ProveTok: Proof-Carrying Budgeted Evidence Tokenization for Grounded 3D CT Report Generation

Anonymous Authors

February 7, 2026

Abstract

3D CT report generation fails most dangerously not when it is stylistically imperfect, but when it produces unsupported clinical assertions under strict inference budgets. Prior systems often treat grounding and citations as post-hoc explanations and refusal as an ad-hoc threshold, making it difficult to jointly audit latency, supportedness, and safety. We present ProveTok, a proof-carrying generation protocol that couples *budgeted evidence tokenization* (BET) and *proof-carrying generation* (PCG) under a unified budget constraint $B = B_{\text{enc}} + B_{\text{gen}}$. Throughout, "proof-carrying" refers to a verifier-checkable evidence chain (citations + trace), not a proof of clinical truth. BET selects evidence tokens where each token is bound to an explicit 3D support region Ω (e.g., a multi-resolution spatial cell) within B_{enc}, while PCG requires each generated clinical statement to carry machine-checkable citations and a verifier trace within B_{gen}; otherwise, the system triggers refinement or calibrated refusal. We implement claim-level proof rules and evaluate under multi-budget, multi-seed protocols using paired bootstrap with Holm correction. On a gold-mask ("real") profile, ProveTokpasses all primary claims across budgets while satisfying hard gates on latency, supportedness, and refusal calibration (Table 2). On cross-dataset stress tests with silver labels, pooled counterfactual perturbations yield statistically significant effects, supporting that citations are not decorative but measurably constrain grounded generation.

1 Introduction

Motivation. In clinical settings, the cost of a generation error is rarely "unnatural phrasing"—it is an unsupported claim that can trigger downstream mis-triage. This risk is amplified in 3D CT: volumes are information-dense but also highly redundant, and naïve tokenization schemes (e.g., fixed grids over all voxels) are computationally expensive. As a result, grounded report generation is inherently a *budget allocation* problem: under a strict inference budget, which spatial regions deserve encoding, which statements must cite evidence, and when should the system refuse due to insufficient support?

Why prior approaches are hard to audit. Existing 3D report generation work has strengthened datasets and baselines for the task [Hamamci et al., 2024b,a, 2023], but the dominant evaluation focus remains "text looks correct" rather than "each clinical statement is supported by localized evidence." Separately, pixel-level grounding datasets make sentence-to-mask evaluation possible [Baharoon et al., 2025], yet grounding is often evaluated post-hoc rather than enforced by the generation protocol. Finally, trustworthy generation and RAG research increasingly emphasizes attribution quality and learning to refuse [Song et al., 2024], but refusal can appear to improve safety by over-refusing unless calibrated under explicit miss-rate constraints. These pieces are rarely tied together under one matched-cost, end-to-end, scriptable audit.

Our reframing: proof-carrying generation under a budget. We propose ProveTok, which rewrites 3D CT report generation as a coupled problem of *budgeted evidence selection* and *proof-carrying generation*. We define a unified compute budget $B = B_{\text{enc}} + B_{\text{gen}}$ and an output contract consisting of *frames*, *citations*, *refusal*, and a *verifier trace*. Here, a *frame* is an atomic, typed clinical statement (e.g., a finding assertion) that is required to cite evidence tokens with explicit 3D support regions Ω. Groundedness is not an after-the-fact visualization: every generated frame must "carry its proof" (a verifier-checkable evidence chain: citations + trace), otherwise it is rejected by a verifier that triggers refinement or calibrated refusal.

Contributions. We emphasize protocol and auditability over yet another backbone. Our contributions are:

- **Proof-carrying protocol under matched budgets.** We introduce BET and PCG as a unified contract that makes evidence, refusal, and latency jointly auditable under $B = B_{\text{enc}} + B_{\text{gen}}$ (Figure 1).

Figure 1. ProveTok Closed-Loop Pipeline

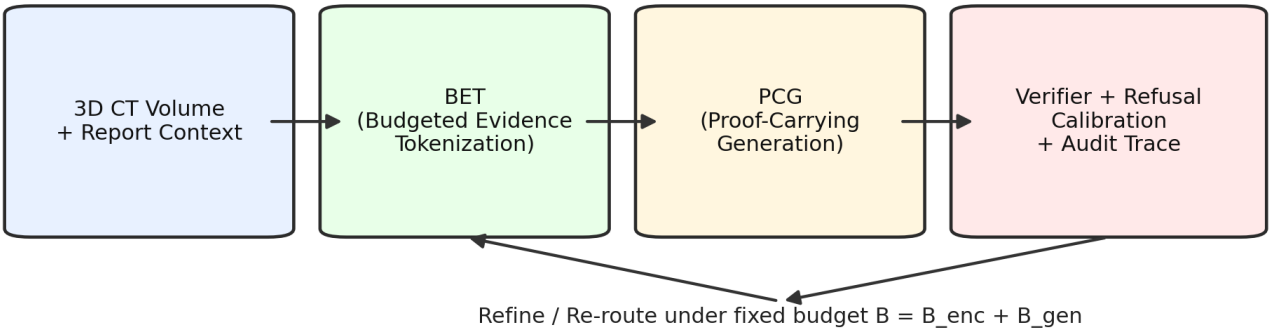

Figure 1: **ProveTokoverview.** Under a unified budget $B = B_{\text{enc}} + B_{\text{gen}}$, BET allocates encoder compute B_{enc} to produce evidence tokens with explicit 3D support regions Ω, and PCG allocates decoder compute B_{gen} to generate report statements with mandatory citations, verifier traces, and calibrated refusal.

- **Claim-level proof rules and statistics.** We implement scriptable claim-level tests with multi-budget, multi-seed evaluation using paired bootstrap and Holm correction (Table 2).
- **Decisive evidence via counterfactual and calibration.** We provide counterfactual stress tests showing non-trivial citation effects and calibrated refusal with explicit anti-silencing gates (Figures 5 and 7).

Paper roadmap. Section 3 defines the budgeted proof-carrying protocol; Section 4 describes matched-cost experiments and statistical tests; Section 5 presents main evidence and stress tests.

