OPEN-ONTOLOGY GAP IN FREE-TEXT MEDICAL SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Free-text medical segmentation lets users request clinical targets in ordinary language, but a plausible mask may represent the wrong concept. Prior work reports category errors and limited generalization, but lacks a controlled test of whether a model selects the requested target when a related alternative is present. We introduce Clinical Contrast Sets, which holds the image fixed, changes the requested concept, and measures whether the mask shifts toward the intended region and away from its competitor. On 40 MS3SEG development patients, Medical-SAM3 and VoxTell preferentially select multiple sclerosis (MS) lesions when asked for the dataset's normal/non-MS white-matter hyperintensities (nWMH). A protocol frozen before evaluating the 20 official-test patients confirms this pattern: mean target-minus-confounder coverage is -0.308 and -0.363, with 0/20 correct canonical prompt pairs for each model. Positive descriptions and volume controls preserve the failure, local size matching confirms it for VoxTell, and published supervised results establish that nWMH is learnable. The failure is not a missing label: VoxTell's published training inventory lists both MS lesions and white matter hyperintensities, and the model switches correctly between documented kidney and tumor concepts, yet it resolves nWMH prompts, and even the generic white-matter-hyperintensity prompt, to MS lesions. Incorrect nWMH masks retain a mean foreground probability of 0.962, and a prompt-overlap warning that flags all 20 official VoxTell failures performs near chance within BraTS. This matters because the failure is not a missing answer but a wrong one: a clinician requesting a benign, non-MS finding instead receives a confident, well-formed mask of MS pathology, with nothing in the model's own output to flag the substitution. We call this mismatch between language flexibility and reliable visual target selection the open-ontology gap, and offer Clinical Contrast Sets as a controlled protocol for auditing it: with two annotated competitors, two prompts, and a superclass reference, developers can identify which clinical distinctions a model does not reliably select and verify whether an adaptation restores correct switching rather than only raising overlap.

1 INTRODUCTION

Text-prompted medical segmentation is moving from fixed label dictionaries toward free-form language interfaces. Recent systems accept anatomical names, disease labels, or clinical descriptions and return 3D masks across multiple modalities (Zhao et al., 2025; Rokuss et al., 2026; Jiang et al., 2026). The appeal is clear: clinicians should be able to express meaningful distinctions without retraining a new network for every label.

A free-text interface accepts any clinical phrase, but accepting a phrase does not ensure that the intended region is selected: a request for one of two coexisting targets may return a plausible mask for the other or for their common category. This raises a complementary evaluation question: when two related clinical targets coexist, does changing the requested concept change which one the model segments?

We address this evaluation gap with Clinical Contrast Sets. Each set contains two clinically distinct regions in one volume. We vary the requested target while holding the patient and image fixed,

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.

then measure overlap with both the intended region and its competitor. Correct grounding requires a directional switch. This adapts contrast sets (Gardner et al., 2020) to dense clinical target selection.

Our motivating case is MS3SEG (Bawil et al., 2026), which annotates both abnormal white-matter hyperintensities corresponding to MS lesions and normal/non-MS white-matter hyperintensities (nWMH), defined as age-related or CSF-contaminated regions. Medical-SAM3 and VoxTell consistently select MS lesions even when prompted for nWMH. That observation alone is not enough: it could reflect class-size bias, awkward terminology, negation failure, generic model incompetence, or an intrinsically unlearnable annotation.

We test these explanations using class-normalized coverage, volume reversal, component-size matching, and positive descriptions without explicit negation. Text-embedding distances probe whether the language encoder distinguishes the prompts. Published training concepts show both MS lesions and white matter hyperintensities are documented for VoxTell; tumor and kidney controls and supervised nWMH results provide context. A protocol frozen before official-test inference specifies held-out confirmation, confidence, MS laterality, and prompt-overlap diagnostics.

The official test confirms the development pattern. VoxTell distinguishes the relevant text embeddings and segments documented control concepts, yet preferentially selects MS lesions for nWMH prompts with high foreground probability. We call this mismatch between language flexibility and reliable visual target selection the open-ontology gap. Our contribution is a controlled diagnosis of substitution, with explicit tests of alternative explanations and limits on detection; category errors and confident failures themselves are established problems.

Our contributions are:

- A controlled clinical grounding protocol. Clinical Contrast Sets uses same-image competing targets, target-minus-confounder margins, superclass nulls, and pairwise target switching to test semantic selection rather than generic localization.
- A frozen official-test confirmation. The untouched 20-patient test reproduces the nWMH family-level failure in two 3D segmenters after class-size normalization, volume reversal, and positive reformulation. Local component matching confirms the effect for VoxTell; Medical-SAM3's matched interval crosses zero.
- A documented training-concept audit. VoxTell's published base inventory lists both MS lesions and white matter hyperintensities, the two concepts this contrast requires, alongside the control concepts, so the substitution occurs between documented concepts; Medical-SAM3 exposes no comparable inventory.
- Learnability, competence, and diagnostic boundaries. Published supervised results establish nWMH learnability; kidney laterality switches reliably whereas hemisphere-qualified MS prompts do not. Confidence and within-dataset detector checks show why neither plausible probabilities nor pooled detection scores establish semantic reliability.

2 Related Work

Text-prompted medical segmentation. MedSAM, SegVol, SAT, BiomedParse, Medical-SAM3, and VoxTell extend promptable segmentation across medical targets and modalities (Ma et al., 2024; Du et al., 2024; Zhao et al., 2025; 2024; Jiang et al., 2026; Rokuss et al., 2026). ReXGrounding documents difficulty grounding chest-CT findings (Baharoon et al., 2025). VoxTell itself reports poor performance on Stanford Knee structures absent from training (Rokuss et al., 2026); our setting differs because MS lesions and white matter hyperintensities are both documented VoxTell concepts, yet the model does not switch between them. We characterize how a requested clinical subclass is replaced by a coexisting competitor even when the output remains confident.

Category errors and visual adaptation. PrAda identifies category assignment as a major source of error despite accurate region localization in specialized domains, and learns visual prototypes for few-shot adaptation (Rosi et al., 2026). Few-Shot Concept Prompt Learning learns continuous prompts from medical image–mask pairs while freezing the segmentation backbone (Venkataramani & Sathish, 2026). These studies establish related limitations and propose adaptations. Our contribution is an evaluation protocol combining same-image clinical competitors, superclass references,

size and wording controls, and frozen held-out confirmation. We do not claim the first observation of incorrect semantic labels or a demonstrated adaptation remedy.

