Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.081416

ARTICLE

Truth-Anchored Evidence-Sensitive Training for Multimodal Radiology LLMs via Dual-Extractor Disagreement and Deterministic Counterfactual Constraints

Xiong Luo*

Department of Information Technology, Uppsala University, Uppsala, Sweden

*Corresponding Author: Xiong Luo. Email: xiong.luo.7609@student.uu.se

Received: 02 March 2026; Accepted: 13 April 2026; Published: 15 June 2026

ABSTRACT: Large multimodal models (LMMs) can produce fluent radiology reports, yet two clinically important error modes remain common: unsupported assertions and missed findings. Optimizing both under open supervision remains difficult because many pipelines still rely on overlapping parser families during training and evaluation. This paper introduces Truth-Anchored Dual-Extractor Counterfactual-Constrained Training (TA-DECT), which combines an ontology-derived atomic finding interface with four coupled objectives: structured prediction, dual-extractor minimax consistency on generated reports, deterministic counterfactual selectivity under evidence removal, and label-anchored completeness. In matched-path internal comparisons across chest radiographs (CheXpert, MIMIC-CXR, MIMIC-CXR-JPG) and chest computed tomography (CT; CT-RATE), TA-DECT improves truth-anchored F1 while reducing both missed-finding and unsupported-assertion rates, with concurrent gains in calibration and selectivity. On held-out region-of-interest (ROI) datasets (MS-CXR, VinDr-CXR), it also improves coarse evidence linkage and intervention-targeted confidence responses under occlusion. In this revision, the strongest claims are kept explicitly anchored to structured labels and ROI references, counterfactual evidence-sensitivity summaries are interpreted with bootstrap uncertainty, and parser-derived report metrics are retained only as supplementary diagnostics.

KEYWORDS: Multimodal radiology; large language models; report generation; counterfactual training; evidence grounding; structured labels

1 Introduction

Radiology report generation is a canonical multimodal task, but practical usefulness depends on image-grounded correctness rather than linguistic fluency alone. Recent encoder–decoder and LMM-based systems produce increasingly fluent prose [1,2]; however, two safety-relevant error modes remain frequent in practice: unsupported assertions and omitted findings. These errors are coupled, because reducing unsupported content can be achieved trivially by suppressing ground-truth findings.

The core methodological challenge is evaluation under open supervision. Many pipelines optimize report likelihood and score outputs with automatic report parsers or lexical surrogates. This is scalable, yet it can blur whether apparent gains reflect better image-grounded reasoning or closer alignment to parser behavior.

Public radiology datasets now expose richer anchors, including structured labels, curated subsets, and region annotations. What remains missing is a unified framework that uses these heterogeneous signals jointly while keeping omission control and evidence sensitivity explicit objectives rather than indirect side effects.

This study addresses this gap with Truth-Anchored Dual-Extractor Counterfactual-Constrained Training (TA-DECT), centered on an ontology-derived atomic finding interface between vision and free-form text. TA-DECT couples structured supervision, dual-extractor minimax consistency, deterministic counterfactual constraints, and label-anchored completeness in one training objective.

This study contributes three elements: (i) a coupled training formulation that jointly targets correctness, completeness, and evidence sensitivity; (ii) deterministic mappings from public label/report sources into a shared tuple schema ⟨concept_id, anatomy_region_id, polarity, uncertainty⟩; and (iii) an evaluation protocol that prioritizes structured labels and ROI anchors, with parser-derived signals used as supplementary diagnostics. Accordingly, the primary claims of this paper are anchor-based rather than parser-based: parser outputs are used as constrained auxiliaries during training and as secondary diagnostics at test time, but not as the main evidence for clinical correctness.

2 Related Work

Report-generation research has advanced from encoder–decoder systems to LMM-based pipelines that condition large text decoders on visual tokens. Representative models such as R2Gen, R2GenGPT, KARGEN, and LLaVA-Med improve fluency and overall report quality [3–6]. However, fluency-oriented progress does not by itself ensure lower unsupported-assertion or missed-finding rates.

A parallel literature studies robustness across institutions and modalities [7,8]. These studies show that in-domain gains can erode under style shift and modality transfer. Yet evaluation is still frequently parser-dominated, making it difficult to separate image-grounded improvement from metric-specific alignment, especially when related extractor families appear in both optimization and evaluation pathways.

Recent multimodal biomedical representation work also highlights the value of structured cross-modal alignment beyond radiology report generation alone. GTP-4o studies modality-prompted heterogeneous graph learning for omni-modal biomedical representation [9], while PV-SSM explores pure visual state-space modeling for high-dimensional medical data analysis [10]. These studies are not direct report-generation baselines, but they are relevant to the broader design questions considered here: multimodal alignment, architecture choice under modality heterogeneity, and structured reasoning over medical evidence.

This work is most closely connected to open-supervision settings that combine heterogeneous public sources, including CheXpert and CheXpert Plus [2,11], MIMIC-CXR [1], MS-CXR [12], VinDr-CXR [13], and CT-RATE [14]. The key distinction here is the joint use of a deterministic atomic interface, anchor-gated dual-extractor minimax consistency, and counterfactual/completeness constraints within one truth-anchored evaluation protocol.

3 Method

3.1 Problem Setting and Design Goals

This paper considers multimodal radiology generation from either a 2D chest radiograph (CXR) or a 3D chest CT volume. The system is required to produce two outputs: (i) a set of structured findings that can be checked against structured labels, and (ii) a free-form impression-style report for human readability and downstream use.

The key design requirement is that the system should be optimized and evaluated under open-access supervision, using released anchors: structured labels, reference-standard subsets when available, and reference regions when available. A second requirement is non-circularity: text/extractor signals may be

used as auxiliary constraints during training, but quantitative correctness evaluation is anchored to held-out structured labels and reference regions under predefined rules.

These goals are implemented by introducing an ontology-derived atomic-finding interface as a bottleneck between vision and free-form text. This interface serves two roles. During training, it provides a structured channel for supervision and for counterfactual constraints that directly manipulate evidence. During evaluation, it provides a consistent target space for measuring correctness and completeness under a shared protocol.

3.2 Ontology-Derived Atomic-Finding Interface

The interface is defined at two levels. The supervised prediction index is an atomic slot

$$f = \langle c, r \rangle, \tag{1}$$

where c is a biomedical concept identifier (e.g., RadLex) and r is an anatomy region identifier. For consistent bookkeeping and cross-dataset conversion, each slot is associated with tuple states

$$a_f = \langle c, r, s, u \rangle, \quad s \in \{\text{pos}, \text{neg}\}, \; u \in \{\text{certain}, \text{uncertain}\}. \tag{2}$$

Thus, polarity and uncertainty are explicit in the tuple schema, while model supervision and primary metrics use the positive projection of each slot.

First, syntactic validity is required to be checkable without dataset-specific heuristics. A tuple record is valid if and only if all four fields are present and each value lies in a fixed, known inventory. This keeps conversions consistent under domain shift in report style.

Second, cross-dataset comparability is needed. Public datasets differ in label definitions and in whether uncertainty is encoded. By mapping each dataset's released structured labels into this common tuple schema, a single slot space $\mathcal{F}$ can be evaluated under multiple anchors with consistent rules.

The inventory of valid concepts and regions is constructed from (i) a public biomedical concept ontology (RadLex) and (ii) the concept inventories covered by two open-source report labelers used as fixed extractors (CheXpert labeler [2] and NegBio [15]). Across datasets, the mapping from released structured labels (e.g., CheXpert observations, VinDr labels, CT-RATE abnormalities) to atomic tuples is predefined.

3.3 Tuple Mapping Protocol

All label-to-tuple and report-to-tuple conversions are rule-based with no learned thresholds beyond the global slot-level decision rule $p_f \geq 0.5$ used in the evaluation protocol below. Table 1 summarizes the complete mapping logic used by training and evaluation.

Mapping ambiguity is accounted for explicitly rather than hidden in dataset-specific scripts. If anatomical region information is absent, the default protocol maps the finding to a thoracic-global identifier so that the study remains countable under the shared slot inventory; if different released sources disagree, a fixed precedence rule is used once and then shared across all methods. Because the same tuple interface is used for TA-DECT, internal baselines, and parser-converted external rows, this mapping layer can shift absolute numbers but does not selectively favor one method family. The experiments therefore include compact sensitivity analyses under alternate region, uncertainty, and conflict-resolution rules.

Table 1: Tuple mapping protocol for training and evaluation.

