SIA-CT: STATE-STRUCTURED ITEM ALIGNMENT FOR OPEN-VOCABULARY 3D CHEST CT REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Radiology reports describe both findings and the states that determine their meaning: where a finding occurs, whether it is present, and how severe or certain it is. Learning from these reports requires supervision that preserves these distinctions. We introduce SIA-CT, a framework for state-structured item alignment in 3D chest CT. Its central principle is that cross-patient item alignment requires agreement in observed states, whereas within-examination item separation requires evidence of distinct findings. We instantiate this principle through coordinated item separation, cross-patient alignment, and false-negative masking. A complementary regional objective aligns organ-supported visual features with report descriptions through shared transformations. Together, these objectives train a representation queried by natural language using only CT and text at inference. On 750 report-derived state questions, SIA-CT reaches 66.8% four-axis macro accuracy, exceeding a matched itemized baseline by 3.5 percentage points. Global and text-conditioned score fusion yields AUROCs of 83.3% on CT-RATE and 75.3% on RadChestCT, while text-conditioned scoring gives stronger aggregate state discrimination. These results establish observed-state relations as a useful source of supervision and show why finding recognition and state discrimination should be evaluated together.

1 INTRODUCTION

A chest CT report communicates more than a list of findings. "Small left pleural effusion," "no pleural effusion," and "possible right pleural effusion" refer to the same finding but express different clinical assertions (Figure 1). A useful CT–text representation should distinguish both the finding and the state in which it is described. Vision–language pre-training offers a scalable route to this goal by learning from paired volumes and reports (Hamamci et al., 2024a; Shui et al., 2025; Zhao et al., 2026). Yet a finding-recognition score alone leaves sensitivity to these modifiers unresolved, much as conventional image–text evaluations can overlook compositional distinctions (Yuksekgonul et al., 2023).

ItemizedCLIP takes an important step toward richer supervision by treating a report as independently meaningful items and learning a text-conditioned visual representation for each (Lyu et al., 2026). This raises a further question: *which relations between report items should a model learn?* Distinct findings within one examination call for distinguishable representations. Rephrasings, overlapping descriptions, and state variants of one finding require a different treatment. Across patients, the same finding name can denote different states, while different sentences can express the same observed assertion. Itemization exposes these relations; the learning objective must decide how to use them.

We propose to make *finding identity and observed state* the basis of item-level supervision. The key distinction is between evidence for alignment and evidence for separation. We align cross-patient items when their finding, observed attributes, and attribute values agree. Within an examination, we apply item separation to eligible distinct findings and exclude semantically dependent pairs. An omitted attribute remains unobserved, preserving the difference between missing information and an explicit clinical statement. Natural sentences provide the text embeddings; structured records determine the training relations.

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.

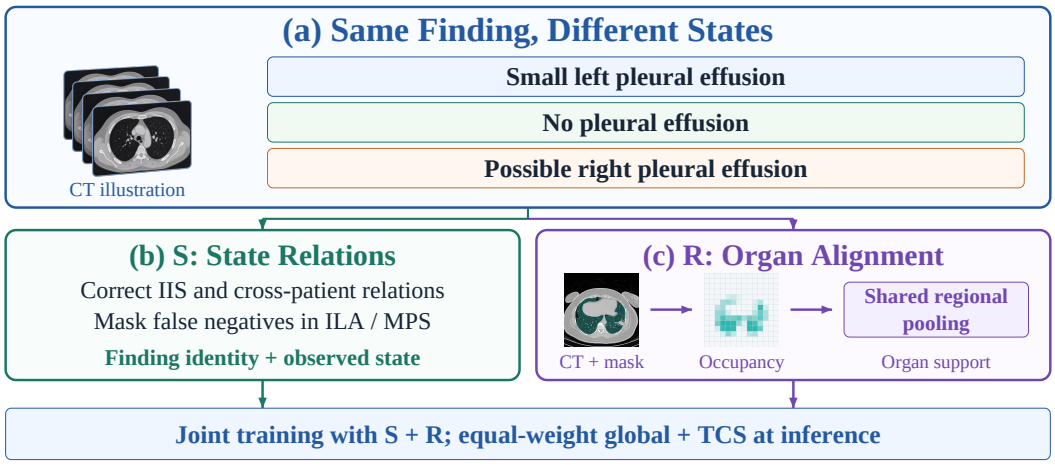

Figure 1: From finding identity to clinical assertions. State relations determine which report items align and which are eligible for separation; organ support supplies complementary regional supervision. Panel (a) is schematic; panel (c) shows a real CT slice, lung mask, and measured token occupancy.

This principle leads to SIA-CT, which combines state-structured relations (S) with organ-region alignment (R). S coordinates within-examination separation, exact-state cross-patient positives, and false-negative masking. R addresses the spatial side of the problem: it supervises organ-pooled features with descriptions of that anatomy. Shared transformations connect this regional signal to the text-conditioned representation used at inference. The two components specify complementary aspects of supervision: which items should relate, and which anatomical features should represent them.

We test this formulation first through same-finding state selection, then through zero-shot finding recognition. SIA-CT improves four-axis state accuracy from 63.3% to 66.8% under a matched text-conditioned readout, with gains from both S and R. The same checkpoint achieves 83.3% and 75.3% disease AUROC on CT-RATE and RadChestCT with global/text-conditioned fusion. The differing effects of fusion on these tasks further motivate evaluating state discrimination directly. Our contributions are:

- A state-structured formulation of item supervision that distinguishes exact observed-state agreement from eligibility for item separation.
- An alignment framework that combines these semantic relations with organ-supported multi-positive supervision through shared visual transformations.
- An evaluation of same-finding state discrimination alongside disease recognition, revealing complementary contributions from S and R and different behavior across readouts.

2 Related Work

Fine-grained language supervision. Report-supervised learning, global–local alignment, and cross-attention have advanced medical image representations (Zhang et al., 2022; Huang et al., 2021; Lai et al., 2024). DreamLIP and FLAIR learn local or text-conditioned features from detailed captions (Zheng et al., 2024; Xiao et al., 2025). ItemizedCLIP formulates item independence and representation completeness for independently meaningful text items (Lyu et al., 2026). We build on its architecture and objectives, extending the treatment of item relations to reports with overlapping findings and partially observed clinical states.

Semantic relations and compositional evaluation. Supervised and debiased contrastive learning study semantic positives and false negatives; MedCLIP applies semantic matching to medical image–text learning (Khosla et al., 2020; Chuang et al., 2020; Wang et al., 2022). CheXpert ex-

plicitly represents report uncertainty (Irvin et al., 2019). ARO demonstrates the value of evaluating attributes and relations separately from conventional image–text performance (Yuksekgonul et al., 2023). Our training relations require agreement in both observed values and observation masks, and our evaluation holds finding identity fixed while varying clinical state.

Volumetric and anatomical alignment. CT-CLIP, BIUD, HLIP, and Merlin extend representation learning to volumetric medical data (Hamamci et al., 2024a; Cao et al., 2024; Zhao et al., 2026; Blankemeier et al., 2026). Particularly relevant is fVLM, which combines anatomy-level alignment with disease-aware false-negative correction (Shui et al., 2025). Our formulation operates on observed item states and coordinates within-examination separation with cross-patient supervision. Regional semantics also motivate locality alignment in general VLMs and anatomical supervision in Chest-OMDL, ARC-CT, and OKA-CT (Covert et al., 2025; Bai et al., 2026; Isik et al., 2026; You et al., 2026). We use organ-supported multi-positive alignment through shared text-conditioned transformations. Appendix F gives further context.

3 METHOD

3.1 ITEMIZED REPRESENTATION LEARNING

Our starting point is ItemizedCLIP, including its backbone, report itemization, multiple text views, and alignment objectives (Lyu et al., 2026). SIA-CT changes the supervision relations through S and adds anatomical supervision through R. Let X_i be a CT volume, $\mathcal{T}_i$ its report items, and p_i its patient identifier. The visual encoder produces a global feature v_i^g and tokens $V_i = \{v_{in}\}_{n=1}^{N}$. The frozen text encoder maps a natural item t to $e(t)$. Text-conditioned summarization (TCS) reads visual tokens using the text query:

$$u_i(t) = \mathrm{TCS}(e(t), V_i), \qquad s_{\mathrm{TCS}}(X_i, t) = \cos(u_i(t), e(t)). \tag{1}$$

Itemized local alignment (ILA) matches text-conditioned features to report items; multi-positive SigLIP (MPS) supervises the global feature with those items (Zhai et al., 2023). Inter-item separation (IIS) encourages distinguishable item representations. S revises the relations used by these objectives, and R supplies an additional organ-level objective (Figure 2).

