Additional file 1: Supplementary Dataset Details, Methods, Extended Results, and Qualitative Panels

Supplementary material for: *Lesion-Grounded Content-Based Retrieval of CT Findings Using Foundation Model Patch Embeddings for Case Review*

Journal: *BMC Medical Imaging*

Author: Zhongsheng Eric Yu

Affiliation: Independent researcher, San Diego, California, USA

Corresponding-author email: peng.sheng.2023@gmail.com

S1. Supplementary Dataset and Implementation Details

S1.1 Evaluated Ontology and Public-Validation Support

Table S1 lists the evaluated ReXGroundingCT ontology and public-validation query support [1]. Support is the number of evaluable lesion queries in the common 407-query cohort used for exact-mask, bounding-box, and loose-box retrieval; it is not a count of source findings or training reference lesions. Parent categories `3-12` in the broader report-extraction ontology are not evaluated because the spatially annotated release used here is restricted to pulmonary and pleural findings.

Table S1. Evaluated ReXGroundingCT label ontology and public-validation query support.

Primary category	Subcategory label	Definition	Validation query support
1 — typically non-focal	1a	Bronchial wall thickening	4
1 — typically non-focal	1b	Bronchiectasis	21
1 — typically non-focal	1c	Emphysema, including centrilobular, paraseptal, and bullous patterns	6
1 — typically non-focal	1d	Septal thickening, including interlobular thickening and reticulation	25
1 — typically non-focal	1e	Micronodules, including centrilobular, tree-in-bud, and perilymphatic patterns	25
1 — typically non-focal	1f	Other non-focal abnormality	9
2 — typically focal	2a	Linear opacity, including subsegmental atelectasis, scarring, and fibrosis	44
2 — typically focal	2b	Atelectasis or consolidation	65
2 — typically focal	2c	Groundglass opacity	113
2 — typically focal	2d	Pulmonary nodules or masses	83
2 — typically focal	2e	Pleural effusion or thickening	7
2 — typically focal	2f	Honeycombing	0
2 — typically focal	2g	Pneumothorax	0
2 — typically focal	2h	Other focal abnormality	5
Total			**407**

S1.2 Retrieval Artifact and Auxiliary-Control Implementation

Each retrieval run records the model name, checkpoint identifier or hash, preprocessing configuration hash, sparse-token schema version, threshold, pooling and normalization policies, query-region policy and version, loose-box parameters, random seed where applicable, embedding artifact version, reference and query splits, fold identifier when applicable, negative-bank policy, and evaluation script version.

Encoder and preprocessing provenance.

Both encoders use the authors' public FlexiCT model definitions and weights [10], released through the FlexiCT source repository and model repository. The primary FlexiCT-3D checkpoint is `3D_final_model.pth` (SHA-256 `640aa0a33c006d89bfdfbd0150b13e37f298b83b938ca02985fae6a06e8ef1c1`), and the ablation checkpoint is `3D_VLM_final_model.pth` (SHA-256 `bb4e0a697380f3636723b76b175cd1d6ca2e72659a4c269be3b3e5d41c4b7403`). The internal implementation identifier `FlexiCTVLM` maps to the public FlexiCT-3D-VLM release rather than to a separate model. Both encoders were frozen and run in evaluation mode without fine-tuning on ReXGroundingCT.

Reference artifacts and exact-mask query artifacts were derived from sparse patch tokens extracted and stored with float16 automatic mixed precision before normalized pooled embeddings were computed in float32. The bounding-box and loose-box query generators instead reran the encoder in float32 and stored normalized pooled query embeddings as float32. In the final eligible training reference bank, the axiswise patch-grid ranges were 13-126 in depth and {64, 96, 128} in both height and width; in public validation, they were 22-65 in depth and {64, 96} in height and width.

For CT preprocessing, a near-constant volume with standard deviation below 1e-6 was mean-centered without division. Padding added at most seven voxels per dimension; when an odd total was required, the additional voxel was placed on the trailing side. The preprocessing sidecar records native shape and spacing, padding offsets, processed shape, and patch-grid shape so that sparse patch records, mask overlays, selected slices, and review-panel coordinates can be mapped back to the native CT grid.

Sparse artifact inventory and schema.

The stored retrieval artifacts include model-level global embeddings; all-disease, lesion-balanced, and finding-balanced volume embeddings; sparse patch-token records for lesion-intersecting patches; lesion and lesion-slice embeddings; mask-derived descriptors; source finding-text links; and preprocessing metadata. A separate validation-only artifact contains annotation-negative candidates; these candidates are not part of the production reference index. Each sparse lesion patch-token record contains:

```
sample_id
lesion_entity_id
finding_id
mask_id
slice_index
patch_grid_index: z, h, w
patch_embedding
mask_fraction_in_patch
processed_patch_bounds
category / subcategory labels
finding_text_id
preprocessing_metadata_id
```

Category labels and finding text are retained for linkage, evaluation, and review displays. They do not contribute to embedding similarity.

Training-only occupancy-threshold selection.

The selected occupancy threshold was re-evaluated using patient-grouped five-fold splits without looking at public-validation performance. The comparison held the candidate thresholds, seed, pooling method, exact-mask query policy, subcategory Recall@10 endpoint, and original coverage rationale fixed. The common comparison cohort contained 11,432 lesions from 2,623 CTs and 2,384 patients; its folds contained 2,276-2,302 query lesions, 520-534 CTs, and 476-478 patients, with no query/reference patient overlap.

Table S2. Training-only patient-grouped occupancy-threshold comparison.

Threshold	R@1	R@5	R@10	R@20	MRR@10	Full-bank lesions / CTs / patients
0.005	0.4844	0.8069	0.8795	0.9292	0.6201	14,557 / 2,992 / 2,717
0.01	0.4852	0.8075	0.8792	0.9290	0.6207	14,551 / 2,992 / 2,717

Threshold	R@1	R@5	R@10	R@20	MRR@10	Full-bank lesions / CTs / patients
0.05	0.4862	0.8059	0.8804	0.9294	0.6212	14,359 / 2,978 / 2,705
0.10	0.4912	0.8076	0.8802	0.9285	0.6243	13,746 / 2,916 / 2,645
0.25	0.4966	0.8086	0.8806	0.9294	0.6283	11,432 / 2,623 / 2,384

Bold marks the retained coverage-specificity choice, not the numerical maximum in every metric. The 0.25-minus-0.05 Recall@10 difference was 0.00017; its five fold-wise differences ranged from -0.00220 to +0.00220 and changed sign. The patient-grouped analysis therefore retained the original 0.05 choice because performance was nearly tied while 0.05 preserved substantially greater lesion and CT coverage.

Query-policy implementation.

The exact-mask, bounding-box, and loose-box implementations are versioned as `exact_lesion_mask_v1`, `bbox_mask_v1`, and `loose_user_like_box_v1`, respectively. For both box policies, inclusive [D, H, W] voxel bounds are mapped to the 8 x 8 x 8 patch grid by floor division and clipped to the valid grid extent. The loose-box implementation applies no center jitter or other random perturbation, consumes no loose-box random seed, and is deterministic across encoders. The `user_like` string is an internal implementation label and does not imply evaluation in a user study. The final FlexiCT-3D and FlexiCT-3D-VLM artifacts contain identical lesion identifiers and bounding-box and loose-box patch counts for all 407 evaluable queries.

Eligibility provenance and positive availability.

