# T3D: Advancing 3D Medical Vision-Language Pre-training by Learning Multi-View Visual Consistency

**Anonymous ACL submission**

## Abstract

While 3D visual self-supervised learning (vSSL) shows promising results in capturing visual representations, it overlooks the clinical knowledge from radiology reports. Meanwhile, 3D medical vision-language pre-training (MedVLP) remains underexplored due to the lack of a large-scale, publicly available 3D medical image-report dataset. To bridge this gap, we introduce **CT-3DVLP**, the first and largest **public** 3D volume-report dataset, establishing a comprehensive benchmark for 3D MedVLP research. Meanwhile, we propose the **T3D** framework, which enhances 3D MedVLP beyond naive CLIP-style alignment that directly pairs volumes with reports but neglects local visual representations. Instead, we introduce **Text-informed Multi-view Alignment (TMA)**, a novel approach that clusters volumetric data while enforcing consistency across different views of the same volume-report pair. TMA integrates textual features into fine-grained visual representations, ensuring contextual coherence across views. We evaluate T3D across multiple downstream tasks in both unimodal and cross-modal settings, including zero-shot and fine-tuned classification, cross-modal retrieval, report generation, and semantic segmentation. Our results show that T3D consistently outperforms existing vSSL and multimodal methods, demonstrating superior zero-shot and fine-tuning capabilities and setting a new benchmark for 3D medical image understanding[1].

## 1 Introduction

Deep learning (DL) has transformed 3D medical image analysis, improving diagnostic accuracy and efficiency. However, supervised DL methods require extensive, high-quality annotations, which are costly and time-consuming. To reduce this dependency, visual self-supervised learning (vSSL) has shown great potential in leveraging large-scale unlabeled medical data. Existing vSSL techniques, including image restoration (IR) and contrastive learning (CL) (Chaitanya et al., 2020; Taleb et al., 2020; Xie et al., 2022a; Haghighi et al., 2022), have demonstrated effectiveness in learning visual representations.

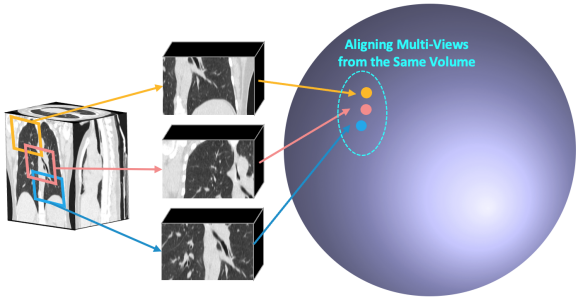


Figure 1: Illustration of the Text-Informed Multi-View Alignment (TMA) method. Multiple local views $V_i^m$ are generated from the same 3D volume, and their embeddings are aligned in the latent space to encourage consistency across views from the same volume-report pair. Each view's embedding is refined by the corresponding report to ensure consistency among all views from the same volume. The details are illustrated in Section 3.3.

IR-based methods reconstruct images from corrupted versions (Vincent et al., 2010; Pathak et al., 2016; Chen et al., 2020; He et al., 2022; Wei et al., 2022; Xie et al., 2022b; Gidaris et al., 2018) but primarily capture low-level features, often overlooking high-level semantics crucial for tasks like disease classification and tumor segmentation (Liu et al., 2023c,b; He et al., 2022). Meanwhile, CL-based vSSL methods (Tang et al., 2022; Zhou et al., 2023; Goncharov et al., 2023) enforce feature similarity between patches from the same image while treating patches from different images as negatives. However, these approaches risk semantic misalignment, as positive pairs may originate from anatomically distinct regions, while negative pairs might share similar structures (Jiang et al., 2023), leading

[1]All data and code will be released upon acceptance.

to suboptimal feature learning.

Medical Vision-Language Pre-training (MedVLP) has emerged as a promising approach to enhance representation learning by aligning medical images with radiology reports, providing clinically relevant supervision and improving feature informativeness in 2D medical imaging tasks (Radford et al., 2021; Liu et al., 2023a; Wang et al., 2022; Wan et al., 2023). However, its application to 3D medical images remains underexplored due to dataset scarcity and the lack of large-scale public benchmarks. Moreover, naive CLIP-style alignment relies solely on language supervision at the whole-volume level, limiting its ability to capture fine-grained 3D visual features.

To address these challenges, we propose **T3D**, a framework designed to enhance 3D MedVLP. Our key contributions include:

- We propose **T3D**, which integrates Global Cross-modal Alignment (GCA) and **Text-informed Multi-view Alignment (TMA)**, a novel mechanism that refines visual representations by leveraging text-informed guidance to enforce consistency across different views while capturing fine-grained visual features.
- To train T3D, we curate **CT-3DVLP**, the first large-scale public dataset including 52,639 paired CT volumes and radiology reports, establishing a comprehensive benchmark for 3D MedVLP research.
- Benefiting from the novel alignment, T3D demonstrates superior performance across various downstream tasks in both unimodal and cross-modal settings, including zero-shot and fine-tuned classification, retrieval, report generation, and segmentation.

## 2 Related Work

**VLP for 2D Medical Images** VLP has been extensively explored for 2D medical imaging to bridge visual and textual modalities. Early works such as ConVIRT (Zhang et al., 2020b) introduced global image-text alignment, later refined by GLoRIA and MGCA (Huang et al., 2021; Wang et al., 2022), which incorporated local alignment for better cross-modal representation learning. Other methods, including Med-UniC (Wan et al., 2023), mitigated language biases, while MedKLIP (Wu et al., 2023a) and KAD (Zhang et al., 2023) leveraged domain-specific knowledge. Additionally, reconstruction-based approaches like MRM (Zhou et al.) and PRIOR (Cheng et al., 2023) utilized image-text token prediction tasks, further improved by (Huang et al., 2023b) through adaptive token weighting. Despite these advancements, 2D VLP methods do not directly transfer to 3D imaging. The volumetric nature of 3D data introduces challenges in aligning 3D scans with textual reports due to high computational costs. While patch-based methods (Tang et al., 2022; Goncharov et al., 2023; Jiang et al., 2023) attempt to retain local information, they often lead to misalignment between cropped sub-volumes and full medical reports. These limitations highlight the need for specialized VLP approaches tailored for 3D medical imaging.

