biomolecules

MDPI

Article

MedicalCLIP: Anomaly-Detection Domain Generalization with Asymmetric Constraints

Liujie Hua [1], Yueyi Luo [2], Qianqian Qi [3] and Jun Long [3,*]

[1] School of Computer Science and Engineering, Central South University, Changsha 410083, China; liujiehua@csu.edu.cn

[2] School of Mathematics and Statistics, Central South University, Changsha 410083, China; luoyueyi@csu.edu.cn

[3] Big Data Institute, Central South University, Changsha 410083, China; qiqianqian@csu.edu.cn

* Correspondence: junlong@csu.edu.cn

Abstract: Medical data have unique specificity and professionalism, requiring substantial domain expertise for their annotation. Precise data annotation is essential for anomaly-detection tasks, making the training process complex. Domain generalization (DG) is an important approach to enhancing medical image anomaly detection (AD). This paper introduces a novel multimodal anomaly-detection framework called MedicalCLIP. MedicalCLIP utilizes multimodal data in anomaly-detection tasks and establishes irregular constraints within modalities for images and text. The key to MedicalCLIP lies in learning intramodal detailed representations, which are combined with text semantic-guided cross-modal contrastive learning, allowing the model to focus on semantic information while capturing more detailed information, thus achieving more fine-grained anomaly detection. MedicalCLIP relies on GPT prompts to generate text, reducing the demand for professional descriptions of medical data. Text construction for medical data helps to improve the generalization ability of multimodal models for anomaly-detection tasks. Additionally, during the text–image contrast-enhancement process, the model's ability to select and extract information from image data is improved. Through hierarchical contrastive loss, fine-grained representations are achieved in the image-representation process. MedicalCLIP has been validated on various medical datasets, showing commendable domain generalization performance in medical-data anomaly detection. Improvements were observed in both anomaly classification and segmentation metrics. In the anomaly classification (AC) task involving brain data, the method demonstrated a 2.81 enhancement in performance over the best existing approach.

Keywords: anomaly detection; multimodal contrastive learning; domain generalization; GPT

check for updates

Citation: Hua, L.; Luo, Y.; Qi, Q.; Long, J. MedicalCLIP: Anomaly-Detection Domain Generalization with Asymmetric Constraints. *Biomolecules* **2024**, *14*, 590. https://doi.org/10.3390/biom14050590

Academic Editor: Mohsin Saleet Jafri

Received: 20 April 2024
Revised: 13 May 2024
Accepted: 14 May 2024
Published: 16 May 2024

1. Introduction

Anomaly detection is widely applied across various sectors including industrial production [1–3], finance, autonomous driving [4], and disease diagnosis [5–10]. In the medical field, anomaly detection can help reduce misdiagnoses and missed diagnoses caused by human error during manual inspections. Compared to the industrial sector, medical data requires a higher degree of specialization. The rarity and diversity of anomaly data make model construction in this context particularly challenging. Traditional methods, which rely on the completeness and availability of data [11,12], often struggle in this context. Constraints related to data privacy and the scarcity of anomaly data further complicate the direct training of anomaly-detection models.

Existing anomaly-detection methods typically train specifically on certain datasets, requiring the construction of multiple models and extensive training to adapt to different application scenarios [13]. This limits the performance of supervised methods and has become a significant bottleneck in medical-data anomaly detection [14–17]. Additionally, in multi-category anomaly-detection tasks, multiple models necessitate substantial computational power and storage resources. Accordingly, the pursuit of a unified generalization model for anomaly detection, applicable across diverse data categories, has become a pivotal area of research. Domain-generalization methods offer new solutions to tackle these challenges, enabling models to perform effectively across various unknown environments by leveraging generalized features that are not tied to the specifics of any single dataset.

Domain generalization [5,18–20] aims to build models for data from unknown domains, which addresses challenges such as data scarcity and inaccessibility in new fields. Given the limitations of available data and the diversity of anomalies, enhancing the domain generalization of models for such tasks is crucial. However, current research in image-based anomaly detection tends to focus excessively on single-modal, domain-specific data, neglecting broader, domain-independent representations. Traditional anomaly detection typically involves only image modality data. Relying exclusively on image data's feature distribution for constructing classification representations, single-modal approaches significantly limit the model's ability to generalize across different domains [21]. Various methods such as distribution-based representation [11], distance optimization [22,23], and adversarial generation [24–26] are employed. However, domain-generalization representations based on single-modal data lack diversity.

Single-type, single-modal anomaly detection relies heavily on the construction of classifiers and the distribution of normal samples, leading to models that are significantly tailored to specific datasets. Significant variations in data distributions across different categories present substantial challenges for domain generalization in models reliant on single-modal data. The development of multimodal contrastive models offers a new research direction for domain generalization in the medical-data anomaly-detection field [14]. Multimodal approaches leverage the strengths of multiple types of data inputs, such as combining image and text data, to enhance the robustness and generalization of anomaly-detection systems across different domains. This integration not only expands the representational diversity but also improves the adaptability of models to new, unseen datasets, overcoming the limitations associated with traditional, single-modal anomaly-detection methods [27].

Medical anomaly detection is a complex and resource-intensive task. It relies heavily on professionals to meticulously annotate features in datasets, a critical step that ensures accuracy during the training process [28]. Establishing orthogonal domain spaces and distributions helps to create clear representational boundaries, which are essential for effective anomaly classification. However, while the construction of medical image anomaly-detection datasets has demonstrated improved model performance due to its orthogonality, this characteristic also poses challenges to the model's generalization ability [29]. Moreover, the distribution of source-data representations is closely related to the representation space of specific domains. Establishing orthogonal domain spaces and distributions helps to create clear representational boundaries, which are essential for effective anomaly classification.

Traditional anomaly-detection models, including self-supervised [6,22,30] and generative models [25], are capable of learning image-representation distributions and normal representations from extensive medical image datasets. However, changes in the detection data lead to corresponding shifts in the distributions of data presentations and normal representations. Relying solely on single-modal data representation makes it difficult to achieve a rich information representation for domain generalization and to distinguish between normal and anomalous conditions effectively.

Therefore, anomaly-detection methods enhance model domain-generalization capabilities by incorporating multimodal approaches. The integrated constraints between multimodal data enhance the model's ability to represent information from the data [31–33]. Image–text models enrich the model's capability for information representation by applying hierarchical representational constraints and building a more diversified information representation [34,35]. The inclusion of linguistic information allows a single description to correspond to multiple objects and categories, as shown in Table 1. For example, the representation of holes is more dispersed and abstract than image representations. The constraints it constructs are more broadly applicable, enhancing the model's ability to relax the representation of broad-domain data. Supervised methods often build orthogonal feature encoding, but this approach is not conducive to domain generalization [13]. Different objects have different characteristic representations and should be measured using different orthogonal representations. In the context of unsupervised feature encoding, the resulting vectors deviate from orthogonality, with distinct objects assuming unique positions that more accurately reflect real-world conditions. The advantage of using natural language for supervision is that it allows for more diverse expansions. Linking language representations with image representations allows for more flexible transformations.

Table 1. Compared to traditional anomaly-detection methods that use a single class and a single model, the method proposed in this paper has extremely strong domain-generalization capability. ✓ indicates possessing the corresponding capability.

