LEARNING HOW MUCH, NOT JUST WHAT: CROSS-PATIENT BURDEN ORDER FOR CT VISION-LANGUAGE PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Volumetric CT vision–language pretraining learns 3D representations from scan–report pairs, but global and anatomy-aware objectives supervise only correspondence: they establish what is present and leave how much unconstrained. Nothing separates a mild from an extensive case of the same finding along a consistent direction, so the graded burden language in reports collapses into a present/absent signal. Longitudinal supervision would supply this order, but patient-matched CT pairs are scarce at scale; cross-sectional cohorts already encode weak burden cues across different patients. We introduce Spectrum, an anatomy-conditioned framework that represents each study at whole-study and organ scopes. For each organ-mapped pathology, a rule-based scorer mines confidence-filtered lower-to-higher pairs of different patients, and Burden-Direction Alignment (BDA) aligns the pathology-conditioned image delta with the report delta at each scope, separating that direction from its reverse. Because the endpoints are different people, a target-conditioned aligner first makes them comparable, so the delta reflects burden rather than between-patient variation. BDA further separates the selected direction from its reverse, anchors it to the observed higher-burden endpoint, and enforces consistency across ordered triplets. Since every pair is drawn within a single pathology, BDA is designed to constrain intra-class structure that image–report contrast alone never touches. Spectrum attains 85.6 zero-shot AUROC on CT-RATE and 72.7 on RAD-ChestCT, with gains in linear probing and retrieval. Weak cross-patient order complements anatomy-aware correspondence, yielding burden-aware CT representations without longitudinal data. The code will be released.

1 INTRODUCTION

Computed tomography is a cornerstone of clinical imaging, yet reading a volumetric study means reviewing hundreds of slices and integrating evidence across anatomical systems (Blankemeier et al., 2026; Gong et al., 2024; Huang et al., 2025; Wang et al., 2026; Chen et al., 2023). Volumetric CT vision–language pretraining (VLP) offers a scalable route to transferable 3D representations, learning from the scan–report pairs that hospitals already produce (Hamamci et al., 2026; Blankemeier et al., 2026; Chen et al., 2023). What makes chest CT reports valuable is not only that they name findings and their locations, but that they grade them: a nodule is recorded as smaller than 5 mm or as 66×55 mm, consolidation as mild or as extensive. This graded language costs nothing to collect and exists at cohort scale. Nevertheless, current objectives compress it into a binary present/absent signal.

The dominant formulation aligns a whole volume with a report (Hamamci et al., 2026; Blankemeier et al., 2026), and recent methods extend this to local or anatomy-aware correspondence (Shui et al., 2025). These objectives answer one question: which visual and textual observations belong together? But correspondence alone is directionless. A model may learn that lung evidence matches a lung snippet with no constraint that "extensive" disease lie beyond "mild" disease along a consistent pathology-burden direction. The cost is a representation organized by category but not by degree: cases sharing a pathology are pulled toward a common region of feature space regardless of how much disease they contain, so anatomy localizes correspondence while burden order within that anatomy is left unconstrained.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

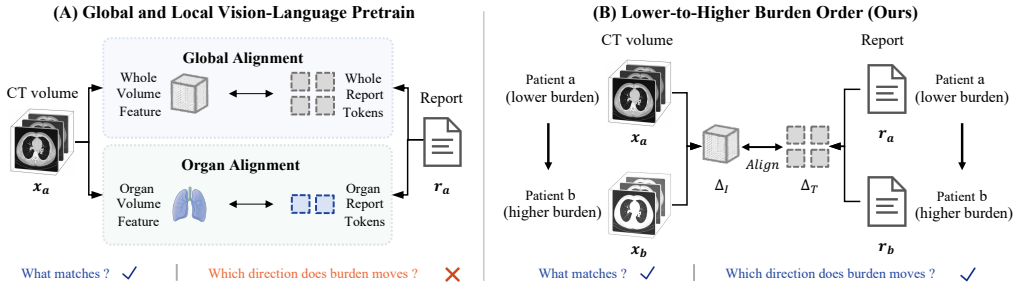

Figure 1: Correspondence versus burden order. (A) Global and organ objectives match a volume to its report at two anatomical scopes. (B) Spectrum mines ordered pairs of different patients, a (lower burden) and b (higher), and aligns image delta Δ_I with report delta Δ_T at both scopes. Correspondence fixes *what matches*; the order supplies *how much*.

Direct longitudinal supervision could impose such order, but repeat CT studies carrying comparable pathology evidence are difficult to assemble at the scale VLP requires. Cross-sectional cohorts, in contrast, are both large and already annotated in the relevant way: labels, negation, size mentions, burden descriptors, and organ-specific report fragments jointly indicate which of two patients carries more of a given pathology. This motivates our central question: *can weak cross-patient order extend anatomy-aware correspondence toward burden-aware representations?*

Our key insight is to use anatomy as the common frame for both correspondence and order, so that each mined relation constrains burden direction within the same scopes in which correspondence is learned. Using two different patients as endpoints, however, raises a difficulty. The difference between their features reflects not only burden but also body habitus, acquisition protocol, and anatomical variation, none of which should define a burden direction. We therefore do not treat the raw feature difference as the supervision target: a target-conditioned aligner first makes the lower-burden representation comparable with its higher-burden counterpart, and only the residual delta is aligned across modalities. Fig. 1 illustrates this shift.

We introduce Spectrum, an anatomy-conditioned VLP framework. Spectrum maintains a whole-study representation and mask-gated organ representations in parallel, grounds them with the whole report and with organ-specific report snippets, and uses a pathology-to-organ map to keep representation learning, burden ordering, and evaluation anatomically consistent. Burden-Direction Alignment (BDA) is designed to supply the missing direction signal. For each organ-mapped pathology, a fixed rule-based scorer mines confidence-filtered lower-to-higher relations from labels and report language across different patients, and each relation is applied at the whole-study scope and, when valid, at the corresponding organ scope. At its core, BDA aligns the pathology-conditioned image and report delta and favors the selected direction over its reverse. Two auxiliary constraints keep that orientation well posed: an endpoint consistency supplies an additional training signal, and a composition constraint keeps pairwise directions coherent across ordered triplets of increasing burden. All BDA modules act in feature space and are discarded after training, leaving inference unchanged.