3.2 LEARNING RELATIONS FROM OBSERVED STATES (S)

Clinical reports provide partial descriptions: absence of an attribute in text is absence of supervision for that attribute. S uses this observation to construct two relations. Exact observed-state agreement admits cross-patient alignment, while semantic independence determines within-examination separation. These relations then govern the corresponding positive and negative terms.

Representing observed state. Each eligible item has a record $z = (f, \mathbf{r}, \mathbf{m})$: f is the finding identity, $\mathbf{r} = (a, p, l, s, u)$ stores anatomy, polarity, laterality, severity, and uncertainty, and $\mathbf{m} \in \{0, 1\}^5$ marks the attributes supported by the item. For example, omitted laterality gives $m_l = 0$, leaving the side unspecified. The text encoder receives the natural sentence; z supplies its supervision relations.

Exact observed-state agreement. For two eligible records, we require the same finding, the same observed attributes, and the same values on those attributes:

$$z_a \equiv z_b \iff f_a = f_b, \quad \mathbf{m}_a = \mathbf{m}_b, \quad \mathbf{r}_a|_{\mathbf{m}_a=1} = \mathbf{r}_b|_{\mathbf{m}_b=1}. \tag{2}$$

The observation mask is part of the assertion. Two descriptions of left pleural effusion can agree despite different wording, provided their remaining observed fields match. Changing left to right breaks value agreement; omitting the side breaks mask agreement. Exact matching thus preserves the level of detail actually supplied by the report (Table 1).

Separating distinct findings. Within an examination, item separation is useful when the items describe distinct findings. Synonyms, hierarchical descriptions, and overlapping statements can instead share visual evidence. We therefore condition IIS on the semantic relation between items. Let $g(t)$ denote a consolidated finding group and $C(t_j, t_k) = 1$ mark a synonym, hierarchy edge,

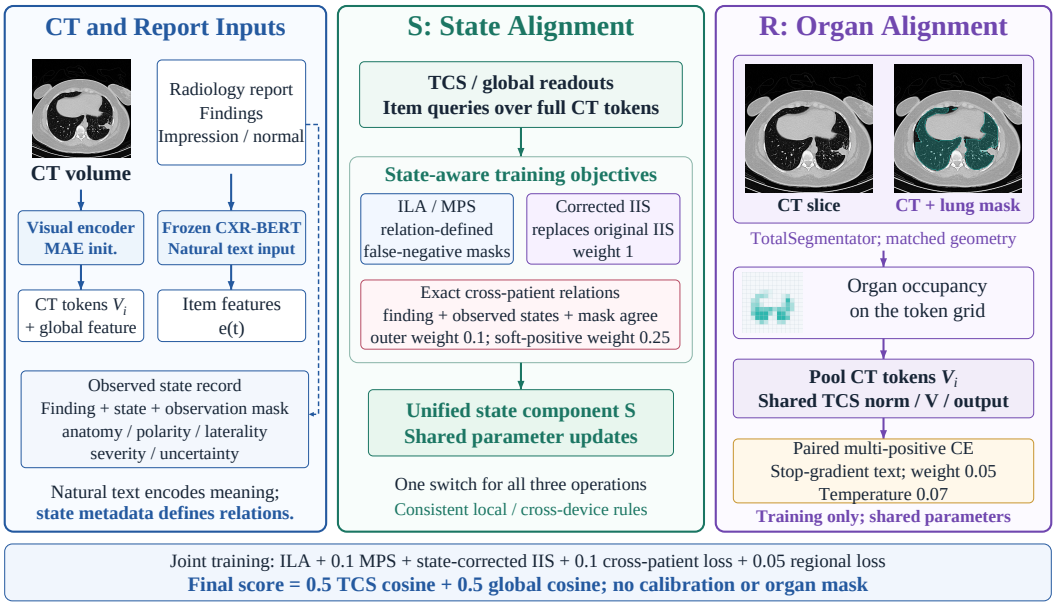

Figure 2: SIA-CT learns from semantic relations and anatomical support. S coordinates item separation, cross-patient positives, and alignment masks. R aligns organ-pooled features with regional descriptions through shared TCS transformations. The CT/mask grid depicts measured anatomical occupancy. Both objectives train the representation queried by TCS/global readouts at inference.

coupled finding, same-finding state variant, attribution ambiguity, or semantic conflict. S restricts separation to

$$\Omega_{\text{sep}}(i) = \{(j,k) : j \neq k,\ g(t_j) \neq g(t_k),\ C(t_j,t_k) = 0\}. \tag{3}$$

Applying the original IIS operator to $\Omega_{\text{sep}}(i)$ gives $\mathcal{L}^S_{\text{IIS}}$, retaining its pair normalization and scale. Same-finding state variants receive no IIS separation term: shared finding evidence makes them ineligible for this independence objective. Their eligibility for cross-patient alignment is determined separately by Equation 2.

Record-pair condition	Within-exam IIS	Cross-patient relation
Same finding, mask, and observed values	Masked	Soft positive
Same finding; left versus right	Masked	Unknown
Same finding; side observed versus omitted	Masked	Unknown
Distinct finding groups with $C = 0$	Eligible	Unknown

Table 1: Illustrative S decisions for eligible records. Unlisted attributes are held equal. Columns refer to separate pairing contexts: IIS operates within an examination; exact-state positives require different patients. Unknown pairs are masked in the cross-patient relation objective.

Cross-patient supervision. Let $\mathcal{E}$ be the eligible item records in the effective training batch and let p_a identify the patient associated with record z_a. The accepted cross-patient relation set is

$$\mathcal{P}_{\text{exact}} = \{(a,b) \in \mathcal{E}^2 : p_a \neq p_b,\ z_a \equiv z_b\}. \tag{4}$$

The resulting objective $\mathcal{L}^S_{\text{cross}}$ extends supervision beyond the original image–report pairing, assigning accepted relations soft-positive weight 0.25. Remaining pairs are unknown and excluded from this auxiliary objective. This asymmetry follows from partial observation: an exact match supplies evidence for alignment, whereas a mismatch can arise from either a changed state or an omitted attribute.

False-negative masking. The same accepted relations also identify false negatives in ILA/MPS. A relation promoted by the cross-patient objective should not simultaneously be repelled as an unpaired negative in these alignment objectives. Let M^-_{ab} indicate a candidate negative in a base

alignment loss and F_{ab} indicate that it is identified as a false negative by the accepted relation construction. The revised negative mask is

$$\widetilde{M}^{-}_{ab} = M^{-}_{ab}(1 - F_{ab}). \tag{5}$$

We apply this mask before loss reduction while retaining paired positives and the original validity rules for other negatives. Consequently, exclusion from $\mathcal{L}^{S}_{\text{cross}}$ alone leaves a pair's ILA/MPS status unchanged. Appendix A records the implementation settings.

Together, these operations express one relation policy across item separation and image–text alignment. Appendix G illustrates how it acts on real report records.

3.3 Aligning Anatomical Support with Report Items (R)

S specifies semantic relations between items. R supplies a spatial constraint: descriptions associated with an organ supervise features pooled from that organ. Sharing the pooling transformations with TCS allows this signal to train the representation later queried by natural text.

Organ priors and token occupancy. Organ labels are generated with TotalSegmentator (Wasserthal et al., 2023); non-thoracic reconstructions are excluded. The organ label map and CT undergo identical geometric transformations. For organ support o and token footprint $\mathcal{P}_n$, the pooling weight is its occupied fraction:

$$a_{ion} = \frac{1}{|\mathcal{P}_n|} \sum_{x \in \mathcal{P}_n} \mathbf{1}[M_i(x) \in o]. \tag{6}$$

Fractional occupancy preserves anatomical support for tokens crossing organ boundaries.

Shared regional pooling. Let ψ_V denote the normalization and value transformation used by TCS and ψ_O its output projection. Regional pooling can be written schematically as

$$h_{io} = \operatorname{norm}\left[\psi_O\left(\frac{\sum_n a_{ion}\, \psi_V(v_{in})}{\sum_n a_{ion}}\right)\right], \qquad \sum_n a_{ion} > 0. \tag{7}$$

The head-wise transformations are shared with TCS, while organ occupancy supplies the pooling weights. Each h_{io} represents an anatomical region shared by its report findings; TCS subsequently provides the query-specific readout.

