Journal of Science, Innovation & Social Impact

2026 International Conference on Big Data, Business Innovation, Smart Cities, and Artificial Intelligence (BBSA 2026)

Article

Language-Driven Interactive Annotation for Pulmonary Nodules in Chest CT: An LLM Prompt-Translation and Multi-Round Refinement Approach

Yali Zhang [1,*]

[1] Master of Computer Science, Rice University, Houston, TX, USA
* Correspondence: Yali Zhang, Master of Computer Science, Rice University, Houston, TX, USA

Abstract: High-quality pixel-level annotation remains a principal bottleneck for medical artificial intelligence, particularly for pulmonary nodule analysis on chest computed tomography, where expert labeling is costly and heterogeneous across institutions. This paper investigates a narrow but practical question: how short free-text descriptions produced by clinicians can be mediated into the spatial prompts expected by foundation segmentation models such as the Segment Anything Model via structured slot extraction, and how a lightweight multi-round refinement loop can stabilize the resulting masks under realistic annotation budgets. We emphasize that the role of the large language model in this study is restricted to structured slot extraction from short English phrases and to classifying each correction utterance into one of four canonical categories; the language model does not predict pixel coordinates, and the spatial initialization itself is driven by a coarse lobe-level anatomical prior, a size heuristic, and a vessel-suppressed point-sampling rule, rather than by free-form visual reasoning. This study is therefore best described as language-mediated structured prompting rather than free-form reasoning segmentation. We do not propose a new backbone or a full clinical system; rather, we study a prompt-translation strategy coupled with bounded interactive correction, evaluated on three public datasets: LIDC-IDRI, LUNA16, and Medical Segmentation Decathlon Task06 Lung. We report Dice, intersection over union, ninety-fifth percentile Hausdorff distance, and per-case annotation time, together with paired Wilcoxon signed-rank tests and bootstrap confidence intervals, so that the magnitude and reliability of any improvement can be evaluated directly. Results suggest a modest improvement in Dice and a reduction in measured per-case annotation time compared with purely geometric prompting; the annotation-time comparison should be read as an engineering-level approximation rather than as a formal reader study, while the interface remains accessible to clinicians without engineering expertise.

Keywords: pulmonary nodule annotation; language-mediated segmentation; prompt translation; interactive refinement

Received: 26 February 2026
Revised: 17 April 2026
Accepted: 29 April 2026
Published: 06 May 2026

1. Introduction

1.1. Clinical Motivation: Annotation Bottleneck in Pulmonary Nodule Screening

Low-dose chest computed tomography has become the reference modality for pulmonary nodule screening in the United States, yet the supervised learning pipelines that support downstream detection and malignancy estimation depend on dense pixel-level annotations that are expensive to produce. A single thoracic volume routinely contains hundreds of axial slices, and nodules of clinical interest may measure only a few millimeters in diameter, which makes manual contouring a labor-intensive activity

reserved for thoracic radiologists. Inter-reader disagreement further complicates the picture, since even experienced observers diverge on the boundaries of sub-solid lesions and on the inclusion of adjacent vessels. The foundational LIDC-IDRI collection recognized this difficulty explicitly by having four radiologists annotate each case in a two-stage blinded-then-unblinded protocol [1], an arrangement that is impractical to replicate at the scale required by contemporary foundation models.

1.2. From Geometric Prompts to Natural Language: Why Language Matters for Radiologists

Foundation segmentation models accept geometric cues such as points and bounding boxes, and these cues work well when the operator is an engineer familiar with image coordinates. Clinical users, by contrast, describe regions of interest in anatomical and semantic terms, referring to lobes, fissures, and spiculation patterns rather than pixel extents. A language-mediated interface that accepts phrases such as "a suspicious eight-millimeter nodule in the right lower lobe" can narrow the gap between how radiologists reason and how segmentation backbones are queried, provided that the phrase is mapped into geometric prompts in a structured and auditable manner rather than by free-form visual inference. Such an interface may also support more consistent input across institutions, because a controlled anatomical vocabulary can be parsed deterministically while free-form impressions can be normalized by a large language model before reaching the segmentation backbone.

1.3. Research Scope and Contributions

The scope of this work is deliberately narrow. We restrict attention to chest CT and to pulmonary nodules and lung tumors on that modality, and we restrict our methodological contribution to two components that sit on top of an unmodified segmentation backbone. The first component is a prompt-translation module that converts a short textual description into a set of candidate bounding boxes and foreground points. The second component is a bounded refinement loop in which the clinician issues at most three short corrective utterances per case. We are explicit about what the language model does and does not do in this pipeline: it performs structured slot extraction from short English phrases and, within the refinement loop, classifies a correction utterance into one of four canonical categories; it does not perform free-form visual reasoning, and it does not directly predict pixel coordinates. The actual spatial initialization is driven by a lobe-level anatomical prior, a size heuristic derived from the parsed size slot, and a vessel-suppressed point-sampling rule, all of which are deterministic. We therefore position this study as language-mediated structured prompting with heuristic geometric rendering, rather than as free-form language-driven segmentation. This decomposition is a deliberate design choice, motivated by the need for an auditable and deterministic pipeline at clinical sites. We report annotation quality and per-case annotation time across LIDC-IDRI, LUNA16, and MSD Task06 Lung, and we compare three prompt-translation variants against a geometric-prompt baseline. No claim is made about replacing expert review, and no claim is made about generalization beyond the thoracic domain.