Source Signal	Condition	Mapping Rule
Structured labels (CheXpert, MIMIC-CXR-JPG, CT-RATE, VinDr)	Label = pos/neg/uncertain/blank	Emit tuple $(c, r_d, \text{pos}, \text{certain})$ for pos; $(c, r_d, \text{neg}, \text{certain})$ for neg; $(c, r_d, \text{pos}, \text{uncertain})$ for uncertain; emit nothing for blank (excluded from supervised/eval counts).
Report parser output (E_1, E_2)	Parser state = positive/uncertain/negative/blank	Convert to binary indicator for minimax and parser-path evaluation: positive or uncertain $\mapsto 1$, negative or blank $\mapsto 0$.
Anatomy region id for finding f	Dataset provides anatomical location/not provided	If provided, map via fixed dataset-to-region dictionary r_d; if unavailable, use fixed thoracic-global region id.
MS-CXR/VinDr ROI to tuple mapping	Multiple boxes/categories/overlaps in one study	Map each ROI category by a fixed dictionary to concept set $C(i)$, map box center to region $r(i)$, then create slots $(c, r(i))$ for all $c \in C(i)$. For repeated boxes with same (c, r) use set-union (deduplicate). For overlapping boxes with different categories, keep all mapped concepts (multi-label). For conflicting polarity states from different sources, apply precedence pos > uncertain > neg.
Intervention region mapping	Grid or ROI intervention i	Map intervention center to one of 9 coarse regions (left/center/right) × (upper/middle/lower). Grid path uses $T_{\text{reg}}(i) = \{f : \text{region}(f) = r(i)\}$. ROI path uses $T_{\text{roi}}(i) = \{f : \text{region}(f) = r(i) \wedge \text{concept}(f) \in C(i)\}$.
Binary finding for metrics	Native model scores/parser outputs binary labels	Native path: $\hat{y}_f = 1$ iff $p_f \geq 0.5$ for slot $f = \langle c, r \rangle$, else 0. Parser path: use converted binary parser output (positive/uncertain $\mapsto 1$, negative/blank $\mapsto 0$).

3.4 Model Architecture

A multimodal encoder–decoder is trained to produce (i) atomic finding probabilities and (ii) a free-form impression-style report. The architecture is intentionally simple in order to isolate the effect of the proposed training objective and evaluation protocol.

The visual backbone is BiomedCLIP's image encoder (`ViT-B/16`, about 86M parameters) [16]. For CXR, inputs are resized to 448 × 448 and encoded into patch tokens. For CT, volumes are clipped to Hounsfield units (HU) $[-1000, 400]$, linearly rescaled, and 32 axial slices are sampled uniformly per study; each slice is encoded by the same `ViT-B/16` and slice tokens are aggregated by a two-layer attention-pooling block. This design is intentionally conservative: it preserves a matched image-to-token interface across CXR and CT and isolates the training objective from a larger architecture search, but it does not claim that 32-slice sampling is the best possible three-dimensional encoder. The revision therefore includes a higher-slice CT variant while keeping the objective fixed.

The atomic head is a multi-label classifier that predicts a probability $p_f(x) \in [0,1]$ for each atomic slot $f = \langle c, r \rangle$. Here $p_f(x)$ denotes confidence for the slot-level positive state used by the label-anchor protocol (with uncertain/blank handled by predefined exclusion/mapping rules). The output of this head is the primary vehicle for truth-anchored correctness and for evidence sensitivity under interventions.

The text decoder is Mistral-7B-v0.1 [17] with added cross-attention blocks that consume visual tokens and an atomic-summary token stream. Decoder adaptation uses LoRA [18] (rank $r = 16$, $\alpha = 32$, dropout 0.05) on attention and MLP projections; base decoder weights remain frozen. This yields about 7.24B total decoder parameters with about 93M trainable decoder-side parameters. In the default implementation, cross-attention adapters are inserted every 4 decoder blocks (8 insertion points over the 32-layer stack), the atomic summary is capped at 64 tokens, and pooled visual conditioning tokens are capped at 256 tokens. Conditioning on the structured predictions encourages report–structure consistency, while the report output remains free-form.

Fig. 1 summarizes the TA-DECT architecture and training objective, highlighting the atomic bottleneck, occlusion interventions, and the detached dual-extractor training auxiliary.

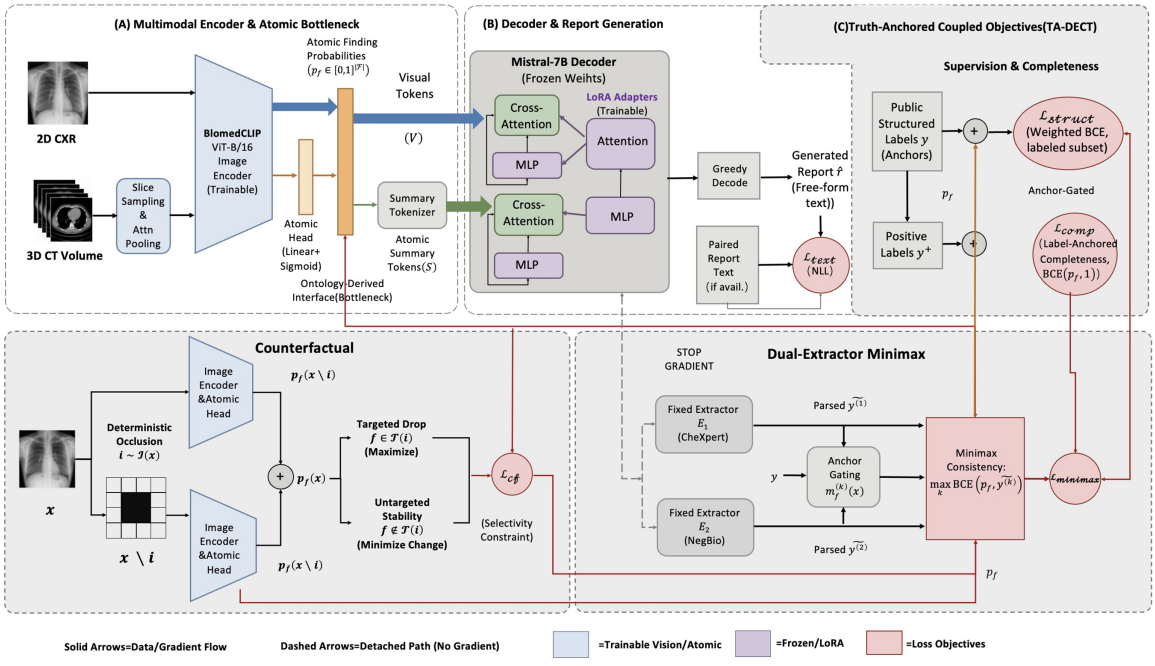

Figure 1: TA-DECT architecture and objective terms. The fixed extractors are E_1 (CheXpert labeler) and E_2 (NegBio). Extractor outputs provide detached targets for $\mathcal{L}_{\text{minimax}}$; dashed/dotted links indicate detached or attention paths rather than parser-through-text gradients.

3.5 TA-DECT Objective

Let x be an image (CXR) or volume (CT), y be public structured labels when available, and r be the generated report. The TA-DECT objective couples four loss terms with an optional report likelihood term. The terms are designed to address distinct but interacting risks: label-supervised correctness, extractor-specific gaming, non-selective evidence sensitivity, and omission.

Training begins with structured supervision when labels are available. On examples with structured labels, define the slot-level positive target $y_f^+ = \mathbb{1}[y_f = \text{pos}]$ on the evaluable label subset $y_f \in \{\text{pos}, \text{neg}\}$ (uncertain/blank excluded by a predefined policy). A weighted multi-label binary cross-entropy over slots is then optimized:

$$\mathcal{L}_{\text{struct}} = \sum_{f \in \mathcal{F}} w_f \, \text{BCE}(p_f(x), \, y_f^+), \tag{3}$$

where labels with state $\{\text{uncertain}, \text{blank}\}$ are excluded under a predefined policy (reported as an excluded fraction). Weights are set to $w_f \propto 1/\sqrt{\pi_f + \epsilon}$ where π_f is the training-set prevalence of slot f (computed on the available structured-label subset for that dataset) and $\epsilon = 10^{-3}$. This weighting reduces the tendency of rare findings to be ignored while preserving a consistent training setup.

To reduce circularity from a single evaluator, a dual-extractor minimax constraint is imposed on generated text. Two fixed report-to-finding extractors E_1 (CheXpert labeler) and E_2 (NegBio) are applied to the generated report r, producing extracted finding indicators $\tilde{y}_f^{(k)}(r) \in \{0, 1\}$. The worst-case mismatch between atomic predictions and extracted indicators is minimized:

$$\mathcal{L}_{\text{minimax}} = \max_{k \in \{1,2\}} \sum_{f \in \mathcal{F}_{\text{parse}}(x)} m_f^{(k)}(x) \, \text{BCE}(p_f(x), \, \tilde{y}_f^{(k)}(r)). \tag{4}$$

Here $\mathcal{F}_{\text{parse}}(x) \subseteq \mathcal{F}$ denotes slots covered by the fixed extractor inventories for the report domain of x. On CT-RATE, minimax updates are restricted to this parser-supported subset; slots outside $\mathcal{F}_{\text{parse}}$ are optimized only by $\mathcal{L}_{\text{struct}}$, $\mathcal{L}_{\text{cf}}$, and $\mathcal{L}_{\text{comp}}$. The mask $m_f^{(k)}(x)$ is an anchor-gating rule:

$$m_f^{(k)}(x) = \begin{cases} 1, & \text{if } y_f \text{ is unavailable,} \\ 1, & \text{if } y_f \in \{\text{pos}, \text{neg}\} \text{ and } \tilde{y}_f^{(k)}(r) = y_f^+, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Hence, on labeled samples, extractor outputs that conflict with the structured anchor do not contribute gradient. In optimization, this term is implemented as a deterministic detached-target step because report decoding and extractor parsing are non-differentiable. For each minibatch, one greedy report $\hat{r}$ is decoded per sample, E_1/E_2 are run on $\hat{r}$, and the resulting $\tilde{y}_f^{(k)}(\hat{r})$ values are treated as constants when optimizing $\mathcal{L}_{\text{minimax}}$. Gradients from $\mathcal{L}_{\text{minimax}}$ update the visual encoder and atomic head (through $p_f(x)$), while decoder parameters are updated by $\mathcal{L}_{\text{text}}$ and by the shared conditioning on the atomic summary. This block-coordinate implementation makes the gradient path explicit: $\mathcal{L}_{\text{minimax}}$ never backpropagates through parser outputs or report text, and it cannot override available structured labels because the anchor gate suppresses conflicting parser targets. Extractor states are converted to binary indicators using the fixed rule in Table 1 (positive/uncertain $\mapsto$ 1, negative/blank $\mapsto$ 0); the maximization over k applies the larger mismatch between the two fixed extractors for each update. Even with this safeguard, the extractors remain part of the training auxiliary loop, so this dependence is stated explicitly in the revised discussion and limitations rather than treated as invisible infrastructure.