**VLP for 3D Medical Images** While VLP has advanced general 3D vision (Xue et al., 2023a; Zeng et al., 2023; Xue et al., 2023b; Chen et al., 2023b), these methods focus on sparse 3D point clouds and are not directly applicable to dense medical volumes like CT scans. Early 3D MedVLP approaches (Chen et al., 2023c,a) attempted to align full medical reports with cropped sub-volumes, introducing misalignment biases. To address this issue, (Wu et al., 2023b; Lei et al., 2023) proposed downsampling high-resolution volumes for report alignment, but this leads to a loss of anatomical details crucial for segmentation and diagnosis.

Additionally, most 3D MedVLP works rely on private datasets (Cao et al., 2024; Shui et al., 2025), limiting reproducibility. They also heavily depend on external annotation tools, such as segmenting each anatomical region and categorically labeling volumes, which introduces additional annotation costs and potential inconsistencies. These limitations underscore the need for a publicly available dataset to advance open 3D MedVLP research.

**vSSL for 3D Medical Imaging** vSSL has been widely explored in 3D medical imaging, with image restoration (IR) and contrastive learning (CL) as dominant strategies. IR-based methods reconstruct corrupted images (He et al., 2022; Wei et al., 2022; Xie et al., 2022b) but primarily capture low-level features, often overlooking high-level semantics crucial for diagnosis (Liu et al., 2023c,b; He et al., 2022). While recent works (Chen et al., 2019; Zhou et al., 2019; Wu et al., 2024b) incorporated anatomical priors, comprehensive semantic understanding remains underexplored. CL-based methods (Chaitanya et al., 2020; Taleb et al., 2020; Xie et al., 2022a) enforce similarity between patches from the same image while treating patches from

different images as negatives. However, positive pairs may originate from distinct anatomical regions, leading to semantic misalignment (Jiang et al., 2023). vox2vec (Goncharov et al., 2023) introduced voxel-level alignment, but treating neighboring voxels as negatives introduces bias. Despite advancements, existing vSSL methods lack clinical knowledge integration, limiting their effectiveness in real-world medical tasks.

# 3 Method

## 3.1 Extracting Visual and Text Features

Let $\mathcal{D} = \{(V_i, R_i)\}_{i=1}^{N}$ be a dataset of $N$ samples, where each 3D medical image

$$V_i \in \mathbb{R}^{1\times H\times W\times S}$$

(with height $H$, width $W$, and slices $S$) is paired with a radiology report $R_i$. We define a 3D visual encoder $f_\theta(\cdot)$ (e.g., 3D ResNet-50) that maps an input volume $V_i$ to a latent feature map $F_i$:

$$F_i = f_\theta(V_i) \in \mathbb{R}^{d_f\times h\times w\times s}$$

where $d_f$ is the number of output channels, and $h, w, s$ are spatial dimensions. A global 1D embedding is obtained via average pooling over all spatial dimensions, followed by a learnable linear projection:

$$\mathbf{z}_i^v = P^v\big(\mathrm{AvgPool}(F_i)\big) \in \mathbb{R}^{768}.$$

For text, we use Med-CPT (Jin et al., 2023) as the text encoder $g_\phi(\cdot)$ to obtain token embeddings from the radiology report $R_i$:

$$\mathbf{T}_i = g_\phi(R_i) \in \mathbb{R}^{L_r\times d_r},$$

where $L_r$ is the token length and $d_r$ is the embedding dimension. We extract the [CLS] token embedding and project it into the shared space:

$$\mathbf{z}_i^r = P^r\big(\mathbf{t}_i^{[\mathrm{CLS}]}\big) \in \mathbb{R}^{768}, \quad \text{where} \quad \mathbf{t}_i^{[\mathrm{CLS}]} \in \mathbb{R}^{d_r}.$$

## 3.2 Global Cross-Modal Alignment

To learn the global cross-modal representation, we align 3D volumes with their corresponding reports using a CLIP loss, as shown in the left part of Figure 2. Given a batch of $B$ samples, the loss $\mathcal{L}_{\mathrm{GCA}}$ is defined as:

$$\mathcal{L}_{\mathrm{GCA}} = -\sum_{i=1}^{B} \log \frac{\exp\big(\mathrm{sim}(\mathbf{z}_i^v, \mathbf{z}_i^r)/\tau\big)}{\sum_{j=1}^{B}\exp\big(\mathrm{sim}(\mathbf{z}_i^v, \mathbf{z}_j^r)/\tau\big)},$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes the dot product similarity, and we set $\tau = 0.07$ following (Radford et al., 2021).

## 3.3 Text-Informed Multi-View Alignment

**Motivation.** While the CLIP loss aligns 3D volumes with their corresponding radiology reports at a global level, it fails to capture fine-grained visual features crucial for understanding 3D medical imaging. To address this, we propose a text-informed multi-view alignment scheme that encourages consistency across multiple local 3D sub-volumes paired with the same report, as illustrated in Figure 1.

**Generating Local Views.** To enable the learning of fine-grained visual features, we generate multiple 3D local views from each $V_i$ by randomly cropping[2] sub-volumes of size $128 \times 128 \times 64$:

$$\{V_i^m\}_{m=1}^{M} = \mathrm{RandomCrop}\big(V_i, 128\times 128\times 64\big),$$

$$V_i^m \in \mathbb{R}^{1\times 128\times 128\times 64}.$$

Passing each $V_i^m$ through the 3D encoder $f_\theta(\cdot)$ yields the corresponding feature map:

$$F_i^m = f_\theta\big(V_i^m\big) \in \mathbb{R}^{d_f\times h'\times w'\times s'},$$
$$h' < h, \quad w' < w, \quad s' < s.$$

**Text-Informed Local Feature Enhancement.** To mitigate biases in local view alignment, we incorporate text-informed features into each local 3D view. Treating views from different volumes as negatives can be problematic, as they may contain similar anatomical regions. Similarly, views from the same volume may originate from distinct regions, making naïve positive pairing unreliable.

Since each volume is paired with a unique radiology report, we leverage textual information to refine local visual representations, as shown in the right part of Figure 2. This integration ensures that semantically similar regions across different volumes are not misclassified as negatives while refining positive associations within the same volume. By conditioning local views on text, we reduce bias in positive and negative pair selection before alignment.

