# GLeVE: Graph-Guided Lesion Grounding with Proposal Verification in 3D CT

Shuo Jiang[1†], Yuhao Hong[1†], Chunbo Jiang[1], Weihong Chen[1]
Huangwei Chen[2], Shenghao Zhu[1], Beining Wu[1], Mingxuan Liu[3], Zhu Zhu[4]
Feiwei Qin[1⋆], Min Tan[1], and Yifei Chen[3⋆]

[1] Zhejiang Key Laboratory of Space Information Sensing and Transmission, Hangzhou Dianzi University, Hangzhou, China
[2] Zhejiang University, Hangzhou, China
[3] Tsinghua University, Beijing, China
[4] Children's Hospital, Zhejiang University School of Medicine, Hangzhou, China
{jiangshuo, qinfeiwei}@hdu.edu.cn, justlfc03@gmail.com

**Abstract.** Grounding radiology report descriptions to 3D CT volumes is essential for verifiable clinical interpretation, yet remains challenging due to the semantic-spatial gap between free-text narratives and volumetric anatomy. Existing report-assisted and vision-language grounding methods typically rely on phrase-level alignment or dense pixel supervision, resulting in limited lesion-wise correspondence and suboptimal localization accuracy. We propose GLeVE, a graph-guided lesion grounding framework with anatomical prior verification and octree-based autoregressive refinement. GLeVE treats each lesion description as an atomic semantic unit and encodes organ attribution, attributes, and inter-lesion relations through relation-aware graph reasoning to produce discriminative lesion-wise queries. Anatomy-aware proposal generation with region-level verification enforces one-to-one text-lesion alignment, while hierarchical octree refinement progressively improves boundary delineation. Experiments on AbdomenAtlas 3.0 demonstrate consistent gains over classical multimodal foundation models and report-supervised baselines in both segmentation accuracy and lesion-level localization. Our source code is available at https://github.com/JSLiam94/GLeVE.

**Keywords:** Report Grounding · Lesion Localization · Graph Reasoning · Anatomical Prior · Octree Refinement

## 1 Introduction

Radiology reports provide structured natural-language descriptions of lesions in 3D Computed Tomography (CT), covering organ attribution, anatomical subregions, and key imaging characteristics such as Hounsfield Units (HU), and thus serve as a central interface between imaging evidence and clinical decision-making [19,8,4]. However, in routine clinical workflows, physicians must repeatedly revisit the original CT and search slice by slice around the reported locations

⋆ Corresponding authors. † Co-first authors.

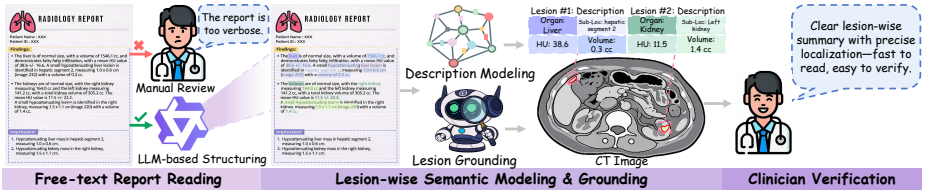


**Fig. 1.** Overview of lesion-wise report grounding, illustrating structured semantic modeling and precise CT localization for efficient clinical verification.

to verify critical findings, making lesion localization highly experience-dependent and inefficient [20,11,3]. This burden is further amplified in multi-organ/multi-lesion scenarios: reports grow substantially longer, salient lesion descriptions are diluted by abundant routine statements, and clinicians are hindered in rapidly retrieving key information from lengthy text and stably mapping it to the corresponding image locations, thereby limiting the practical utility of reports for precise decision-making and verifiable interpretation [1,26,9].

Existing general-purpose segmentation models (e.g., SwinUNETR [25] and MedSAM3 [21]) produce global CT masks, hindering interpretable lesion-by-lesion report alignment and requiring dense pixel-level supervision [10]. R-Super [5] uses reports as auxiliary supervision to reduce annotation reliance, but without explicit inference-time matching, it struggles to enforce one-to-one "text-lesion" correspondence. Studies on medical text grounding aim to align report descriptions into the global image coordinate system for explainable localization: Ichinose *et al.* [17] leverage anatomical-structure segmentation and structured report parsing to improve abnormality localization, but their reliance on cross-modal attention alignment still limits spatial precision; Nützel *et al.* [23] employ latent diffusion models for bimodal bias fusion, and Vilouras [27] performs feature selection via cross-attention maps for medical phrase grounding. Nevertheless, these methods mainly perform phrase-level alignment and cannot reliably treat a full lesion description as a coherent semantic unit for precise localization. Furthermore, Multimodal Large Language Model-based methods introduce parameter-efficient fine-tuning and uncertainty-aware multi-hypothesis mechanisms [15,32] to enhance robustness; Yang *et al.* [30] propose a contrastive learning framework with multi-level semantic alignment, but it remains highly sensitive to fine-grained report quality and shows limited cross-scenario generalization. Overall, while prior work has advanced bounding-box-level localization, substantial gaps remain in achieving reliable, pixel-accurate localization of multiple low-contrast small lesions, falling short of core clinical requirements.