Next, a counterfactual selectivity constraint is added, which directly tests whether predicted confidences respond selectively to the removal of evidence. An intervention family $i \sim \mathcal{I}(x)$ removes evidence by occluding either (a) a fixed image grid region (used in training and evaluation) or (b) a reference bounding box (evaluation only on held-out ROI splits).

Grid occlusion (train+eval). Training uses a fixed 3×3 grid: for 2D CXR, exactly one grid cell is occluded by replacing all pixels in that cell with the per-image mean intensity; for 3D CT, the same cell index is occluded across all axial slices. The 3×3 design is used in training because it is aligned to the coarse 9-region anatomy partition used by the atomic interface: (left/center/right) × (upper/middle/lower). In the revision experiments, a targeted retraining variant additionally mixes 3×3 and 7×7 interventions while keeping the rest of the objective unchanged.

ROI occlusion (eval-only). For MS-CXR and VinDr-CXR, given a held-out ground-truth box b, the pixels inside b are occluded (again replaced by the per-image mean intensity). ROI occlusion is never used for training.

Intervention-to-region and ROI-category-to-concept mappings follow the predefined rules in Table 1. For grid interventions, the targeted set is $T_{\text{reg}}(i) = \{f \in \mathcal{F} : \text{region}(f) = r(i)\}$, and for ROI interventions with category labels (MS-CXR and VinDr-CXR), the targeted set is $T_{\text{roi}}(i) = \{f \in \mathcal{F} : \text{region}(f) = r(i) \ \wedge \ \text{concept}(f) \in C(i)\}$.

Intervention distribution. For the default training setup (3×3), $\mathcal{I}(x)$ is uniform over the 9 grid cells and one cell is sampled per example. At evaluation, metrics are averaged over all interventions in the tested grid family (3×3, 7×7, or 16×16) or over all held-out ROI boxes (ROI occlusion). Interventions with $|T(i)| = 0$ are excluded from both training and evaluation averages.

The counterfactual term enforces selective decreases for targeted findings while stabilizing untargeted findings:

$$\mathcal{L}_{\text{cf}} = \mathbb{E}_{i \sim \mathcal{I}(x)} \Big[\underbrace{\frac{1}{|T(i)|} \sum_{f \in T(i)} \max\big(0,\, m - (p_f(x) - p_f(x \backslash i))\big)}_{\text{targeted drop}} \tag{6}$$

$$+ \lambda \underbrace{\frac{1}{|\mathcal{F} \backslash T(i)|} \sum_{f \notin T(i)} |p_f(x) - p_f(x \backslash i)|}_{\text{untargeted stability}} \Big], \tag{7}$$

with margin $m > 0$. The experiments use $m = 0.05$ and $\lambda = 1.0$ throughout. Intuitively, the first term demands a minimum decrease for findings logically linked to the occluded region, while the second term penalizes global instability that would be consistent with non-specific confidence collapse.

Finally, to explicitly control omission, a label-anchored completeness constraint is added on examples with public structured labels:

$$\mathcal{L}_{\text{comp}} = \sum_{f \in \mathcal{F}} \alpha \, y_f^+ \, \text{BCE}(p_f(x), 1). \tag{8}$$

α is set to 1 for all findings, and the overall strength is controlled via δ in the total objective. This term adds a direct penalty on false negatives under the structured anchor: low-confidence predictions on positive labels increase this loss.

On datasets with paired report text, negative log-likelihood $\mathcal{L}_{\text{text}}$ is also optimized. The full objective is

$$\mathcal{L} = \mathcal{L}_{\text{struct}} + \beta\mathcal{L}_{\text{minimax}} + \gamma\mathcal{L}_{\text{cf}} + \delta\mathcal{L}_{\text{comp}} + \eta\mathcal{L}_{\text{text}}. \tag{9}$$

Unless explicitly varied in ablations, weights are set to $\beta = 0.5$, $\gamma = 1.0$, $\delta = 0.5$, and $\eta = 1.0$ on paired-text datasets; for image-only datasets, $\eta = 0$. Text losses are applied only when paired text exists; on image-only datasets (e.g., VinDr-CXR), training uses only atomic losses.

3.6 Truth-Anchored Evaluation Metrics

The evaluation protocol is designed to make correctness and completeness jointly visible under a consistent rule set, while keeping extractor-based signals diagnostic. The evaluation emphasizes micro-averaged, label-anchored outcomes because they directly quantify omitted positives and unsupported positives in a form that remains interpretable across datasets.

Label-anchored indicators (micro-averaged) are used on held-out truth anchors. For each example x and slot f with ground-truth $y_f(x) \in \{\text{pos}, \text{neg}\}$, predicted-positive is defined as $\hat{y}_f(x) = \mathbb{1}[p_f(x) \geq 0.5]$. The default threshold $p_f \geq 0.5$ is fixed once and shared across datasets, modalities, and internal methods so that comparisons do not inherit post hoc threshold tuning. Because different threshold choices can materially change the miss–unsupported balance, the revision experiments include a dedicated robustness sweep rather than treating 0.5 as intrinsically optimal.

$$\text{TP} = \sum_{x,f} \mathbb{1}\left[y_f^+ = 1 \wedge \hat{y}_f = 1\right], \ \text{FP} = \sum_{x,f} \mathbb{1}\left[y_f^+ = 0 \wedge \hat{y}_f = 1\right], \tag{10}$$

$$\text{FN} = \sum_{x,f} \mathbb{1}\left[y_f^+ = 1 \wedge \hat{y}_f = 0\right]. \tag{11}$$

Here TP, FP, and FN denote true positives, false positives, and false negatives. Missed-finding rate = $\text{FN}/(\text{TP} + \text{FN})$, unsupported-assertion rate = $\text{FP}/(\text{TP} + \text{FP})$, and F1 = $2\text{TP}/(2\text{TP} + \text{FP} + \text{FN})$. For anchors with explicit uncertain/blank states (CheXpert-style labels), uncertain/blank are excluded under a fixed policy. The excluded fraction (Excl.) is computed once from ground-truth labels on each anchor split and is therefore method-invariant within that split. This policy improves comparability under a binary slot projection, but it may hide clinically relevant uncertainty failures; it is therefore treated as a limitation rather than as a claim that uncertain labels are unimportant in practice.

For counterfactual evidence sensitivity, the study reports the selectivity metric. For input x and intervention i, let $p_f(x)$ be confidence for finding f, and $p_f(x\backslash i)$ after evidence removal. Define

$$\text{targeted\_drop}(x, i) = \frac{1}{|T(i)|} \sum_{f \in T(i)} \max(0, p_f(x) - p_f(x\backslash i)), \tag{12}$$

$$\text{untargeted\_change}(x, i) = \frac{1}{|\mathcal{F}\backslash T(i)|} \sum_{f \notin T(i)} |p_f(x) - p_f(x\backslash i)|, \tag{13}$$

and report selectivity $S = \mathbb{E}[\text{targeted\_drop} - \text{untargeted\_change}]$. For grid interventions, $T(i) = T_{\text{reg}}(i)$; for ROI interventions with category labels, $T(i) = T_{\text{roi}}(i)$. The subtraction makes the metric sensitive to two common failure patterns: (i) insufficient response for targeted findings and (ii) non-specific instability that changes many untargeted findings.

Evaluation pathways follow method family. Methods with a native atomic head are evaluated directly from $p_f(x)$ (including calibration). Report-only baselines are converted to binary findings using the same

dual-parser rule in Table 1; because this pathway does not expose calibrated $p_f(x)$, expected calibration error (ECE) is not reported for report-only external baselines.

Calibration is evaluated using expected calibration error (ECE, 15 bins) for methods with native atomic probabilities, while extractor disagreement on generated reports is reported separately as a diagnostic signal.