Methods	Anomaly Score	Anomaly Segmentation	Model Unification	Domain Generalization
Traditional methods	✓	✓		
Few-shot methods	✓	✓		
Lvlms	✓	✓	✓	
Ours	✓	✓	✓	✓

The purpose of anomaly detection is to identify data that does not conform to the normal distribution [36,37]. Due to the scarcity of anomalous data, a common approach is to distinguish anomalies by learning only the feature distribution of normal data [11,22]. This requires the extracted data representations to be highly orthogonal, and the model is constructed using a single type of data. Image generation methods generate images of the normal distribution for specific categories of data and reconstruct feature distributions [25]. Common detection methods include classification-based [2,7] and generation-based methods [38]. Given the scarcity of anomalous data, learning the feature distribution of normal data is key to effective detection. Data representations are typically orthogonal and tailored to specific data categories [1,22]. On the other hand, image generation methods focus on reconstructing the distribution of normal samples for specific categories.

Domain-generalization techniques are employed to enhance anomaly detection in image-based texture and surface defect identification [31]. Large-language pre-trained models (such as VLP [39], ALIGN [40]) demonstrate great adaptability in feature consistency expression and model generalization capabilities, achieving cross-modal information interaction through global information expression and similarity comparison. Multimodal large models have shown good results in anomaly detection through contrastive learning [41], generative learning, and large model denoising methods. The multimodal invariant representation anomaly-detection method improves model generalization performance by learning domain-invariant representations [42]. For more granular detection, image patch and meta-learning-based methods are applied in anomaly detection.

Based on the issues discussed, MedicalCLIP explores a unified medical anomaly-detection model with strong generalization capabilities. By utilizing spatially consistent representations of multimodal data, the model achieves not only unification but also an enhancement of its generalization capabilities. Through image–text comparative learning, a category-independent model for domain generalization in anomaly detection is implemented [43]. Constructing a unified anomaly-detection model with robust domain-generalization capabilities is of significant practical value for improving model efficiency and achieving model generalization. Figure 1 demonstrates the classification capabilities of MedicalCLIP.

For anomaly-detection tasks, through the comprehensive comparative representations within and between modalities, we find that the intramodal representation constraints of different modal data have varying impacts. Common anomaly-detection methods involve contrastive learning among normal samples, but they often overlook the model limitations caused by spatial differences. A unified anomaly-detection model should be able to extract more applicable representational data, utilizing multimodal data integration and multitask loss for adaptive feature extraction. In the processing of single-image modal data, models primarily focus on the visual representation of the image, lacking guidance from other modal data, making it difficult to extract logical information from the image. Introducing textual descriptions, such as "This is a bottle with cracks", enables the accurate identification and extraction of key elements in the image, such as the number of bottles and defects, by leveraging semantic information. Multimodal integrated contrastive encoding, when compared to image data, enhances the textual representation's guidance on image representation. MedicalCLIP explores a medical-data anomaly-detection method with strong generalization capabilities that uses multi-angle, cross-modal reasoning. This prevents any single modality from dominating the entire model-training process, therefore learning more universal information representations and enhancing the associative capabilities between different modal data.

Movtivation: For medical-data anomaly-detection tasks, this paper proposes an asymmetric constraint MedicalCLIP domain-generalization method for anomaly detection and segmentation. By implementing intramodal constraints within image and text modalities, the consistency of image modality representations and the constraints within and between modalities are enhanced, thus balancing the model's domain generalization and detection effectiveness.

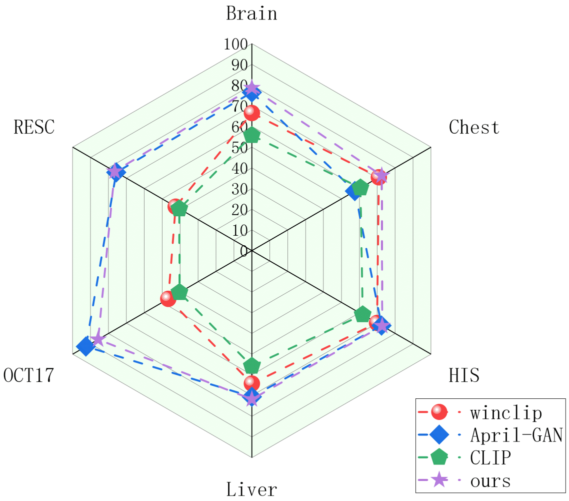


Figure 1. Domain-generalization model for zero-sample anomaly classification method for different data. Compared to the existing methods, our method shows competitive results.

2. Materials and Methods

The purpose of MedicalCLIP is to train a unified model capable of achieving anomaly detection through zero-shot learning. For the given training data $\mathcal{D}_{train} = \{\mathcal{I}_{train}, \mathcal{Y}_{train}\}$, the images $x^{\mathcal{I}} \in \mathcal{I}_{train}$ and label of images $\mathcal{Y}_{train} \in \{0,1\}$. The test dataset $\mathcal{D}_{test} = \{\mathcal{D}^1_{test}, \mathcal{D}^2_{test}, \ldots, \mathcal{D}^m_{test}\}$, m is the classes of test dataset, $\mathcal{D}_{train} \cap \mathcal{D}_{test} = \varnothing$. The model is trained by the given dataset $\mathcal{D}train$, resulting in superior anomaly-detection performance on the test set $\mathcal{D}test$. There are information differences between different categories of data, i.e., $\mathcal{D}^1_{test}, \mathcal{D}^2_{test}$, making it difficult to distinguish between different types of data in the feature space. We employ a multimodal approach to develop a unified model capable of handling multiclass anomaly detection. For the given image, we introduce textual $\mathcal{T}$ to guide the representation of the image data and conduct multimodal contrastive analysis based on textual information. The calculation method for anomalies is to compute the similarity between data of different modalities. We characterize the data using the image encoder $f_\phi(x^{\mathcal{I}})$ and the text encoder $g_\theta(x^{\mathcal{T}})$. The intermodal contrast losses are as follows:

$$\mathcal{L}(\mathcal{I}, \mathcal{T}) = min < f_\phi\left(x_i^{\mathcal{I}}\right), g_\theta\left(x_j^{\mathcal{T}}\right) > \quad (1)$$

where $\mathcal{L}_{CL}$ denotes the contrast loss, $f_\phi(x_i^{\mathcal{I}})$ and $g_\theta(x_j^I)$ denotes text and images of the same category, and$f_\phi(x_i^I)$, $g_\theta(x_j^I)$ denotes text and images of different categories.

$$\mathcal{L}_{cl,\mathcal{I}\to\mathcal{T}} = \frac{-1}{n}\sum_{i=1}^{n}\log\frac{\exp(\langle f_\phi(x_i^{\mathcal{I}}), g_\theta(x_i^{\mathcal{T}})/\tau)}{\sum_{j\in[n]}\exp(\langle f_\phi(x_i^{\mathcal{I}}), g_\theta\left(x_j^{\mathcal{T}}\right)/\tau)} \tag{2}$$

The purpose of MedicalCLIP is to establish fine-grained constraints within modalities for comprehensive representation and to achieve more fine-grained information filtering using comprehensive constraints $\mathcal{L}$ between image-image $f_\phi(x^{\mathcal{I}})$ and text-text $g_\theta(x^{\mathcal{T}})$. The CLIP model possesses powerful feature extraction capabilities by comparing text and image contrast representations. MedicalCLIP optimizes the fine-tuning of the training data through contrast embedding between image modalities and textual modalities and guides the image learning anomaly-detection data representation by adaptively generating textual cues.