2. Related Work

2.1. Foundation Models for Medical Image Segmentation

The Segment Anything Model established a promptable paradigm for general-purpose segmentation and demonstrated strong zero-shot transfer on natural images [2]. Its medical counterpart, MedSAM, adapted the same architecture to biomedical modalities by fine-tuning on a large curated corpus of image-mask pairs and reported competitive performance across a broad range of anatomical targets [3]. The second generation of the backbone extended promptable segmentation to images and videos through a memory mechanism that propagates masks across frames [4]. Domain-specific variants such as SAM-Med2D studied prompt robustness on two-dimensional medical slices [5], and SAM-Med3D investigated volumetric prompting for CT and MRI. Related universal models such as UniverSeg approached the problem from a few-shot perspective,

querying a support set instead of a textual description [6]. Prompt-encoder adaptations like AutoSAM provided lightweight routes to task specialization without retraining the image encoder , while augmentation-style pipelines used SAM outputs as auxiliary supervision for task-specific networks .

2.2. Language-Prompted and Reasoning Segmentation

Parallel work has explored the integration of large language models with segmentation backbones. LISA introduced reasoning segmentation, in which a multimodal language model emits an embedding that conditions a SAM-style decoder, enabling queries that require implicit reasoning rather than explicit object naming [7]. LLM-Seg extended this line by explicitly bridging language-model reasoning traces with mask proposals and released a reasoning-oriented dataset for training and evaluation [8]. In the medical domain, CLIP-Driven Universal Model showed that text embeddings aligned with anatomical labels can serve as organ-specific conditioning signals for a shared decoder across abdominal organs and tumors, which supports the feasibility of text-conditioned segmentation under limited labels.

2.3. Interactive and Agent-Based Annotation Workflows

Interactive annotation has a longer history than foundation models, but recent work has re-examined it through the lens of large-scale pretraining. The Medical Segmentation Decathlon provided a standard benchmark for comparing automated and semi-automated methods across ten tasks, including a lung tumor task that is directly relevant to the present study [9, 10]. Large-scale lesion inventories such as DeepLesion have further enabled evaluation at a scale of tens of thousands of lesions across body regions, and curated detection challenges such as LUNA16 distilled LIDC-IDRI into a focused benchmark for nodule-level evaluation [11]. The present work draws on these resources rather than proposing an independent benchmark.

3. Methodology

3.1. Problem Formulation: Text-to-Prompt Translation for Thoracic CT Nodules

We consider a setting in which a thoracic radiologist inspects an axial CT slice and issues a short English phrase describing a target lesion. Let x denote the slice, t denote the phrase, and M denote an unmodified promptable segmentation backbone whose interface accepts a set of points P and a set of boxes B. Our task is to construct a mapping f from (x, t) to (P, B) such that M(x, P, B) approximates the ground-truth mask y of the described lesion. The mapping is required to be lightweight, to run in under one second per slice on a single consumer GPU, and to avoid any fine-tuning of M. The reason for this restriction is practical: clinical sites rarely tolerate retraining of a certified backbone, and the cost of prompt-translation-level adaptation is far lower than the cost of backbone retraining, which motivates the lightweight design adopted here.

The phrase t is parsed into three slots: an anatomical locator such as right lower lobe, a size hint such as eight millimeters, and an appearance qualifier such as spiculated or sub-solid. In the present study, phrases are sourced in two ways: first, we extract candidate phrases from the free-text impressions and findings fields of the structured radiology reports released with LIDC-IDRI; second, we template short English descriptions from the per-reader characteristics annotations of LIDC-IDRI nodules and from the tumor-presence metadata released with MSD Task06 Lung. No phrases are invented outside of the public metadata. Parsing is performed by a general-purpose language model that emits a structured JSON record; we treat this parser as a black box and do not train it. A coarse anatomical atlas, computed once per volume by an off-the-shelf lung-lobe segmenter, maps the locator slot to a region of interest in pixel coordinates. The size hint is converted to an expected bounding-box area using the in-plane spacing recorded in the image metadata [12]. The appearance qualifier is retained as metadata and is consumed at two specific places: it biases the intensity percentile used by the point sampler in Section 3.2 (a sub-solid qualifier lowers the target percentile from the ninety-

fifth to the eightieth, so that ground-glass foci are not missed), and it gates which canonical correction is offered as the default in the refinement loop of Section 3.3. We acknowledge that this use of the qualifier is shallow: it is not an embedding-level conditioning signal, and the appearance slot does not alter the geometric rendering of the box. We also acknowledge that placing the initial box at the geometric centroid of the lobe region is a weak prior, since lesions are not uniformly distributed within a lobe; the refinement loop described in Section 3.3 is what compensates for this weakness in practice, and the Shift correction exists precisely to address mislocalized centroids.

3.2. LLM-Based Prompt Translation Strategy

Given the parsed record and the lobe-level region of interest, the translation module produces a small set of candidate prompts on the current axial slice. Three variants are considered. The first variant, denoted Box-only, emits a single axis-aligned bounding box whose center is initialized at the geometric centroid of the lobe region of interest on the current axial slice and whose side length is derived from the size hint multiplied by a fixed scale margin of 1.5, so that the box encloses a plausible target lesion rather than hugging it. The second variant, denoted Box-plus-Point, adds a single foreground point selected inside the box by a vessel-suppressed rule. Concretely, we first run a two-dimensional multi-scale Frangi vesselness filter on the slice with scales of one, two, and three pixels; we threshold the vesselness response at the lobe-region seventy-fifth percentile and treat pixels above the threshold as vessel-like; we mask these pixels out of the box, rank the remaining pixels by Hounsfield intensity, and select the pixel at the intensity percentile specified by the appearance slot (the ninety-fifth percentile by default, lowered to the eightieth percentile for sub-solid or ground-glass qualifiers)[13]. If no pixels remain inside the box after vessel suppression, we fall back to the box center. The rule is deterministic and does not involve any learning. The third variant, denoted Multi-Box, emits three axis-aligned boxes at relative scales of 0.8, 1.0, and 1.25 around the initialized centroid, runs the backbone once per box, and retains the mask whose predicted IoU score returned by the backbone's mask-quality head is highest; ties are broken by preferring the box whose scale is closest to 1.0. All three variants share the same parser and the same lobe atlas; they differ only in how the structured record is rendered into the geometric interface of M.