4 Experiments

4.1 Datasets, Splits, and Protocol

TA-DECT is evaluated on public datasets chosen to cover paired image–report supervision, structured-label anchors, region-level evidence anchors, and cross-modality transfer. Where available, the study reports three complementary channels: label-anchored finding quality, region-of-interest (ROI)-based evidence linkage, and counterfactual selectivity.

Table 2 summarizes, for each dataset, modality/text availability, applied text losses, evaluation anchors, reference-region availability, intervention type, and train/test role.

Table 2: Datasets and evaluation channels matrix. Modality is abbreviated as Modality in the column header, and paired report text availability is abbreviated as Text. Y* indicates sentence-level text exists but is not used for training here; Y+ indicates paired report text provided by CheXpert Plus on the CheXpert domain. Interventions distinguish grid occlusions (train+evaluation) from ROI occlusion (evaluation-only on held-out ROI splits).

Dataset	Mod.	Text?	Text Losses	Primary Anchor	Secondary Anchor	Regions?	Interventions	Shift Role
CheXpert	I + T	Y+	NLL + Minimax	Struct. labels (held-out)	Rad. benchmark (diag.)	N	Grid (Tr + Te)	Tr + Te
MIMIC-CXR	I + T	Y	NLL + Minimax	Man. cur. (JPG)	CheX/NegBio (JPG)	Y (MS-CXR)	Grid (Tr)	Tr
MIMIC-CXR-JPG	I + Tab	N	–	Man. cur.	CheX/NegBio	N	Grid (Te only)	Te
MS-CXR	ROI + T	Y*	–	Box cats	–	Y	ROI (Te only)	Te
VinDr-CXR	I + Tab	N	–	Rad. labels	–	Y	ROI (Te only)	Te
CT-RATE	I + T + Tab	Y	NLL + Minimax	Struct. labels	–	N	Grid (Tr + Te)	Tr + Te

CheXpert provides uncertainty-aware labels and an official radiologist benchmark subset. In this manuscript, primary CheXpert counting uses a held-out split of released structured labels aligned to the full atomic slot inventory; the smaller radiologist benchmark (validation 200, test 500 in the official protocol) is used as a secondary diagnostic reference. When text losses are applied in the CheXpert domain, paired reports are sourced from CheXpert Plus and aligned by released image paths.

MIMIC-CXR serves as the primary paired CXR training domain and as a source/target for report-style shift. Structured-label evaluation for MIMIC uses released MIMIC-CXR-JPG benchmark labels (`mimic-cxr-2.1.0-test-set-labeled.csv`); consistent with the official PhysioNet release, some auxiliary metadata files in v2.1.0 retain `2.0.0` prefixes, and the released naming is kept unchanged. Train/validation/test partitions follow the official MIMIC-CXR-JPG split file, with paired free-text reports joined from MIMIC-CXR. Because MS-CXR is derived from MIMIC-CXR, the study enforces study-level de-overlap: any MIMIC study appearing in the MS-CXR evaluation pool is removed from MIMIC training and validation for all compared methods. MS-CXR and VinDr-CXR boxes are then used only for held-out ROI evaluation (evidence linkage and ROI occlusion), never as training supervision. CT-RATE provides paired 3D CT volumes, reports, and structured labels, and is used for both in-domain testing and cross-modality shift experiments.

4.2 Implementation and Training Details

All internal variants use the same backbone family and decoding policy unless a specific ablation requires a structural change: BiomedCLIP `ViT-B/16` with Mistral-7B-v0.1 and LoRA adaptation (rank 16, $\alpha = 32$, dropout 0.05). CXR inputs are resized to 448×448; CT volumes are clipped to HU $[-1000, 400]$, normalized, and sampled at 32 axial slices. This matched-backbone design is used to attribute performance differences to objective terms rather than to capacity changes.

Optimization uses AdamW (2×10^{-4} base LR), cosine decay, and 2k warmup. Batch sizes are 64 (CXR) and 8 (CT, with gradient accumulation to effective 64). Training runs for 80k steps on CXR and 40k on CT. Mixed-precision BF16, gradient checkpointing, and distributed sharding (FSDP-style) are used under the same configuration for all internal runs. Validation and checkpoint selection follow a single fixed rule across methods: select the checkpoint with highest validation anchor F1 under the same deterministic thresholding and exclusion policy used at test time.

For internal comparisons, preprocessing, augmentation, decoding, and intervention settings are shared across all rows. Inference uses beam size 3 and max report length 192 tokens, and atomic findings are positive at $p_f \geq 0.5$. Calibration settings (ECE with 15 bins), intervention families, and region-mapping rules are unchanged across datasets. Unless explicitly stated otherwise, reported point estimates are the mean of five independent seeds and confidence intervals use two-stage bootstrap resampling. The same uncertainty protocol is used for aggregate counterfactual and localization summaries; when a compact table reports point estimates only, the corresponding values are interpreted as descriptive operating-point summaries rather than standalone significance claims. Training hardware is 8×A100-80GB (or equivalent).

Optimization remained stable across the reported internal runs under the same warmup, sharding, and checkpoint-selection protocol. The main additional training cost of TA-DECT relative to the direct-to-text baseline comes from one greedy decode per sample plus two fixed extractor passes for the detached minimax target; the inference architecture itself is unchanged. Wall-clock comparisons are not reported here because they are cluster-dependent, but the later experiments isolate the extra components by matched-backbone ablations and targeted retraining variants.

4.3 Baselines and Ablations

Baselines are organized into two groups: internal matched-backbone baselines and external reference baselines. The internal group preserves backbone capacity, data splits, and decoding policy, and changes only objective components. The primary comparator is direct-to-text, which removes atomic conditioning while retaining the same visual and decoder family. Additional internal controls include Atomic+report without dual minimax, single-extractor consistency (E1 only), and a grounding-only auxiliary model that emphasizes localization without explicit semantic completeness constraints.

External baselines provide reference points against established methods. The comparison includes R2Gen, R2GenGPT, KARGEN, LLaVA-Med, CT2Rep, and continual tuning (all introduced and cited in Section 2). Because several external methods are report-only and do not expose calibrated slot probabilities, label-anchor outcomes are computed via the same dual-parser conversion used for all report-only rows, and ECE is omitted in those rows. The internal matched-backbone comparisons are therefore treated as the strongest causal evidence for the proposed objective, while the external rows serve as practical reference comparisons under a necessarily less symmetric evaluation interface.

Ablation design follows a one-factor-at-a-time protocol around the full TA-DECT objective. The ablation study removes each major term ($\mathcal{L}_{\text{minimax}}$, $\mathcal{L}_{\text{cf}}$, $\mathcal{L}_{\text{comp}}$), tests single-extractor replacement for minimax,

and tests removal of anchor masking in minimax. This protocol is intended to reveal whether gains arise from one dominant component or from interaction among complementary constraints.

4.4 Truth-Anchored Finding Quality

Table 3 reports micro-averaged finding quality on the primary structured-label anchors (CheX-hold., MIMIC-CXR-JPG curated labels, and CT-RATE structured labels), including F1, precision, recall, missed-finding rate, unsupported-assertion rate, ECE (when available), and exclusion fraction. Secondary report-derived rows are included for scale and extractor-drift diagnostics.

Table 3: Main truth-anchored finding results (point estimate ±95% bootstrap confidence interval (CI)).

