Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Zhongsheng Eric Yu

peng.sheng.2023@gmail.com

Research Article

Keywords: content-based image retrieval, computed tomography, lesion retrieval, foundation models, medical imaging informatics, patch embeddings

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10854370/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

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

Author: Zhongsheng Eric Yu

Affiliation: Independent researcher, San Diego, California, USA

Corresponding author: Zhongsheng Eric Yu (peng.sheng.2023@gmail.com)

Abstract

Background: Computed tomography (CT) education benefits from prior examinations containing related localized findings. Whole-volume retrieval can be driven by global anatomy or unrelated findings, while database-specific retrieval models limit reuse. This study evaluates whether patch tokens from a frozen volumetric CT foundation encoder can support a reusable lesion-grounded retrieval framework without task-specific retrieval fine-tuning.

Methods: Spatial patch tokens were pooled into lesion representations linked to their parent examination, annotation, and source finding text. ReXGroundingCT served as the demonstration benchmark. The primary evaluation used FlexiCT-3D to search 14,359 reference lesions with 407 public-validation queries from 50 CT examinations. Exact masks, bounding boxes, and loose boxes tested sensitivity to query-region precision. Retrieval was evaluated against ontology-based relevance, analytic chance expectations, and annotation-informed descriptor controls.

Results: For exact-mask queries, exact-subcategory Recall@1 was 0.484, more than three times the analytic chance expectation of 0.154, while Recall@10 was 0.848 versus 0.729. Bounding-box and loose-box queries gave Recall@10 of 0.840 and 0.821, respectively, but less precise query regions lowered early-rank performance. Under both box policies, embedding retrieval outperformed the descriptor-only control.

Conclusions: A reusable lesion-retrieval layer can be built from frozen volumetric foundation features without task-specific retrieval training. The layer retains links to source finding text without using text for ranking and can

organize compatible medical-image databases into lesion-centered ranked lists and localized review panels for educational case comparison.

Keywords

content-based image retrieval; computed tomography; lesion retrieval; foundation models; medical imaging informatics; patch embeddings

1. Background

Educational review of computed tomography (CT) findings can benefit from access to prior cases containing related localized abnormalities. A representation optimized for overall study similarity can be too coarse when the finding of interest is a specific lesion, nodule, or localized abnormality, because normal anatomy, acquisition properties, background disease burden, or unrelated findings may dominate the comparison.

This study presents and evaluates a lesion-grounded content-based retrieval framework for CT volumes. Instead of representing only an entire CT examination, the framework treats each annotated reference lesion as a searchable entity while preserving its link to the parent CT case, finding record, mask, and clinical description. A frozen volumetric CT foundation encoder produces spatial patch-token features. Lesion masks are projected to the encoder patch grid, lesion-associated patch tokens are pooled into reusable lesion-level embeddings, and these embeddings form the reference retrieval bank. The indexing and retrieval design is not tied to one dataset. It can be applied to other medical-image databases that provide compatible images, a patch-token encoder, and either localized reference annotations or an upstream localization method.

The target application is educational, lesion-centered inter-patient case comparison. Given a query image and a user-provided region of interest, the system retrieves related reference lesions and their parent CT cases, so a user can compare lesion-centered slices, masks, descriptors, and finding text side by side. This is an intended educational workflow rather than a demonstrated educational benefit.

The framework uses query and reference regions asymmetrically. Query regions identify the finding of interest; precise reference masks define stable indexed lesion entities. Exact masks, tight boxes, and loose boxes test how retrieval changes with localization precision. Exact masks provide an annotation-derived upper bound, tight boxes

approximate regions obtained from automated segmentation or detection, and loose boxes approximate coarse manual selection. These controlled conditions do not prescribe a deployed query interface. Here, regions serve as retrieval inputs and evaluation annotations, not diagnostic predictions. The system has not been designed or validated for autonomous diagnosis, treatment decisions, or fully automated lesion classification. The study asks whether spatial features from a frozen volumetric CT foundation encoder, without task-specific retrieval training, retain enough fine-grained information for category-consistent lesion retrieval.

The study's contributions center on the lesion-indexed retrieval framework and its evaluation:

1. A lesion-indexed CT retrieval framework that converts patch tokens from a frozen volumetric foundation encoder into lesion-level embeddings without task-specific fine-tuning, uses a localized finding as the retrieval query, and groups retrieved lesions by parent CT case for review.
2. A sparse artifact pipeline that derives lesion, volume, and lesion-slice representations from lesion-overlapping patch tokens without repeated encoder inference.
3. A controlled robustness evaluation using exact masks, tight bounding boxes, and loose boxes as progressively less precise query-localization conditions.
4. A validation suite with patient-grouped development folds and CT-grouped retrieval evaluation, including public-validation testing, descriptor and volume baselines, lesion-slice retrieval, failure analysis, and a matched reference-partition sensitivity analysis.
5. An encoder ablation comparing the primary image-only FlexiCT-3D encoder with FlexiCT-3D-VLM to test whether report-aligned VLM pretraining changes retrieval behavior.

ReXGroundingCT provides the study benchmark because it links free-text chest CT findings to lesion- or entity-level three-dimensional segmentation masks rather than only study-level labels [1]. This makes it possible to test whether a query finding retrieves spatially grounded reference lesions with related labels, rather than only whether a retrieved examination contains a broadly similar label somewhere in the volume.

2. Related Work

Four strands of prior work motivate this study: the appropriate retrieval unit; reusable localized 3D features from frozen foundation models; representations of indexed findings and query regions; and educational similar-case

retrieval with reader-based evaluation. The following sections compare methods according to these design questions and the gaps they leave.

2.1 From Global Medical CBIR to Lesion-Level Retrieval

CBIR ranks database items by similarity between query and candidate representations. Deep learning has shifted the field from handcrafted color, texture, and shape descriptors toward learned representations, feature aggregation, attention, and region selection [2]. In medical imaging, clinically meaningful similarity may depend on subtle localized findings. The appropriate retrieval unit may therefore be a lesion or region rather than an entire image or examination [3,4]. Reviews identify several recurring challenges. Global descriptors can dilute small abnormalities, key-slice methods can lose volumetric context, precise ROI definition increases interaction or annotation burden, and heterogeneous datasets complicate evaluation [3,4].

These limitations motivate indexing a localized finding while retaining its volumetric context. The key representation question is whether useful regional embeddings require database-specific retrieval learning. Localization precision is evaluated as a secondary robustness factor, not prescribed as an interaction design.

2.2 Volumetric Foundation Models and Local Patch Representations

3D-MIR compared slice, volume, and multimodal retrieval across four CT anatomies [5]. CT-RATE and CT-CLIP established image- and text-query whole-volume retrieval on paired chest CTs and reports [6], while RadIR added report-derived, anatomy-conditioned similarity [7]. These systems operate at examination level rather than indexing findings specified by a user-defined three-dimensional region.

Frozen foundation representations can support CBIR without task-specific fine-tuning, although pathological retrieval remains harder and specialist models may remain stronger [8]. BiomedCLIP extends such cross-modal retrieval across two-dimensional biomedical images [9]. For volumetric CT, FlexiCT progresses from anatomical pretraining to report-aligned vision-language learning and supports several downstream tasks [10]. These include tumor-phenotype and semantic image-report retrieval. CT-FM independently demonstrates broad transfer from image-only volumetric pretraining [11]. These studies motivate CT-native frozen features but do not establish lesion-level retrieval from their spatial tokens.

Localized pooling provides a bridge between global and lesion-level representations. In chest radiography, bounding-box-restricted patch tokens can preserve small-lesion information that global aggregation suppresses [12]. Studies of CT adaptation and DINOv3 further show that foundation-feature transfer remains modality dependent [13,14]. This motivates testing whether CT-native 3D tokens overlapping a lesion can form a reusable lesion-retrieval layer without task-specific training. FlexiCT-3D serves as the primary image-only encoder, while FlexiCT-3D-VLM provides the report-aligned ablation. Text is excluded from retrieval scoring.

2.3 Lesion-Level and Region-Based Retrieval Methods

Earlier pulmonary-nodule CBIR systems retrieved segmented CT nodules using engineered shape, margin, and texture features with class-aware similarity strategies [15]. Nodule2vec later represented 3D pulmonary nodules as learned low-dimensional embeddings that preserved semantic information for similarity search [16]. It provided a task-specific bridge between engineered descriptors and more recent foundation representations. RadiomicsRetrieval learned a multimodal representation that aligns point-prompted tumor image features with radiomics and anatomical-position embeddings [17]. It supports image, location, and partial-feature queries in lung CT and brain MRI. ReXGroundingCT provides a complementary resource in which report-derived findings are linked to category metadata and 3D masks [1]. Together, these studies establish the value of spatially grounded retrieval entities. They do not determine whether a frozen CT representation can support lesion-level image-to-image retrieval without retrieval-specific training or radiomics in the primary score.

Exact-mask, bounding-box, and loose-box query conditions provide controlled tests of sensitivity to localization precision. They do not prescribe how a deployed query interface should be designed. The retrieval setting remains asymmetric: query regions identify the finding of interest, while precise reference masks define stable indexed entities. Geometry is kept separate as a transparent baseline or reranking signal, in contrast to methods that learn anatomical position inside the retrieval embedding [17].

2.4 Educational and Human-Centered Similar-Case Review

Human relevance is more complex than label agreement. A perceptual study of focal CT liver lesions found substantial inter-reader variability in judgments of visual similarity [18]. Category-based metrics offer reproducible proxies but do not fully capture perceived similarity.

Medical image repositories have long been considered resources for interpatient comparison, teaching, and research [4]. In a controlled reader study, top-three deep-learning-based CT retrieval improved diagnostic accuracy and inter-reader agreement for interstitial lung disease [19]. An ROI-based similar-patient search linked to clinical reference content also improved residents' performance when reviewing rare chest CT diseases [20]. This result supports region-based case review as a workflow but does not establish benefit for the present system. A later radiologist evaluation of ROI-based HRCT retrieval reported perceived educational and workflow benefits [21]. Retrieval accuracy nevertheless depended on the number and placement of supplied ROIs. Radiologist-ranked retrieval studies have additionally used nDCG and inter-reader comparisons to evaluate whether algorithmic rankings align with clinical similarity judgments [22].

These studies motivate inspectable top-k case review. They do not, however, validate this system or show that automatic retrieval metrics predict reader benefit. Within this human-centered literature, reference banks are generally indexed at the examination or ROI level, not as individually annotated lesion entities linked to parent cases. The present study contributes such a lesion-indexed representation and evaluates it automatically. Determining whether its ranked lists improve education or reader performance requires a separate user study.

2.5 Synthesis and Study Positioning

Prior work spans whole-volume and anatomy-conditioned CT retrieval [5-7], off-the-shelf and CT-native foundation representations [8-14], lesion- and tumor-level retrieval [15-17], and educational and human-centered relevance and interaction [18-22]. However, it has not established a reusable lesion-indexed image-to-image retrieval layer derived from frozen CT-native 3D patch tokens. The layer developed here retains links to parent cases and contextual information and can be adapted to compatible medical-image databases without retrieval-specific training. The evaluation covers lesion and volume retrieval, robustness across regional query conditions, and image-only versus report-aligned encoders. Sparse multi-granularity artifacts are implementation components of this system, not a separate claim about the literature.