2.1. Overview of Framework

As illustrated in Figure 2, the framework of MedicalCLIP is structured into four primary components: 1. Promote Embedding. GPT [44] is utilized to generate textual corpora $\mathcal{T}$, and leveraging the Contrastive Language-Image Pre-training (CLIP) [45] template, we craft both standard and anomalous textual descriptions tailored for diverse image categories. 2. Hierarchical image representation Constructing an image–text corpus $\mathcal{X} = \{x_i^{\mathcal{T}}, x_i^{\mathcal{I}}\}_{i=1}^{n}$ from the generated text, we construct comprehensive comparison methods based on the same modality and different modalities. 3. In-Modal Learning By integrating an asymmetric image–text constraint $\mathcal{L}$, we bolster the synergy between modalities, ensuring the model offers a harmonized representation of anomalous data. Image $\mathcal{T}$ and textual $\mathcal{I}$ intermodal comparison module, which deepens cross-modal understanding by comparing feature differences between text and images; Text and image intramodal comparison module, which focuses on feature comparisons within their respective modalities to improve the model's detailed representation of the data, as is shown in Figure 3.

Multimodal comparative learning The vision–language model (VLM), which aims to maximize consistency between $x^{\mathcal{T}}$ and $x^{\mathcal{I}}$, has a limited ability to characterize details. For AD tasks, anomalies are often not obvious, and therefore, more detailed image representations need to be extracted. Relying on intermodal contrast constraints alone is not sufficiently capable of characterizing the lower-order information of an image. We propose a novel irregular multimodal constraint(IRC) technique to improve the model's understanding of the distribution of image and text modalities through intramodal data constraints.

$$\mathcal{L}_{IRC} =< f_\phi(x_i^{\mathcal{I}}, f_\phi(x_j^{\mathcal{I}}) > -\lambda < f_\phi(x_i^{\mathcal{T}}, f_\phi(x_j^{\mathcal{T}}) >, \tag{3}$$

λ is the asymmetry factor. The representation of multimodal semantic information can bolster the invariance of features during the domain-generalization process. Within multimodal information, the capability of text $f_\phi(x^I)$ to represent data information is enhanced through adaptive text generation. Given the limited capacity of intermodal contrastive representations for detailed information, we employ irregular constraints to improve the model's ability to represent data across different modalities. Moreover, in the process of anomaly segmentation, models constrained by local intermodal interactions can focus more on local detail information. For these irregular constraints, we consider two types of comprehensive constraints across different modalities.

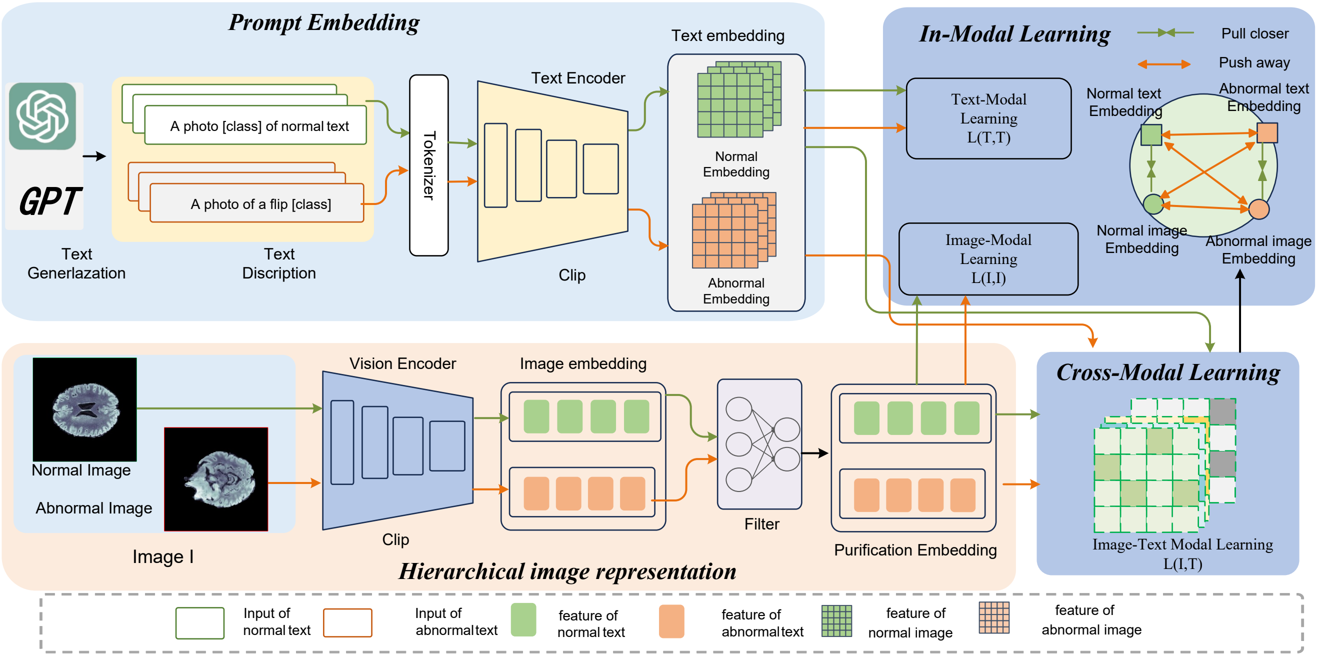


Figure 2. Overall framework of MedicalCLIP. The prompt embedding section includes prompt generation and text feature extraction. The hierarchical image-representation section preserves features at different levels of the image feature extraction representation and refines the image representation through a filtering module. The asymmetric constraint module contains cross-modal constraints and modal content.

2.2. Promote Embedding

In multimodal contrastive models, textual information $\mathcal{T}$ serves as an anchoring guide for image representation $\mathcal{I}$. We achieve a comprehensive representation of images $f_\phi(x^{\mathcal{I}})$ through adaptive generation, facilitating a more thorough information portrayal. During the text-generation process, we construct two textual generation strategies: specific image descriptions spd and category-agnostic descriptions cad. For specific image descriptions, text is generated based on the acquired image category information **[cls]**, containing more detailed image-specific details. In contrast, category-agnostic representations lean towards general textual descriptions unrelated to specific categories. Using text as an anchor point for contrastive optimization allows for the acquisition of more generalized, domain-invariant information representations.

$$\mathcal{T} = \mathcal{T}(spd) + \mathcal{T}(cad) \tag{4}$$

Prompt information generation The foundational corpus of the CLIP model was specifically designed for classification tasks, encompassing both the template $M = m_1, \ldots, m_n$ and the state $D = d1, d2, \ldots, d_l$. Recognizing the multifaceted nature of anomalies, we aimed to curate a corpus tailored for anomaly detection. For a given image x_i, we first obtain the category information of the image and combine it with the template information for image description generation. Subsequently, leveraging an automated prompt generation strategy, we sculpt a corpus apt for anomaly detection. The crafted template for anomaly-detection resonates with the structure "`A photo of a state object`", exemplified by "`A photo of a healthy brain`". For the descriptor ensemble $D = d1, \ldots di$, the embedding is achieved via the prompt template. The illustration delineates both the normal and anomalous data, showcasing their respective textual representations.

spd: A [domain] photo of a [state] [class] .

For anomaly detection in medical imaging, each characterization is multifaceted, with each facet boasting an array of templates and descriptor terms D, such as ``‘‘normal’’``, ``‘‘enhanced contrast’’``, and ``‘‘intact structure’’`` for typical annotations. In the context of atypical scans, descriptors might encompass phrases like ``‘‘presence of lesions’’`` or ``‘‘indications of calcifications’’``. Harnessing the robust text-generation prowess of GPT, we are equipped to craft intricate textual categorizations for distinct diagnostic categories.