A query lesion is eligible for automatic ranking metrics when its mask has nonzero voxel support, preprocessing succeeds, the category label is available, the selected occupancy threshold leaves at least one occupied patch, and at least one valid positive exists after exclusions. Of the 429 annotated public-validation entities, 407 meet the 0.05 patch-occupancy requirement; the other 22 have nonzero masks and valid labels but no patch with occupancy at least 0.05. All 407 queries have at least one positive among the 14,359 reference lesions under both relevance rules. Primary category and exact subcategory each therefore have 407 total and 407 evaluable queries, with zero queries without positives. The mean and median reference-positive counts per query are 9,277.4 and 10,941 for primary category and 2,217.1 and 2,384 for subcategory, respectively.

The reference pool excludes the same sample identifier as the query. Queries come from the ReXGroundingCT public-validation split, and references come from its training split [1]. The retained dataset identifiers show that, at the underlying CT-RATE level, all 50 public-validation query CTs belong to the CT-RATE training partition [1,6]. Resolving CT-RATE patient identifiers gives 49 patients across the 50 query CTs and 2,717 across the 2,992 reference CTs, with zero patients in common. Train five-fold evaluation is patient-grouped: patient identifiers are namespaced by partition, and all eligible scans belonging to one patient are assigned to the same fold. The 0.05 cohort contains 2,705 patients across 2,978 CTs; 202 patients contribute multiple eligible scans, covering 475 CTs, and every fold has zero query/reference patient overlap. Ranking remains lesion-level: every eligible reference lesion is a distinct candidate, multiple lesions from one reference CT can appear separately, and parent-volume grouping is applied only to case-level presentation rather than lesion-level ranks or denominators.

The reference embedding workflow has two stages. Stage 1 runs the encoder, temporarily holds the full patch grid, projects lesion masks to the patch grid, extracts sparse lesion-associated tokens, stores mask-fraction metadata, writes global embeddings, and discards the full grid unless explicitly requested. Stage 2 reads the sparse token bank and derives lesion embeddings, lesion-slice embeddings, and all-disease, lesion-balanced, and finding-balanced volume embeddings under versioned threshold and pooling policies.

Volume-embedding definitions.

Let $F_i^p(p)$ be the patch token at location p in CT i, let w_{ijp} be the occupied voxel fraction of that patch under lesion mask j, and let ρ_{ij} be the normalized lesion embedding defined in Section 4.5 of the main manuscript. The global baseline uses the encoder's global representation: the CLS embedding for FlexiCT-3D and the released projected image embedding for FlexiCT-3D-VLM. For all-disease pooling, one row is retained per patch even when multiple masks intersect it, with $w_{ip}^{\text{disease}} = \max_j w_{ijp}$. Applying the selected threshold $\tau = 0.05$ gives

$$f_{i,\,\text{disease}-\text{all}}^{v} = \text{norm}\left(\frac{\sum_{p:\, w_{ip}^{\text{disease}} \geq \tau} w_{ip}^{\text{disease}} F_i^p(p)}{\sum_{p:\, w_{ip}^{\text{disease}} \geq \tau} w_{ip}^{\text{disease}} + \epsilon}\right).$$

Because this representation can be dominated by large lesions, the lesion-balanced embedding instead averages normalized lesion embeddings and normalizes the result:

$$f^{v}_{i,\text{lesion}-\text{balanced}} = \text{norm}\left(\frac{1}{K_i}\sum_{j=1}^{K_i}\rho_{ij}\right).$$

Finding-balanced pooling first aggregates lesions linked to finding c,

$$\phi_{ic} = \text{norm}\left(\frac{1}{|\mathcal{K}_{ic}|}\sum_{j\in\mathcal{K}_{ic}}\rho_{ij}\right),$$

and then pools across findings:

$$f^{v}_{i,\text{finding}-\text{balanced}} = \text{norm}\left(\frac{1}{C_i}\sum_{c=1}^{C_i}\phi_{ic}\right),$$

where K_i is the number of eligible lesions in CT i, $\mathcal{K}_{ic}$ is the set of lesion entities linked to finding c, and C_i is the number of eligible findings in CT i. All disease-aware representations are L2-normalized before cosine retrieval. Lesion-balanced pooling averages individually normalized lesion vectors before final normalization; finding-balanced pooling additionally normalizes each within-finding mean before the final normalized across-finding mean. Each loaded primitive embedding matrix is row-wise L2-normalized before similarity calculation. Invalid zero placeholders for CTs without an eligible patch at the selected threshold are excluded rather than indexed.

Descriptor controls, reranking, and negative-bank protocol.

The geometry descriptors are computed from the exact annotation corresponding to each lesion entity, including for the bounding-box and loose-box query experiments; they are therefore auxiliary controls rather than a claim about a deployed reranker with automatically estimated lesion geometry. Equivalent diameter is computed as $2(3V/4\pi)^{1/3}$ from physical lesion volume V. Positive patch count includes every encoder patch with lesion mask fraction greater than zero. Size buckets are assigned independently within the evaluable query and reference tables using the 33.3333rd and 66.6667th diameter percentiles. For public validation, the query cut points are 8.576 and 17.387 mm, and the reference cut points are 10.333 and 22.390 mm. Positive axial-slice count was an implemented optional descriptor, but its source field was absent for all 407 query and 14,359 reference lesions in the final validation tables, so availability normalization removed it from every descriptor pair. Native query slice counts were later reconstructed from the source masks solely for the post hoc strata in Section S2.4; they were not used in retrieval or reranking.

For a continuous descriptor, missing values are excluded pairwise; otherwise the query and reference values use the reference-table median and interquartile range, with a unit fallback when the interquartile range is zero. Continuous similarity is $s_j(q,r) = \exp(-|z_{qj} - z_{rj}|)$, and the bucket score is binary agreement. With $a_j(q,r)$ indicating pairwise availability and w_j the component weight, the availability-normalized score is

$$D_w(q,r) = \frac{\sum_j w_j a_j(q,r) s_j(q,r)}{\sum_j w_j a_j(q,r)}.$$

The descriptor-only ranking applies weights 1.0 to diameter similarity, 1.0 to patch-count similarity, 0.5 to the unavailable axial-slice-count similarity, and 0.5 to exact size-bucket agreement. Thus, for the executed tables, its denominator is 2.5 rather than 3.0. No oracle label component is active.

Reranking starts from the 20 candidates per query stored by the embedding-retrieval run. The embedding component is min-max normalized separately for each query across those 20 candidates; a constant candidate-score vector maps to ones. The descriptor component applies weights 0.25, 0.25, 0.10, and 0.10 to diameter, patch count, axial-slice count, and size-bucket agreement, respectively, and is normalized over available components, giving an executed denominator of 0.60 because axial-slice count is missing. With the resulting components denoted $E(q,r) \in [0, 1]$ and $D_{\text{rerank}}(q,r) \in [0, 1]$, the executed fusion is

$$S_{\text{rerank}}(q,r) = \frac{E(q,r) + D_{\text{rerank}}(q,r)}{2}.$$

All 20 candidates are reordered, and the first 10 are used for the reported metrics. Semantic labels and finding text are used only to evaluate the resulting order, never to compute the reported descriptor or reranker scores.