Dataset	Label Anchor	Method	F1	Prec.	Rec.	Miss	Unsup.	ECE	Excl.
			CheXpert (internal)						
CheXpert	Struct. hold-out	TA-DECT	0.577 ± 0.011	0.602	0.554	0.446	0.398	0.063	0.28
CheXpert	Struct. hold-out	Direct-to-text (same bb.)	0.529 ± 0.018	0.540	0.519	0.481	0.460	0.079	0.28
CheXpert	Struct. hold-out	Atomic+report (no dual)	0.566 ± 0.014	0.590	0.544	0.456	0.410	0.069	0.28
CheXpert	Struct. hold-out	Cycle-cons. (single extr.)	0.560 ± 0.020	0.569	0.551	0.449	0.431	0.084	0.28
CheXpert	Struct. hold-out	Grounding-only aux.	0.524 ± 0.024	0.551	0.500	0.500	0.449	0.094	0.28
			CheXpert (external)						
CheXpert	Struct. hold-out	R2Gen (ext.)	0.501 ± 0.026	0.516	0.486	0.514	0.484	–	0.28
CheXpert	Struct. hold-out	R2GenGPT (ext.)	0.538 ± 0.015	0.566	0.513	0.487	0.434	–	0.28
CheXpert	Struct. hold-out	KARGEN (ext.)	0.536 ± 0.018	0.553	0.521	0.479	0.447	–	0.28
CheXpert	Struct. hold-out	LLaVA-Med (ext.)	0.509 ± 0.031	0.497	0.522	0.478	0.503	–	0.28
CheXpert	Struct. hold-out	Continual tuning (ext.)	0.554 ± 0.013	0.579	0.532	0.468	0.421	–	0.28
			MIMIC-CXR-JPG (internal)						
MIMIC-CXR-JPG	Man. curated	TA-DECT	0.557 ± 0.007	0.578	0.537	0.463	0.422	0.069	0.00
MIMIC-CXR-JPG	Man. curated	Direct-to-text (same bb.)	0.524 ± 0.015	0.543	0.506	0.494	0.457	0.081	0.00
MIMIC-CXR-JPG	Man. curated	Atomic+report (no dual)	0.548 ± 0.010	0.571	0.527	0.473	0.429	0.071	0.00
MIMIC-CXR-JPG	Man. curated	Cycle-cons. (single extr.)	0.545 ± 0.017	0.561	0.531	0.469	0.439	0.081	0.00
MIMIC-CXR-JPG	Man. curated	Grounding-only aux.	0.527 ± 0.022	0.541	0.514	0.486	0.459	0.089	0.00
			MIMIC-CXR-JPG (external)						
MIMIC-CXR-JPG	Man. curated	R2Gen (ext.)	0.512 ± 0.023	0.523	0.501	0.499	0.477	–	0.00
MIMIC-CXR-JPG	Man. curated	R2GenGPT (ext.)	0.526 ± 0.014	0.561	0.495	0.505	0.439	–	0.00
MIMIC-CXR-JPG	Man. curated	KARGEN (ext.)	0.534 ± 0.017	0.552	0.517	0.483	0.448	–	0.00
MIMIC-CXR-JPG	Man. curated	LLaVA-Med (ext.)	0.517 ± 0.026	0.505	0.529	0.471	0.495	–	0.00
MIMIC-CXR-JPG	Man. curated	Continual tuning (ext.)	0.546 ± 0.012	0.574	0.520	0.480	0.426	–	0.00
			CT-RATE (internal)						
CT-RATE	Struct. labels	TA-DECT	0.471 ± 0.013	0.485	0.458	0.542	0.515	0.088	0.00
CT-RATE	Struct. labels	Direct-to-text (same bb.)	0.440 ± 0.011	0.454	0.426	0.574	0.546	0.096	0.00
			CT-RATE (external)						
CT-RATE	Struct. labels	CT2Rep (ext.)	0.461 ± 0.014	0.482	0.443	0.557	0.512	–	0.00
CT-RATE	Struct. labels	Continual tuning (ext.)	0.468 ± 0.012	0.490	0.450	0.550	0.499	–	0.00
			Secondary report-derived diagnostic rows						
CheXpert	Rep.-derived	TA-DECT	0.676 ± 0.015	0.692	0.661	0.339	0.308	0.056	0.22
CheXpert	Rep.-derived	Direct-to-text (same bb.)	0.653 ± 0.017	0.671	0.636	0.364	0.329	0.071	0.22
MIMIC	Rep.-derived	TA-DECT	0.691 ± 0.013	0.709	0.674	0.326	0.291	0.055	0.15
MIMIC	Rep.-derived	Continual tuning (ext.)	0.686 ± 0.014	0.701	0.671	0.329	0.299	–	0.15

The central question is whether unsupported-assertion reductions can be obtained without increasing omissions. Across matched-path internal rows, TA-DECT improves F1 while reducing both missed-finding and unsupported-assertion rates relative to direct-to-text and other internal controls. For the primary matched-backbone comparison against direct-to-text, the 95% bootstrap CIs for F1 are non-overlapping on all three primary anchors (CheX-hold., MIMIC-cur-hold., and CT-RATE), indicating that the observed F1 gains exceed the estimated run-and-sample uncertainty under the shared protocol. Missed-finding and

unsupported-assertion rates remain part of the same operating-point description rather than separately tuned hypothesis-test endpoints. The CheX-hold. anchor has a non-zero excluded fraction under the fixed uncertain/blank policy, whereas MIMIC-cur-hold. and CT-RATE are binary anchors in this protocol and therefore have Excl. = 0.00. As noted above, these internal rows provide the strongest evidence for the proposed objective because backbone family, split files, decoding policy, and intervention settings are matched.

4.5 Sample Size, Raw Counts, and Seed Variation

Table 4 provides TP/FP/FN counts from a fixed reference checkpoint (seed 1) on the held-out primary anchors. Seed variation is reported as across-seed SD (five runs). Confidence intervals in Table 3 use two-stage bootstrap resampling.

Table 4: Raw micro-counts and seed variation on primary anchors. TP/FP/FN are from a reference checkpoint (seed 1); Seed SD is computed across five independent seeds under the same protocol. N_{pos} = TP + FN and N_{pred+} = TP + FP. Anchor composition (study/image counts, evaluable slot totals, and rater protocol) is reported in Table 5.

Anchor	Method	TP	FP	FN	N_{pos}	N_{pred+}	Seed SD (F1)
CheX-hold.	TA-DECT	8901	5891	7179	16,080	14,792	0.0079
CheX-hold.	Direct-to-text (same bb.)	8347	7113	7733	16,080	15,460	0.0118
MIMIC-cur-hold.	TA-DECT	10,983	8015	9455	20,438	18,998	0.0062
MIMIC-cur-hold.	Direct-to-text (same bb.)	10,337	8705	10,101	20,438	19,042	0.0101
CT-RATE	TA-DECT	7002	7421	8288	15,290	14,423	0.0086
CT-RATE	Direct-to-text (same bb.)	6521	7850	8769	15,290	14,371	0.0094

Anchor composition and counting scope are summarized in Table 5.

Table 5: Primary-anchor composition and counting scope. Evaluable slots are the total slot-level labels entering TP/FP/FN computation after exclusion rules. Positives and negatives are ground-truth slot counts and are therefore method-invariant within an anchor. For CheX-hold., raw structured-label slots are 4800 × 14 = 67,200 and evaluable slots after uncertain/blank exclusion are 48384. CheX-hold. is a structured-label hold-out split and is not the radiologist benchmark split.

Anchor (Table Shorthand)	Studies	Images/ Volumes	Evaluable Slots	Positive Slots	Negative Slots	Rater Protocol
CheXpert structured-label hold-out (CheX-hold.)	4800	4800	48,384	16,080	32,304	Released CheXpert structured labels (single label vector per study)
MIMIC-CXR-JPG manually curated hold-out (MIMIC-cur-hold.)	3269	3269	45,766	20,438	25,328	`mimic-cxr-2.1.0-test-set-labeled.csv` (released benchmark labels)
CT-RATE structured-label hold-out (CT-RATE)	1530	1530	45,870	15,290	30,580	Released CT-RATE structured labels

Fig. 2 provides a complementary view of the omission–hallucination trade-off, plotting missed-finding rate vs. unsupported-assertion rate with uncertainty intervals.

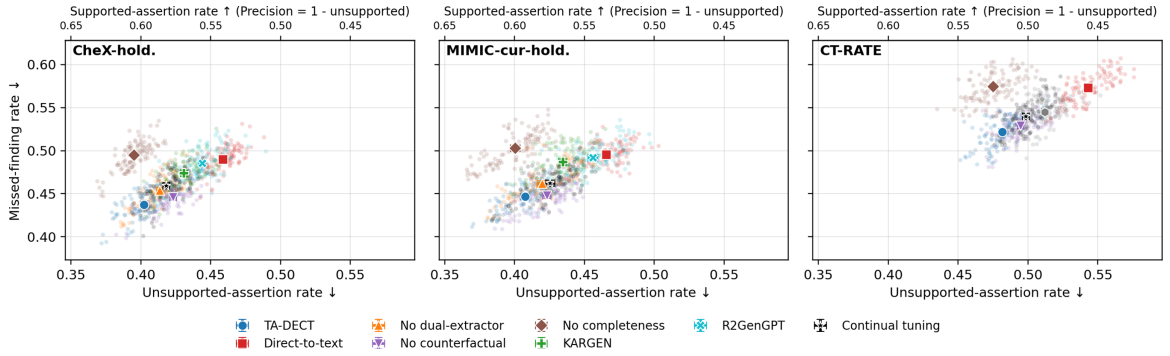

Figure 2: Supported findings vs. missed findings trade-off. Faint points denote per-(site, finding, run) replicates; solid markers denote means; error bars denote 95% CIs. The top axis reports supported-assertion rate (Precision = 1 – unsupported).

4.6 Revision Analyses

To keep the revision compact, the newly requested sensitivity and retraining analyses are summarized in one table. The first block re-scores existing checkpoints under alternate mapping and threshold rules without retraining. The second block reports the two targeted retraining variants requested by the reviewers: mixed-granularity counterfactual training and a higher-slice CT pathway.

These additions are intentionally narrow. They are included to test whether the main directional conclusions persist under reasonable protocol changes, not to introduce a new benchmark or a broader architecture search. Across the re-scoring settings, the default protocol remains a stable operating point with the expected miss-vs.-unsupported trade-off under threshold sweeps. Table 6 summarizes both the re-scoring sensitivity analyses and the targeted retraining variants.

Table 6: Compact revision analyses. The first block re-scores existing checkpoints without retraining. The second block reports the two targeted retraining variants added in this revision. Metrics are micro-averaged F1, missed-finding rate (Miss), unsupported-assertion rate (Unsup.), selectivity *S*, and expected calibration error (ECE) where applicable. In the CT rows, the "3 × 3 only" and "32 slices" entries share the same reference checkpoint and are therefore numerically identical.