2 Related Work

3D CT report generation and multimodal CT resources. Recent efforts have significantly advanced data resources and baselines for 3D CT report generation and multimodal CT foundation modeling [Hamamci et al., 2024b,a, 2023]. These works primarily target generation quality, scalability, and task coverage. In contrast, ProveTokis a protocol and evaluation contribution: rather than introducing a new backbone, we focus on an auditable generation *contract* that couples compute budgets with evidence binding, verification outcomes, and refusal behavior under matched-cost comparisons.

Pixel-level grounding in 3D CT. Grounding datasets such as ReXGroundingCT explicitly connect free-text findings to pixel-level 3D segmentations, enabling sentence-level evaluation of evidence localization [Baharoon et al., 2025]. This evaluation substrate is essential for turning "explainability" into a measurable contract. Where prior work enables evaluation, ProveTokadditionally enforces grounding at generation time: citations are required, mapped to explicit 3D support regions, and checked by a verifier as part of the output protocol.

Trustworthy generation, attribution, and refusal. In trustworthy generation and RAG, attribution quality and learning to refuse are increasingly treated as part of the objective, not a post-processing step [Song et al., 2024]. However, refusal without explicit miss-rate and calibration constraints can reduce apparent error by over-refusing. ProveTokbuilds on this perspective but targets a different failure mode and setting: 3D CT evidence is spatial and budgeted, and refusal must be audited jointly with grounding, latency, and supportedness under matched budgets.

Positioning. Across these lines of work, the missing piece is a unified, scriptable contract that ties *budget*, *evidence binding*, *verification*, and *refusal calibration* into one end-to-end protocol. ProveTokaims to close that gap.

3 Method

3.1 Problem Setup

Let V be a 3D CT volume and let the model generate a report consisting of clinical statements. We define a unified inference budget

$$B = B_{\text{enc}} + B_{\text{gen}}, \tag{1}$$

where B_{enc} bounds the compute used for evidence tokenization/selection and B_{gen} bounds the compute used for proof-carrying generation and verifier interactions. The system output is a structured object containing (i) textual frames/statements, (ii) citations to evidence tokens, (iii) a refusal/uncertainty decision when evidence is insufficient, and (iv) a verifier trace that makes failures machine-auditable. Formally, BET produces evidence tokens $\{t_i\}$, where each token is associated with an explicit 3D support region Ω_i and metadata (e.g., resolution level and score). PCG generates frames $\{y_k\}$ and citations $C_k \subseteq \{t_i\}$, and a verifier $g(y_k, C_k)$ returns a pass/fail decision plus a failure type used for refinement or refusal.

Evidence cells and support regions. We represent spatial evidence with a multi-resolution dyadic grid. A *cell* $c = (\ell, i_x, i_y, i_z)$ at level ℓ partitions the volume into 2^ℓ bins per axis, yielding an octree-like refinement family. We use a deterministic mapping ϕ that maps a cell id to voxel indices, defining the support region $\Omega(c) = \phi(c; \text{shape}(V)) \subseteq [1..D] \times [1..H] \times [1..W]$. An evidence token is a tuple $t_i = (e_i, c_i, \ell_i, s_i, u_i)$ where e_i is an embedding pooled over $\Omega(c_i)$ from a 3D encoder feature map (mean pooling in our scaffold), $s_i \in [0, 1]$ is an evidence score, and u_i is an uncertainty proxy. This representation makes citations machine-checkable: a citation points to a token id, which deterministically implies a 3D support region.

3.2 BET: Budgeted Evidence Tokenization

BET constructs a set of evidence tokens $\{t_i\}$ under budget B_{enc}. Conceptually, BET answers: *which spatial regions are worth spending encoder compute on*, when each additional token consumes part of B_{enc}? In our implementation, encoding one active cell yields one token; budgets are expressed either directly as a token count (b_{enc}) or via FLOPs/latency-matched accounting (Section 4).

Full-cover initialization to avoid early misses. A greedy refinement policy that starts from a single root cell can "commit" too early and miss lesions entirely under strict budgets. We therefore initialize with a full-covering grid at an initial level ℓ_0 such that $|\mathcal{C}_0| = (2^{\ell_0})^3 \leq b_{\text{enc}}$, and only then refine within this grid until reaching the budget.

Refinement loop. At step s, given the current active cell set $\mathcal{C}_s$, we (i) encode tokens for all $c \in \mathcal{C}_s$, (ii) run PCG to produce frames and citations, (iii) run the verifier to obtain a typed issue list with evidence traces, and (iv) select one cell $c^\star \in \mathcal{C}_s$ to split into its eight children. The selection is driven by either:

- **Verifier-driven greedy.** If issues exist, prioritize cells referenced by the verifier evidence traces for failing frames (severity-weighted); otherwise split the cell with highest uncertainty u.
- **EvidenceHead ranking.** A small head predicts an expected issue reduction $\Delta(c)$ from the cell embedding and the current issue context, and we split the cell maximizing $\Delta(c)$ with a small ϵ-greedy exploration.

This makes allocation auditable: when refinement happens, the system can point to which verifier failure type triggered it and which cited cells were implicated.

3.3 PCG: Proof-Carrying Generation

PCG generates statements under budget B_{gen} with a hard contract: every asserted statement must cite evidence tokens. We use a structured output space for auditability: the report is emitted as a set of typed *frames* y_k (finding type + slots such as polarity and laterality) together with citations C_k and an accept probability q_k.

Frames and mandatory citations. PCG uses query-based attention from a small set of learnable finding queries to token embeddings, producing per-frame features and attention weights. The top-k attended tokens (or a score-interleaved variant) are recorded as citations C_k. Slot classifiers then read the frame features to generate the typed statement.

Support probability and refusal. PCG also outputs an accept probability $q_k \in [0, 1]$ intended to represent *evidence-backed support* for the frame (not necessarily clinical correctness). In our scaffold, q_k is predicted by a small MLP from the frame feature together with citation strength proxies (e.g., maximum cited score) and decoding uncertainty (e.g., slot entropy). Frames with insufficient support are either refined (by requesting additional evidence tokens or revising the statement) or refused under a calibrated threshold (next subsection).