To address these issues, we propose **G**raph-Guided **Le**sion Grounding with Proposal **Ve**rification framework (**GLeVE**). **GLeVE** encodes composite semantics in lesion descriptions, including organ attribution, size, and HU, through graph-structured representation learning, and incorporates organ-segmentation priors to construct geometry-aware anatomical embeddings. A multi-candidate verification mechanism is subsequently introduced to suppress boundary noise

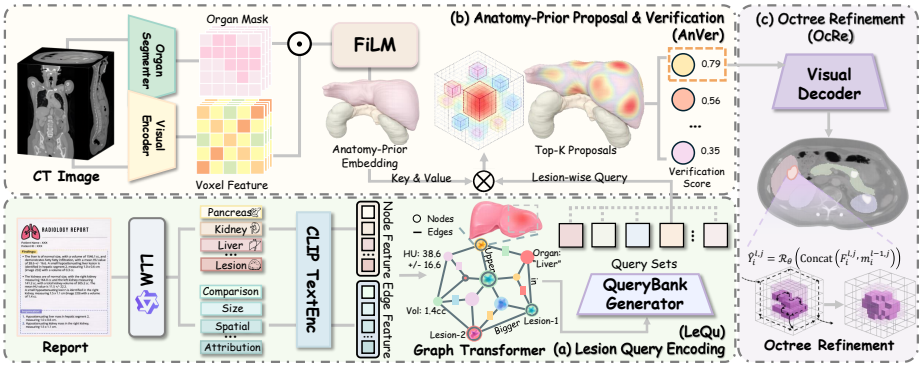


**Fig. 2.** Architecture of **GLeVE**, including (a) lesion semantic graph modeling, (b) anatomy-aware proposal verification, and (c) octree-based autoregressive refinement.

and spurious responses. This is followed by an octree-based autoregressive refinement strategy that enhances pixel-level localization sensitivity for small lesions. As illustrated in Fig. 1, by optimizing semantic consistency together with spatial precision in complex anatomy, **GLeVE** delivers verifiable pixel-level lesion localization with one-to-one alignment between key report descriptions and imaging evidence, yielding traceable, text-grounded results for reliable clinical interpretation and quantitative evaluation.

## 2 Proposed Method

### 2.1 Model Architecture

As illustrated in Fig. 2, **GLeVE** comprises three modules: (i) a graph-structured semantic query generator that converts each lesion description into lesion-wise queries while preserving organ attribution, attributes, and inter-lesion relations; (ii) anatomy-prior feature modulation with voxel-level similarity to generate high-recall candidates, followed by region-level verification for one-to-one text-lesion alignment; and (iii) octree-based autoregressive refinement within selected candidates for high-resolution segmentation, yielding interpretable masks consistent with global spatial coordinates.

### 2.2 Lesion Semantic Modeling and Querying (LeQu)

For lesion-level semantic completeness and precise cross-modal matching, we use each full lesion description as the atomic reasoning unit, encode its attribute constraints and spatial relations in an explicit graph, and generate corresponding lesion-wise queries for stable, interpretable localization in multi-lesion settings.

**Structured Report Parsing and Lesion Unit Construction.** Given a report $R$, we employ a frozen Qwen3-8B [29] to extract a lesion set $\{r_i\}$ with

structured fields (organ attribution $o(i)$, sub-location, size, and HU), and incorporate organ-level statistics to explicitly model lesion-background contrast.

**Lesion Semantic Graph Construction.** The extracted fields are organized into a semantic graph $G_i = (\mathcal{V}_i, \mathcal{E}_i)$ comprising lesion, anatomical, and attribute nodes. Edges encode lesion-organ attribution, lesion-attribute associations, and intra-organ relations, enabling cross-lesion comparison to disambiguate multi-lesion cases and enhance lesion-level localization queries.

**Node Initialization and Relation Encoding.** For each node $v \in \mathcal{V}_i$, we initialize its embedding $h_v^0 \in \mathbb{R}^d$ with a CLIP text encoder; for each edge $(u \to v) \in \mathcal{E}_i$, we assign a learnable embedding $r_{uv}$ to encode relation types.

**Relation-Aware Graph Transformer Reasoning.** We adopt a relation-type-aware attention mechanism for message passing, where relation embeddings are injected into the graph reasoning process as attention biases:

$$\begin{aligned} \alpha_{uv}^{\ell} &= \mathrm{softmax}_{u \in \mathcal{N}(v)} \left( \frac{\left(W_q h_v^{\ell-1}\right)^{\top} \left(W_k h_u^{\ell-1} + W_r r_{uv}\right)}{\sqrt{d}} \right), \\ h_v^{\ell} &= \mathrm{FFN}\left(h_v^{\ell-1}\right) + \sum_{u \in \mathcal{N}(v)} \alpha_{uv}^{\ell} \, W_v h_u^{\ell-1}, \end{aligned} \tag{1}$$

where $h_v^{\ell} \in \mathbb{R}^d$ is the $\ell$-th layer embedding of node $v$, $\mathcal{N}(v)$ denotes its neighborhood, $W_q, W_k, W_v \in \mathbb{R}^{d \times d}$ are the query/key/value projections, and $W_r \in \mathbb{R}^{d \times d_r}$ projects relation embeddings; $d$ is the feature dimension, and $\mathrm{FFN}(\cdot)$ is a residual Feed-Forward Network. This facilitates relation-aware message passing and enhances lesion-level semantic discrimination in multi-lesion settings.