3. Materials and Dataset

3.1 Dataset

ReXGroundingCT provides the demonstration dataset. It is a gated-access chest CT grounding dataset that links free-text radiology findings to pixel-level 3D segmentation masks [1]. ReXGroundingCT was curated from CT-RATE, which pairs 25,692 non-contrast chest CT scans from 21,304 patients with radiology reports [6]. ReXGroundingCT contains 3,142 scans, 8,028 finding-segmentation pairs, and 16,301 annotated entities across its reported train, public-validation, and private-test partitions [1]. Acquisition characteristics vary widely. Slice thickness ranges from 0.50 to 5.00 mm, axial pixel spacing from 0.30 to 0.98 mm, and volume depth from 104 to 1005 slices [1].

ReXGroundingCT is well suited to lesion-centered retrieval because each selected finding is linked to text metadata and a 3D spatial mask. Its findings include both focal and non-focal pulmonary and pleural abnormalities. Approximately 79% are focal and 21% are non-focal. The 14-category ontology includes localized findings such as pulmonary nodules or masses and broader abnormalities such as emphysema [1]. These properties allow the study to test retrieval of localized findings rather than only study-level disease labels.

Annotation, metadata, and segmentation files were obtained from the gated Rajpurkar Lab ReXGroundingCT repository. Corresponding CT volumes were obtained separately from the gated CT-RATE repository. Both resources were used solely for non-commercial research and are cited accordingly [1,6].

CT-RATE volume identifiers follow the pattern `split_patientID_scanID_reconstructionID`. ReXGroundingCT retains these identifiers, so the original CT-RATE partition can be recovered for every scan. All 50 public-validation scans come from the CT-RATE train partition. The ReXGroundingCT train split contains 2,578 scans from the CT-RATE train partition and 414 from the CT-RATE validation partition. This composition affects two later analyses. First, the public-validation query volumes belong to the partition used to pretrain the frozen encoder and therefore do not provide external validation (Section 4.3). Second, the 414 train-split scans originating from CT-RATE validation support the reference-side exposure analysis in Section 4.9; 413 of them contribute eligible reference lesions. Patient identifiers are also recoverable. They are keyed by partition because `train_13082` and `valid_13082` denote different patients.

The processing pipeline distinguishes source-dataset annotations from task-specific retrieval artifacts. A *finding* is a report-derived textual observation paired with one segmentation channel. An *entity* is a separately labeled spatial instance within that channel, so one finding can contain multiple lesion entities [1]. Each unique positive integer label within a finding channel defines one lesion entity and can produce at most one lesion embedding. When the metadata entity count disagrees with the labels present in a mask, the observed mask labels serve as the source of truth for spatial processing. The source masks contain 14,557 train entities and 429 public-validation entities. All 2,992 train CTs were processed successfully. At the selected 0.05 occupancy threshold, 2,978 CTs contributed 14,359 eligible reference embeddings. The remaining 14 CTs contained no eligible lesion embedding. The public-validation set contains 407 eligible query embeddings. Thus, the source-mask counts and retrieval denominators describe different processing stages, not conflicting dataset sizes.

The experiments use only the ReXGroundingCT train and public-validation splits. Private-test data and leaderboard labels are not used. The public-validation set contains 50 CT scans and 407 evaluated lesion queries. The train reference bank contains 14,359 derived lesion embeddings from 2,978 CTs. For five-fold development, the train split is divided into patient-grouped folds. Finding text is retained for metadata and review panels. It does not contribute to the primary image-retrieval score. Table 1 summarizes the dataset inventory and derived retrieval artifacts at each counting stage.

Table 1. Dataset and retrieval artifact summary.

Inventory / artifact	CTs	Findings / entities / artifacts	Counting stage
Published ReXGroundingCT dataset	3,142	8,028 findings; 16,301 entities	All reported partitions; source-dataset statistics [1]
Train source-mask inventory	2,992	14,557 observed entities	Unique positive mask labels before retrieval filtering
Public-validation source-mask inventory	50	429 observed entities	Unique positive mask labels before retrieval filtering
Train lesion reference bank	2,978	14,359 derived lesion embeddings	Eligible at threshold 0.05; 14 processed train CTs contributed no eligible lesion row

Inventory / artifact	CTs	Findings / entities / artifacts	Counting stage
Public-validation query set	50	407 evaluable lesion queries	Queries scored under each of the three query policies

Retrieval eligibility requires a nonzero mask, successful preprocessing and embedding generation, an available category label, at least one retained patch, and at least one valid reference positive. The resulting 2,978-CT lesion reference bank is a thresholded subset of the 2,992 processed train CTs, not evidence of extraction failures. The global-embedding volume baseline retains all 2,992 CTs. Mask-derived volume baselines use the 2,978 CTs with at least one eligible lesion embedding.

3.2 ReXGroundingCT Annotation Protocol and Implications

ReXGroundingCT combines report-derived finding text, hierarchical abnormality categories, and volumetric segmentations. Its construction pipeline first standardized CT-RATE reports. Large language models then extracted and categorized findings under radiologist-guided quality control. Selected pulmonary and pleural findings were then segmented in 3D [1]. The dataset also includes anatomical reasoning annotations, but these are not used in the present retrieval experiments. Board-certified radiologists annotated the public-validation set. Professional annotators and supervised medical students annotated the train split under radiologist quality-control procedures [1].

This annotation design has direct implications for retrieval evaluation. Public-validation cases provide higher-confidence query masks. These cases were excluded from threshold selection and reference-bank construction and serve as the fixed evaluation set. The term *retrieval-held-out* refers only to this separation from retrieval development. All 50 public-validation scans belong to the CT-RATE partition used to pretrain the frozen encoder, so they are not held out from encoder pretraining. The train split provides a large searchable reference bank but may have partial instance coverage. Train annotations include at most three representative instances per finding, whereas public-validation annotations are exhaustive under the ReXGroundingCT protocol [1]. Metrics therefore measure performance against the available train annotations. They do not assume that every potentially matching lesion in CT-RATE has been indexed.

3.3 Labels and Evaluation Levels

ReXGroundingCT provides hierarchical finding categories and text descriptions [1]. Metrics are reported at two levels: primary category and exact subcategory. Primary-category metrics measure broad agreement but provide limited information because the benchmark contains only two imbalanced parent categories. Exact-subcategory metrics define the stricter fine-grained target and are emphasized in the main results.

The evaluated primary categories are `1`, *typically non-focal lung, airway, or pleural abnormalities*, and `2`, *typically focal lung, airway, or pleural opacities*. A primary-category positive shares the first digit of the category code with the query. An exact-subcategory positive shares the complete two-character code. The focality descriptors are inherited ontology group names rather than case-specific judgments about every lesion. The benchmark contains 14 child labels. Parent categories `3-12` from the broader ReXGroundingCT report-extraction ontology are not evaluated because the spatially annotated release is restricted to pulmonary and pleural findings [1]. See Additional file 1, Section S1.1 and Table S1 for definitions and public-validation support for all evaluated labels.

The public-validation distribution is highly imbalanced. Groundglass opacity (`2c`) and pulmonary nodules or masses (`2d`) account for 196 of the 407 queries. Five represented subcategories contain fewer than 10 queries. Honeycombing (`2f`) and pneumothorax (`2g`) have no public-validation queries. Estimates for rare labels are unstable, and validation performance cannot be estimated for the two zero-support labels. These labels remain present in the train reference ontology. See Additional file 1, Section S2.6 for category-specific support and results. Broad primary-category performance and rare-label estimates are not interpreted as evidence of uniform performance across the ontology.

Category labels were inherited from ReXGroundingCT and were not independently re-adjudicated for this retrieval study [1]. The dataset authors report categorization precision and recall of 0.99 across the full 12-category ontology [1]. For 100 audited findings from the two parent categories evaluated here, they report a subcategory F1 score of 0.92. They also identify residual false negatives and descriptor omissions in earlier report-processing stages [1]. Remaining uncertainty may arise from report translation, automated finding extraction and categorization, omitted descriptors, and visually or conceptually overlapping classes. The `2a` label includes subsegmental atelectasis, scarring, and fibrosis, whereas `2b` separately includes atelectasis or consolidation. Distinguishing micronodules

(1e) from pulmonary nodules or masses (2d) also depends on pattern and scale. These distinctions may not always be recoverable from local appearance alone. Exact-subcategory agreement is a standardized benchmark relevance label, not proof of diagnostic equivalence or radiologist-assessed similarity.

4. Methods

4.1 Overview

The framework builds reusable lesion representations from frozen volumetric patch tokens and retrieves lesions using a localized query region. During offline indexing, each reference CT is encoded once. Its lesion masks are then projected onto the patch grid, and the overlapping tokens are pooled into normalized lesion embeddings. Each embedding remains linked to its source CT and annotation metadata. At query time, the method pools tokens inside the specified query region and compares the resulting embedding with the indexed reference-lesion embeddings using cosine similarity. Retrieved lesions remain linked to their parent CTs and localized slices for case review.

Figure 1 separates the indexing pass over the reference bank from the retrieval pass performed for each query. Lesion-level retrieval is the primary task. Volume retrieval provides case-level baselines, while lesion-slice retrieval supports localized review panels. Descriptor controls, geometry reranking, hard-negative analysis, and annotation-negative retrieval show what information the image embeddings capture.

Fig. 1 Lesion-grounded retrieval workflow. (a) Each reference CT is encoded once by frozen FlexiCT-3D. Exact lesion masks are projected onto the 8 x 8 x 8-voxel patch grid, and qualifying tokens are pooled into normalized lesion embeddings. (b) Tokens within an exact- or box-defined query region are pooled from the encoded query CT and compared with the exact-mask reference embeddings by cosine similarity. Rankings are lesion-level; retrieved lesions can be grouped by parent CT for presentation. No component is fine-tuned or used for learned metric reranking. Figure 2 details the mask projection

4.2 CT Preprocessing and Mask Projection

CT volumes were processed on their native NIfTI voxel grids without spatial resampling or cropping. Intensities were converted to float32, clipped to [-1000, 1000] HU, and z-score normalized within each volume. Each spatial dimension was then padded symmetrically to the next multiple of eight. Padding used the minimum normalized

intensity in that volume as its constant value. See Additional file 1, Section S1.2 for edge-case handling and recorded preprocessing metadata.

Segmentation masks were required to match the native CT shape. They were converted from the stored [H, W, D] array order to the model-facing [D, H, W] order and padded with the same offsets as the corresponding CT. Each encoder patch contains 8 x 8 x 8 voxels. Lesion occupancy was defined as the proportion of those 512 voxels belonging to the lesion. Because the CT volumes were not resampled, the physical size represented by one patch varies across scans.

Figure 2 summarizes the projection from a binary entity mask to occupancy-weighted patch-token pooling. Its occupancy values are illustrative; the threshold and pooling rule correspond to the primary retrieval configuration.