3.4 Verifier Taxonomy

We implement a lightweight, rule-based verifier that categorizes failures into machine-checkable types (e.g., missing citation, insufficient spatial coverage, citation irrelevance) and emits a trace. The verifier is deterministic and fixed for a paper artifact (versioned), and it is not trained to agree with the generator. Therefore, "supportedness" in our protocol is defined with respect to this verifier, and should not be conflated with clinical truth.

Unsupportedness and overclaim. For unsupportedness (U1), key checks include: positive claims must have citations; cited evidence must exceed minimum score / maximum uncertainty thresholds; and citations must cover non-trivial spatial extent. We approximate cited coverage by the sum of dyadic volume fractions, $\sum_{t \in C_k} 8^{-\ell(t)}$, and flag insufficient coverage below a fixed ratio. For citation relevance, we optionally score tokens with a deterministic query attention and require that citations recover the top-k attended evidence with sufficient attention mass (Appendix). Overclaim (O1) captures specificity mismatches, e.g., a laterality-specific statement backed only by coarse evidence levels.

3.5 Refusal Calibration with Anti-Silencing Gates

Each frame y_k is assigned a support probability $q_k \in [0, 1]$ intended to estimate *evidence-backed support* (not necessarily clinical correctness). We calibrate a refusal threshold τ_{refuse} on development data under explicit safety constraints, then freeze it for test evaluation. Crucially, refusal is not allowed to "game" supportedness by staying silent: we enforce hard gates on (i) *critical miss-rate* (critical findings present in ground-truth but incorrectly refused), (ii) refusal calibration error (ECE/reliability), and (iii) refusal rate.

We use a fixed, audited critical finding set (e.g., pneumothorax, effusion, consolidation, nodule) for anti-silencing accounting. For ECE/reliability, we treat "correct" as "supported by the verifier" (i.e., no unsupported issues) on asserted positive frames, matching the semantics of q_k.

3.6 Claim-level Proof Rules

Rather than relying on cherry-picked qualitative examples, we define a bounded set of claims and implement scriptable pass/fail rules. Each claim points to a reproducible artifact and a statistical protocol (paired bootstrap and Holm correction). Table 2 provides the minimal decisive evidence set for oral-level defensibility.

4 Experiments

4.1 Datasets and Evaluation Profiles

We separate a gold-mask ("real") profile used for primary claims from cross-dataset silver-label stress tests. Table 1 summarizes the profiles.

4.2 Baselines and Matched-Budget Protocol

We compare tokenization strategies (e.g., fixed-grid and ROI-style variants) under a matched-cost protocol. All comparisons are performed across a fixed set of budgets (2×10^6 to 7×10^6 in six steps) with either FLOPs-matched or latency-aware matching depending on the claim. We report quality and grounding together with latency and supportedness, and we enforce hard gates where appropriate (e.g., latency p95 and supportedness deltas). To prevent post-hoc threshold tuning, calibration thresholds (e.g., τ_{refuse}) are selected once on development data and frozen for test evaluation.

4.3 Metrics

Our evaluation emphasizes what is auditable:

- **Grounding.** When gold masks are available (real profile), we report sentence-level grounding metrics (e.g., IoU) that connect citations to spatial support.
- **Supportedness.** We report an *unsupported rate* defined as the fraction of frames flagged by the verifier as unsupported (issue U1).
- **Clinical correctness proxies.** We report frame-level correctness on structured findings (frame F1). For safety-critical findings, we additionally report *critical-present recall* averaged over studies with at least one ground-truth critical present frame (the critical set is fixed and audited).
- **Refusal.** We report refusal rate and refusal calibration (ECE/reliability), and we enforce an anti-silencing constraint via *critical miss-rate* (ground-truth critical findings that are refused).

4.4 Statistical Protocol

We follow paired bootstrap testing (typically n_{boot}=20,000) with a one-sided alternative when comparing a method against a baseline, and we apply Holm correction over the relevant family (e.g., multiple budgets or counterfactual variants) to control family-wise error. We report confidence intervals and explicitly track the number of random seeds used for each artifact.

Implementation details and artifact paths are provided in Table 2 and Appendix tables.

Concretely, we use:

- **Budget families.** For multi-budget claims (e.g., C0001/C0004), we treat budgets as one family and apply Holm correction over the budget sweep.
- **Counterfactual families.** For counterfactual non-triviality (e.g., C0003 and V0003), we treat perturbation variants (e.g., no_cite, cite_swap, ω-perm) as one family and apply Holm correction accordingly.

5 Results

5.1 Main Evidence: Minimal Decisive Set

We present the minimal decisive evidence set in Table 2. Each row corresponds to a script-audited claim with a fixed protocol and a reproducible artifact path, enabling end-to-end verification without manual cherry-picking. For readability, we summarize the intent of each claim here: C0001 tests multi-budget Pareto tradeoffs under matched cost; C0002 tests whether allocation is predictable via regret; C0003 uses counterfactuals to test non-triviality of the citation channel; C0004 tests pixel-level grounding improvements where gold masks exist; C0005 tests calibrated refusal with anti-silencing gates; C0006 tests baseline coverage and audited cost accounting; V0003 probes cross-dataset stress tests with pooled ω-permutation.

5.2 Multi-budget Tradeoffs

Figure 2 shows multi-budget curves under matched-cost evaluation, highlighting quality/grounding improvements while satisfying latency and supportedness constraints.

5.3 Clinical Correctness Proxy: Critical-Present Recall

Supportedness is a necessary but not sufficient condition for clinical correctness. As a minimal correctness proxy, Figure 3 reports critical-present recall over a fixed, audited set of safety-critical findings, evaluated under the same matched-budget protocol.

5.4 Allocation is Non-trivial: Regret

Figure 4 quantifies allocation regret, supporting that budget allocation can be modeled and evaluated beyond post-hoc configuration selection.

5.5 Counterfactual Non-triviality (Omega Permutation)

We stress-test the citation channel using counterfactual perturbations. Figure 5 summarizes pooled significance for ω-permutation, supporting that perturbing evidence support degrades grounding under a statistically guarded protocol.

Seed-level stability. Figure 6 shows the per-seed effect sizes and confidence intervals for ω-permutation. The direction is consistent across seeds (19/20 positive), but single-seed tests can be underpowered; this motivates pooled, family-wise controlled protocols rather than relying on a small number of seeds.