**Lesion-Level Query Generation.** After graph reasoning, the final representation of the lesion node $h_{v_i^{\mathrm{les}}}^{L}$ is treated as the semantic summary vector $z_i$ for the $i$-th lesion. A QueryBank Generator ($\Psi_{\mathrm{QB}}$) then maps $z_i$ to a set of lesion queries, $\left\{q_i^{(m)}\right\}_{m=1}^{M} = \Psi_{\mathrm{QB}}(z_i)$, where $M$ denotes the number of queries associated with each lesion.

### 2.3 Anatomy-Prior Lesion Proposal and Verification (AnVer)

To incorporate anatomical constraints, we combine organ embeddings with feature modulation to render voxel features organ-aware prior to matching, thereby enforcing consistency in candidate generation and region-level verification.

**Anatomical Prior Soft Modulation.** The visual encoder produces features $F$. An organ segmenter provides binary masks $\{M_k\}_{k=1}^{K_o}$, where $K_o$ denotes the total number of organ categories, and we Gaussian-smooth them to obtain soft maps $\{\widetilde{M}_k\}_{k=1}^{K_o}$. For each organ, we learn an embedding $e_k$ and form a voxel-wise anatomy token that conditions Feature-wise Linear Modulation (FiLM):

$$\begin{aligned} \hat{E}(x) &= \phi_o\left(\sum_{k=1}^{K_o} \widetilde{M}_k(x)\, e_k\right) \in \mathbb{R}^C, \\ (\gamma(x), \beta(x)) &= \psi\Big(\hat{E}(x)\Big), \qquad \widetilde{F}(x) = \gamma(x) \odot F(x) + \beta(x), \end{aligned} \tag{2}$$

where $C$ denotes the channel dimension of the visual feature map, $\phi_o$ is a linear projection from the organ-embedding space to $\mathbb{R}^C$, and $\psi$ is a lightweight MLP that predicts the FiLM parameters $\gamma(x), \beta(x) \in \mathbb{R}^C$. This channel-wise modulation enhances organ-consistent responses while suppressing irrelevant regions, injecting anatomical awareness without altering global spatial structure.

**Voxel-Level Similarity Response and Candidate Generation.** For the $i$-th lesion, its query set $\{q_i^{(m)}\}_{m=1}^{M}$ is matched against the organ-aware feature map $\widetilde{F}$ via voxel-level cosine similarity, yielding a response map $S_i$:

$$S_i(x) = \max_m \frac{\left\langle q_i^{(m)},\, \widetilde{F}(x) \right\rangle}{\left\| q_i^{(m)} \right\| \left\| \widetilde{F}(x) \right\|}. \tag{3}$$

Based on $S_i$, adaptive thresholding and 3D connected-component analysis are performed to select the Top-$K_p$ high-response components, forming a high-recall candidate set $\mathcal{P}_i = \{P_i^{(k)}\}_{k=1}^{K_p}$. Each candidate $P_i^{(k)}$ retains its spatial indices in the global CT coordinate system and serves as a localization anchor for subsequent verification and segmentation refinement.

**Region Consistency Verification.** Since voxel-level similarity peaks are prone to noise, we verify candidates at the region level. For lesion description $i$ with candidates $\{P_i^{(k)}\}_{k=1}^{K_p}$, we crop and pool the feature map $\widetilde{F}$ to obtain region features $v_i^{(k)}$, and construct evidence vectors $\xi_i^{(k)}$ encoding anatomical coverage and report-attribute consistency. Specifically, $\xi_i^{(k)}$ includes the overlap ratio with $\widetilde{M}_{o(i)}$ and region statistics from $P_i^{(k)}$, i.e., volume $V_i^{(k)}$ and mean HU $\mu_i^{(k)}$, which are compared with reported volume and HU to derive matching metrics:

$$s_i^{(k)} = \sigma\left(W_s\left[v_i^{(k)}; z_i\right]\right) + \boldsymbol{\eta}^\top \xi_i^{(k)}, \qquad P_i^* = \arg\max_k s_i^{(k)}, \tag{4}$$

where $z_i$ denotes the lesion-level semantic vector, $\sigma(W_s[\cdot])$ produces a scalar semantic matching score, and $W_s$ and $\boldsymbol{\eta}$ are learnable parameters.