Candidate and positive sets. Candidates for (i, o) are restricted to descriptions with the same organ support. Multiple descriptions from the same examination and region are positives. Completely repeated text from a different patient is also positive within that candidate support. Other examinations from the same patient are excluded. These rules distinguish anatomical support from patient identity and prevent repeated positive descriptions from becoming negatives merely because they appear in another patient.

Explicitly negated report items also describe the corresponding image. Thus, "no pleural effusion" can be a positive image–text pairing: its clinical polarity is part of the content being aligned (Figure 3).

Paired regional objective. For candidate t_b, the regional logit is

$$\ell_{io,b} = \frac{\cos(h_{io}, \operatorname{sg}[e(t_b)])}{\tau_R}, \qquad \tau_R = 0.07. \tag{8}$$

The objective $\mathcal{L}_R$ applies multi-positive cross-entropy over these candidate and positive sets. Stop-gradient text features direct its updates through the visual tokens and shared transformations. The resulting organ-level supervision complements the item relations in S.

3.4 Joint Objective and Inference

Writing $\mathcal{L}^{S}_{\text{ILA}}$ and $\mathcal{L}^{S}_{\text{MPS}}$ for the inherited objectives with S-defined false-negative masks, the complete objective is

$$\mathcal{L}_{\text{SIA}} = \mathcal{L}^{S}_{\text{ILA}} + 0.1\mathcal{L}^{S}_{\text{MPS}} + \mathcal{L}^{S}_{\text{IIS}} + 0.1\mathcal{L}^{S}_{\text{cross}} + 0.05\mathcal{L}_R. \tag{9}$$

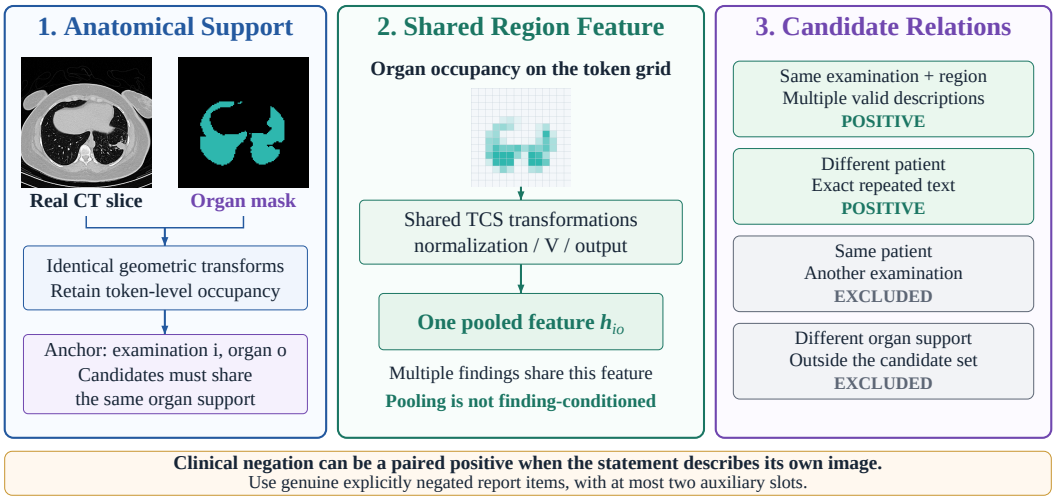

Figure 3: Regional supervision with the paired CT/mask from Figure 2. Candidates share organ support. Same-examination regional descriptions and exact text repetitions across patients are positives; other examinations of the same patient are excluded. Clinical negation can be a paired positive. Several findings share one organ feature.

The first three terms provide itemized local alignment, global alignment, and item separation; the final two add exact-state cross-patient and regional supervision. Baseline + S omits $\mathcal{L}_R$, while Baseline uses the original ILA/MPS/IIS objectives. At inference, the model scores any natural-language query through TCS or the global feature. For disease recognition we average their raw cosine similarities:

$$s_{\text{final}}(X_i, t_c) = \tfrac{1}{2} s_{\text{TCS}}(X_i, t_c) + \tfrac{1}{2} s_{\text{global}}(X_i, t_c). \tag{10}$$

The same fixed weights apply to every CT and finding. State evaluation reports both TCS and fusion; component comparisons hold the TCS readout fixed. All inference routes require only CT and text, since S and R act through training supervision.

4 Experiments

4.1 Experimental Setup

We evaluate two aspects of the learned representation: sensitivity to states within a finding, and recognition across findings. Controlled component comparisons use a shared backbone and readout; comparison with published models provides broader context.

Tasks and metrics. State selection uses 750 report-derived questions from CT-RATE validation scans. Each question holds the finding fixed and asks the model to rank descriptions that differ in state. We report accuracy for four state axes and a conflict-selection task (Section 4.2). Disease recognition uses 3,039 CT-RATE scans with 16 findings and 3,630 external RadChestCT scans with 14 findings, following the splits used by CT-CLIP, HLIP, and ItemizedCLIP (Hamamci et al., 2024a; Draelos et al., 2021; Zhao et al., 2026; Lyu et al., 2026). We report AUROC and average precision (AP), alongside accuracy (Acc), weighted F1 (wF1), recall, and precision. Threshold-dependent metrics use the same prior-study threshold-selection procedure across local models (Shui et al., 2025; Lyu et al., 2026). Appendix C specifies the finding lists.

Training and comparisons. Baseline, Baseline + S, and SIA-CT share report itemization, MAE-initialized HLIP-style ViT-B, and a frozen CXR-BERT text encoder (He et al., 2022; Zhao et al., 2026; Boecking et al., 2022). We compare epoch-60 checkpoints with common prompts and TCS; readout ablations use the same SIA-CT checkpoint. Published-model disease results are reproduced from ItemizedCLIP and labeled separately. Appendix A gives input geometry, optimization, and component settings.

Model	Readout	Pol.	Lat.	Sev.	Unc.	Four-axis mean	Conflict R@1	Five-task mean
ItemizedCLIP	Official	58.0	64.0	72.0	64.0	64.5	**54.7**	62.5
HLIP	Official	57.3	66.0	58.7	**70.0**	63.0	48.0	60.0
CT-CLIP	Official	47.3	43.3	16.7	65.3	43.2	43.3	43.2
Baseline	TCS	56.0	63.3	76.7	57.3	63.3	51.3	60.9
Baseline + S	TCS	**62.0**	63.3	78.0	58.0	65.3	53.3	62.9
SIA-CT	TCS	**62.0**	65.3	**79.3**	60.7	**66.8**	52.7	**64.0**
SIA-CT	Global + TCS	58.0	**69.3**	76.0	60.7	66.0	46.7	62.1

Table 2: State readout on 750 questions, with 150 per task (%). Pol., Lat., Sev., and Unc. denote polarity, laterality, severity, and uncertainty. Official denotes each external model's official evaluation scheme. Local checkpoints are from epoch 60; Global + TCS uses equal weights. Five-task mean equals all-question accuracy before rounding.

	CT-RATE		RadChestCT	
Model	AUROC	wF1	AUROC	wF1
CT-CLIP	73.3	70.8	63.3	64.7
BIUD	71.3	71.6	62.9	65.2
Merlin	72.8	70.9	64.4	66.3
fVLM	77.8	75.1	68.0	68.8
HLIP-RA	77.7	74.7	72.3	72.1
HLIP-SA	78.7	75.5	71.7	71.4
ItemizedCLIP	83.2	**78.9**	74.7	73.2
SIA-CT	**83.3**	78.8	**75.3**	**73.6**

Table 3: Final model and published results (%). Comparison values are from the chest CT results reported by ItemizedCLIP (Lyu et al., 2026). All metrics are shown to one decimal place. Our epoch-60 SIA-CT model uses equal-weight global and TCS readouts. Bold marks the highest reported value per column.

4.2 Discriminating States within the Same Finding

The state evaluation directly tests the distinction motivating SIA-CT. Its five tasks contain 150 questions each: polarity, laterality, severity, uncertainty, and selection among conflicting state descriptions. Models rank identical candidate sets by image–text similarity. Appendix B details construction, semantic filtering, and the identifier-overlap check.