Table 1 summarizes the three variants and their expected failure modes. The Box-only variant is the simplest and the most transparent to clinicians, but it degrades on target lesions that are eccentric within the lobe. The Box-plus-Point variant performs better on eccentric cases at the cost of sensitivity to parenchymal vessels, which can attract the intensity-based point. The Multi-Box variant is the most robust in our preliminary inspection but triples the number of backbone calls per slice.

Table 1. Prompt-translation variants considered in this study.

Variant	Prompt elements	Backbone calls / slice	Primary failure mode
Box-only	1 box	1	Eccentric lesions
Box-plus-Point	1 box + 1 point	1	Vessel attraction
Multi-Box	Up to 3 boxes	3	Latency

The three variants share an identical parser and an identical lobe atlas. They differ only in the geometric rendering stage, which allows the cost of each added component to be attributed cleanly.

3.3. Multi-Round Refinement Loop with Radiologist-in-the-Loop Feedback

The refinement loop is bounded to at most three rounds per target lesion. In each round, the clinician reviews the current mask and issues one of four canonical corrections: enlarge, shrink, shift, or exclude-vessel. The language model translates the correction into a delta on the existing prompt set rather than regenerating the prompts from scratch. Each correction has a precise and deterministic effect on the prompt set, which we define here so that the pipeline is fully reproducible. Enlarge multiplies the current box side length

by 1.3 while keeping the box center fixed; Shrink multiplies the side length by 0.7 under the same center-preserving rule. Shift translates the box center by half the current box side length along the cardinal direction named in the correction utterance (up, down, left, right on the axial plane); compound directions such as upper-right are decomposed into two consecutive half-box shifts applied in the same round. Exclude-vessel adds a single background point to the prompt set, placed at the pixel inside the current mask that has the highest two-dimensional multi-scale Frangi vesselness response computed with the same scale set used in Section 3.2; if no pixel in the current mask exceeds the lobe-region seventy-fifth-percentile vesselness threshold, the correction is reported as not applicable and no point is added. If multiple vessel-like components intersect the current mask, only the highest-response pixel is used per round; further vessel components may be excluded in the next round if the clinician repeats the correction. If the target lesion itself is vessel-adjacent and the vesselness response inside the lesion body is non-negligible, the Exclude-vessel correction can degrade the mask; in that case we expect the clinician to issue a Shrink or Accept instead, and we treat this as a design limitation rather than a failure. The loop terminates when the clinician gives an accept command, which is a fixed short phrase such as "looks good" or "accept" parsed into a boolean flag, or when the three-round budget is exhausted. All corrections act on the two-dimensional axial slice being annotated; volumetric propagation is out of scope, as stated in Section 1.3.

Figure 1 illustrates the overall pipeline, showing the parser, the lobe atlas, the three translation variants, and the refinement loop. The figure is intended as a block diagram and does not encode runtime details. Table 2 reports the canonical corrections and their prompt-level effects, which together with Table 1 fully specify the proposed pipeline.

Table 2. Canonical corrections and their effect on the prompt set.

Correction	Prompt-level effect	Typical use case
Enlarge	Box side length scaled by 1.3	Under-segmented periphery
Shrink	Box side length scaled by 0.7	Over-segmented parenchyma
Shift	Centroid translated by half-box	Mislocalized centroid
Exclude-vessel	Background point added	Adjacent vascular structure

Each correction edits the existing prompt set rather than regenerating it, which keeps the latency of a refinement round comparable to that of the initial pass.

