ONE LESION, MANY VIEWS: DISENTANGLED VISION-LANGUAGE PRETRAINING FOR MULTI-DIMENSIONAL LESION UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Pretraining (VLP) has emerged as a powerful paradigm for medical image understanding by leveraging unstructured reports. However, existing medical VLP methods often fall short in fine-grained lesion understanding, primarily due to the inherent sparsity of lesion-specific information within free-text reports and the lack of explicit lesion-level anchoring during alignment. To address these limitations, we propose Lesion-CLIP, a disentangled VLP framework for multi-dimensional lesion understanding. First, for what to align, we introduce multi-dimensional data decoupling by decomposing 3D CT scans and radiology reports into five distinct semantic dimensions. Second, for where to align, we develop dimension-specific space decoupling by projecting image features into distinct subspaces. Third, for how to align, we design asymmetric direction decoupling to perform hierarchical dimension-level and lesion-level alignment. To this end, we curate LiLA, a pathology-calibrated 3D CT lesion dataset spanning type, size, location, phase, and attribute to facilitate fine-grained training and evaluation. Extensive experiments on LiLA and the public MSWAL dataset demonstrate that Lesion-CLIP achieves state-of-the-art performance across both discriminative and generative tasks. Notably, zero-shot Lesion-CLIP consistently outperforms established fully supervised models across all core dimensions on LiLA, achieving an absolute accuracy gain of 7.56% in lesion classification.

1 INTRODUCTION

Vision-Language Pretraining (VLP) has emerged as a cornerstone of medical image understanding by leveraging unstructured reports to learn transferable multimodal representations. This paradigm has evolved rapidly along two main trajectories: spatial dimensionality and semantic granularity. To capture the volumetric and spatial context missed by early 2D models (Zhang et al., 2022; Wang et al., 2022b), recent advancements have transitioned to 3D VLP, training volumetric foundation models on large-scale CT data (Hamamci et al., 2026; Blankemeier et al., 2026). Concurrently, medical VLP has progressed from coarse, scan-level matching (Lin et al., 2023; Zhang et al., 2025) to fine-grained, anatomy-level grounding (Shui et al., 2025; Cao et al., 2025). By segmenting scans into predefined anatomical structures and pairing them with localized report descriptions, these frameworks have effectively mitigated the spatial ambiguity inherent in global matching. This continuous drive toward spatial and semantic localization presents an unprecedented opportunity to transition from whole-scan screening to fine-grained, lesion-focused medical image understanding.

Despite these advancements, existing medical VLP methods face a critical bottleneck when applied to fine-grained lesion understanding, constrained by a dual dilemma of methodological entanglement and benchmark scarcity. Specifically, current models fail to capture localized pathologies due to the inherent sparsity of lesion-specific information within free-text reports and the lack of explicit lesion-level anchoring during alignment. In clinical practice, evaluating a lesion is essentially multi-dimensional (Ganeshan et al., 2018; Nobel et al., 2022), requiring the simultaneous assessment of its diagnostic type, physical size, anatomical location, imaging phase, and phenotypic attribute, as illustrated in Fig. 1. Existing global or anatomy-aligned VLP methods compress these heterogeneous and sparse semantic dimensions into a single embedding space for holistic alignment, causing severe representation interference. Resource-wise, existing lesion-level benchmarks are highly frag-

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

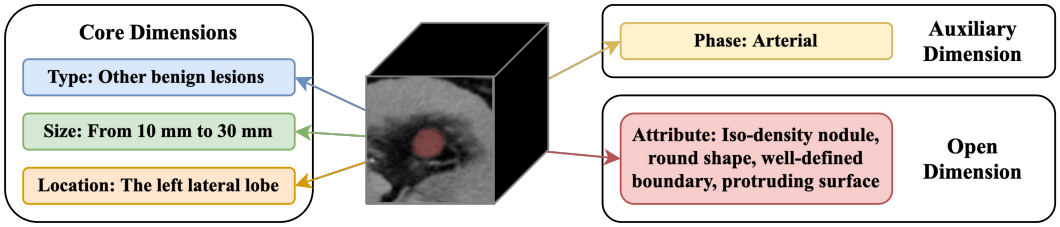


Figure 1: Schematic of the multi-dimensional nature of lesions.

mented. As illustrated in Appendix B.1, they either restrict annotations to 2D slices (Zhu et al., 2026) or provide 3D masks without unstructured text narratives (Wu et al., 2025), leaving a critical void of a high-quality, multi-dimensional 3D lesion dataset for fine-grained training and evaluation.