We extract text token embeddings $\mathbf{T}_i$ from the text encoder $g_\phi(\cdot)$ and reshape the local 3D feature

[2]We implement the random cropping using the MONAI package https://docs.monai.io/en/stable/transforms.html

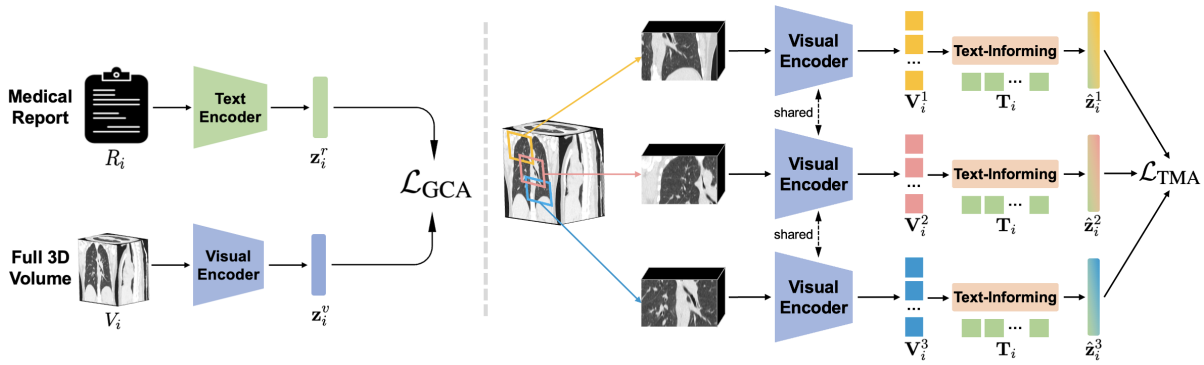


Figure 2: The T3D framework for learning multi-level 3D visual representations from corresponding medical reports. **Left:** To learn global cross-modal representations, we align the full 3D volume $V_i$ with its corresponding medical report $R_i$ using the loss function $\mathcal{L}_{\text{GCA}}$. The output embeddings $\mathbf{z}_i^v$ and $\mathbf{z}_i^r$ are optimized to encourage the matching of paired visual and textual features. **Right:** To further capture fine-grained visual representations, we first generate $M$ local views $V_i^m$ from the same volume using random cropping. The same visual encoder, as used in the GCA framework, is applied to obtain the embeddings for these local views. We then refine these embeddings using the report embedding $\mathbf{T}_i$, encouraging the local views from the same volume-report pair to become more similar in the latent space by minimizing the loss $\mathcal{L}_{\text{TMA}}$.

map $F_i^m$ into a sequence:

$$\mathbf{V}_i^m = \text{Reshape}(F_i^m) \in \mathbb{R}^{L_v \times d_f},$$

$$L_v = h' \times w' \times s'.$$

A single-layer Transformer block $\mathcal{F}_\psi(\cdot)$ refines $\mathbf{V}_i^m$ using $\mathbf{T}_i$ as keys and values, followed by average pooling to obtain the text-informed local embedding:

$$\hat{\mathbf{z}}_i^m = \text{AvgPool}\big(\mathcal{F}_\psi(\mathbf{V}_i^m, \mathbf{T}_i)\big) \in \mathbb{R}^{d_f}.$$

**Multi-View Alignment.** Since multiple local views $\{V_i^m\}_{m=1}^M$ are generated per volume-report pair, a naïve contrastive loss is unsuitable as it assumes one-to-one positive pairings. Instead, we assign each local view to one of $B$ cluster labels, where $B$ is the batch size, and each cluster label corresponds to a specific volume-report pair in the batch. A linear projection layer is used to predict the cluster assignment probability:

$$\mathcal{L}_{\text{TMA}} = -\sum_{i=1}^{B}\sum_{m=1}^{M} \log \frac{\exp\big(f(\hat{\mathbf{z}}_i^m)_{c_i}/\tau\big)}{\sum_{j'} \exp\big(f(\hat{\mathbf{z}}_i^m)_{j'}/\tau\big)}.$$

where $f(\cdot)$ is a linear projection function, and $c_i$ is the assigned cluster label for the corresponding volume-report pair. This objective encourages views from the same pair to cluster together while distinguishing them from those of different pairs.

### 3.4 Overall Objective

The final optimization objective of T3D aims to learn both global and local representations through:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GCA}} + \mathcal{L}_{\text{TMA}}.$$

## 4 Experiments

**Pre-training Dataset.** To construct the largest publicly available dataset for 3D MedVLP, we curate data from three public resources: BIMCV-R (Chen et al., 2024), CT-RATE (Hamamci et al., 2024a), and INSPECT (Huang et al., 2023a). We include all available samples from these datasets for pretraining, except for the official test set of CT-RATE. Additionally, we split the test set from BIMCV-R for later cross-modal tasks, following (Chen et al., 2024). For preprocessing, we follow the RadGenome-CT (Zhang et al., 2024) pipeline to extract unique CT-report pairs. In total, we obtain 52,639 samples, with 6,548 samples from BIMCV-R, 25,691 samples from CT-RATE, and 20,400 samples from INSPECT for pre-training. All CT volumes are resampled to a spacing of [1,1,4] mm, resized to $256 \times 256 \times 128$, and normalized to the range [0,1] after truncating Hounsfield unit (HU) values to $[-1000, +1000]$.

**Pre-training Implementation.** We use a 3D ResNet50[3] as the visual encoder and MedCPT (Jin et al., 2023) as the text encoder. The AdamW optimizer is employed with a learning rate of $1 \times 10^{-3}$

[3]We use the implementation from https://docs.monai.io/en/stable/networks.html

and a cosine annealing scheduler. We pretrain for 50 epochs with a 5-epoch warmup. The batch size is set to 32 per GPU, and we implement our training on 8 A100-80G GPUs, resulting in a total effective batch size of 256. No data augmentation is applied to the volumes to preserve spatial integrity and intensity. We only use random cropping to generate three local views of size $128 \times 128 \times 64$. In this study, we generate three local views for each 3D volume-report pair.

### 4.1 Downstream Tasks Configuration

We evaluate T3D on a variety of downstream tasks:
**Classification:** We assess zero-shot and fine-tuned classification on the CT-RATE (Hamamci et al., 2024a) and CC-CCII (Zhang et al., 2020a) official test set.
**Cross-modal Retrieval:** We evaluate zero-shot image-to-text and text-to-image retrieval on the BIMCV-R dataset following (Chen et al., 2024).
**Report Generation:** We evaluate this task on the official test set of CT-RATE (Hamamci et al., 2024a), implementing it based on the LLaVA architecture (Liu et al., 2024). We use Qwen2.5-7B-Instruct [4] as the LLM backbone and employ the visual encoder from our work and baselines to extract the visual representation.
**Segmentation:** We perform multi-organ segmentation on the AMOS (Ji et al., 2022) dataset and lung tumor segmentation on the MSD-Lung (Antonelli et al., 2022a) dataset, following the protocols in VoCo (Wu et al., 2024a,b).