Fig. 2 Mask-to-patch projection. Lesion masks are aligned with the 8 x 8 x 8-voxel patch grid. Tokens from patches with lesion occupancy of at least 0.05 are pooled using occupancy weights and normalized. The two-dimensional schematic represents the three-dimensional operation; category labels do not enter the embedding score

4.3 Encoders

FlexiCT-3D [10], a volumetric image encoder produced through 3D anatomical pretraining, serves as the primary frozen encoder. This choice allows the framework to operate on databases containing CT images and lesion annotations without requiring paired reports. The image encoder from FlexiCT-3D-VLM provides an ablation of report-aligned pretraining. Both encoders use the released model definitions and weights, remain frozen in evaluation mode, and receive no fine-tuning on ReXGroundingCT. See Additional file 1, Section S1.2 for checkpoint identifiers and implementation details.

Each encoder receives a single-channel CT tensor with shape [B, 1, D, H, W], preprocessed and padded as specified in Section 4.2, without model-specific resizing, resampling, or cropping. The retrieval experiments use visual outputs only. For FlexiCT-3D-VLM, the three-dimensional vision tower and image-projection layer are loaded, but the text tower is neither instantiated nor scored. ReXGroundingCT text is retained for metadata display and label-based evaluation but does not contribute to retrieval scores.

Both vision backbones produce normalized 864-dimensional CLS and spatial patch tokens from non-overlapping 8 x 8 x 8-voxel patches. Occupancy-weighted pooling of spatial tokens therefore produces 864-dimensional lesion and lesion-slice embeddings for both encoders. FlexiCT-3D uses its normalized 864-dimensional CLS token as its global representation. FlexiCT-3D-VLM concatenates its CLS token with the mean spatial token. The released image-projection layer converts this combination into a normalized 1,024-dimensional global image embedding. This representation is used only for the corresponding volume-level baseline, not for lesion pooling.

Because native CT dimensions are retained, the patch grid is sample-dependent. For a padded processed shape (D, H, W), the grid shape is (D/8, H/8, W/8). Preprocessing metadata maps each token back to the native CT and lesion masks. See Additional file 1, Section S1.2 for numerical execution details and observed grid ranges.

4.4 Sparse Reference Embedding Artifacts

During reference indexing, the full patch grid was retained only long enough to extract lesion-associated records and global embeddings. Each sparse record stores the patch embedding, lesion occupancy, and grid location. It also contains identifiers linking the patch to its lesion, finding, mask, parent CT, and preprocessing metadata. The canonical retrieval identifier is `lesion_entity_id`; `finding_id`, `mask_id`, and `sample_id` preserve the relationships among lesions, annotations, and parent CTs.

The full patch grid was then discarded by default. The stored sparse records can regenerate lesion, volume, and lesion-slice embeddings under different thresholds and pooling policies without another encoder pass. This design keeps the retrieval and pooling steps auditable without storing dense whole-volume tokens. It does not support dense feature-based registration, which lies outside this study. See Additional file 1, Section S1.2 for the complete artifact inventory and field-level schema.

4.5 Lesion and Volume Embedding Construction

For each lesion, the method selects patches with occupancy at or above $\tau = 0.05$. Occupancy-weighted mean pooling followed by L2 normalization produces the lesion embedding. This representation is fixed for every reference lesion across the exact-mask, bounding-box, and loose-box experiments. Exact-mask queries use the same construction, while the two box policies use the support and weighting defined in Section 4.6.

Let F_i^p be the patch grid for CT volume i:

$$F_i^p \in \mathbb{R}^{Z_i \times H_i \times W_i \times D_p},$$

where Z_i, H_i, and W_i are patch-grid dimensions rather than raw voxel dimensions. Let M_{ij} denote the projected mask for lesion j in CT i. Let w_{ijp} denote the fraction of patch p occupied by that lesion. The lesion embedding is computed as a normalized weighted mean:

$$\rho_{ij} = \mathrm{norm}\left(\frac{\sum_{p:w_{ijp} \geq \tau} w_{ijp}\, F_i^p(p)}{\sum_{p:w_{ijp} \geq \tau} w_{ijp} + \epsilon}\right), \qquad \tau = 0.05.$$

Here, norm denotes L2 normalization and ϵ is a small numerical stabilizer.

Only threshold-passing patches contribute to the derived embedding. The sparse source artifact retains `mask_fraction_in_patch`, allowing alternative thresholds and pooling strategies to be evaluated without rerunning the encoder.

The 0.05 occupancy threshold was first selected on training data, without looking at public-validation performance. After the patient-identifier audit, it was re-evaluated with patient-grouped five-fold splits while keeping the candidate thresholds, endpoint, pooling method, seed, query policy, and coverage rationale fixed. Performance was nearly tied. Subcategory Recall@10 was only 0.0002 higher at 0.25 than at 0.05. However, 0.25 retained 11,432 lesions from 2,623 CTs, whereas 0.05 retained 14,359 lesions from 2,978 CTs. The patient-grouped analysis therefore retained 0.05 as the coverage-specificity choice. See Additional file 1, Section S1.2 for the complete training-only threshold matrix.

Volume retrieval compares the global encoder representation with four lesion-aware alternatives. These are all-disease pooling over unique lesion-intersecting patches, lesion-balanced pooling over normalized lesion embeddings, finding-balanced pooling that first aggregates lesions linked to one finding, and fusion of the global and lesion-balanced embeddings. All representations are L2-normalized before cosine retrieval. CTs without an eligible mask-derived embedding are excluded from the corresponding index. See Additional file 1, Section S1.2 for the complete pooling equations and normalization contract.

4.6 Query Policies

Three query-region policies evaluate sensitivity to progressively coarser localization:

1. **Exact lesion mask:** an oracle upper-bound query using the ground-truth segmentation mask, which tests embedding capacity under ideal localization.
2. **Bounding-box mask:** the tight three-dimensional bounding box around the lesion mask, used as a controlled approximation of region marking without contour tracing.
3. **Loose box:** a deterministically expanded three-dimensional box that includes additional surrounding tissue and stress-tests coarser localization.

Let $b_{\min}$ and $b_{\max}$ denote the inclusive [D, H, W] bounds of the lesion's bounding box. Its axiswise size is $s = b_{\max} - b_{\min} + 1$. The loose-box policy adds a margin $m = \lceil 0.25s \rceil$ to both sides of each axis. The resulting bounds are clipped to the processed CT, mapped to the 8 x 8 x 8 patch grid, and pooled with unit weight over every enclosed patch. The 0.05 occupancy threshold determines query eligibility but does not filter patches inside either box.

No center jitter or other random perturbation is applied; query geometry is deterministic and identical across encoders. Exact masks define the best-case benchmark. Bounding-box and loose-box queries measure sensitivity to progressively less precise query regions. These controlled policies do not reproduce unaided user localization. See Additional file 1, Section S1.2 for version identifiers and complete coordinate-mapping details.

The query/reference asymmetry is intentional. Query regions may be coarse, but every indexed reference embedding remains derived from an exact lesion mask. This design tests whether a coarse query region can retrieve precisely indexed lesions of the same category; mask reconstruction is outside its scope. Figure 3 applies the three policies to one worked example and shows how pooled extent and patch weighting change between them.

Fig. 3 Controlled query regions in decreasing localization precision. (a) Exact masks use occupancy-weighted patches meeting the 0.05 threshold. (b) Bounding boxes approximate region marking without contour tracing. (c) Loose boxes add $m = \lceil 0.25s \rceil$ voxels per side. Both box policies pool enclosed patches with unit weight. The schematic is two-dimensional, but all operations are three-dimensional. Query policies affect only the query; indexed references always use exact lesion masks

4.7 Retrieval Task, Eligibility, and Scoring

The primary retrieval unit is one lesion entity rather than one CT examination. This distinction is necessary because a CT can contain multiple findings, and retrieving a relevant examination does not mean the intended lesion was retrieved.

For lesion-oriented retrieval, query and reference embeddings are computed as weighted means over patch tokens. The pooled support and weights depend on the query policy. Let M_q denote the query region projected onto the query patch grid, and let $w_q(x)$ denote its occupancy in patch x. Let $M_{r,k}$ denote the exact mask for reference lesion k in volume r. Its occupancy in patch y is $w_{r,k}(y)$. Both embeddings take the common form

$$\rho_q = \mathrm{norm}\left(\frac{\sum_{x\in\Omega_q} a_q(x) F_q^p(x)}{\sum_{x\in\Omega_q} a_q(x) + \epsilon}\right), \qquad \rho_{r,k} = \mathrm{norm}\left(\frac{\sum_{y\in\Omega_{r,k}} a_{r,k}(y) F_r^p(y)}{\sum_{y\in\Omega_{r,k}} a_{r,k}(y) + \epsilon}\right),$$

where x and y index three-dimensional patch-grid locations. Reference support and weighting remain fixed across all query policies: $\Omega_{r,k} = \{y : w_{r,k}(y) \geq \tau\}$, $a_{r,k}(y) = w_{r,k}(y)$, and $\tau = 0.05$. Query support and weighting depend on the policy in Section 4.6:

- Exact-mask queries use $\Omega_q = \{x : w_q(x) \geq \tau\}$ and $a_q(x) = w_q(x)$.
- Bounding-box and loose-box queries use $\Omega_q = B_q$ and $a_q(x) = 1$, so every enclosed patch contributes equally and the threshold governs eligibility rather than pooled support.

This formulation places coarse query regions and precisely segmented reference lesions in the same feature space. The box policies differ from the exact mask in both pooled extent and weighting.

A query is eligible when preprocessing succeeds, a label is available, its mask is nonzero, at least one patch meets the occupancy threshold, and at least one valid reference positive remains after exclusions. The common set contains 407 eligible queries and 14,359 reference lesions. Section 3.1 and Additional file 1, Section S1.2 reconcile these counts with the 429 source entities.

The reference pool excludes the query sample. The 50 validation query CTs contain 49 patients. The 2,992 processed train CTs contain 2,717 patients, none of whom appears in the query set. Ranking remains lesion-level, so

distinct lesions from one reference CT can appear separately. Parent-volume grouping is used only for presentation. Patient-grouped development folds are specified in Section 5.1.

Each query is compared with the reference lesion bank using cosine similarity between normalized embeddings. The primary score uses neither category labels nor finding text. Mask descriptors are evaluated only in the descriptor-control and reranking experiments in Section 4.9.

Relevance follows the hierarchy in Section 3.3. Primary-category positives share the first category-code digit, whereas exact-subcategory positives share the complete code. Both are reported, but exact-subcategory agreement is emphasized because the primary-category task is coarse and class-imbalanced.

As a non-image chance control, the analytic chance expectation of Recall@k is computed for uniform random sampling without replacement. For query q with P_q positive lesions among $N = 14{,}359$ references, the expectation is

$$R_q^{\text{random}}@k = 1 - \frac{\binom{N-P_q}{k}}{\binom{N}{k}},$$

and is averaged across the 407 queries. Expected primary-category Recall@1/5/10/20 is 0.646/0.943/0.985/0.999. The corresponding exact-subcategory values are 0.154/0.531/0.729/0.859. High random recall at larger k reflects the prevalence of category positives, not image similarity. This expectation uses the observed prevalence but does not use labels to order candidates. A label-informed ranker is not included because it would require the ground-truth query category as an input.

