FALSIFIABLE CLAIM–REGION GROUNDING IN 3D CT

Anonymous authors
Paper under double-blind review

ABSTRACT

3D CT grounding usually evaluates where a supported finding is located, while verification evaluates whether a claim matches a scan. These separate tests miss a failure in which a real finding is displayed as support for a false claim. If a CT shows a finding in the left lower lobe but the claim places it in the left upper lobe, the finding is real yet does not support that claim. Positive-only grounding may still return its region, whereas a global verifier may reject the sentence without identifying the conflicting region. This paper introduces ReX-Verify, which keeps the CT and annotated source region fixed as location and finding edits change that region from support to annotation-relative counter evidence. The accompanying Claim-conditioned Evidence Grounding for Verification (CEG-Verify) model predicts separate support and counter fields, suppresses support for non-supported claims, and uses the predicted relation to select the returned field. Matched comparisons show that false-support supervision reduces exposed support after relation errors, mainly for cross-case unsupported claims, while also reducing returned maps for supported claims. Outcome decomposition attributes the reduction to map blocking after false admission. Exact-text and region analyses show weak image-conditioned discrimination and coarse counter localization.

1 INTRODUCTION

A chest CT claim names a finding and its location, so grounding requires both a support decision and a region with the right evidential role. If the location or finding in the sentence changes while the scan stays fixed, the original region may cease to support it. Its mask can still overlap a real finding while assigning that finding the wrong role for the new claim. A global verifier may reject the mismatch without locating it, while positive grounding may still display the original region as support. Relation accuracy and positive-target overlap miss this false spatial support.

Scan–text encoders align CT volumes with reports (Hamamci et al., 2026; Blankemeier et al., 2026; Zhao et al., 2026). Region-linked datasets and models localize positively described findings (Lin et al., 2024; Zhang et al., 2025; Baharoon et al., 2026). These methods test compatibility or positive correspondence, leaving the region's role after a claim edit unexamined. Rejection-oriented grounding represents absent referents with empty outputs (Liu et al., 2023; Xia et al., 2024), and visual entailment assigns a global relation (Vu et al., 2018; Yanaka et al., 2023). Neither output tracks the changed role of a fixed region under a paired claim edit.

This paper makes that role change testable by holding a CT and annotated region fixed while editing the claim. In Figure 1, a nodular-consolidation region in the superior segment of the left lower lobe supports its source sentence. Replacing the location with *left upper lobe* invalidates that region as support. A grounding head that still exposes it as support makes a measurable false-support error. For this edited claim, the source annotation instead supplies a counter target. The judgment concerns this region even if another finding elsewhere in the CT matches the new wording.

ReX-Verify combines supported claims, controlled location and finding edits, and cross-case unsupported claims. A four-way relation distinguishes these cases. The proposed model conditions separate support and counter fields on the claim, penalizes support on non-supported pairs, and lets the predicted relation control which field can be returned, as detailed in Figure 2. ReX-Verify then distinguishes a correct rejection from a false admission whose support map is blocked or exposed. These choices lead to three contributions:

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

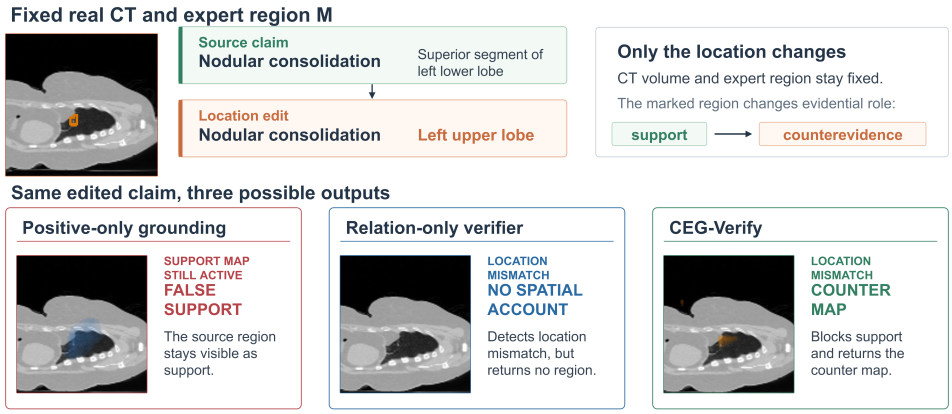


Figure 1: Fixed CT and region after a location edit. Positive-only grounding exposes false support; CEG-Verify blocks it and returns counter evidence. Map panels show real outputs; the verifier is schematic.

- We define fixed-region edits that change a region's evidential role. False-support exposure detects errors missed by relation accuracy and positive overlap.
- We couple a four-way relation with separate support and counter fields. Paired ablation shows that direct map suppression reduces false exposure from .130 to .032.
- We separate relation rejection, map blocking, and localization. Frozen-prediction analysis shows that blocked false admissions account for the reduction in exposed support.