SIA-CT with TCS reaches 66.8% four-axis macro accuracy and 64.0% accuracy over all questions, the highest aggregate results in Table 2. The gains over Baseline are 3.5 and 3.1 percentage points, respectively; compared with ItemizedCLIP, they are 2.3 and 1.5 points. All four axes improve over Baseline, including a polarity increase from 56.0% to 62.0%. Holding finding identity fixed therefore exposes a benefit of state-structured supervision that a disease-level label does not measure. The axis results also locate the remaining difficulty: HLIP leads on uncertainty, and ItemizedCLIP leads on conflict selection.

4.3 Zero-Shot Finding Recognition

SIA-CT also achieves strong finding recognition, with AUROCs of 83.3% on CT-RATE and 75.3% on RadChestCT (Table 3). These are the highest listed values, exceeding the published ItemizedCLIP results by 0.1 and 0.6 points. External-set wF1 increases from 73.2% to 73.6%, while CT-RATE wF1 is 78.8% versus 78.9%. The main empirical gain is thus in same-finding state discrimination, accompanied by competitive disease recognition and a modest improvement on the external dataset. Per-finding results appear in Appendix C.

Model	AUROC	AP	Acc	wF1	Recall	Precision	Spec.
CT-RATE							
Baseline	83.1	55.1	**76.8**	**79.1**	77.3	**43.4**	**76.5**
Baseline + S	**83.2**	55.3	76.1	78.6	**78.5**	42.7	75.5
SIA-CT	**83.2**	**55.5**	76.1	78.6	78.1	42.8	75.9
RadChestCT							
Baseline	74.3	49.4	69.0	72.5	68.7	41.1	68.9
Baseline + S	74.4	49.5	69.2	72.7	**69.5**	40.9	68.1
SIA-CT	**74.8**	**49.9**	**69.3**	**72.8**	68.4	**41.3**	**69.7**

Table 4: Epoch-60 component ablations using TCS and a common prompt for all models (%). CT-RATE uses 16 findings and RadChestCT uses 14. Values and changes are rounded to one decimal place after calculation. Bold indicates the best displayed value within each dataset. Recall is sensitivity; Acc is ordinary accuracy.

Readout	AUROC	AP	Acc	wF1	Recall	Precision
CT-RATE (16 classes)						
TCS	83.2	55.5	76.1	78.6	78.1	**42.8**
Global + TCS	**83.3**	**55.8**	**76.3**	**78.8**	**78.2**	42.6
RadChestCT (14 classes)						
TCS	74.8	49.9	69.3	72.8	68.4	41.3
Global + TCS	**75.3**	**50.8**	**70.3**	**73.6**	**68.8**	**41.8**

Table 5: Readout ablation for SIA-CT at epoch 60 (%). Global + TCS averages raw cosine scores with fixed weights of 0.5 (Equation 10). Metrics are shown to one decimal place; changes are computed before rounding. Bold marks the higher value for each metric. Recall is sensitivity; Acc is ordinary accuracy.

5 Analysis

5.1 How State Relations and Anatomical Support Contribute

Tables 2 and 4 connect the supervision design to state and disease performance. All three local models use epoch-60 checkpoints and TCS, with common prompts for disease recognition.

State relations improve assertion discrimination. Adding S raises four-axis state accuracy from 63.3% to 65.3% and overall accuracy from 60.9% to 62.9%, with the largest gain in polarity (+6.0 points). Disease AUROC and AP also increase on both datasets. The larger change in state accuracy is consistent with S's purpose: refining the supervision attached to each assertion, beyond its finding identity.

Anatomical support adds complementary supervision. Adding R to S increases four-axis state accuracy by a further 1.5 points, reaching 66.8%, and improves overall accuracy from 62.9% to 64.0%. It also raises AP by 0.3 points on CT-RATE and 0.4 points on RadChestCT. Thus, supervising shared features within organ support adds value to the semantic relation policy. The threshold metrics characterize the accompanying trade-off: CT-RATE Acc and wF1 are lower than Baseline, while both are higher on RadChestCT.

5.2 Finding Recognition and State Selection Favor Different Readouts

Global/TCS fusion improves disease AUROC and AP on both datasets; on RadChestCT, they rise from 74.8% to 75.3% and from 49.9% to 50.8%, respectively (Table 5). With the representation fixed, the two readouts provide complementary information for finding recognition.

State selection behaves differently. Fusion raises laterality accuracy to 69.3%, but lowers conflict R@1 from 52.7% to 46.7% and overall state accuracy from 64.0% to 62.1% (Table 2). This divergence is informative: a readout that improves disease recognition can weaken discrimination

between assertions about the same finding. We therefore use TCS for aggregate state comparisons and report fusion for disease recognition, retaining both state results explicitly.

5.3 Per-Finding Performance and Prompt Sensitivity

Per-finding results expose the distribution of gains rather than only their macro average (Appendix C). We also evaluate a shared expanded text pool: the full model retains higher AP than Baseline on both datasets under both prompt settings, although expansion reduces absolute performance (Appendix D).

6 Discussion and Limitations

Relations are part of the supervision. The results suggest that the usefulness of a report item depends on how it relates to other items as well as on its content. S encodes observed-state agreement and semantic dependence in the training objective; R adds anatomical support to the same representation. Their combined benefit is clearest when the evaluation varies state while keeping finding identity fixed. The readout results reinforce this distinction: disease recognition and assertion discrimination capture different aspects of a CT–text representation.

Scope and open questions. The state evaluation measures agreement with parsed reports and generated candidates. Its annotations inherit report omissions and parser errors; radiologist-adjudicated image assessment remains an important next step. Organ priors provide coarse anatomical support shared by several findings, leaving finding-specific localization to future evaluation. Our ablations measure S jointly and the incremental contribution of R given S. Separating S's operations, evaluating R alone, and testing multiple training seeds with confidence intervals would clarify their individual effects and variability. RadChestCT supplies the external disease evaluation; broader state evaluation across institutions would test transfer of the central result.

7 Conclusion

SIA-CT treats observed clinical state as part of the supervision relation between report items. Exact-state alignment and selective item separation organize semantic supervision, while organ-supported alignment trains complementary regional features. Together they improve same-finding state discrimination and retain strong zero-shot disease recognition. The contrast between state and disease readouts highlights a broader requirement for open-vocabulary CT learning: evaluating which finding a representation recognizes and which assertion about that finding it supports.

AI Use Statement

We used generative AI tools to assist with writing and language polishing, retrieve and identify related literature, and assist in constructing report-derived synthetic textual data for the state-readout evaluation. For data construction, a multi-agent workflow combined programmatic extraction with multiple rounds of validation to structure findings from existing CT reports; candidate descriptions were then generated from these structured records (Appendix B). The authors take responsibility for the final manuscript, including its claims, references, and AI-assisted text and artifacts.

Ethics Statement

This study concerns representation learning from medical images and reports, not autonomous diagnosis. Report-derived states and organ priors may propagate annotation, documentation, and population biases. Use of datasets and derived annotations must follow their access, licensing, and privacy requirements. Clinical use requires independent validation.

REFERENCES

Xuguang Bai, Mingxuan Liu, Yifei Chen, Hongjia Yang, and Qiyuan Tian. Chest-OMDL: Organ-specific multidisease detection and localization in chest computed tomography using weakly supervised deep learning from free-text radiology report. In *Proceedings of The 8th International Conference on Medical Imaging with Deep Learning*, volume 301 of *Proceedings of Machine Learning Research*, pp. 59–81, 2026. URL https://proceedings.mlr.press/v301/bai26a.html.

Louis Blankemeier, Ashwin Kumar, Joseph Paul Cohen, Jiaming Liu, Longchao Liu, Dave Van Veen, Syed Jamal Safdar Gardezi, Hongkun Yu, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Robbie Holland, Cesar Truyts, Christian Bluethgen, et al. Merlin: A computed tomography vision–language foundation model and dataset. *Nature*, 652(8112): 1318–1328, 2026. doi: 10.1038/s41586-026-10181-8.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the most of text semantics to improve biomedical vision–language processing. In *Computer Vision – ECCV 2022*, pp. 1–21. Springer Nature Switzerland, 2022. doi: 10.1007/978-3-031-20059-5_1.

Weiwei Cao, Jianpeng Zhang, Yingda Xia, Tony C. W. 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. doi: 10.1109/CVPR52733.2024.01068.