Compositionality, negation, and semantic faithfulness. Vision-language models struggle with compositional reasoning (Thrush et al., 2022; Ma et al., 2023; Zeng et al., 2024) and negation (Alhamoud et al., 2025). CAFE tests segmentation faithfulness through counterfactual visual changes (Liang et al., 2026). In robotics, LangGap holds tabletop layouts fixed while varying instruction semantics (Hou & Zhao, 2026), sharing our broader principle of separating language sensitivity from scene cues. We instantiate that principle for clinical masks with coexisting annotated targets, size and wording controls, and a documented-concept audit.

Confident failures and monitoring. Beyond Uncertainty studies high-confidence surgical segmentation errors under acquisition degradation and combines confidence, shape, temporal consistency, and image quality for monitoring (Pham et al., 2026). Our study changes the requested clinical meaning while keeping acquisition fixed. Confident errors are not our novelty; our evidence concerns concept substitution and a prompt-swap diagnostic whose poor within-BraTS discrimination we report.

Training vocabulary and ontology coverage. Documented vocabulary is neither necessary nor sufficient for grounding, but it constrains one explanation: when a requested concept is documented in training, failure to select it cannot be attributed to a missing label. We audit VoxTell's published inventory for this purpose in Section 4.3 and Appendix K.

3 Clinical Contrast Sets

3.1 Same-image contrastive grounding

For an image volume x, let A and B denote two disjoint, coexisting clinical regions. A text-conditioned segmenter f produces a binary mask $P_q = f(x, q)$ for prompt q. Prompts q_A and q_B request A and B, respectively. A faithful model should satisfy two directional conditions: P_{q_A} should overlap A more than B, and P_{q_B} should overlap B more than A.

$$\Delta_{\mathrm{Dice}}(q_A) = \mathrm{Dice}(P_{q_A}, A) - \mathrm{Dice}(P_{q_A}, B), \tag{1}$$

with signs reversed for q_B. Because Dice depends on target size, our primary size-normalized statistic is

$$\Delta_{\mathrm{Cov}}(q_A) = \frac{|P_{q_A} \cap A|}{|A|} - \frac{|P_{q_A} \cap B|}{|B|}, \tag{2}$$

again reversing target/confounder for q_B. Positive values indicate greater fractional coverage of the requested target; each class has its own denominator, although size may still affect prediction difficulty. We also report pair success: both canonical prompts must have positive Δ_{Cov} for the same case. This is a directional criterion, not a guarantee of high Dice; a broad mask covering both classes equally has zero margin.

3.2 Superclass and hypothesis controls

For each contrast set we include a superclass prompt q_S such as "white matter hyperintensity," "meningioma," or "kidney." It provides a reference for the model's unqualified bias between A and B. For a child prompt, we therefore also measure its directional shift relative to the superclass.

The protocol separates four explanations: (i) prevalence/base-rate bias, where a generic mask favors the larger target; (ii) prompt compression or compositional failure, where the linguistic distinction is weakly represented; (iii) generic localization failure, where the model cannot segment the relevant anatomy/pathology; and (iv) an open-ontology gap, where the interface accepts the expression but the requested subclass is not reliably bound to a distinct visual target.

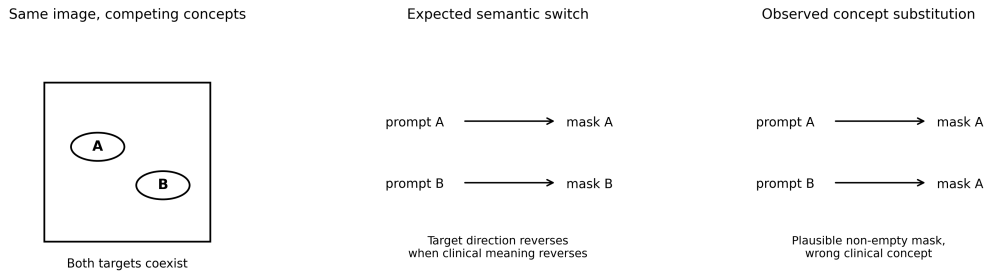


Figure 1: Clinical Contrast Sets. Two clinically distinct targets coexist in the same image. Correct grounding requires the selected region to switch when the requested clinical concept switches. Concept substitution returns a plausible non-empty mask while remaining on the competing concept.

Table 1: Development and external contrast sets. A/B volume ratio is the median case-wise ratio after preprocessing; the MS3SEG row describes the 40 development patients.

Dataset	**Concept A**	**Concept B**	**Median** $\lvert A\rvert/\lvert B\rvert$
MS3SEG	MS lesion	normal/non-MS WMH	0.89
BraTS MEN	enhancing tumor	non-enhancing core	89.19
TotalSegmentator	left kidney	right kidney	1.02

4 Experimental Design

4.1 Datasets and contrast sets

MS3SEG. MS3SEG contains 100 patients with T1, T2, and FLAIR MRI and tri-mask annotations for ventricles, nWMH, and abnormal WMH/MS lesions (Bawil et al., 2026). We use the provided preprocessed FLAIR volumes. Of the 80 development patients, separate cohorts were used during discovery; the main cross-model analysis uses a fixed 40-patient cohort subsequently reused for mechanism and control analyses. The official 20-patient test set remained blocked during development. Before opening it, we froze the exact IDs, prompt bank, models, preprocessing, endpoints, and analysis rules, then evaluated each released model once on all 20 official patients without changing those choices. We report development and confirmation cohorts separately. Here, nWMH is the dataset's annotation-defined class, not a claim that every non-MS hyperintensity is biologically normal.

BraTS 2023 Meningioma. BraTS MEN provides multi-sequence MRI with enhancing tumor, non-enhancing tumor core, and surrounding T2/FLAIR hyperintensity labels (LaBella et al., 2024). We select 40 training cases containing at least 20 voxels of both enhancing and non-enhancing tumor core, use T1 post-contrast MRI, and allow at most one exam per patient. Selection is deterministic from a frozen hash salt and occurs before model inference. Because the control concepts are represented in VoxTell's published training vocabulary, we interpret this as a concept-competence/transfer control, not as prospective out-of-distribution validation.

TotalSegmentator. TotalSegmentator provides dense anatomical labels on CT, including separate left and right kidneys (Wasserthal et al., 2023). We deterministically select 40 of the 102 cases in the small v2.0.1 release for which both kidney masks are present, resample to 2-mm isotropic resolution, and preserve the full field of view. For VoxTell, left/right kidney are documented training concepts; this experiment therefore tests whether the model can execute a contrastive target switch when both concepts are demonstrably grounded.

4.2 Models and prompts

We evaluate two independently developed 3D free-text segmenters: Medical-SAM3 (Jiang et al., 2026) and VoxTell (Rokuss et al., 2026), using released checkpoints and official inference imple-

mentations. Exact repository revisions and runtime versions are recorded in the supplementary manifests.