We evaluate Spectrum on CT-RATE and RAD-ChestCT through zero-shot abnormality diagnosis, linear probing, image and report retrieval, controlled component and aligner ablations, and qualitative burden-order analysis. Together, these experiments assess in-domain recognition, external transfer, representation quality, retrieval alignment, and the contribution of individual BDA components.

Our contributions are:

- We introduce Spectrum, an anatomy-conditioned CT vision–language framework that learns not only which findings a report describes but how much disease is present, by mining weak cross-patient ordered relations without requiring longitudinal data.
- We formulate Burden-Direction Alignment, which conditions both modalities on a pathology embedding, removes between-patient variation with a target-conditioned aligner, and aligns image and report deltas under directional, composition and auxiliary endpoint constraints.
- We show that Spectrum reaches 85.6 AUROC on CT-RATE and 72.7 AUROC on external RAD-ChestCT, with consistent gains in frozen-encoder linear probing and retrieval, and we isolate the contribution of each constraint and of the aligner through controlled ablations.

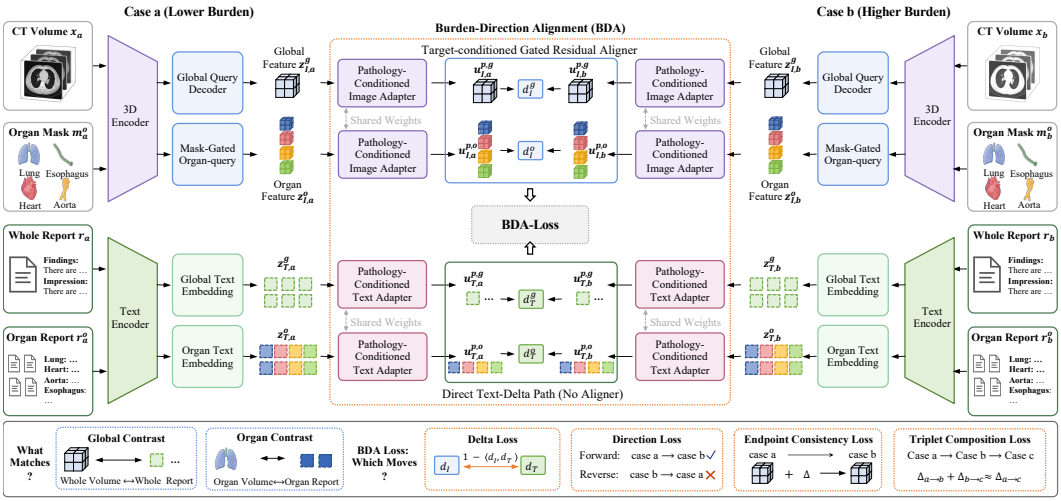

Figure 2: Architecture of Spectrum. A mined relation orders two different patients, a (lower burden) and b (higher), for pathology p. Shared encoders embed both studies at whole-study and mask-gated organ scopes, where paired contrastive objectives fix correspondence (*what matches*). Burden-Direction Alignment (BDA) adds order (*which moves*), aligning image and report delta directions d_I and d_T at each scope. BDA is training-only.

2 Related Work

Medical VLP as correspondence learning. Contrastive VLP learns transferable representations from paired or semantically matched image–text data (Radford et al., 2021; Zhai et al., 2022; Zhang et al., 2022; Wang et al., 2022b; Tiu et al., 2022; Gong et al., 2021; 2022). Medical models exploit report semantics through radiology-specific text modeling, global/local or multi-granular alignment, and structured entity/knowledge supervision (Boecking et al., 2022; Huang et al., 2021; Wang et al., 2022a; Müller et al., 2022; Wu et al., 2023; Jain et al., 2021). Broader biomedical models scale this paradigm to millions of scientific image–text pairs (Zhang et al., 2025; Lin et al., 2023). Despite these advances, existing VLP objectives primarily organize *what matches*, rather than explicitly constraining a consistent lower-to-higher burden direction among cases sharing the same pathology.

Volumetric and anatomy-aware CT VLP. Volumetric CT methods establish whole-volume–report alignment, transfer radiographic knowledge, or incorporate structured, generative, and image-only objectives for representation learning (Hamamci et al., 2026; Bai et al., 2024; Cao et al., 2024; Blankemeier et al., 2026; Lai et al., 2025; Wald et al., 2025). CT-GLIP and fVLM localize correspondence by constructing organ-level image–text pairs or extracting mask-guided organ representations (Lin et al., 2024; Shui et al., 2025). These methods expose anatomy-specific evidence.

Relative order and temporal supervision. Pairwise ranking and severity-aware representation learning impose order from preferences or explicit severity annotations (Burges et al., 2005; Nguyen et al., 2024). Longitudinal medical VLP derives change supervision from patient-matched prior/current radiographs and their reports (Bannur et al., 2023; Yang & Shen, 2025). These methods obtain order from explicit severity supervision.