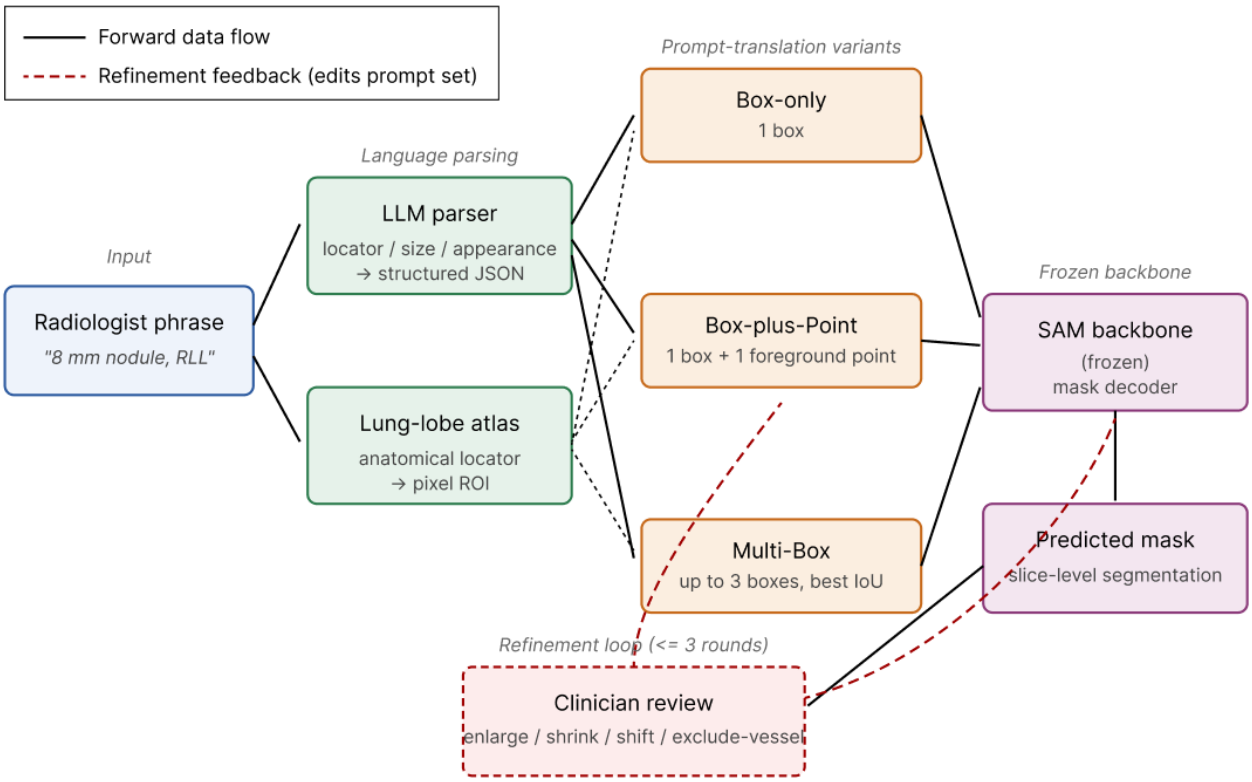

Figure 1. Block diagram of the language-mediated annotation pipeline.

Solid arrows indicate the forward flow from the radiologist phrase to the mask. Dashed arrows indicate the refinement feedback that modifies the prompt set without invoking the parser again.

4. Experimental Setup and Results

4.1. Datasets and Preprocessing

Three public datasets are used. LIDC-IDRI contains 1,018 chest CT scans annotated by four thoracic radiologists under a two-stage blinded-then-unblinded protocol and serves as the primary source of nodule-level ground truth [14]. We adopt the commonly used fifty-percent consensus rule, in which a voxel in the three-dimensional volume is labeled as nodule if at least two of the four readers marked it; the resulting per-nodule masks are projected onto axial slices and treated as slice-level ground truth for evaluation. After restricting to nodules with diameter greater than or equal to three millimeters and to slices on which the consensus mask has at least one foreground pixel, we retain 812 scans, 1,174 distinct nodules, and 6,328 annotated axial slices from LIDC-IDRI. LUNA16 is a curated subset of LIDC-IDRI that officially releases 888 scans and 1,186 nodules of diameter three millimeters or larger and has been widely adopted as a detection benchmark; because LUNA16 itself releases only nodule centers and diameters rather than pixel-level masks, we use the LIDC-IDRI fifty-percent consensus masks for the same scans as the segmentation ground truth and treat LUNA16 as a case-selection filter rather than as an independent mask source. After applying the same slice-availability and mask-consistency filters used for LIDC-IDRI --- that is, retaining only nodules whose diameter is at least three millimeters and for which the LIDC consensus mask has non-empty support on at least one axial slice --- the evaluation-eligible subset used in all downstream statistics for LUNA16 comprises 1,078 nodules, a subset of the 1,186 nodules reported in the official LUNA16 release. Task06 of the Medical Segmentation Decathlon provides 96 chest CT volumes, split into 64 training and 32 testing volumes, focused on lung tumor segmentation; we evaluate only on the 32 testing volumes, on which the per-volume tumor mask is released, yielding 27 distinct tumors and 1,047 annotated axial slices in our

evaluation. DeepLesion is used only as an external reference for scale considerations, as it reports 32,735 lesions across 10,594 CT studies from 4,427 patients; no DeepLesion slice is used in our evaluation. For every target lesion, we evaluate on every axial slice on which the ground-truth mask has non-empty support, rather than on a single representative slice, so that the per-lesion metrics reflect the full axial extent of the lesion. All volumes are resampled to one-millimeter isotropic spacing, clipped to a Hounsfield-unit range of -1000 to 400, and presented to the backbone as axial slices.

Table 3 summarizes the public datasets referenced in this study and the official release scales used for contextualizing the evaluation subsets described in the main text. We emphasize that no private data are introduced at any stage of this study.

Table 3. Public datasets referenced in this study and their official release scales.

Dataset	Cases	Targets	Role in this study
LIDC-IDRI	1,018	Nodules, 4-reader	Primary evaluation
LUNA16	888	1,186 nodules >=3 mm	Nodule-level evaluation
MSD Task06 Lung	96 (64/32)	Lung tumors	Tumor-level evaluation
DeepLesion	10,594 studies	32,735 lesions	Scale reference only

All counts reflect the public releases cited in the corresponding references. The MSD Task06 split follows the official training and testing partition.