For MS3SEG, the original frozen families contain three MS prompts and three nWMH prompts, plus the generic "white matter hyperintensity" superclass. Four positive descriptions test terminology/negation, including "age-related or CSF-contaminated white matter hyperintensity." The confirmation protocol adds "left-hemisphere multiple sclerosis lesion" and its right-hemisphere counterpart as a same-disease modifier control. A secondary development-only "periventricular cap" prompt is reported in the appendix. BraTS uses two enhancing prompts, two non-enhancing prompts, and "meningioma"; TotalSegmentator uses "left kidney," "right kidney," and "kidney." Appendix A gives every exact prompt. Family-level statistics average the designated prompts within each patient before averaging across patients.

4.3 Published concept-coverage and learnability audits

VoxTell's published base inventory explicitly lists multiple sclerosis lesion, white matter hyperintensities, white matter lesions, enhancing and non-enhancing tumor, and left/right kidney (Rokuss et al., 2026); Appendix K gives the counts. The MS3SEG contrast therefore involves two documented concepts: MS lesions and white matter hyperintensities, the category to which the dataset's normal/non-MS WMH belong. The question is whether a model that documents both concepts selects the requested one when both coexist. Medical-SAM3 reports 33 training datasets across 10 modalities without a comparable concept list (Jiang et al., 2026); its concept membership is marked unknown.

The MS3SEG dataset paper reports nWMH Dice 0.6452 ± 0.0091 for supervised U-Net in a five-fold multi-class experiment (Bawil et al., 2026). This independent result establishes learnable signal in the annotation under that protocol. It is not an exact-split upper bound or a matched comparator for the evaluated models.

4.4 Statistical protocol and integrity

The development analysis and external-cohort rules were frozen in JSON before the corresponding new inference. A separate confirmation protocol was frozen before official-test inference, with primary endpoints covering MS/nWMH family directions, canonical pair success, positive descriptions, volume reversal, and component-matched direction. Canonical JSON hashes match the execution manifests (Appendix H). Confidence intervals use 5,000 case-level bootstrap resamples unless otherwise stated; components are clustered by patient. Intervals are endpoint-wise, not multiplicityadjusted. A protocol freeze documents fixed choices, not an externally registered study.

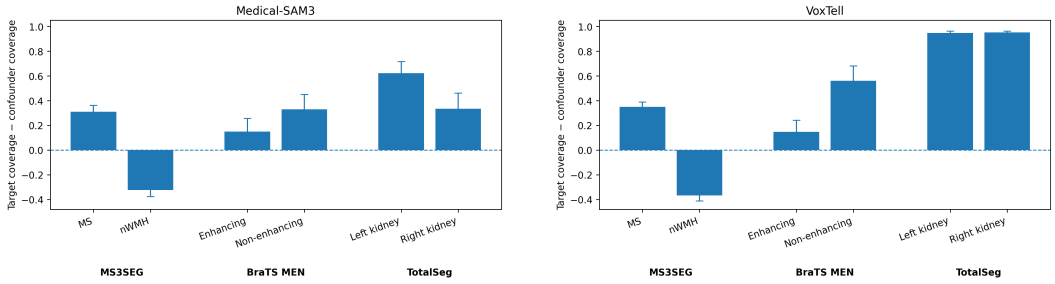
Three diagnostics were also fixed before confirmation. The laterality control splits canonical-RAS MS masks at the nonzero-FLAIR bounding-box midline and requires at least 20 MS voxels per hemisphere after excluding MS/nWMH overlap. VoxTell confidence uses native pre-threshold logits. PromptSwapDice measures Dice overlap between opposing canonical-prompt masks; larger values are flagged. The score needs no labels at application, but its threshold is calibrated with success/failure labels on 120 pre-existing pairs per model (40 per dataset), maximizing balanced accuracy, then applied unchanged to the official test. We additionally report a post-hoc within-BraTS discrimination audit without threshold refitting.

5 Results

5.1 The frozen official test confirms the MS3SEG failure

On the fixed 40-patient development cohort, both models select the expected side for MS prompts but move in the wrong direction for nWMH prompts (Figure 2). Medical-SAM3 has mean class-normalized margins +0.310 for the MS family and -0.324 [-0.375, -0.274] for the nWMH family; VoxTell has +0.349 and -0.368 [-0.411, -0.323]. Neither model achieves a correct canonical MS/nWMH prompt pair in any development patient (0/40 each).

The one-shot official test reproduces this pattern closely (Table 2, Figure 3). Medical-SAM3 yields +0.310 [0.254, 0.363] for MS and -0.308 [-0.359, -0.254] for nWMH; VoxTell yields +0.335



(a) Medical-SAM3 (b) VoxTell

Figure 2: Development-cohort directional grounding. Bars show target coverage minus confounder coverage with 95% bootstrap intervals. MS3SEG nWMH is systematically negative, whereas BraTS and kidney-laterality controls are positive on average. The external contrasts establish competence on documented anatomical and tumor concepts; they are not additional failing contrast sets.

[0.266, 0.416] and −0.363 [−0.446, −0.291]. Canonical pair success is again 0/20 for both models. Thus the central result replicates on patients not used in discovery, model comparison, prompt selection, or threshold selection.

Larger MS lesions alone do not explain the result. In the development cohort, 23/40 patients have at least as much nWMH as MS-lesion volume; their mean nWMH margins remain −0.282 for Medical-SAM3 and −0.376 for VoxTell, with intervals entirely below zero. Among the nine official-test patients satisfying this pre-specified volume reversal, the corresponding means are −0.315 [−0.404, −0.220] and −0.406 [−0.550, −0.296]. In development data, VoxTell's nWMH direction is essentially uncorrelated with the log MS:nWMH volume ratio (ρ = 0.004, p = 0.981); Medical-SAM3 shows some dependence (ρ = −0.348, p = 0.028), but its mean remains negative in every volume tertile.

We further match MS and nWMH connected components within each patient by component volume (maximum 2:1 ratio). In development data, the canonical nWMH prompt covers the matched MS component more strongly: mean nWMH-minus-MS component coverage is −0.082 [−0.117, −0.046] over 877 Medical-SAM3 pairs and −0.147 [−0.192, −0.097] over 878 VoxTell pairs. On the official test, the direction remains negative over 462 and 458 pairs, respectively: −0.052 [−0.135, 0.016] for Medical-SAM3 and −0.102 [−0.198, −0.010] for VoxTell. The Medical-SAM3 interval crosses zero, whereas the primary VoxTell result remains significant after local size matching.