The validation-negative candidate bank is separate from the production retrieval index. It is built during validation embedding extraction, before the full patch grid is discarded, and contains only the executed near/far annotation-negative patch-grid-plane candidates; foreground and anatomy-matched policies were not used in the reported results. A patch-grid depth plane is lesion-positive if the maximum union-mask fraction across its in-plane patches is greater than zero; all other planes are eligible. With a target of eight candidates per CT, the selector takes the four eligible planes with the smallest distance to any lesion-positive

plane and the four with the largest distance, preserves this near-then-far order, and removes duplicates. Forty-eight CTs contribute eight candidates and two contribute seven, yielding 398 candidates from 50 CTs. Each candidate is the L2-normalized mean of all patch embeddings in one depth plane and retains its CT identifier, plane index, distance to the nearest lesion-positive plane, and a deterministic per-CT seed derived from base seed 17 and the sample identifier. Query and bank embeddings are L2-normalized before the full 407 x 398 cosine comparison. The analysis records the 20 highest-scoring negatives per query and flags whether the best exceeds the query's top-1 lesion score or falls within 0.05 of it. Same-CT negatives are retained, and the bank's annotation-negative status does not establish that a plane is clinically normal.

S2. Extended Retrieval Protocol and Result Tables

Section S2 reports extended support, stratification, and uncertainty analyses for the completed retrieval experiments. The underlying per-query and per-result tables retain the identifiers and provenance fields described in Section S1.2. The manuscript tables summarize the endpoints needed to interpret the main claims.

S2.1 Patient-Grouped Five-Fold Training-Set Query Coverage and Positive Availability

The patient-grouped fold assignment is shared by FlexiCT-3D and FlexiCT-3D-VLM, so the following support statistics apply to both encoders. Table S3 reports fold-level coverage: every query is evaluable and has at least one positive in the complementary four-fold reference bank under both relevance rules. The five folds contain 540-542 query patients and 592-605 query CTs each, with zero query/reference patient overlap.

Table S3. Patient-grouped five-fold training-set query coverage and positive availability.

Fold	Total/evaluable queries	Without primary positives	Primary positives, mean/median	Without subcategory positives	Subcategory positives, mean/median
0	2,872/2,872	0	7,226.0/8,811	0	1,676.7/1,885
1	2,872/2,872	0	7,363.0/8,725	0	1,739.6/1,941
2	2,872/2,872	0	7,278.2/8,779	0	1,718.7/1,915
3	2,872/2,872	0	7,271.7/8,783	0	1,705.1/1,916
4	2,871/2,871	0	7,452.8/8,666	0	1,752.2/1,879

S2.2 Public-Validation Primary-Category Encoder Ablation

Primary-category metrics are reported for completeness but are secondary because the benchmark contains only two broad, imbalanced primary categories. Table S4 reports the per-encoder rows and Table S5 the paired differences, using the same CT-grouped 10,000-resample procedure as Section 5.3 of the main manuscript.

Table S4. Public-validation primary-category retrieval by encoder and query policy.

Query policy	Encoder	Primary R@1	R@5	R@10	R@20	MRR@10
Exact mask	FlexiCT-3D	0.828	0.968	0.985	0.998	0.890
Exact mask	FlexiCT-3D-VLM	0.828	0.961	0.988	0.990	0.883
Bounding box	FlexiCT-3D	0.781	0.961	0.980	0.995	0.861
Bounding box	FlexiCT-3D-VLM	0.794	0.966	0.988	0.995	0.867
Loose box	FlexiCT-3D	0.693	0.946	0.975	0.995	0.798
Loose box	FlexiCT-3D-VLM	0.747	0.951	0.983	0.998	0.834

Table S5. Paired primary-category encoder differences, FlexiCT-3D-VLM minus FlexiCT-3D.

Query policy	Paired R@1 difference, VLM - 3D (95% CI)	Paired MRR@10 difference (95% CI)
Exact mask	0.000 (-0.029 to +0.025)	-0.007 (-0.029 to +0.011)
Bounding box	+0.012 (-0.029 to +0.049)	+0.006 (-0.019 to +0.026)
Loose box	+0.054 (+0.022 to +0.082)	+0.036 (+0.016 to +0.054)

The loose-box primary-category intervals exclude zero, but this exploratory endpoint is susceptible to class imbalance, measures agreement between only two broad categories, and is one of several examined metrics. It does not establish improved fine-grained retrieval; all corresponding subcategory early-rank intervals include zero.

S2.3 FlexiCT-3D Query-Policy Uncertainty

Intervals are CT-grouped percentile 95% bootstrap intervals from 10,000 resamples with seed 17. Point estimates remain lesion-level averages, while resampling keeps lesions from the same CT together. Table S6 reports the per-policy intervals. Table S7 reports paired differences calculated as the first-named policy minus the second-named policy on the identical 407-query cohort.

Table S6. FlexiCT-3D query-policy subcategory metrics with CT-grouped bootstrap intervals.

Query policy	Subcategory R@1 (95% CI)	R@5 (95% CI)	R@10 (95% CI)	R@20 (95% CI)	MRR@10 (95% CI)
Exact mask	0.484 (0.408-0.569)	0.786 (0.704-0.872)	0.848 (0.783-0.914)	0.926 (0.892-0.962)	0.614 (0.538-0.695)
Bounding box	0.430 (0.355-0.517)	0.759 (0.685-0.839)	0.840 (0.787-0.896)	0.916 (0.878-0.955)	0.568 (0.498-0.646)
Loose box	0.356 (0.286-0.436)	0.693 (0.630-0.764)	0.821 (0.776-0.868)	0.897 (0.854-0.935)	0.500 (0.437-0.573)

Table S7. Paired FlexiCT-3D query-policy differences for subcategory retrieval. Differences are first-named minus second-named; every metric cell reports the paired difference and CT-grouped 95% confidence interval.

Paired comparison	Difference in R@1 (95% CI)	R@5	R@10	R@20	MRR@10
Bounding box - exact mask	-0.054 (-0.098 to -0.007)	-0.027 (-0.065 to +0.006)	-0.007 (-0.049 to +0.040)	-0.010 (-0.034 to +0.015)	-0.046 (-0.079 to -0.014)
Loose box - exact mask	-0.128 (-0.171 to -0.086)	-0.093 (-0.154 to -0.036)	-0.027 (-0.077 to +0.033)	-0.029 (-0.066 to +0.005)	-0.114 (-0.155 to -0.074)
Loose box - bounding box	-0.074 (-0.113 to -0.038)	-0.066 (-0.105 to -0.030)	-0.020 (-0.051 to +0.011)	-0.020 (-0.045 to +0.005)	-0.068 (-0.094 to -0.043)

The intervals show a consistent distinction between early ranking and top-k coverage. Coarser query regions produce clear losses in Recall@1 and MRR@10, whereas the paired Recall@10 and Recall@20 intervals include zero. Thus, the evaluated box conditions retain similar top-k point estimates without establishing equivalence to exact-mask retrieval or performance under real user localization.

S2.4 Exploratory Retrieval Strata

Strata use the common 407-query FlexiCT-3D cohort. Point estimates are lesion-level subcategory metrics; Recall@10 parentheses contain CT-grouped percentile 95% intervals from 10,000 resamples with seed 17. Lesion-size buckets are the query-cohort diameter tertiles defined in Section S1.2. Positive patch count is the number of encoder patches with nonzero exact-mask occupancy, including patches below the 0.05 occupancy threshold. Tables S8, S9, and S10 report the size-tertile, patch-count, and slice-support strata, respectively.