4.8 Volume and Lesion-Slice Retrieval

Volume retrieval evaluates case-level similarity using the representations defined in Section 4.5. It is an easier task than lesion retrieval because each query can have many positive reference CTs.

Lesion-slice retrieval supports localized browsing and review-panel construction. The method creates one candidate for each lesion-bearing axial patch-grid plane. Each candidate is pooled only from tokens intersecting that lesion. For lesion k in reference volume r on plane z, with projected slice mask $M_{r,k,z}$, the embedding is

406 $$\ell_{r,k,z}^{\text{mean}} = \text{norm}\left(\frac{\sum_{h,w} M_{r,k,z}(h,w) F_r^p[z,h,w]}{\sum_{h,w} M_{r,k,z}(h,w) + \epsilon}\right).$$

407 Exact lesion embeddings serve as queries against 45,402 such reference candidates. This auxiliary experiment
408 measures category-consistent slice selection for localized display; spatial registration is outside its scope.

409 **4.9 Baselines and Auxiliary Analyses**

410 Auxiliary analyses test whether retrieval can be explained by lesion geometry, describe high-scoring failures, and
411 assess sensitivity to documented reference-side pretraining exposure.

412 The annotation-informed descriptor control ranks lesions by equivalent diameter, positive patch count, and diameter-
413 tercile agreement without image appearance or semantic labels. The geometry reranker begins with the top 20 cosine
414 candidates for each query. It combines normalized cosine and descriptor scores, reorders those candidates, and
415 evaluates the resulting top 10. It cannot recover a lesion that was absent from the original candidate set. See
416 Additional file 1, Section S1.2 for standardization, weights, cut points, missing-feature handling, and scoring
417 equations.

418 The separate annotation-negative bank contains 398 normalized depth-plane embeddings selected near to and far
419 from annotated lesions across the 50 validation CTs. For each query, its highest score is compared with the top-
420 ranked lesion score and a 0.05 margin. “Negative” denotes non-overlap with available annotations, not clinically
421 confirmed normal tissue. Lesion hard negatives are analyzed separately as high-ranked reference lesions with
422 mismatched evaluation labels. See Additional file 1, Section S1.2 for the complete selection protocols.

423 The reference-partition sensitivity analysis tests whether retrieval performance varies with documented pretraining
424 exposure of the reference volumes. FlexiCT reports 47,149 CT-RATE volumes in its pretraining corpus [10]. This
425 count matches the 47,149 reconstructed volumes in the documented CT-RATE training split [6]. The FlexiCT
426 exposure table also identifies the CT-RATE validation split as held out. On this basis, reference lesions from
427 `train_*` scans are classified as exposed and those from `valid_*` scans as unexposed. Exposure is documented
428 at the partition level; no released per-volume FlexiCT pretraining manifest was audited.

The unexposed bank contains 2,174 eligible lesions from 413 CT-RATE validation CTs. The exposed bank contains 12,185 eligible lesions from 2,565 CT-RATE training CTs. The primary design matches lesion count and subcategory frequency. This equalizes chance recall but represents more distinct cases than the unexposed bank. The whole-CT sensitivity design better preserves case structure but only approximately matches lesion count and label composition. Neither design matches both properties. Both designs are repeated over 500 draws, and each reported point estimate is averaged across those draws.

The primary 95% confidence interval is a two-level percentile bootstrap targeting the draw-averaged effect. Each replicate independently resamples, with replacement, the 50 query CTs and the 500 matched-pool draws. It then averages the resampled pools. The 95% single-pool sensitivity interval instead uses one matched pool per replicate. It describes the variability expected when constructing one new comparison bank rather than uncertainty in the draw-averaged effect. Both intervals condition on the complete eligible unexposed bank as fixed. See Additional file 1, Section S2.10 for complete estimand definitions and implementation details.

Differences are reported as exposed-matched minus unexposed. A positive difference would indicate better performance with exposed references, but it would not by itself demonstrate instance-level memorization. CT-RATE partition membership may also capture broader differences in representation quality or data distribution. This analysis addresses the reference side only. Because every public-validation query volume belongs to the CT-RATE training partition, query-side exposure cannot be assessed in this set.

Use of generative AI. OpenAI was used as an AI-assisted tool for implementation support, code review, and language refinement. The author executed the analyses, inspected the outputs, verified the code, numerical results, citations, and final text, made all scientific decisions, and accepts full responsibility for the work.

5. Experimental Design and Statistical Analysis

5.1 Five-Fold Training-Set Representation Benchmark

The five-fold exact-mask training benchmark tests whether FlexiCT-3D embeddings support lesion-category retrieval within the training split. FlexiCT-3D-VLM serves as an encoder ablation rather than a co-primary model. In each fold, held-out patients serve as queries and the remaining patients as references. Patients are greedily

assigned to balance primary-category representation, lesion count, and CT count, using seed 17. All scans and lesions from each patient remain together, and both encoders use the same assignments. This is a development and robustness benchmark; public validation remains the primary retrieval-held-out evaluation.

Each query fold contains 2,871-2,872 lesions from 540-542 patients and 592-605 CTs. Every one of the 14,359 lesions serves once as a query, and every query has at least one reference positive under both relevance rules. Across folds, primary-category positive counts per query have means of 7,226.0-7,452.8 and medians of 8,666-8,811. Exact-subcategory positive counts have means of 1,676.7-1,752.2 and medians of 1,879-1,941. Additional file 1, Section S2.1 reports fold-level coverage and positive availability. Rare-subcategory results should be interpreted in light of their label-specific support.

Results are summarized as unweighted means and sample standard deviations across the five patient-grouped folds. Primary-category metrics are secondary because they reflect only two broad, imbalanced classes.

5.2 Public Validation Lesion Retrieval

The primary public-validation experiment compares exact-mask, bounding-box, and loose-box queries using 407 validation lesions and 14,359 training reference lesions. Validation lesions serve as retrieval-held-out queries, while training lesions form the searchable reference bank. This arrangement reflects the intended workflow in which a new query CT is compared with a prebuilt reference bank.

Metrics are computed separately for primary-category and exact-subcategory relevance. The primary endpoints are Recall@1, Recall@10, and MRR@10. Additional file 1 reports Recall@5 and Recall@20 in the complete endpoint grids and Precision@10 for selected auxiliary comparisons. The analytic chance expectation is defined in Section 4.7. CT-grouped percentile 95% bootstrap intervals use 10,000 resamples of the 50 query CTs with seed 17; all lesions from a sampled CT remain together. Additional file 1, Section S2.3 reports intervals for individual FlexiCT-3D query policies and their paired differences. These descriptive intervals are not adjusted for multiple endpoints, policies, or strata. An isolated interval excluding zero is not treated as confirmatory evidence.

Post hoc exploratory analyses stratify the common 407-query set by lesion-size tertile, positive patch count (1-4, 5-16, 17-64, or at least 65 patches), and native axial-slice support. A single-slice lesion contains positive voxels on exactly one native axial slice, as determined directly from the source segmentation rather than bounding-box depth.

The same CT-grouped bootstrap is applied within each stratum. Because lesion size, patch count, slice support, and category composition are related, these analyses characterize failure patterns rather than independent causal effects.

5.3 Encoder Ablation: FlexiCT-3D Versus FlexiCT-3D-VLM

FlexiCT-3D-VLM is evaluated as an encoder ablation in both the five-fold training benchmark and the public-validation experiment. The comparison tests whether report-aligned pretraining changes lesion retrieval relative to the primary image-only FlexiCT-3D encoder. Its scope is limited to two checkpoints from one model family, not a survey of encoder architectures.

The same 10,000-resample CT-grouped percentile bootstrap compares the encoders on identical resampled validation cohorts, using seed 17. Additional file 1, Section S2.2 reports public-validation primary-category results and paired early-rank intervals.

5.4 Descriptor Baseline and Reranking

Descriptor-only retrieval is an annotation-informed geometry control that tests whether simple geometric and metadata descriptors can explain retrieval performance. It uses annotation-derived lesion geometry but no image-appearance information or semantic labels. Geometry reranking tests whether the same descriptors improve embedding-based rankings. Both analyses use bounding-box and loose-box queries, with paired CT-grouped bootstrap intervals constructed as in Section 5.2. Additional file 1, Section S2.7 reports uncertainty and category-wise differences. Section S1.2 records the descriptor and negative-bank protocols.

5.5 Volume-Level Retrieval

Volume-level baselines test whether global or lesion-aware embeddings retrieve CT volumes containing findings from the query subcategory. This task is easier than lesion-level retrieval because each query has many positive reference CTs: a mean of 1,430.0 and a median of 1,360.5 among the 2,978 mask-derived candidates.

The analytic chance expectation applies the hypergeometric formula in Section 4.7 to the volume candidate pools and averages it across the 50 query CTs. It is not an executed ranker, and random MRR@10 was not computed. The CLS pool contains all 2,992 processed training CTs, whereas the mask-aware pools contain the 2,978 CTs with at

least one eligible lesion embedding at the 0.05 threshold. This 14-CT difference is intrinsic to mask-aware indexing, so the CLS contrasts are not perfectly inventory-matched. Each contrast reflects both aggregation policy and the 0.5% inventory difference; the two contributions were not separated. Rounded chance expectations are still identical for both pools: 0.480, 0.916, and 0.971 at ranks 1, 5, and 10. Recall@10 is saturated because so many candidate CTs are positives, so Recall@1 and MRR@10 are the primary interpretive endpoints and are presented first. Paired intervals in Additional file 1, Section S2.9 compare each mask-aware method only with the CLS baseline, not with the other mask-aware methods.

5.6 Lesion-Slice Retrieval

Lesion-slice retrieval evaluates exact-mask lesion queries against 45,402 reference lesion-slice candidates, each representing one 8-voxel patch-depth slab of a reference lesion (Section 4.8). This experiment evaluates localized slice selection for review panels rather than lesion-entity ranking.

5.7 Failure Analysis

Lesion-index hard-negative analysis identifies high-scoring retrieved lesions whose primary-category or subcategory labels differ from the query. A separate negative-bank analysis tests whether annotation-negative patch-grid-plane candidates score above or near lesion matches. Its 398 near- and far-from-lesion candidates were stored while the complete validation patch grids were available. This bank is an evaluation artifact and is not part of the production lesion index.

6. Results

6.1 Coarser Query Regions Reduce Early-Rank Performance, while Recall@10 Differences Remain Uncertain

Table 2 reports the primary FlexiCT-3D results and paired query-policy differences. Exact-mask queries achieve exact-subcategory Recall@1 of 0.484 (CT-grouped 95% CI, 0.408-0.569) and Recall@10 of 0.848 (0.783-0.914), compared with analytic chance expectations of 0.154 and 0.729. MRR@10 is 0.614 (0.538-0.695). Bounding-box

and loose-box queries retain Recall@10 point estimates of 0.840 and 0.821 but show progressively lower Recall@1 and MRR@10.