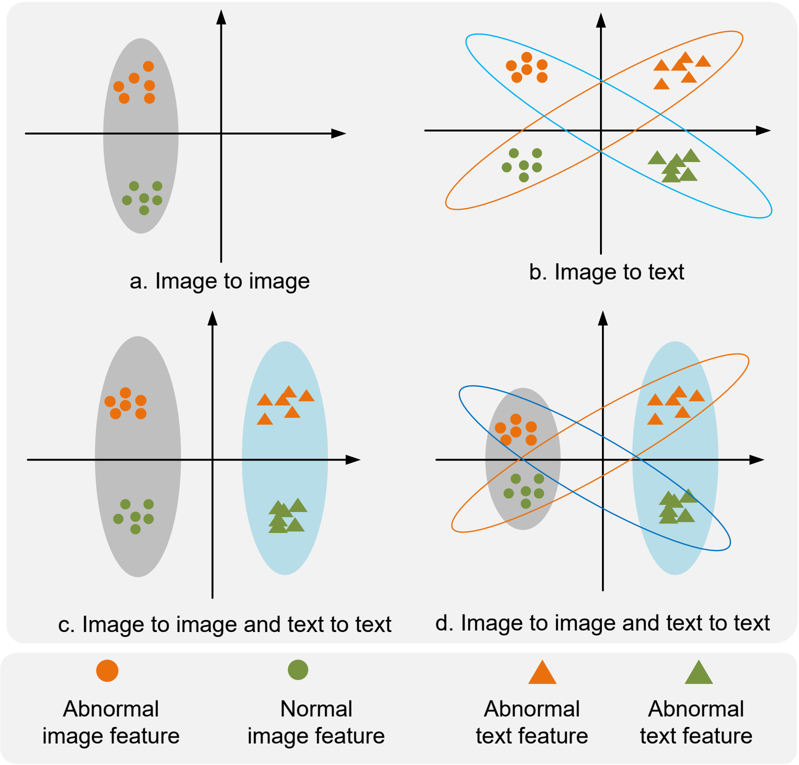


Figure 3. Irregular constraints. (**a**) represents the constraints for image category data. Crossing constraints between graphics are represented in (**b**). (**c**) represents constraints between image and image, text and text. (**d**) represents the multimodal irregular constraint method. The blue and orange circles indicate the sample constraints for different modes, respectively

cad: An [image] photo of a [state] .

$$x^{\mathcal{T}} = \sum_{i=1}^{N} FillTemplate\{m_i, d_i\} \tag{5}$$

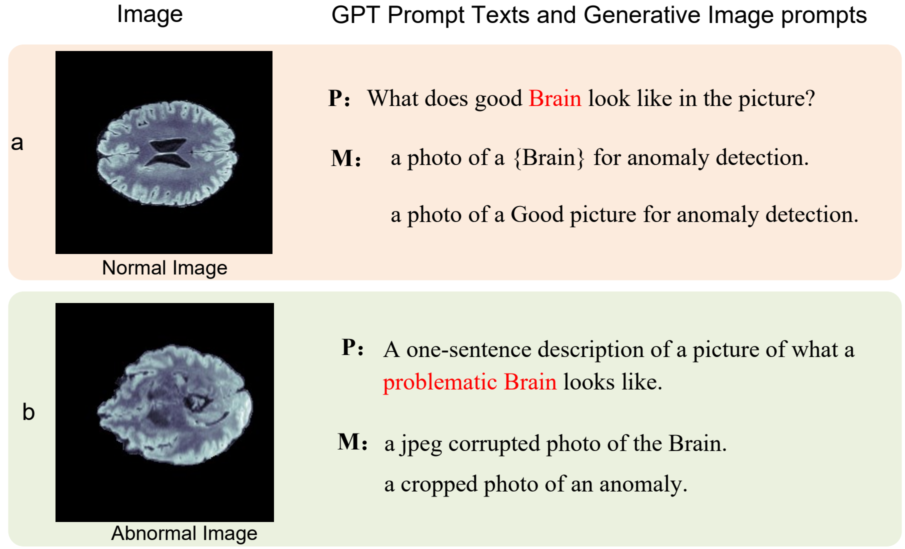
Merging templates with descriptor terms empowers us to formulate textual portrayals for images. When assimilating unknown data, pinpointing the image’s category and integrating it into the query template suffices for automated text generation. In juxtaposition with handcrafted corpora, this automated narrative is notably more exhaustive and intricate, particularly for datasets demanding niche expertise.

$$f_{normal}, f_{abnormal} = textencoder(prompt^{n}_{\text{normal}}, prompt^{m}_{\text{abnormal}}) \tag{6}$$

For the given text feature representation, it consists of a combination of normal and anomalous representations. An example of the generated text representation is shown in Figure 4.

$$f_{\phi} x^{\mathcal{I}} = \{f_{normal}, f_{abnormal}\} \tag{7}$$

$$\mathcal{T} = Generate\{GPT(class, model)\} \tag{8}$$



P: What does good Brain look like in the picture?

M: a photo of a {Brain} for anomaly detection.

a photo of a Good picture for anomaly detection.

P: A one-sentence description of a picture of what a problematic Brain looks like.

M: a jpeg corrupted photo of the Brain.

a cropped photo of an anomaly.

Figure 4. Image Text Generation. Leveraging textual cues to generate textual representations that conform to the template. (**a**) is a textual description of a normal image; (**b**) is a textual description of an abnormal image. **P** is the prompt message, and **M** denotes the text generated according to the template form.

2.3. Hierarchical Feature Representation

In our method, image anomaly classification is achieved through global representations, while anomaly segmentation is accomplished using local feature representations. By employing multi-scale image feature representations, more fine-grained feature extraction is achieved.

$$f_\phi x^{\mathcal{I}}, f_\phi x^{\mathcal{IP}} = imageencoder\{image, patch(image)\} \tag{9}$$

2.4. Image Feature Adaptation

The CLIP model, as a visual-language model, is primarily designed for classification tasks. Classification models, through training, make data of different categories cluster in the feature space, displaying clear, orthogonal boundaries. However, this method of data representation fundamentally differs from what is required in anomaly-detection tasks. In the context of anomaly detection, anomalies are typically sparser compared to normal data, leading to blurred boundaries and potentially non-orthogonal representations in the feature space. Confronting this challenge, we present a feature adapter, E_ψ, aimed at adjusting and aligning text and image features so that they can better adapt to the needs of anomaly-detection tasks.

$$o^{\mathcal{T}}, o^{\mathcal{I}} = E_\psi\{(g_\theta(x^{\mathcal{T}}), f_\phi(x^{\mathcal{I}})\} \tag{10}$$

In the field of anomaly detection, the problem of domain generalization focuses on detecting and locating anomalies within normal images and generalizing them to untrained target domains. This primarily addresses the challenge of limited training data during the production process. For given source domain data $I^{sl} = \{i^{sl}_m, y^{sl}_m\}_{m=1}^{N_{sl}}$, where $i_m \in I$, $y \in \{0,1\}$, and assuming all training data are normal, the objective is to train on this normal data to achieve anomaly detection, and then apply this detection capability to data in untrained target domains. The source domain data includes image data I and the generated textual data T. The aim is to use normal image data in conjunction with adaptively generated textual data. By contrasting the representations of text and image data, a comprehensive representation and analysis of the source domain data is achieved. Within the source domain, the generation of textual data adaptively incorporates domain expert knowledge, resulting in well-characterized textual data. By contrasting data intra-modally and inter-modally, detection efficacy is enhanced. During the domain-generalization process, bridging the gap between the representations of normal and anomalous data and enhancing adaptability to the source domain data is crucial.

The vision–language model demonstrates good performance in cross-modal contrastive learning. For the given image encoder f_i and text encoder f_t, given an image $x_I \in \mathcal{I}$ and text data $x_T \in \mathcal{T}$, the encoders $f_\phi(\boldsymbol{x}^I)$ and $g_\theta(\boldsymbol{x}^T)$ represent the image and text encoders, respectively.

