau_2)} \right) \right] \quad (9)$$

$$\mathcal{L}_{vf} = -\frac{1}{2b} \sum_{i=1}^b \left[\frac{1}{N} \sum_{j=1}^N \left(\log \frac{\exp(\phi(q_i^j, v_i^j)/\tau_2)}{\sum_{k=1}^N \exp(\phi(q_i^j, v_i^k)/\tau_2)} + \log \frac{\exp(\phi(v_i^j, q_i^j)/\tau_2)}{\sum_{k=1}^N \exp(\phi(v_i^j, q_i^k)/\tau_2)} \right) \right] \quad (10)$$

Here, t_i^j and v_i^j denote the j -th text token embedding and image patch embedding for the i -th image-text pair, respectively. p_i^j and q_i^j represent the j -th TGV and PGL embeddings, respectively. The IGA loss aims to maximize the

similarity between each token or patch embedding and its corresponding TGV or PGL embedding, while minimizing its similarity with other TGV or PGL embeddings within the same sequence, and vice versa.

Secondly, we perform grouped alignment between the TGV embeddings and PGL embeddings. To explicitly match and align cross-modal grouped representations between medical images and radiology reports, we adopt an efficient Bidirectional Cross-modal Grouped Alignment (BCGA) module. This module employs a cross-attention mechanism Vaswani et al. [2017] to compute soft alignments between the generated TGV and PGL embeddings. Specifically, for the i -th image-text pair, the generated TGV embeddings and PGL embeddings are first normalized, resulting in $P_i = \{p_i^1, p_i^2, \dots, p_i^{M_i}\}$ and $Q_i = \{q_i^1, q_i^2, \dots, q_i^N\}$, where $p_i \in \mathbb{R}^d$, $q_i \in \mathbb{R}^d$. For each TGV embedding p_i^j , we attend over all PGL embeddings Q_i , and compute the corresponding cross-modal TGV embedding u_i^j ,

$$u_i^j = \sum_{k=1}^N o\left(\beta_i^{j2k} (Vq_i^k)\right), \quad (11)$$

$$\beta_i^{j2k} = \text{softmax}\left(\frac{(Qp_i^j)^T (Kq_i^k)}{\sqrt{d}}\right)$$

Here, $Q \in \mathbb{R}^{d \times d}$, $K \in \mathbb{R}^{d \times d}$ and $V \in \mathbb{R}^{d \times d}$ denote learnable projection matrices. After that, we apply the grouped language-to-vision alignment loss \mathcal{L}_{gla} , which encourages each TGV embedding p_i^j to be close to its corresponding cross-modal TGV embedding u_i^j , while pushing it away from other cross-modal TGV embeddings. This objective effectively maximizes a lower bound on the group-level cross-modal mutual information within each image-report pair Oord et al. [2018]. The \mathcal{L}_{gla} loss is formulated as follows:

$$\mathcal{L}_{gla} = -\frac{1}{2b} \sum_{i=1}^b \left[\frac{1}{M_i} \sum_{j=1}^{M_i} \left(\log \frac{\exp(\phi(p_i^j, u_i^j)/\tau_3)}{\sum_{k=1}^{M_i} \exp(\phi(p_i^j, u_i^k)/\tau_3)} \right. \right. \\ \left. \left. + \log \frac{\exp(\phi(u_i^j, p_i^j)/\tau_3)}{\sum_{k=1}^{M_i} \exp(\phi(u_i^j, p_i^k)/\tau_3)} \right) \right] \quad (12)$$

Similarly, we obtain the cross-modal PGL embeddings w_i^j , and define the grouped vision-to-language alignment loss \mathcal{L}_{gva} as follows:

$$\mathcal{L}_{gva} = -\frac{1}{2b} \sum_{i=1}^b \left[\frac{1}{N} \sum_{j=1}^N \left(\log \frac{\exp(\phi(q_i^j, w_i^j)/\tau_3)}{\sum_{k=1}^N \exp(\phi(q_i^j, w_i^k)/\tau_3)} \right. \right. \\ \left. \left. + \log \frac{\exp(\phi(w_i^j, q_i^j)/\tau_3)}{\sum_{k=1}^N \exp(\phi(w_i^j, q_i^k)/\tau_3)} \right) \right] \quad (13)$$

The final loss function is defined as follows:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_g + \frac{\lambda_2}{2} (\mathcal{L}_{tf} + \mathcal{L}_{vf}) + \frac{\lambda_3}{2} (\mathcal{L}_{gla} + \mathcal{L}_{gva}) \quad (14)$$

where λ_1 , λ_2 , and λ_3 are hyperparameters that balance different components of the alignment process.

4 Experiments

We first introduce the paired dataset used for contrastive pretraining, the datasets employed for downstream task evaluation, and the baseline methods for comparison.