2 Related Work

Volumetric alignment and positive-target localization. CLIP-style contrastive learning underlies many scan–text representations (Radford et al., 2021). CT-RATE, Merlin, and HLIP extend this approach to volumetric retrieval and abnormality recognition (Hamamci et al., 2026; Blankemeier et al., 2026; Zhao et al., 2026); M3D and VividMed support volumetric instructions and grounded outputs (Bai et al., 2024; Luo et al., 2025). More localized correspondence comes from generated medical terms (Chen et al., 2025), organ-level pairs, structured reports, region-linked text, or expert finding masks (Lin et al., 2024; Ichinose et al., 2023; Zhang et al., 2025; Baharoon et al., 2026). Weakly supervised localization and proposal verification further improve maps or filter lesions before refinement (Huy et al., 2025; Jiang et al., 2026). These evaluations localize positive descriptions, leaving the changed role of a fixed region untested.

Answerability and grounding hallucination. A missing referent can be represented by a pseudo region (Kim et al., 2022), an empty target (Liu et al., 2023), or segmentation and rejection tokens (Xia et al., 2024). False-premise correction can precede segmentation (Wu et al., 2024); GroundingME tests rejection of ungroundable expressions in general-domain images (Li et al., 2026a). Counterfactual Segmentation Reasoning edits the image and tests abstention (Li et al., 2026b), while counterfactual referring comprehension pairs localization with labels for edited attributes (Yu & Li, 2024). Editing the claim while holding both the image and observed region fixed instead tests whether that region changes from support to counter evidence.

Cross-modal verification and localized evidence. Controlled text substitutions have been used for image–caption mismatch classification and correction (Shekhar et al., 2017). Visual entailment predicts a global image-conditioned relation (Vu et al., 2018), including for clinical text and images (Yanaka et al., 2023); M2-Verify broadens consistency evaluation to multimodal scientific evidence and includes medical anatomical shifts (Ansari et al., 2026). Recent hallucination detectors recover localized evidence after generation by comparing factual and perturbed entities (Song et al., 2026) or by post-hoc bidirectional verification (Zhou et al., 2026). The fixed-region test connects a claim's relation to the spatial role of an observed region.

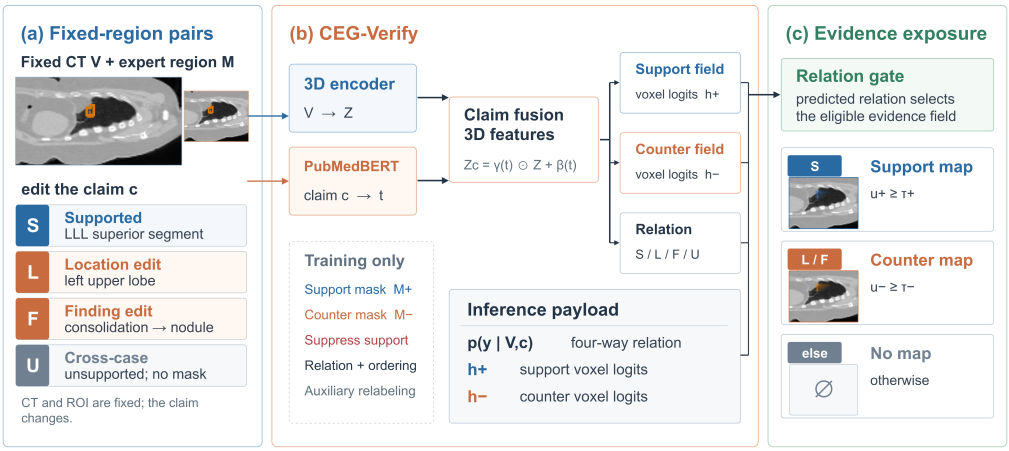


Figure 2: Fixed-region relations and claim-conditioned evidence. **(a)** A source claim, two structured edits with counter targets, and a cross-case claim share one output interface. **(b)** The CT and claim produce a four-way relation and separate support and counter fields under spatial and suppression supervision. **(c)** The predicted relation and validation-selected activation thresholds return one eligible field or no map.

3 Method

3.1 Problem Formulation

Let V be a chest CT on voxel domain Ω and c a finding claim. Their relation $y \in \mathcal{Y} = \{S, L, F, U\}$ denotes support, location mismatch, finding mismatch, or cross-case unsupported. Let $\mathcal{M} = \{L, F\}$ denote the two structured mismatches. A source claim has an annotated region $M \subseteq \Omega$. In the losses below, M and its derived masks also denote binary voxel indicators. For $y = S$, its support target is $M^+ = M$. Editing the location or finding leaves V and M fixed and assigns $M^- = M$ as the counter target for $y \in \mathcal{M}$. Cross-case pairs have no spatial target.

The model predicts a relation $\hat{y}$ and two voxel maps, $\hat{M}^+$ and $\hat{M}^-$. The relation selects the role eligible for output:

$$e(y) = \begin{cases} (1,0), & y = S, \\ (0,1), & y \in \mathcal{M}, \\ (0,0), & y = U. \end{cases} \tag{1}$$

At inference, $e(\hat{y})$ selects support, counter, or no spatial output, followed by a map-activation test. The counter target is the source region under an edited claim. The source mask supplies training and evaluation targets and is hidden at inference.

3.2 Fixed-Region Claim Pairs

Each supported example consists of a CT, a report-derived finding claim, and its annotated source mask. A location edit changes the anatomical phrase in that claim; a finding edit changes its finding term. Both retain the CT and source mask. A cross-case claim comes from a different CT case and contributes only a relation label. Thus, the two within-case edits have source-region counter targets, whereas a cross-case claim has no localized counter target.