Sihong Chen, Kai Ma, and Yefeng Zheng. Med3D: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8765–8775, 2020.

Ian Covert, Tony Sun, James Zou, and Tatsunori Hashimoto. Locality alignment improves vision-language models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ce6326ac4794bb04d5eb16f597446baf-Abstract-Conference.html.

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. doi: 10.1016/j.media.2020.101857.

Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Omer Faruk Durugol, Bastian Wittmann, Tamaz Amiranashvili, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, Abdullah Alanbay, Anjany Sekuboyina, Berkan Lafci, Christian Bluethgen, Mehmet Kemal Ozdemir, and Bjoern Menze. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834*, 2024a.

Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. CT2Rep: Automated radiology report generation for 3d medical imaging. In *Medical Image Computing and Computer Assisted Intervention*, pp. 476–486. Springer, 2024b.

Ibrahim Ethem Hamamci, Sezgin Er, Anjany Sekuboyina, Enis Simsar, Alperen Tezcan, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Furkan Almas, Irem Dogan, Muhammed Furkan Dasdelen, et al. GenerateCT: Text-conditional generation of 3d chest ct volumes. In *Computer Vision – ECCV 2024*, pp. 126–143. Springer Nature Switzerland, 2024c. doi: 10.1007/978-3-031-72986-7_8.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

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.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 590–597, 2019.

Huseyin Umut Isik, Mehmet Alp Ozaydin, Sila Kurugol, and Seyda Ertekin. ARC-CT: Anatomy-routed contrastive vision-language learning for 3d chest ct. *arXiv preprint arXiv:2608.28455*, 2026.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 2021.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020.

Haoran Lai, Qingsong Yao, Zihang Jiang, Rongsheng Wang, Zhiyang He, Xiaodong Tao, and S. Kevin Zhou. CARZero: Cross-attention alignment for radiology zero-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11137–11146, 2024. doi: 10.1109/CVPR52733.2024.01059.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42: 60–88, 2017.

Yiwei Lyu, Chenhui Zhao, Soumyanil Banerjee, Shixuan Liu, Akshay Rao, Akhil Kondepudi, Honglak Lee, and Todd C. Hollon. Learning complete and explainable visual representations from itemized text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21110–21120, 2026.

Jonggwon Park, Kyoyun Choi, Byungmu Yoon, Hong Geun Cho, and Bumcheol Hwang. RadZero3D: Bridging self-supervised video models and medical vision-language alignment for zero-shot chest ct interpretation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 6801–6808, 2025. doi: 10.1109/ICCVW69036.2025.00702.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Zhongyi Shui, Jianpeng Zhang, Weiwei Cao, Sinuo Wang, Ruizhe Guo, Le Lu, Lin Yang, Xianghua Ye, Tingbo Liang, Qi Zhang, and Ling Zhang. Large-scale and fine-grained vision-language pre-training for enhanced ct image understanding. In *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. doi: 10.1038/s41551-022-00936-9.

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, 2022. doi: 10.18653/v1/2022.emnlp-main.256.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Thomas Heye, Daniel Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023. doi: 10.1148/ryai.230024.

Rui Xiao, Sanghwan Kim, Mariana-Iuliana Georgescu, Zeynep Akata, and Stephan Alaniz. FLAIR: VLM with fine-grained language-informed image representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24884–24894, 2025.

Guoliang You, Hongming Li, Yuanwang Zhang, and Yong Fan. Learning anatomy-grounded ct vision-language representations with organ-hierarchical report knowledge. *arXiv preprint arXiv:2607.10953*, 2026.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.01936.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

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 *Proceedings of Machine Learning for Healthcare*, volume 182 of *Proceedings of Machine Learning Research*, pp. 2–25, 2022.

Chenhui Zhao, Yiwei Lyu, Asadur Chowdury, Edward Harake, Akhil Kondepudi, Akshay Rao, Xinhai Hou, Honglak Lee, and Todd C. Hollon. Towards scalable language-image pre-training for 3d medical imaging. *Transactions on Machine Learning Research*, 2026.

Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. DreamLIP: Language-image pre-training with long captions. *arXiv preprint arXiv:2403.17007*, 2024. URL https://arxiv.org/abs/2403.17007.

Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B. Gotway, and Jianming Liang. Models genesis: Generic autodidactic models for 3d medical image analysis. In *Medical Image Computing and Computer Assisted Intervention*, pp. 384–393. Springer, 2019.

A Implementation and Training Details

Code availability. The anonymous supplementary archive includes the core S and paired R implementations in `code/`, with interfaces, configuration, and synthetic tests documented in `code/README.md`. This component release does not include the full training pipeline, datasets, or model weights.

CT inputs are padded/cropped to $112 \times 336 \times 336$; $8 \times 24 \times 24$ patches yield 14^3 local tokens. The frozen CXR-BERT specialized encoder uses training context length 256. Baseline is trained through epoch 60 and the augmented models through epoch 70; epoch 60 is the latest common checkpoint. The RadChestCT main evaluation retains coronary-artery calcification and excludes the additional arterial-wall query. Component ablations also report specificity. Equal-weight inference fusion uses raw cosine scores without additional calibration.

The S and S + R runs use four GPUs, batch size 16 per GPU, and eight forward accumulations, giving an effective study batch of 512. AdamW uses learning rate 10^{-4}, weight decay 0.5, $\beta = (0.9, 0.98)$, and $\epsilon = 10^{-6}$. A 100-step linear warm-up precedes cosine decay. Mixed precision, gradient checkpointing, at most ten finding groups per study, and a TCS token keep rate of 0.95 are retained across the controlled runs.

Setting	Configuration
State-corrected IIS	Replaces original IIS; weight 1
Cross-patient relations	Exact mode; outer weight 0.1; soft-positive weight 0.25
False-negative masks	ILA/MPS; consistent local/cross-device rules
Organ-label source	TotalSegmentator; non-thoracic reconstructions excluded
Regional loss	Paired multi-positive CE; weight 0.05; temperature 0.07
Regional transformations	Shared TCS normalization, value and output projections
Regional text	Stop-gradient
Auxiliary negated items	Genuine report statements; at most two slots
Additional regional losses	Relative and localization objectives disabled
Inference	Equal-weight global + TCS for final results; TCS for component ablations; no organ mask or regional head

Table 6: Full SIA-CT configuration. S is a joint switch for its three operations.

B Construction of the State-Readout Evaluation

Evaluation set and source reports. The evaluation contains 750 questions constructed from CT-RATE validation reports, with 150 questions for each of five tasks: polarity, laterality, severity, uncertainty, and conflict selection. Each question pairs a CT scan with candidate descriptions of the same finding. The target is the description supported by the report. Balancing applies to tasks; multiple questions can come from the same scan, so 750 denotes questions rather than distinct scans or patients. Table 2 reports results on this set.

Structured report representation. Report statements are converted into structured finding records through a multi-agent workflow combining programmatic extraction with multiple rounds of validation. Each structured item records finding, anatomy, polarity, laterality, severity, uncertainty, and its source sentence. Candidate sentences are generated from these fields. Scan and patient/examination identifiers retain the link to the source report. The generator selects entries marked `test` in the structured metadata for the CT-RATE validation source. The available patient and examination identifiers were checked against the training manifest, with no overlap found within those identifier namespaces.

State-specific candidate construction. Each task keeps finding identity fixed and targets a particular distinction:

- **Polarity:** contrast affirmative and negated descriptions, with a possible-presence alternative when generated and retained by filtering. This task can therefore involve both presence and certainty, rather than a uniformly binary polarity reversal.

- **Laterality:** exchange explicit left and right values. Items whose anatomy name already contains left/right lung are excluded to avoid retaining a conflicting side cue in the anatomy text.
- **Severity:** replace the severity expression using predefined contrasts, including small/large, mild/severe, and minimal/severe.
- **Uncertainty:** contrast definite and uncertain descriptions, including possible, suspicious, and cannot-exclude expressions.
- **Conflict selection:** combine a report-supported description with state-modified alternatives from the existing candidate pool. Alternatives in one question can target different state fields while referring to the same finding.

The first four tasks organize evaluation by the intended state distinction. Polarity and uncertainty can overlap through possible-presence descriptions. Conflict selection tests competition among alternative assertions about one finding.