Asymmetric Image-Text Constraints In anomaly detection, the generated textual information is categorized into two types: normal class descriptions and anomaly class descriptions. All images are represented as vectors of these two classes, and the textual information serves as an anchor point for calculating the loss between images and textual information. The specific process is shown in Algorithm 1.

Algorithm 1 Feature self augmentation process

```
# I Input Image
# T Input Text
# F ← Image_Encoder()
# T ← Feature_Extractor()
# A ← Adaptor()
# N ← Feature_Filter()
pretrain_init(F)
for each x in data_loader do
# Asymmetry constraint
# extract feature representations of different modes
    I_f = image_encoder(I)
    T_f = text_encoder(T)
# Loss function
    loss_cl = cross_entropy_loss (I_f, T_f)
    loss_IRC = cross_entropy_loss (I_f, I_f) - $\beta$ cross_entorpy_loss (T_f, T_f)
    loss = loss_(cl) +loss_(IRC)
    F ← F.detach()
    update(T, D)
end for
```

Contrastive Segmentation For image-segmentation tasks, the model needs to focus more on the fine-grained details of images. To enhance the model's attention to image details, local representations of the image are extracted and compared with textual representations. Furthermore, during the feature extraction process of images, features at different levels contain various types of image information. Therefore, this paper implements a hierarchical representation of images to achieve a diversified association between image and text. For the given anomaly-detection method, we propose an asymmetric clip constraint domain-generalization method for anomaly detection and segmentation, which is used to perform anomaly detection and segmentation. The asymmetric constraint improves the consistency of image modality representation and the constraints within and between modal representations, enhancing the model's balance between domain generalization and detection effectiveness.

In a manner similar to the feature similarity calculation method used in CLIP, we first obtain the hierarchical representation x of images and the representations t of various types of text. Then, we compare the similarity of intra-class features for each modal data. The input image data x^T, x^I is first constrained by the original CLIP model method, where the original model establishes the $x^T \odot x^I$ clip constraint by constraining the data between different modalities.

$$\mathcal{L}_{\text{C}-\text{clip}} = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{N}\Big(\langle x_j^I, x_i^T\rangle - \lambda\langle x_k^I, x_j^T\rangle\Big)^2. \tag{11}$$

Constraints between different modalities tend to focus more on the representation of semantic information. In contrast, modal content comparison constraints are more about the representation of fine-grained information. As for irregular constraints within modalities, they are as follows:

$$\mathcal{L}_{\text{I}-\text{clip}} = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{N}\Big(\langle x_j^I, x_k^I\rangle - \beta\langle x_k^T, x_j^T\rangle\Big)^2. \tag{12}$$

β is the irregular factor in irregular constraints. We use the input image and enhanced image to perform intermodal contrast constraints. The intermodal constraints are enforced in the form of element-wise dot products, x^T, x^I, and the intramodal constraints are enforced by associating various types of image representations with text representations, enhancing the model's understanding of a single modality. Beyond the constraints between images, there are also constraints between texts represented by $\mathcal{L}_{clip}$. Since image data tend to represent more detail-oriented features, while text data have a more nuanced understanding of higher-order information, the intermodal constraints for images and texts are looser for images and tighter for texts.

3. Results

Datasets To demonstrate the wide adaptability of the MedicalCLIP model in the field of anomaly detection, we conducted validations on datasets such as brain, chest, and liver, covering various areas including multiple types of data within the medical field. During the training process, the training data only contains normal data, while the testing data includes both normal data and annotated anomaly data for evaluating the model's performance. Moreover, through extensive ablation studies and comparative experiments across datasets from different domains, we further confirmed the generalization ability of our method in anomaly detection.

Metrics The performance of our model is evaluated on our medical public datasets. These datasets include brain and chest, and include normal and abnormal data, as well as annotated segmentation data. Figure 5 shows the sample image of the medical dataset. To measure the performance of the model in the process of anomaly classification (AC) and anomaly segmentation (AS), the anomaly-detection task is set to evaluate the model performance. The AUROC metrics for image anomaly detection and anomaly segmentation are used to evaluate the classification results. In anomaly detection and segmentation tasks, data categories are often imbalanced. Although anomaly-detection segmentation tasks involve pixel-level segmentation, the ultimate goal remains to distinguish between abnormal and normal areas. AUROC can evaluate the model's ability to detect anomalies at different thresholds, and it is not affected by the imbalance in data categories.

Implementation In the image and text feature extraction models, we employ the CLIP model pre-trained by OpenAI, and the text is automatically generated by leveraging the GPT-3.5 model for template construction. The pre-training model is ViT-L/14 [46] as the MedicalCLIP backbone, and the obtained image representations and text representations are compared. CLIP's original text features and image representations are trained for classification tasks. To align better with anomaly detection and segmentation tasks, this paper introduces an adaptive network layer for feature adaptation. The feature adaptive layer utilizes a shallow network to facilitate task adaptation for image and text representations.

In hierarchical feature representation, we use 6, 12, 18, 24 layers of representations for extraction. The software pytorch-2.1.1 used in the experiments was run using a single NVIDIA V100 32 GB GPU(The equipment was sourced from NVIDIA Corporation, located in Santa Clara, CA, USA), with an epoch setting of 50, a batch size of 16, and a learning rate of 0.0002. Throughout the model-training process, the irregular constraint β was set at 0.4

to obtain optimal results. During the training process, we adopted a multi-category unified training approach, inputting multiple different categories into the model for simultaneous training. Concurrently, while constructing a multi-category unified model, we aimed to establish an anomaly-detection model with strong domain-generalization capabilities.

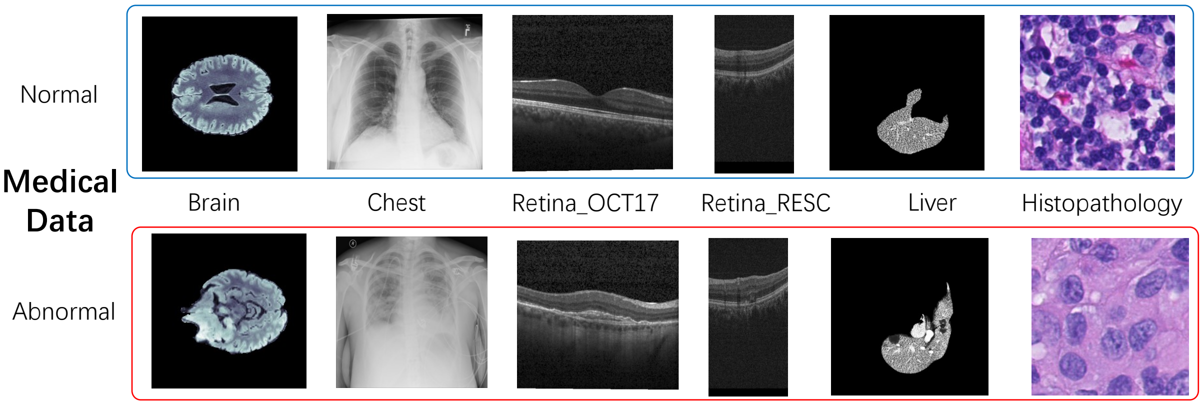


Figure 5. Partial samples from the Medical datasets are presented, where blue boxes indicate normal samples, while red boxes denote anomalous samples.

3.1. Domain Adaptation Anomaly Detection

To evaluate the generalization ability of the model for different classes of data, we validate it on 6 different medical data. The anomaly-detection results for different categories and methods are shown in Table 2. The different models show some domain-generalization ability in medical data. Among all the methods, MedicalCLIP shows better generalization ability and superior domain-generalization performance for data such as the brain and liver.

Table 2. In domain generalization, different methods are compared. The main criteria for evaluation are the AUC for anomaly classification and anomaly segmentation. Bold indicates the best result, and underline indicates the second-best result.