Re-Scoring Sensitivity on Primary Anchors									
Protocol Variant	**CheX F1**	**CheX Miss**	**CheX Unsup.**	**MIMIC F1**	**MIMIC Miss**	**MIMIC Unsup.**	**CT F1**	**CT Miss**	**CT Unsup.**
Default protocol	0.577	0.446	0.398	0.557	0.463	0.422	0.471	0.542	0.515
Drop slots with missing region	0.565	0.458	0.410	0.549	0.471	0.430	0.465	0.550	0.520
Parser uncertain ↦ 0	0.566	0.468	0.395	0.555	0.469	0.419	0.469	0.548	0.510
Conflict precedence: pos > neg > uncertain	0.574	0.449	0.401	0.555	0.465	0.423	0.470	0.545	0.514
Global threshold 0.3	0.553	0.385	0.498	0.528	0.391	0.534	0.461	0.472	0.591

(Continued)

Table 6 (continued)

Re-Scoring Sensitivity on Primary Anchors									
Protocol Variant	**CheX F1**	**CheX Miss**	**CheX Unsup.**	**MIMIC F1**	**MIMIC Miss**	**MIMIC Unsup.**	**CT F1**	**CT Miss**	**CT Unsup.**
Global threshold 0.5	0.577	0.446	0.398	0.557	0.463	0.422	0.471	0.542	0.515
Global threshold 0.7	0.545	0.545	0.320	0.530	0.562	0.328	0.443	0.634	0.438
Class-wise validation thresholds	0.579	0.438	0.403	0.562	0.451	0.425	0.478	0.526	0.518

Targeted retraining analyses						
Setting	**Variant**	**F1**	**Miss**	**Unsup.**	**Sel. *S***	**ECE**
MIMIC-cur-hold. training grid	3 × 3 only	0.557	0.463	0.422	0.096	0.069
MIMIC-cur-hold. training grid	Mixed 3 × 3/7 × 7	0.558	0.458	0.426	0.118	0.072
CT-RATE training grid	3 × 3 only	0.471	0.542	0.515	0.087	0.088
CT-RATE training grid	Mixed 3 × 3/7 × 7	0.472	0.535	0.520	0.102	0.091
CT-RATE slices	32 slices	0.471	0.542	0.515	0.087	0.088
CT-RATE slices	64 slices	0.480	0.521	0.518	0.091	0.093

4.7 Counterfactual Selectivity and Evidence Linkage

Label-anchor metrics quantify correctness and completeness, whereas counterfactual tests quantify evidence sensitivity. Fig. 3 reports selectivity for held-out ROI occlusion (MS-CXR, VinDr-CXR) and grid occlusion (CT-RATE). These intervention results are intended as coarse evidence-sensitivity diagnostics rather than lesion-level causal localization tests. Because targeted sets are defined by fixed region/category mappings, improved selectivity should be read as a more appropriate confidence response under the released anchors, not as evidence that every fine-grained localization failure is removed.

To connect selectivity with spatial evidence behavior, patch-level gradient saliency maps are computed and localization is evaluated on held-out ROI data using pointing game accuracy (PGA) and intersection-over-union at threshold 0.1 (IoU@0.1). For each target finding, the corresponding slot logit is backpropagated to the visual patch tokens, the absolute gradient is taken, the result is averaged across the channel dimension, normalized to $[0,1]$, and upsampled to image resolution. PGA counts whether the peak response lies inside the reference ROI, and IoU@0.1 thresholds the normalized heatmap at 0.1 before measuring overlap with the held-out ROI mask. Table 7 reports localization and counterfactual outcomes jointly.

The later failure review makes this limit visible: anatomy mislocalization/region mismatch and uncertainty-collapse cases remain even when coarse selectivity and PGA/IoU improve.

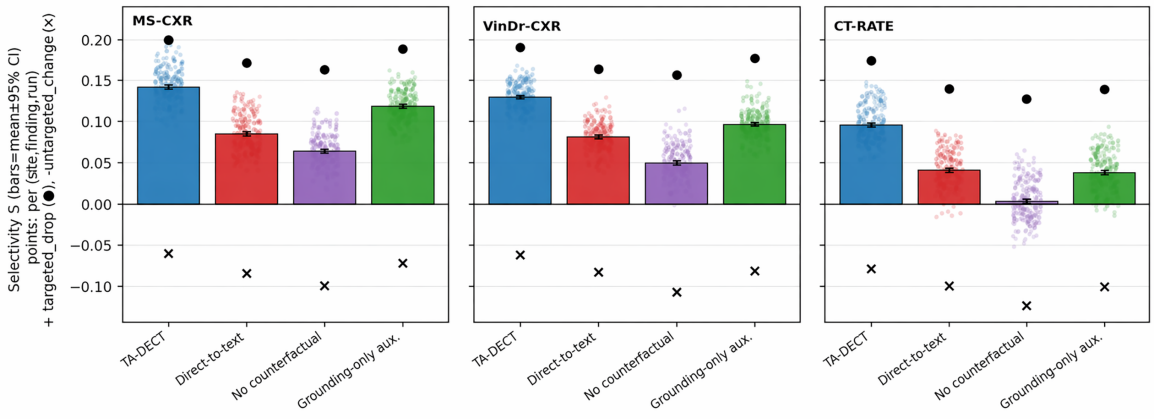

Figure 3: Counterfactual evidence sensitivity. Bars summarize selectivity aggregated at the (site, finding, run) level; black markers show mean targeted_drop (•) and −untargeted_change (×). This aggregation differs from sample-level summaries in some tables and can yield different absolute values while preserving direction.

Table 7: Evidence linkage and counterfactual results on held-out ROI splits. Localization score is computed from model-produced evidence maps with a shared post-processing rule. For ROI datasets, targeted sets are concept-and-region matched to ROI category mappings; for grid datasets, targeted sets are region-matched. Untgt. stab. is 1 − untargeted_change. Supp. rate is the fraction of positive anchor labels predicted positive (micro-averaged).

Dataset	Method	Loc. Metric	Loc. Score	Tgt. Drop	Untgt. stab.	Selectivity	Label Anchor	Supp. Rate
MS-CXR	TA-DECT	PGA↑	0.404	0.202	0.940	0.142	Box cats	0.591
MS-CXR	Direct-to-text	PGA↑	0.307	0.170	0.915	0.085	Box cats	0.533
MS-CXR	No counterfactual	PGA↑	0.381	0.163	0.901	0.064	Box cats	0.565
MS-CXR	Grounding-only aux.	PGA↑	0.425	0.189	0.929	0.118	Box cats	0.551
VinDr-CXR	TA-DECT	IoU@0.1↑	0.346	0.191	0.938	0.130	Img labels	0.561
VinDr-CXR	Direct-to-text	IoU@0.1↑	0.263	0.164	0.917	0.081	Img labels	0.514
VinDr-CXR	No counterfactual	IoU@0.1↑	0.322	0.157	0.893	0.050	Img labels	0.542
VinDr-CXR	Grounding-only aux.	IoU@0.1↑	0.369	0.177	0.919	0.096	Img labels	0.527

4.8 Objective Ablations

Table 8 isolates each objective component. Removing the counterfactual term consistently lowers selectivity, showing that targeted evidence sensitivity does not emerge reliably from label supervision and report likelihood alone. Removing completeness increases missed-finding rates, consistent with under-reporting behavior.

Weakening dual-extractor constraints (removing dual minimax, replacing minimax with single-extractor consistency, or removing anchor masking) increases extractor disagreement and reduces label-anchor quality. Together, these ablations are consistent with complementary roles across objective terms rather than a single dominant term.

Table 8: Objective ablations across datasets. Dual: dual-extractor minimax term. CF: counterfactual term. Comp: label-completeness term. Sel.: selectivity score under the intervention family used in each setting (grid occlusions for these ablations). Extr. dis.: disagreement rate between the two fixed extractors on generated reports (diagnostic only), computed as the fraction of finding slots for which $E_1(r) \neq E_2(r)$ on the generated report r. "w/o anchor mask" removes the label-consistency gate $m_f^{(k)}(x)$ in $\mathcal{L}_{\text{minimax}}$. Excl. is computed from ground-truth anchor labels within each setting.