To bridge this gap, we propose Lesion-CLIP, a disentangled VLP framework for multi-dimensional lesion understanding. Unlike prior approaches that rely primarily on holistic alignment, our method explicitly reformulates the lesion-centric VLP along three axes: (1) *what* to align, via multi-dimensional data decoupling that decomposes 3D CT scans and radiology reports into five distinct semantic dimensions; (2) *where* to align, through dimension-specific space decoupling which projects image features into distinct subspaces; (3) *how* to align, using asymmetric direction decoupling to perform hierarchical dimension-level and lesion-level alignment. To support this framework, we curate LiLA, a pathology-calibrated 3D CT lesion dataset spanning type, size, location, phase, and attribute. Extensive experiments on LiLA and the public MSWAL dataset demonstrate that Lesion-CLIP achieves state-of-the-art performance across both discriminative and generative tasks. Notably, zero-shot Lesion-CLIP consistently outperforms established fully supervised models across all core dimensions on LiLA, achieving an absolute improvement of 7.56% in lesion classification accuracy. Our contributions can be summarized as follows:

- We propose Lesion-CLIP, a disentangled VLP framework for multi-dimensional lesion understanding, which shifts the medical VLP focus from holistic anatomical matching to decoupled lesion analysis.
- We formulate *what*, *where*, and *how* to align by introducing multi-dimensional data decoupling, dimension-specific space decoupling, and asymmetric direction decoupling.
- We curate LiLA, a pathology-calibrated multi-dimensional 3D CT lesion dataset providing rich, localized cross-modal supervision.
- Extensive experiments on LiLA and the public MSWAL dataset demonstrate that Lesion-CLIP achieves state-of-the-art performance across both discriminative and generative tasks.

2 Related Work

2.1 Medical Vision-Language Pretraining

Early medical VLP frameworks primarily focused on 2D images, notably chest radiographs or those sourced from general biomedical literature. For instance, ConVIRT (Zhang et al., 2022) pioneered bidirectional contrastive learning between X-ray images and clinical reports, laying the foundation for subsequent large-scale 2D models like BiomedCLIP (Zhang et al., 2025). However, these 2D models cannot capture the rich volumetric and spatial context inherent in 3D medical scans (Shui et al., 2025). To tackle this challenge, recent advancements have transitioned from 2D to 3D VLP. RadCLIP (Lu et al., 2025) incorporates a slice-pooling mechanism to analyze volumetric radiology data, while CT-CLIP (Hamamci et al., 2026) aligns 3D CT volumes with text reports to enable zero-shot multi-abnormality detection. Furthermore, Merlin (Blankemeier et al., 2026) represents a massive 3D foundation model trained on millions of CT scans with comprehensive electronic health records and radiology reports. Despite their success, these models generally align whole 3D volumes with reports globally, falling short in capturing localized features.

To overcome the limitations of global image-text matching, medical VLP has evolved from coarse-grained to fine-grained alignment. Early coarse-grained methods, such as MedCLIP (Wang et al.,

2022b) and PMC-CLIP (Lin et al., 2023), rely on global representation matching between entire images and text reports. To model localized semantics, implicit fine-grained models like GLoRIA (Huang et al., 2021) and MGCA (Wang et al., 2022a) leverage a cross-attention mechanism to capture correspondences between regions and words. Moving beyond these approaches, explicit fine-grained methods have been proposed to anchor medical concepts to explicit regions. For example, fVLM (Shui et al., 2025) explicitly segments CT scans into predefined anatomical regions produced by TotalSegmentator (Wasserthal et al., 2023) and pairs them with corresponding sentences in radiology reports. Building upon this organ-aware paradigm, ViSD-Boost (Cao et al., 2025) further enhances the visual semantic density of these anatomical regions by incorporating disease-level contrastive learning and anatomical normality modeling. However, these methods focus primarily on anatomy-level alignment, largely overlooking the localized nature of individual lesions. In contrast, our proposed Lesion-CLIP achieves fine-grained lesion-level anchoring.