The two models should not be weighted equally as competence evidence. On development data, Medical-SAM3 has low absolute MS3SEG accuracy: canonical MS Dice is 0.198, while the "normal WMH" prompt produces an even higher MS Dice of 0.239 but only 0.026 Dice on nWMH. We therefore treat Medical-SAM3 as corroboration. VoxTell is the primary evidence: it attains 0.552 MS Dice under the canonical MS prompt yet gives 0.179 nWMH Dice under "normal WMH" while preferentially covering MS. Published supervised U-Net nWMH Dice of 0.6452 shows that the free-text failure is not equivalent to an intrinsically unlearnable target.

5.2 Positive descriptions do not rescue nWMH grounding

If the problem were merely that "normal WMH" is awkward or that negation is mishandled, positive descriptions should move the prediction toward nWMH. They do not. Averaging the four pre-specified positive descriptions, development-cohort nWMH direction is −0.322 [−0.374, −0.271] for Medical-SAM3 and −0.332 [−0.378, −0.283] for VoxTell. The frozen official test reproduces the result: −0.312 [−0.364, −0.257] and −0.340 [−0.429, −0.263].

On the official test, the generic "white matter hyperintensity" prompt is likewise MS-biased: −0.307 for Medical-SAM3 and −0.341 for VoxTell. Since white matter hyperintensities is a documented VoxTell concept, even that documented prompt resolves to MS lesions here. The descriptive family differs from this superclass direction by only −0.005 [−0.018, 0.009] and +0.001 [−0.008, 0.011]. More clinically explicit wording therefore does not recover the requested subtype.

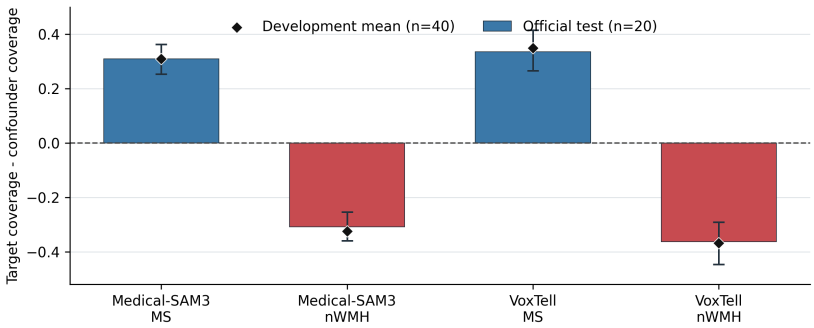


Figure 3: Frozen official-test confirmation. Bars and 95% bootstrap intervals show official-test family directions; black diamonds show the corresponding development-cohort means. The correct MS direction remains positive, while the requested nWMH direction remains negative for both models.

5.3 Confidence and the limits of prompt-overlap warnings

VoxTell's mean probability inside the predicted-positive mask is 0.9616 for the misdirected nWMH prompt versus 0.9658 for the MS prompt. The paired difference is -0.0042 [-0.0085, -0.0008]; the mean within-patient 95th percentile is 1.000 for both. Thus incorrect semantic selection can coexist with high foreground probability. These uncalibrated summaries do not establish a confidence-based detector or rule out other uncertainty scores.

PromptSwapDice initially appears informative when datasets are pooled: calibration AUROC/average precision is 0.916/0.917 for VoxTell and 0.595/0.771 for Medical-SAM3. The frozen thresholds flag 20/20 and 11/20 official MS3SEG failures, respectively. However, all MS3SEG calibration pairs fail, while all VoxTell kidney pairs succeed, confounding dataset/concept identity with failure status.

The within-BraTS audit exposes this limitation. Among 21 failures and 19 successes, VoxTell has AUROC 0.461 [0.276, 0.649] and average precision 0.603; the unchanged threshold flags 9/21 failures and 6/19 successes. Medical-SAM3 has AUROC 0.473 [0.301, 0.655]. Thus the pooled score does not establish within-dataset failure discrimination. Because every official MS3SEG pair fails, that cohort estimates alert sensitivity but cannot estimate false-positive rate or AUROC. We report prompt overlap as a setting-dependent diagnostic, not a validated safeguard (Appendix I).

5.4 Text geometry separates two failure modes

For Medical-SAM3's contextual token mean, the MS-related/non-MS polarity distance is 0.087 versus an average within-family paraphrase distance of 0.125 (ratio 0.69). Together with its low MS3SEG accuracy, this is consistent with prompt compression contributing to failure. VoxTell behaves differently: its Qwen3 embedding distance is 0.120 versus a mean paraphrase distance of 0.108 (ratio 1.11), while the corresponding output-mask distance ratio is 1.57. Thus VoxTell is measurably responsive to the linguistic distinction even though the selected visual target remains wrong. Related-pair-only text/output distance correlation is high for VoxTell (ρ = 0.943 over six pre-specified pairs), but we treat this as descriptive because the number of pairs is small.

5.5 Known-concept controls establish selective competence

BraTS and TotalSegmentator should be interpreted narrowly. VoxTell's published base inventory explicitly contains enhancing/non-enhancing tumor concepts and left/right kidney, so success does not demonstrate out-of-distribution compositional generalization. Instead, it establishes that the same model can execute the contrast-set operation for documented anatomical and tumor concepts.

On the full 40-case BraTS cohort, both models have positive normalized directional margins for both enhancing and non-enhancing prompt families despite a median enhancing:non-enhancing volume ratio of 89.2:1. Medical-SAM3 achieves +0.149 [0.043, 0.256] for enhancing and +0.329 [0.212,

Table 2: Primary contrast-set results. Direction values are class-normalized target-minus-confounder coverage. Pair success requires both canonical prompt directions to be positive for the same case. Development and official MS3SEG cohorts are reported separately.

Model	Cohort	Contrast	Dir. A	Dir. B	Pair success
Medical-SAM3	Dev.	MS / nWMH	+0.310	-0.324	0/40
Medical-SAM3	Official	MS / nWMH	+0.310	-0.308	0/20
VoxTell	Dev.	MS / nWMH	+0.349	-0.368	0/40
VoxTell	Official	MS / nWMH	+0.335	-0.363	0/20
Medical-SAM3	Control	Enh. / non-enh.	+0.149	+0.329	12/40
VoxTell	Control	Enh. / non-enh.	+0.146	+0.561	19/40
Medical-SAM3	Control	Left / right kidney	+0.621	+0.335	30/40
VoxTell	Control	Left / right kidney	+0.948	+0.951	40/40
Medical-SAM3	Official	Left / right MS	-0.055	+0.094	1/19
VoxTell	Official	Left / right MS	+0.031	+0.255	10/19

0.450] for non-enhancing tumor; VoxTell achieves +0.146 [0.058, 0.243] and +0.561 [0.434, 0.681]. Canonical pair success is 12/40 and 19/40.