To promote a single dominant match per description, we minimize the entropy of the temperature-scaled candidate distribution:

$$\mathcal{L}_{\text{uni}} = -\sum_{i=1}^{N}\sum_{k=1}^{K_p} p_i^{(k)} \log\left(p_i^{(k)} + \epsilon\right), \qquad p_i^{(k)} = \frac{\exp\left(s_i^{(k)}/\tau\right)}{\sum_{j=1}^{K_p}\exp\left(s_i^{(j)}/\tau\right)}, \tag{5}$$

where $p_i^{(k)}$ is the temperature-scaled soft assignment, $\tau$ controls distribution sharpness, and $\epsilon$ is a small constant for numerical stability.

### 2.4 Octree Autoregressive Lesion Refinement (OcRe)

We propose an octree-driven autoregressive refinement framework for coarse-to-fine, hierarchical optimization of lesion masks, improving boundary delineation for small lesions with blurred details while remaining computationally efficient.

**Octree Construction and Hierarchical Refinement.** From the selected candidate $P_i^*$, we build an octree $\mathcal{T}_i$ with root cube $\Omega_i$ enclosing $P_i^*$; recursive subdivision yields multi-level blocks $\Omega_i^{\ell,j}$ across spatial scales, progressively narrowing the region of interest in a coarse-to-fine, top-down refinement hierarchy.

**Autoregressive Mechanism and Conditional Modulation.** At each refinement level, we employ a conditional autoregressive strategy, where the prediction from the previous level serves as an explicit structural prior for progressive segmentation refinement. The initial coarse mask $\hat{Y}_i^0$ is initialized from the selected candidate $P_i^*$. For each octree node $\Omega_i^{\ell,j}$, we crop local organ-aware features, $F_i^{\ell,j} = \mathrm{Crop}(\widetilde{F}, \Omega_i^{\ell,j})$, and use the upsampled prediction from the preceding level as a conditional input, $m_i^{\ell-1,j} = \mathrm{Crop}\Big(\mathrm{Up}(\hat{Y}_i^{\ell-1}), \Omega_i^{\ell,j}\Big)$. The concatenated representation is then fed into a parameter-shared refinement decoder:

$$\hat{Y}_i^{\ell,j} = R_\theta\Big(\mathrm{Concat}(F_i^{\ell,j},\, m_i^{\ell-1,j})\Big), \tag{6}$$

where $R_\theta$ is a lightweight 3D residual refinement head shared across levels and nodes. The node-level predictions $\{\hat{Y}_i^{\ell,j}\}_j$ are merged into $\hat{Y}_i^\ell$ for coarse-to-fine boundary refinement, especially around small and low-contrast lesions.

**Loss Under Missing Mask Annotations.** Since many clinical samples lack pixel-level masks, we optimize the model on unlabeled data via report-driven weak supervision: for lesion $i$, the prediction yields $V_i$ and $\mu_i$ to match $V_i^r$ and $\mu_i^r$, while $\mathcal{L}_{\text{sep}}$ penalizes inter-lesion overlap; the weak loss is defined as:

$$\mathcal{L}_{\text{weak}} = \lambda_{\text{uni}}\mathcal{L}_{\text{uni}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{sep}}\mathcal{L}_{\text{sep}}, \tag{7}$$

where $\mathcal{L}_{\text{con}} = \mathcal{L}_{\text{attr}} + \mathcal{L}_{\text{org}}$, $\mathcal{L}_{\text{attr}} = \sum_{a\in\{V,\mu\}} \sum_{i=1}^{N} \left(\frac{|a_i - a_i^r|}{|a_i^r| + \epsilon}\right)^2$, $\mathcal{L}_{\text{org}}$ penalizes predicted lesion mass outside the binary target-organ mask $M_{o(i)}$ and $\lambda_{\text{uni}}, \lambda_{\text{con}}, \lambda_{\text{sep}}$ weight unimodality, consistency, and separation.

For samples with Ground Truth (GT) masks, a combined Dice and cross-entropy loss $\mathcal{L}_{\text{seg}}$ is employed. Let $\mathcal{D}_m$ denote the subset of samples with pixel-level labels. For each sample $x$, the total loss is defined as:

$$\mathcal{L}_{\text{total}} = \delta\, \lambda_{\text{seg}}\, \mathcal{L}_{\text{seg}} + \lambda_{\text{weak}}\, \mathcal{L}_{\text{weak}}, \quad \text{s.t.} \quad \delta = \mathbb{I}[x \in \mathcal{D}_m]. \tag{8}$$

## 3 Experiments

### 3.1 Dataset and Implementation

**Dataset.** We conduct experiments on AbdomenAtlas 3.0 [6], comprising 9,262 cases across 138 institutions. Each case includes a report reviewed by 12 board-certified radiologists and voxel-aligned tumor masks for abdominal tumors, including kidney (K, 2,122), liver (L, 1,472), and pancreas (P, 361). We randomly split the cases into training, evaluation, and test sets at an 8:1:1 ratio.