3 Method

3.1 Problem Formulation and Overview

Each study i contains a CT volume x_i, organ masks $m_i = \{m^o_i\}_{o\in\mathcal{O}}$, a whole report r_i, organ-specific report snippets $\{r^o_i\}_{o\in\mathcal{O}}$, and CT-RATE abnormality labels $y_i \in \{0,1\}^{18}$. We use four organ groups, $\mathcal{O} = \{\text{lung}, \text{heart}, \text{esophagus}, \text{aorta}\}$, together with a global scope g. Here, a *scope* $q \in \{g\} \cup \mathcal{O}$ denotes paired image and text representations at one anatomical level. A fixed pathology-to-organ map σ assigns 16 of the 18 labels to one organ: 11 pulmonary findings to the lung, 3 cardiac findings

to the heart, hiatal hernia to the esophagus, and arterial wall calcification to the aorta. BDA is applied to these 16 organ-mapped pathologies, while medical material and lymphadenopathy remain global-only. A selected relation (a, b, p) encodes a lower-to-higher order between two different patients for pathology p. Fig. 2 summarizes the framework. Spectrum organizes each study into anatomy-conditioned scopes. A whole-study representation retains global context, while four mask-gated organ queries produce organ-specific representations; paired image–text objectives align each visual representation with its whole-report or organ-snippet counterpart. These scopes define *where* correspondence is learned; by aligning pathology-conditioned image and text deltas, BDA defines *how* lower-to-higher pathology-burden order is imposed at the global and organ scopes.

3.2 Anatomy-Conditioned Scopes

Global and organ-specific representations. A 3D encoder produces the feature map $F_i = E_\theta(x_i)$. A learned global query decoder G aggregates the feature map into the whole-study feature $g_i = G(F_i)$. For each organ o, the mask m_i^o is downsampled to the feature-map resolution and used as an attention mask; the mask-gated organ-query aggregator H, with one learned query per organ, pools the valid anatomical features into $h_i^o = H(F_i, m_i^o)$. A shared text encoder maps the whole report and organ-specific report snippets to $t_i^g = E_\phi(r_i)$ and $t_i^o = E_\phi(r_i^o)$, respectively. A global visual projection head P_I^g, an organ-scope visual projection head P_I^o, and a shared text projection head P_T map these features to the embeddings used by the correspondence objectives:

$$z_{I,i}^g = P_I^g(g_i), \;\; z_{T,i}^g = P_T(t_i^g), \;\; z_{I,i}^o = P_I^o(h_i^o), \;\; z_{T,i}^o = P_T(t_i^o). \tag{1}$$

Invalid organ masks are excluded by a validity indicator v_i^o. The organ visual projection P_I^o is shared across organs, while P_T is shared by whole reports and organ snippets.

Global and organ correspondence. Let Z_I^g and Z_T^g stack a mini-batch of projected whole-study representations. Global correspondence is learned with a symmetric image–text contrastive loss using diagonal image–report pairs. For each organ o, let $\mathcal{B}_o = \{i \in \mathcal{B} : v_i^o = 1\}$ denote the valid studies, and let Z_I^o and Z_T^o stack their organ image and snippet representations. We define the organ-level similarity matrix as $S^o = Z_I^o Z_T^{o\top}/\tau_o$. The global and organ correspondence objectives are

$$\mathcal{L}_{\text{global}} = \tfrac{1}{2}\Big[\mathrm{CE}(Z_I^g Z_T^{g\top}, I) + \mathrm{CE}(Z_T^g Z_I^{g\top}, I)\Big], \; \mathcal{L}_{\text{organ}} = \frac{1}{2|\mathcal{O}_\mathcal{B}|}\sum_{o\in\mathcal{O}_\mathcal{B}}\Big[\mathrm{CE}(S^o, I_o) + \mathrm{CE}\Big((S^o)^\top, I_o\Big)\Big], \tag{2}$$

where I_o is the diagonal pairing target over $\mathcal{B}_o$, and $\mathcal{O}_\mathcal{B}$ contains organs with sufficient valid image–snippet pairs in the mini-batch. Together, $\mathcal{L}_{\text{global}}$ and $\mathcal{L}_{\text{organ}}$ establish *what matches* at the whole-study and organ scopes, while cross-patient lower-to-higher order is introduced only by BDA.

3.3 Burden-Direction Alignment

Rule-based burden scoring and cross-patient relation mining. For each organ-mapped pathology p, a fixed, non-learned scorer ϕ_p combines the label $y_{i,p}$ with the whole report and corresponding organ-specific report snippet to produce a burden-evidence score s_i^p and confidence ρ_i^p. Both $s_i^p \in \mathbb{R}_{\geq 0}$ and $\rho_i^p \in \mathbb{R}_{\geq 0}$ are additive, non-normalized scores rather than probabilities. Different patients form an eligible candidate relation when

$$(s_i^p, \rho_i^p) = \phi_p(y_{i,p}, r_i, r_i^{\sigma(p)}), \quad (a, b, p) \in \widetilde{\mathcal{P}} \iff \begin{array}{c} a \neq b, \quad s_b^p - s_a^p \geq \delta, \\ s_a^p, s_b^p \geq s_{\min}, \quad \rho_a^p, \rho_b^p \geq \eta. \end{array} \tag{3}$$

Here, δ requires sufficient separation between the two burden-evidence scores, $s_{\min}$ requires both endpoints to contain adequate evidence for pathology p, and η filters low-confidence relations. The scorer parses pathology mentions, negation patterns, burden or severity descriptors, size or comparative cues, and uncertainty language. Its score supports relative ordering. Within mini-batches, eligible candidates from all active pathologies are assigned the reliability weight $w_{ab}^p = (s_b^p - s_a^p) \times \min(\rho_a^p, \rho_b^p)$, and the top-$K$ candidates form the selected relation set $\mathcal{P} \subseteq \widetilde{\mathcal{P}}$. Each selected triple defines a cross-patient relation $a \to b$, in which study a carries weaker burden cues than study b for pathology p.