Because a 20-voxel non-enhancing minimum could make coverage unstable, we additionally performed a post-hoc sensitivity analysis requiring at least 200 non-enhancing voxels. Nineteen cases remain, with median enhancing:non-enhancing ratio 22.5:1. Coverage directions remain positive for both families: Medical-SAM3 +0.296 [0.148, 0.447] / +0.495 [0.353, 0.624], and VoxTell +0.298 [0.160, 0.448] / +0.592 [0.400, 0.758]. Canonical pair success rises to 10/19 and 15/19. Dice-direction sensitivity is also positive except that VoxTell's non-enhancing Dice-margin interval crosses zero; full values are reported in Appendix L. We therefore use BraTS as a directional competence control, not as evidence of perfect fine-grained tumor segmentation.

Laterality is the cleanest competence control. Kidney volumes are approximately balanced (median left:right ratio 1.02). Medical-SAM3 has positive left/right margins (+0.621, +0.335) and 30/40 pair success. VoxTell has +0.948 and +0.951, with 40/40 pair success and high requested-target Dice (0.919 left, 0.917 right).

Hemisphere-qualified MS is a compositional probe. Among 19 eligible official-test patients, Medical-SAM3 succeeds for both hemisphere prompts in 1/19 and VoxTell in 10/19. VoxTell's left/right mean directions are +0.031/ + 0.255; Medical-SAM3's are -0.055/ + 0.094 (Table 2). VoxTell's left direction has a post-hoc 95% interval of $[-0.045, 0.103]$, versus [0.177, 0.335] for the right. It does not establish reliable same-disease laterality: switching that is reliable for kidney laterality does not extend to a hemisphere qualifier on MS lesions. It is not a second independent nWMH-like failure, because it reuses MS3SEG, has partial success, and differs from kidney in anatomy, modality, and target geometry; we do not attribute the difference to training coverage.

6 The Open-Ontology Gap

The evidence is consistent with a limit on which clinical distinctions these models reliably select, even among documented concepts. The nWMH effect persists after class-size normalization, volume reversal, positive descriptions, and VoxTell component matching, and replicates on the untouched official test. Both MS lesions and white matter hyperintensities are documented VoxTell training concepts, so the substitution cannot be explained by a missing label: the model represents both categories yet resolves nWMH prompts, and the generic WMH prompt itself, to MS lesions. Our observations do not establish a causal mechanism or generality across other concepts.

We use open-ontology gap for the mismatch between lexical flexibility and reliable visual concept selection. Concept substitution describes a requested subclass yielding a mask preferentially aligned with a competing concept or superclass. A parsimonious alternative description is soft retrieval over learned labels: flexible text resolves to the most salient familiar concept, here MS lesions, rather than the requested one. Our observations are compatible with that account, but do not establish a literal lookup or nearest-neighbor mechanism. The contribution is to make this boundary measurable,

separate it from size and negation explanations, and confirm it on held-out patients; none of our controls establishes arbitrary compositional generalization.

The confidence result identifies a risk already recognized in surgical segmentation monitoring (Pham et al., 2026): incorrect masks can retain high probability. Here, the error concerns the selected clinical concept on a fixed image. Opposing-prompt overlap flags all official VoxTell failures, but its within-BraTS discrimination is near chance. It can reveal prompt insensitivity in this setting; it cannot certify semantic correctness or serve as a general safeguard.

The models suggest different points of failure. Medical-SAM3 compresses the relevant text distinction and is weak on MS3SEG, so language conditioning may contribute. VoxTell preserves the distinction in its text embeddings and has substantially better MS segmentation, yet still selects the competing region for nWMH prompts. Thus distinct text representations alone do not ensure that the annotation-defined subclass is correctly selected in the image.

Implications for evaluation. For a proposed clinical distinction, annotate both competing regions, run both target prompts and a superclass prompt on the same image, and inspect directional margins alongside absolute Dice. Then test positive rephrasings and size controls, document training-concept coverage to rule out a missing label, and confirm the result on held-out patients. This workflow turns a vague concern about language understanding into a measurable target-selection requirement that can also score adaptation methods. Failure detectors separately require successes and failures within each dataset and evaluation beyond their calibration cohorts.

Implications for model design. PrAda's visual prototypes and Few-Shot Concept Prompt Learning's learned embeddings offer relevant adaptation approaches (Rosi et al., 2026; Venkataramani & Sathish, 2026); whether they resolve MS/nWMH substitution remains untested. Clinical Contrast Sets supplies an endpoint for that question: improved overlap should be accompanied by correct switching between both targets. Our training-free Contrastive Vote Exclusion baseline worsened macro Dice for both models (Appendix G). The present solution is therefore an evaluation procedure; a reliable model-level repair remains future work.

7 Limitations

The decisive failure concerns one annotation-defined subclass in one dataset. Hemisphere-qualified MS adds a compositional probe, but its mixed result is not an independent second failing contrast. Our successful controls do not establish unrestricted composition, and the kidney comparison changes anatomy and modality. The published supervised result uses a different five-fold protocol; we provide no usable exact-split supervised bound. The inventory lists concept names, not annotation conventions, so VoxTell's training definition of white matter hyperintensities may differ from MS3SEG's nWMH convention; no training intervention isolates a causal mechanism. Medical-SAM3's low MS accuracy and inconclusive official component-matched interval make it corroborating evidence; VoxTell provides the stronger diagnosis. Development patients were reused during analysis design, with independent confirmation confined to the official 20-patient cohort. BraTS size-floor sensitivity, within-dataset detector analysis, and laterality intervals are post-hoc. Foreground probabilities are uncalibrated, pooled detector performance is confounded by dataset, and an all-failure MS3SEG cohort cannot establish specificity.

8 Conclusion

Clinical Contrast Sets addresses the evaluation gap between accepting a clinical phrase and selecting its intended visual target. Both models repeat the MS/nWMH substitution on the frozen official test, with 0/20 correct prompt pairs each. Wording and size controls, published learnability evidence, and documented-concept competence narrow the possible explanations, with the strongest diagnosis from VoxTell. Confidence and pooled detector scores cannot establish semantic reliability. The practical contribution is a controlled test for identifying clinical distinctions a model does not reliably select, even between documented concepts, and for evaluating whether future adaptations restore correct target selection alongside segmentation accuracy.

ETHICS STATEMENT

We use publicly released, de-identified research datasets: MS3SEG, BraTS 2023 Meningioma, and TotalSegmentator. We recruit no participants, attempt no re-identification, and make no clinical decisions from model outputs. The study evaluates a safety-relevant failure mode: plausible masks attached to the wrong clinical concept. We make no claim that the evaluated systems are ready for autonomous clinical use.