4.1 Experimental datasets

4.1.1 Datasets for Pretraining

MIMIC-CXR Johnson et al. [2019]: We utilize the second version of the publicly available MIMIC-CXR dataset, a large scale and openly accessible collection of chest radiographs paired with corresponding radiology reports. Consistent

with prior work Li et al. [2023], Tanida et al. [2023], Yang et al. [2022], we use the *findings* section of the raw radiology reports as reference texts. After preprocessing, the training, validation, and test sets contain 270,742 (152142), 2130 (1196), and 3858 (2347) image-report pairs, respectively. The numbers in parentheses indicate splits based on the unique “study_id”.

SMTs: We collected a private dataset consisting of image-report pairs related to submucosal tumors (SMTs) of the gastrointestinal tract from the Third People’s Hospital of Chengdu. The study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the Third People’s Hospital of Chengdu on September 25, 2024 (IRB No. 2023-S-48-1). Due to the retrospective nature of the study, the requirement for informed consent was waived. To ensure patient privacy, all personally identifiable information has been removed. The dataset includes EUS (Endoscopic Ultrasound) images and corresponding textual reports collected from five different hospitals. We used data from four hospitals to construct our pretraining dataset, comprising 2455 (547), 266 (68), and 600 (120) EUS image-report pairs for the training, validation, and test sets, respectively. The numbers in parentheses indicate the quantities grouped by patient ID. Data from the same patient were assigned exclusively to a single split to ensure fair and non-overlapping partitioning. The data from the remaining hospital were reserved for downstream task evaluation.

4.1.2 Downstream Task Datasets

CheXpert 5×200: The dataset contains five abnormality categories sampled from the CheXpert dataset Irvin et al. [2019], namely Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion, with 200 image-report pairs per category Huang et al. [2021], Zhang et al. [2022a]. Each instance in the dataset corresponds to a single abnormality category.

RSNA Pneumonia Shih et al. [2019]: We use the training set from the second phase version of this dataset, as test set labels are not available. This subset contains approximately 26700 frontal chest X-rays. The task is to classify whether each chest image exhibits Lung Opacity.

SMTs 3×200: We use the SMTs pretraining test set as a downstream task subset, which includes three tumor categories (gastrointestinal stromal tumors (GISTs), neuroendocrine tumors (NETs) and Leiomyomas) randomly sampled from four hospitals, with 200 image-text pairs per category. Each instance in the dataset belongs to a single tumor category.

SMTs SN: This dataset is an image-text dataset with the same structure as the SMTs dataset, obtained from a single hospital. It contains 584 pairs of data from 120 patients, covering four categories: GISTs, NETs, leiomyomas, and others. Each example belongs to a single abnormality category.

4.2 Experimental results

4.2.1 Implementation details

The experiments are conducted on a platform with four NVIDIA GeForce RTX 3090 GPUs running Ubuntu 20.04, using Python 3.7.16 and PyTorch 1.12.1. The maximum number of training epochs is set to 50, with a batch size of 48. The optimizer used is AdamW, with an initial learning rate of 5×10^{-5} for both the SMTs and MIMIC-CXR datasets. For the SMTs and MIMIC datasets, the momentum hyperparameters γ_{tg} (or γ_{vg}) are set to 0.99 and 0.999, respectively. Following the practice in contrastive learning Zhang et al. [2022a], Taleb et al. [2022], the embedding dimension d is set to 128, and the temperature hyperparameters τ_1 , τ_2 and τ_3 are set to 0.3, 0.3, and 0.1, respectively. The loss weights λ_1 , λ_2 and λ_3 are all set to 0.5 by default. For comparison, we select the classical methods ConVIRT Zhang et al. [2022a], Gloria Huang et al. [2021], MGCA Wang et al. [2022], and SPARC Bica et al. [2024]. These methods are reproduced on the same datasets using identical image and text encoders to ensure a fair comparison of different alignment strategies. We pretrain these models separately on the two pretraining datasets and apply them to their corresponding downstream tasks for evaluation.

4.2.2 Image-text retrieval

We first evaluate the effectiveness of our representation learning framework on image-text retrieval using three datasets: CheXpert 5×200, SMTs 3×200, and SMTs SN. Following the setup in Huang et al. [2021], given an image as the input query, the goal is to retrieve the target report by computing similarity scores between the query image and all candidate reports using the learned representations. By checking whether the retrieved report belongs to the same category as the query image, we evaluate retrieval accuracy using the Precision@ K metric. The top- K precision scores are reported for $K = 5, 10$, and 100.