Relation mining is scope-independent: each triple (a, b, p) is selected once and then evaluated at the whole-study scope $q = g$ and the organ scope $q = \sigma(p)$. At $q = g$, the pathology embedding e_p

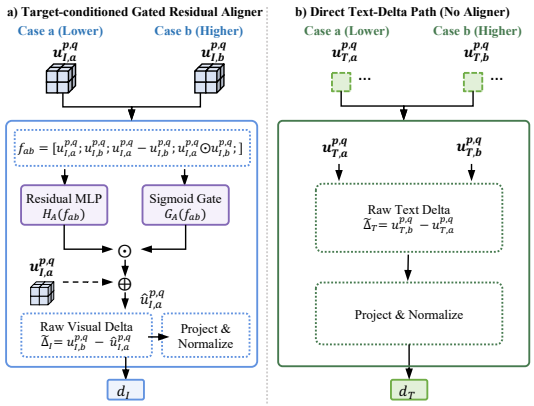

Figure 3: Delta construction in BDA for pathology p at scope q. A target-conditioned aligner maps endpoint a (lower burden) toward b (higher) before differencing, giving d_I, while the aligner-free text path gives d_T. The endpoints are different patients, so only the residual difference isolates burden from habitus and protocol.

conditions the whole-study representations, yielding a pathology-indexed global burden direction rather than a generic whole-study burden axis. At $q = \sigma(p)$, the same relation constrains the corresponding organ-specific representations, grounding the burden direction in the mapped anatomy.

Pathology-conditioned delta construction. For each selected relation (a, b, p) and scope $q \in \{g, \sigma(p)\}$, we first condition the corresponding image and text representations on a learned pathology embedding e_p using modality-specific gated residual adapters:

$$u^{p,q}_{M,i} = C_M(z^q_{M,i}; e_p), \qquad M \in \{I, T\}, \quad i \in \{a, b\}. \tag{4}$$

For each modality, C_M concatenates the selected representation with e_p; modality-specific two-layer update and sigmoid-gate networks produce a gated residual that is added to the input. The image and text adapters are shared across studies, anatomical scopes, and pathologies, with p changing only the embedding e_p rather than selecting an independent network. Each relation is therefore conditioned within one active whole-study or organ-local scope.

Before measuring visual difference, a target-conditioned aligner produces a target-aware visual source, as shown in Fig. 3(a). The aligned visual source, pre-projection image and text differences, and normalized projected directions d_M for $M \in \{I, T\}$ are

$$\widehat{u}^{p,q}_{I,a} = A(u^{p,q}_{I,a}, u^{p,q}_{I,b}), \;\; \widetilde{\Delta}^{a\to b,p,q}_I = u^{p,q}_{I,b} - \widehat{u}^{p,q}_{I,a}, \widetilde{\Delta}^{a\to b,p,q}_T = u^{p,q}_{T,b} - u^{p,q}_{T,a}, \;\; d^{a\to b,p,q}_M = \mathrm{norm}\big(\widetilde{\Delta}^{a\to b,p,q}_M\big). \tag{5}$$

The aligner A concatenates the visual source, target, their signed difference, and element-wise interaction. A two-layer MLP predicts a residual update modulated by a learned sigmoid gate and added to the source. For the forward relation $a \to b$, only the source is transformed, while the target remains unchanged. The text branch bypasses the aligner and uses the difference between pathology-conditioned report representations, retaining it as a direct semantic reference. The modality-specific projectors map the two differences into a shared delta space, and norm denotes ℓ_2 normalization.

Direction-aware and geometric constraints. The deltas above define pathology-conditioned cross-patient difference. The core BDA term makes this difference cross-modal by aligning the unit image and text directions:

$$\mathcal{L}_\Delta = 1 - \left\langle d^{a\to b,p,q}_I, d^{a\to b,p,q}_T \right\rangle. \tag{6}$$

Because both directions are ℓ_2-normalized, $\mathcal{L}_\Delta$ performs cosine direction matching rather than batch-level contrastive learning; $\mathcal{L}_{\text{global}}$ and $\mathcal{L}_{\text{organ}}$ preserve the absolute image–text correspondence. We omit the common indices (p, q) in the following auxiliary terms. To distinguish the selected orientation from undirected similarity, let $s_{\text{fwd}} = \langle d^{a\to b}_I, d^{a\to b}_T \rangle$ and $s_{\text{rev}} = \langle d^{b\to a}_I, d^{a\to b}_T \rangle$. Here, $d^{b\to a}_I$ is recomputed by swapping the visual source and target through the same aligner and delta projector, rather than being defined as $-d^{a\to b}_I$. The auxiliary constraints are

$$\mathcal{L}_{\text{dir}} = [\mu - (s_{\text{fwd}} - s_{\text{rev}})]_+ ,$$
$$\mathcal{L}_{\text{end}} = \mathrm{MSE}\Big(\Psi\Big([\widehat{u}_{I,a}; \widetilde{\Delta}^{a\to b}_I]\Big), u_{I,b}\Big), \mathcal{L}_{\text{comp}} = \mathrm{MSE}\Big(d^{a\to c}_I, \mathrm{norm}\Big(d^{a\to b}_I + d^{b\to c}_I\Big)\Big). \tag{7}$$

Here, $[x]_+ = \max(0, x)$. The direction term enforces $s_{\text{fwd}} \geq s_{\text{rev}} + \mu$ while keeping the textual reference fixed in the selected $a \to b$ direction. The endpoint head Ψ maps $[\widehat{u}_{I,a}; \widetilde{\Delta}_I^{a\to b}] \in \mathbb{R}^{2D}$ to the observed higher-burden visual endpoint, so $\mathcal{L}_{\text{end}}$ serves as an auxiliary latent endpoint-consistency regularizer rather than an image-reconstruction. For an eligible same-pathology triplet with $s_a^p < s_b^p < s_c^p$, $\mathcal{L}_{\text{comp}}$ encourages the direct $a \to c$ direction to agree with the normalized two-step path $a \to b \to c$.

Using the reliability weight w_{ab}^p defined during relation mining, let $\mathcal{P}_q$ denote the selected relations whose representations are valid at scope q. The pairwise and scope-level objectives are

$$\ell_{ab}^{p,q} = \mathcal{L}_\Delta + \lambda_{\text{dir}}\mathcal{L}_{\text{dir}} + \lambda_{\text{end}}\mathcal{L}_{\text{end}}, \quad \mathcal{L}_{\text{BDA}}^q = \frac{\sum_{(a,b,p)\in\mathcal{P}_q} w_{ab}^p \ell_{ab}^{p,q}}{\sum_{(a,b,p)\in\mathcal{P}_q} w_{ab}^p} + \lambda_{\text{comp}}\overline{\mathcal{L}}_{\text{comp}}^q. \tag{8}$$

$\overline{\mathcal{L}}_{\text{comp}}^q$ averages eligible same-pathology triplets at scope q and is zero when none exists; a scope with $\mathcal{P}_q = \varnothing$ is omitted from the batch objective. Thus, the same mined relation is supervised at the whole-study scope and, when valid, at the corresponding pathology-mapped organ scope.