Table S8. Exploratory strata by lesion-size tertile.

Size stratum	Queries	Query policy	R@1	R@10 (95% CI)	MRR@10
Small	136	Exact mask	0.471	0.801 (0.725-0.880)	0.590
Small	136	Bounding box	0.390	0.824 (0.764-0.890)	0.546
Small	136	Loose box	0.338	0.801 (0.727-0.890)	0.488
Medium	135	Exact mask	0.430	0.822 (0.696-0.944)	0.568
Medium	135	Bounding box	0.437	0.807 (0.740-0.886)	0.544
Medium	135	Loose box	0.363	0.807 (0.731-0.888)	0.503
Large	136	Exact mask	0.551	0.919 (0.852-0.971)	0.683
Large	136	Bounding box	0.463	0.890 (0.798-0.954)	0.614
Large	136	Loose box	0.368	0.853 (0.793-0.911)	0.510

Table S9. Exploratory strata by positive patch count.

Positive patches	Queries	Query policy	R@1	R@10 (95% CI)	MRR@10
1-4	60	Exact mask	0.517	0.783 (0.662-0.902)	0.619
1-4	60	Bounding box	0.467	0.833 (0.733-0.922)	0.603
1-4	60	Loose box	0.367	0.750 (0.603-0.889)	0.497
5-16	120	Exact mask	0.408	0.800 (0.702-0.889)	0.540
5-16	120	Bounding box	0.367	0.800 (0.705-0.891)	0.498
5-16	120	Loose box	0.300	0.825 (0.735-0.912)	0.458
17-64	116	Exact mask	0.457	0.845 (0.725-0.947)	0.599
17-64	116	Bounding box	0.466	0.845 (0.769-0.924)	0.592
17-64	116	Loose box	0.405	0.853 (0.781-0.927)	0.552
At least 65	111	Exact mask	0.577	0.937 (0.871-0.984)	0.706
At least 65	111	Bounding box	0.441	0.883 (0.778-0.959)	0.599
At least 65	111	Loose box	0.360	0.820 (0.734-0.900)	0.494

Table S10. Exploratory strata by query-lesion native axial-slice support.

Native axial support	Queries	Query policy	R@1	R@10 (95% CI)	MRR@10
Single-slice	14	Exact mask	0.429	0.714 (0.462-0.929)	0.520
Single-slice	14	Bounding box	0.429	0.714 (0.462-0.929)	0.504
Single-slice	14	Loose box	0.429	0.714 (0.462-0.929)	0.494
Multi-slice	393	Exact mask	0.486	0.852 (0.787-0.920)	0.617
Multi-slice	393	Bounding box	0.430	0.845 (0.791-0.901)	0.570
Multi-slice	393	Loose box	0.354	0.824 (0.779-0.872)	0.500

Single-slice lesions occur in only nine CTs, so their intervals are particularly wide. The size and positive-patch strata are not controlled comparisons: they differ in category composition and morphology, and positive patch count is mechanically related to lesion size and shape.

S2.5 Patient-Grouped Five-Fold Training-Set Subcategory Metrics

Table S11 shows fold-wise results under exact-mask queries for both encoders. The identical query counts reflect the shared patient-grouped fold assignment. These rows underlie the mean and sample-standard-deviation summaries in Table 3 of the main manuscript.

Table S11. Patient-grouped five-fold training-set subcategory metrics by encoder and fold.

Encoder	Fold	Queries	R@1	R@5	R@10	R@20	MRR@10
FlexiCT-3D	0	2,872	0.481	0.779	0.866	0.920	0.607
FlexiCT-3D	1	2,872	0.512	0.818	0.893	0.934	0.643
FlexiCT-3D	2	2,872	0.492	0.791	0.867	0.921	0.618
FlexiCT-3D	3	2,872	0.497	0.811	0.882	0.933	0.631
FlexiCT-3D	4	2,871	0.492	0.814	0.894	0.938	0.628
FlexiCT-3D-VLM	0	2,872	0.474	0.782	0.864	0.919	0.605
FlexiCT-3D-VLM	1	2,872	0.520	0.816	0.883	0.931	0.646
FlexiCT-3D-VLM	2	2,872	0.500	0.795	0.868	0.914	0.625
FlexiCT-3D-VLM	3	2,872	0.512	0.805	0.886	0.932	0.638
FlexiCT-3D-VLM	4	2,871	0.490	0.816	0.895	0.937	0.630

S2.6 Public-Validation Category Support and Retrieval

Table S12 reports per-subcategory support and retrieval. Every represented subcategory has at least one training-reference positive, and all its validation queries are evaluable under every query policy. *Training-reference positives* is the number of eligible lesions in the common 14,359-lesion reference bank; it is constant for every query with that subcategory. The two zero-query labels retain training references but cannot be evaluated on public validation.

Table S12. Public-validation subcategory support and retrieval. Query-policy cells report Recall@1 / Recall@10 / MRR@10.

Label	Queries/evaluable	Training-reference positives	Exact mask	Bounding box	Loose box
1a	4/4	467	0.250/1.000/0.479	0.000/0.750/0.271	0.000/0.750/0.333
1b	21/21	571	0.524/0.857/0.630	0.571/0.905/0.714	0.476/0.905/0.628
1c	6/6	1,031	0.667/1.000/0.806	0.667/1.000/0.792	0.667/1.000/0.833
1d	25/25	415	0.000/0.440/0.097	0.040/0.360/0.094	0.000/0.440/0.067
1e	25/25	639	0.120/0.520/0.224	0.160/0.640/0.289	0.120/0.520/0.210
1f	9/9	295	0.333/0.556/0.411	0.444/0.556/0.460	0.444/0.778/0.495
2a	44/44	1,883	0.500/0.909/0.659	0.500/0.886/0.636	0.455/0.864/0.596
2b	65/65	2,384	0.369/0.800/0.507	0.308/0.877/0.480	0.200/0.892/0.403
2c	113/113	3,282	0.637/0.991/0.784	0.558/0.947/0.697	0.496/0.885/0.621
2d	83/83	2,914	0.614/0.916/0.731	0.494/0.904/0.646	0.373/0.867/0.548
2e	7/7	328	0.857/1.000/0.929	0.571/0.714/0.643	0.429/0.857/0.521
2f	0/0	32	—	—	—
2g	0/0	21	—	—	—
2h	5/5	97	0.000/0.200/0.067	0.000/0.200/0.100	0.200/0.200/0.200

Rare-label point estimates are unstable. For example, the apparently perfect Recall@10 for 1a, 1c, and 2e under selected policies is based on only 4, 6, and 7 queries. The larger 2c and 2d strata provide more stable evidence; for both labels, Recall@1 and MRR@10 decrease as query regions become coarser.

S2.7 Descriptor Controls and Geometry Reranking

Category-wise cells in Table S13 report the reranker minus embedding-only change in Recall@1 / Recall@10 / MRR@10. Geometry reranking is heterogeneous: its largest positive early-rank change occurs for 2d, whereas 1f decreases under both policies. Sparse categories do not support inference of systematic benefit.

Table S13. Category-wise geometry-reranking differences relative to embedding-only retrieval.