Table 1 presents the image-to-text retrieval results on the downstream CheXpert 5×200 dataset after pretraining the models on the MIMIC-CXR dataset. As shown in the table, our proposed AGA framework consistently outperforms

other baselines. Compared to methods incorporating local contrastive losses, such as Gloria Huang et al. [2021], MGCA Wang et al. [2022], and SPARC Bica et al. [2024], AGA achieves superior performance, reaching a Precision@5 of 50.28. This demonstrates that the group alignment strategy provides more effective structured feature representations than conventional fine-grained alignment approaches. Table 2 reports the image-to-text retrieval results on the downstream SMTs 3×200 and SMTs SN datasets, following pretraining on the SMTs dataset. Our model achieves Precision@5 scores of 55 and 42.43, respectively. These results further highlight the competitiveness of the proposed model in capturing discriminative cross-modal representations under varying clinical data distributions.

Table 1: Results of image-to-text retrieval on the CheXpert 5×200 dataset.

Method	CheXpert 5×200		
	Prec@5	Prec@10	Prec@100
ConVIRTZhang et al. [2022a]	48.30	47.92	42.00
GloriaHuang et al. [2021]	45.56	43.78	39.26
MGCAWang et al. [2022]	49.54	49.10	42.25
SPARCBica et al. [2024]	47.44	47.36	42.73
AGA	50.28	49.84	43.80

4.2.3 Image Classification

We further evaluate the learned feature representations on two distinct image classification tasks: supervised classification and zero-shot classification.

Supervised Classification. For supervised classification, following the setting in Zhang et al. [2022a], we attach a linear layer to the pretrained image encoder and train the model using varying proportions of labeled data (1%, 10%, and 100%) to evaluate the data efficiency of global image representations. Table 3 reports the area under the ROC curve (AUC) on the downstream CheXpert 5×200 and RSNA Pneumonia datasets. As the amount of training data increases, the AUC also improves. When fine-tuned with only 1% of the training data, our model achieves AUC scores of 56.1 and 68.91 on CheXpert 5×200 and RSNA Pneumonia, respectively, both surpassing the baseline methods and demonstrating superior data efficiency. Due to the limited sample size in the private SMTs 3×200 and SMTs SN datasets, which is insufficient to support a 1% fine-tuning scenario, we merge them into a unified dataset, SMTs 3×200-SN, for the supervised classification task. Table 4 shows the results on the SMTs 3×200-SN dataset, where we observe a similar trend as in Table 3. When fine-tuned with 10%, 50%, and 100% of the training data, our model achieves AUC scores of 57.04, 59.28, and 83.71, respectively, consistently outperforming baseline methods.

Zero-shot classification. For zero-shot classification, we use an image as input with the objective of predicting the corresponding label, even though the model is not explicitly trained with class labels. Inspired by Huang et al. [2021], we convert each classification category into a textual prompt. Specifically, for datasets involving chest diseases, we adopt the prompt engineering strategy from Huang et al. [2021] to generate representative textual descriptions for each category, capturing possible subtypes, severity levels, and anatomical locations of the medical conditions. Subsequently, all category prompts and the input image are projected into a shared multimodal embedding space using the pretrained representation learning model, and the label associated with the prompt that has the highest similarity score is selected as the prediction.

Table 5 reports the zero-shot classification results on the CheXpert 5×200 and RSNA Pneumonia datasets, with our method achieving accuracies of 63.6 and 51.1, respectively, outperforming the baseline models. Compared to existing fine-grained alignment approaches such as GLORIA and MGCA, our group-wise alignment strategy captures fine-grained semantic relationships between image regions and textual descriptions more effectively. Table 6 presents the results on the private SMTs 3×200 and SMTs SN datasets, where our model achieves accuracies of 56.5 and 61.5, respectively, further demonstrating its effectiveness.

Table 2: Results of image-to-text retrieval on the SMTs 3×200 and SMTs SN datasets.

Method	SMTs 3×200			SMTs SN		
	Prec@5	Prec@10	Prec@100	Prec@5	Prec@10	Prec@100
ConVIRTZhang et al. [2022a]	48.83	48.67	43.14	40.68	39.35	37.23
GloriaHuang et al. [2021]	54.17	52.33	42.99	34.93	34.93	38.60
MGCAWang et al. [2022]	48.83	50.83	44.58	36.44	36.03	38.06
SPARCBica et al. [2024]	44.33	45.00	41.40	27.40	32.81	32.35
AGA	55.00	54.42	43.16	42.43	45.07	37.40

Table 3: Results of linear classification on CheXpert 5×200 and RSNA with 1%, 10%, 100% training data.

Method	CheXpert 5×200 (AUC)			RSNA Pneumonia (AUC)		
	1%	10%	100%	1%	10%	100%
ConVIRTZhang et al. [2022a]	46.62	56.67	83.69	66.20	73.08	86.40
GloriaHuang et al. [2021]	54.75	57.81	84.20	67.84	74.51	86.92
MGCAWang et al. [2022]	55.21	59.21	83.92	67.28	76.12	87.76
SPARCBica et al. [2024]	54.12	58.12	83.65	68.21	75.23	87.47
AGA	56.10	61.31	84.32	68.91	75.78	87.92

Table 4: Results of linear classification on SMTs 3×200-SN with 1%, 10%, 100% training data.