3.4 Training Objective

Before cross-modal training, we warm up the visual branch with $\mathcal{L}_{\text{warm}} = \mathcal{L}_{\text{anchor}} + \lambda_{\text{rank}}\mathcal{L}_{\text{rank}}$. $\mathcal{L}_{\text{anchor}}$ applies label-supervised prediction at the whole-study scope and at valid pathology-mapped organ scopes. At both scopes, $\mathcal{L}_{\text{rank}}$ uses a margin hinge so that a label-positive study scores above a label-negative study for the same pathology. This stage establishes coarse disease-sensitive representations by separating pathology presence from absence, but does not order different burden states among studies sharing the same pathology; the lower-to-higher relations used by BDA are subsequently mined from labels and report cues. During image–text training, the 3D backbone is frozen, while the global decoder, organ-query module, projection layers, text encoder, and BDA-specific modules remain trainable. The training objective combines paired whole-study and organ-local correspondence with cross-patient burden order:

$$\mathcal{L}_{\text{train}} = \underbrace{\mathcal{L}_{\text{global}} + \lambda_{\text{organ}}\mathcal{L}_{\text{organ}}}_{\text{correspondence}} + \underbrace{\lambda_{\text{BDA}} \frac{1}{|\mathcal{Q}_\mathcal{B}|} \sum_{q\in\mathcal{Q}_\mathcal{B}} \mathcal{L}_{\text{BDA}}^q}_{\text{cross-patient burden order}}. \tag{9}$$

Here, $\mathcal{Q}_\mathcal{B} = \{q : \mathcal{P}_q \neq \varnothing\}$ contains the whole-study and pathology-mapped organ scopes with at least one selected relation valid in the current mini-batch.

4 Experiments

4.1 Datasets and Metrics

Datasets. We evaluate Spectrum on two public datasets: CT-RATE and RAD-ChestCT. CT-RATE contains 25,692 non-contrast 3D scans from 21,304 patients, expanded to 50,188 reconstructed volumes with paired reports and abnormality labels (Hamamci et al., 2026). RAD-ChestCT is a non-contrast chest CT abnormality dataset, comprising 36,316 volumes collected between 2012 and 2017 with 83 abnormality labels (Draelos et al., 2021). Its public subset contains 3,630 scans.

Metrics. We consider four tasks: zero-shot abnormality diagnosis, frozen linear probing, image–image retrieval, and report–image retrieval. For zero-shot diagnosis evaluation, we follow the protocol of CT-CLIP Hamamci et al. (2026). These tasks assess zero-shot diagnostic prediction, linear separability, and retrieval-based representation quality. For zero-shot diagnosis, we report AUROC, ACC, F1, and precision; for linear probing, we report AUPRC and AUROC; for image–image retrieval, we report MAP@5/10/50; and for report–image retrieval, we report Recall@5/10/50/100. Reported baseline results follow their published protocols, and missing metrics are marked as "–". We additionally inspect cross-patient burden order through one pathology-specific three-study chain.

4.2 Implementation Details

The CT encoder is a 3D ResNet-18. Organ masks are grouped into lung, heart, esophagus, and aorta and used to select organ-specific evidence from the 3D feature map. A four-query global decoder and

Method	CT-RATE↑				RAD-ChestCT↑			
	AUROC	ACC	F1	Prec.	AUROC	ACC	F1	Prec.
CT-Net (Draelos et al., 2021)	60.3	58.1	63.1	23.9	54.4	54.0	58.7	28.5
CT-CLIP (Hamamci et al., 2026)	73.1	66.8	70.7	32.3	62.9	59.5	64.2	33.6
BIUD (Cao et al., 2024)	71.3	68.1	71.6	33.8	62.9	60.6	65.2	33.7
Merlin (Blankemeier et al., 2026)	72.8	67.2	70.9	33.7	64.4	61.9	66.3	34.8
COLIPRI-C† (Wald et al., 2025)	76.3	–	–	–	69.1	–	–	–
fVLM (Shui et al., 2025)	77.8	71.8	75.1	37.9	68.0	64.7	68.8	37.4
Spectrum (Ours)	**85.6**	**79.0**	**81.2**	**46.2**	**72.7**	**67.7**	**71.9**	**39.7**

Table 1: Zero-shot diagnosis. No downstream classifier or evaluation-set tuning is used. AUROC, ACC (accuracy), F1 and Prec. (precision), in %; "–": not reported; †: contrastive-only variant.

Method	CT-RATE		RAD-ChestCT	
	AUPRC	AUROC	AUPRC	AUROC
Curia (Dancette et al., 2025)	45.9	78.0	39.2	65.7
CT-CLIP	–	75.1	–	64.7
CT-FM (Pai et al., 2025)	53.5	82.1	42.4	68.5
Merlin	54.8	82.6	45.3	70.9
COLIPRI-C	55.1	82.6	46.3	71.2
Spectrum (Ours)	**64.7**	**87.5**	**60.72**	**77.97**

Table 2: Frozen-encoder linear probing. A linear classifier is fitted on frozen image features. AUPRC and AUROC in %; "–": not reported. Higher is better.

four mask-gated organ queries produce 512-dimensional visual representations, while CXR-BERT encodes whole reports and organ-specific report snippets as 768-dimensional text features. Across visual warm-up and image–text training, we set $\lambda_{\text{rank}} = \lambda_{\text{organ}} = 0.10$, $\lambda_{\text{BDA}} = 0.64$, $\lambda_{\text{dir}} = 0.50$, $\lambda_{\text{end}} = 0.25$, $\lambda_{\text{comp}} = 0.03$, and $\tau_o = 0.07$. During image–text training, we use AdamW with batch size 4 and learning rates of 10^{-5} and 10^{-4} for the text encoder and projection modules, respectively. BDA uses $\delta = \eta = s_{\min} = 0.75$, $K = 32$, at most eight eligible triplets per mini-batch, a 256-dimensional delta space, and margin $\mu = 0.10$. BDA losses are averaged equally across active whole-study and organ scopes. All experiments were conducted using NVIDIA A100 GPUs.