5.6 Refusal Calibration (Anti-silencing)

Figure 7 reports calibrated refusal with hard gates on critical miss-rate, ECE, and refusal rate, and demonstrates supportedness improvements without "silence for safety".

5.7 Qualitative Case Studies

Figure 8 shows representative cases with evidence tokens, citations, and verifier outcomes.

6 Discussion and Limitations

Why these results are defensible. Our evidence is organized around auditable, claim-level proof rules rather than narrative examples. The protocol enforces matched-cost comparisons across multiple budgets and seeds, and it uses paired bootstrap testing with Holm correction to guard against cherry-picking across budgets or counterfactual families. Importantly, refusal is evaluated with explicit anti-silencing gates (critical miss-rate and calibration constraints), preventing apparent safety gains from over-refusal.

Limitations. Primary grounding claims rely on gold masks available in the real profile; cross-dataset stress tests use silver labels (automatic or pseudo masks) and should not be interpreted as replacing gold-mask evidence. Our "proof-carrying" contract provides verifier-checkable evidence binding (citations + trace), but it does not prove clinical truth: supportedness is a necessary condition, not a sufficient one. Our counterfactual perturbations are designed to test the non-triviality of the citation channel; they do not fully capture clinical correctness beyond the audited (and proxy) metrics. Finally, while our claim-level suite aims to be minimal and decisive, extending it to broader clinical settings will require additional datasets and stronger public baselines.

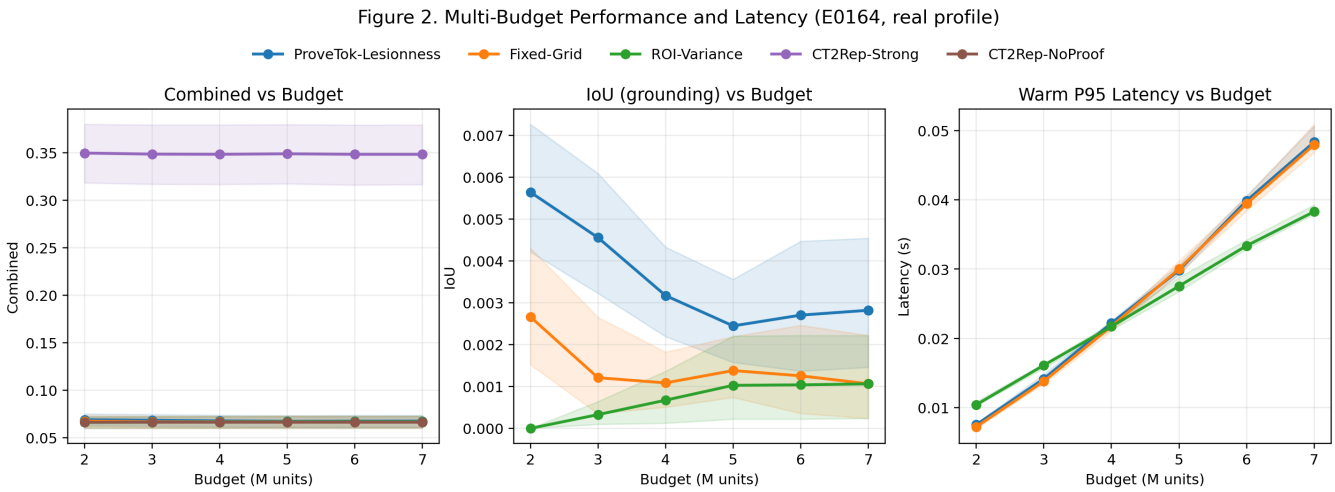

Figure 2: Multi-budget tradeoffs under matched-cost evaluation (data: `outputs/E0164-full/baselines_curve_multiseed.json`)

Future work. We plan to (i) expand cross-domain evaluation with gold-mask subsets, (ii) strengthen public baseline coverage under matched latency constraints, and (iii) integrate asset regeneration and proof gates into continuous integration to keep artifacts synchronized with code.

Two-sentence oral summary. Under a fixed inference budget, ProveTokturns grounded generation into a contract: every clinical statement must carry machine-checkable spatial evidence and a verifier trace, otherwise it is refined or refused under calibrated, anti-silencing gates. Across multi-budget, multi-seed evaluations with paired bootstrap and Holm correction, we obtain decisive, reproducible evidence that citations and refusal constraints measurably shape grounded 3D CT report generation (Table 2).

7 Reproducibility

All paper figures and tables are generated from repository artifacts. To regenerate the paper assets used in this draft, run:

```
python scripts/paper/build_readme_figures.py --out-dir docs/paper_assets/figures
python scripts/paper/build_readme_tables.py --out-dir docs/paper_assets/tables
```

To run the claim-level proof gate:

```
python scripts/proof_check.py --profile real
python scripts/oral_audit.py --sync --out outputs/oral_audit.json --strict
```

Tables in this paper include artifact paths (Appendix) to enable direct verification of protocols and statistics.

References

Mohammed Baharoon, Luyang Luo, Michael Moritz, Abhinav Kumar, Sung Eun Kim, Xiaoman Zhang, Miao Zhu, Mahmoud Hussain Alabbad, Maha Sbayel Alhazmi, Neel P. Mistry, Lucas Bijnens, Kent Ryan Kleinschmidt, Brady Chrisler, Sathvik Suryadevara, Sri Sai Dinesh Jaliparthi, Noah Michael Prudlo, Mark David Marino, Jeremy Palacio, Rithvik Akula, Di Zhou, Hong-Yu Zhou, Ibrahim Ethem Hamamci, Scott J. Adams, Hassan Rayhan AlOmaish, and Pranav Rajpurkar. Rexgroundingct: A 3d chest ct dataset for segmentation of findings from free-text reports. *arXiv preprint arXiv:2507.22030*, 2025.

Ibrahim Ethem Hamamci, Sezgin Er, Anjany Sekuboyina, Enis Simsar, Alperen Tezcan, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Furkan Almas, Irem Dogan, Muhammed Furkan Dasdelen, Chinmay Prabhakar, Hadrien Reynaud, Sarthak Pati, Christian Bluethgen, Mehmet Kemal Ozdemir, and Bjoern Menze. Generatect: Text-conditional generation of 3d chest ct volumes. *arXiv preprint arXiv:2305.16037*, 2023.

Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. Ct2rep: Automated radiology report generation for 3d medical imaging. *arXiv preprint arXiv:2403.06801*, 2024a.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, Weicheng Dai, Murong Xu, Hadrien Reynaud, Muhammed Furkan Dasdelen, Bastian Wittmann, Tamaz Amiranashvili, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, Abdullah

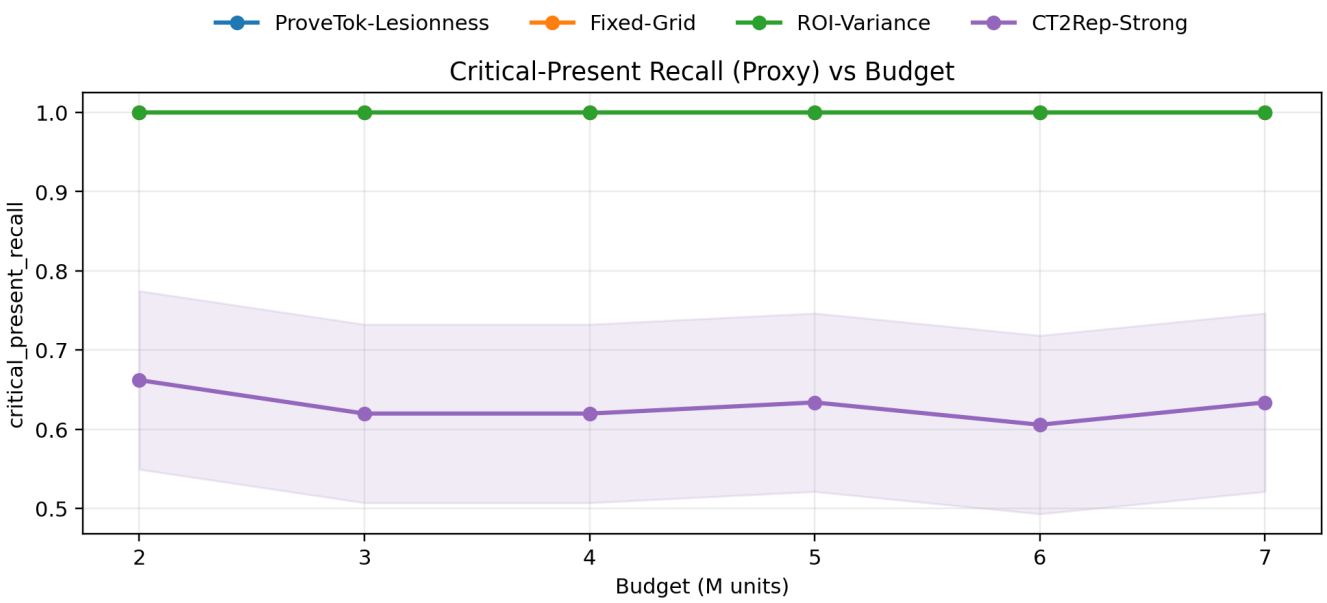

Figure 3: Critical-present recall vs. budget (data: `outputs/E0171-full3/baselines_curve_multiseed.json`).

Alanbay, Anjany Sekuboyina, Berkan Lafci, Ahmet Kaplan, Zhiyong Lu, Malgorzata Polacin, Bernhard Kainz, Christian Bluethgen, Kayhan Batmanghelich, Mehmet Kemal Ozdemir, and Bjoern Menze. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834*, 2024b.

Maojia Song, Shang Hong Sim, Rishabh Bhardwaj, Hai Leong Chieu, Navonil Majumder, and Soujanya Poria. Measuring and enhancing trustworthiness of llms in rag through grounded attributions and learning to refuse. *arXiv preprint arXiv:2409.11242*, 2024.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel T. Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023. doi: 10.1148/ryai.230024. Preprint: arXiv:2208.05868.

A Additional Tables

B Verifier Ruleset (Abridged)

C Notes on Evidence Profiles

We use the gold-mask real profile for primary claims, and we treat cross-dataset silver-label profiles as stress tests. This separation is necessary to avoid over-claiming based on automatically generated masks (e.g., TotalSegmentator [Wasserthal et al., 2023]) or pseudo masks.

Figure 3. Allocation Regret Sweep (E0161)

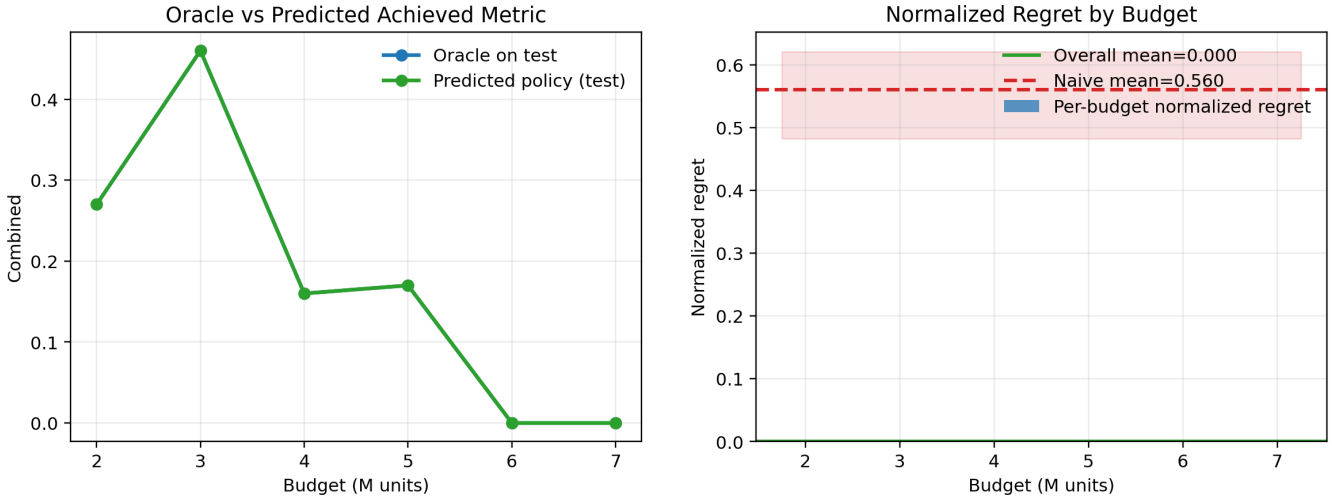

Figure 4: Allocation regret sweep with bootstrap confidence intervals (data: `outputs/E0161-full/fig3_regret_sweep.json`).