Method	SMTs 3×200-SN(AUC)		
	10%	50%	100%
ConVIRTZhang et al. [2022a]	49.91	54.96	83.63
GloriaHuang et al. [2021]	50.75	59.17	80.78
MGCAWang et al. [2022]	55.67	55.51	83.00
SPARCBica et al. [2024]	45.94	49.59	80.55
AGA	57.04	59.28	83.71

4.2.4 Ablation studies

This section evaluates the contribution of different components in our method. We conduct ablation studies on three variants of the pretraining setup: (1) Only Global Alignment, (2) Removing the BCGA module, and (3) Using a Fixed Threshold. These experiments are designed to assess the effectiveness of our group alignment strategy, the IGA Loss, and the dynamic threshold gating mechanism. For the fixed threshold setting, we set the parameters to $\sigma_{tg} = 1/361$ and $\sigma_{vg} = 1/97$, where 361 corresponds to the default number of image patch embeddings and 97 is the default maximum number of text tokens. This ensures that each text token and image patch receives a corresponding group representation.

Table 7 presents the image-to-text retrieval results on the CheXpert 5×200 dataset under different pretraining settings. Compared to our full model, the Only Global Alignment setting shows a performance drop of approximately 4%, indicating that our group alignment strategy is effective in capturing fine-grained semantics beyond global features. Notably, the Removing the BCGA module setting yields the largest performance degradation, highlighting the importance of inter-group alignment. Without the BCGA module, the model relies solely on IGA to perform intra-group alignment between text tokens and its TGV embeddings. However, the absence of inter-group alignment leads to semantically disjoint group representations, which hinders the model’s ability to reason over the global context. Additionally, the groups constructed through IGA may contain noisy or ambiguous associations that cannot be refined without BCGA. As a result, the model’s performance significantly degrades, even falling below the level of only global alignment, which emphasizes the critical role of the joint use of BCGA and IGA in achieving fine-grained semantic integration and contextual consistency. The Fixed Threshold variant shows the smallest performance drop (approximately 3%), which first demonstrates the overall robustness of our framework in capturing fine-grained information. Moreover, the adaptive threshold gating module further improves overall performance by dynamically adjusting the grouping thresholds, validating the effectiveness of the adaptive mechanism. Table 8 reports the image-to-text retrieval performance on the SMTs 3×200 and SMTs SN datasets under the same settings. Similar trends are observed, which further supports the consistency and generalization of our findings.

Table 5: Results of zero-shot image classification on the CheXpert 5×200 and RSNA datasets.

Method	CheXpert 5×200			RSNA Pneumonia		
	ACC	F_1	ROC	ACC	F_1	ROC
ConVIRTZhang et al. [2022a]	63.4	25.9	53.0	50.3	36.5	53.5
GloriaHuang et al. [2021]	62.4	22.9	50.8	50.3	36.0	54.7
MGCAWang et al. [2022]	62.3	25.1	50.3	50.2	40.9	54.9
SPARCBica et al. [2024]	61.0	23.6	48.8	49.8	38.6	54.5
AGA	63.6	26.0	52.4	51.1	37.6	55.3

Table 6: Results of zero-shot image classification on the SMTs 3×200 and SMTs SN datasets.

Method	SMTs 3×200			SMTs SN		
	ACC	F_1	ROC	ACC	F_1	ROC
ConVIRTZhang et al. [2022a]	55.2	29.5	44.9	61.0	27.5	49.5
GloriaHuang et al. [2021]	56.4	30.4	48.6	61.3	25.8	46.2
MGCAWang et al. [2022]	55.6	32.7	42.8	61.0	28.2	53.6
SPARCBica et al. [2024]	53.0	30.9	46.0	58.5	25.8	45.9
AGA	56.5	28.9	49.3	61.5	24.2	55.4

Table 7: Ablation results on the image-to-text retrieval task on the CheXpert 5×200 dataset.