REPRODUCIBILITY STATEMENT

Appendix A lists all prompts; Appendix H documents cohort selection and protocol integrity; Appendices I–J specify detector and laterality analyses. The supplementary evidence package includes the unmodified confirmation protocol, execution manifests, patient-level official results, detector calibration rows, and code reproducing the post-hoc within-dataset audit. Development summary tables and the experiment-status ledger distinguish prior evidence, confirmation endpoints, and unsuccessful auxiliary work. The official test was opened only after the confirmation protocol was frozen; no official-test result was used to select prompts or detection thresholds.

AI USE STATEMENT

AI tools were used to assist code development and debugging, literature retrieval, and manuscript drafting/editing. The authors reviewed the final manuscript and take responsibility for all claims and content.

REFERENCES

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip H. S. Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-language models do not understand negation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 29612–29622, 2025.

Mohammed Baharoon, Luyang Luo, Michael Moritz, Abhinav Kumar, Sung Eun Kim, Xiaoman Zhang, Miao Zhu, Kent Kleinschmidt, Sri Sai Dinesh Jaliparthi, Sathvik Suryadevara, Rithvik Akula, Mark Marino, Wenhui Lei, Ibrahim Ethem Hamamci, and Pranav Rajpurkar. State-of-the-art text-prompted medical segmentation models struggle to ground chest CT findings. In Proceedings of the 10th Machine Learning for Healthcare Conference, volume 298 of Proceedings of Machine Learning Research, 2025.

Mahdi Bashiri Bawil, Mousa Shamsi, Aydin Ghalehasadi, Ali Fahmi Jafargholkhanloo, and Abolhassan Shakeri Bavil. A multiple sclerosis MRI dataset with tri-mask annotations for lesion segmentation. Scientific Data, 13:867, 2026. doi: 10.1038/s41597-026-07184-5.

Yuxin Du, Fan Bai, Tiejun Huang, and Bo Zhao. SegVol: Universal and interactive volumetric medical image segmentation. In Advances in Neural Information Processing Systems, volume 37, 2024.

Matt Gardner, Yoav Artzi, Victoria Basmova, et al. Evaluating models' local decision boundaries via contrast sets. In Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 1307–1323, 2020.

Yuchen Hou and Lin Zhao. LangGap: Diagnosing and closing the language gap in vision-language-action models. arXiv preprint arXiv:2603.00592, 2026.

Chongcong Jiang, Tianxingjian Ding, Chuhan Song, Jiachen Tu, Ziyang Yan, Yihua Shao, Zhenyi Wang, Yuzhang Shang, Tianyu Han, and Yu Tian. Medical SAM3: A foundation model for universal prompt-driven medical image segmentation. arXiv preprint arXiv:2601.10880, 2026.

Dominic LaBella, Ujjwal Baid, Omaditya Khanna, et al. Analysis of the BraTS 2023 intracranial meningioma segmentation challenge. arXiv preprint arXiv:2405.09787, 2024.

Shuang Liang, Zeqing Wang, Yuxian Li, Xihui Liu, and Han Wang. From pixels to concepts: Do segmentation models understand what they segment? arXiv preprint arXiv:2605.09591, 2026.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. Nature Communications, 15(1):654, 2024.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. CREPE: Can vision-language foundation models reason compositionally? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10910–10921, 2023.

Hieu D. Pham, Dang P. M. Cao, and Thanh Trung Huynh. Beyond uncertainty: Generalizable failure monitoring for surgical segmentation under acquisition degradation. arXiv preprint arXiv:2608.16748, 2026.

Maximilian Rokuss, Moritz Langenberg, Yannick Kirchhoff, Fabian Isensee, Benjamin Hamm, Constantin Ulrich, Sebastian Regnery, Lukas Bauer, Efthimios Katsigiannopulos, Tobias Norajitra, and Klaus Maier-Hein. VoxTell: Free-text promptable universal 3D medical image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 37538–37557, 2026.

Gabriele Rosi, Fabio Cermelli, Carlo Masone, and Barbara Caputo. PrAda: Few-shot visual adaptation for text-prompted segmentation. arXiv preprint arXiv:2605.19623, 2026.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5238–5248, 2022.

Rahul Venkataramani and Rachana Sathish. Few-shot concept prompt learning for segmentation foundation models via visual grounding. arXiv preprint arXiv:2608.01663, 2026.

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.

Yunan Zeng, Yan Huang, Jinjin Zhang, Zequn Jie, Zhenhua Chai, and Liang Wang. Investigating compositional challenges in vision-language models for visual grounding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14141–14151, 2024.

Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, et al. BiomedParse: A biomedical foundation model for image parsing of everything everywhere all at once. arXiv preprint arXiv:2405.12971, 2024.

Ziheng Zhao, Yao Zhang, Chaoyi Wu, Xiaoman Zhang, Xiao Zhou, Ya Zhang, Yanfeng Wang, and Weidi Xie. Large-vocabulary segmentation for medical images with text prompts. npj Digital Medicine, 8:566, 2025.

A Exact prompt banks

MS3SEG original prompts. MS family: “multiple sclerosis lesion,” “multiple sclerosis demyelinating lesion,” “MS-related white matter hyperintensity.” nWMH family: “normal white matter hyperintensity,” “white matter hyperintensity unrelated to multiple sclerosis,” “non-MS white matter hyperintensity.” Superclass: “white matter hyperintensity.”

MS3SEG positive descriptive prompts. Primary: “age-related or CSF-contaminated white matter hyperintensity,” “age-related white matter hyperintensity,” “nonspecific age-related white matter hyperintensity,” “white matter hyperintensity due to CSF partial volume.” Secondary: “periventricular cap.”

MS3SEG same-disease laterality control. “left-hemisphere multiple sclerosis lesion” and “right-hemisphere multiple sclerosis lesion.”

BraTS MEN. Enhancing family: "enhancing meningioma," "contrast-enhancing meningioma." Non-enhancing family: "non-enhancing meningioma tumor core," "nonenhancing meningioma core." Superclass: "meningioma."

TotalSegmentator. "left kidney," "right kidney," and superclass "kidney."

B FULL MS3SEG TARGET-OVERLAP MATRIX

Table 3 reports the complete canonical/superclass overlap values used to contextualize directional margins. In particular, Medical-SAM3's "normal WMH" prompt has higher MS Dice than its canonical MS prompt while nWMH Dice remains 0.026.

Table 3: Mean Dice and class coverage on the fixed 40-patient MS3SEG development cohort.