**Implementation Details.** All experiments are implemented in PyTorch 2.4 and conducted on an NVIDIA H100 GPU. We use MedFormer [12] as the visual

**Table 1.** Quantitative comparison under varying mask ratios across **GLeVE** and other methods. Red and blue cells denote the best and second-best results, respectively.

| Method | Dice$_{std}$ (%)↑ | | | HD95$_{std}$ (mm)↓ | | | LR$_{std}$↑ | LLS$_{std}$ (%)↑ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | K | L | P | K | L | P | Avg | K | L | P |
| **Multimodal Foundation Models** | | | | | | | | | | |
| M3D-LaMed$_{100\%}$ [2] | $35.3_{32.7}$ | $13.6_{26.7}$ | $19.9_{26.1}$ | $95.9_{66.2}$ | $103_{45.9}$ | $42.1_{32.8}$ | $34.6_{41.6}$ | $10.4_{14.1}$ | $2.62_{5.94}$ | $10.6_{15.6}$ |
| Merlin$_{100\%}$ [7] | $47.7_{34.7}$ | $27.5_{30.8}$ | $18.4_{20.9}$ | $130_{128}$ | $121_{87.6}$ | $70.0_{36.4}$ | $62.9_{41.1}$ | $21.7_{22.1}$ | $19.6_{18.4}$ | $5.77_{9.99}$ |
| SAT$_{100\%}$ [31] | $51.6_{26.4}$ | $34.9_{21.3}$ | $17.6_{25.5}$ | $96.7_{76.2}$ | $82.3_{84.7}$ | $54.9_{68.1}$ | $57.9_{33.9}$ | $23.7_{15.7}$ | $28.9_{21.3}$ | $11.3_{10.5}$ |
| R1Seg-3D$_{100\%}$ [14] | $28.9_{29.7}$ | $15.6_{22.6}$ | $6.47_{15.0}$ | $276_{264}$ | $196_{177}$ | $237_{146}$ | $43.7_{39.6}$ | $12.5_{18.1}$ | $15.3_{17.8}$ | $2.30_{4.83}$ |
| CT-CLIP$_{100\%}$ [13] | $48.9_{28.1}$ | $28.6_{24.5}$ | $16.4_{19.8}$ | $127_{86.8}$ | $93.7_{73.1}$ | $81.6_{56.4}$ | $48.7_{36.8}$ | $18.9_{17.3}$ | $23.2_{20.4}$ | $9.38_{11.5}$ |
| **Universal Segmentation Methods** | | | | | | | | | | |
| UNet$_{100\%}$ [24] | $51.3_{30.9}$ | $31.7_{24.6}$ | $12.7_{16.3}$ | $180_{152}$ | $234_{164}$ | $105_{151}$ | $66.2_{38.7}$ | $30.2_{25.4}$ | $36.3_{22.2}$ | $6.83_{10.8}$ |
| nnUNetV2$_{100\%}$ [18] | $55.7_{34.2}$ | $34.8_{28.1}$ | $21.1_{22.3}$ | $152_{129}$ | $72.7_{65.1}$ | $109_{85.4}$ | $71.4_{40.2}$ | $42.2_{29.0}$ | $27.2_{26.3}$ | $11.6_{12.3}$ |
| STUNet$_{100\%}$ [16] | $53.0_{35.3}$ | $30.8_{25.6}$ | $22.6_{23.2}$ | $136_{112}$ | $119_{104}$ | $111_{101}$ | $65.5_{40.2}$ | $33.8_{28.3}$ | $33.2_{26.5}$ | $12.8_{15.6}$ |
| SwinUNETR$_{100\%}$ [25] | $56.5_{32.2}$ | $34.2_{30.2}$ | $22.2_{19.5}$ | $130_{115}$ | $153_{133}$ | $105_{79.9}$ | $71.4_{37.5}$ | $36.6_{26.3}$ | $31.7_{25.6}$ | $13.4_{13.5}$ |
| Medformer$_{100\%}$ [12] | $59.8_{28.0}$ | $44.3_{25.2}$ | $16.4_{22.3}$ | $71.0_{149}$ | $90.1_{127}$ | $36.5_{30.2}$ | $73.2_{39.0}$ | $39.2_{22.6}$ | $35.3_{27.8}$ | $15.9_{17.0}$ |
| Med-SAM3$_{100\%}$ [21] | $34.5_{25.4}$ | $28.1_{28.9}$ | $15.1_{17.9}$ | $141_{71.5}$ | $120_{67.8}$ | $127_{79.3}$ | $68.5_{40.1}$ | $23.3_{19.6}$ | $25.9_{23.6}$ | $11.8_{13.3}$ |
| **Advanced Methods** | | | | | | | | | | |
| R-Super$_{10\%}$ [5] | $53.3_{32.4}$ | $37.5_{27.8}$ | $15.1_{19.6}$ | $162_{193}$ | $128_{139}$ | $99.6_{85.3}$ | $73.9_{36.9}$ | $32.9_{25.5}$ | $33.5_{19.8}$ | $6.71_{12.5}$ |
| R-Super$_{25\%}$ [5] | $54.3_{32.0}$ | $39.6_{31.3}$ | $20.5_{25.3}$ | $101_{128}$ | $74.4_{73.5}$ | $85.1_{73.2}$ | $74.4_{36.4}$ | $37.2_{26.8}$ | $37.1_{23.6}$ | $15.3_{14.8}$ |
| R-Super$_{100\%}$ [5] | $63.9_{28.5}$ | $49.2_{26.9}$ | $22.6_{25.5}$ | $81.1_{118}$ | $66.6_{53.2}$ | $33.4_{26.9}$ | $74.6_{38.3}$ | $42.9_{23.7}$ | $40.9_{28.7}$ | $15.5_{17.4}$ |
| **GLeVE (Ours)**$_{10\%}$ | $53.9_{32.1}$ | $38.2_{29.4}$ | $18.4_{23.0}$ | $113_{148}$ | $96.1_{131}$ | $62.7_{92.4}$ | $74.1_{40.1}$ | $35.4_{25.1}$ | $33.9_{23.7}$ | $12.8_{18.4}$ |
| **GLeVE (Ours)**$_{25\%}$ | $56.6_{29.8}$ | $42.9_{27.1}$ | $19.3_{21.4}$ | $98.4_{132}$ | $83.3_{118}$ | $54.6_{85.7}$ | $74.9_{37.8}$ | $38.1_{24.5}$ | $37.7_{21.4}$ | $15.5_{18.6}$ |
| **GLeVE (Ours)**$_{100\%}$ | $65.8_{26.2}$ | $48.7_{23.9}$ | $27.8_{19.6}$ | $69.5_{109}$ | $60.2_{96.4}$ | $29.5_{64.8}$ | $76.2_{35.3}$ | $44.3_{21.2}$ | $38.9_{20.4}$ | $17.8_{14.8}$ |