Label	Queries	Bounding-box delta	Loose-box delta
1a	4	+0.000/+0.000/+0.062	+0.250/+0.000/+0.083
1b	21	-0.048/+0.000/-0.052	+0.095/+0.000/+0.025
1c	6	+0.167/+0.000/+0.125	+0.000/+0.000/-0.042
1d	25	+0.000/+0.040/+0.008	+0.000/-0.040/-0.001
1e	25	+0.040/+0.000/+0.027	+0.000/+0.000/+0.002
1f	9	-0.333/+0.000/-0.272	-0.333/-0.111/-0.270
2a	44	+0.000/+0.023/+0.018	-0.045/+0.023/-0.004
2b	65	-0.015/-0.046/-0.007	+0.062/-0.046/+0.038
2c	113	+0.000/+0.027/-0.003	+0.009/+0.018/+0.001
2d	83	+0.145/+0.000/+0.099	+0.217/+0.012/+0.156
2e	7	+0.000/+0.143/+0.016	-0.143/+0.000/-0.033
2h	5	+0.000/+0.000/+0.000	+0.000/+0.000/+0.000

Overall paired differences in Table S14 use 10,000 CT-grouped resamples with seed 17. Each interval compares the named auxiliary method with embedding-only retrieval on the identical 407 queries. The descriptor-only deficits exclude zero,

whereas every geometry-reranker interval shown below includes zero.

Table S14. Overall paired differences of descriptor controls and geometry reranking relative to embedding-only retrieval on the 407-query cohort.

Query policy	Auxiliary method	Difference in R@1 (95% CI)	Difference in R@10 (95% CI)	Difference in MRR@10 (95% CI)
Bounding box	Descriptor-only	-0.177 (-0.276 to -0.081)	-0.111 (-0.183 to -0.048)	-0.167 (-0.246 to -0.089)
Bounding box	Geometry reranker	+0.022 (-0.019 to +0.065)	+0.007 (-0.012 to +0.026)	+0.016 (-0.012 to +0.044)
Loose box	Descriptor-only	-0.103 (-0.192 to -0.011)	-0.091 (-0.172 to -0.019)	-0.099 (-0.171 to -0.027)
Loose box	Geometry reranker	+0.049 (-0.003 to +0.102)	-0.002 (-0.025 to +0.018)	+0.033 (-0.0004 to +0.066)

S2.8 Category-Wise Lesion-Index Hard Negatives

Table S15 reports the fraction of queries whose first-ranked result mismatches the broad primary category or exact subcategory. It is derived from the same 407-query, rank-20 hard-negative analysis used in Section 5.7 of the main manuscript. The subcategory mismatch column equals one minus subcategory Recall@1, but presenting it beside broad-category mismatch distinguishes cross-family errors from fine-grained within-family confusion. The less selective endpoint of *any* subcategory mismatch within the top 20 occurs for 99.5% of bounding-box queries and 100% of loose-box queries, so it is not repeated by category.

Table S15. Category-wise top-ranked lesion-index mismatch rates by query policy.

Label	Queries	Bounding-box primary mismatch	Bounding-box subcategory mismatch	Loose-box primary mismatch	Loose-box subcategory mismatch
1a	4	0.000	1.000	0.000	1.000
1b	21	0.190	0.429	0.190	0.524
1c	6	0.333	0.333	0.167	0.333
1d	25	0.200	0.960	0.280	1.000
1e	25	0.480	0.840	0.560	0.880
1f	9	0.333	0.556	0.333	0.556
2a	44	0.159	0.500	0.295	0.545
2b	65	0.154	0.692	0.200	0.800
2c	113	0.177	0.442	0.301	0.504
2d	83	0.301	0.506	0.410	0.627
2e	7	0.000	0.429	0.286	0.571
2h	5	0.200	1.000	0.000	0.800

Fine-grained confusion matches or exceeds broad-category confusion for every represented label. Loose boxes increase primary-category mismatch most visibly for 2a, 2c, and 2d, consistent with additional surrounding tissue weakening lesion-specific ordering.

S2.9 Volume-Baseline Uncertainty

Each row of Table S16 is the named method minus global CLS on the identical 50 public-validation CT queries. Intervals use a paired 10,000-resample percentile bootstrap with seed 17. Every mask-aware method improves Recall@1 and MRR@10 relative to CLS with intervals excluding zero, whereas Recall@10 is saturated and does not distinguish the methods. CLS searches all 2,992 processed train CTs, while mask-aware methods search the 2,978 CTs with at least one eligible lesion embedding; the comparison therefore reflects both aggregation policy and this 14-CT inventory difference.

Table S16. FlexiCT-3D subcategory volume-retrieval differences relative to global CLS. Values are the named alternative minus CLS on the identical 50-query cohort.

Alternative to CLS	Difference in R@1 (95% CI)	Difference in R@10 (95% CI)	Difference in MRR@10 (95% CI)
All-disease pooling	+0.200 (+0.020 to +0.380)	+0.000 (-0.060 to +0.060)	+0.120 (+0.011 to +0.227)
Lesion-balanced pooling	+0.240 (+0.080 to +0.400)	+0.000 (-0.060 to +0.060)	+0.154 (+0.057 to +0.251)

Alternative to CLS	Difference in R@1 (95% CI)	Difference in R@10 (95% CI)	Difference in MRR@10 (95% CI)
Finding-balanced pooling	+0.240 (+0.080 to +0.400)	+0.000 (-0.060 to +0.060)	+0.144 (+0.043 to +0.244)
CLS/lesion-balanced fusion	+0.220 (+0.080 to +0.360)	+0.000 (-0.060 to +0.060)	+0.133 (+0.049 to +0.217)

The paired analyses in Sections S2.7 and S2.9 compare each auxiliary method with its specified embedding-only or CLS baseline. They do not test differences among the mask-aware volume methods or among lesion subcategories.

S2.10 Reference-Partition Sensitivity Analysis

The analysis described in Section 4.9 of the main manuscript evaluates the fixed 407-query cohort against two reference pools. The unexposed pool is every eligible reference lesion from a CT-RATE `valid_*` scan: 2,174 lesions from 413 CTs and 383 patients. The exposed pool is the 12,185 eligible reference lesions from CT-RATE `train_*` scans, spanning 2,565 CTs and 2,322 patients, from which matched samples are drawn under two modes over 500 independent draws each, summarized in Table S17.

Table S17. Construction of matched exposed-reference pools across 500 draws by matching mode. Values are means with ranges in parentheses where applicable.

Matching mode	Lesions per draw	CTs per draw	Patients per draw	Subcategory match
Subcategory-frequency (primary)	2,174	1,364 (1,322-1,418)	1,257 (1,216-1,315)	Exact, no shortfall
Whole-CT (sensitivity)	2,177 (2,174-2,198)	458 (420-504)	448 (406-488)	Approximate

Neither mode matches both subcategory frequency and case structure, so both are reported. No reference lesion failed exposure classification in either mode.

Pool construction contributes materially to the single-pool variability: across the 500 primary-mode draws, exposed exact-mask subcategory Recall@10 has a standard deviation of 1.8 percentage points.

Two interval types are reported because they answer different questions. The **95% CI** is a two-level percentile bootstrap for the draw-averaged effect. Each of 10,000 replicates (seed 17) resamples the 50 query CTs, retains all lesions from a sampled CT, independently resamples the 500 pool draws with replacement, and averages those draws within the replicate. This matches the estimand of the draw-averaged point estimate. The **95% single-pool sensitivity interval** instead uses one pool draw per replicate. It describes the spread expected if a single newly constructed matched pool were used rather than the average of many. It is a design-sensitivity interval, not a confidence interval for the averaged effect, and is wider by construction. Across the 30 endpoints, the sensitivity interval is typically about 1.8 times wider than the confidence interval (median 1.77, interquartile range 1.64-1.85, full range 1.37-2.07).