Filtering and balanced selection. Candidate texts are normalized and deduplicated. The target must be supported by the scan's structured report, and any distractor also supported by that report is removed. A question is retained when the target and at least one distinct alternative remain. Within each task, eligible questions are grouped by finding, shuffled within groups, and selected in round-robin order across findings until the 150-question quota is reached. This limits domination by frequent findings without imposing equal counts per finding. Candidate order is shuffled and the answer index updated accordingly.

The five task quotas are applied during dataset generation, yielding the final 5×150 questions. The saved records contain the task, finding, source sentence, candidate texts and states, answer index, and source identifiers. Candidate counts can vary after construction and filtering.

Scoring and aggregation. All models receive the same candidate sets. Given CT X_q and candidates $\{t_{qj}\}_{j=1}^{K_q}$, prediction selects the highest image–text similarity, $\hat{y}_q = \arg\max_j s(X_q, t_{qj})$, and is correct when $\hat{y}_q$ equals the report-derived answer index. No classification threshold is required. ItemizedCLIP, HLIP, and CT-CLIP use their respective official evaluation schemes. Local models use epoch-60 TCS; the fixed equal-weight global/TCS readout is reported separately for SIA-CT.

We compute question-level accuracy separately for polarity, laterality, severity, and uncertainty. Their unweighted average is the four-axis mean. Conflict R@1 is the top-one accuracy of the fifth task. Because each task contains 150 questions, the five-task mean equals accuracy over all 750 questions before rounding. These are question-level aggregates, without patient weighting. Variable candidate counts imply that a single binary-chance reference does not apply to every task.

Interpretation. The evaluation measures agreement with report-derived assertions, including the uncertainty expressed in the report. Filtering uses structured report information; image-level adjudication by radiologists is not part of this protocol. Report omissions, parsing errors, and coexisting findings can leave ambiguous candidates. These annotation limitations apply when interpreting the state-readout results.

C PER-FINDING DISEASE RECOGNITION

Figures 4 and 5 report AUROC and AP for every finding using the same 16-finding CT-RATE and 14-finding RadChestCT scopes as the main evaluation. The component columns use epoch-60 TCS and the common prompt "Imaging findings are consistent with {finding}." The lung-opacity query uses "opacity" for all models. The final column uses the fixed equal-weight global/TCS readout from the same SIA-CT checkpoint. Calcification on RadChestCT uses the coronary-artery query; the additional arterial-wall query is excluded from its macro average. All values are percentages rounded after calculation.

CT-RATE: per-finding recognition (16 findings)
Cell text: score (%) | Cell color: change from Baseline TCS

	(a) AUROC				(b) AP			
	Baseline TCS	Baseline + S TCS	SIA-CT TCS	SIA-CT Fusion	Baseline TCS	Baseline + S TCS	SIA-CT TCS	SIA-CT Fusion
Emphysema	80.5	80.5	80.4	81.0	54.4	54.8	54.3	55.0
Atelectasis	76.9	75.9	76.8	75.8	51.7	49.9	52.0	51.7
Lung nodule	63.6	64.8	63.5	66.2	57.2	59.1	57.4	59.9
Lung opacity	81.1	81.1	80.5	81.1	68.0	68.6	67.6	68.7
Pulmonary fibrotic sequela	69.4	69.0	70.3	69.9	45.4	45.0	46.1	46.5
Pleural effusion	96.6	96.7	96.6	96.5	86.6	87.0	86.5	85.8
Mosaic attenuation pattern	88.6	88.5	88.0	87.6	45.1	45.2	45.5	45.2
Peribronchial thickening	80.4	81.4	81.2	82.0	38.4	40.6	39.1	40.7
Consolidation	90.6	90.8	91.0	90.5	65.0	66.0	65.7	65.1
Bronchiectasis	78.2	78.1	78.9	77.9	38.8	36.7	39.2	39.1
Interlobular septal thickening	87.3	87.7	87.2	87.4	38.8	40.2	39.6	40.0
Cardiomegaly	94.3	94.5	94.4	94.3	68.7	69.7	69.0	67.1
Pericardial effusion	86.1	85.7	85.8	86.1	40.9	40.6	42.0	43.5
Coronary artery calcification	89.9	90.2	90.0	90.5	71.4	70.7	72.4	73.0
Hiatal hernia	75.9	76.5	75.7	75.2	33.1	33.6	34.6	33.1
Arterial wall calcification	90.6	90.6	90.7	91.0	77.5	76.5	77.8	78.3
Macro mean	**83.1**	**83.2**	**83.2**	**83.3**	**55.1**	**55.3**	**55.5**	**55.8**

-5 0 +5 Change vs. Baseline (pp)

-12 0 +12 Change vs. Baseline (pp)

Figure 4: CT-RATE per-finding recognition. Cell labels report AUROC or AP (%); color indicates change from Baseline TCS in percentage points, computed before rounding. Blue indicates an increase and orange a decrease; the baseline column is neutral. The first three columns use TCS. Fusion is the fixed mean of global and TCS scores from the same SIA-CT checkpoint.

External evaluation. Figure 5 shows the variation across findings on RadChestCT. For example, emphysema AP rises from 54.7% for Baseline to 60.7% for SIA-CT with TCS and 65.6% with fusion. Other findings show smaller or negative changes, making the per-finding view complementary to the macro average.

RadChestCT: per-finding recognition (14 findings)
Cell text: score (%) | Cell color: change from Baseline TCS

	(a) AUROC				(b) AP			
	Baseline TCS	Baseline + S TCS	SIA-CT TCS	SIA-CT Fusion	Baseline TCS	Baseline + S TCS	SIA-CT TCS	SIA-CT Fusion
Emphysema	77.3	79.3	80.0	81.3	54.7	60.3	60.7	65.6
Atelectasis	68.9	69.5	68.9	70.0	51.8	52.6	50.9	53.5
Lung nodule	63.3	63.6	64.9	64.2	85.3	85.2	86.0	85.6
Lung opacity	62.3	61.1	62.9	64.7	63.7	62.8	63.9	65.9
Pulmonary fibrotic sequela	79.6	80.7	80.0	80.7	51.0	48.5	50.1	50.4
Pleural effusion	90.0	89.6	90.4	91.3	77.0	76.9	76.4	78.9
Peribronchial thickening	63.2	63.9	65.0	67.2	14.1	14.6	16.6	17.4
Consolidation	79.2	78.9	79.7	80.2	40.8	39.7	40.6	42.3
Bronchiectasis	75.0	76.0	75.1	74.5	41.2	42.5	42.8	39.8
Interlobular septal thickening	82.6	82.5	82.0	82.0	25.0	24.3	24.2	24.2
Cardiomegaly	88.6	88.3	89.0	88.9	51.6	50.6	51.4	52.8
Pericardial effusion	68.6	68.5	68.1	67.8	30.1	30.4	30.2	29.8
Calcification	74.4	74.2	74.7	74.9	85.0	85.2	85.6	85.6
Hiatal hernia	67.2	66.3	66.6	65.9	20.3	19.7	19.3	19.0
Macro mean	**74.3**	**74.4**	**74.8**	**75.3**	**49.4**	**49.5**	**49.9**	**50.8**

-5 0 +5
Change vs. Baseline (pp)

-12 0 +12
Change vs. Baseline (pp)

Figure 5: RadChestCT per-finding recognition under the 14-finding protocol. Scores and color semantics follow Figure 4, with identical color limits for each metric across datasets. Calcification denotes the coronary-artery query. The maps show the distribution of gains and losses alongside the macro averages.

D SENSITIVITY TO AN EXPANDED PROMPT POOL

We evaluate all models with two shared prompt settings: the single prompt "Imaging findings are consistent with {finding}." (using "opacity" for lung opacity) and an expanded pool of previously explored descriptions. Both settings use the 16-finding CT-RATE and 14-finding RadChestCT scopes from Appendix C. Each text embedding in the pool is normalized, embeddings are averaged with equal weight, and the resulting prototype is normalized again before querying TCS. This is text-prototype averaging, distinct from the global/TCS score fusion used for final disease recognition. Each model uses the same text pool and evaluation scans. The expanded pool was assembled from historical prompt exploration; this experiment measures sensitivity to that pool rather than performance on a prospectively specified set of unseen templates.