2.2 Multi-Dimensional Lesion Understanding

Multi-dimensional lesion analysis requires highly detailed and annotated datasets. Early efforts mainly focused on harvesting structured datasets. For instance, DeepLesion (Yan et al., 2018) mined thousands of universal lesion annotations from Picture Archiving and Communication System (PACS) bookmarks, capturing bounding boxes and physical measurements for lesion analysis in 2D CT slices. To better capture the spatial complexity of anatomical structures, subsequent research shifted from 2D key slices to 3D volumetric datasets. KiTS (Heller et al., 2021) and LiTS (Bilic et al., 2023) provided voxel-level 3D annotations for specific organs and tumors. Subsequently, MSWAL (Wu et al., 2025) scaled this effort to the entire abdominal cavity, curating a multi-organ lesion dataset to capture a broader spectrum of clinical pathologies. However, these lesion datasets lack unstructured, natural language narratives, such as attribute descriptions, which are crucial for providing rich diagnostic context. Consequently, recent benchmarks have transitioned towards integrating structured annotations with unstructured text. A notable milestone is CT-Bench (Zhu et al., 2026), which extends DeepLesion by pairing over 20,000 CT lesion annotations with both structured physical measurements and unstructured text descriptions. Despite these advances, there is a critical lack of a multi-dimensional 3D benchmark that integrates localized lesions with both structured and unstructured annotations. To bridge this gap, we curate LiLA, a pathology-calibrated 3D CT lesion dataset spanning type, size, location, phase, and attribute, thereby providing a dedicated foundation for training and evaluating multi-dimensional lesion understanding.

3 Method

In this section, we present Lesion-CLIP, a disentangled VLP framework for multi-dimensional lesion understanding. As illustrated in Fig. 2, Lesion-CLIP relies on three core mechanisms: multi-dimensional data decoupling (Sec. 3.1), dimension-specific space decoupling (Sec. 3.2), and asymmetric direction decoupling (Sec. 3.3).

3.1 Multi-Dimensional Data Decoupling

Targeting the semantic sparsity in free-text radiology reports manifested primarily in missing contexts, descriptive irregularities, and diagnostic speculation, we first establish *what* to align by decomposing the raw data into five distinct semantic dimensions: type, size, location, phase, and attribute. Unlike prior works that rely solely on noisy report parsing (Shui et al., 2025; Cao et al., 2025), our decoupling pipeline leverages multi-source clinical and physical priors to construct high-fidelity supervision signals while providing reliable anchors for subsequent text grounding.

Radiology Annotation and Pathology Calibration. The foundation of our data decoupling lies in systematic clinical annotation and validation. For each case, a radiologist independently drafts the findings and impressions based on the original CT scan. Subsequently, a second radiologist manually annotates the 3D lesion mask on the venous phase volume, with access to the written report to facilitate precise spatial localization. The mask is then propagated to the remaining phases via 3D deformable registration (Mok et al., 2024). This sequential, report-informed annotation protocol ensures spatial consistency between the visual mask and the text narrative while preserving the independence of the original diagnostic interpretation. Crucially, the lesion type is not derived from