Table S18 reports the primary subcategory-frequency-matched mode and Table S19 the whole-CT sensitivity mode, both under subcategory relevance. Both interval types condition on the unexposed side, which is the complete eligible unexposed bank in this benchmark rather than a sample drawn from it. Differences are exposed-matched minus unexposed, so a positive value indicates an advantage associated with documented reference-side pretraining exposure.

Table S18. Reference-partition subcategory results under subcategory-frequency-matched sampling. Differences are exposed-matched minus unexposed.

Query policy	Metric	Unexposed	Exposed-matched	Difference (95% CI)	95% single-pool sensitivity interval
Exact mask	Recall@1	0.432	0.456	+0.024 (-0.022 to +0.069)	(-0.054 to +0.102)
Exact mask	Recall@5	0.762	0.781	+0.020 (-0.016 to +0.064)	(-0.041 to +0.093)
Exact mask	Recall@10	0.835	0.866	+0.031 (-0.012 to +0.077)	(-0.031 to +0.100)
Exact mask	Recall@20	0.894	0.922	+0.028 (-0.015 to +0.067)	(-0.031 to +0.080)
Exact mask	MRR@10	0.569	0.594	+0.025 (-0.006 to +0.058)	(-0.030 to +0.084)
Bounding box	Recall@1	0.410	0.410	+0.000 (-0.044 to +0.050)	(-0.080 to +0.083)

Query policy	Metric	Unexposed	Exposed-matched	Difference (95% CI)	95% single-pool sensitivity interval
Bounding box	Recall@5	0.762	0.747	-0.015 (-0.046 to +0.019)	(-0.076 to +0.049)
Bounding box	Recall@10	0.850	0.848	-0.002 (-0.029 to +0.028)	(-0.058 to +0.051)
Bounding box	Recall@20	0.894	0.915	+0.020 (-0.016 to +0.057)	(-0.031 to +0.071)
Bounding box	MRR@10	0.554	0.554	-0.000 (-0.034 to +0.039)	(-0.060 to +0.062)
Loose box	Recall@1	0.386	0.359	-0.027 (-0.073 to +0.020)	(-0.107 to +0.052)
Loose box	Recall@5	0.713	0.706	-0.006 (-0.043 to +0.031)	(-0.074 to +0.063)
Loose box	Recall@10	0.816	0.821	+0.006 (-0.031 to +0.047)	(-0.055 to +0.069)
Loose box	Recall@20	0.894	0.906	+0.011 (-0.023 to +0.048)	(-0.041 to +0.063)
Loose box	MRR@10	0.522	0.505	-0.017 (-0.053 to +0.022)	(-0.079 to +0.047)

Table S19. Reference-partition subcategory results under whole-CT sensitivity sampling. Differences are exposed-matched minus unexposed.

Query policy	Metric	Unexposed	Exposed-matched	Difference (95% CI)	95% single-pool sensitivity interval
Exact mask	Recall@1	0.432	0.453	+0.021 (-0.024 to +0.065)	(-0.058 to +0.102)
Exact mask	Recall@5	0.762	0.765	+0.003 (-0.031 to +0.043)	(-0.062 to +0.079)
Exact mask	Recall@10	0.835	0.853	+0.018 (-0.021 to +0.057)	(-0.047 to +0.090)
Exact mask	Recall@20	0.894	0.913	+0.019 (-0.021 to +0.056)	(-0.043 to +0.074)
Exact mask	MRR@10	0.569	0.586	+0.016 (-0.014 to +0.048)	(-0.041 to +0.079)
Bounding box	Recall@1	0.410	0.402	-0.009 (-0.051 to +0.038)	(-0.088 to +0.073)
Bounding box	Recall@5	0.762	0.729	-0.033 (-0.064 to -0.0001)	(-0.098 to +0.034)
Bounding box	Recall@10	0.850	0.832	-0.018 (-0.047 to +0.013)	(-0.078 to +0.040)
Bounding box	Recall@20	0.894	0.904	+0.010 (-0.024 to +0.044)	(-0.047 to +0.062)
Bounding box	MRR@10	0.554	0.540	-0.014 (-0.047 to +0.024)	(-0.077 to +0.050)
Loose box	Recall@1	0.386	0.342	-0.044 (-0.087 to +0.004)	(-0.124 to +0.041)
Loose box	Recall@5	0.713	0.684	-0.029 (-0.066 to +0.009)	(-0.102 to +0.045)
Loose box	Recall@10	0.816	0.803	-0.013 (-0.048 to +0.026)	(-0.078 to +0.053)
Loose box	Recall@20	0.894	0.892	-0.002 (-0.034 to +0.031)	(-0.059 to +0.051)
Loose box	MRR@10	0.522	0.486	-0.036 (-0.073 to +0.004)	(-0.101 to +0.034)

Of the 30 confidence intervals, 29 include zero. One unadjusted confidence interval narrowly excludes zero: bounding-box Recall@5 under whole-CT sampling is -0.033 (95% CI, -0.064 to -0.0001). The direction indicates lower, not higher, performance for exposed references. Because this is an isolated marginal result among 30 correlated endpoints without multiplicity adjustment, the result is treated as exploratory; its 95% single-pool sensitivity interval crosses zero (-0.098 to +0.034), showing that the sign is not robust to construction of a single comparison bank. All 30 95% single-pool sensitivity intervals include zero.

A third conditional interval was also computed but is not tabulated. It holds the draw-averaged pool fixed and targets the same estimand as the reported confidence interval while omitting the Monte Carlo uncertainty of the pool average, which is negligible at 500 draws. As stated in Sections 6.6 and 8.5 of the main manuscript, this analysis supports only the conclusion that no consistent reference-side exposure advantage was detected. It does not establish the absence of an effect and does not address query-side exposure.

S2.11 Additional Endpoint Grids and Paired Encoder Intervals

Tables S20-S26 provide additional endpoint grids and paired fine-grained encoder intervals corresponding to the summaries in Tables 3 and 4 of the main manuscript. They use the same cohorts, relevance definitions, and methods described in the main text; no values were recomputed for the reorganization.

Table S20. Patient-grouped five-fold training-set exact-mask retrieval.

Model	Level	R@1	R@5	R@10	R@20	MRR@10
FlexiCT-3D	Primary category	0.812 ± 0.009	0.957 ± 0.003	0.980 ± 0.004	0.991 ± 0.003	0.876 ± 0.007
FlexiCT-3D	Subcategory	0.495 ± 0.011	0.802 ± 0.017	0.880 ± 0.014	0.929 ± 0.008	0.625 ± 0.014
FlexiCT-3D-VLM	Primary category	0.815 ± 0.010	0.956 ± 0.005	0.980 ± 0.003	0.990 ± 0.002	0.878 ± 0.007
FlexiCT-3D-VLM	Subcategory	0.499 ± 0.018	0.803 ± 0.014	0.879 ± 0.013	0.927 ± 0.010	0.629 ± 0.016

Table S21. Public-validation encoder ablation with subcategory relevance.