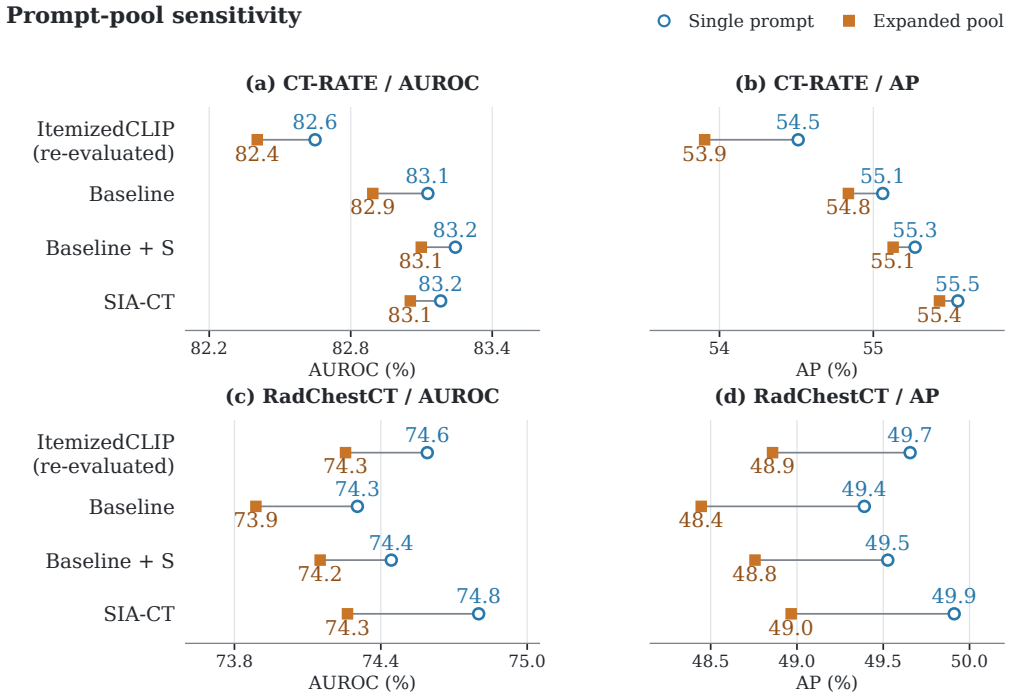

Figure 6: Prompt-pool sensitivity using TCS. Each line connects a single-prompt score (open circle) to an expanded-pool score (filled square) for the same model. All scores are labeled in percent; axes use focused ranges. ItemizedCLIP (re-evaluated) denotes our local evaluation of the released checkpoint with these shared prompts; Table 3 instead quotes the original paper. AUROC and AP are averaged over findings; AP is computed from our evaluation scores. Our trained models use epoch 60.

Figure 6 shows that the expanded pool lowers AUROC and AP for all four models on both datasets. Nevertheless, SIA-CT retains higher AP than Baseline under both settings: 55.5 versus 55.1 and 55.4 versus 54.8 on CT-RATE, and 49.9 versus 49.4 and 49.0 versus 48.4 on RadChestCT. Additional text variants therefore do not automatically improve recognition, while the AP advantage of S + R persists across these two settings.

E IMAGE-TO-IMAGE RETRIEVAL

We additionally evaluate image-to-image retrieval on CT-RATE. Figure 7 reports mean average precision at ranks 5, 10, and 50, including six external comparison models and the three local epoch-60 checkpoints.

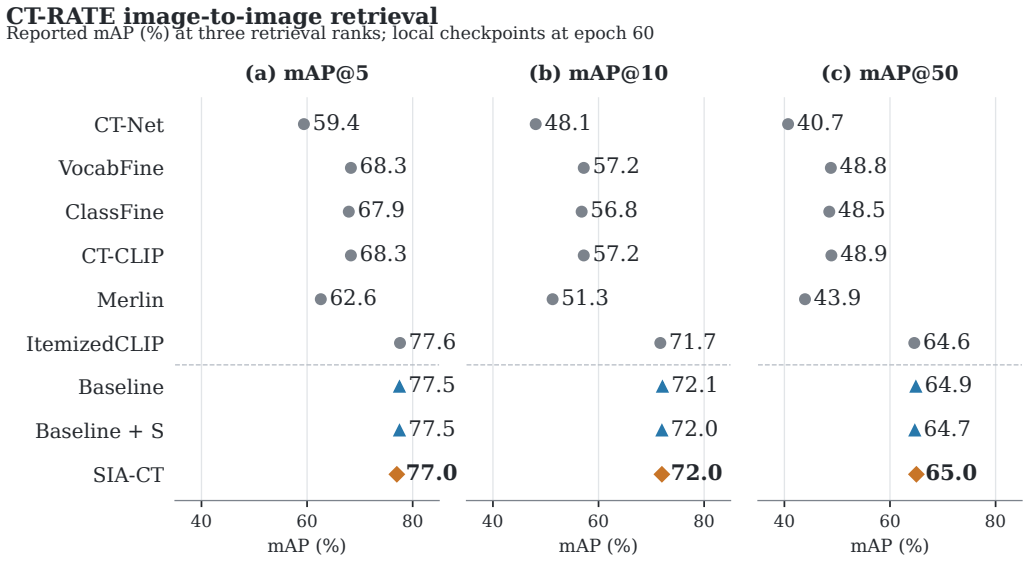

Figure 7: CT-RATE image-to-image retrieval (%). All panels share the same horizontal scale. External-model values (gray circles) are reproduced from the earlier comparison table; Baseline and Baseline + S (blue triangles) and SIA-CT (orange diamonds) use the current epoch-60 checkpoints. Labels preserve the reported one-decimal values.

Among the models in Figure 7, SIA-CT achieves the highest mAP@50, Baseline the highest mAP@10, and ItemizedCLIP the highest mAP@5. Retrieval performance remains close across the three local models, with rank-dependent changes. Compared with Baseline, SIA-CT changes mAP@5/10/50 by $-0.5/-0.1/+0.1$ percentage points, computed before rounding. These results show broadly retained retrieval performance alongside the state-readout improvements.

F Extended Related Work

Image–text learning and medical supervision. Deep learning has enabled diverse medical-image analysis tasks, with annotation availability remaining a central constraint (Litjens et al., 2017). CLIP and ALIGN established scalable image–text representation learning (Radford et al., 2021; Jia et al., 2021); SigLIP introduced a pairwise sigmoid objective (Zhai et al., 2023). ConVIRT learns from paired medical images and reports, while GLoRIA combines global alignment with local image–word correspondences (Zhang et al., 2022; Huang et al., 2021). BioViL studies biomedical text semantics, CheXzero learns radiograph recognition from image–report pairs, and BiomedCLIP scales biomedical pre-training using scientific image–text data (Boecking et al., 2022; Tiu et al., 2022; Zhang et al., 2023). CARZero further explores cross-attention-based radiology alignment (Lai et al., 2024). These studies motivate preserving fine-grained textual information when reports supervise visual representations.

Volumetric CT representations and generative tasks. Models Genesis studies self-supervised volumetric pre-training, and Med3D studies transferable 3D medical-image representations (Zhou et al., 2019; Chen et al., 2019). CT-Net examines multiple-abnormality prediction from full CT volumes (Draelos et al., 2021). CT-CLIP and CT-RATE support language-based chest CT recognition; BIUD transfers radiograph knowledge to CT, and fVLM develops fine-grained CT–text pre-training (Hamamci et al., 2024a; Cao et al., 2024; Shui et al., 2025). HLIP addresses scalable 3D language–image learning, while Merlin learns volumetric representations using CT, reports, and structured clinical data (Zhao et al., 2026; Blankemeier et al., 2026). RadZero3D explores transfer from self-supervised video models to CT language alignment (Park et al., 2025). Related generative tasks include CT2Rep for report generation and GenerateCT for text-conditioned CT synthesis (Hamamci et al., 2024b;c). Our evaluation concentrates on recognition, report-state discrimination, and image retrieval using a shared representation.

Subcaptions, itemization, and relation supervision. DreamLIP uses multiple subcaptions to supervise local visual content, and FLAIR uses text-conditioned attention pooling to obtain query-specific representations (Zheng et al., 2024; Xiao et al., 2025). ItemizedCLIP studies multiple independently meaningful items in one visual input, including report findings, and coordinates local alignment, item separation, and representation coverage (Lyu et al., 2026). These developments make individual descriptions accessible to the visual representation. Our work addresses the relations between those descriptions: independent findings, overlapping statements, and exact observed-state matches supply different supervision. Debiased contrastive learning, supervised contrastive learning, and MedCLIP establish the importance of semantic relations in positive and negative construction (Chuang et al., 2020; Khosla et al., 2020; Wang et al., 2022). Here, observation masks distinguish partial descriptions from explicit state agreement and connect this decision to item separation and alignment losses.