The detailed configurations and implementations for these downstream tasks are provided in Appendix A.

### 4.2 Baseline Selection

We compare T3D with several state-of-the-art (SOTA) visual representation learning methods via vSSL and language supervision:
**3DMAE (Chen et al., 2023d):** A vSSL-based model that reconstructs pixel-level features from masked volumes to learn low-level visual representations.
**VoCo (Wu et al., 2024b):** A vSSL-based model that crops sub-volumes and predicts their locations in the original volume, learning relative local visual features.
**MRM (Zhou et al.):** A 2D MedVLP method that applies masked image and text modeling, leveraging cross-modal reconstruction to learn joint representations.
**IMITATE (Liu et al., 2023a):** A 2D MedVLP method that aligns multi-level visual features with different sections of the report.
**CT-CLIP (Hamamci et al., 2024a):** A 3D MedVLP model that aligns the entire volume with text using the original CLIP loss.
**Merlin (Blankemeier et al., 2024):** A 3D MedVLP model trained on in-house data using CLIP-style alignment.

For 3DMAE (Chen et al., 2023d), we use their official code to reimplement them on our curated dataset since they do not release official pretrained weights. For MRM (Zhou et al.) and IMITATE (Liu et al., 2023a), we replace their 2D visual encoder with a 3D version for a fair comparison and use 3D input. For VoCo (Wu et al., 2024b), Merlin (Blankemeier et al., 2024), and CT-CLIP (Hamamci et al., 2024a), we use their official pretrained weights to ensure a fair comparison.

## 5 Results

We evaluate T3D across a range of downstream tasks, comparing its performance against several state-of-the-art models. Our results are presented for each task in terms of standard evaluation metrics, and we discuss the performance of T3D in comparison to the baselines.

### 5.1 Zero-shot and Fine-tuned Classification

For zero-shot classification and the fine-tuning setting, T3D outperforms all baselines, achieving the highest accuracy across both the CT-RATE (Hamamci et al., 2024a) and CC-CCII (Zhang et al., 2020a) datasets. Notably, it surpasses visual SSL methods and language supervision methods, demonstrating the superiority of our proposed framework in terms of precision, AUC, and F1-score, as shown in Table 1. Furthermore, in the fine-tuning setting, all language supervision baselines reach or even outperform the visual SSL methods (Chen et al., 2023d; Wu et al., 2024b) across both datasets. This demonstrates the necessity of designing a VLP method to learn more representative 3D visual features.

### 5.2 Cross-modal Retrieval

In zero-shot cross-modal retrieval, we implement both image-to-text and text-to-image retrieval tasks to evaluate how well the image and text representations are aligned. T3D outperforms all baselines on

[4] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

| Method | Zero-Shot Classification | | | | | | | | Fine-Tune Classification | | | | | | | | Cross-Modal Retrieval | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CT-RATE | | | | CC-CCII | | | | CT-RATE | | | | CC-CCII | | | | BIMCV-R (Text to Image) | | | BIMCV-R (Image to Text) | | |
| | Prec. | AUC | ACC | F1 | Prec. | AUC | ACC | F1 | Prec. | AUC | ACC | F1 | Prec. | AUC | ACC | F1 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| **Visual SSL only** | | | | | | | | | | | | | | | | | | | | | | |
| 3DMAE (Chen et al., 2023d) | / | / | / | / | / | / | / | / | 30.1 | 70.4 | 64.7 | 64.8 | 82.7 | 88.4 | 87.6 | 85.4 | / | / | / | / | / | / |
| VoCo (Wu et al., 2024b) | / | / | / | / | / | / | / | / | 32.0 | 72.0 | 68.1 | 69.4 | 87.9 | 90.9 | 90.83* | 88.7 | / | / | / | / | / | / |
| **Language Supervision** | | | | | | | | | | | | | | | | | | | | | | |
| MRM (Zhou et al.) | 27.6 | 67.3 | 61.4 | 65.2 | 65.2 | 82.1 | 78.5 | 80.0 | 32.4 | 74.8 | 67.5 | 68.6 | 85.2 | 90.7 | 88.0 | 88.5 | 3.0 | 7.2 | 21.3 | 3.2 | 7.6 | 20.9 |
| IMITATE (Liu et al., 2023a) | 29.5 | 68.9 | 63.6 | 66.4 | 68.6 | 83.7 | 80.2 | 81.5 | 33.0 | 74.3 | 68.2 | 69.7 | 86.4 | 91.5 | 89.2 | 89.7 | 3.1 | 7.9 | 21.5 | 3.6 | 7.8 | 21.7 |
| CT-CLIP (Hamamci et al., 2024a) | 30.6* | 70.4* | 65.1* | 69.1* | 71.6 | 84.3 | 82.3 | 83.0 | 34.2* | 75.0* | 69.2* | 72.8* | 90.8 | 92.0 | 91.4 | 90.3 | 3.9 | 8.3 | 22.4 | 3.7 | 8.5 | 22.9 |
| Merlin (Blankemeier et al., 2024) | 33.7* | 72.8* | 67.2* | 70.9* | 73.2 | 86.4 | 85.0 | 85.9 | 37.1 | 76.2 | 71.0 | 75.0 | 91.5 | 91.9 | 91.5 | 89.6 | 4.0 | 8.7 | 23.5 | 4.1 | 8.9 | 23.4 |
| **T3D (Ours)** | **35.1** | **73.7** | **69.0** | **72.5** | **75.0** | **89.4** | **88.3** | **87.2** | **39.5** | **80.2** | **76.3** | **77.8** | **93.1** | **93.2** | **92.7** | **92.1** | **4.7** | **10.0** | **25.6** | **4.9** | **10.4** | **25.9** |

Table 1: Performance comparison of visual SSL and language supervision methods on zero-shot classification, fine-tune classification, and cross-modal retrieval tasks. '/' indicates that visual SSL methods are unable to perform cross-modal tasks since they only learn representations from images. '*' denotes results directly cited from (Shui et al., 2025; Wu et al., 2024b). The best results in each column are highlighted in **bold**.