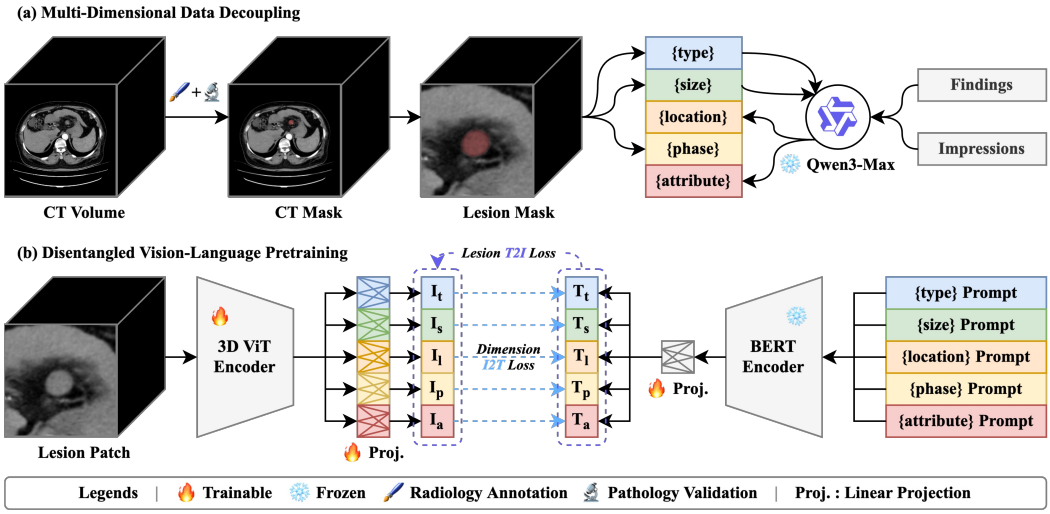


Figure 2: Overview of the proposed Lesion-CLIP. (a) Multi-dimensional data decoupling: 3D CT scans and radiology reports are decomposed into five semantic dimensions. (b) Disentangled VLP: Image features are projected into distinct subspaces for hierarchical dimension-level Image-to-Text (I2T) and lesion-level Text-to-Image (T2I) alignment with text representations.

noisy radiology reports but is calibrated against histopathological examination results to establish a reliable clinical reference standard. Specifically, this calibration is applied to suspected malignant lesions with histopathological confirmation, whereas typical benign lesions are verified based on established radiological criteria. This guarantees that the primary classification signal is free from the hallucination or ambiguity often present in uncalibrated imaging-only labels.

Physical Measurement and Patch-Level Extraction. To avoid the inconsistency of free-form natural language size descriptions, we derive physical measurements directly from the masks and discretize them into ordinal bins to define the size dimension. Subsequently, we resample each volume to an isotropic resolution of 1 mm and perform cropping to extract $96 \times 96 \times 96$ voxel patches, each centered at the geometric center of the lesion isolated via connected component analysis, enabling precise visual anchoring. Additionally, the imaging phase of each patch is extracted automatically from the metadata, providing essential temporal context for contrast-enhanced studies.

LLM-Driven Text Grounding. To ground the anatomical location and phenotypic attribute in radiology reports while maintaining alignment with the visual evidence, we employ a Large Language Model (LLM), Qwen3-Max (Yang et al., 2025), as an information extractor rather than a generator. Conditioned on the pathology-calibrated type, measured size, and metadata-derived phase, the LLM localizes the target lesion within the radiologist's findings, using the impressions only as auxiliary context. It then extracts the anatomical location mapped to a fixed vocabulary and the phenotypic attribute description strictly from the text span describing the matched lesion. This grounding strategy promotes spatial and semantic consistency between the extracted text descriptions and the cropped lesion patches, effectively distilling dense supervision from sparse narratives. All extracted results are reviewed by a radiologist, with identified errors manually corrected. Notably, the initial LLM grounding error rate is below 1%. The detailed prompt design is provided in Appendix B.3.

3.2 Dimension-Specific Space Decoupling

Having established five distinct semantic dimensions, we next address *where* to align. Projecting heterogeneous concepts into a single shared embedding space inevitably causes representation interference. To mitigate this, we introduce dimension-specific space decoupling, which learns dedicated visual subspaces for each dimension while maintaining a unified text representation space for semantic coherence. While the text inputs are already decoupled by dimension, the image features remain highly entangled, thus requiring separate visual projections to isolate distinct semantics.