Metrics	**Methods**	**Brain**	**Chest**	**Histo Pathology**	**Liver**	**Retina OCT2017**	**Retina RESC**	**Metrics**	**Brain**	**Liver**	**Retina RESC**
AC AUROC	WinCLIP [41]	66.49	70.86	69.85	64.20	46.64	42.51	AS AUROC	85.99	**96.20**	80.56
	April-GAN [47]	76.43	57.49	72.36	70.57	**92.61**	75.67		91.79	97.05	85.23
	CLIP [45]	55.63	60.62	61.87	55.78	40.42	40.46		80.11	82.35	76.46
	+ Prompt ens. [45]	55.95	61.45	62.53	58.62	41.78	41.32		91.26	89.86	79.65
	CoOp [21]	73.26	65.83	71.09	65.89	68.93	66.54		90.53	88.56	77.85
	Ours	**78.61**	**72.51**	**72.73**	**71.79**	85.79	**76.54**		**92.67**	95.63	**86.33**

Analysis of results Traditional anomaly-detection evaluation metrics include two primary tasks: anomaly detection and anomaly segmentation. Intramodal constraints aim to achieve a more compact feature representation within the same type of modal data. In contrast to classification models, the distributions of normal and abnormal data are largely represented within a similar data space, with only local outliers deviating from the typical representations. By constructing intramodal data contrasts between images, such as the comparison of similarity between $x^{\mathcal{T}}$ and $x^{\mathcal{I}}$, the focus tends to be more on semantic representation, often overlooking details. In the course of generating text, establishing a more diverse linguistic representation enriches the types of data representation.

Table 2 indicates that our method demonstrates excellent anomaly detection and segmentation performance on most datasets. However, it does not perform optimally on the OCT17 dataset. From an overall perspective, the MedicalCLIP model exhibits relatively balanced generalization capabilities across different domains, but it falls short in higher-precision detection tasks, such as on the OCT17 dataset, indicating room for improvement.

For instance, on the OCT dataset, the April-GAN shows superior detection performance, yet its overall generalization ability remains limited.

Given the frequent changes in data distribution, anomaly detection becomes crucial for identifying distributional anomalies within normal data representations. The primary challenge lies in enhancing the model's sensitivity to shifts in data distribution while simultaneously improving its capacity to detect such changes within similar types of data. In the process of enforcing intermodal constraints, we enhance the model's capability for diversity transfer. Concurrently, applying intramodal constraints boosts the model's relevance transfer. The improvement in classification and segmentation results is shown in the Figure 6.

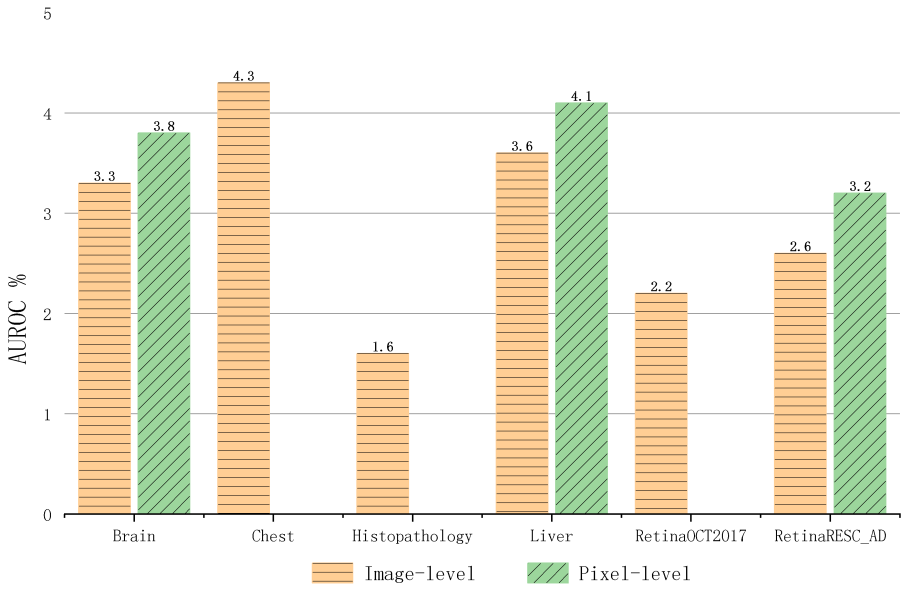
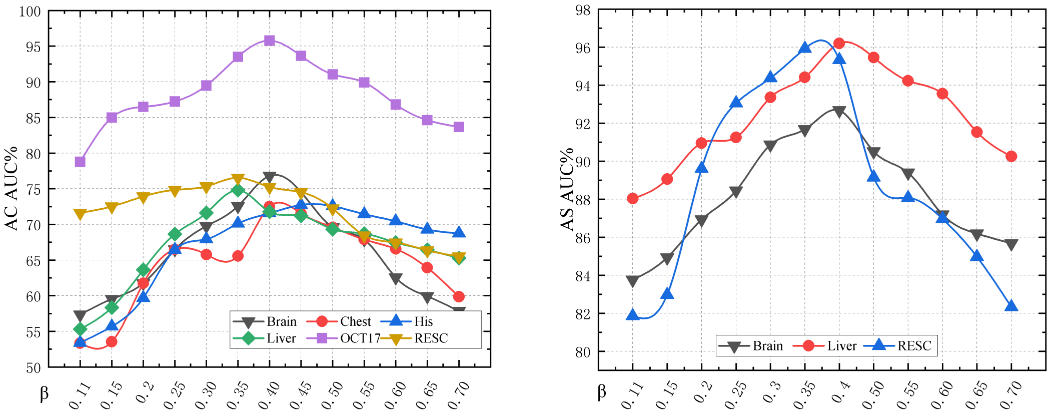


Figure 6. Introducing the irregular constraint, overall results demonstrate that the improvement in performance due to this constraint is more significant in segmentation tasks compared to classification tasks. The yellow color indicates an increase in the classification (AC) effect, and the green color indicates an increase in the segmentation (AS) effect.

3.2. Irregular Constraints

In the context of irregular constraints, different degrees of constraints have varying impacts on the effectiveness of detection. During the process of anomaly segmentation, the model pays more attention to the details of the image, thus imposing stronger constraints on the image.

The given text discusses the concept of constraint rate β in models, specifically in the context of handling modal data. It states that different constraint rates reflect the model's tolerance towards data within a particular mode. For image data, the constraint rate, represented by β, is crucial. The text finds that a constraint rate of 0.4 yields the best results for image tolerance in the model, as is shown in Figure 7. Furthermore, it suggests that constraining image data is particularly beneficial for enhancing the model's ability to represent image features effectively. This insight indicates that adjusting the constraint rate can significantly impact the model's performance, especially regarding image data.



(**a**) (**b**)

Figure 7. The figure shows that different constraint values have varying impacts on the results. Both anomaly classification (AC) and anomaly segmentation (AS) are similarly affected. (**a**) The figure demonstrates the impact of different constraint values on anomaly classification (AC) results. The optimal results are achieved when the constraint value is set to 0.4. (**b**) The results of anomaly segmentation (AS) are different for different constraint values. The results demonstrate the importance of irregular constraints for anomaly segmentation.

3.3. Text Prompt

MedicalCLIP explores the relationship between the domain-generalization capabilities of the model and the types of data categories. For models pertaining to different data categories, we examined the variation of the model in relation to changes in data categories. Within the brain dataset, we conducted separate validations to assess the impact of varying numbers of categories and different types of object categories on the model's performance.