Method	CheXpert 5×200		
	Prec@5	Prec@10	Prec@100
Only global alignment	45.34	45.04	40.96
No BCGA	34.28	33.83	31.21
AGA(fixed)	48.54	47.12	39.14
AGA	50.28	49.84	43.80

5 Discussion

5.1 Variation of Grouping Thresholds

We visualize the variation of grouping thresholds during model pretraining using line plots. As shown in Fig. 3, the curves illustrate the evolution of the language grouping threshold σ_{tg} and the visual grouping threshold σ_{vg} over optimization steps on the MIMIC-CXR dataset. It can be observed that σ_{tg} stabilizes around 0.15, while σ_{vg} stabilizes around 0.22, with a noticeable upward trend in the later stages. The fact that σ_{vg} surpasses σ_{tg} indicates that each image patch in a chest X-ray tends to align with an declining number of textual tokens. One possible explanation is that in the MIMIC-CXR dataset, findings related to specific anatomical regions are often described across multiple sentences. The language tends to be more loosely structured and unstructured. For example, the statements “There is a right lower lobe opacity.” and “This may represent pneumonia.” are independent sentences, yet together they describe the same abnormality in the right lower lobe.

In contrast, Fig. 4 also presents the line plots of the thresholds σ_{tg} and σ_{vg} over optimization steps on the SMTs dataset. It can be observed that σ_{tg} stabilizes around 0.45 and σ_{vg} around 0.43, with both values being nearly equal. This indicates a strong correlation between individual textual tokens and single image patches. This phenomenon can be interpreted in the dataset context as the private SMTs dataset typically features sentences that comprehensively describe multiple attributes of a single region. The descriptions are more structured and focused, for example: “A hypoechoic mass can be seen in the lesion, which is oval in shape, protruding into and outside the cavity, with...”. Coreference is rare, and the association between words and image regions is strong, which contrasts with the more loosely structured and distributed descriptions seen in the MIMIC-CXR dataset.

5.2 Visualization of attention weights

Fig. 5 presents a qualitative visualization of the learned word-to-region correspondences facilitated by our AGA framework. The top row shows the original medical images, including both chest X-rays (CXR) and endoscopic ultrasound (EUS) images. The bottom row displays the corresponding heatmaps generated by our model, where warmer colors denote higher activation weights, indicating stronger associations between specific image regions and the given medical concepts. For Atelectasis and Pneumonia, the model focuses on appropriate pulmonary regions, demonstrating strong localization aligned with radiological pathology. For the SMTs domain, terms like low-echoic

Table 8: Ablation results on the image-to-text retrieval task on the SMTs 3×200 and SMTs SN datasets.

Method	SMTs 3×200			SMTs SN		
	Prec@5	Prec@10	Prec@100	Prec@5	Prec@10	Prec@100
Only global alignment	48.83	48.67	43.14	40.68	39.35	37.23
No BCGA	45.33	44.58	42.15	39.55	37.84	38.09
AGA(fixed)	51.00	52.42	42.41	38.84	38.49	36.54
AGA	55.00	54.42	43.16	42.43	45.07	37.40

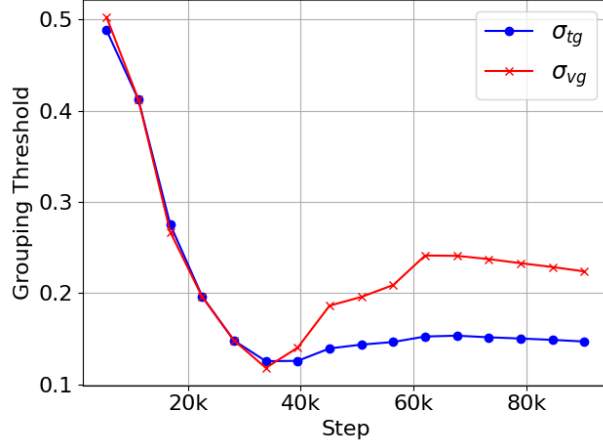


Figure 3: Line plot of language grouping threshold σ_{tg} and visual grouping threshold σ_{vg} on the MIMIC-CXR dataset during pretraining.

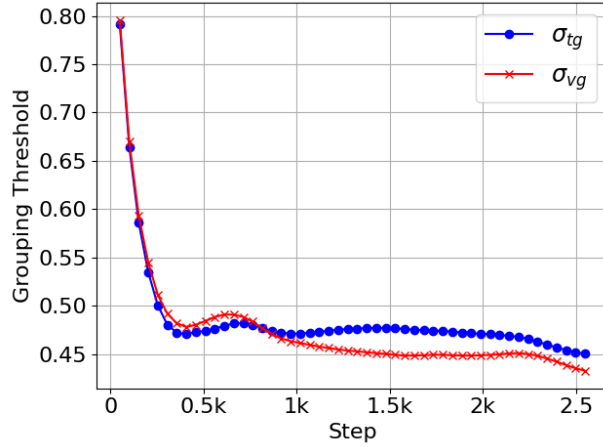


Figure 4: Line plot of language grouping threshold σ_{tg} and visual grouping threshold σ_{vg} on the SMTs dataset during pretraining.

mass and protruded into the cavity activate precisely the relevant interior structures of the lesion in EUS images. The distinct and interpretable activation patterns validate the effectiveness of our AGA mechanism in achieving fine-grained multimodal alignment.