Formally, let $f_v(\cdot)$ and $f_t(\cdot)$ denote the 3D vision encoder and the frozen text encoder, with output feature sizes of D_v and D_t, respectively. For each semantic dimension $m \in \{1, \ldots, M\}$, we learn a dedicated visual projection matrix $\mathbf{W}_v^m \in \mathbb{R}^{D \times D_v}$ to map the global image features $f_v(I)$ into a dimension-specific subspace in $\mathbb{R}^D$. Conversely, a shared text projection $\mathbf{W}_t \in \mathbb{R}^{D \times D_t}$ is applied across all dimensions to project the text features $f_t(T^m)$ into the same joint embedding space in $\mathbb{R}^D$, thereby preserving the inter-dimensional semantic relationships inherent in natural language. The normalized embeddings for the m-th dimension are computed as follows:

$$\mathbf{z}_v^m = \frac{\mathbf{W}_v^m f_v(I)}{\|\mathbf{W}_v^m f_v(I)\|_2}, \quad \mathbf{z}_t^m = \frac{\mathbf{W}_t f_t(T^m)}{\|\mathbf{W}_t f_t(T^m)\|_2}, \tag{1}$$

where I denotes the input lesion patch and T^m represents the text prompt templated for dimension m, as detailed in Appendix B.4. This decoupling design fosters structural independence among the visual subspaces, allowing the model to capture dimension-specific semantics.

3.3 Asymmetric Direction Decoupling

Finally, we optimize *how* to align by acknowledging the intrinsic asymmetry between visual perception and linguistic description. Image features of a lesion are holistic and multi-dimensional, whereas text descriptions are typically sequential and dimension-specific. Therefore, we propose asymmetric direction decoupling, which applies distinct alignment strategies for Image-to-Text (I2T) and Text-to-Image (T2I) directions.

Specifically, the I2T direction is conceptualized as a dense-to-sparse analytical process. Due to the highly dense and redundant volumetric details within a 3D CT lesion patch, mapping this visual signal into the decoupled semantic subspaces $\mathbf{z}_v^m$ serves as an effective divide-and-conquer strategy to mitigate cross-dimensional representation interference. Conversely, the T2I direction is modeled as a sparse-to-dense integrative process. To impose a global cohesive constraint over the decoupled subspaces, we treat the lesion as a unified entity with clinically correlated semantic dimensions, leveraging lesion-level alignment of the integrated representations to capture their joint distribution.

Dimension-Level I2T Alignment. To leverage the fine-grained nature of space decoupling, the I2T direction enforces alignment exclusively within each distinct subspace. This objective encourages the dimension-specific visual projector $\mathbf{W}_v^m$ to isolate features dedicated solely to its corresponding semantic dimension. For a training batch of size B, let $\mathbf{z}_{v,i}^m$ and $\mathbf{z}_{t,i}^m$ denote the normalized visual and text embeddings of the i-th instance for dimension m. The dimension-level I2T contrastive loss for dimension m is formulated as follows:

$$\ell_{\text{I2T}}^m = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp\left(\tau_m \langle \mathbf{z}_{v,i}^m, \mathbf{z}_{t,i}^m \rangle\right)}{\sum_{j=1}^{B} \exp\left(\tau_m \langle \mathbf{z}_{v,i}^m, \mathbf{z}_{t,j}^m \rangle\right)}, \tag{2}$$

where $\tau_m = \exp(\theta_m)$ is a learnable logit scale parameter for dimension m, conceptually acting as the inverse temperature. This dimension-specific optimization mitigates the risk of dominant dimensions overshadowing subtler ones.

Lesion-Level T2I Alignment. In the T2I direction, we perform lesion-level alignment between the integrated representations of both modalities. Beyond capturing cross-dimensional correlations, this global constraint serves as a gravitational anchor to regularize the shared 3D vision encoder $f_v(\cdot)$. While the independent I2T losses pull the shared visual backbone in divergent gradient directions, the global T2I loss ensures that their integrated visual representation remains structurally bound to the same physical lesion, thereby mitigating potential subspace drift.