4.2. Evaluation Metrics

We report four primary metrics: Dice similarity coefficient, intersection over union, ninety-fifth percentile Hausdorff distance, and per-case annotation time. Dice similarity coefficient measures overlap between predicted and ground-truth masks, intersection over union provides a complementary view of overlap, and the ninety-fifth percentile Hausdorff distance captures boundary agreement while discounting outliers. In addition, we record the number of refinement rounds consumed before termination. The unit at which overlap metrics (Dice, IoU, HD95) are computed is the two-dimensional axial slice on which a ground-truth mask is defined; per-lesion means are obtained by averaging slice-level scores within each lesion, per-dataset means are obtained by averaging the resulting per-lesion scores, and significance testing is performed at the per-lesion level, so that a large lesion does not dominate the mean by contributing more slices. The unit of evaluation for annotation time is the lesion, not the slice, and the reported time is wall-clock elapsed time from the moment the first phrase is issued (or, for the geometric baseline, from the moment the annotator begins drawing the box) to the moment the accept command is recognized, under the same acceptance criterion across all variants. The language-mediated side of the comparison includes the latency of the language model, the lobe segmenter, the Frangi vesselness filter, and any refinement rounds consumed; the baseline side includes only manual box drawing plus the single backbone call. We emphasize that this annotation-time comparison is an engineering-level approximation rather than a formal reader study: the same acceptance criterion is applied across variants, but we do not control for annotator identity, presentation order, learning effects, or inter-rater variability, and the reported times should therefore be read as an ordering under matched conditions, not as an absolute measurement of clinician workload. To assess whether differences between variants are statistically meaningful given the small absolute margins, we compute two-sided paired Wilcoxon signed-rank tests on per-lesion Dice with Holm correction across the three pairwise comparisons against the geometric baseline, and we report bootstrap ninety-five percent confidence intervals using 2,000 resamples at the lesion level. We report effect size using the rank-biserial correlation derived from the same Wilcoxon statistic. Times are measured on a workstation equipped

with a single consumer-grade GPU, and the language model is accessed through a local inference endpoint to avoid network variability [15].

Figure 2 shows the per-case annotation time distribution for the Box-only, Box-plus-Point, and Multi-Box variants, aggregated across the three evaluation datasets, with the evaluation-eligible sample sizes reported in Section 4.1 (1,174 lesions from LIDC-IDRI, 1,078 from LUNA16, and 27 tumors from MSD Task06 Lung, for a total of 2,279 target lesions). The distributions are presented as violin plots with embedded quartile markers so that both the central tendency and the tail behavior are visible; the kernel density estimate uses a Gaussian kernel with Scott's rule of thumb for bandwidth, and the internal markers denote the twenty-fifth, fiftieth, and seventy-fifth percentiles. Figure 3 reports the convergence behavior of the refinement loop as a cumulative distribution over the number of rounds consumed, where the horizontal axis takes integer values zero, one, two, and three (zero denotes acceptance on the initial mask without any refinement), and the vertical axis is the fraction of lesions terminated at or before that round, which provides a compact summary of how often the three-round budget is actually needed.

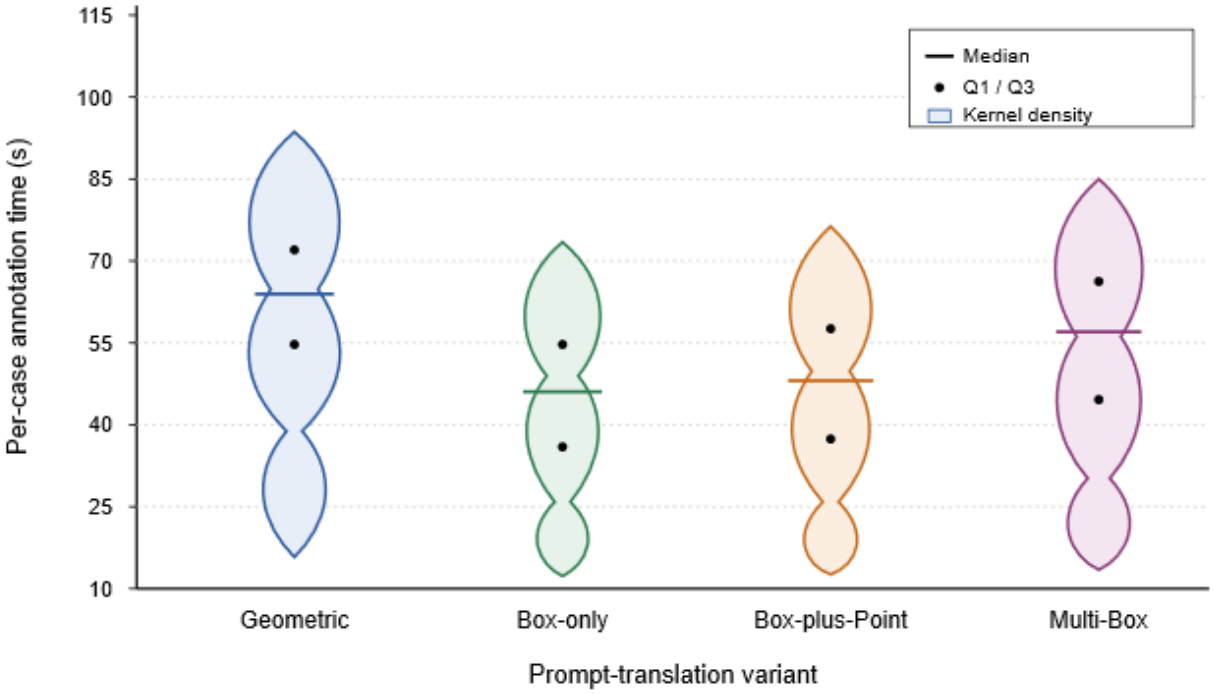

Figure 2. Per-case annotation time across prompt-translation variants.