Table 2. Primary public-validation query-policy results with FlexiCT-3D and subcategory relevance. Policy rows contain point estimates with CT-grouped percentile 95% confidence intervals. Contrast rows subtract the second named policy from the first on the identical 407-query set.

Policy or paired contrast	R@1 (95% CI)	R@10 (95% CI)	MRR@10 (95% CI)
Exact mask	0.484 (0.408-0.569)	0.848 (0.783-0.914)	0.614 (0.538-0.695)
Bounding box	0.430 (0.355-0.517)	0.840 (0.787-0.896)	0.568 (0.498-0.646)
Loose box	0.356 (0.286-0.436)	0.821 (0.776-0.868)	0.500 (0.437-0.573)
Bounding box - exact mask	-0.054 (-0.098 to -0.007)	-0.007 (-0.049 to +0.040)	-0.046 (-0.079 to -0.014)
Loose box - exact mask	-0.128 (-0.171 to -0.086)	-0.027 (-0.077 to +0.033)	-0.114 (-0.155 to -0.074)
Loose box - bounding box	-0.074 (-0.113 to -0.038)	-0.020 (-0.051 to +0.011)	-0.068 (-0.094 to -0.043)

All paired Recall@1 and MRR@10 intervals exclude zero, whereas all paired Recall@10 intervals include zero. Exact masks provide the annotation-derived upper-bound benchmark. Among the simulated coarse-query conditions, bounding boxes retain stronger early ordering than loose boxes, which serve as the localization stress test. See Additional file 1, Section S2.3 for full Recall@5 and Recall@20 results.

The exploratory strata further characterize this aggregate result. Under exact-mask queries, large lesions achieve Recall@10 of 0.919, compared with 0.801 for small lesions. Lesions occupying at least 65 positive patches achieve 0.937, compared with 0.783 for lesions occupying 1-4 patches. These gradients are less consistent under bounding-box and loose-box queries, indicating that patch support alone does not explain coarse-query behavior.

Only 14 queries are single-slice. Their exact-mask Recall@10 is 0.714 (95% CI, 0.462-0.929), compared with 0.852 (0.787-0.920) for the 393 multi-slice queries. The wide single-slice interval and the overlap in lesion characteristics prevent a firm conclusion about slice support. Additional file 1, Section S2.4 reports all three query policies.

6.2 Five-Fold Results Are Stable, with No Clear Encoder Difference

Table 3, panel A, summarizes the five-fold training benchmark and public-validation encoder ablation. Panel B is discussed in Section 6.3. Five-fold exact-mask performance varies little across the patient-grouped folds. On public validation, FlexiCT-3D-VLM has higher exact-subcategory Recall@1 and MRR@10 point estimates under all three query policies, but every paired exact-subcategory interval for the encoder differences includes zero. These results do not establish an advantage for either checkpoint and do not support conclusions about unrelated encoder families.

Table 3. Encoder and descriptor controls for the three primary endpoints. Panel A reports FlexiCT-3D / FlexiCT-3D-VLM. Training values are unweighted fold means ± sample standard deviations; public-validation values are point estimates. In panel B, rows represent query policies, and each cell lists descriptor-only / embedding-only / geometry-reranked retrieval in that order.

Panel A. Encoder results.

Evaluation	Query policy	R@1	R@10	MRR@10
Train five-fold, subcategory	Exact mask	0.495 ± 0.011 / 0.499 ± 0.018	0.880 ± 0.014 / 0.879 ± 0.013	0.625 ± 0.014 / 0.629 ± 0.016
Public validation, subcategory	Exact mask	0.484 / 0.506	0.848 / 0.857	0.614 / 0.625
Public validation, subcategory	Bounding box	0.430 / 0.445	0.840 / 0.833	0.568 / 0.580
Public validation, subcategory	Loose box	0.356 / 0.398	0.821 / 0.813	0.500 / 0.526

Panel B. Descriptor controls.

Query policy	R@1	R@10	MRR@10
Bounding box	0.253 / 0.430 / 0.452	0.730 / 0.840 / 0.848	0.401 / 0.568 / 0.584
Loose box	0.253 / 0.356 / 0.405	0.730 / 0.821 / 0.818	0.401 / 0.500 / 0.533

See Additional file 1, Sections S2.2, S2.5, and S2.11 for full endpoint grids, primary-category results, fold rows, and paired encoder intervals.

6.3 Embedding Retrieval Outperforms Descriptor-Only Retrieval; Geometry Reranking Shows No Reliable Improvement

Under bounding-box queries, descriptor-only Recall@10 is 0.730, compared with 0.840 for embedding retrieval; under loose-box queries, the corresponding values are 0.730 and 0.821 (Table 3, panel B). Paired deficits in Recall@1, Recall@10, and MRR@10 exclude zero under both policies. Geometry reranking produces several favorable point estimates, but every paired interval includes zero. The analysis does not establish an improvement over embedding-only retrieval. See Additional file 1, Sections S2.7 and S2.11 for the full endpoint grid, paired intervals, and category-wise differences.

6.4 Lesion-Aware Pooling Improves Early Volume Ranking, and Lesion-Slice Retrieval Supports Localized Selection

Table 4, panel A, summarizes volume and lesion-slice retrieval. Panel B reports the failure analysis discussed in Section 6.5.

Table 4. Volume and lesion-slice retrieval with overall failure rates for FlexiCT-3D.

Panel A. Volume and lesion-slice retrieval. Volume rows use subcategory relevance and candidate pools of 2,992 CTs for CLS or 2,978 CTs for mask-derived methods; the analytic chance expectations round identically for both pools. Lesion-slice rows use the stated relevance level and a 45,402-slab candidate pool.

Retrieval unit	Method or level	R@1	R@10	MRR@10
Volume	Analytic chance expectation	0.480	0.971	—
Volume	Global CLS	0.620	0.980	0.763
Volume	All-disease pooling	0.820	0.980	0.883
Volume	Lesion-balanced pooling	0.860	0.980	0.917
Volume	Finding-balanced pooling	0.860	0.980	0.907

Retrieval unit	Method or level	R@1	R@10	MRR@10
Volume	CLS/lesion-balanced fusion	0.840	0.980	0.896
Lesion slice	Primary category	0.806	0.980	0.864
Lesion slice	Subcategory	0.459	0.806	0.577

Panel B. Overall failure rates.

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

Lesion- and finding-balanced volume pooling improve early ordering relative to global CLS. Both have paired Recall@1 differences of +0.240 (95% CI, +0.080 to +0.400); their MRR@10 differences are +0.154 (+0.057 to +0.251) and +0.144 (+0.043 to +0.244), respectively. The comparison is not perfectly inventory-matched because CLS searches 2,992 CTs and the mask-aware methods search 2,978. Each contrast reflects both aggregation policy and a 0.5% inventory difference; the two contributions were not separated. Recall@10 is prevalence-saturated (analytic chance expectation, 0.971), so interpretation relies on early ordering rather than top-k coverage. See Additional file 1, Sections S2.9 and S2.11 for full volume point estimates and paired intervals.

Against 45,402 lesion-slice candidates, exact-mask queries achieve exact-subcategory Recall@1 of 0.459, Recall@10 of 0.806, and MRR@10 of 0.577. These results concern category-consistent localized slice selection; spatial registration was not evaluated. See Additional file 1, Section S2.11 for the complete lesion-slice results.

6.5 Failure Analysis Is Dominated by Fine-Grained Category Confusion

Fine-grained confusion exceeds broad-category confusion under both box policies (Table 4, panel B). For bounding-box queries, the top-ranked result mismatches the exact subcategory for 57.0% of queries but the primary category for 21.9%. Under loose-box queries, the corresponding rates are 64.4% and 30.7%. No annotation-negative candidate exceeds the top lesion match for bounding-box queries, compared with 7.4% of loose-box queries.

Because the bank is annotation-negative rather than clinically adjudicated normal tissue, these values describe ranking pressure rather than clinical false positives. Category-wise rates in Additional file 1, Section S2.8 show greater loss of category-specific ordering under loose boxes, consistent with their role as a localization stress test.

6.6 Documented Reference-Side Pretraining Exposure Shows No Consistent Advantage

Neither matching mode shows a consistent exposed-reference advantage. Under the primary equal-lesion-count, subcategory-frequency-matched design, matching is exact, with no subcategory shortfall and no unclassified reference lesion. Under exact-mask queries, exact-subcategory Recall@10 is 0.835 for the unexposed pool and 0.866 for the draw-averaged exposed pool, a difference of +3.1 percentage points (95% CI, -1.2 to +7.7). The corresponding MRR@10 difference is +2.5 points (-0.6 to +5.8). Across all three query policies and all five metrics, all 15 confidence intervals include zero. Point estimates also change sign between policies rather than consistently favoring the exposed pool.

Constructing the exposed comparison bank is itself a source of variability. Across the 500 matched draws, exact-mask exact-subcategory Recall@10 for the exposed arm has a standard deviation of 1.8 percentage points and a 9.8-point range. The single-pool sensitivity interval defined in Section 4.9 is typically about 1.8 times wider than the corresponding confidence interval. It describes sensitivity to constructing one comparison bank rather than uncertainty in the draw-averaged effect.

The subcategory-matched exposed sample represents more distinct cases than the unexposed pool. Each draw contains 2,174 lesions from a mean of 1,364 CTs and 1,257 patients, compared with 413 CTs and 383 patients in the unexposed pool. Whole-CT sampling reduces most of this asymmetry, drawing a mean of 2,177 lesions from 458 CTs and 448 patients. It produces the same overall pattern: the exact-mask Recall@10 difference is +1.8 points (95% CI, -2.1 to +5.7), and the loose-box MRR@10 difference is -3.6 points (-7.3 to +0.4).

One of the 30 unadjusted confidence intervals narrowly excludes zero: bounding-box Recall@5 under whole-CT sampling is 3.3 percentage points lower for exposed references (95% CI, -6.4 to -0.01). This direction is opposite to an exposure advantage. Because it is an isolated marginal result among 30 correlated endpoints and its single-pool sensitivity interval crosses zero (-9.8 to +3.4), it is treated as exploratory. Additional file 1, Section S2.10 reports both matching modes in full.

As a descriptive comparison only, the unexposed arm retrieved a same-subcategory lesion within the top 10 for 83.5% of exact-mask queries while searching 2,174 references. The complete 14,359-lesion bank yielded 84.8% in Section 6.1. Because the banks differ in size and composition, this is not a formal check of either result.

The matched-pool analyses detect no consistent reference-side advantage associated with documented pretraining exposure. This is a bounded result, not evidence of no effect: across endpoints, confidence-interval bounds extend to approximately nine percentage points from zero, and single-pool sensitivity bounds extend to approximately twelve percentage points. Both matching modes leave one design dimension unmatched, exposure is assigned at the partition level rather than verified per volume, and query-side exposure cannot be assessed in this set. These constraints are restated in Section 8.5.

7. Qualitative Retrieval Examples

Figures 4 and 5 illustrate two behaviors quantified in Section 6: successful tight-box retrieval and fine-grained confusion under a loose box. Exact-mask retrieval and lesion-slice retrieval are illustrated in Additional file 1, Figures S1 and S2.