State sensitivity and anatomical support. CheXpert represents uncertainty in report-derived labels (Irvin et al., 2019); our item records also track anatomy, polarity, laterality, and severity. The need to test these distinctions relates to ARO's evaluation of attributes and relations in general vision–language models (Yuksekgonul et al., 2023). Our questions isolate clinical state by keeping finding identity fixed. On the visual side, locality alignment studies how to improve regional semantics in general-purpose backbones (Covert et al., 2025). In CT, fVLM combines anatomy-level matching with disease-aware correction of false negatives (Shui et al., 2025). SIA-CT adds an observed-state criterion for cross-patient item relations and conditions within-examination separation on finding dependence. TotalSegmentator, Chest-OMDL, ARC-CT, and OKA-CT provide further context for anatomical supervision (Wasserthal et al., 2023; Bai et al., 2026; Isik et al., 2026; You et al., 2026). R uses organ-supported multi-positive alignment through shared TCS transformations to complement these state relations.

G A Worked Example of State and Region Supervision

We trace S and R using report items, relation metadata, CT, and organ labels from a real training examination, denoted Case A. Three cross-patient items illustrate the relation rules. The candidate set below is assembled for explanation from stored records; it is not a recovered training minibatch. A separate held-out example then shows the saved epoch-60 inference scores.

G.1 From Report Items to State Relations

The source report for Case A contains "Paraseptal emphysematous changes are observed in both lungs." Its itemized text pool includes the corresponding view A1, a nodule description A2, and another emphysema description A3 (Table 7). The text encoder receives the natural item; S uses its stored relation metadata. All six records in the table have observed lung anatomy, positive polarity, and bilateral laterality. The remaining observed fields are shown explicitly. A dash means unobserved, even when a default value is present in storage. The serialized records also retain an unobserved temporal-change field, which does not distinguish these examples.

ID	Stored natural report item	Severity	Uncertainty
A1	Paraseptal emphysematous changes are present in both lungs.	–	–
A2	A few subpleural millimetric nonspecific nodules in both lungs .	–	–
A3	Minor centrilobular emphysematous changes are present in both lungs.	mild	definite
B	Emphysematous changes are present in both lungs.	–	–
C	Emphysema is present in both lungs.	–	definite
D	Paraseptal emphysematous changes are present in both lungs.	–	–

Table 7: Real items and their stored observation metadata. A1–A3 belong to one examination; B, C, and D each belong to another patient. A2 maps to pulmonary nodule; the other rows map to emphysema. The fields describe parser output, not independently adjudicated clinical truth.

Within-examination IIS. The stored relation matrix marks A1–A2 and A2–A3 as eligible separation pairs (-1), whereas A1–A3 is masked (0): the latter two are views of the same finding group. This does not make their states equal. The unique-finding sampler subsequently selects at most one emphysema view per examination, so A1 and A3 are alternatives rather than simultaneous sampled inputs. If A1 and A2 are selected, they retain their distinct-finding IIS term.

Exact cross-patient relations. A1 and B have equal finding identities, observation masks, and state signatures, despite different wording. They receive soft-positive weight 0.25 in the cross-patient relation objective, and the corresponding accepted match is masked from ILA/MPS negatives. D has the same signature and receives the same S treatment. C is not accepted: its uncertainty field is marked observed, whereas A1's is unobserved. This example exposes the dependence of exact matching on extracted metadata; semantic similarity of the sentences alone does not override the stored observation masks.

Pair	Exact-state positive	S false-negative mask	Paired R positive
A1–B	Yes	Yes	No
A1–C	No	No from this pair	No
A1–D	Yes	Yes	Yes

Table 8: Different relation scopes in S and R. All three remote items share the lung support of A1 and are allowed regional candidates. S accepts exact state signatures; paired R accepts cross-patient exact text repetitions. A missing S match does not label the pair as a known clinical contradiction.

G.2 From Organ Labels to Regional Supervision

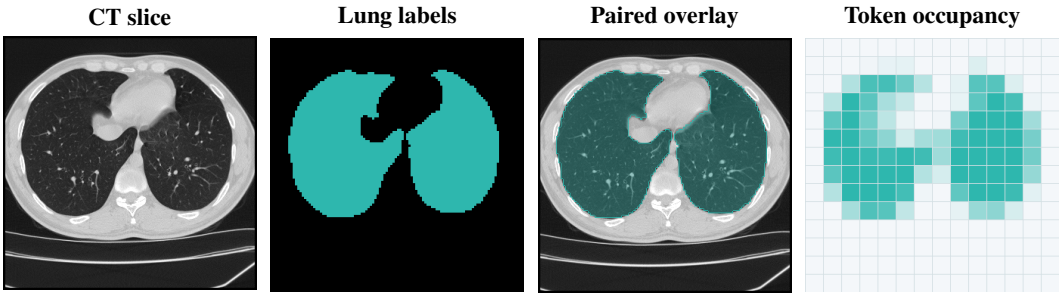

Figure 8: Case A's CT and TotalSegmentator lung support, after the same center pad/crop to $112 \times 336 \times 336$. The displayed slice is selected by maximal lung-mask area, independently of model scores. The grid shows measured occupancy in its eight-slice token block, from empty (light) to full (teal). These are organ labels and geometric occupancy, not lesion annotations or learned attention.

Geometry and pooling. The source tensor has shape $117 \times 330 \times 330$. For this display, three voxels are padded on each in-plane side, and the depth crop starts at index 2. This retains 99.9% of the lung labels. The union of the five lung-lobe labels yields 895 nonempty tokens: 213 fully occupied and 682 partially occupied, among 14^3 tokens. For example, one boundary token contains 549 lung voxels out of 4,608, giving 11.9% occupancy. Equation 7 normalizes these occupancy weights and pools through the shared TCS transformations. The display uses a fixed center crop; training uses its sampled CT/mask geometry and rejects regions retaining less than 90.0% of their labeled volume.

Multi-positive regional targets. Consider the illustrative candidate set $\mathcal{C} = \{\mathrm{A1}, \mathrm{A2}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$. Every candidate has the same lung-label support. A1 and A2 are positives for Case A's lung feature because they describe the same examination and region, even though they refer to different findings. D is also positive because its text exactly repeats A1 across patients. B and C remain contrastive alternatives in paired R. Thus, $\mathcal{P}_R = \{\mathrm{A1}, \mathrm{A2}, \mathrm{D}\}$, and the regional cross-entropy for this candidate set is

$$\mathcal{L}_{R,A} = -\frac{1}{3} \sum_{b \in \mathcal{P}_R} \log \frac{\exp(\cos(h_{A,\mathrm{lung}}, \mathrm{sg}[e(t_b)])/0.07)}{\sum_{c \in \mathcal{C}} \exp(\cos(h_{A,\mathrm{lung}}, \mathrm{sg}[e(t_c)])/0.07)}. \tag{11}$$

The same lung feature serves A1 and A2; it is not a finding-specific lesion readout. The implementation corrects repeated-region anchor weights before aggregation, and R enters the joint objective with weight 0.05. S and paired R apply different positive criteria, as Table 8 makes explicit.

G.3 Saved Scores on a Held-Out State Question

A separate validation report states: "When examined in the lung parenchyma window; Density reduction compatible with mild emphysema is observed in both lungs." Its severity question contrasts the two descriptions in Table 9. This was the first emphysema severity question in the inspected benchmark order; it was selected without filtering on prediction correctness. Both saved B60 readouts rank the report-supported mild description first. Fusion is computed from the unrounded cosine scores. This single example illustrates scoring, while aggregate state performance is reported in Table 2.

Candidate description	TCS	Global	Equal fusion
There is likely severe bilateral emphysema.	20.1	8.0	14.0
There is likely mild bilateral emphysema.	**37.6**	**21.2**	**29.4**

Table 9: Saved epoch-60 scores on the held-out question, displayed as $100\times$ cosine similarity to one decimal place. These values are similarities, not probabilities. This scan is separate from training Case A; inference uses CT and candidate text without organ masks.