encoder-decoder backbone and adopt TotalSegmentator [28] for organ segmentation. We set the batch size to 1 and train for 100 epochs with the random seed of 2026. Model parameters are optimized using AdamW [22] with an initial learning rate of $1 \times 10^{-4}$ and cosine annealing for learning-rate decay.

**Evaluation metrics.** For evaluation, we report Dice and HD95 and additionally adopt Lesion Recall (LR) and Lesion Localization Score (LLS). We extract lesion instances via connected-component analysis on prediction and GT masks, declare a match when Dice $\geq 0.1$, and use the Hungarian algorithm for optimal assignment, yielding $N_{\text{matched}}$ and $N_{\text{gt}}$. LLS explicitly penalizes centroid displacement, yielding greater sensitivity to lesion-level localization errors.

$$\text{LR} = \frac{N_{\text{matched}}}{N_{\text{gt}}}, \qquad \text{LLS} = \frac{1}{N_{\text{gt}}} \sum_{i=1}^{N_{\text{gt}}} \text{Dice}(i, \pi(i)) \exp\left(-\frac{d_i}{d_0}\right), \tag{9}$$

where $d_i$ is the centroid distance for the $i$-th match, $d_0 = 20$ mm is the decay scale, and $\text{Dice}(i, \pi(i)) = 0$ for unmatched GT lesions.

### 3.2 Experiment Results

**Comparative Experiments.** As shown in Table 1 and Fig. 3, with full mask annotations and report supervision, **GLeVE** consistently surpasses multimodal foundation models (e.g., CT-CLIP [13]), universal segmentation methods (e.g., MedFormer [12]), and an advanced report-assisted method R-Super [5]. Averaged across organs, it increases Dice by 7.26% over the best universal segmentation method and reduces HD95 by 7.3 mm relative to the advanced method,

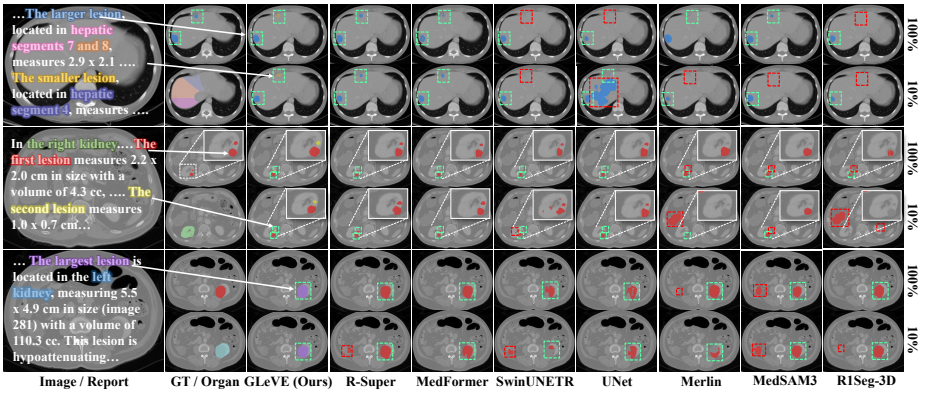


**Fig. 3.** Qualitative comparison of grounding results under 10% and 100% mask supervision, highlighting localization accuracy and multi-lesion disambiguation.