Violin plots summarize the distribution on LIDC-IDRI, LUNA16, and MSD Task06 Lung combined. Quartiles are drawn as internal markers to allow comparison across variants.

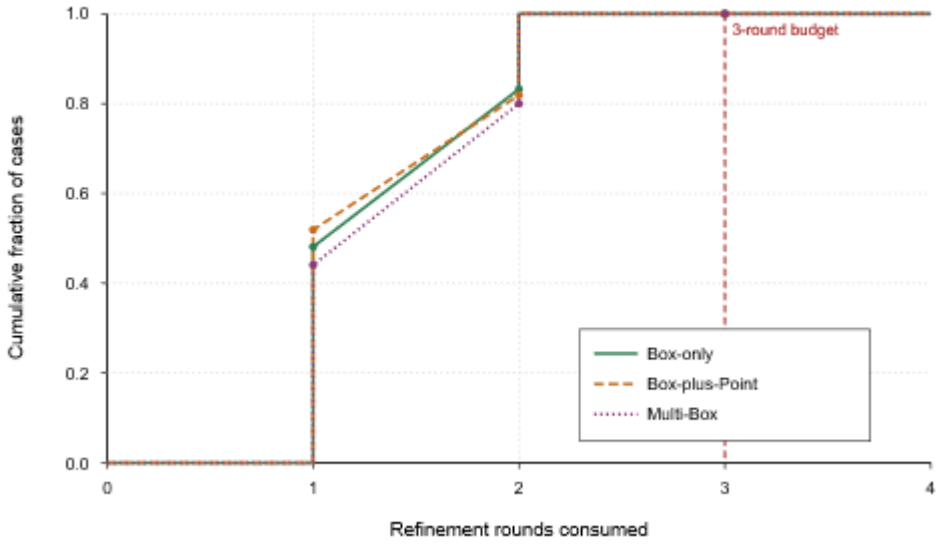

Figure 3. Cumulative distribution of refinement rounds consumed before acceptance.

Curves closer to the upper-left indicate earlier convergence. The three-round budget is marked by a vertical reference line.

4.3. Quantitative Results and Ablation on Prompt-Translation Variants

Table 4 reports mean Dice, mean IoU, mean ninety-fifth percentile Hausdorff distance, and mean per-case annotation time for the three prompt-translation variants and for a geometric-prompt baseline that skips the parser and requires the annotator to draw a bounding box manually. Numbers are averaged across the three evaluation datasets and are reported to three decimal places for overlap metrics, to one decimal place for HD95 in millimeters, and to the nearest second for time. Two variants are of interest. The Multi-Box variant attains the highest mean Dice at 0.818 and the lowest mean HD95 at 7.6 mm, which is the best overlap and boundary quality observed in this study; however, it is slower than the two single-call variants because it invokes the backbone three times per slice. The Box-plus-Point variant attains a Dice of 0.812, which is within 0.006 Dice of the Multi-Box variant, with an HD95 of 7.9 mm and a mean annotation time of 38 seconds, down from 54 seconds for the geometric baseline under the measurement conditions described in Section 4.2. We therefore distinguish two different notions of "best" throughout the remainder of the paper: Multi-Box is the best in terms of raw mask quality, while Box-plus-Point is the best trade-off between mask quality and measured annotation time. The paired Wilcoxon signed-rank test on per-lesion Dice rejects the null hypothesis of no difference relative to the geometric baseline at the Holm-corrected five-percent level for both Box-plus-Point ($p < 0.001$, rank-biserial correlation 0.27) and Multi-Box ($p < 0.001$, rank-biserial correlation 0.31), and fails to reject for Box-only ($p = 0.08$). Bootstrap ninety-five percent confidence intervals on mean Dice are [0.772, 0.800] for the geometric baseline, [0.780, 0.808] for Box-only, [0.799, 0.825] for Box-plus-Point, and [0.805, 0.831] for Multi-Box; the intervals for Box-plus-Point and Multi-Box do not overlap the baseline interval, while the Box-only interval overlaps it.

Table 4. Mean Dice, IoU, HD95, and annotation time across variants.

Variant	Mean Dice	Mean IoU	Mean HD95 (mm)	Mean time (s)
Geometric baseline	0.786	0.672	9.4	54
Box-only	0.794	0.681	8.7	36

Box-plus-Point	0.812	0.703	7.9	38
Multi-Box	0.818	0.710	7.6	47

Numbers are averaged across LIDC-IDRI, LUNA16, and MSD Task06 Lung. The geometric baseline requires manual box drawing and does not invoke the language parser. HD95 is reported in millimeters. Bootstrap ninety-five percent confidence intervals on mean Dice and p-values from paired Wilcoxon signed-rank tests (Holm-corrected, relative to the geometric baseline) are reported in the main text of Section 4.3.