Setting	Variant	Dual	CF	Comp	F1	Excl.	Miss	Unsup.	Sel.	Extr. dis.
MIMIC-cur-hold.	TA-DECT	Y	Y	Y	0.557	0.00	0.463	0.422	0.096	0.14
MIMIC-cur-hold.	- Dual-extr.	N	Y	Y	0.548	0.00	0.475	0.431	0.081	0.19
MIMIC-cur-hold.	- Counterfactual	Y	N	Y	0.553	0.00	0.468	0.425	0.029	0.16
MIMIC-cur-hold.	- Completeness	Y	Y	N	0.533	0.00	0.519	0.409	0.091	0.14
MIMIC-cur-hold.	Single-extr. (E1)	(E1)	Y	Y	0.545	0.00	0.483	0.436	0.067	0.23
MIMIC-cur-hold.	Minimax w/o anchor mask	Y	Y	Y	0.542	0.00	0.486	0.438	0.072	0.18
MIMIC-cur-hold.	Direct-to-text (no atomic cond.)	Y	Y	Y	0.524	0.00	0.494	0.457	0.044	0.20
CheX-hold.	TA-DECT	Y	Y	Y	0.577	0.28	0.446	0.398	0.112	0.11
CheX-hold.	- Dual-extr.	N	Y	Y	0.566	0.28	0.455	0.410	0.094	0.17
CheX-hold.	- Counterfactual	Y	N	Y	0.576	0.28	0.446	0.416	0.018	0.13
CheX-hold.	- Completeness	Y	Y	N	0.551	0.28	0.496	0.389	0.102	0.11
CheX-hold.	Single-extr. (E1)	(E1)	Y	Y	0.560	0.28	0.463	0.417	0.073	0.19
CheX-hold.	Minimax w/o anchor mask	Y	Y	Y	0.558	0.28	0.468	0.420	0.081	0.16
CheX-hold.	Direct-to-text (no atomic cond.)	Y	Y	Y	0.529	0.28	0.481	0.460	0.051	0.21
CT-RATE	TA-DECT	Y	Y	Y	0.471	0.00	0.542	0.515	0.087	0.16
CT-RATE	- Counterfactual	Y	N	Y	0.469	0.00	0.546	0.519	0.017	0.17
CT-RATE	- Completeness	Y	Y	N	0.448	0.00	0.586	0.501	0.079	0.16
CT-RATE	Direct-to-text (no atomic cond.)	Y	Y	Y	0.440	0.00	0.574	0.546	0.038	0.18

4.9 Robustness under Pre-Specified Shifts

The study evaluates a pre-specified shift suite: MIMIC→CheXpert, CheXpert→MIMIC, and two CXR/CT modality shifts via CT-RATE. Fig. 4 decomposes drift into label, evidence, and extractor components; Table 9 reports deltas relative to each method's in-domain baseline.

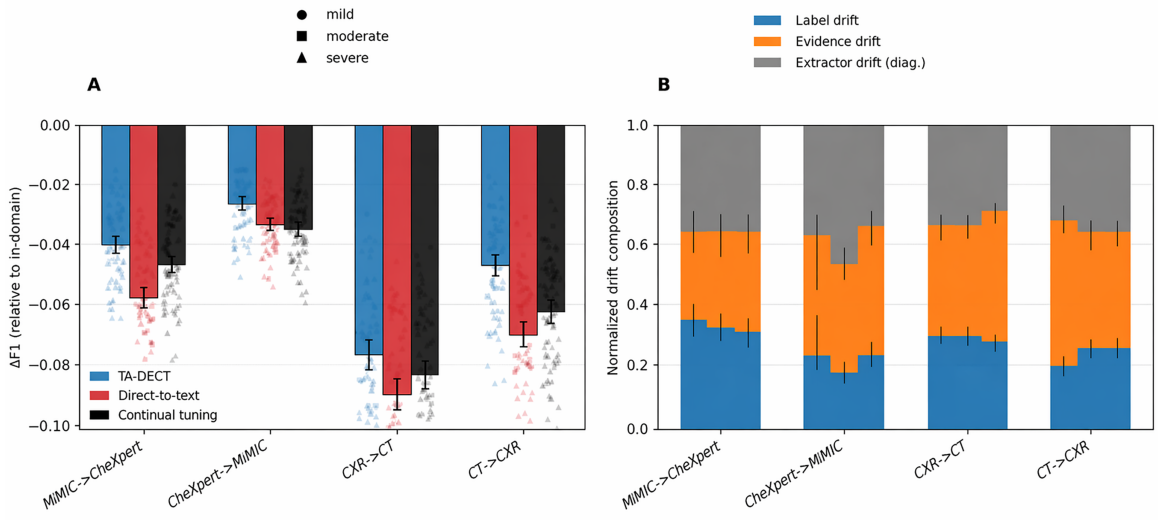

Figure 4: Pre-specified robustness and drift decomposition. (**A**) Shift-stratum robustness summary: circles/squares/triangles denote mild/moderate/severe shift strata from the fixed pre-registered severity multipliers (0.78/1.00/1.28) used in the stress suite generator; bars are means with 95% CIs. (**B**) Normalized drift composition with quantile spread markers.

Table 9: Pre-specified shift suite results. Deltas are computed against each method's in-domain test performance using identical decoding and the same deterministic intervention family. Drift src. is chosen by a fixed rule: the largest normalized magnitude among label drift (|ΔF1|), evidence drift (|ΔS|), and extractor drift (|ΔExtr.dis.|). Extractor drift is diagnostic only. Continual tuning follows Sun et al. [7].

Train	Test	Shift	Method	Anchor	ΔF1	ΔS	ΔExtr.dis.	Drift src.
MIMIC	CheX-hold.	Style	TA-DECT	Struct. hold-out	−0.037	−0.019	+0.031	Label
MIMIC	CheX-hold.	Style	Direct-to-text	Struct. hold-out	−0.061	−0.017	+0.047	Label
MIMIC	CheX-hold.	Style	Continual tuning	Struct. hold-out	−0.045	−0.028	+0.039	Label
CheXpert	MIMIC-cur-hold.	Style	TA-DECT	Man. curated hold-out	−0.031	−0.010	+0.034	Extractor
CheXpert	MIMIC-cur-hold.	Style	Direct-to-text	Man. curated hold-out	−0.029	−0.026	+0.061	Extractor
CheXpert	MIMIC-cur-hold.	Style	Continual tuning	Man. curated hold-out	−0.039	−0.021	+0.036	Label
MIMIC	CT-RATE	Modality	TA-DECT	Struct.	−0.074	−0.041	+0.050	Label
MIMIC	CT-RATE	Modality	Direct-to-text	Struct.	−0.082	−0.033	+0.069	Label
MIMIC	CT-RATE	Modality	Continual tuning	Struct.	−0.079	−0.055	+0.053	Label
CT-RATE	CheX-hold.	Modality	TA-DECT	Struct. hold-out	−0.052	−0.039	+0.058	Extractor
CT-RATE	CheX-hold.	Modality	Direct-to-text	Struct. hold-out	−0.066	−0.047	+0.062	Label
CT-RATE	CheX-hold.	Modality	Continual tuning	Struct. hold-out	−0.058	−0.031	+0.060	Extractor

All methods degrade under shift, but the decomposition clarifies where degradation concentrates. In several settings, TA-DECT shows smaller label-anchor drift than direct-to-text under identical decoding and intervention rules. Extractor drift remains diagnostic rather than primary evidence, but it helps separate text-parser instability from anchor-level performance change.

4.10 Metric Validity and Sanity Checks

Fig. 5 tests whether anchor metrics can be improved by degenerate policies such as forced short reports, uncertainty-heavy phrasing, and global confidence shrinkage ($p_f \leftarrow \kappa p_f$, fixed $\kappa < 1$).

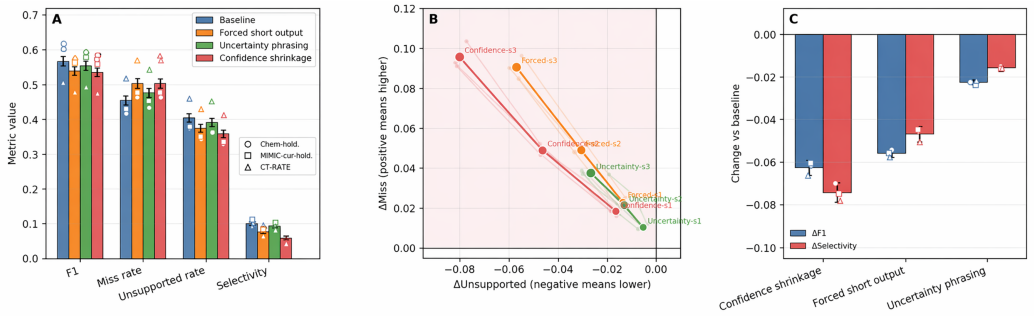

Figure 5: Metric validity under automatic gaming stress tests. (**A**) Compact cross-metric comparison across the gaming policies. (**B**) Direction-specific delta summary for unsupported-vs-miss changes. (**C**) Direction-specific delta summary for F1-vs-selectivity changes.

Across settings, these stress tests produce compensating failures: lower unsupported-assertion rates are typically accompanied by higher miss rates or weaker selectivity. This supports interpreting the main endpoint as a coupled profile (F1, miss, unsupported, selectivity) rather than any single scalar.

4.11 Case Studies

Fig. 6 provides representative qualitative examples that complement the aggregate metrics. Each panel summarizes supported/unsupported/missed findings together with confidence changes before and after ROI or grid occlusion under the same mapping rules used in the quantitative protocol.