4.3 Main Results

Zero-shot abnormality diagnosis. Table 1 compares Spectrum with existing CT methods on CT-RATE and RAD-ChestCT. Here, zero-shot denotes fixed-prompt diagnosis without downstream adaptation, consistent with Merlin Blankemeier et al. (2026). On CT-RATE, Spectrum achieves the highest result among the compared methods for every metric, reaching 85.6 AUROC, 79.0 ACC, 81.2 F1, and 46.2 precision. Relative to fVLM, these scores improve by 7.8, 7.2, 6.1, and 8.3 points; Spectrum also surpasses the contrastive-only COLIPRI-C variant by 9.3 AUROC points. On RAD-ChestCT, Spectrum obtains 72.7 AUROC, 67.7 ACC, 71.9 F1, and 39.7 precision, exceeding the prior AUROC in the table by 3.6 points and outperforming fVLM by 4.7 AUROC points. The gains across both datasets demonstrate stronger in-domain diagnosis and external zero-shot transfer.

Frozen-encoder linear probing. Table 2 evaluates the discriminative quality of the frozen image representation under a linear classifier. On CT-RATE, Spectrum achieves 64.7 AUPRC and 87.5 AUROC, surpassing the strongest baseline by 9.6 and 4.9 points, respectively. On RAD-ChestCT, it reaches 60.72 AUPRC and 77.97 AUROC. These results show that Spectrum produces more linearly separable features for both in-domain and external abnormality classification.

Image–image retrieval. Table 3 shows that Spectrum achieves mAP@5/10/50 of 71.2/61.6/54.2. Compared with CT-CLIP, these results improve by 2.9, 4.4, and 5.3 points, respectively. The increasing margin at deeper retrieval cutoffs demonstrates that the learned image space preserves abnormality-aware neighborhood structure beyond the nearest scans. Together with the diagnosis and

Method	Img2Img MAP@5/10/50↑	Rpt2Img Recall@5/10/50/100↑
CT-Net	59.4 / 48.1 / 40.7	– / – / – / –
Merlin	62.6 / 51.3 / 43.9	1.5 / 2.7 / 7.7 / 12.7
CT-CLIP	68.3 / 57.2 / 48.9	2.9 / 5.0 / 18.0 / 28.7
– ClassFine	67.9 / 56.8 / 48.5	– / – / – / –
fVLM	49.1 / 36.8 / 26.0	0.8 / 1.5 / 4.9 / 8.3
Spectrum (Ours)	**71.2 / 61.6 / 54.2**	**18.8 / 26.3 / 51.8 / 64.3**

Table 3: CT-RATE retrieval. Image–image MAP measures abnormality-overlap retrieval, while report–image Recall evaluates paired cross-modal retrieval.

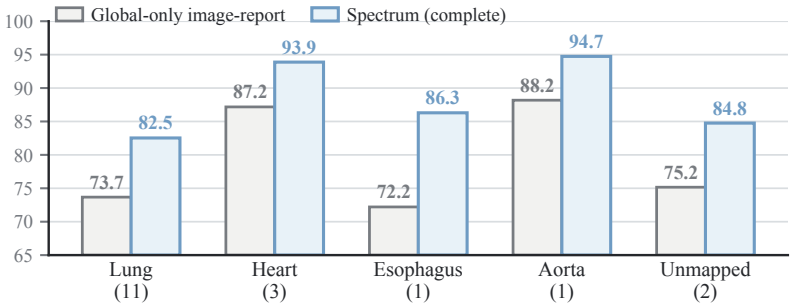

Figure 4: Global-prediction AUROC by organ group on CT-RATE. Spectrum is compared with global-only image–report alignment, a matched internal control rather than a prior method. Groups follow the pathology-to-organ map, with pathologies per group in parentheses and "unmapped" the two global-only labels.

linear-probing results, this confirms that anatomy-conditioned burden-order learning improves both category discrimination and the organization of CT representations.

Report–image retrieval. For report–image retrieval, Spectrum achieves R@5/10/50/100 of 18.8/26.3/51.8/64.3, compared with 2.9/5.0/18.0/28.7 for CT-CLIP. These correspond to gains of 15.9, 21.3, 33.8, and 35.6 points, respectively. The substantial improvements across retrieval depths demonstrate that Spectrum preserves strong report–scan correspondence while learning anatomy-conditioned pathology-burden directions. BDA therefore complements the global and organ-local correspondence objectives and strengthens the resulting cross-modal representation space.

4.4 Ablation and Analysis

BDA loss ablation. Table 4 compares the correspondence-only baseline with the cumulative addition of the four BDA losses under the same representation architecture and optimization protocol. Without BDA supervision, the model obtains 76.7 AUROC. Adding the core image–text delta-alignment loss $\mathcal{L}_{\Delta}$ raises performance to 83.9, a gain of 7.2 points. Forward–reverse discrimination $\mathcal{L}_{\text{dir}}$ and endpoint consistency $\mathcal{L}_{\text{end}}$ further improve it to 84.3 and 84.5, respectively. The complete objective reaches 85.6 AUROC. In particular, the 1.1-point gain from $\mathcal{L}_{\text{comp}}$ shows that pairwise delta alignment benefits substantially from path consistency across ordered $a \to b \to c$ triplets.