Model	Prompt	Dice MS	Dice nWMH	Cov. MS	Cov. nWMH
Medical-SAM3	MS canonical	.198	.026	.363	.062
Medical-SAM3	normal WMH	.239	.026	.406	.076
Medical-SAM3	non-MS WMH	.154	.017	.392	.077
Medical-SAM3	WMH superclass	.170	.018	.403	.086
VoxTell	MS canonical	.552	.177	.550	.201
VoxTell	normal WMH	.509	.179	.600	.232
VoxTell	non-MS WMH	.552	.191	.597	.229
VoxTell	WMH superclass	.541	.180	.546	.204

C VOLUME-STRATIFIED AND COMPONENT-MATCHED CONTROLS

On MS3SEG, 23/40 patients have $|\mathrm{nWMH}| \geq |\mathrm{MS}|$. In that subgroup, mean nWMH normalized direction is -0.282 [-0.341, -0.218] for Medical-SAM3 and -0.376 [-0.430, -0.321] for VoxTell. Across low, middle, and high MS:nWMH volume-ratio tertiles, all nWMH family means remain negative for both models.

For component matching, we use 18-connectivity, discard components smaller than 3 voxels, and solve a within-patient Hungarian assignment minimizing absolute log-volume ratio. The primary analysis retains matches with at most a 2:1 volume ratio. The canonical nWMH prompt yields patient-clustered mean nWMH-minus-MS coverage of -0.082 [-0.117, -0.046] over 877 pairs for Medical-SAM3 and -0.147 [-0.192, -0.097] over 878 pairs for VoxTell.

Nine of 20 official-test patients satisfy $|\mathrm{nWMH}| \geq |\mathrm{MS}|$; their nWMH directions are -0.315 [-0.404, -0.220] and -0.406 [-0.550, -0.296]. Official component matching yields -0.052 [-0.135, 0.016] over 462 pairs for Medical-SAM3 and -0.102 [-0.198, -0.010] over 458 pairs for VoxTell. The former interval crosses zero and is not claimed as confirmatory on its own.

D DESCRIPTIVE NWMH PROMPTS

All four pre-specified primary positive descriptions remain directionally negative. Medical-SAM3 normalized margins range from -0.314 to -0.329; VoxTell ranges from -0.297 to -0.353. The secondary "periventricular cap" prompt approaches zero (Medical-SAM3: -0.032; VoxTell: +0.003) but is not annotation-complete and is therefore not used as primary evidence.

On the official test, the four-description family mean is -0.312 [-0.364, -0.257] for Medical-SAM3 and -0.340 [-0.429, -0.263] for VoxTell. Relative to the generic-WMH superclass, the shifts are -0.005 [-0.018, 0.009] and +0.001 [-0.008, 0.011], respectively.

E TEXT GEOMETRY DETAILS

For Medical-SAM3 contextual-token means, the MS-related/non-MS polarity distance is 0.087, the within-MS paraphrase distance is 0.125, and the within-nWMH paraphrase distance is 0.125. For

VoxTell's Qwen3 embedding, the corresponding values are 0.120, 0.041, and 0.175. Distances from different embedding spaces are never compared numerically across models. After excluding unrelated-anatomy and external-control pairs, the text/output-distance Spearman correlations are 0.886 for Medical-SAM3 contextual-token means and 0.943 for VoxTell, each over six pre-specified related pairs; these are descriptive rather than inferential endpoints.

F External full prompt-target matrix

Table 4 reports canonical-prompt Dice against both competing targets. In BraTS, A is enhancing tumor and B is non-enhancing core; for kidneys, A is left and B is right. These are absolute-overlap competence measurements, not tests of unseen concept composition.

Table 4: External mean Dice for the requested target and competing target. For compactness, one canonical prompt per child concept is shown.

Dataset	Model	Prompt	Dice A	Dice B
BraTS	Medical-SAM3	enhancing	.816	.025
BraTS	Medical-SAM3	non-enhancing	.116	.240
BraTS	VoxTell	enhancing	.943	.033
BraTS	VoxTell	non-enhancing	.212	.360
TotalSeg	Medical-SAM3	left kidney	.706	.225
TotalSeg	Medical-SAM3	right kidney	.236	.571
TotalSeg	VoxTell	left kidney	.919	.000
TotalSeg	VoxTell	right kidney	.000	.917

G Negative post-hoc mitigation

Before the ontology analysis, we evaluated a frozen training-free Contrastive Vote Exclusion (CVE) baseline. Three MS-family and three nWMH-family masks vote at each voxel; one shared threshold $\tau \in \{1, 2, 3\}$ is selected on a 20-patient calibration split across both models/classes and evaluated once on a disjoint 20-patient split. CVE does not solve the problem. Relative to canonical prompting, macro Dice changes by -0.104 $[-0.138, -0.072]$ for Medical-SAM3 and -0.306 $[-0.341, -0.271]$ for VoxTell. We include this negative result because it argues against a trivial mask-arithmetic repair; it is not presented as a successful method.

H Protocol integrity and cohort selection

The prompt/control analysis was frozen before new descriptive and external inference, including prompt banks, target definitions, volume controls, metric definitions, and cohort-selection salts. The external cohort contains 40 BraTS MEN cases and 40 TotalSegmentator cases. BraTS selection requires at least 20 voxels in both enhancing and non-enhancing target masks and keeps at most one exam per patient. TotalSegmentator selection requires both kidney masks with at least 100 voxels before 2-mm resampling. No model output is used for cohort selection.

The executed Medical-SAM3 external inference records repository revision b610242c...07fc; VoxTell records revision d557d1a6...32f2 with PyTorch 2.8.0+cu126. The external analysis asserts prediction/ground-truth shape equality for every row, verifies RAS reorientation preserves foreground counts, and contains no execution errors in the completed notebooks.

The development analysis hash is ab2e9b3f...62fa, and the external-cohort hash is 5d5e18c0...099c. The confirmation JSON was frozen before official-test inference. Its canonical SHA256, also recorded in both model execution manifests, is: 47d8eea14c14ca535cf1962389d09287cc7bbde467bb035906cbd46383aeaf21

47d8eea14c14ca535cf1962389d09287cc7bbde467bb035906cbd46383aeaf21

Canonicalization removes the stored hash field and serializes the remaining JSON with sorted keys and compact separators before SHA256 hashing. The official cohort contains exactly the 20 pre-specified IDs, all contrast eligible; the analysis contains 440 model–patient–prompt rows for the 11 non-laterality prompts. The two laterality prompts are analyzed separately. No model, prompt, threshold, or primary endpoint was changed after official-test results were opened.

The frozen protocol also included an auxiliary supervised replication that did not yield a usable bound. It is recorded in the supplementary experiment-status ledger and is not used as learnability evidence. The within-BraTS detector audit and added laterality intervals are explicitly post-hoc analyses of existing outputs, not newly frozen confirmation endpoints.