**Table 2.** Ablation analysis of key components under 100% mask supervision.

| Method | Dice$_{std}$ (%)↑ | | | HD95$_{std}$ (mm)↓ | | | LR$_{std}$↑ | LLS$_{std}$ (%)↑ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | K | L | P | K | L | P | Avg | K | L | P |
| w/o LeQu | $59.5_{28.4}$ | $42.5_{25.7}$ | $23.5_{20.9}$ | $84.6_{124}$ | $72.0_{108}$ | $39.3_{76.5}$ | $73.3_{37.2}$ | $40.9_{22.1}$ | $32.8_{20.7}$ | $16.4_{16.2}$ |
| w/o AnVer | $55.9_{31.6}$ | $39.7_{28.9}$ | $20.3_{23.1}$ | $89.8_{134}$ | $75.1_{121}$ | $47.6_{84.9}$ | $69.3_{40.5}$ | $34.6_{24.6}$ | $30.5_{22.9}$ | $13.9_{17.8}$ |
| w/o OcRe | $63.7_{26.9}$ | $44.3_{24.1}$ | $25.4_{19.8}$ | $74.8_{112}$ | $64.0_{99.6}$ | $33.0_{69.2}$ | $75.1_{36.6}$ | $42.2_{21.7}$ | $33.9_{19.8}$ | $17.6_{15.1}$ |
| **GLeVE (Ours)** | $65.8_{26.2}$ | $48.7_{23.9}$ | $27.8_{19.6}$ | $69.5_{109}$ | $60.2_{96.4}$ | $29.5_{64.8}$ | $76.2_{35.3}$ | $44.3_{21.2}$ | $38.9_{20.4}$ | $17.8_{14.8}$ |

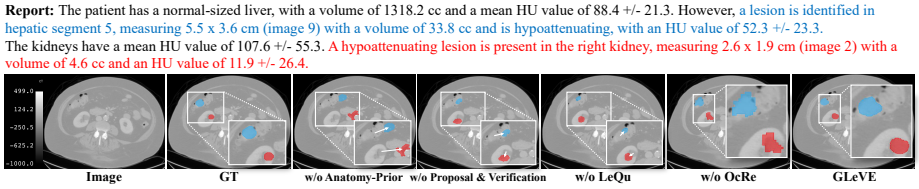


**Fig. 4.** Qualitative ablation results of GLeVE, showing the effect of removing Anatomical-Prior, Proposal & Verification, LeQu, and OcRe on lesion grounding and boundary refinement. Arrows indicate localization center offsets.

indicating improved global geometric consistency and fewer extreme boundary errors. Lesion-level performance is likewise strong, reaching 76.2% LR and 33.7% LLS. Under weak supervision with only 10% or 25% masks, **GLeVE** remains competitive and often outperforms most alternatives, highlighting robustness to annotation scarcity and improved data efficiency.

**Ablation Study.** As reported in Table 2 and Fig. 4, replacing LeQu with a text encoder applied to structured descriptions reduces LR by 2.9%, indicating that explicit relational modeling preserves lesion semantics and enhances representation discriminability. Disabling AnVer reduces the organ-averaged LLS by 7.3%, indicating that candidate filtering and verification are critical for seman-

tic alignment. Removing OcRe increases the organ-averaged HD95 by 4.2 mm, highlighting the contribution of hierarchical recursive refinement to accurate boundary delineation. Overall, the three modules are complementary in semantic modeling, verification, and refinement, collectively enabling robust localization.

## 4 Conclusion

In this study, we propose **GLeVE** for verifiable, pixel-accurate lesion grounding from radiology reports in 3D CT. By modeling each lesion description as an atomic unit, **GLeVE** learns relation-aware lesion queries, enforces anatomy-consistent one-to-one text-lesion matching via soft modulation and region-level verification, and refines masks with octree-based autoregressive coarse-to-fine decoding; a report-driven objective further enables training under missing masks. Experiments on AbdomenAtlas 3.0 demonstrate consistent improvements over advanced methods, supporting robust grounding in challenging multi-lesion settings. Future work will explore extending the framework to broader anatomical regions and more diverse disease categories.

**Acknowledgements** This work was supported by the National Natural Science Foundation of China (No. 62472133) and the Fundamental Research Funds for the Provincial Universities of Zhejiang (No. GK259909299001-006).

**Disclosure of Interests** The authors declare no competing interests.

## References

1. Alabed, S., Anderson, A., Maiter, A., et al.: Large language models for simplifying radiology reports: A systematic review and meta-analysis of patient, public, and clinician evaluations. The Lancet Digital Health (2026)
2. BAI, F., Du, Y., Huang, T., q.-h. Meng, M., Zhao, B.: M3D: Advancing 3D medical image analysis with multi-modal large language models. In: International Conference on Learning Representations (2024)
3. Bai, X., Liu, M., Chen, Y., Yang, H., Tian, Q.: Chest-OMDL: Organ-specific multidisease detection and localization in chest computed tomography using weakly supervised deep learning from free-text radiology report. In: Medical Imaging with Deep Learning (2025)
4. Bai, X., Liu, M., Song, T., et al.: EXACT: an explainable anomaly-aware vision foundation model for analysis of 3D chest CT. arXiv preprint arXiv:2604.24146 (2026)
5. Bassi, P.R., Li, W., Chen, J., et al.: Learning segmentation from radiology reports. In: Proceedings of Medical Image Computing and Computer Assisted Intervention. pp. 305–315 (2025)
6. Bassi, P.R., Yavuz, M.C., Hamamci, I.E., Er, S., Chen, X., Li, W., Menze, B., Decherchi, S., Cavalli, A., Wang, K., et al.: RadGPT: Constructing 3D image-text tumor datasets. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23720–23730 (2025)