5.3 Visualization of encoded image representations

To qualitatively assess the effectiveness of our AGA framework, we visualize the learned image representations using t-SNE on three datasets: CheXpert 5×200, SMTs 3×200, and SMTs SN. The results are shown in Fig. 6. In subfigure (a), despite the limited supervision and complex semantics of chest X-ray images, our model captures meaningful intra-class patterns, with moderate separation between disease categories such as Atelectasis, Cardiomegaly, and Pleural Effusion. Some overlaps are observed between semantically similar conditions, which may stem from the ambiguity in clinical labels. In contrast, subfigure (b) shows more distinct and compact clusters on the SMTs 3×200 dataset, suggesting that the structured and domain-specific textual annotations in our private dataset enhance the semantic alignment between visual and textual modalities. Subfigure (c) further demonstrates the generalization ability of our method on SMTs SN, a more challenging dataset with a broader category distribution. The representation space remains clearly structured, and additional disease types such as "Other" are accommodated without disrupting the separability of core classes like GISTs, Leiomyoma, and NETs. These results confirm that our grouping-based strategy enabled by the IGA loss and BCGA module, effectively promotes intra-class cohesion and inter-class discrimination. The visualizations provide

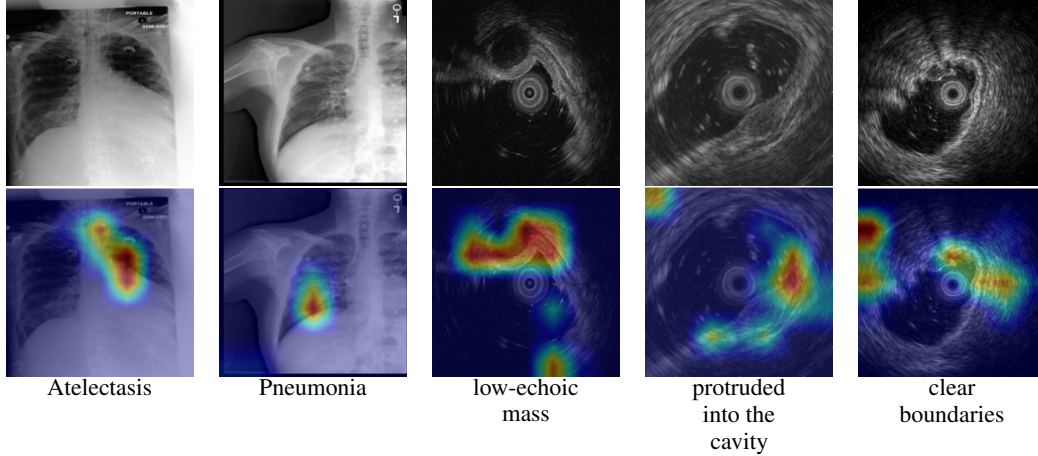


Figure 5: Visualization of learned token correspondence by our AGA. Highlighted pixels represent higher activation weights by corresponding word.

strong qualitative evidence that AGA learns semantically grounded and context-consistent representations, which serve as a robust foundation for downstream tasks such as classification and retrieval.

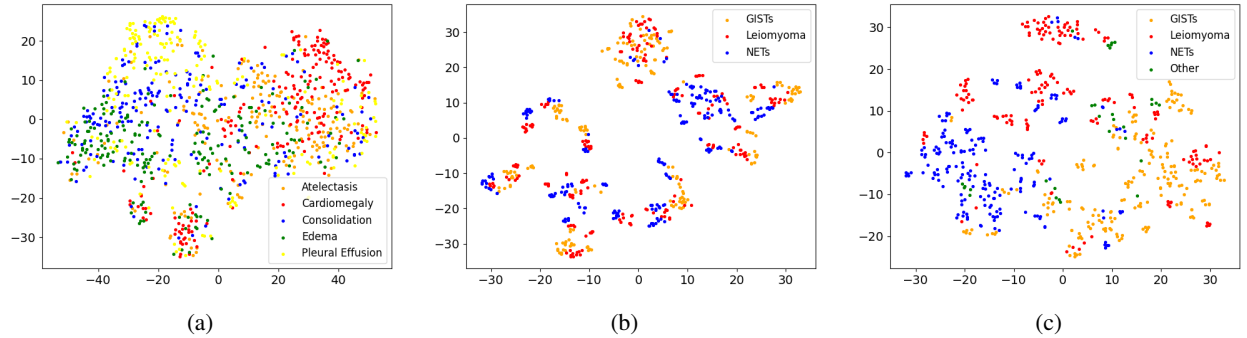


Figure 6: . T-SNE visualizations of encoded image representations. Colors represent the ground truth disease types. Subfigures (a), (b), and (c) correspond to the results on the CheXpert 5×200, SMTs 3×200, and SMTs SN datasets, respectively.