| Method | AMOS | MSD-Lung |
|---|---|---|
| | Dice | Dice |
| **Visual SSL only** | | |
| 3DMAE (Chen et al., 2023d) | 82.71* | 65.32 |
| VoCo (Wu et al., 2024b) | 88.06* | 68.99* |
| **Language Supervision** | | |
| MRM (Zhou et al.) | 85.12 | 65.67 |
| IMITATE (Liu et al., 2023a) | 84.51 | 67.31 |
| CT-CLIP (Hamamci et al., 2024a) | 83.44 | 68.37 |
| Merlin (Blankemeier et al., 2024) | 84.74 | 68.89 |
| **T3D (Ours)** | **89.83** | **70.12** |

Table 2: Semantic segmentation performance comparison of visual SSL and language supervision methods on AMOS and MSD-Lung datasets. Dice scores are reported for both datasets. '*' denotes results directly cited from (Wu et al., 2024b). The best results in each column are highlighted in **bold**.

the R@1, R@5, and R@10 metrics in both tasks. This demonstrates the superiority of our framework and the benefits of the multi-view alignment strategy in enhancing cross-modal representation learning.

### 5.3 Report Generation

For the report generation task, as shown in Table 3, we use lexical metrics such as BLEU-1 to BLEU-4 and ROUGE-1, ROUGE-2, and ROUGE-L to evaluate the quality of generated reports. Additionally, we utilize clinical efficacy metrics, including Precision, Recall, and F1, following (Hamamci et al., 2024b), to assess the relevance and accuracy of the reports. On both types of metrics, T3D outperforms all baselines, demonstrating the superiority of our framework in generating high-quality reports across both lexical and clinical dimensions.

Furthermore, all language supervision-based visual encoders outperform the visual SSL-only methods, as shown in Table 3. This highlights that, on the report generation task, the multimodal representations learned through language supervision result in better performance, benefiting the task's specific requirements. Sample generated reports are shown in Figure 4. As shown, our method detects the correct patterns, including subtle ones such as lymph nodes.

### 5.4 Semantic Segmentation

We further evaluate the dense visual representations learned from our method using multi-organ segmentation on the AMOS (Ji et al., 2022) and MSD-Lung tumor segmentation (Antonelli et al., 2022b) datasets. The results, as shown in Table 2, demonstrate that our method, T3D, outperforms all other methods. Although VoCo (Wu et al., 2024b) substantially outperforms other language supervision methods on the organ segmentation task, it does not achieve the same advantage on the tumor segmentation task. This highlights the limitations of visual SSL methods, which may not fully capture the complexities of tumor segmentation. However, our method, T3D, still surpasses VoCo (Wu et al., 2024b), which can be attributed to our multi-view alignment approach. This approach allows for learning multi-level visual features, significantly benefiting the segmentation task.

## 6 Analysis

**Loss Function Ablation:** We ablate the $\mathcal{L}$GCA and $\mathcal{L}$TMA losses, finding that the best performance is achieved when both losses are used. Removing $\mathcal{L}$GCA reduces performance on the classification task due to the lack of global representation, while removing $\mathcal{L}$TMA significantly harms segmentation and report generation tasks due to the loss of local visual feature learning. These results suggest that using both losses is essential to boost performance across all tasks, as shown in Table 4(a).

| Method | Report Generation on CT-RATE | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Lexical Metrics | | | | | | | Clinical Efficacy Metric | | | | |
| | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE-1 | ROUGE-2 | ROUGE-L | Precision | Recall | F1 | GREEN | RaTEScore |
| **Visual SSL only** | | | | | | | | | | | | |
| 3DMAE (Chen et al., 2023d) | 26.5 | 17.9 | 13.2 | 10.3 | 24.0 | 19.4 | 21.2 | 13.1 | 9.4 | 11.0 | 26.9 | 35.3 |
| VoCo (Wu et al., 2024b) | 30.3 | 23.7 | 18.5 | 15.2 | 30.1 | 25.1 | 27.4 | 16.3 | 13.2 | 14.6 | 31.7 | 39.6 |
| **Language Supervision** | | | | | | | | | | | | |
| MRM (Zhou et al.) | 34.7 | 26.9 | 20.9 | 16.7 | 33.4 | 27.2 | 29.6 | 20.4 | 15.3 | 17.5 | 36.7 | 44.9 |
| IMITATE (Liu et al., 2023a) | 41.2 | 31.7 | 25.1 | 21.0 | 37.9 | 31.1 | 32.4 | 25.8 | 17.2 | 20.7 | 40.2 | 49.1 |
| CT-CLIP (Hamamci et al., 2024a) | 44.4 | 34.4 | 27.9 | 23.6 | 40.1 | 33.8 | 30.9 | 31.7 | 18.1 | 25.3 | 42.9 | 53.1 |
| Merlin (Blankemeier et al., 2024) | 47.9 | 35.6 | 28.2 | 24.1 | 41.5 | 35.0 | 36.0 | 33.1 | 19.3 | 25.8 | 46.4 | 56.5 |
| **T3D (Ours)** | **50.1** | **38.3** | **30.4** | **26.2** | **43.8** | **36.7** | **37.8** | **35.5** | **20.7** | **27.4** | **49.2** | **59.6** |

Table 3: Comparison of methods on the report generation task on the CT-RATE official test set using both lexical and clinical efficacy metrics. Lexical metrics include BLEU-1 to BLEU-4 and ROUGE-1, ROUGE-2, and ROUGE-L scores, while clinical metrics include Precision, Recall, and F1 following (Hamamci et al., 2024b), as well as GREEN (Ostmeier et al., 2024) and RaTEScore (Zhao et al., 2024). The best results in each column are highlighted in **bold**.