Query policy	Encoder	R@1	R@5	R@10	R@20	MRR@10	Interpretation
Exact mask	FlexiCT-3D	0.484	0.786	0.848	0.926	0.614	Primary image-only result
Exact mask	FlexiCT-3D-VLM	0.506	0.776	0.857	0.916	0.625	No clear paired difference
Bounding box	FlexiCT-3D	0.430	0.759	0.840	0.916	0.568	Controlled coarse-query result
Bounding box	FlexiCT-3D-VLM	0.445	0.747	0.833	0.921	0.580	No clear paired difference
Loose box	FlexiCT-3D	0.356	0.693	0.821	0.897	0.500	Localization stress-test result
Loose box	FlexiCT-3D-VLM	0.398	0.690	0.813	0.907	0.526	No clear paired difference

Table S22. Public-validation descriptor control and geometry reranking with FlexiCT-3D and subcategory relevance.

Query policy	Method	R@1	R@5	R@10	MRR@10	Precision@10
Bounding box	Descriptor-only	0.253	0.597	0.730	0.401	0.280
Bounding box	Embedding-only	0.430	0.759	0.840	0.568	0.397
Bounding box	Geometry reranker	0.452	0.767	0.848	0.584	0.427
Loose box	Descriptor-only	0.253	0.597	0.730	0.401	0.280
Loose box	Embedding-only	0.356	0.693	0.821	0.500	0.354
Loose box	Geometry reranker	0.405	0.703	0.818	0.533	0.378

Table S23. FlexiCT-3D volume-level retrieval baselines with subcategory relevance.

Baseline	Subcategory R@1	R@5	R@10	MRR@10
Analytic chance expectation	0.480	0.916	0.971	—
Global CLS	0.620	0.960	0.980	0.763
All-disease pooling	0.820	0.960	0.980	0.883
Lesion-balanced pooling	0.860	0.980	0.980	0.917

Baseline	Subcategory R@1	R@5	R@10	MRR@10
Finding-balanced pooling	0.860	0.980	0.980	0.907
CLS/lesion-balanced fusion	0.840	0.940	0.980	0.896

Table S24. FlexiCT-3D lesion-slice retrieval against 45,402 candidate slabs.

Level	R@1	R@5	R@10	R@20	MRR@10	Precision@10
Primary category	0.806	0.948	0.980	0.995	0.864	0.793
Subcategory	0.459	0.740	0.806	0.880	0.577	0.453

Table S25. Overall FlexiCT-3D box-query failure rates from the lesion-index and annotation-negative-bank analyses.

Query policy	Top-1 primary mismatch	Top-1 subcategory mismatch	Negative above top-1 lesion	Negative within margin
Bounding box	0.219	0.570	0.000	0.017
Loose box	0.307	0.644	0.074	0.111

Table S26. Paired public-validation subcategory encoder differences, FlexiCT-3D-VLM minus FlexiCT-3D.

Query policy	Difference in R@1 (95% CI)	Difference in MRR@10 (95% CI)
Exact mask	+0.0221 (-0.0196 to +0.0634)	+0.0115 (-0.0198 to +0.0398)
Bounding box	+0.0147 (-0.0257 to +0.0569)	+0.0121 (-0.0210 to +0.0453)
Loose box	+0.0418 (-0.0098 to +0.0912)	+0.0263 (-0.0083 to +0.0593)

S3. Supplementary Qualitative Retrieval Panels

Figures S1 and S2 provide the two supplementary examples from the four-case qualitative set introduced in Section 7 of the main manuscript. Figure S1 illustrates exact-mask lesion retrieval from Section 6.1, while Figure S2 illustrates lesion-slice retrieval from Section 6.4.

All four cases were selected from recorded results before image inspection. Their queries belong to the common 407-query public-validation cohort, each rank-1 result comes from a CT different from the query CT, and the set spans four subcategories (`2d`, `1b`, `2b`, and `2c`) and two lesion-size strata. Query panels mark query regions, and reference panels mark indexed lesions. Frames and text labels indicate subcategory agreement or mismatch. Lesion-level panels use the axial slice with the largest mask area. Lesion-slice candidates use the largest-mask-area slice within the retrieved 8-voxel slab.

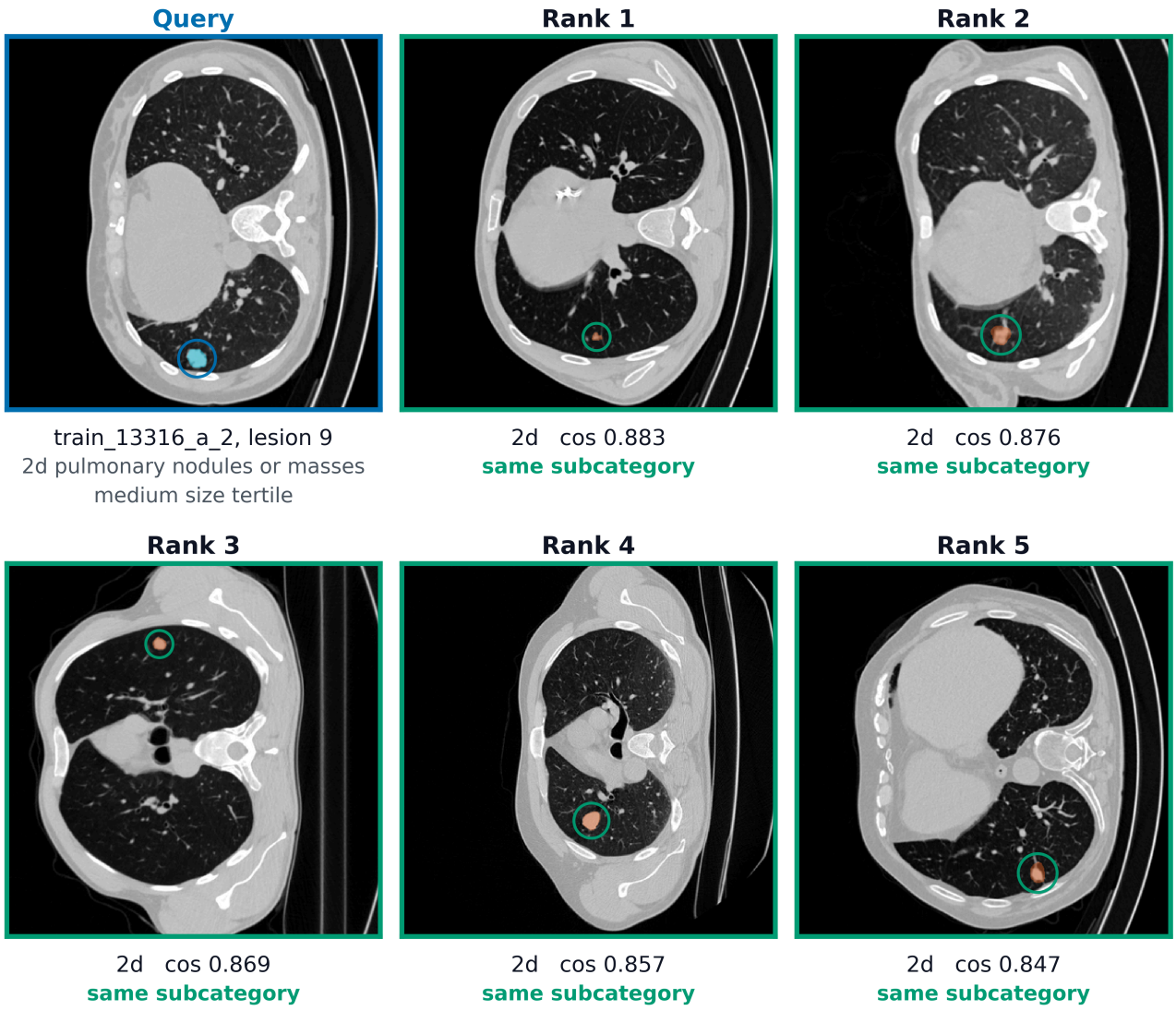


Fig. S1 Exact-mask retrieval with five same-subcategory results. A medium-size 2d pulmonary-nodule query (lesion 9, `train_13316_a_2`) retrieves five same-subcategory lesions from five distinct CTs, with cosine scores from 0.883 to 0.847. Both the query and reference embeddings are derived from exact masks; thin rings locate small masks. This figure adapts CT-RATE images from Hamamci et al. [6] and is distributed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/); the author windowed the images and added ReXGroundingCT-derived overlays and labels [1]