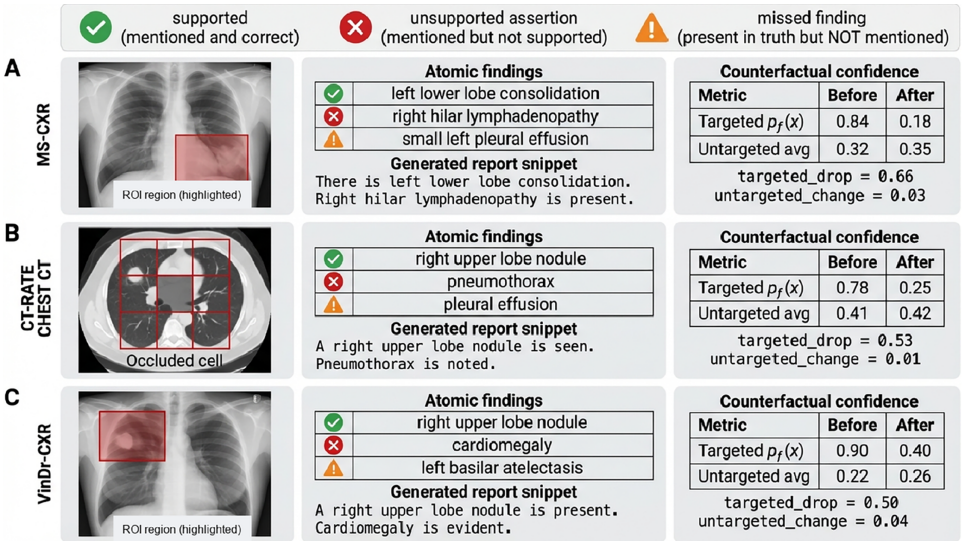

Figure 6: Qualitative grounded report cases with counterfactual interventions. (**A**) MS-CXR ROI-occlusion example. (**B**) VinDr-CXR ROI-occlusion example. (**C**) CT-RATE grid-occlusion example. Each panel summarizes supported/unsupported/missed findings together with confidence changes before and after the intervention; for visualization clarity, MS-CXR and VinDr-CXR highlight ROI intervention regions, whereas CT-RATE shows the occluded grid cell directly.

These examples are illustrative diagnostics rather than primary evidence. They are included to make mixed error behavior within a single study visually explicit while keeping the main claims anchored to structured-label and ROI-based results.

4.12 Systematic Failure Taxonomy

To make residual safety-relevant errors more explicit, and to expose failure modes that can be obscured by coarse region partitions or fixed uncertain-label exclusion, this revision reports a compact manual review of 20 errorful studies per primary anchor (CheX-hold., MIMIC-cur-hold., CT-RATE), where "errorful" means at least one false positive or false negative under the default threshold. Each sampled study receives one primary label: omission-dominant, unsupported-assertion-dominant, anatomy mislocalization/region mismatch, or uncertainty/wording collapse. Table 10 summarizes the resulting error categories within each reviewed anchor.

Table 10: Systematic failure taxonomy from manual review of 20 errorful studies per anchor. Percentages are within-anchor proportions over the reviewed sample.

Anchor	Omission-Dominant	Unsupported-Assertion-Dominant	Anatomy Mislocalization/ Region Mismatch	Uncertainty or Wording Collapse
CheX-hold.	25%	45%	15%	15%
MIMIC-cur-hold.	35%	30%	20%	15%
CT-RATE	50%	15%	30%	5%

5 Discussion and Limitations

Across the reported settings, a structured bottleneck combined with dual-extractor consistency and counterfactual/completeness constraints is associated with overall gains in correctness and completeness on the primary anchors, rather than a collapse into a single error-mode trade-off. Gains on selectivity and localization further suggest that the improvement is not confined to surface report style, while parser signals remain auxiliary and diagnostic rather than the main evidence for correctness.

The revision also clarifies the main scope limits. Deterministic tuple conversion remains a real design commitment, so alternate region mappings, uncertainty policies, or conflict-resolution rules can shift absolute numbers even when ranking is stable. The evidence-sensitivity protocol still relies on a coarse 9-region partition and limited held-out ROI subsets, saliency-based localization remains only a supportive faithfulness probe, and the CT pathway remains a deliberately simple volumetric approximation. Accordingly, the localization claim in this paper is limited to coarse anchor-aligned evidence sensitivity rather than fine-grained lesion-level causal verification. In addition, the fixed exclusion of uncertain and blank labels keeps the binary anchor protocol comparable across methods but can hide clinically relevant uncertainty behavior.

6 Conclusion

TA-DECT couples atomic structured prediction, dual-extractor report consistency, and counterfactual training in a unified objective. Across CXR and CT anchors, the reported results support a coupled correctness profile: higher truth-anchored F1 with simultaneous reductions in missed-finding and unsupported-assertion rates, together with improved coarse selectivity under evidence removal.

The study also clarifies supervision roles within open-data settings: structured labels and held-out ROI annotations support the primary correctness and evidence-linkage claims, whereas parser-derived report signals are most informative for consistency and drift diagnostics.

More broadly, the results suggest that omission control and evidence sensitivity are worth optimizing directly, rather than treating them as by-products of fluent text generation. A practical next step is to expand the atomic inventory, test finer-grained intervention schemes, and evaluate stronger CT backbones while preserving the shared mapping and anchor-accounting framework for cross-domain comparison.

Acknowledgement: The author thanks the creators and maintainers of the publicly released datasets and open-source tools used in this study.

Funding Statement: The author received no specific funding for this study.

Availability of Data and Materials: All datasets used in this study are publicly released from their official sources, including CheXpert, MIMIC-CXR, MIMIC-CXR-JPG, MS-CXR, VinDr-CXR, and CT-RATE. Access and usage follow each dataset's license and terms; for PhysioNet-hosted datasets, credentialing and data-use agreement requirements apply.

Ethics Approval: This study used only publicly released, de-identified datasets obtained under their respective access terms. No new patient recruitment, intervention, or identifiable data collection was conducted by the author; therefore, additional ethics approval and informed consent were not required for this study.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Johnson AE, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng CY, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data. 2019;6(1):317. doi:10.1038/s41597-019-0322-0.
2. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. Proc AAAI Conf Artif Intell. 2019;33(1):590–7.
3. Chen Z, Song Y, Chang TH, Wan X. Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020 Nov 16–20; Online. p. 1439–49.
4. Wang Z, Liu L, Wang L, Zhou L. R2GenGPT: radiology report generation with frozen LLMS. Meta-Radiology. 2023;1(3):100033.
5. Li Y, Wang Z, Liu Y, Wang L, Liu L, Zhou L. Kargen: knowledge-enhanced automated radiology report generation using large language models. In: Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2024; 2024 Oct 6–10; Marrakesh, Morocco. Cham, Switzerland: Springer; 2024. p. 382–92.
6. Li C, Wong C, Zhang S, Usuyama N, Liu H, Yang J, et al. Llava-med: training a large language-and-vision assistant for biomedicine in one day. Adv Neural Inf Process Syst. 2023;36:28541–64.
7. Sun Y, Khor HG, Wang Y, Wang Z, Zhao H, Zhang Y, et al. Continually tuning a large language model for multi-domain radiology report generation. In: Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2024; 2024 Oct 6–10; Marrakesh, Morocco. Cham, Switzerland: Springer; 2024. p. 177–87.
8. Hamamci IE, Er S, Menze B. Ct2rep: automated radiology report generation for 3D medical imaging. In: Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2024; 2024 Oct 6–10; Marrakesh, Morocco. Cham, Switzerland: Springer; 2024. p. 476–86.
9. Li C, Liu X, Wang C, Liu Y, Yu W, Shao J, et al. GTP-4o: modality-prompted heterogeneous graph learning for omni-modal biomedical representation. In: European conference on computer vision. Berlin/Heidelberg, Germany: Springer; 2024. p. 168–87.
10. Wang C, Liu X, Li C, Liu Y, Yuan Y. PV-SSM: exploring pure visual state space model for high-dimensional medical data analysis. In: Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2024 Dec 3–6; Lisbon, Portugal. 2024. p. 2542–9.
11. Chambon P, Delbrouck JB, Sounack T, Huang SC, Chen Z, Varma M, et al. Chexpert plus: hundreds of thousands of aligned radiology texts, images and patients. arXiv:240519538. 2024.
12. Boecking B, Usuyama N, Bannur S, Castro DC, Schwaighofer A, Hyland S, et al. Making the most of text semantics to improve biomedical vision–language processing. In: European conference on computer vision. Berlin/Heidelberg, Germany: Springer; 2022. p. 1–21.
13. Nguyen HQ, Lam K, Le LT, Pham HH, Tran DQ, Nguyen DB, et al. VinDr-CXR: an open dataset of chest X-rays with radiologist's annotations. Sci Data. 2022;9(1):429.
14. Hamamci IE, Er S, Wang C, Almas F, Simsek AG, Esirgun SN, et al. Generalist foundation models from a multimodal dataset for 3D computed tomography. Nat Biomed Eng. 2026:1–19. doi:10.1038/s41551-025-01599-y.
15. Peng Y, Wang X, Lu L, Bagheri M, Summers R, Lu Z. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. AMIA Summits Transl Sci Proc. 2018;2018:188.
16. Zhang S, Xu Y, Usuyama N, Xu H, Bagga J, Tinn R, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv:230300915. 2023.
17. Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, de las Casas D, et al. Mistral 7B. arXiv:2310.06825. 2023. doi:10.48550/arXiv.2310.06825.
18. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. Lora: low-rank adaptation of large language models. Iclr. 2022;1(2):3.