Formally, the integrated representations $\bar{\mathbf{z}}_{v,i}$ and $\bar{\mathbf{z}}_{t,i}$ for the i-th instance are derived via average pooling followed by ℓ_2 normalization across all M dimensions:

$$\bar{\mathbf{z}}_{v,i} = \frac{\sum_{m=1}^{M} \mathbf{z}_{v,i}^m}{\left\| \sum_{m=1}^{M} \mathbf{z}_{v,i}^m \right\|_2}, \quad \bar{\mathbf{z}}_{t,i} = \frac{\sum_{m=1}^{M} \mathbf{z}_{t,i}^m}{\left\| \sum_{m=1}^{M} \mathbf{z}_{t,i}^m \right\|_2}. \tag{3}$$

Using the geometric mean of dimension-specific logit scales to derive the global inverse temperature,

$$\tau_h = \left(\prod_{m=1}^{M} \tau_m \right)^{1/M} = \left(\prod_{m=1}^{M} \exp\left(\theta_m\right) \right)^{1/M} = \exp\left(\frac{1}{M} \sum_{m=1}^{M} \theta_m \right), \tag{4}$$

which simplifies to the exponential of the arithmetic mean of the learnable log scale parameters θ_m, the lesion-level T2I contrastive loss is defined as follows:

$$\ell^h_{\text{T2I}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(\tau_h\langle\bar{\mathbf{z}}_{v,i},\bar{\mathbf{z}}_{t,i}\rangle\right)}{\sum_{j=1}^{B}\exp\left(\tau_h\langle\bar{\mathbf{z}}_{v,j},\bar{\mathbf{z}}_{t,i}\rangle\right)}. \tag{5}$$

Total Pretraining Objective. The final VLP objective is given by the following equation:

$$\mathcal{L}_{\text{VLP}} = \frac{1}{M+1}\left(\sum_{m=1}^{M}\ell^m_{\text{I2T}} + \ell^h_{\text{T2I}}\right). \tag{6}$$

This hierarchical formulation harmonizes dense-to-sparse analysis with sparse-to-dense integration.

4 Experiments and Discussion

4.1 Experimental Setup

Datasets. To evaluate the efficacy of multi-dimensional lesion understanding, we conduct experiments on two complementary 3D CT lesion datasets: **LiLA**, our curated dataset providing fine-grained, pathology-calibrated information across five dimensions for in-depth analysis, and **MSWAL** (Wu et al., 2025), a large-scale public multi-organ lesion dataset used to assess the model's robustness and generalization across diverse clinical scenarios. Detailed dataset comparison and statistics for LiLA and MSWAL are provided in Appendix B.1 and Appendix B.2.

LiLA is an in-house pathology-calibrated 3D CT lesion dataset curated for multi-dimensional lesion understanding, which contains 6,136 scans and 30,134 liver lesions spanning 11 distinct lesion types. Each lesion is paired with a 3D mask and multi-dimensional information including its type, size, location, phase, and attribute. The lesions are divided into training, validation, and test sets at an 8:1:1 ratio at the patient level.

MSWAL is a public 3D whole-abdominal CT lesion dataset initially constructed for lesion segmentation, which contains 694 scans. In this work, we extracted 3,424 lesions from the 484 publicly available scans. Each lesion is paired with a 3D mask and its corresponding type label. The lesions are divided into training, validation, and test sets at an 8:1:1 ratio at the patient level.

Evaluation Metrics. For discriminative tasks, we employ Accuracy (ACC), macro-averaged Area Under the ROC Curve (AUC), and Mean Absolute Error (MAE) specifically for the ordinal size dimension. For generative tasks, we evaluate performance using the clinical semantic metric GREEN (Ostmeier et al., 2024), alongside standard lexical metrics including BLEU-1/2/3/4 (Papineni et al., 2002) and ROUGE-L (Lin, 2004). Furthermore, computational cost is assessed by Parameter Counts (Params), Floating-Point Operations (FLOPs), and Video Random Access Memory (VRAM).

Implementation Details. Implementation details are available in Appendix D.