Figure 4. Counterfactual Pooled Test (E0167R2) | omega_perm mean=0.0026, p1=0.0001, p_holm=0.0006

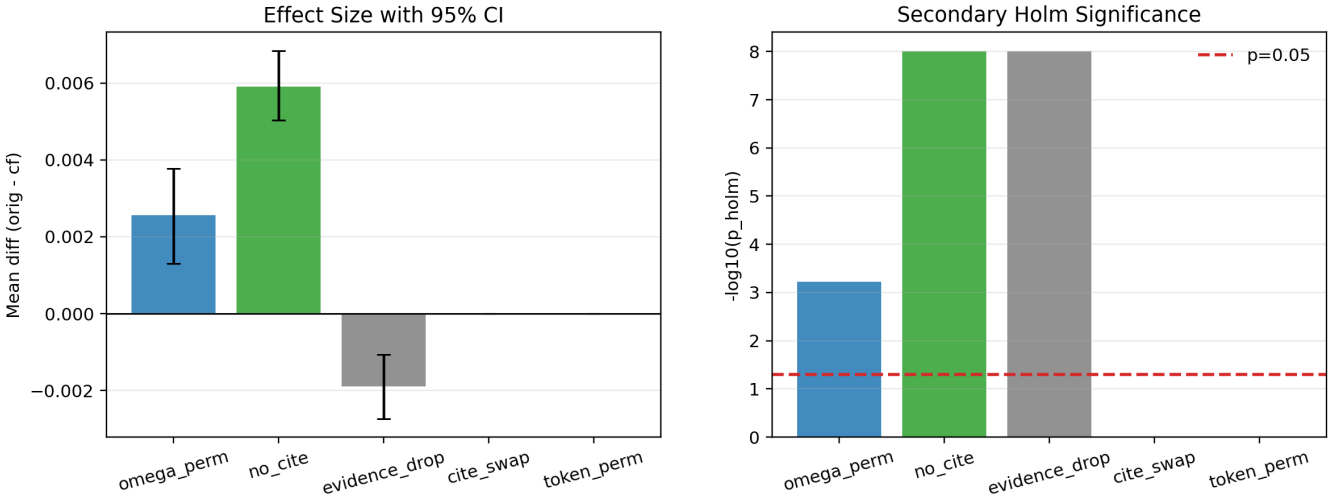

Figure 5: Pooled counterfactual power and significance for ω-permutation (data: `outputs/E0167R2-ct_rate-tsseg-effusion-counterfactual-power-seed20/omega_perm_power_report.json`).

Table 1: Datasets and evaluation profiles. We use gold-mask profiles for primary claims and silver-label profiles only for cross-dataset stress tests.

Profile	Dataset	Mask type	Role / Notes
Real (gold)	ReXGroundingCT-100g (test=231) [Baharoon et al., 2025]	finding-level 3D masks	Primar grounding and counterfactual evaluation; used for main claims.
Real (gold)	ReXGroundingCT-mini (test=57) [Baharoon et al., 2025]	finding-level 3D masks	Fast iteration and regression (smoke,
Silver (auto)	CT-RATE TS-Seg eval-only (test=38) [Hamamci et al., 2024b, Wasserthal et al., 2023]	automatic masks	Cross-dataset sanity and stress tests; *not* a replacement for gold evidence.
Silver (pseudo)	CT-RATE pseudo-mask (test=30)	pseudo masks	Cross-dataset stress tests built from saliency to-mask pipelin used to probe robustness.

Table 2: Oral minimal evidence set (paper-grade). Each item is audited by a fixed protocol and points to a reproducible artifact.

Item	Verdict	Key numbers	Protocol	Evidence
C0001	Pass	6/6 budgets pass for combined and grounding; latency and unsupported constraints satisfied	6 budgets, 5 seeds, paired bootstrap (n_{boot}=20,000) with Holm over budgets; latency p95 $\leq$ +5%, unsupported $\Delta \leq 0.05$	`outputs/E0164-full/baselines_curve_mult`
C0002	Pass	dev→test regret CI small; beats naive policy	model fit (AIC/BIC), paired bootstrap CI; requires CI upper bound < 0.15 and improvement over naive	`outputs/E0161-full/fig3_regret_sweep.js`
C0003	Pass	no_cite breaks grounding; cite_swap breaks supportedness (Holm-significant)	paired bootstrap + Holm over counterfactual family	`outputs/E0162-full_retry3/figX_counterf`
C0004	Pass	grounding IoU improves across baselines over multiple budgets (Holm-significant)	one-sided (H_1>0) + Holm over budgets; n_{boot}=20,000	`outputs/E0156-grounding_proof_100g_sali`
C0005	Pass	τ_{refuse}=0.002 meets miss/ECE/refusal constraints; supportedness improves across budgets	hard gates per budget; requires improvement on supportedness in $\geq 2/3$ budgets	`outputs/E0144-full/figX_refusal_calibra`
C0006	Pass	audited cost accounting and strong baselines are present and non-degenerate	baseline coverage + audited cost accounting + reproducible strong baseline	`outputs/E0164-full/baselines_curve_mult`
V0003 (ω-perm)	Pass	pooled mean diff = 0.0026, 95% CI [0.0013, 0.0038]; one-sided $p = 10^{-4}$; Holm $p = 6 \times 10^{-4}$; positive 19/20	pooled one-sided test + secondary Holm over counterfactual family	`outputs/E0167R2-ct_rate-tsseg-effusion-`