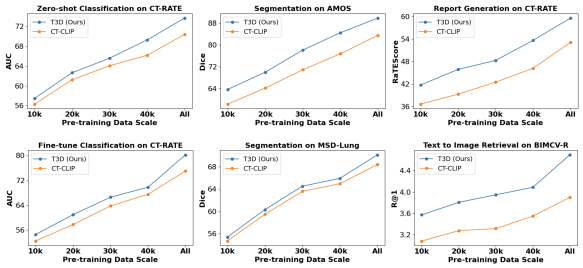


Figure 3: Comparison of T3D (Ours) and CT-CLIP (Hamamci et al., 2024a) across six tasks, showing AUC, Dice, RaTES score, and R@1 for varying pre-training data scales from 10k to the full dataset. T3D consistently outperforms CT-CLIP across all data scales and tasks, particularly with larger datasets.

| Loss Components | | Zero-shot Classification | Segmentation | Report Generation |
|---|---|---|---|---|
| $\mathcal{L}_{\text{GCA}}$ | $\mathcal{L}_{\text{TMA}}$ | CT-RATE (AUC) | AMOS (Dice) | CT-RATE (RaTEScore) |
| ✓ | | 71.8 | 85.0 | 56.3 |
| | ✓ | 71.4 | 85.2 | 55.9 |
| ✓ | ✓ | **73.7** | **89.8** | **59.6** |

(a) Loss Function Ablation.

| $\mathcal{L}_{\text{TMA}}$ | Zero-shot Classification | Segmentation | Report Generation |
|---|---|---|---|
| | CT-RATE (AUC) | AMOS (Dice) | CT-RATE (RaTEScore) |
| w/ Text-Informing | **73.7** | **89.8** | **59.6** |
| w/o Text-Informing | 72.2 | 86.5 | 57.6 |

(b) Effect of Text-Informing on $\mathcal{L}_{\text{TMA}}$.

| Number of Cropped Views | Zero-shot Classification | Segmentation | Report Generation |
|---|---|---|---|
| | CT-RATE (AUC) | AMOS (Dice) | CT-RATE (RaTEScore) |
| 1 | 70.4 | 86.4 | 56.5 |
| 2 | 71.8 | 87.9 | 57.1 |
| 3 | **73.7** | **89.8** | **59.6** |
| 4 | 73.0 | 88.3 | 59.1 |

(c) Effect of Number of Cropped Local Views.

Table 4: Ablation study results for T3D. (a) Comparison of loss functions $\mathcal{L}_{\text{GCA}}$ and $\mathcal{L}_{\text{TMA}}$. (b) Impact of text-informed alignment in $\mathcal{L}_{\text{TMA}}$. (c) Effect of the number of cropped local views used during pre-training. Best results are bolded.

**Text-informed Alignment:** We investigate the impact of incorporating the text-informing strategy in TMA, as shown in Table 4(b). Removing text-informing results in a significant drop in performance, particularly for tasks like report generation and segmentation, which rely on learning fine-grained visual features. This decline may be due to the absence of report information, causing the local view embeddings to become ambiguous and harder to associate with their source volume. Without text-informing, the model may confuse regions from different volumes.

**Number of Cropped Local Views:** We investigate the impact of the number of local views used during training. Reducing the number of views from 3 to 2 or 1 leads to a decrease in performance, particularly for multi-organ segmentation and report generation tasks. This suggests that a higher number of cropped local views encourages the model to learn more comprehensive spatial features. When increasing the number of views to 4, no further improvement is observed. Based on this, we select 3 local views as the optimal choice for training, as shown in Table 4(c).

**Model Architecture Hyperparameters:** We ablate the number of transformer layers in the text-informed block $\mathcal{F}_{\psi}(\cdot)$, varying the layers from 1 to 3. The results show that performance saturates after a single transformer layer, with minimal improvement observed by adding more layers. This suggests that a single transformer layer is sufficient for text-informed alignment, and further layers do not contribute significantly to the model's performance, as detailed in Table 5.

**Model and Data Scalability:** We evaluate the impact of both model and data scale on T3D's performance. As shown in Figure 3, we analyze the effect of varying pre-training data scales on T3D

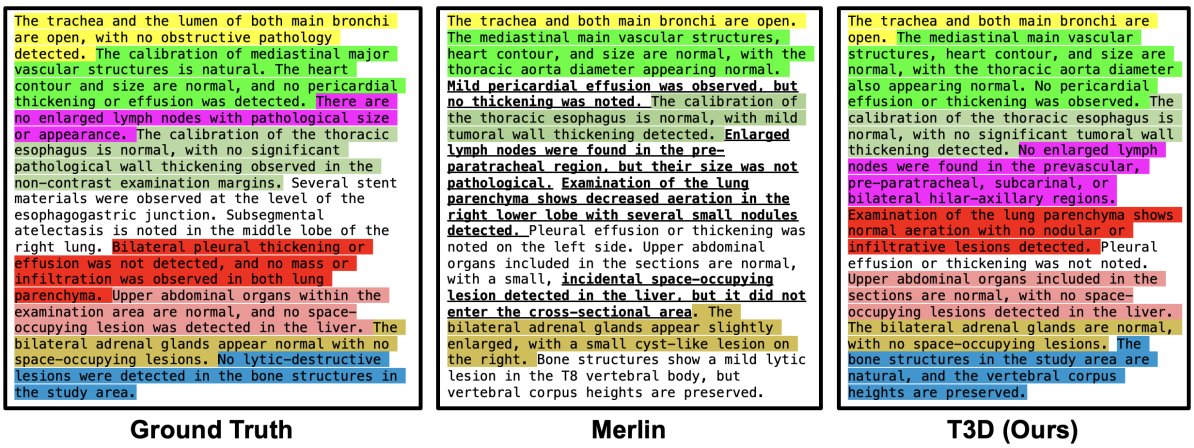


Figure 4: Report generation results of Merlin (Blankemeier et al., 2024) and **T3D (Ours)**. Text highlighted in the same color indicates correct predictions, while bold and underlined text marks incorrect parts. Merlin shows incorrect patterns in various areas, whereas T3D provides more accurate results, particularly in the detection of lymph nodes and other pathologies.

| Number of Transformer Layers | Zero-shot Classification | Segmentation | Report Generation |
|---|---|---|---|
| | CT-RATE (AUC) | AMOS (Dice) | CT-RATE (RaTEScore) |
| 1 | **73.7** | **89.8** | **59.6** |
| 2 | 73.5 | 89.3 | 59.4 |
| 3 | 73.3 | 89.5 | 59.3 |

Table 5: Performance comparison of models with different transformer layer counts during text-informing for $\mathcal{L}_{\text{TMA}}$. Best performance for each task is bolded.