For a structured edit, M^- identifies the original finding even when another finding elsewhere in the CT may match the edited text. The relation label is therefore defined relative to the paired annotation. Cases are assigned to disjoint train, validation, and test partitions before their claim pairs enter evaluation. Relation and map decisions are evaluated on held-out CTs without giving the model the source mask.

3.3 Claim-Conditioned Support and Counter Fields

With C channels and feature-grid size (D', H', W'), the image encoder E_θ maps V to $Z \in \mathbb{R}^{C\times D'\times H'\times W'}$. PubMedBERT maps c to a d-dimensional mean-pooled vector $t = T_\phi(c) \in \mathbb{R}^d$ (Gu et al., 2021). Learned projections $\gamma, \beta : \mathbb{R}^d \to \mathbb{R}^C$ scale and shift the image channels, broadcast over the feature grid:

$$Z_c = \gamma(t) \odot Z + \beta(t).$$

Two spatial heads map Z_c to support and counter logits $h^+, h^- \in \mathbb{R}^{|\Omega|}$. The sigmoid σ gives voxel fields $\hat{M}^+ = \sigma(h^+)$ and $\hat{M}^- = \sigma(h^-)$. A relation head uses mean-pooled Z_c and t to produce probabilities $p = (p_S, p_L, p_F, p_U)$. A training-only auxiliary head uses the same features for logits q, with $r(S) = \text{none}$, $r(L) = \text{location}$, $r(F) = \text{type}$, and $r(U) = \text{unsupported}$.

For a voxel-logit field h and fraction $\rho \in (0, 1]$, let $K_\rho(h) \subseteq \Omega$ contain the largest ρ fraction of its voxel responses and write $\mu_\rho(h) = |K_\rho(h)|^{-1} \sum_{v\in K_\rho(h)} h_v$. The signed field score is

$$s^\pm = \mu_\rho(h^\pm), \qquad s = s^+ - s^-.$$

A second fraction $\rho_{\text{fs}} \in (0, 1]$ identifies high support responses on non-supported pairs.

3.4 Training Objective

For logit field h and binary mask $B \in \{0, 1\}^{|\Omega|}$, BCE is voxel-wise binary cross-entropy and BCE_w is its class-weighted form. With Dice smoothing $\epsilon > 0$, the mask loss is

$$\ell_{\text{mask}}(h, B) = \text{BCE}_w(h, B) + 1 - \frac{2\langle\sigma(h), B\rangle + \epsilon}{\|\sigma(h)\|_1 + \|B\|_1 + \epsilon}. \tag{2}$$

Averaging this loss on (h^+, M^+) for supported pairs gives $\mathcal{L}_+$, and on (h^-, M^-) for structured mismatches gives $\mathcal{L}_-$.

For non-supported pairs, the objective suppresses the full support field and its largest responses. Let 0 denote the all-zero mask:

$$\begin{aligned}\mathcal{L}_{\text{fs}}^{\text{dense}} &= \mathbb{E}_{y\neq S}[\text{BCE}(h^+, 0)],\\ \mathcal{L}_{\text{fs}}^{\text{top}} &= \mathbb{E}_{y\neq S}\left[\frac{1}{|K_{\rho_{\text{fs}}}(h^+)|}\sum_{v\in K_{\rho_{\text{fs}}}(h^+)} \log(1 + e^{h_v^+})\right].\end{aligned} \tag{3}$$

With multiclass cross-entropy CE, the relation and auxiliary losses are $\mathcal{L}_{\text{rel}} = \text{CE}(p, y)$ and $\mathcal{L}_{\text{err}} = \text{CE}(q, r(y))$. For margin $m > 0$, the signed score gives

$$\mathcal{L}_{\text{ord}} = \begin{cases}\max(0, m - s), & y = S,\\ \max(0, m + s), & y \neq S.\end{cases} \tag{4}$$

Cross-case pairs contribute to relation, ordering, and suppression losses but have no counter-mask target. With nonnegative weights $\lambda_r, \lambda_e, \lambda_+, \lambda_-, \lambda_o, \lambda_d, \lambda_t$, the objective is

$$\begin{aligned}\mathcal{L} = \lambda_r\mathcal{L}_{\text{rel}} + \lambda_e\mathcal{L}_{\text{err}} + \lambda_+\mathcal{L}_+ + \lambda_-\mathcal{L}_-\\ + \lambda_o\mathcal{L}_{\text{ord}} + \lambda_d\mathcal{L}_{\text{fs}}^{\text{dense}} + \lambda_t\mathcal{L}_{\text{fs}}^{\text{top}}.\end{aligned} \tag{5}$$

The mask terms localize the two field roles where targets exist; suppression trains the support field to remain inactive on all non-supported pairs. Dense suppression lowers responses across the support field, while the top-response term targets the voxels most likely to survive the activation threshold. Both act before the relation-controlled output rule. A mistaken support relation can therefore be blocked by an inactive support field, which is the mechanism measured by FSR.