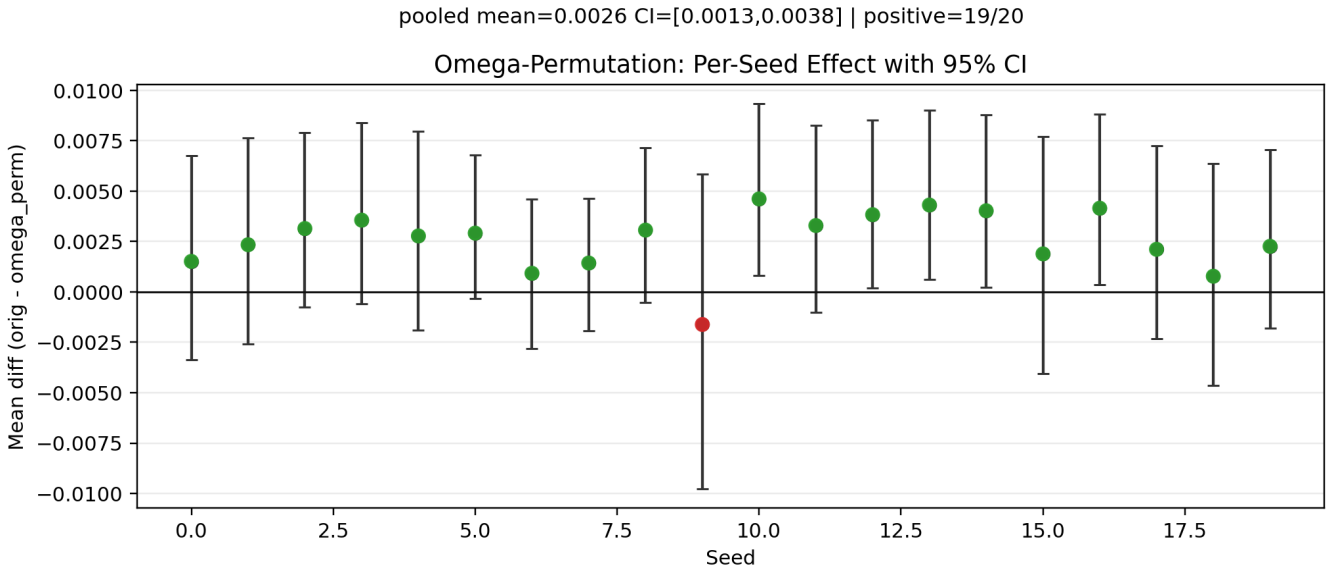

Figure 6: Seed-level ω-permutation effect sizes with 95% confidence intervals (data: `outputs/E0167R2-ct_rate-tsseg-effusion-counterfactual-power-seed20/omega_perm_power_report.json`).

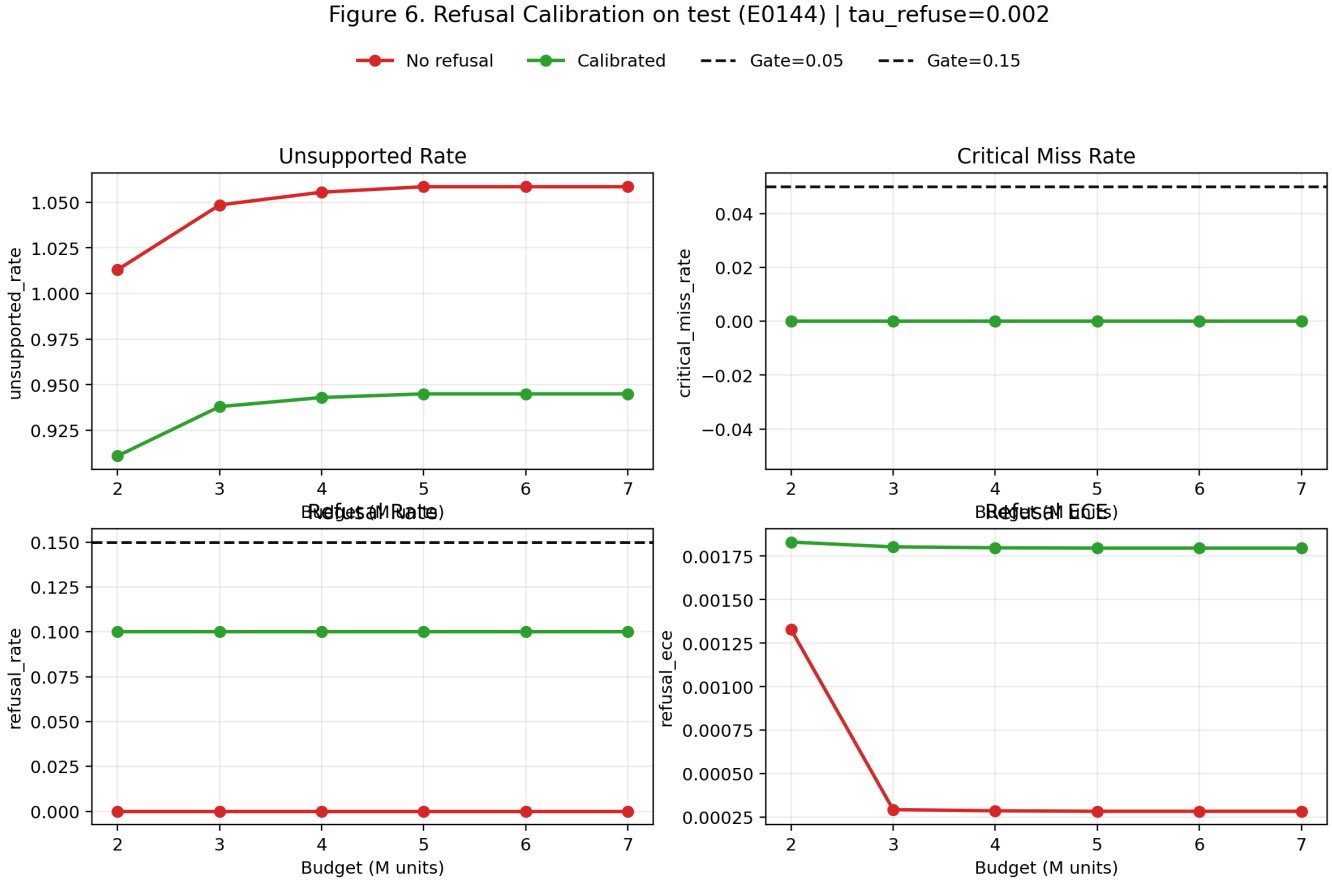

Figure 7: Refusal calibration with hard gates (data: `outputs/E0144-full/figX_refusal_calibration.json`).