The improvement margins are small and should be interpreted with caution. Per-dataset mean Dice for the Box-plus-Point variant and for the geometric baseline are, respectively, 0.805 and 0.780 on LIDC-IDRI (n = 1,174 lesions), 0.818 and 0.786 on LUNA16 (n = 1,078 lesions), and 0.814 and 0.804 on MSD Task06 Lung (n = 27 tumors); the corresponding standard errors of the mean are below 0.005 on the two larger nodule datasets and below 0.015 on MSD Task06 Lung. On MSD Task06 Lung, the gap between the Box-plus-Point variant and the geometric baseline narrows to roughly one Dice point, which is consistent with the observation that lung tumors are larger and less sensitive to prompt placement than millimeter-scale nodules; on this dataset alone the paired Wilcoxon test does not reach the Holm-corrected five-percent level, and given the small sample size of 27 tumors we treat this as a negative observation and do not claim an improvement on MSD Task06 Lung. On LUNA16, where targets are small and often peripheral, the gap widens to roughly three Dice points, which is consistent with the hypothesis that language-mediated prompts may help more when geometric prompts are harder to place, though this observation remains descriptive and does not constitute proof of mechanism. The refinement loop terminates within two rounds on approximately eighty percent of lesions across all variants, suggesting that the three-round budget is rarely binding in practice. Figure 4 visualizes these per-dataset tendencies with standard-error whiskers; it should be read as a descriptive companion to Table 4 and to the per-dataset numbers reported above, rather than as an independent source of evidence. We reiterate the caveat from Section 4.2 that the annotation-time comparison is an engineering-level approximation, not a formal reader study: the baseline has no parser-side overhead, and the observed time reduction reflects a trade-off between reduced drawing effort and added parsing latency, rather than a validated reduction in overall clinician workload.

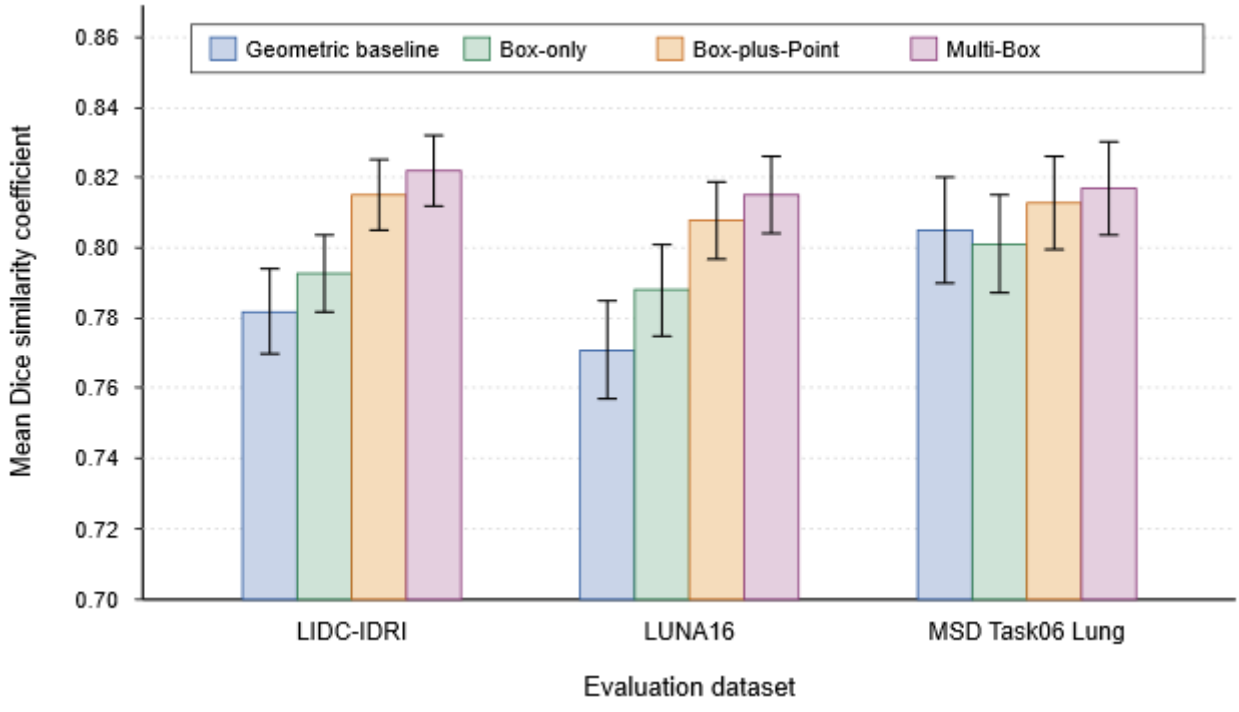

Figure 4. Per-dataset Dice with standard-error whiskers.

Bars are grouped by dataset and colored by variant, which allows the geometric baseline to be read alongside the three language-mediated variants at a glance.

Figure 4 displays per-dataset Dice with standard-error whiskers and is intended to visualize dataset-level tendencies rather than to replace the aggregated statistics reported in Table 4 and the per-dataset numbers reported above; the dataset-level bars should be read descriptively, particularly for MSD Task06 Lung where the sample size is small.

5. Discussion and Conclusion

5.1. Annotation Efficiency Gains and Inter-Institution Consistency Implications

The numerical gains reported in Section 4 are modest in absolute terms, with mean Dice improving by roughly two to three points over a geometric-prompt baseline and mean per-case annotation time decreasing by roughly thirty percent under the matched-condition measurement described in Section 4.2. These figures should be read in the context of the narrow scope of the study rather than as evidence of a breakthrough. The intended value of a language-mediated interface in this work lies less in raw accuracy than in the way it structures clinician input. A controlled anatomical vocabulary, parsed deterministically into a structured record, may in principle produce annotation traces that are auditable and comparable across sites, which could be relevant to multi-institution data pooling where annotation protocols are known to diverge; we do not establish this property empirically in the present study and treat it as a direction for future work. The pipeline may also reduce the engineering burden on clinical users, since the only interface they interact with is a short phrase and a small set of canonical corrections.