3.5 Inference

Let τ_r be the support-decision threshold. The inference rule sets $\hat{y} = S$ when $p_S \geq \tau_r$ and otherwise chooses the most probable label in $\{L, F, U\}$. Let $u^\pm = \max_{v\in\Omega} \sigma(h_v^\pm)$ be peak map activations,

with corresponding thresholds $\tau_{\pm}$. The output is

$$\mathcal{O}(V, c) = \begin{cases} \hat{M}^{+}, & \hat{y} = S \text{ and } u^{+} \geq \tau_{+}, \\ \hat{M}^{-}, & \hat{y} \in \mathcal{M} \text{ and } u^{-} \geq \tau_{-}, \\ \varnothing, & \text{otherwise.} \end{cases} \tag{6}$$

The gate returns at most one field. Global validation-selected thresholds control relations, map activation, and overlap binarization, and remain frozen on test. Raw-map disagreement checks pre-gate activation against the role in Equation 1.

3.6 Relation Errors and Spatial Exposure

Let $A = \{\hat{y} = S\}$ be support admission, $H = \{u^{+} \geq \tau_{+}\}$ support-map activation, and $N = \{y \neq S\}$ a non-supported query. False admission is $\mathrm{FA} = P(A \mid N)$. For $\mathrm{FA} > 0$, the conditional probability rule gives

$$\mathrm{FSR} = P(A \cap H \mid N) = \mathrm{FA}\, P(H \mid A, N) \leq \mathrm{FA}. \tag{7}$$

When $\mathrm{FA} = 0$, $\mathrm{FSR} = 0$. On non-supported queries, rejection A^{c}, blocked admission $A \cap H^{c}$, and exposure $A \cap H$ form disjoint outcomes. Relation metrics measure admission; FSR measures admitted claims whose support map remains active. Supported recall $R_S = P(A \mid y = S)$ and supported-map exposure $C_S = P(A \cap H \mid y = S)$ measure support admission and map availability alongside FSR. This parallels risk–coverage evaluation (Geifman & El-Yaniv, 2019), with coverage defined by an exposed spatial map.

At similar false-admission rates, a lower FSR requires less map activation after admission. The support field also serves valid claims, so suppressing it may reduce C_S. Correct rejection, blocked admission, false exposure, and supported-map exposure distinguish these effects.

3.7 Validation Policy Selection

For a task-trained model and seed, let $\mathrm{F1}_{\mathrm{macro}}$ denote validation four-way relation macro-F1 and $\mathrm{IoU}_{+}, \mathrm{IoU}_{-}$ denote validation support and counter overlap. Candidate models and map thresholds must meet validation $R_S \geq 0.5$. Among eligible candidates, the validation criterion maximizes

$$J = \mathrm{F1}_{\mathrm{macro}} + \mathrm{IoU}_{+} + \mathrm{IoU}_{-} - 0.25\,\mathrm{FSR}. \tag{8}$$

After selecting the model and map thresholds, the relation threshold is set nearest validation $R_S = 0.5$. The selected policy is frozen before test evaluation. The same criterion applies to every task-trained model and ablation, with zero counter overlap for models without a counter field. Realized test recall is reported under this frozen policy.

4 Experiments

4.1 Experimental Setup

Data and comparison. ReX-Verify pairs ReXGroundingCT claims and masks with CT-RATE scans. The case-disjoint split has 13,292 training records, 718 validation records, and 1,591 test records: 529 supported, 208 location-mismatched, 325 finding-mismatched, and 529 cross-case unsupported. Task-trained comparators use the same split, training budget, and three random seeds. The language-conditioned U-Net predicts one support field, whereas the role-separated U-Net predicts support and counter fields. Single+gate uses the full model's backbone and verifier with one support head; independent dual heads retain both field types without the full coupling. The multi-scale single head provides a higher-capacity reference. All task-trained rows use the same relation labels and validation rule.

Selection and measures. Section 3.7 defines the validation-only selection criterion and target supported recall of 0.5. Reported measures are realized test recall, four-way macro-F1, support/counter IoU, and FSR, which counts non-supported queries admitted as supported with an active support map. Seed summaries use sample standard deviation; paired differences retain the common seed assignment.

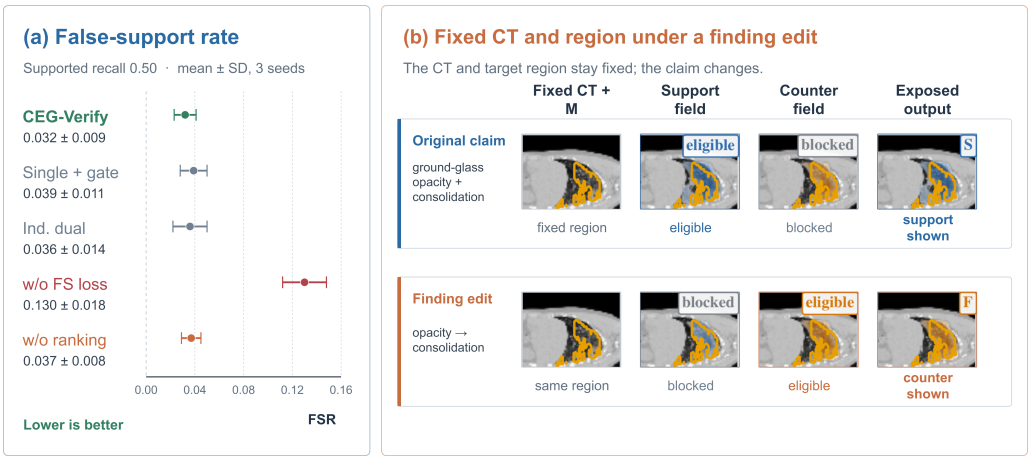


Figure 3: False-support exposure and one fixed-region finding edit. **(a)** Validation-frozen FSR at target supported recall .5, shown as mean ± standard deviation across three seeds. **(b)** For the same CT and annotated region, the source claim returns support and the edited claim returns counter evidence. Both raw fields are displayed; only the role-eligible field is exposed.