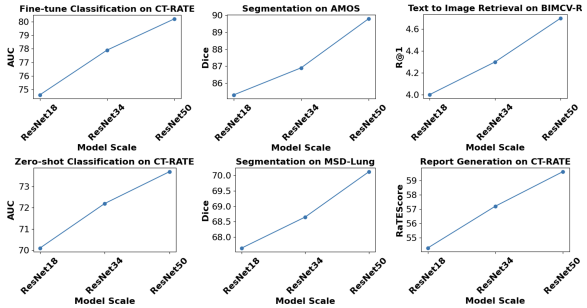


Figure 5: Performance of **T3D** pre-trained on the proposed **CT-3DVLP** dataset across six tasks, with varying model scales: ResNet18, ResNet34, and ResNet50. The results show consistent performance improvement as the model scale increases.

and CT-CLIP (Hamamci et al., 2024a). Our method consistently surpasses CT-CLIP (Hamamci et al., 2024a) from 10k to the full pre-training dataset, demonstrating the effectiveness of T3D across different data scales. Additionally, we evaluate the scalability of our model by testing different ResNet architectures (ResNet18, ResNet34, and ResNet50) as visual encoders. As visualized in Figure 5, T3D shows consistent performance improvements as the model scale increases, highlighting its ability to leverage larger models for better performance across multiple tasks.

## 7 Conclusion

In this work, we present the first and largest publicly available 3D medical VLP dataset, named **CT-3DVLP**, curated entirely from public resources. We also introduce the **T3D** framework, which leverages both global alignment and a novel **text-informed multi-view alignment** strategy to enhance learning and improve performance across various tasks. We demonstrate the effectiveness of T3D on six downstream tasks, including both uni-modal and cross-modal tasks, and show that it outperforms existing methods, such as vSSL and other language supervision approaches that rely on in-house data. Additionally, we highlight the scalability of our method. We believe that T3D, alongside the CT-3DVLP dataset, will make a significant contribution to advancing research in the 3D medical VLP domain.

## Limitations

While we propose the T3D framework and the largest publicly available CT-3DVLP dataset, there are several limitations. Even though we have collected nearly all publicly available 3D medical image-report pairs, the dataset still remains limited in size compared to the large-scale datasets used in models like CLIP (Radford et al., 2021). With only 50k samples, it falls short of the million-level datasets typically used in such models. Additionally, due to the complexities of 3D medical data, it is impractical to directly leverage powerful 2D visual encoders, limiting the performance of our model. Computational constraints also led us to use ResNet-50 as the visual encoder, rather than more advanced or larger vision models. These limitations point to areas for future work, such as dataset expansion and the integration of more sophisticated 3D vision encoders.

## References


Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. 2022a. The medical segmentation decathlon. *Nature communications*, 13(1):4128.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. 2022b. The medical segmentation decathlon. *Nature communications*, 13(1):4128.

Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar Gardezi, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Cesar Truyts, et al. 2024. Merlin: A vision language foundation model for 3d computed tomography. *Research Square*, pages rs–3.

Weiwei Cao, Jianpeng Zhang, Yingda Xia, Tony CW Mok, Zi Li, Xianghua Ye, Le Lu, Jian Zheng, Yuxing Tang, and Ling Zhang. 2024. Bootstrapping chest ct image understanding by distilling knowledge from x-ray expert models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11238–11247.

Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. 2020. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in neural information processing systems*, 33:12546–12558.

Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. 2019. Self-supervised learning for medical image analysis using image context restoration. *Medical image analysis*, 58:101539.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. 2020. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR.

Qiuhui Chen, Xinyue Hu, Zirui Wang, and Yi Hong. 2023a. Medblip: Bootstrapping language-image pre-training from 3d medical images and texts. *arXiv preprint arXiv:2305.10799*.

Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Gang Yu, and Tao Chen. 2023b. End-to-end 3d dense captioning with vote2cap-detr. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11124–11133.

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

Yinda Chen, Che Liu, Xiaoyu Liu, Rossella Arcucci, and Zhiwei Xiong. 2024. Bimcv-r: A landmark dataset for 3d ct text-image retrieval. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 124–134. Springer.

Zekai Chen, Devansh Agarwal, Kshitij Aggarwal, Wiem Safta, Mariann Micsinai Balan, and Kevin Brown. 2023d. Masked image modeling advances 3d medical image analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1970–1980.

Pujin Cheng, Li Lin, Junyan Lyu, Yijin Huang, Wenhan Luo, and Xiaoying Tang. 2023. Prior: Prototype representation joint learning from medical images and reports. *arXiv preprint arXiv:2307.12577*.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. 2018. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

Mikhail Goncharov, Vera Soboleva, Anvar Kurmukov, Maxim Pisov, and Mikhail Belyaev. 2023. vox2vec: A framework for self-supervised contrastive learning of voxel-level representations in medical images. *arXiv preprint arXiv:2307.14725*.

Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Michael B Gotway, and Jianming Liang. 2022. Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20824–20834.

Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Bastian Wittmann, Enis Simsar, Mehmet Simsar, et al. 2024a. A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. *arXiv preprint arXiv:2403.17834*.

Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. 2024b. Ct2rep: Automated radiology report generation for 3d medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 476–486. Springer.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

Shih-Cheng Huang, Zepeng Huo, Ethan Steinberg, Chia-Chun Chiang, Matthew P Lungren, Curtis P Langlotz, Serena Yeung, Nigam H Shah, and Jason A Fries. 2023a. Inspect: a multimodal dataset for pulmonary embolism diagnosis and prognosis. *arXiv preprint arXiv:2311.10798*.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. 2021. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951.

Weijian Huang, Hongyu Zhou, Cheng Li, Hao Yang, Jiarun Liu, and Shanshan Wang. 2023b. Enhancing representation in radiography-reports foundation model: A granular alignment algorithm using masked contrastive learning. *arXiv preprint arXiv:2309.05904*.

Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. 2022. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732.

Yankai Jiang, Mingze Sun, Heng Guo, Ke Yan, Le Lu, and Minfeng Xu. 2023. Anatomical invariance modeling and semantic alignment for self-supervised learning in 3d medical image segmentation. *arXiv preprint arXiv:2302.05615*.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Jiayu Lei, Lisong Dai, Haoyun Jiang, Chaoyi Wu, Xiaoman Zhang, Yao Zhang, Jiangchao Yao, Weidi Xie, Yanyong Zhang, Yuehua Li, et al. 2023. Unibrain: Universal brain mri diagnosis with hierarchical knowledge-enhanced pre-training. *arXiv preprint arXiv:2309.06828*.

Che Liu, Sibo Cheng, Miaojing Shi, Anand Shah, Wenjia Bai, and Rossella Arcucci. 2023a. Imitate: Clinical prior guided hierarchical vision-language pre-training. *arXiv preprint arXiv:2310.07355*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yuan Liu, Songyang Zhang, Jiacheng Chen, Kai Chen, and Dahua Lin. 2023b. Pixmim: Rethinking pixel reconstruction in masked image modeling. *arXiv preprint arXiv:2303.02416*.