6 Conclusion

In this work, we propose an AGA framework for cross-modal medical visual representation learning. Our approach introduces an IGA loss that captures fine-grained associations between textual tokens and their corresponding group representations, supported by a BCGA module to refine group interactions and promote semantic consistency. To address the granularity inconsistency across datasets and enable adaptive grouping, we introduce a dynamic grouping threshold gate that learns to adjust thresholds during training, facilitating more flexible alignment based on the data. We conduct extensive evaluations across image-to-text retrieval, supervised classification, and zero-shot classification tasks on both public and private datasets. The results demonstrate that our model outperforms existing methods, especially in low-data regimes. Visualization analyses including feature scatter plots and activation heatmaps further confirm that the learned representations exhibit strong intra-class cohesion and precise semantic grounding. Compared with MIMIC-CXR, the SMTs datasets exhibit more structured descriptions, leading to nearly identical visual and textual grouping thresholds and stronger word-region correlations.

Since our work primarily focuses on medical visual representation learning, we did not evaluate performance on text-based downstream tasks, which can be regarded as a limitation of this study. In future work, we plan to extend the grouping strategy to the sample level to enable alignment across groups of samples. We also intend to integrate our approach with generation-based pre-training methods to facilitate joint learning of image and textual features.

Declaration of Interests

Authors declare that they have no conflict of interest.

Acknowledgments

This work is partially supported by National Natural Science Foundation of China (62376231), Sichuan Science and Technology Program (24NSFSC1070), Fundamental Research Funds for the Central Universities (2682025ZTPY052, 2682023ZDPY001).

References

- Geeticka Chauhan, Ruizhi Liao, William Wells, Jacob Andreas, Xin Wang, Seth Berkowitz, Steven Horng, Peter Szolovits, and Polina Golland. Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. *Medical Image Computing and Computer Assisted Intervention-MICCAI 2020: 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part II* 23. Springer International Publishing, pages 529–539, 2020.
- Zizhao Zhang, Pingjun Chen, Manish Sapkota, and Lin Yang. Tandemnet: Distilling knowledge from medical images using diagnostic reports as optional semantic references. *Medical Image Computing and Computer Assisted Intervention-MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III* 20. Springer International Publishing, pages 320–328, 2017.
- Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *Machine learning for healthcare conference. PMLR*, pages 2–25, 2022a.
- Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. Learning to exploit temporal structure for biomedical vision-language processing. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15016–15027, 2023.
- Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3942–3951, 2021.
- Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in Neural Information Processing Systems*, 35:33536–33549, 2022.
- Yan Gao, Zhiwei Ni, Wentao Liu, Liping Ni, Ling Xin, Linbo Hu, and Li Zhang. Abnormal-region-aware multi-modal feature fusion for medical report generation. *Knowledge-Based Systems*, 318:113538, 2025.
- Pujin Cheng, Li Lin, Junyan Lyu, Yijin Huang, Wenhan Luo, and Xiaoying Tang. Prior: Prototype representation joint learning from medical images and reports. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21361–21371, 2023.
- Ke Zhang, Yan Yang, Jun Yu, Hanliang Jiang, Jianping Fan, and Qingming Huang. Multi-task paired masking with alignment modeling for medical vision-language pre-training. *IEEE Transactions on Multimedia*, 26:4706–4721, 2023a.
- Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *Advances in neural information processing systems*, 33:21798–21809, 2020.
- R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.
- Ji Zhang, Jingkuan Song, Lianli Gao, Nicu Sebe, and Hengtao Shen. Reliable few-shot learning under dual noises. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–17, 2025.
- Ji Zhang, Jingkuan Song, Lianli Gao, Ye Liu, and Hengtao Shen. Progressive meta-learning with curriculum. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):5916–5930, 2022b.
- Shihuan He, Zhihui Lai, Ruxin Wang, and Heng Kong. Prototype contrastive consistency learning for semi-supervised medical image segmentation. *arXiv preprint arXiv:2502.06650*, 2025.