4.2 Validation-Targeted Recall Comparison

Under the shared validation-selected policy in Table 1, the full model has the lowest observed mean FSR among task-trained models. Single+gate has the highest relation F1 and stronger support overlap, while the multiscale single head has the highest support overlap. Independent dual heads have similar FSR and slightly higher counter overlap, with no observed localization gain from the coupling under this policy. The full-minus-Single+gate FSR difference is $-.0069 \pm .0173$ across paired seeds and favors the full model in two of three seeds. Its two-sided Student-t interval, $[-.0498, .0360]$, includes zero. The same-backbone comparison assesses field design, while the within-architecture ablation below isolates the suppression loss. Realized supported recall is .318 for the full model and .311 for Single+gate, below their common validation target. Because R_S counts support admission before map activation, the shortfall arises at the relation decision. Supported-map exposure C_S then measures the further loss of spatial output after admission. Reading both alongside FSR separates missed support from exposed false support. Figure 3 places the FSR comparison beside a fixed-region finding edit.

Table 1: Case-disjoint three-seed results at validation target supported recall .5. R_S is realized test recall, FSR false-support rate, and S/C-IoU support/counter overlap. Bold compares the rows above the divider. MS single is a capacity-scaled reference. Additional classification metrics are in Table 4.

Model	F1↑	R_S	FSR↓	S-IoU↑	C-IoU↑
Lang. U-Net	$0.581 \pm .004$	$0.344 \pm .136$	$0.170 \pm .055$	$0.010 \pm .005$	–
Role U-Net	$0.583 \pm .024$	$0.409 \pm .144$	$0.068 \pm .024$	$0.006 \pm .003$	$0.013 \pm .003$
Single+gate	$\mathbf{0.653 \pm .009}$	$0.311 \pm .017$	$0.039 \pm .011$	$\mathbf{0.012 \pm .001}$	–
Ind. dual	$0.636 \pm .016$	$0.303 \pm .038$	$0.036 \pm .014$	$0.009 \pm .001$	$\mathbf{0.020 \pm .001}$
CEG-Verify	$0.636 \pm .007$	$0.318 \pm .041$	$\mathbf{0.032 \pm .009}$	$0.010 \pm .001$	$0.018 \pm .002$
MS single†	$0.651 \pm .001$	$0.296 \pm .053$	$0.062 \pm .004$	$0.022 \pm .003$	–

The three-seed ablation in Table 2 isolates the training terms. Removing false-support supervision raises mean FSR from .032 to .130. The full-minus-no-FS-loss paired difference is $-.0976 \pm .0255$, with a two-sided Student-t interval of $[-.1609, -.0343]$. The ablation supports a within-architecture suppression effect; the full-versus-Single+gate interval does not establish a field-design advantage. Without false-support supervision, mean relation F1 rises slightly from .636 to .645 while FSR increases, so the exposure change is not explained by worse relation classification. Ranking removal has a smaller effect on FSR. Removing the auxiliary head leaves mean FSR unchanged but raises raw relation–map disagreement from .311 to .338.

Table 2: ReX-Verify component ablations under the same three-seed protocol and validation-selected thresholds. FS denotes false-support supervision.

Setting	F1	FSR	Dis.
Full	0.636±.007	0.032±.009	0.311±.046
w/o FS loss	0.645±.021	0.130±.018	0.347±.012
w/o ranking	0.635±.014	0.037±.008	0.304±.045
w/o auxiliary	0.630±.014	0.032±.004	0.338±.080

Non-supported outcomes in Table 3 separate correct rejection, blocked false admission, and exposed false support. Correct rejection is .777 for the full model and .775 without false-support supervision, while conditional map exposure after false admission changes from .579 to .147. The blocked-admission rate rises from .095 to .191. With little change in rejection, the shift from exposed to blocked admissions places the FSR reduction after the relation decision. False-support supervision also lowers supported-map exposure C_S from .340 to .124. Across all three paired seeds, both FSR and C_S decrease, indicating that suppression attenuates maps on supported as well as non-supported claims.

Table 3: Non-supported outcomes across three seeds, as mean (sample standard deviation). Block is false admission with an inactive support map; E/FA is conditional exposure, and C_S is support-map exposure on supported claims.

Model	Reject	Block	Expose↓	E/FA↓	C_S
Full	.777(.027)	.191(.035)	**.032(.009)**	**.147(.057)**	.124(.020)
w/o FS loss	.775(.034)	.095(.022)	.130(.018)	.579(.056)	.340(.018)
Single+gate	.797(.016)	.164(.019)	.039(.011)	.192(.052)	.152(.030)
MS single	.805(.044)	.134(.040)	.062(.004)	.323(.052)	.205(.015)

Supported recall measures admission by the relation decision; C_S additionally requires an active support map. The full model and its no-FS-loss ablation have similar relation F1 and rejection but differ in false and supported-map exposure. Single+gate and the multiscale single head return maps for more supported claims under their validation-selected policies, while the multiscale model also exposes more false support. The full model therefore trades supported-map availability for lower false exposure.