Yuan Liu, Songyang Zhang, Jiacheng Chen, Zhaohui Yu, Kai Chen, and Dahua Lin. 2023c. Improving pixel-based mim by reducing wasted modeling capability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5361–5372.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, et al. 2024. Green: Generative radiology report evaluation and error notation. *arXiv preprint arXiv:2405.03595*.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. 2016. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Zhongyi Shui, Jianpeng Zhang, Weiwei Cao, Sinuo Wang, Ruizhe Guo, Le Lu, Lin Yang, Xianghua Ye, Tingbo Liang, Qi Zhang, et al. 2025. Large-scale and fine-grained vision-language pre-training for enhanced ct image understanding. *arXiv preprint arXiv:2501.14548*.

Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 2020. 3d self-supervised methods for medical imaging. *Advances in neural information processing systems*, 33:18158–18172.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. 2022. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20730–20740.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12).

Zhongwei Wan, Che Liu, Mi Zhang, Jie Fu, Benyou Wang, Sibo Cheng, Lei Ma, César Quilodrán-Casas, and Rossella Arcucci. 2023. Med-unic: Unifying cross-lingual medical vision-language pre-training by diminishing bias. *arXiv preprint arXiv:2305.19894*.

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. 2022. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *arXiv preprint arXiv:2210.06044*.

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. 2022. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023a. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21372–21383.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023b. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*.

Linshan Wu, Jiaxin Zhuang, and Hao Chen. 2024a. Large-scale 3d medical image pre-training with geometric context priors. *arXiv preprint arXiv:2410.09890*.

Linshan Wu, Jiaxin Zhuang, and Hao Chen. 2024b. Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22873–22882.

Yutong Xie, Jianpeng Zhang, Yong Xia, and Qi Wu. 2022a. Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier. In *European Conference on Computer Vision*, pages 558–575. Springer.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. 2022b. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663.

Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. 2023a. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1179–1189.

Le Xue, Ning Yu, Shu Zhang, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. 2023b. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. *arXiv preprint arXiv:2305.08275*.

Yihan Zeng, Chenhan Jiang, Jiageng Mao, Jianhua Han, Chaoqiang Ye, Qingqiu Huang, Dit-Yan Yeung, Zhen Yang, Xiaodan Liang, and Hang Xu. 2023. Clip2: Contrastive language-image-point pretraining from real-world point cloud data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15244–15253.

Kang Zhang, Xiaohong Liu, Jun Shen, Zhihuan Li, Ye Sang, Xingwang Wu, Yunfei Zha, Wenhua Liang, Chengdi Wang, Ke Wang, et al. 2020a. Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography. *Cell*, 181(6):1423–1433.

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

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Jiayu Lei, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Radgenome-chest ct: A grounded vision-language dataset for chest ct analysis. *arXiv preprint arXiv:2404.16754*.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2020b. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Ratescore: A metric for radiology report generation. *arXiv preprint arXiv:2406.16845*.

Hong-Yu Zhou, Chenyu Lian, Liansheng Wang, and Yizhou Yu. Advancing radiograph representation learning with masked record modeling. In *The Eleventh International Conference on Learning Representations*.

Hong-Yu Zhou, Chixiang Lu, Chaoqi Chen, Sibei Yang, and Yizhou Yu. 2023. A unified visual information preservation framework for self-supervised pre-training in medical image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B Gotway, and Jianming Liang. 2019. Models genesis: Generic autodidactic models for 3d medical image analysis. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 384–393. Springer.

## A Downstream Task Details

**Zero-shot Classification:** We use the CT-RATE dataset (Hamamci et al., 2024a) for zero-shot classification, following the protocol in (Hamamci et al., 2024a). T3D is applied without fine-tuning, using the pretrained model for direct classification with the disease name as the category prompt. Evaluation metrics include precision, AUC, accuracy, and F1 score. Image preprocessing is the same as during pretraining.

**Fine-tuned Classification:** For fine-tuned classification, we follow the fine-tuning procedure from (Hamamci et al., 2024a) on the CT-RATE dataset. The images are preprocessed as in pretraining, and T3D is fine-tuned on the training set and evaluated on the test set. Metrics include accuracy, precision, and recall. We use a batch size of 32, a learning rate of $1 \times 10^{-3}$, epochs as 50, and cosine learning rate decay. Experiments are run on a single A100-80GB GPU.

**Zero-shot Cross-modal Retrieval:** For zero-shot cross-modal retrieval, we use the BIMCV-R dataset (Chen et al., 2024) and follow (Chen et al., 2024). Both image and report are embedded into a latent space, and cosine similarity is computed to identify the top-K matches. Retrieval performance is measured using recall@1-10. Image preprocessing is consistent with the pretraining implementation.

**Report Generation:** For report generation, we use the official training set from the CT-RATE dataset (Hamamci et al., 2024a) and the official test set for evaluation. Following LLaVA (Liu et al., 2024), we use Qwen2.5-7b-Instruct as the LLM backbone and the pretrained visual encoder to extract image embeddings. A two-layer MLP serves as the connector, and training is done in two stages: first training the connector, then freezing the ViT and fine-tuning both the connector and the LLM. Generated reports are evaluated using BLEU-1 to 4 and ROUGE-1, 2, L scores. Additional clinical efficacy metrics are adopted from (Hamamci et al., 2024b), with further evaluation using GREEN and RaTEScore (Ostmeier et al., 2024; Zhao et al., 2024).

**Segmentation Tasks:** For multi-organ segmentation, we use the AMOS (Ji et al., 2022) dataset, following the protocols in (Wu et al., 2024a). For lung tumor segmentation, we use the MSD-Lung tumor dataset (Antonelli et al., 2022b). A 3D U-Net architecture is employed, with a pretrained visual encoder and a randomly initialized decoder. Input volumes are normalized to a spacing of 1mm along the three axes, with voxel intensities truncated within the HU range of [-1000, 1000] and normalized to [0,1]. During training, the entire volume is used, with augmentations applied at probabilities of 0.5 for random flipping, 0.3 for rotation, 0.1 for intensity scaling, and 0.1 for shifting. The Dice score is used as the evaluation metric, adhering to the fine-tuning procedure from the official VoCo repository [5].

[5] https://github.com/Luffy03/VoCo