All four queries were selected from recorded results before image inspection. They span four subcategories and two lesion-size strata. Each panel shows the query and five retrieved candidates with lesion-visible axial slices, masks, ranks, cosine scores, category labels, and subcategory-agreement indicators. The query panel marks the query region, and reference panels mark the indexed lesions. Frames and text labels indicate subcategory agreement or mismatch.

Fig. 4 Tight-box retrieval success. A medium-size `1b` bronchiectasis query (lesion 4, `train_13252_b_2`) retrieves five same-subcategory lesions from five distinct CTs. The query box uses unit-weighted pooling, whereas the reference lesions use exact masks. Because the query box is derived from the annotation, this panel illustrates controlled box-query behavior rather than unaided localization. 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. 5 Within-parent ontology confusion under a loose box. For a large `2b` atelectasis/consolidation query (lesion 0, `train_13177_a_2`), expanding the query box by 25% per axis replaces an exact-mask rank-1 success with five top-ranked subcategory mismatches from four CTs. All five candidates remain within the same primary category; four are `2a` linear opacity, whose definition also includes subsegmental atelectasis (Section 3.3). The panel therefore illustrates the ontology-boundary confusion quantified in Section 6.5, not reader-assessed irrelevance. 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]

The four panels illustrate behaviors quantified in Section 6 but provide no additional statistical evidence. They do not establish radiologist-assessed similarity or educational benefit.

8. Discussion

8.1 Principal Findings

Spatial patch tokens from a frozen volumetric CT foundation encoder were sufficient to build a reusable lesion-retrieval layer. Each indexed lesion retains links to its parent examination, annotation, localized slices, and source finding text. The same layer handles lesion-, lesion-slice-, and volume-level retrieval without task-specific retrieval training. The main contribution is a practical method for converting an existing volumetric representation and an annotated medical-image database into an inspectable resource for lesion-centered case comparison.

The primary evaluation shows that these frozen representations preserve fine-grained localized information. Exact-mask Recall@1 is 0.484, compared with an analytic chance expectation of 0.154, and Recall@10 is 0.848, compared with 0.729. Embedding retrieval also outperforms the annotation-informed geometry control under both box policies. Coarser query regions weaken early ordering, while their Recall@10 differences remain uncertain. FlexiCT-3D-VLM shows similar fine-grained behavior, so the results do not establish that report-aligned pretraining is necessary or superior.

Earlier studies established engineered, task-specific, and multimodal lesion-retrieval methods [15-17]. This study adds a lesion-indexed image-to-image retrieval framework derived directly from frozen CT-native patch tokens, with coarse-query/precise-reference asymmetry and reusable multi-granularity artifacts.

8.2 Implications of Query-Region Precision

Query-region precision affects early ordering more strongly than top-10 coverage. Exact masks provide the strongest retrieval. Among the controlled box conditions, tight boxes retain stronger Recall@1 and MRR@10 than loose boxes, while the paired Recall@10 differences remain uncertain. A coarse region may therefore still populate a review set with category-consistent cases even when it ranks them less sharply. This interpretation is limited to the simulated conditions: both box policies were generated from exact annotations and do not establish performance with automated localization or manual user input.

8.3 Why Top-k Retrieval Matters for Case Review

The framework returns an inspectable set of related cases, not a single diagnostic prediction. Although top-1 exact-subcategory retrieval remains moderate, especially for box queries, top-10 retrieval identifies a category-consistent lesion for more than 80% of queries under all three policies. Because each retrieved lesion remains linked to its parent examination, localized image context, and source finding text, the ranked set can be used for comparative review as well as label matching. Prior CT reader studies have likewise presented small sets of retrieved cases for inspection [19,20]. Those studies motivate the workflow, but the automatic evaluation reported here does not establish educational benefit or radiologist-assessed relevance.

8.4 Value of Multi-Granularity Retrieval

The reusable patch-token artifacts serve three complementary retrieval units. Lesion-level retrieval provides the strictest test of localized finding similarity. Lesion-aware volume representations organize parent CT cases and improve early ordering relative to global CLS, although volume retrieval is an easier, prevalence-rich task and the candidate inventories differ by 0.5%. Lesion-slice retrieval selects localized reference slabs for review-panel construction. Together, these experiments show how one indexed representation enables lesion search, case grouping, and localized display without repeated encoder inference.

8.5 Limitations

Public validation comprises 50 CT examinations and 407 evaluable lesion queries, with no external cohort. ReXGroundingCT contains lung and pleural findings, so the results do not establish performance for other thoracic or abdominal abnormalities [1].

FlexiCT was pretrained on public CT corpora that include CT-RATE [10], from which ReXGroundingCT is derived [1,6]. Because all public-validation query volumes belong to the CT-RATE train partition, they cannot provide external validation. The reference-side sensitivity analysis found no consistent advantage from documented exposure. It is not an equivalence test, however: the intervals remain compatible with differences of about nine percentage points in at least one direction. Additional file 1, Section S2.10 reports the matching designs, the two interval constructions, and the single isolated marginal endpoint. Exposure was assigned at the partition level rather than verified from a per-volume pretraining manifest, and query-side exposure cannot be tested without an unexposed query set. Only evaluation on a lesion dataset outside the FlexiCT pretraining corpus can resolve this limitation.

The reference annotations and relevance labels impose further constraints. Training annotations include at most three representative instances per finding and may omit visible reference lesions [1]. The filtered benchmark has only two imbalanced primary categories, while subcategory labels can be noisy, visually ambiguous, and incomplete proxies for perceived similarity [18]. Query geometry, reference labels, and category-based evaluation also originate from the same annotation system. Same-subcategory retrieval measures consistency with the ReXGroundingCT ontology, not independently adjudicated visual or clinical similarity. No reader study was conducted, and the automatic metrics do not establish case-review benefit or diagnostic utility.

The query-policy experiments are controlled simulations. Because exact annotations generate both box conditions and the geometry descriptors, the results measure sensitivity to region precision rather than performance of an automated localization system or manual query interface. The encoder's 8 x 8 x 8-voxel patch resolution may dilute very small lesions and destabilize their lesion-slice localization. The post hoc size, patch-support, and slice-support strata are confounded by category and morphology; only 14 queries are single-slice, and coarse-query patterns are not uniformly monotonic.

Annotation-negative planes are not clinically adjudicated normal tissue and cannot define clinical false positives. Registration was outside the retrieval evaluation and should be treated only as a possible visualization aid.

Recorded checkpoint hashes, preprocessing settings, seeds, and core command histories support procedural reproducibility. Exact runtime, peak memory, a fully pinned software environment, and complete hardware provenance were not preserved. The study makes no claim of computational efficiency or bitwise reproduction.

8.6 Future Work

Future work should first evaluate the framework on a CT lesion set outside the FlexiCT pretraining corpus. This would test external generalization and resolve the query-side exposure limitation. A complementary encoder study could repeat the benchmark with a volumetric model whose pretraining corpus excludes CT-RATE. Query-region experiments should then use regions produced by an automated localization method or supplied through a real user interface rather than annotation-derived simulations. If resources permit, a reader study could assess radiologist relevance rankings, educational value, and comparison with ordinary case search using ranking-sensitive measures such as nDCG [18,22]. Methodological extensions include improved handling of fine-grained ontology confusion, learned integration of geometry signals, and richer lesion-centered review panels.

9. Conclusions

This study demonstrates a lesion-grounded retrieval framework that converts spatial patch tokens from a frozen volumetric CT foundation encoder into reusable lesion representations without task-specific retrieval training. On ReXGroundingCT public validation, exact-subcategory Recall@1 was 0.484 compared with an analytic chance expectation of 0.154, while Recall@10 was 0.848 compared with 0.729. Bounding-box and loose-box queries retained Recall@10 above 0.82 but produced weaker early ordering, and embedding retrieval outperformed the annotation-informed geometry control under both box policies. The same indexed artifacts supported lesion retrieval, case-level volume retrieval, and localized lesion-slice selection. By retaining links from retrieved lesions to their parent CTs, localized images, annotations, and source finding text, the framework can organize compatible medical-image databases into inspectable ranked case sets for educational comparison. These results establish the feasibility of category-consistent retrieval, but not reader-assessed relevance, educational benefit, diagnostic utility, or external generalization.

List of abbreviations

CBIR: content-based image retrieval; CI: confidence interval; CLS: classification token; CT: computed tomography; HDF5: Hierarchical Data Format version 5; MRR@10: mean reciprocal rank truncated at 10; ROI: region of interest; R@k: Recall at rank *k*; VLM: vision-language model.

Declarations

Ethics approval and consent to participate

This study was a secondary computational analysis of previously collected and anonymized human imaging data and involved no new participant contact or data collection. The present secondary analysis was conducted in accordance with the ethical principles of the World Medical Association Declaration of Helsinki. The original CT-RATE study reports approval by the Clinical Research Ethics Committee at Istanbul Medipol University for the study and open release of CT-RATE (E-10840098-772.02-6841; October 27, 2023) [6]. ReXGroundingCT provides derived annotations from CT-RATE [1]. The author had no access to direct participant identifiers. Consent to participate in the present secondary analysis was therefore not applicable.

Consent for publication

Not applicable. The manuscript contains no identifying participant information. The displayed CT images are de-identified examples from CT-RATE and are reproduced in accordance with the dataset license and written clarification received from a dataset author.

Availability of data and materials

The datasets analyzed during this study are available through the gated ReXGroundingCT repository (https://huggingface.co/datasets/rajpurkarlab/ReXGroundingCT) and CT-RATE repository (https://huggingface.co/datasets/ibrahimhamamci/CT-RATE), subject to their respective access conditions [1,6]. The aggregate results needed to interpret the study are included in this article and Additional file 1.

Competing interests

The author declares that he has no competing interests.

Funding

The author declares that no funds, grants, or other support were received for this research or the preparation of this manuscript.

Authors' contributions

ZEY conceived and designed the research project, carried out the study, implemented the core retrieval and analysis software, prepared and analyzed the data, interpreted the results, prepared the figures, wrote and revised the manuscript, and approved the submitted version.

Acknowledgements

The author acknowledges the CT-RATE and ReXGroundingCT teams for making the imaging data and annotations available for research, and the FlexiCT team for releasing the model code and pretrained weights.

Additional file

Additional file 1 (.pdf): Supplementary Dataset Details, Methods, Extended Results, and Qualitative Panels. Evaluated ontology, implementation information, complete supporting analyses, and the exact-mask and lesion-slice retrieval panels.