4.3 Relation and Map Agreement

Relation AUROC, false admission, and raw relation–map disagreement are compared in Table 4. Single+gate has the highest relation AUROC and the lowest false admission and disagreement among the same-scale models. The full model has lower FSR despite weaker relation AUROC and higher raw disagreement than Single+gate. Raw disagreement is measured before the relation gate, while FSR counts maps exposed after gating. This pattern locates the full model's lower false exposure in its output behavior rather than in a better relation decision. The auxiliary-head ablation in Table 2 reinforces this distinction: raw disagreement rises while FSR remains unchanged. Its effect is visible before gating but not in the returned-map error.

Table 4: Additional case-disjoint metrics from Table 1. FA is false admission; Dis. is raw relation–map disagreement. Three-seed means and standard deviations are shown.

Model	AUC↑	FA↓	Dis.↓
Lang. U-Net	0.723±.003	0.261±.106	0.420±.053
Role U-Net	0.726±.003	0.312±.109	0.414±.017
Single+gate	**0.766±.002**	**0.203±.016**	**0.289±.022**
Ind. dual	0.751±.010	0.212±.032	0.301±.050
CEG-Verify	0.747±.015	0.223±.027	0.311±.046
MS single†	0.770±.004	0.195±.044	0.294±.049

4.4 Exposure by Claim Source

Overall FSR averages over location edits, finding edits, and cross-case unsupported claims. The full model and its paired no-FS-loss ablation are compared within each source under the validation-frozen policies in Table 5. False-support supervision lowers exposure in all three groups. After weighting by test counts, cross-case claims account for about .095 of the overall .098 decrease. Location and finding edits contribute less because the full model already rejects .954 and .985 of them, respectively, as detailed in Table 6. Cross-case claims have the lowest correct-rejection rate and the highest blocked-admission rate in Table 6, accounting for most admissions on which map suppression can act. Structured edits directly test whether the source region changes evidential role, but their high rejection rates leave few exposed support maps to suppress. Cross-case claims have no counter target; they test whether unsupported claims elicit spatial support at all. The aggregate FSR change is therefore driven mainly by that latter failure mode.

Table 5: False-support rate by non-supported source under the same three paired seeds and validation-frozen policies. Δ is w/o FS loss minus full; entries are mean $\pm$ sample standard deviation.

Source	n	Full	w/o FS loss	Δ
Location edit	208	.005±.000	.016±.007	.011±.007
Finding edit	325	.003±.003	.006±.003	.003±.003
Cross-case	529	.060±.017	.250±.040	.190±.053

Table 6: Full-model non-supported outcomes by relation, as three-seed means. Reject is correct rejection; Block is false admission without an active support map.

Relation	Records	Reject	Block	Expose
Location edit (L)	208	.954	.042	.005
Finding edit (F)	325	.985	.012	.003
Cross-case (U)	529	.579	.360	.060

4.5 External Models and Shortcut Controls

Native CT-CLIP (Hamamci et al., 2026) and M3D-CLIP (Bai et al., 2024) scores give relation AUROCs of .541 and .528; VoxTell (Rokuss et al., 2026) returns one prompt-conditioned map with support IoU .020. Their native outputs cover different parts of ReX-Verify: the global score models supply no spatial field, while VoxTell supplies no typed counter field. Table 7 also shows higher support overlap for VoxTell than for the full model. The native score and map results characterize what each model returns; the task-trained rows in Table 1 provide the matched test of relation and exposure behavior. A phrase-matched structured-edit control compares text-only and frozen image-conditioned scores in Table 8.

Table 7: Native public-model outputs without ReX-Verify fine-tuning. Dashes indicate absent output types.

Model	Rel. AUC	S-IoU	C-IoU
CT-CLIP	0.541	–	–
M3D-CLIP	0.528	–	–
VoxTell	–	0.020	–
CEG-Verify	0.747	0.010	0.018

Table 8: Phrase-matched structured-edit pairs. Text is a TF-IDF control; Model is a frozen three-seed ensemble.

Edit	n	Text AUC	Model AUC
Location	186	0.730	0.826
Finding	153	0.939	0.964

Identical supported and cross-case claim strings remove replacement-word cues: the text input is the same, and only the CT differs. Across 470 strings and 885 scan pairs, the three seeded runs yield 0.561 ± 0.031 AUROC and 0.557 ± 0.026 within-text pair accuracy. Because each pair shares the claim, the modest above-chance separation reflects variation in the scan. This control removes lexical differences within each pair, whereas the structured-edit AUROCs can still reflect systematic

differences between source and replacement phrases. The relation test measures scan conditioning. Source-mask overlap separately evaluates where the returned field lies.

Finding edits retain stronger text cues than location edits. On phrase-matched pairs in Table 8, text-only and frozen-model AUROCs are both high for finding edits, whereas location edits show a larger difference between the two scores. Replacing a scan with another from the same edit category changes the target probability by .003 for finding edits and .054 for location edits, with relation flips of 2.7% and 14.4%. The small scan-swap response for finding edits is consistent with their strong text-only classification.