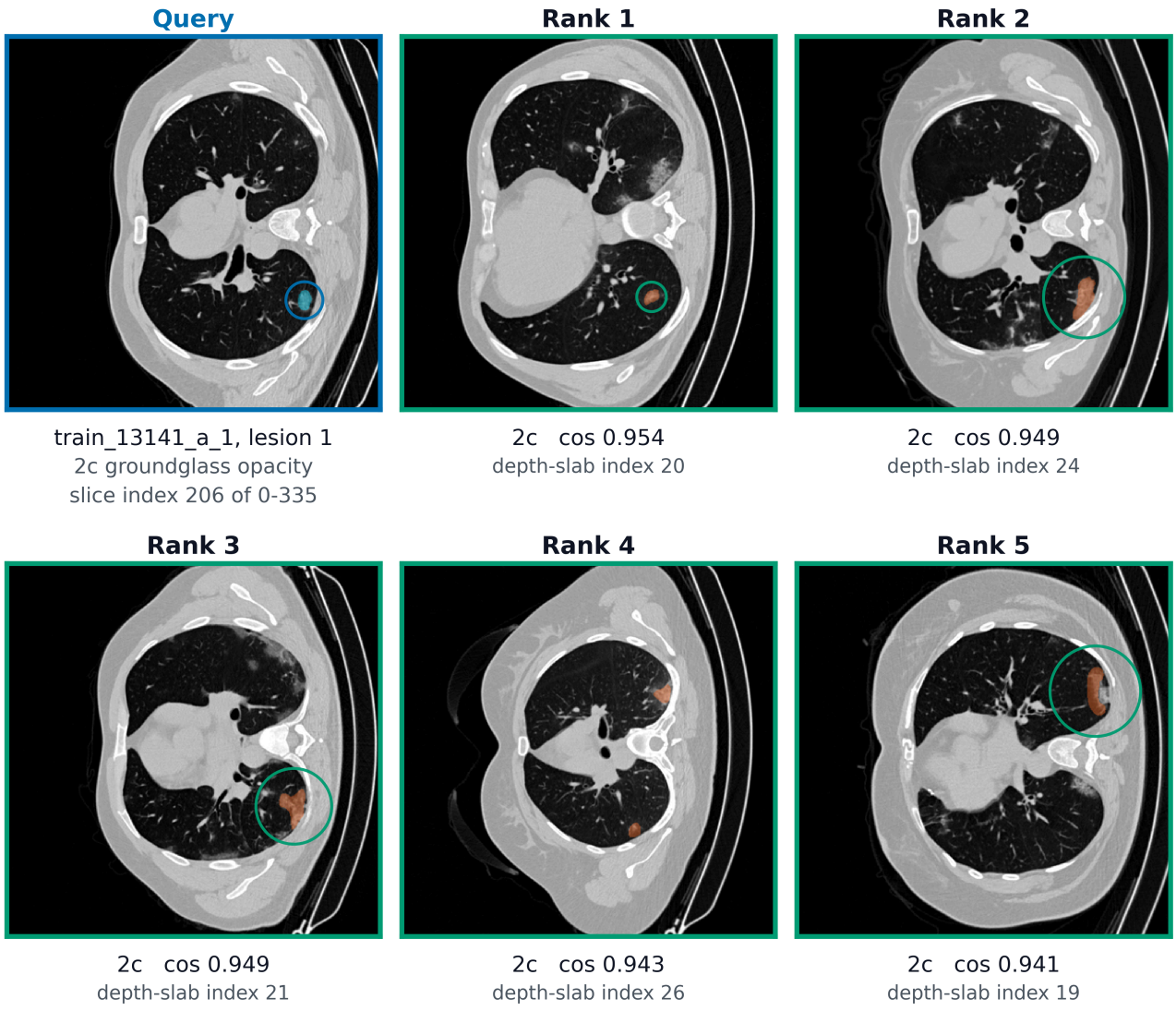


Fig. S2 Exact-mask lesion-slice retrieval. An exact-mask query for a `2c` groundglass-opacity lesion (lesion 1, `train_13141_a_1`) uses native axial slice 206 (zero-based range, 0-335). Its five highest-ranked candidates share the query subcategory and come from five CTs within a bank of 45,402 lesion-slice candidates. Each candidate represents one 8-voxel depth slab and is displayed on the slice with the largest mask area within that slab. All slice and depth-slab indices are zero-based. The spatially dispersed rank-4 mask is shown without a locator ring. This figure adapts CT-RATE images from Hamamci et al. [6] and is distributed under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/); the author windowed the images and added ReXGroundingCT-derived overlays and labels [1]

These panels illustrate already quantified behavior but provide no additional statistical evidence. They do not establish radiologist-assessed similarity or educational benefit.

References

Reference numbering follows the main manuscript.

1. Baharoon M, Luo L, Moritz M, Kumar A, Kim SE, Zhang X, Zhu M, Alabbad MH, Alhazmi MS, Mistry NP, Bijnens L, Kleinschmidt KR, Chrisler B, Suryadevara S, Jaliparthi SSD, Prudlo NM, Marino MD, Palacio J, Akula R, Zhou D, Zhou HY, Hamamci IE, Adams SJ, AlOmaish HR, Rajpurkar P: ReXGroundingCT: a 3D chest CT dataset for segmentation of findings from free-text reports. arXiv preprint arXiv:2507.22030, 2025. https://doi.org/10.48550/arXiv.2507.22030

6. Hamamci IE, Er S, Wang C, Almas F, Simsek AG, Esirgun SN, Dogan I, Durugol OF, Hou B, Shit S, Dai W, Xu M, Reynaud H, Dasdelen MF, Wittmann B, Amiranashvili T, Simsar E, Simsar M, Erdemir EB, Alanbay A, Sekuboyina A, Lafci B, Kaplan A, Lu Z, Polacin M, Kainz B, Bluethgen C, Batmanghelich K, Ozdemir MK, Menze B: Generalist foundation models from a multimodal dataset for 3D computed tomography. Nat Biomed Eng, DOI: 10.1038/s41551-025-01599-y, February 12, 2026

10. Li Y, Gao Y, Dong H, Lai Y, Wang S, Safari M, Baciak JE, Yang X: Universal CT representations from anatomy to disease phenotype through agglomerative pretraining. arXiv preprint arXiv:2605.21906, 2026. https://doi.org/10.48550/arXiv.2605.21906