To validate the impact of different text-generation methods on model performance, we conducted comparative tests for both category-dependent and category-independent text-generation. The results indicate that category-dependent textual descriptions enhance the model's ability to guide anomaly-detection tasks. Conversely, category-independent text formulations exert a greater influence on the model's domain-generalization capabilities.The results are shown in the Table 3.

Table 3. The impact of different types of text generation on model outcomes. Compared to category-specific generated text, category-independent text effectively enhances the model's generalization ability for anomaly-detection tasks. Overall comparisons show that text generation has a greater impact on classification tasks than on segmentation tasks. Red indicates × that the relevant description was not used, green indicates ✓ that it was used.

Textpormpt		**AC AUROC**						**AS AUROC**		
spd	**cad**	**Brain**	**Chest**	**Histo Pathology**	**Liver**	**Retina 2017**	**Retina RESC**	**Brain**	**Liver**	**Retina RESC**
✓	×	76.45	68.28	68.75	68.93	81.98	72.93	89.34	92.53	83.49
×	✓	77.56	69.93	69.91	71.36	83.47	74.84	91.45	93.56	85.28
✓	✓	78.61	72.51	72.73	71.79	85.79	76.54	92.67	95.63	86.33

Feature Adaptation The initial CLIP model was primarily designed for classification tasks and has limited adaptability to anomaly-detection tasks. This paper refines the feature selection and adaptation through fine-tuning of image representations. The feature representation adaptive module employs a shallow network architecture for fine-tuning. As illustrated in Figure 8, after processing through the feature adaptive layer, the model is capable of delineating clearer boundaries. The results are shown in Table 4.

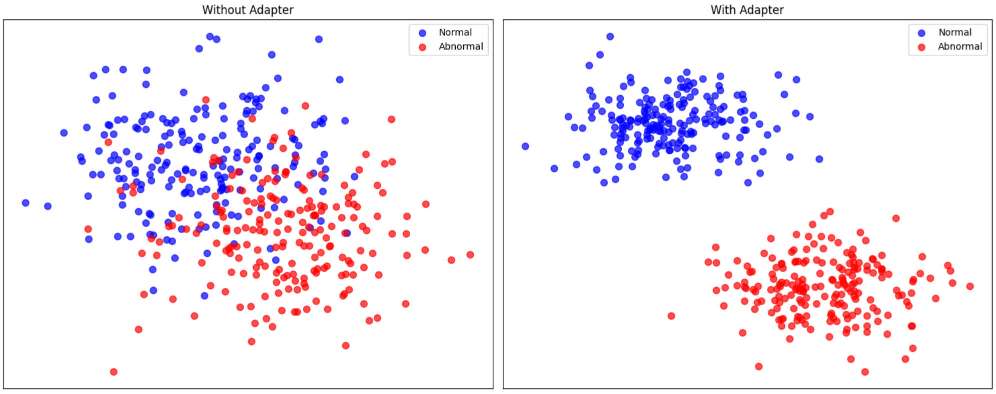


Figure 8. A fine-tuned adaptor layer allows filtering of features for anomaly-detection tasks. Filtering and constraints on image features can improve the model's ability to discriminate between representations.

Table 4. The effect of whether or not fine-tuning is performed on the model for which the image was acquired. The fine-tuned image representations are more suitable for anomaly-detection tasks. A more clearly differentiated representation of normal and abnormal can be created. AC denotes the anomaly classification indicator, and AS denotes the anomaly segmentation indicator.

Anomaly Classification	**Withadaptor**		**Noadaptor**	
Class	**AC**	**AS**	**AC**	**AS**
Brain	78.61	92.67	75.38	88.46
Chest	72.51	-	69.26	-
Histopathology	72.73	-	66.17	-
Liver	71.79	95.63	59.36	84.4
RetinaOCT2017	85.79	-	78.09	-
RetinaRESC	76.54	86.33	68.58	82.53

4. Discussion

Limited by the CLIP model's understanding of semantic objects, it demonstrates weaker performance in anomaly-detection tasks. Through the customization of text prompts for anomaly detection, WinCLIP shows improved results. WinCLIP boosts performance using customized text prompts that are manually set, with their effectiveness critically dependent on the thoroughness of their text prompt. The textual information representation learned by CoOp relies more on training data, which, to some extent, limits the model's generalization ability for unknown data. To make CLIP lean more towards semantic representation and enhance the model's performance in segmentation tasks, a hierarchical and image block form of information representation is adopted. By using automatically generated text prompts, the restrictions of text prompts on visual representation are reduced, therefore improving generalization capability. Furthermore, to capture the details of image data, this paper establishes an asymmetric constraint-based intermodal contrast method. MedicalCLIP can perform fine-grained anomaly segmentation on different types of medical data, showcasing its ability to handle detailed information. This provides a new perspective for the application of the CLIP model in medical image anomaly-detection tasks.

5. Conclusions

In conclusion, this paper introduces an image–text irregular constraint method applied to medical image anomaly detection to achieve the ability to generalize the anomaly detection of different categories of data. A more professional and comprehensive description of textual information is established by generating textual hints through GPT. In this paper, modal content constraints are applied to text and images, and hierarchical information representation of image information is used to achieve more fine-grained textual semantic guidance and obtain more detailed anomaly-detection information. Combined with the multimodal contrast learning strategy, the method can be flexibly generalized to different types of data. This method provides new research ideas for anomaly detection of different categories of data due to existing methods in domain-generalization anomaly detection.

Author Contributions: Conceptualization, L.H. and J.L.; Software, L.H. and Y.L.; Validation, Y.L.; Formal analysis, L.H. and Q.Q.; Data curation, L.H. and Q.Q.; Writing and editing, L.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Systemically significant projects of China National Railway Group Co., I LTD (P2023S006.) and the Systemically significant projects of China National Railway Group Co., I LTD (P2023W002.), National Natural Science Foundation of China under Grant 62302062, National 745 Natural Science Foundation of China (62202501, U2003208), the Hunan Provincial Natural Science Foundation of China under Grant Number 2024JJ5448 and High-Performance Computing Center of Central South University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Acknowledgments: High-Performance Computing Center of Central South University.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liu, Z.; Zhou, Y.; Xu, Y.; Wang, Z. Simplenet: A simple network for image anomaly detection and localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 20402–20411.
2. Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; Gehler, P. Towards total recall in industrial anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 14318–14328.
3. Jiang, H.; Dang, Z.; Wei, Z.; Xie, J.; Yang, J.; Salzmann, M. Robust Outlier Rejection for 3D Registration with Variational Bayes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 1148–1157.
4. Yao, Y.; Wang, X.; Xu, M.; Pu, Z.; Wang, Y.; Atkins, E.; Crandall, D.J. DoTA: Unsupervised detection of traffic anomaly in driving videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 444–459. [CrossRef] [PubMed]
5. Su, J.; Shen, H.; Peng, L.; Hu, D. Few-shot domain-adaptive anomaly detection for cross-site brain images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *46* , 1819–1935. [CrossRef] [PubMed]
6. Madan, N.; Ristea, N.C.; Ionescu, R.T.; Nasrollahi, K.; Khan, F.S.; Moeslund, T.B.; Shah, M. Self-supervised masked convolutional transformer block for anomaly detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *46*, 525–542. [CrossRef] [PubMed]
7. Xiang, T.; Zhang, Y.; Lu, Y.; Yuille, A.L.; Zhang, C.; Cai, W.; Zhou, Z. SQUID: Deep Feature In-Painting for Unsupervised Anomaly Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 23890–23901.
8. Liu, J.; Zhang, Y.; Chen, J.N.; Xiao, J.; Lu, Y.; A Landman, B.; Yuan, Y.; Yuille, A.; Tang, Y.; Zhou, Z. Clip-driven universal model for organ segmentation and tumor detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 21152–21164.
9. Aladhadh, S.; Almatroodi, S.A.; Habib, S.; Alabdulatif, A.; Khattak, S.U.; Islam, M. An Efficient Lightweight Hybrid Model with Attention Mechanism for Enhancer Sequence Recognition. *Biomolecules* **2022**, *13*, 70. [CrossRef] [PubMed]
10. Huang, C.; Jiang, A.; Feng, J.; Zhang, Y.; Wang, X.; Wang, Y. Adapting Visual-Language Models for Generalizable Anomaly Detection in Medical Images. *arXiv* **2024**, arXiv:2403.12570.
11. Defard, T.; Setkov, A.; Loesch, A.; Audigier, R. Padim: A patch distribution modeling framework for anomaly detection and localization. In Proceedings of the International Conference on Pattern Recognition, Kolkata, India, 15–18 December 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 475–489.
12. Yi, J.; Yoon, S. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In Proceedings of the Asian Conference on Computer Vision, Kyoto, Japan, 30 November–4 December 2020.
13. Huang, C.; Guan, H.; Jiang, A.; Zhang, Y.; Spratling, M.; Wang, Y.F. Registration based few-shot anomaly detection. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 303–319.