4.6 Counter-Field Localization

Counter-field localization is evaluated on the two structured edits, which retain a source mask; cross-case claims have no counter-mask target. Contact means that the thresholded counter map intersects the source mask, while IoU measures the overlap relative to the union. Location/finding contact is .370/.345, but IoU is .021/.015. The median structured counter target contains 128 voxels. In the fixed example of Figure 3, the returned field changes role after the claim edit, while the aggregate overlap results show that this behavior is not yet precise localization. The support-overlap comparison in Table 1 likewise favors the multiscale single head over the full model. The contact–IoU gap indicates that counter fields often touch the source region without delineating it.

5 Conclusion

A fixed CT region can change evidential role when a finding claim is edited, a change that positive localization and global verification do not jointly evaluate. This paper introduces ReX-Verify to test that role change and CEG-Verify to couple separate support and counter fields with relation-controlled output. In matched comparisons, false-support supervision reduces exposed support after incorrect admissions, chiefly for cross-case unsupported claims, while also returning fewer maps for supported claims. A single-head comparator remains stronger on relation classification and support overlap. Outcome decomposition places the measured effect in map suppression. Exact-text and region analyses show limited image-conditioned discrimination and counter localization. The fixed-region benchmark exposes this distinction between relation classification and spatial evidence. Balanced claim–CT pairs with independently assessed region relations are a next test for more precise image-conditioned evidence.

Reproducibility Statement

The Method defines the fixed-region labels, training objective, and output rule. Experiments specify the case-disjoint split, validation-only selection, common seeds, and paired uncertainty.

Ethics Statement

ReX-Verify reuses existing CT volumes, reports, and released region annotations under the source access terms. Controlled claim edits and cross-case unsupported labels derive from annotations, reports, and metadata; no new radiologist adjudication was collected. The study evaluates retrospective model outputs and does not assess clinical deployment.

AI Use Statement

Generative AI was used for English translation and polishing and for adjustments to manuscript formatting.

REFERENCES

Abolfazl Ansari, Delvin Ce Zhang, Zhuoyang Zou, Wenpeng Yin, and Dongwon Lee. M2-Verify: A large-scale multidomain benchmark for checking multimodal claim consistency, 2026. URL https://arxiv.org/abs/2604.01306.

Mohammed Baharoon, Luyang Luo, Michael Moritz, et al. ReXGroundingCT: A 3D chest CT dataset for segmentation of findings from free-text reports. *NEJM AI*, 3(7):AIdbp2501220, 2026. doi: 10.1056/AIdbp2501220. URL https://ai.nejm.org/doi/full/10.1056/AIdbp2501220.

Fan Bai, Yuxin Du, Tiejun Huang, Max Q.-H. Meng, and Bo Zhao. M3D: Advancing 3D medical image analysis with multi-modal large language models, 2024. URL https://arxiv.org/abs/2404.00578.

Louis Blankemeier, Ashwin Kumar, Joseph Paul Cohen, Jiaming Liu, et al. Merlin: A computed tomography vision-language foundation model and dataset. *Nature*, 652(8112):1318–1328, 2026. doi: 10.1038/s41586-026-10181-8.

Yanyuan Chen, Dexuan Xu, Yu Huang, Songkun Zhan, Hanpin Wang, Dongxue Chen, Xueping Wang, Meikang Qiu, and Hang Li. MIMO: A medical vision language model with visual referring multimodal input and pixel grounding multimodal output. In *Proceedings of CVPR*, pp. 24732–24741, 2025. doi: 10.1109/CVPR52734.2025.02303.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):2:1–2:23, 2021. doi: 10.1145/3458754.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, et al. Generalist foundation models from a multimodal dataset for 3D computed tomography. *Nature Biomedical Engineering*, 2026. doi: 10.1038/s41551-025-01599-y.

Ta Duc Huy, Duy Anh Huynh, Yutong Xie, Yuankai Qi, Qi Chen, Phi Le Nguyen, Sen Kim Tran, Son Lam Phung, Anton van den Hengel, Zhibin Liao, Minh-Son To, Johan W. Verjans, and Vu Minh Hieu Phan. Seeing the trees for the forest: Rethinking weakly-supervised medical visual grounding. In *Proceedings of ICCV*, pp. 24445–24455, 2025. doi: 10.1109/ICCV51701.2025.02266.

Akimichi Ichinose, Taro Hatsutani, Keigo Nakamura, Yoshiro Kitamura, Satoshi Iizuka, Edgar Simo-Serra, Shoji Kido, and Noriyuki Tomiyama. Visual grounding of whole radiology reports for 3D CT images. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14224, pp. 611–621, 2023. doi: 10.1007/978-3-031-43904-9_59.

Shuo Jiang, Yuhao Hong, Chunbo Jiang, Weihong Chen, Huangwei Chen, Shenghao Zhu, Beining Wu, Mingxuan Liu, Zhu Zhu, Feiwei Qin, Min Tan, and Yifei Chen. GLeVE: Graph-guided lesion grounding with proposal verification in 3D CT, 2026. URL https://arxiv.org/abs/2605.22619.

Yongmin Kim, Chenhui Chu, and Sadao Kurohashi. Flexible visual grounding. In *Proceedings of the ACL Student Research Workshop*, pp. 285–299, 2022. doi: 10.18653/v1/2022.acl-srw.22. URL https://aclanthology.org/2022.acl-srw.22/.