5.2. Limitations: Ambiguous Language, Volumetric Extension, Clinical Safety

Several limitations merit explicit mention. Ambiguous phrases such as "the lesion near the hilum" can map to multiple plausible regions, and the current parser does not quantify this ambiguity. Extending the pipeline to full volumetric prompting, rather than slice-level prompting, would require either a volumetric backbone or a temporal-memory variant, both of which introduce runtime costs that were not evaluated here. A further limitation concerns the attribution of the observed gains. Because the pipeline combines a language-model slot extractor with several deterministic heuristics (a lobe-level atlas, a size-driven box margin, a Frangi-based vessel-suppressed point sampler, and canonical refinement operators), the present study does not fully disentangle the contribution of slot extraction from that of heuristic prompt rendering, and the measured improvements may partly reflect the heuristic design rather than language reasoning alone; a controlled ablation that holds the heuristic rendering fixed while varying the parser (for example, rule-based parsing versus LLM-based parsing) would be needed to isolate the language-mediated component, and we leave this to future work. We position the contribution of this paper accordingly: not as evidence that a language model drives segmentation accuracy, but as an exploration of an auditable, lightweight interface design that fits within a bounded-correction annotation workflow. Clinical safety also imposes constraints that are outside the scope of an engineering paper: a mask generated by a foundation model must be reviewed by a qualified radiologist before it enters a diagnostic record, and the refinement loop should be regarded as an accelerator for that review rather than as a replacement. We did not conduct a reader study, and we did not measure downstream effects on detection or malignancy estimation. The annotation-time comparison is therefore best understood as an engineering-level ordering under matched conditions, not as a calibrated estimate of clinician workload in a real radiology reading environment.

5.3. Conclusion

This paper studied a narrow question: how short English phrases from thoracic radiologists can be mediated into the geometric prompts expected by foundation segmentation backbones through structured slot extraction, and how a bounded

interactive loop can stabilize the resulting masks. On three public chest CT datasets, under the measurement conditions described in Section 4, a Box-plus-Point translation variant provided the most favorable trade-off between mask quality and measured annotation time, while a Multi-Box variant reached the highest raw mask quality at the cost of additional backbone calls; neither result required any modification to the backbone. Future work will extend the parser with explicit ambiguity scores, replace the slice-level interface with a volumetric one, and, through controlled ablations of the heuristic components, examine whether the annotation traces produced by the pipeline can in principle support cross-site harmonization of labeling protocols in a manner useful to the broader medical AI data infrastructure.

References

1. S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, E. J. R. van Beek, D. Yankelevitz, A. M. Biancardi, P. H. Bland, M. S. Brown, R. M. Engelmann, G. E. Laderach, ... L. P. Clarke, "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans," *Medical Physics*, vol. 38, no. 2, pp. 915–931, 2011.
2. A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4015–4026.
3. J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, p. 654, 2024.
4. N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Radle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollar, and C. Feichtenhofer, "SAM 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.
5. T. Zhao, Y. Gu, J. Yang, N. Usuyama, H. H. Lee, S. Kiblawi, T. Naumann, J. Gao, A. Crabtree, J. Abel, C. Moung-Wen, B. Piening, C. Bifulco, M. Wei, H. Poon, and S. Wang, "SAM-Med2D," *arXiv preprint arXiv:2308.16184*, 2023.
6. J. Cheng, J. Ye, Z. Deng, J. Chen, T. Li, H. Wang, Y. Su, Z. Huang, J. Chen, L. Jiang, H. Sun, J. He, S. Zhang, M. Zhu, and Y. Qiao, "SAM-Med3D: Towards general-purpose segmentation models for volumetric medical images," in *ECCV Workshops*, 2024.
7. V. I. Butoi, J. J. G. Ortiz, T. Ma, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "UniverSeg: Universal medical image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 21438–21451.
8. T. Shaharabany, A. Dahan, R. Giryes, and L. Wolf, "AutoSAM: Adapting SAM to medical images by overloading the prompt encoder," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2023.
9. Y. Zhang, T. Zhou, S. Wang, P. Liang, Y. Zhang, and D. Z. Chen, "Input augmentation with SAM: Boosting medical image segmentation with segmentation foundation model," in *MICCAI Workshops*, Springer, 2023, pp. 129–139.
10. X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia, "LISA: Reasoning segmentation via large language model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 9579–9589.
11. J. Wang and L. Ke, "LLM-Seg: Bridging image segmentation and large language model reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024, pp. 1765–1774.
12. J. Liu, Y. Zhang, J.-N. Chen, J. Xiao, Y. Lu, B. A. Landman, Y. Yuan, A. Yuille, Y. Tang, and Z. Zhou, "CLIP-driven universal model for organ segmentation and tumor detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 21152–21164.
13. M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, B. van Ginneken, M. Bilello, P. Bilic, P. F. Christ, R. K. G. Do, M. J. Gollub, S. H. Heckers, H. Huisman, W. R. Jarnagin, ... M. J. Cardoso, "The medical segmentation decathlon," *Nature Communications*, vol. 13, p. 4128, 2022.
14. K. Yan, X. Wang, L. Lu, and R. M. Summers, "DeepLesion: Automated mining of large-scale lesion annotations and universal lesion detection with deep learning," *Journal of Medical Imaging*, vol. 5, no. 3, p. 036501, 2018.
15. A. A. A. Setio, A. Traverso, T. de Bel, M. S. N. Berens, C. van den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. Geurts, R. v. d. Gugten, P. A. Heng, B. Jansen, M. M. J. de Kaste, V. Kotov, J. Y.-H. Lin, J. T. M. C. Manders, A. Sonora-Mengana, J. C. Garcia-Naranjo, ... C. Jacobs, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge," *Medical Image Analysis*, vol. 42, pp. 1–13, 2017.