4.2 Discriminative Evaluation

We evaluate the discriminative capability of Lesion-CLIP across zero-shot and linear probing settings to assess both its pretrained representation quality and downstream transferability.

Zero-Shot Multi-Dimensional Lesion Understanding. Table 1 summarizes the comparison of zero-shot Lesion-CLIP with fully supervised models on the LiLA dataset. Notably, Lesion-CLIP outperforms 3D ResNet-50 (He et al., 2016), 3D DenseNet-121 (Huang et al., 2017), and CT-ViT (Hamamci et al., 2024) across all three core dimensions. Specifically, it achieves 62.21% accuracy and 84.35% AUC on the primary type dimension. For ordinal size grading, it yields the lowest MAE of 0.39. In terms of anatomical location, it establishes a leading accuracy of 72.84% and an AUC of 80.34%. Furthermore, the dimension ablation results demonstrate that removing either attribute or phase supervision degrades performance, suggesting that our VLP framework effectively leverages auxiliary phase cues and open attribute semantics for lesion understanding. Extended analyses in Appendices E.1, E.2, E.3, and E.4 present category-wise comparisons, per-phase performance profiles, per-class performance breakdowns, and representative case studies, respectively.

Table 1: Comparison of zero-shot Lesion-CLIP with fully supervised models for multi-dimensional lesion understanding on the LiLA dataset, where all fully supervised models are jointly trained with three task-specific heads. Best results are highlighted as **first**, second, and third.

Method	Type		Size		Location	
	ACC (%)	AUC (%)	ACC (%)	MAE	ACC (%)	AUC (%)
3D ResNet-50	53.13	80.97	61.38	0.44	70.39	74.50
3D DenseNet-121	52.24	79.93	54.65	0.52	71.18	75.59
CT-ViT	54.65	78.70	62.17	0.49	72.24	75.37
Lesion-CLIP	**62.21**	**84.35**	**67.64**	**0.39**	**72.84**	**80.34**
w/o Attribute	57.60	81.31	64.52	0.43	68.67	75.94
w/o Phase	55.22	80.23	61.68	0.47	70.69	76.10
w/o Both	54.95	79.22	60.62	0.49	68.60	75.54

Table 2: Comparison of different methods for lesion classification under linear probing settings with varying training data ratios on the LiLA dataset. Best results are highlighted as **first**, second, and third.

Method	Source	1% Data		10% Data		100% Data	
		ACC (%)	AUC (%)	ACC (%)	AUC (%)	ACC (%)	AUC (%)
Radiomics	Cancer Res.$_{17}$	44.39	70.43	50.55	75.12	54.95	79.38
CLIP	ICML$_{21}$	30.31	62.14	35.01	69.55	39.65	74.53
BiomedCLIP	NEJM AI$_{25}$	27.92	59.83	34.15	69.37	35.91	70.14
fVLM	ICLR$_{25}$	30.47	54.27	33.42	65.57	36.63	68.70
RadCLIP	TNNLS$_{25}$	30.44	61.73	36.47	68.64	41.47	73.52
ViSD-Boost	ICCV$_{25}$	30.71	61.46	32.06	64.11	32.36	66.61
CT-CLIP	Nature BME$_{26}$	27.96	55.77	32.03	64.08	34.68	67.51
Merlin	Nature$_{26}$	30.08	58.09	41.27	70.87	44.92	74.91
Lesion-CLIP	**Ours**	**62.54**	**83.90**	**63.86**	**84.69**	**64.16**	**85.11**

Table 3: Comparison of different methods for lesion classification under linear probing settings with varying training data ratios on the MSWAL dataset. Best results are highlighted as **first**, second, and third.