Rang Li, Lei Li, Shuhuai Ren, Hao Tian, Shuhao Gu, Shicheng Li, Zihao Yue, Yudong Wang, Wenhan Ma, Zhe Yang, Jingyuan Ma, Zhifang Sui, and Fuli Luo. GroundingME: Exposing the visual grounding gap in MLLMs through multi-dimensional evaluation. In *Proceedings of CVPR*, pp. 2412–2422, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_GroundingME_Exposing_the_Visual_Grounding_Gap_in_MLLMs_through_Multi-Dimensional_CVPR_2026_paper.html.

Xinzhuo Li, Adheesh Juvekar, Jiaxun Zhang, Xingyou Liu, Muntasir Wahed, Kiet A. Nguyen, Yifan Shen, Tianjiao Yu, and Ismini Lourentzou. Counterfactual segmentation reasoning: Diagnosing and mitigating pixel-grounding hallucination. In *Proceedings of CVPR Findings*, pp. 7450–7460, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026F/html/Li_Counterfactual_Segmentation_Reasoning_Diagnosing_and_Mitigating_Pixel-Grounding_Hallucination_CVPRF_2026_paper.html.

Jingyang Lin, Yingda Xia, Jianpeng Zhang, Ke Yan, Le Lu, Jiebo Luo, and Ling Zhang. CT-GLIP: 3D grounded language-image pretraining with CT scans and radiology reports for full-body scenarios, 2024. URL https://arxiv.org/abs/2404.15272.

Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In *Proceedings of CVPR*, pp. 23592–23601, 2023. doi: 10.1109/CVPR52729.2023.02259.

Lingxiao Luo, Bingda Tang, Xuanzhong Chen, Rong Han, and Ting Chen. VividMed: Vision language model with versatile visual grounding for medicine. In *Proceedings of NAACL*, pp. 1800–1821, 2025. doi: 10.18653/v1/2025.naacl-long.89. URL https://aclanthology.org/2025.naacl-long.89/.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Maximilian Rokuss, Moritz Langenberg, Yannick Kirchhoff, et al. VoxTell: Free-text promptable universal 3D medical image segmentation. In *Proceedings of CVPR*, pp. 37538–37557, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Rokuss_VoxTell_Free-Text_Promptable_Universal_3D_Medical_Image_Segmentation_CVPR_2026_paper.html.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. FOIL it! find one mismatch between image and language caption. In *Proceedings of ACL*, pp. 255–265, 2017. doi: 10.18653/v1/P17-1024. URL https://aclanthology.org/P17-1024/.

Xiao Song, Haonan Qin, Zhaoxu Zhang, Jiong Zhang, Yuqi Fang, and Caifeng Shan. Detecting clinical hallucinations in LVLMs via counterfactual visual grounding uncertainty, 2026. URL https://arxiv.org/abs/2606.28520.

Hoa Trong Vu, Claudio Greco, Aliia Erofeeva, Somayeh Jafaritazehjan, Guido Linders, Marc Tanti, Alberto Testoni, Raffaella Bernardi, and Albert Gatt. Grounded textual entailment. In *Proceedings of COLING*, pp. 2354–2368, 2018. URL https://aclanthology.org/C18-1199/.

Tsung-Han Wu, Giscard Biamby, David Chan, Lisa Dunlap, Ritwik Gupta, Xudong Wang, Joseph E. Gonzalez, and Trevor Darrell. See, say, and segment: Teaching LMMs to overcome false premises. In *Proceedings of CVPR*, pp. 13459–13469, 2024. doi: 10.1109/CVPR52733.2024.01278.

Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. GSVA: Generalized segmentation via multimodal large language models. In *Proceedings of CVPR*, pp. 3858–3869, 2024. doi: 10.1109/CVPR52733.2024.00370.

Hitomi Yanaka, Yuta Nakamura, Yuki Chida, and Tomoya Kurosawa. Medical visual textual entailment for numerical understanding of vision-and-language models. In *Proceedings of ClinicalNLP*, pp. 8–18, 2023. doi: 10.18653/v1/2023.clinicalnlp-1.2. URL https://aclanthology.org/2023.clinicalnlp-1.2/.

Zhihan Yu and Ruifan Li. Revisiting counterfactual problems in referring expression comprehension. In *Proceedings of CVPR*, pp. 13438–13448, 2024. doi: 10.1109/CVPR52733.2024.01276.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Jiayu Lei, Weiwei Tian, Ya Zhang, Weidi Xie, and Yanfeng Wang. Development of a large-scale grounded vision language dataset for chest CT analysis. *Scientific Data*, 12(1):1636, 2025. doi: 10.1038/s41597-025-05922-9.

Chenhui Zhao, Yiwei Lyu, Asadur Zaman Chowdury, Edward S. Harake, Akhil Kondepudi, Akshay T. Rao, Xinhai Hou, Honglak Lee, and Todd C. Hollon. Towards scalable language-image pre-training for 3d medical imaging. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=WxHf4EcBWA.

Nan Zhou, Ke Zou, Meng Liu, Linchao He, Jiaqi Zhu, Yi Zhang, Hu Chen, and Huazhu Fu. Hallucination detection and correction in medical VLMs via counter-evidence verification, 2026. URL https://arxiv.org/abs/2606.18609.