I CONFIDENCE AND WITHIN-DATASET DETECTION

For VoxTell, mean probability inside the predicted-positive mask is 0.9658 for the canonical MS prompt and 0.9616 for the canonical nWMH prompt. The paired nWMH-minus-MS difference is -0.0042 [-0.0085, -0.0008]; the within-patient 95th percentile equals 1.000 for both prompts. These native scores are not clinically calibrated probabilities.

PromptSwapDice is the Dice overlap between opposing canonical-prompt masks. Failure means that at least one canonical target-minus-confounder coverage margin is nonpositive. Larger scores are more suspicious. On 120 calibration pairs per model, the threshold maximizes balanced accuracy over unique scores, including a candidate above the maximum; ties choose the smallest threshold. The frozen threshold is 0.1946805594 for VoxTell and 0.4064609450 for Medical-SAM3, with alerts when score $\geq$ threshold. No threshold is refitted in the within-dataset audit.

Table 5 separates pooled and within-dataset results. Average precision is the non-interpolated precision–recall summary used by the original analysis. AUROC is undefined for cohorts containing only successes or only failures; consequently, no within-MS3SEG AUROC or false-positive rate is reported. The post-hoc 95% AUROC intervals use 5,000 case-bootstrap samples stratified by failure status (seed 20260919); a direct positive–negative pair comparison independently reproduces the point estimates. BraTS is a subset of the calibration cohort, so this is a confounding audit, not an independent detector test.

Table 5: Prompt-overlap diagnostic audit. "Flag fail/succ" gives alerts among failures/successes at the unchanged model-specific threshold. AP is average precision. Dashes indicate metrics not reported for single-class cohorts. Official MS3SEG is held out; the other rows reuse calibration data.

Model	Cohort	Fail/succ	AUROC	AP	Flag fail/succ
Medical-SAM3	Pooled	78/42	.595	.771	28/6
Medical-SAM3	BraTS	28/12	.473	.759	1/0
Medical-SAM3	Kidney	10/30	.843	.778	7/6
Medical-SAM3	MS3SEG dev.	40/0	—	—	20/0
Medical-SAM3	MS3SEG official	20/0	—	—	11/0
VoxTell	Pooled	61/59	.916	.917	49/6
VoxTell	BraTS	21/19	.461	.603	9/6
VoxTell	Kidney	0/40	—	—	0/0
VoxTell	MS3SEG dev.	40/0	—	—	40/0
VoxTell	MS3SEG official	20/0	—	—	20/0

Within BraTS, VoxTell sensitivity is 9/21 = 42.9% and false-positive rate is 6/19 = 31.6%; Medical-SAM3 sensitivity is 1/28 = 3.6% with 0/12 false alerts. Neither within-BraTS AUROC interval excludes 0.5: VoxTell [0.276, 0.649], Medical-SAM3 [0.301, 0.655]. The pooled association can therefore be driven substantially by between-dataset/concept separation, and official-test alert sensitivity alone cannot validate a reliable abstention rule.

J HEMISPHERE-QUALIFIED MS GROUNDING

Nineteen of 20 official patients contain at least 20 MS-lesion voxels per hemisphere after removal of annotation overlap. The fixed split uses the midpoint of the nonzero-FLAIR bounding box in canonical RAS coordinates, not a manually delineated anatomical midline. Medical-SAM3 has mean left/right directions -0.055/ + 0.094 and 1/19 pair success; VoxTell has +0.031/ + 0.255 and 10/19. VoxTell is correctly directed for the left prompt in 11/19 cases and the right prompt in 18/19; Medical-SAM3 is correct in 7/19 and 11/19, respectively. Post-hoc patient-bootstrap intervals (5,000 samples, seed 20260919) are $[-0.045, 0.103]$/$[0.177, 0.335]$ for VoxTell and $[-0.160, 0.045]$/$[-0.019, 0.214]$ for Medical-SAM3.

The left/right pair endpoint was fixed before official-test inference; its comparison with kidney laterality is exploratory. It shows incomplete hemisphere composition on the same disease, not a uniformly negative target direction or an independent second pathology failure. Differences in region size, anatomy, and modality prevent attributing the kidney–MS contrast to any single factor, and we do not attribute it to training coverage.

K PUBLISHED TRAINING-CONCEPT AUDIT

Table 6 summarizes concepts in the published base inventory of Rokuss et al. (2026). Counts are reported instances, not independent subjects. Both concepts required for the MS3SEG contrast are listed: multiple sclerosis lesion, and white matter hyperintensities, the category to which the dataset's normal/non-MS WMH belong. Hemisphere-qualified MS lesion labels are not listed as base concepts; the 9,682 rewritten variants are not enumerated, so non-listing does not establish absence. We use the audit to document what is listed, not to explain behavioral outcomes by listing or non-listing.

Table 6: Published VoxTell base-concept audit relevant to this study.

Published base concept	**Reported count**
multiple sclerosis lesion	366
white matter hyperintensities	60
white matter lesions	30
enhancing tumor	4681
non-enhancing tumor core	1943
left kidney	5666
right kidney	5603
normal/non-MS WMH (MS3SEG nWMH)	covered by the listed white matter hyperintensities category
left/right-hemisphere MS lesion	not listed as a base label (rewritten variants not enumerated)

For Medical-SAM3, public materials report training on 33 datasets spanning 10 modalities but do not expose a complete concept-level inventory comparable to VoxTell's supplement. We therefore do not classify its tested concepts as seen or unseen.

L BRATS 200-VOXEL SENSITIVITY ANALYSIS

The primary BraTS cohort was pre-specified with a 20-voxel minimum for each child target. To assess sensitivity to very small non-enhancing cores, we repeated the analysis post hoc after requiring $|B| \geq 200$ voxels. Nineteen of 40 cases remain; their median enhancing:non-enhancing ratio is 22.53:1.

Table 7: Post-hoc BraTS sensitivity analysis with at least 200 non-enhancing voxels. Coverage and Dice values are target-minus-confounder family means with 95% patient bootstrap intervals.

Model	Family	Coverage direction	Dice direction
Medical-SAM3	enhancing	.296 [.148, .447]	.792 [.688, .867]
Medical-SAM3	non-enhancing	.495 [.353, .624]	.276 [.082, .455]
VoxTell	enhancing	.298 [.160, .448]	.887 [.858, .914]
VoxTell	non-enhancing	.592 [.400, .758]	.162 [-.109, .415]

Canonical pair success is 10/19 for Medical-SAM3 and 15/19 for VoxTell. The VoxTell non-enhancing Dice-direction interval crosses zero, so we interpret this analysis as support for directional coverage robustness, not as proof of high absolute non-enhancing segmentation accuracy.