7. Blankemeier, L., Cohen, J.P., Kumar, A., et al.: Merlin: A vision language foundation model for 3D computed tomography. Research Square pp. rs–3 (2024)
8. Busch, F., Hoffmann, L., Dos Santos, D.P., et al.: Large language models for structured reporting in radiology: Past, present, and future. European Radiology **35**(5), 2589–2602 (2025)
9. Chen, H., Chen, Y., Yan, Z., Ding, M., Li, C., Zhu, Z., Qin, F.: MMLNB: Multimodal learning for neuroblastoma subtyping classification assisted with textual description generation. arXiv preprint arXiv:2503.12927 (2025)
10. Chen, Y., Zou, B., Guo, Z., et al.: SCUNet++: Swin-UNet and CNN bottleneck hybrid architecture with multi-fusion dense skip connection for pulmonary embolism CT image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7759–7767 (2024)
11. Dong, F., Nie, S., Chen, M., et al.: Keyword-based AI assistance in the generation of radiology reports: A pilot study. NPJ Digital Medicine **8**(1), 490 (2025)
12. Gao, Y., Zhou, M., Liu, D., Yan, Z., Zhang, S., Metaxas, D.N.: A data-scalable Transformer for medical image segmentation: Architecture, model efficiency, and benchmark. arXiv preprint arXiv:2203.00131 (2023)
13. Hamamci, I.E., Er, S., Wang, C., Almas, F., et al.: Generalist foundation models from a multimodal dataset for 3D computed tomography. Nature Biomedical Engineering pp. 1–19 (2026)
14. Hao, Q., Yu, L., Tian, S., Ye, X., Zhang, L.: R1Seg-3D: Rethinking reasoning segmentation for medical 3D CTs. In: Proceedings of Medical Image Computing and Computer Assisted Intervention. pp. 415–425 (2025)
15. He, J., Li, P., Liu, G., Zhong, S.: Parameter-efficient fine-tuning medical multimodal large language models for medical visual grounding. In: International Symposium on Biomedical Imaging. pp. 1–5. IEEE (2025)
16. Huang, Z., Wang, H., Deng, Z., Ye, J., Su, Y., Sun, H., et al.: STU-Net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. arXiv preprint arXiv:2304.06716 (2023)
17. Ichinose, A., Hatsutani, T., Nakamura, K., et al.: Visual grounding of whole radiology reports for 3D CT images. In: Proceedings of Medical Image Computing and Computer Assisted Intervention. pp. 611–621 (2023)
18. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
19. Kulsoom, U., Glavin, F.G., Bendechache, M.: Natural language processing and machine learning for analysis of radiology reports-A systematic review. IEEE Access **13**, 112215–112254 (2025)
20. Li, Y., Kong, C., Zhao, G., Zhao, Z.: Automatic radiology report generation with deep learning: a comprehensive review of methods and advances. Artificial Intelligence Review **58**(11), 344 (2025)
21. Liu, A., Xue, R., Cao, X.R., et al.: MedSAM3: Delving into segment anything with medical concepts. arXiv preprint arXiv:2511.19046 (2025)
22. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2017)
23. Nützel, F., Dombrowski, M., Kainz, B.: Generate to ground: Multimodal text conditioning boosts phrase grounding in medical vision-language models. In: Medical Imaging with Deep Learning (2025)
24. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Proceedings of Medical Image Computing and Computer Assisted Intervention. pp. 234–241 (2015)

25. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3D medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20730–20740 (2022)
26. Tejani, A.S., Klontzas, M.E., Gatti, A.A., et al.: Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. Radiology: Artificial Intelligence **6**(4), e240300 (2024)
27. Vilouras, K., Sanchez, P., O'Neil, A.Q., Tsaftaris, S.A.: Zero-shot medical phrase grounding with off-the-shelf diffusion models. IEEE Journal of Biomedical and Health Informatics **29**(12), 9051–9059 (2025)
28. Wasserthal, J., Breit, H.C., Meyer, M.T., et al.: TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images. Radiology: Artificial Intelligence **5**(5), e230024 (2023)
29. Yang, A., Li, A., Yang, B., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
30. Yang, H., Zhou, H.Y., Liu, J., et al.: A multimodal vision–language model for generalizable annotation-free pathology localization. Nature Biomedical Engineering pp. 1–15 (2026)
31. Zhao, Z., Zhang, Y., Wu, C., Zhang, X., Zhou, X., Zhang, Y., Wang, Y., Xie, W.: Large-vocabulary segmentation for medical images with text prompts. NPJ Digital Medicine **8**(1), 566 (2025)
32. Zou, K., Bai, Y., Liu, B., et al.: Uncertainty-aware medical diagnostic phrase identification and grounding. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(12), 11315–11329 (2025)