14. Fernando, T.; Gammulle, H.; Denman, S.; Sridharan, S.; Fookes, C. Deep learning for medical anomaly detection–a survey. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–37. [CrossRef]
15. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
16. Schlegl, T.; Seeböck, P.; Waldstein, S.M.; Schmidt-Erfurth, U.; Langs, G. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In Proceedings of the International Conference on Information Processing in Medical Imaging, Boone, NC, USA, 25–30 June 2017; Springer: Berlin/Heidelberg, Germany, 2017; pp. 146–157.
17. Wu, P.; Zhou, X.; Pang, G.; Zhou, L.; Yan, Q.; Wang, P.; Zhang, Y. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 6074–6082.
18. Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; Loy, C.C. Domain generalization: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 4396–4415. [CrossRef]
19. Wang, J.; Lan, C.; Liu, C.; Ouyang, Y.; Qin, T.; Lu, W.; Chen, Y.; Zeng, W.; Yu, P. Generalizing to unseen domains: A survey on domain generalization. *IEEE Trans. Knowl. Data Eng.* **2022**, *35*, 8052–8072. [CrossRef]
20. Guo, J.; Qi, L.; Shi, Y. Domaindrop: Suppressing domain-sensitive channels for domain generalization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 19114–19124.
21. Zhou, K.; Yang, J.; Loy, C.C.; Liu, Z. Learning to prompt for vision-language models. *Int. J. Comput. Vis.* **2022**, *130*, 2337–2348. [CrossRef]
22. Li, C.L.; Sohn, K.; Yoon, J.; Pfister, T. CutPaste: Self-Supervised Learning for Anomaly Detection and Localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9664–9674.
23. Bergman, L.; Cohen, N.; Hoshen, Y. Deep nearest neighbor anomaly detection. *arXiv* **2020**, arXiv:2002.10445.
24. Akçay, S.; Abarghouei, A.A.; Breckon, T.P. Skip-GANomaly: Skip Connected and Adversarially Trained Encoder-Decoder Anomaly Detection. In Proceedings of the International Joint Conference on Neural Networks, IJCNN 2019, Budapest, Hungary, 14–19 July 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–8.
25. Perera, P.; Nallapati, R.; Xiang, B. OCGAN: One-Class Novelty Detection Using GANs With Constrained Latent Representations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 15–20 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 2898–2906.
26. Zavrtanik, V.; Kristan, M.; Skočaj, D. DRAEM-A discriminatively trained reconstruction embedding for surface anomaly detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 8330–8339.
27. Qiu, S.; Ye, J.; Zhao, J.; He, L.; Liu, L.; Bicong, E.; Huang, X. Video anomaly detection guided by clustering learning. *Pattern Recognit.* **2024**, 110550. [CrossRef]
28. Sarhadi, V.K.; Armengol, G. Molecular biomarkers in cancer. *Biomolecules* **2022**, *12*, 1021. [CrossRef] [PubMed]
29. Zavrtanik, V.; Kristan, M.; Skočaj, D. Dsr–a dual subspace re-projection network for surface anomaly detection. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2020; Springer: Berlin/Heidelberg, Germany, 2022; pp. 539–554.
30. Bozorgtabar, B.; Mahapatra, D. Attention-conditioned augmentations for self-supervised anomaly detection and localization. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 14720–14728.
31. Chen, S.F.; Liu, Y.M.; Liu, C.C.; Chen, T.P.C.; Wang, Y.C.F. Domain-Generalized Textured Surface Anomaly Detection. In Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 18–22 July 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6.
32. Salehi, M.; Sadjadi, N.; Baselizadeh, S.; Rohban, M.H.; Rabiee, H.R. Multiresolution knowledge distillation for anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14902–14912.
33. Yang, T.; Huang, Y.; Xie, Y.; Liu, J.; Wang, S. MixOOD: Improving Out-of-distribution Detection with Enhanced Data Mixup. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 1–18. [CrossRef]
34. Li, Y.; Goodge, A.; Liu, F.; Foo, C.S. PromptAD: Zero-Shot Anomaly Detection Using Text Prompts. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 1093–1102.
35. Wu, P.; Liu, J.; He, X.; Peng, Y.; Wang, P.; Zhang, Y. Toward Video Anomaly Retrieval From Video Anomaly Detection: New Benchmarks and Model. *IEEE Trans. Image Process.* **2024**, *33*, 2213–2225. [CrossRef] [PubMed]
36. Pang, G.; Shen, C.; Cao, L.; van den Hengel, A. Deep Learning for Anomaly Detection: A Review. *ACM Comput. Surv.* **2021**, *54*, 1–38. [CrossRef]
37. Jiang, M.; Hou, C.; Zheng, A.; Hu, X.; Han, S.; Huang, H.; He, X.; Yu, P.S.; Zhao, Y. Weakly supervised anomaly detection: A survey. *arXiv* **2023**, arXiv:2302.04549.
38. Zhang, X.; Li, S.; Li, X.; Huang, P.; Shan, J.; Chen, T. DeSTSeg: Segmentation Guided Denoising Student-Teacher for Anomaly Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 3914–3923.

39. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the International Conference on Machine Learning, PMLR, Honolulu, HI, USA, 23–29 July 2023; pp. 19730–19742.
40. Song, F.; Yu, B.; Li, M.; Yu, H.; Huang, F.; Li, Y.; Wang, H. Preference ranking optimization for human alignment. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 18990–18998.
41. Jeong, J.; Zou, Y.; Kim, T.; Zhang, D.; Ravichandran, A.; Dabeer, O. Winclip: Zero-/few-shot anomaly classification and segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 19606–19616.
42. Yang, Z.; Soltani, I.; Darve, E. Anomaly detection with domain adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 2957–2966.
43. Yao, X.; Bai, Y.; Zhang, X.; Zhang, Y.; Sun, Q.; Chen, R.; Li, R.; Yu, B. Pcl: Proxy-based contrastive learning for domain generalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 7097–7107.
44. Biswas, S.S. Role of chat gpt in public health. *Ann. Biomed. Eng.* **2023**, *51*, 868–869. [CrossRef] [PubMed]
45. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 8748–8763.
46. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
47. Chen, X.; Han, Y.; Zhang, J. A Zero-/Few-Shot Anomaly Classification and Segmentation Method for CVPR 2023 VAND Workshop Challenge Tracks 1&2: 1st Place on Zero-shot AD and 4th Place on Few-shot AD. *arXiv* **2023**, arXiv:2305.17382.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.