Aligner controls. Table 5 keeps the complete BDA objective fixed and changes only how the lower-burden image representation is aligned before computing its delta. Removing the higher-burden representation from the aligner input reduces AUROC from 85.6 to 84.5, while bypassing the aligner entirely yields 84.2. Thus, target context contributes 1.1 points over the target-blind variant, and the complete target-conditioned aligner contributes 1.4 points over direct source-to-target subtraction.

Organ-group behavior. Fig. 4 groups the 18 abnormalities predicted from the global representation by their pathology-to-organ assignment. Compared with global-only image–report alignment, the

L_Δ	L_{dir}	L_{end}	L_{comp}	AUROC↑
–	–	–	–	76.7
✓	–	–	–	83.9
✓	✓	–	–	84.3
✓	✓	✓	–	84.5
✓	✓	✓	✓	**85.6**

Table 4: BDA loss ablation on CT-RATE. Columns add delta alignment, forward–reverse direction, endpoint consistency and ordered-triplet composition in turn.

Aligner	Target input	Learned	AUROC↑
Identity	–	–	84.2
Target-blind	–	✓	84.5
Target-conditioned	✓	✓	**85.6**

Table 5: Visual-aligner controls on CT-RATE. With the complete BDA objective fixed, rows compare identity, target-blind, and target-conditioned alignment before computing the image delta. Target input denotes conditioning on the higher-burden representation.

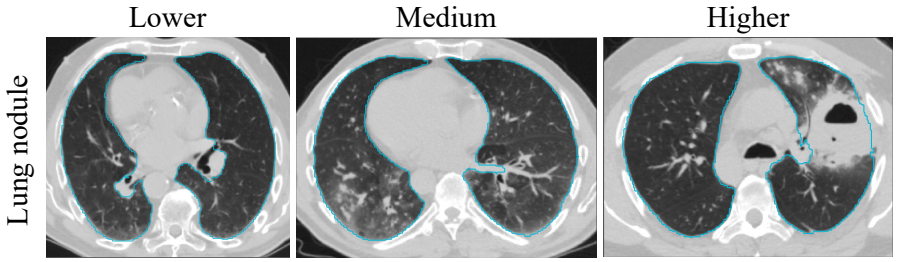

Figure 5: Model-supported burden order for lung nodule. BDA mines these three studies, one per patient, as a lower–medium–higher chain; the report size cues rise accordingly, and cyan contours delineate the lungs. All three forward–reverse similarity gaps are positive, so the trained model prefers the displayed direction to its reverse. The order is cross-patient, not longitudinal.

complete framework improves lung, heart, esophagus, aorta, and the two unmapped pathologies by 8.8, 6.7, 14.1, 6.5, and 9.6 AUROC points, raising the overall score from 76.8 to 85.6. The positive gain in every group demonstrates that the improvement is not driven by a single dominant organ or pathology family. Notably, the two unmapped abnormalities also improve, showing that the anatomy-conditioned and burden-order training signals are encoded beyond their directly mapped organ scopes in the shared whole-study representation.

Model-derived burden-order geometry. For each ordered pair, we report the forward–reverse similarity difference $s_{\text{fwd}} - s_{\text{rev}}$, averaged over the whole-study and organ scopes. A positive value indicates that the visual transition agrees more strongly with the selected report direction than with its swapped reverse. For the lung-nodule triplet in Fig. 5, the lower-to-medium, medium-to-higher, and lower-to-higher gaps are $+0.034$, $+0.011$, and $+0.060$, respectively. A least-squares one-dimensional fit jointly summarizes the three pairwise gaps; within-triplet min–max normalization yields coordinates $0.000/0.716/1.000$, recovering the lower–medium–higher order.

5 Discussion and Conclusion

Spectrum advances volumetric CT VLP from anatomy-aware correspondence to anatomy-conditioned burden direction. By mining confidence-filtered lower-to-higher relations from cross-sectional labels and reports, BDA aligns pathology-conditioned image and text deltas at the whole-study and pathology-mapped organ scopes, converting weak cross-patient cues into structured feature-space order. The complete framework raises CT-RATE AUROC from 76.8 to 85.6 over matched global-only image–report alignment and achieves 72.7 AUROC on external RAD-ChestCT; consistent gains in frozen-encoder linear probing, retrieval, and pathology-group analysis further demonstrate stronger and more transferable representations. More broadly, Spectrum moves CT VLP beyond learning only *what matches* toward encoding consistent lower-to-higher pathology-burden directions from routine cross-sectional data, without requiring longitudinal CT. Future work will investigate its utility for broader downstream tasks, including report generation and disease risk prediction.

AI USE STATEMENT

Generative AI tools were used only to assist with manuscript preparation, including language editing, shortening passages, grammar checking, LaTeX formatting, submission-compliance checks, and routine coding assistance for repetitive tasks. They were not used to generate data, design the method, or make scientific decisions. The authors reviewed every AI-assisted revision and checked all technical content against the experimental records and source files. The authors take full responsibility for the final content of this work, including any AI-assisted text or code.

ETHICS STATEMENT

This study uses the publicly available CT-RATE dataset and the released RAD-ChestCT subset. It involves no prospective participant recruitment, clinical intervention, or collection of new patient data. We used these data solely for research and in accordance with their access and usage conditions. Results are reported in aggregate, and the displayed examples contain no direct patient identifiers. Models trained on retrospective clinical data may inherit demographic, institutional, and reporting biases. The proposed method is intended for research use and requires further external validation and appropriate clinical evaluation before deployment.

REPRODUCIBILITY STATEMENT

The main paper specifies the datasets, evaluation tasks, model architecture, optimization settings, objective weights, and compute platform. The supplementary material documents the fixed pathology-to-organ map, deterministic burden scorer, relation and triplet mining procedure, thresholds, and qualitative audit. We will release the implementation code, and all evaluations use the released data cited in the paper. Together, these materials are intended to support independent reproduction of the training and evaluation protocol.

REFERENCES

Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15016–15027, 2023.