References

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

2. Dubey SR: A decade survey of content based image retrieval using deep learning. IEEE Trans Circuits Syst Video Technol 32:2687–2704, 2022. https://doi.org/10.1109/TCSVT.2021.3080920

3. Akgül CB, Rubin DL, Napel S, Beaulieu CF, Greenspan H, Acar B: Content-based image retrieval in radiology: current status and future directions. J Digit Imaging 24:208–222, 2011. https://doi.org/10.1007/s10278-010-9290-9

4. Kumar A, Kim J, Cai W, Fulham M, Feng D: Content-based medical image retrieval: a survey of applications to multidimensional and multimodality data. J Digit Imaging 26:1025–1039, 2013. https://doi.org/10.1007/s10278-013-9619-2

5. Ben Abacha A, Santamaria-Pang A, Lee HH, Merkow J, Cai Q, Devarakonda ST, Islam A, Gong J, Lungren MP, Lin T, Codella NC, Tarapov I: 3D-MIR: a benchmark and empirical study on 3D medical image retrieval in radiology. arXiv preprint arXiv:2311.13752, 2023. https://doi.org/10.48550/arXiv.2311.13752

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

7. Zhang T, Zhao Z, Wu C, Zhou X, Zhang Y, Wang Y, Xie W: RadIR: a scalable framework for multi-grained medical image retrieval via radiology report mining. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2025. Lect Notes Comput Sci 15964:508–518, 2025. https://doi.org/10.1007/978-3-032-04971-1_48

8. Denner S, Zimmerer D, Bounias D, Bujotzek M, Xiao S, Stock R, Kausch L, Schader P, Penzkofer T, Jäger PF, Maier-Hein K: Leveraging foundation models for content-based image retrieval in radiology. Comput Biol Med 196:110640, 2025. https://doi.org/10.1016/j.compbiomed.2025.110640

9. Zhang S, Xu Y, Usuyama N, Xu H, Bagga J, Tinn R, Preston S, Rao R, Wei M, Valluri N, Wong C, Tupini A, Wang Y, Mazzola M, Shukla S, Liden L, Gao J, Crabtree A, Piening B, Bifulco C, Lungren MP, Naumann T, Wang S, Poon H: A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI 2:AIoa2400640, 2025. https://doi.org/10.1056/AIoa2400640

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

11. Pai S, Hadzic I, Bontempi D, Bressem K, Kann BH, Fedorov A, Mak RH, Aerts HJWL: Vision foundation models for computed tomography. arXiv preprint arXiv:2501.09001, 2025. https://doi.org/10.48550/arXiv.2501.09001

12. Muthyala R, Yin Z, Jilla A, Li F, Dapamede T, Khosravi B, Chavoshi M, Gichoya J, Purkayastha S: Frozen foundation-model embeddings discard small-lesion signal in chest radiography: implications for pre-deployment evaluation. arXiv preprint arXiv:2606.11606, 2026. https://doi.org/10.48550/arXiv.2606.11606

13. Li Y, Wu Y, Lai Y, Hu M, Yang X: MedDINOv3: how to adapt vision foundation models for medical image segmentation? arXiv preprint arXiv:2509.02379, 2025. https://doi.org/10.48550/arXiv.2509.02379

14. Liu C, Chen Y, Shi H, Lu J, Jian B, Pan J, Cai L, Wang J, Yu J, Gao Z, Zhang X, Bai L, Zhang Y, Li J, Bercea CI, Ouyang C, Chen C, Xiong Z, Wiestler B, Wachinger C, Duncan JS, Rueckert D, Bai W, Arcucci R: Does DINOv3 set a new medical vision standard? Benchmarking 2D and 3D classification, segmentation, and registration. arXiv preprint arXiv:2509.06467, 2025. https://doi.org/10.48550/arXiv.2509.06467

15. Mehre SA, Dhara AK, Garg M, Kalra N, Khandelwal N, Mukhopadhyay S: Content-based image retrieval system for pulmonary nodules using optimal feature sets and class membership-based retrieval. J Digit Imaging 32:362–385, 2019. https://doi.org/10.1007/s10278-018-0136-1

16. Kravets I, Heletz T, Greenspan H: Nodule2vec: a 3D deep learning system for pulmonary nodule retrieval using semantic representation. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2020. Lect Notes Comput Sci 12266:599–608, 2020. https://doi.org/10.1007/978-3-030-59725-2_58

17. Na I, Rue N, Chung J, Park H: RadiomicsRetrieval: a customizable framework for medical image retrieval using radiomics features. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2025. Lect Notes Comput Sci 15960:556–566, 2025. https://doi.org/10.1007/978-3-032-04927-8_53

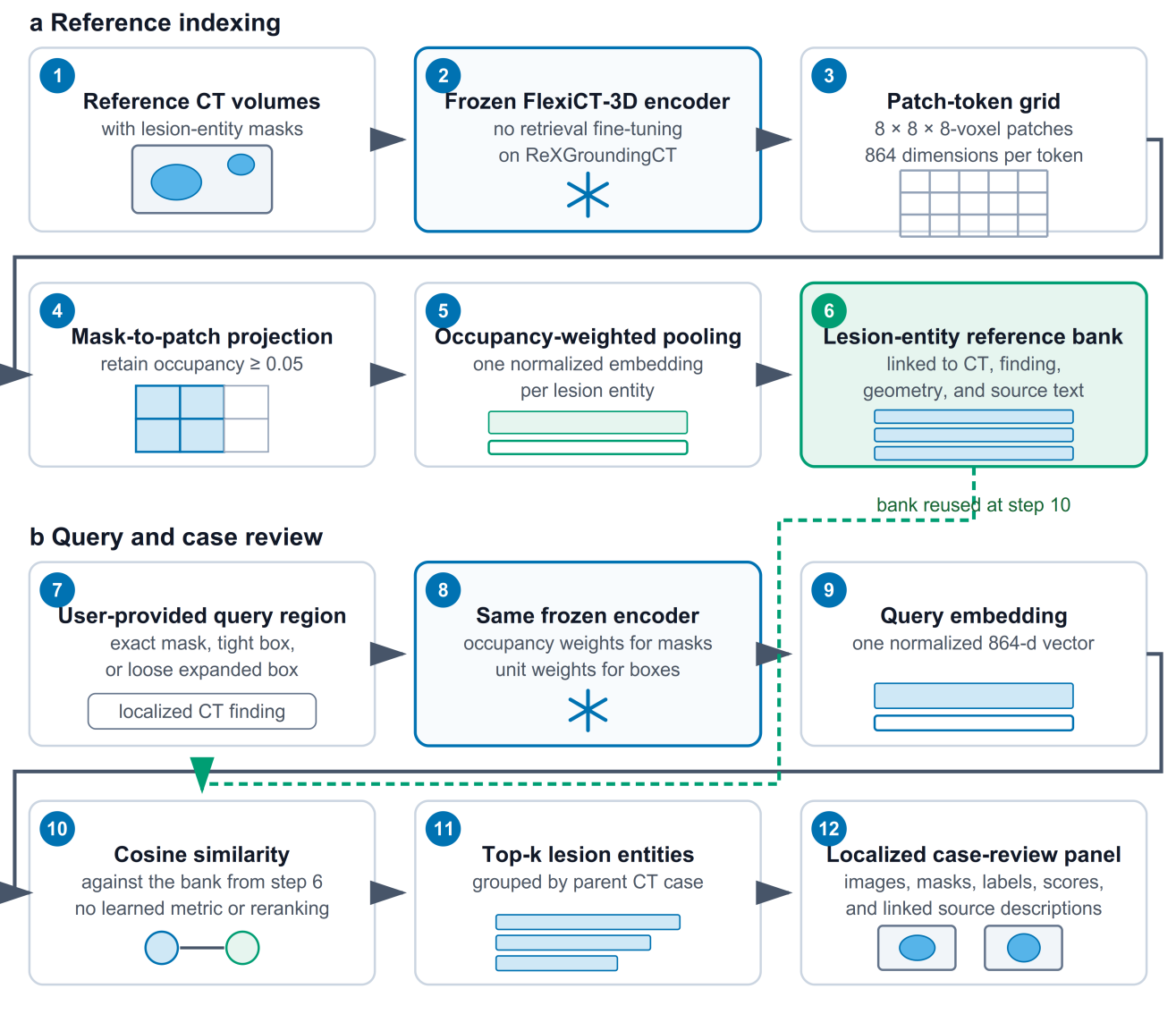
18. Faruque J, Beaulieu CF, Rosenberg J, Rubin DL, Yao D, Napel S: Content-based image retrieval in radiology: analysis of variability in human perception of similarity. J Med Imaging (Bellingham) 2:025501, 2015. https://doi.org/10.1117/1.JMI.2.2.025501

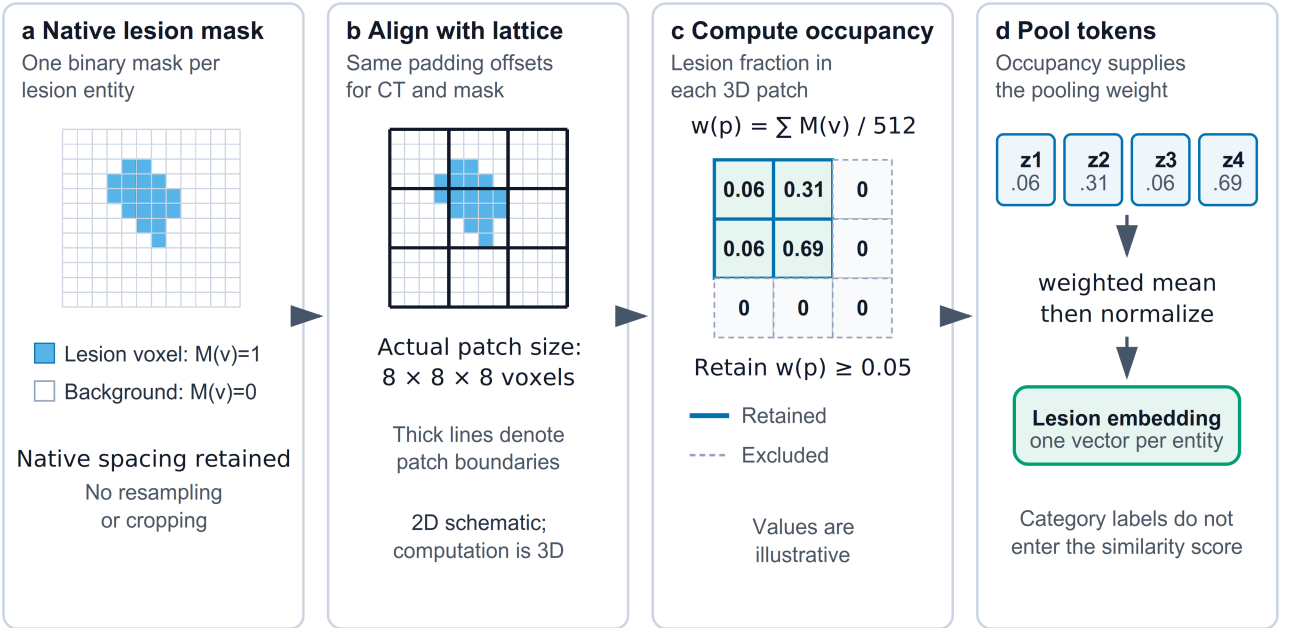
19. Choe J, Hwang HJ, Seo JB, Lee SM, Yun J, Kim MJ, Jeong J, Lee Y, Jin K, Park R, Kim J, Jeon H, Kim N, Yi J, Yu D, Kim B: Content-based image retrieval by using deep learning for interstitial lung disease diagnosis with chest CT. Radiology 302:187–197, 2022. https://doi.org/10.1148/radiol.2021204164