- Yumin Zhang, Hongliu Li, Yajun Gao, Haoran Duan, Yawen Huang, and Yefeng Zheng. Prototype correlation matching and class-relation reasoning for few-shot medical image segmentation. *IEEE Transactions on Medical Imaging*, 43(11):4041–4054, 2024.
- Jie Wang, Tianrui Li, Yan Yang, Shiqian Chen, and Wanming Zhai. Diagllm: multimodal reasoning with large language model for explainable bearing fault diagnosis. *Science China Information Sciences*, 68(6):160103, 2025a.
- Ioana Bica, Anastasija Ilić, Matthias Bauer, Goker Erdogan, Matko Bošnjak, Christos Kaplanis, Alexey A Gritsenko, Matthias Minderer, Charles Blundell, Razvan Pascanu, and Jovana Mitrović. Improving fine-grained understanding in image-text pre-training. *arXiv preprint arXiv:2401.09865*, 2024.
- Haiwei Pan, Shuning He, Kejia Zhang, Bo Qu, Chunling Chen, and Kun Shi. Amam: An attention-based multimodal alignment model for medical visual question answering. *Knowledge-Based Systems*, 255:109763, 2022.
- Hongzhao Li, Hongyu Wang, Xia Sun, Hua He, and Jun Feng. Context-enhanced framework for medical image report generation using multimodal contexts. *Knowledge-Based Systems*, 310:112913, 2025.
- Lin Fan, Xun Gong, Cenyang Zheng, Xuli Tan, Jiao Li, and Yafei Ou. Cycle-vqa: A cycle-consistent framework for robust medical visual question answering. *Pattern Recognition*, 165:111609, 2025.
- Ji Zhang, Lianli Gao, Bingguang Hao, Hao Huang, Jingkuan Song, and Hengtao Shen. From global to local: Multi-scale out-of-distribution detection. *IEEE Transactions on Image Processing*, 32:6115–6128, 2023b.
- Ruizhi Liao, Daniel Moyer, Miriam Cha, Keegan Quigley, Seth Berkowitz, Steven Horng, Polina Golland, and William M Wells. Multimodal representation learning via maximization of local mutual information. *Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference, Strasbourg, France, September 27-October 1, 2021, Proceedings, Part II 24. Springer International Publishing*, pages 273–283, 2021.
- Constantin Seibold, Simon Reiß, M Saquib Sarfraz, Rainer Stiefelhagen, and Jens Kleesiek. Breaking with fixed set pathology recognition through report-guided contrastive training. *International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland*, pages 690–700, 2022.
- Anil Palepu and Andrew Beam. Tier: Text-image entropy regularization for medical clip-style models. *Machine Learning for Healthcare Conference. PMLR*, 219:548–564, 2023.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. in international conference on machine learning. *International conference on machine learning. PMLR*, 119:1597–1607, 2020.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.
- Ruibin Feng, Zongwei Zhou, Michael B Gotway, and Jianming Liang. Parts2whole: Self-supervised contrastive learning via reconstruction. *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4-8, 2020, Proceedings 2. Springer International Publishing*, pages 85–95, 2020.
- Jean-Bastien Grill, Florian Strub, Florent Althé, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.
- Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in neural information processing systems*, 33: 12546–12558, 2020.
- Yan Han, Chongyan Chen, Ahmed Tewfik, Ying Ding, and Yifan Peng. Pneumonia detection on chest x-ray using radiomic features and contrastive learning. *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 247–251, 2021.
- Aihm Taleb, Matthias Kirchler, Remo Monti, and Christoph Lippert. Contig: Self-supervised multimodal contrastive learning for medical imaging with genetics. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20908–20921, 2022.
- Gang Wang, Yajun Du, and Yurui Jiang. Lspcl: Label-specific supervised prototype contrastive learning for multi-label text classification. *Knowledge-Based Systems*, 309:112887, 2025b.
- Xudong Wang, Ziwei Liu, and Stella X Yu. Unsupervised feature learning by cross-level instance-group discrimination. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12586–12595, 2021.

- Yuanfan Guo, Minghao Xu, Jiawen Li, Bingbing Ni, Xuanyu Zhu, Zhenbang Sun, and Yi Xu. Hcsc: Hierarchical contrastive selective coding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9706–9715, 2022.
- Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. *Proceedings of the AAAI conference on artificial intelligence*, 35(10):8547–8555, 2021.
- Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33: 9912–9924, 2020.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Feifei Li. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*, 1:4171–4186, 2019.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *International conference on machine learning. PMLR*, 139:4904–4916, 2021.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *International conference on machine learning. PMLR*, 139:8748–8763, 2021.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- Alistair E. W Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.
- Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. Dynamic graph enhanced contrastive learning for chest x-ray report generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3334–3343, 2023.
- Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable regionguided radiology report generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442, 2023.
- Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. Knowledge matters: chest radiology report generation with general and specific knowledge. *Medical image analysis*, 80:102510, 2022.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghighi, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A Mong, Safwan S Halabi, Jesse K Sandberg, Ricky Jones, David B Larson, Curtis P Langlotz, Bhavik N Patel, Matthew P Lungren, and Andrew Y Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI conference on artificial intelligence*, 33(1):590–597, 2019.
- George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, Ritu R Gill, Myrna C.B Godoy, Stephen Hobbs, Jean Jeudy, Archana Laroia, Palmi N Shah, Dharshan Vummidi, Kavitha Yaddanapudi, and Anouk Stein. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019.