Louis Blankemeier, Ashwin Kumar, Joseph Paul Cohen, Jiaming Liu, Longchao Liu, Dave Van Veen, Syed Jamal Safdar Gardezi, Hongkun Yu, Magdalini Paschali, Zhihong Chen, et al. Merlin: a computed tomography vision–language foundation model and dataset. *Nature*, 652(8112): 1318–1328, 2026.

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

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pp. 89–96, 2005.

Weiwei Cao, Jianpeng Zhang, Yingda Xia, Tony CW Mok, Zi Li, Xianghua Ye, Le Lu, Jian Zheng, Yuxing Tang, and Ling Zhang. 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*, pp. 11238–11247, 2024.

Jieneng Chen, Yingda Xia, Jiawen Yao, Ke Yan, Jianpeng Zhang, Le Lu, Fakai Wang, Bo Zhou, Mingyan Qiu, Qihang Yu, Mingze Yuan, Wei Fang, Yuxing Tang, Minfeng Xu, Jian Zhou, Yuqian Zhao, Qifeng Wang, Xianghua Ye, Xiaoli Yin, Yu Shi, Xin Chen, Jingren Zhou, Alan Yuille,

Zaiyi Liu, and Ling Zhang. Cancerunit: Towards a single unified model for effective detection, segmentation, and diagnosis of eight major cancers using a large collection of ct scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21327–21338, 2023.

Corentin Dancette, Julien Khlaut, Antoine Saporta, Helene Philippe, Elodie Ferreres, Baptiste Callard, Théo Danielou, Léo Alberge, Léo Machado, Daniel Tordjman, et al. Curia: A multi-modal foundation model for radiology. *arXiv preprint arXiv:2509.06830*, 2025.

Rachel Lea Draelos, David Dov, Maciej A Mazurowski, Joseph Y Lo, Ricardo Henao, Geoffrey D Rubin, and Lawrence Carin. Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. *Medical image analysis*, 67:101857, 2021.

Haifan Gong, Guanqi Chen, Sishuo Liu, Yizhou Yu, and Guanbin Li. Cross-modal self-attention with multi-task pre-training for medical visual question answering. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, pp. 456–460. ACM, 2021.

Haifan Gong, Guanqi Chen, Mingzhi Mao, Zhen Li, and Guanbin Li. Vqamix: Conditional triplet mixup for medical visual question answering. *IEEE Transactions on Medical Imaging*, 41(11): 3332–3343, 2022.

Haifan Gong, Wenhao Huang, Huan Zhang, Yu Wang, Xiang Wan, Hong Shen, Guanbin Li, and Haofeng Li. Intensity confusion matters: An intensity-distance guided loss for bronchus segmentation. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2024.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, et al. Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature Biomedical Engineering*, pp. 1–19, 2026.

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

Wenhao Huang, Haifan Gong, Huan Zhang, Yu Wang, Xiang Wan, Guanbin Li, Haofeng Li, and Hong Shen. Bcnet: Bronchus classification via structure guided representation learning. *IEEE Transactions on Medical Imaging*, 44(1):489–498, 2025.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

Haoran Lai, Zihang Jiang, Qingsong Yao, Rongsheng Wang, Zhiyang He, Xiaodong Tao, Weifu Lv, Wei Wei, and Shaohua Kevin Zhou. Bridged semantic alignment for zero-shot 3d medical image diagnosis. *IEEE Journal of Biomedical and Health Informatics*, 2025.

Jingyang Lin, Yingda Xia, Jianpeng Zhang, Ke Yan, Kai Cao, Le Lu, Jiebo Luo, and Ling Zhang. Ct-glip: 3d grounded language-image pretraining with ct scans and radiology reports for full-body scenarios. *arXiv preprint arXiv:2404.15272*, 2024.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 525–536. Springer, 2023.

Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rueckert. Joint learning of localized representations from medical images and reports. In *European conference on computer vision*, pp. 685–701. Springer, 2022.

Hong Nguyen, Hoang Nguyen, Melinda Chang, Hieu Pham, Shrikanth Narayanan, and Michael Pazzani. Conpro: learning severity representation for medical images using contrastive learning and preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5105–5112, 2024.

Suraj Pai, Ibrahim Hadzic, Dennis Bontempi, Keno Bressem, Benjamin H Kann, Andriy Fedorov, Raymond H Mak, and Hugo JWL Aerts. Vision foundation models for computed tomography. *arXiv preprint arXiv:2501.09001*, 2025.

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

Zhongyi Shui, Jianpeng Zhang, Weiwei Cao, Sinuo Wang, Ruizhe Guo, Le Lu, Lin Yang, Xianghua Ye, Tingbo Liang, Qi Zhang, et al. Large-scale and fine-grained vision-language pre-training for enhanced ct image understanding. In *The Thirteenth International Conference on Learning Representations*, 2025.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature biomedical engineering*, 6(12):1399–1406, 2022.

Tassilo Wald, Ibrahim Ethem Hamamci, Yuan Gao, Sam Bond-Taylor, Harshita Sharma, Maximilian Ilse, Cynthia Lo, Olesya Melnichenko, Anton Schwaighofer, Noel CF Codella, et al. Comprehensive language-image pre-training for 3d medical image understanding. *arXiv preprint arXiv:2510.15042*, 2025.

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in neural information processing systems*, 35:33536–33549, 2022a.

Senmao Wang, Haifan Gong, Runmeng Cui, Boyao Wan, Zhonglin Hu, Haiqing Yang, Jingyang Zhou, Haiyue Jiang, and Lin Lin. Costal cartilage segmentation with topology guided deformable mamba: Method and benchmark. *Expert Systems with Applications*, 300:130085, 2026.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3876–3887, 2022b.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 21372–21383, 2023.

Zhuoyi Yang and Liyue Shen. Tempa-vlp: Temporal-aware vision-language pretraining for longitudinal exploration in chest x-ray image. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4625–4634. IEEE, 2025.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18123–18133, 2022.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *Nejm Ai*, 2(1):AIoa2400640, 2025.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine learning for healthcare conference*, pp. 2–25. PMLR, 2022.