20. Haubold J, Zeng K, Farhand S, Stalke S, Steinberg H, Bos D, Meetschen M, Kureishi A, Zensen S, Goeser T, Maier S, Forsting M, Nensa F: AI co-pilot: content-based image retrieval for the reading of rare diseases in chest CT. Sci Rep 13:4336, 2023. https://doi.org/10.1038/s41598-023-29949-3

21. Böttcher B, van Assen M, Fari R, von Knebel Doeberitz PL, Kim EY, Berkowitz EA, Meinel FG, De Cecco CN: Evaluation of a content-based image retrieval system for radiologists in high-resolution CT of interstitial lung diseases. Eur Radiol Exp 9:4, 2025. https://doi.org/10.1186/s41747-024-00539-w

22. Silva W, Gonçalves T, Härmä K, Schröder E, Obmann VC, Barroso MC, Poellinger A, Reyes M, Cardoso JS: Computer-aided diagnosis through medical image retrieval in radiology. Sci Rep 12:20732, 2022. https://doi.org/10.1038/s41598-022-25027-2

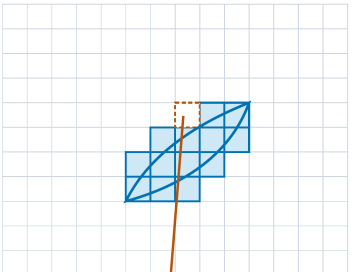

a Reference indexing
1
Reference CT volumes
with lesion-entity masks
2
Frozen FlexiCT-3D encoder
no retrieval fine-tuning
on ReXGroundingCT
3
Patch-token grid
8 × 8 × 8-voxel patches
864 dimensions per token
4
Mask-to-patch projection
retain occupancy ≥ 0.05
5
Occupancy-weighted pooling
one normalized embedding
per lesion entity
6
Lesion-entity reference bank
linked to CT, finding,
geometry, and source text
bank reused at step 10
b Query and case review
7
User-provided query region
exact mask, tight box,
or loose expanded box
localized CT finding
8
Same frozen encoder
occupancy weights for masks
unit weights for boxes
9
Query embedding
one normalized 864-d vector
10
Cosine similarity
against the bank from step 6
no learned metric or reranking
11
Top-k lesion entities
grouped by parent CT case
12
Localized case-review panel
images, masks, labels, scores,
and linked source descriptions



Schematic: the displayed grid is 2D; mask projection and token pooling use the full 3D patch lattice.

a Exact lesion mask

`exact_lesion_mask_v1`

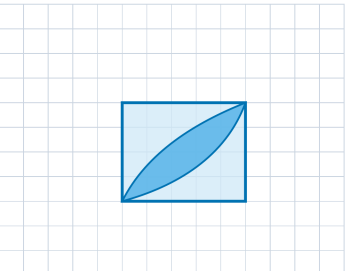


13 patches pooled

Occupancy weights; threshold 0.05

b Tight bounding box

`bbox_mask_v1`

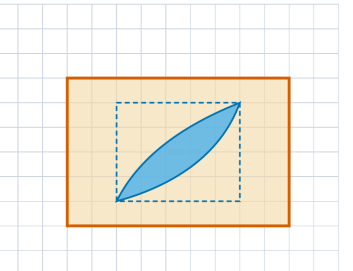


20 patches pooled

Unit weights across the whole box

c Loose expanded box

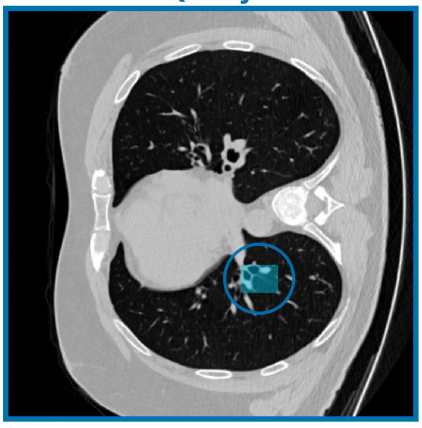
`loose_user_like_box_v1`



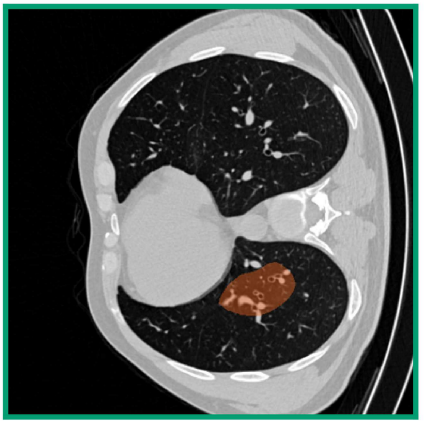
54 patches pooled

Unit weights; surrounding tissue included

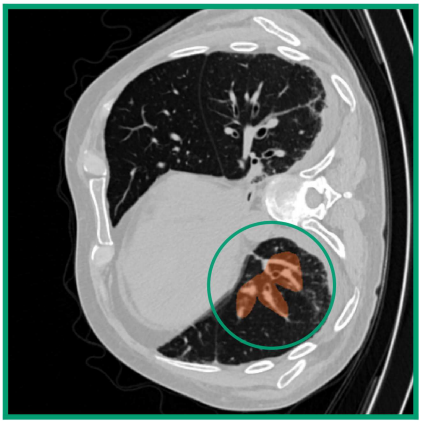
decreasing localization precision



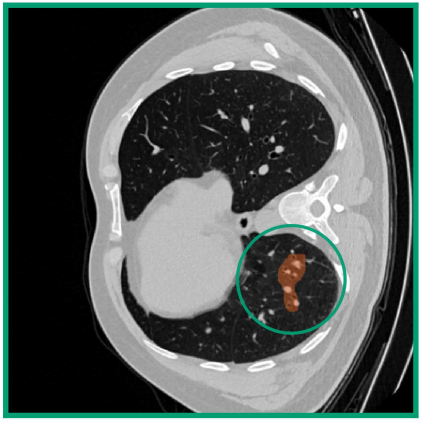
train_13252_b_2, lesion 4
1b bronchiectasis
medium size tertile



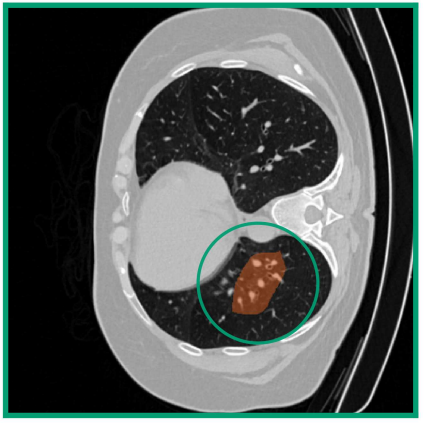
1b cos 0.936
same subcategory



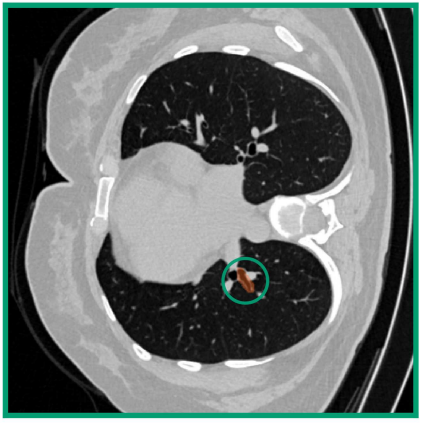
1b cos 0.935
same subcategory



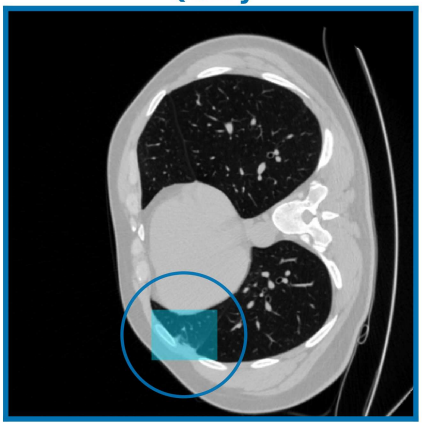
1b cos 0.921
same subcategory



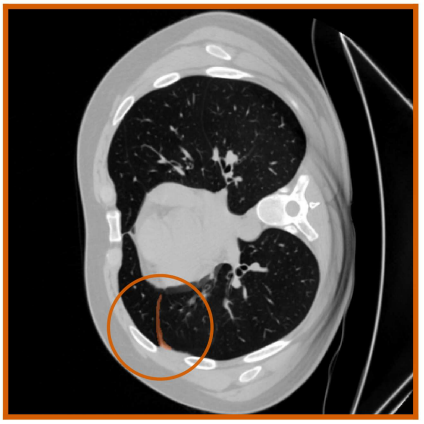
1b cos 0.920
same subcategory



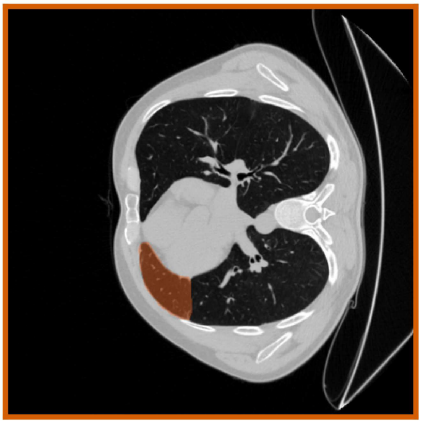
1b cos 0.919
same subcategory



train_13177_a_2, lesion 0
2b atelectasis or consolidation
large size tertile



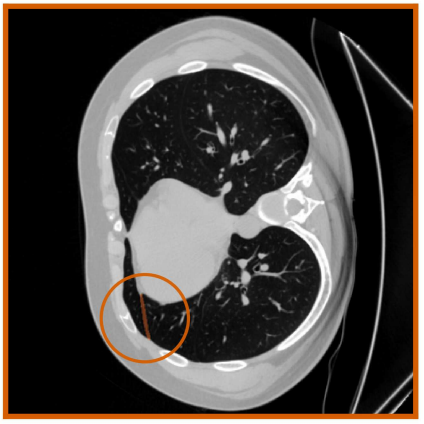
2a cos 0.813
different subcategory
same parent category



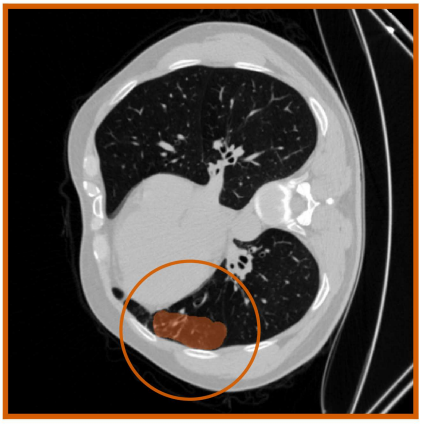
2a cos 0.808
different subcategory
same parent category



2a cos 0.802
different subcategory
same parent category



2a cos 0.797
different subcategory
same parent category



2c cos 0.797
different subcategory
same parent category

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.pdf