Figure 5. Qualitative Cases from outputs/E0163-full-v3

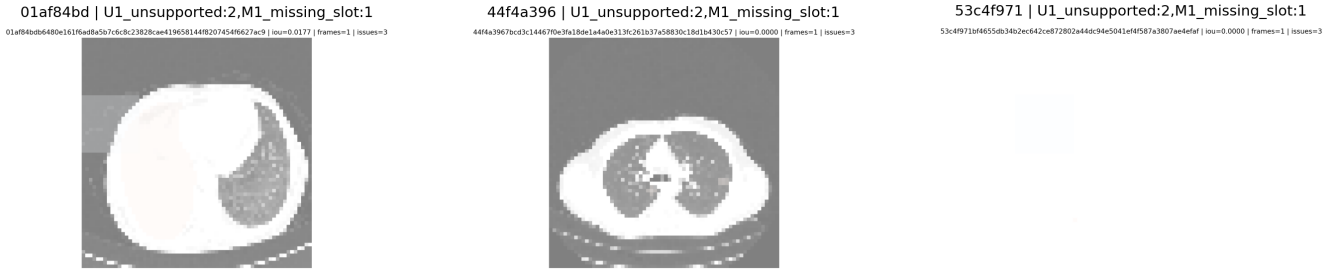

Figure 8: Qualitative case studies (data: `outputs/E0163-full-v3/`).

Table 3: Claim-level machine verdict (real profile).

Claim	Status	Summary	Primary evidence
C0001	Pass	ProveTokimproves combined quality and grounding across budgets under matched cost, while satisfying latency and supportedness constraints.	`outputs/E0164-full/baselines_curve_multiseed.json`
C0002	Pass	Dev→test regret is statistically bounded and improves over a naive allocation policy in the real pipeline.	`outputs/E0161-full/fig3_regret_sweep.json`
C0003	Pass	Counterfactual tests show that removing citations breaks grounding and swapping citations breaks supportedness.	`outputs/E0162-full_retry3/figX_counterfactual_20260206_10`
C0004	Pass	Citation-grounding improves over fixed-grid and ROI-style baselines with Holm-significant gains across multiple budgets.	`outputs/E0156-grounding_proof_100g_saliency_full/figX_gro`
C0005	Pass	τ_{refuse}=0.002 meets miss/ECE/refusal constraints and reduces unsupported rates across budgets.	`outputs/E0144-full/figX_refusal_calibration.json`
C0006	Pass	Baseline suite and cost accounting are present, with reproducible strong baselines.	`outputs/E0164-full/baselines_curve_multiseed.json`

Table 4: V0003 cross-dataset grounding and counterfactual summary (silver-label stress tests).

Item	Scope	Key result	Verdict
E0166 grounding vs ROI	TS-Seg eval-only, budgets $2 \times 10^6..7 \times 10^6$	IoU_union positive 6/6, Holm-significant 6/6	Pass
E0166 grounding vs Fixed-Grid	TS-Seg eval-only, budgets $2 \times 10^6..7 \times 10^6$	IoU_union positive 5/6, Holm-significant 4/6	Partial
E0167 seeds 0..2 no_cite	counterfactual	mean diff (avg)=0.0059, Holm-significant 3/3	Pass
E0167 seeds 0..2 ω-perm	counterfactual	mean diff (avg)=0.0023, Holm-significant 0/3	Not sig.
E0167R pooled	seeds 0..9	mean diff=0.0020, 95% CI $[0.0001, 0.0037]$, one-sided $p = 0.0187$, Holm $p = 0.1122$	Primary only
E0167R2 pooled	seeds 0..19	mean diff=0.0026, 95% CI $[0.0013, 0.0038]$, one-sided $p = 10^{-4}$, Holm $p = 6 \times 10^{-4}$	Pass

Table 5: ω-permutation variant search (seed 0).

Variant	Setting	ω diff	p_{Holm}	no_cite diff	p_{Holm}
BASE	score + topk=3 (baseline)	0.0015	1	0.0059	0
RA	+ score_to_uncertainty	0.0015	1	0.0059	0
RD	+ score_level_power=1.0	0.0015	1	0.0059	0
RC	score_interleave citations	0.0006	1	0.0132	0
RB	baseline with topk=1	0.0005	1	0.0017	0.0145

Table 6: **Rule-based verifier taxonomy (abridged).** The verifier is deterministic and versioned for audited artifacts. Each triggered rule emits an issue with an `evidence_trace` containing cited token ids/cell ids and rule-specific intermediate values (Appendix schema). Thresholds shown are default audited settings.

Rule	Type	Check (applies to asserted frames unless noted)	Key thresholds / outputs
U1.0	unsupported	Positive claim without any citations	severity=3; `token_ids=[]`
U1.1	unsupported	Low evidence strength: $\max_{t \in C_k} s(t) <$ `min_score`	`min_score=0.35`; `scores,max_score`
U1.2	unsupported	High uncertainty: $\min_{t \in C_k} u(t) >$ `max_uncertainty`	`max_uncertainty=0.7`; `uncertainties,min`
U1.3	unsupported	Insufficient spatial coverage: $\sum_{t \in C_k} 8^{-\ell(t)} <$ `min_coverage_ratio`	`min_coverage_ratio=0.1`; `levels,coverage`
U1.4	unsupported	Citation relevance proxy: cited tokens fail `recall@k` or `attention mass` under a deterministic query attention	`min_recall@k=0.5`, `min_att_mass=0.2`; `topk_token_ids,recall@k,att_mass`
O1.0	overclaim	Specificity mismatch: laterality specified but only coarse evidence levels cited	`max_coarse_level=0`; `levels,max`
O1.1	overclaim	High-confidence serious finding lacks strong evidence (score mismatch)	`avg_evidence_score`, `confidence`