Method	Source	1% Data		10% Data		100% Data	
		ACC (%)	AUC (%)	ACC (%)	AUC (%)	ACC (%)	AUC (%)
Radiomics	Cancer Res.$_{17}$	43.16	66.07	54.71	75.01	65.65	88.70
CLIP	ICML$_{21}$	49.85	64.46	51.67	78.65	61.40	86.16
BiomedCLIP	NEJM AI$_{25}$	44.38	66.03	45.59	76.97	63.53	89.84
fVLM	ICLR$_{25}$	47.11	62.59	54.41	78.29	67.48	89.32
RadCLIP	TNNLS$_{25}$	43.77	53.61	52.89	66.47	60.49	80.77
ViSD-Boost	ICCV$_{25}$	41.64	64.45	49.54	73.80	54.10	82.74
CT-CLIP	Nature BME$_{26}$	28.57	44.61	46.20	66.93	58.36	82.92
Merlin	Nature$_{26}$	35.26	64.17	49.54	75.60	68.69	90.22
Lesion-CLIP	**Ours**	**51.67**	**68.28**	**57.14**	**78.77**	**70.21**	**90.51**

Linear Probing and Data Efficiency. To evaluate representation transferability, we conduct linear probing experiments across varying supervision scales. In Tables 2 and 3, we compare Lesion-CLIP with Radiomics (Van Griethuysen et al., 2017), 2D VLP-based CLIP (Radford et al., 2021) and BiomedCLIP (Zhang et al., 2025), and 3D VLP-based fVLM (Shui et al., 2025), RadCLIP (Lu et al., 2025), ViSD-Boost (Cao et al., 2025), CT-CLIP (Hamamci et al., 2026), and Merlin (Blankemeier et al., 2026). On LiLA, linear probing of Lesion-CLIP with merely 1% of the training data achieves an accuracy of 62.54%, further elevating the performance ceiling beyond the zero-shot baseline, while on MSWAL, Lesion-CLIP consistently outperforms competing methods across all training

Table 4: Comparison of different VLP methods for lesion attribute generation on the LiLA dataset. Best results are highlighted as **first**, second, and third.

Method	GREEN (%)	BLEU-1 (%)	BLEU-2 (%)	BLEU-3 (%)	BLEU-4 (%)	ROUGE-L (%)
CLIP	27.23	30.32	20.84	16.50	13.63	34.95
BiomedCLIP	29.80	26.77	18.24	14.14	11.35	35.05
fVLM	26.88	31.65	21.63	17.09	14.13	34.29
RadCLIP	25.37	34.52	23.19	18.61	15.98	30.76
ViSD-Boost	17.08	22.79	14.19	10.72	8.48	28.33
CT-CLIP	22.92	29.70	19.55	15.23	12.48	32.20
Merlin	28.02	24.19	16.29	12.52	9.71	34.04
Lesion-CLIP	**45.59**	**53.06**	**43.39**	**38.13**	**34.67**	**46.67**
Text-Only	22.38	26.80	17.55	13.30	10.06	31.42
Random Tokens	22.05	32.87	20.67	14.46	10.37	28.42
Shuffled Images	21.08	31.55	19.04	13.22	9.43	26.78

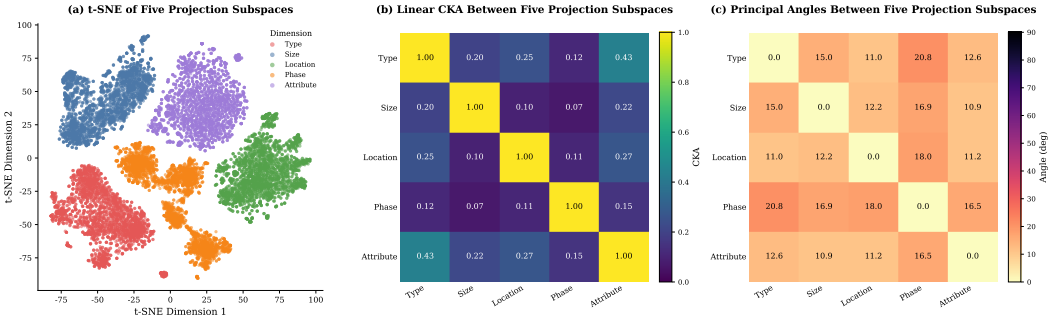


Figure 3: Geometry of the five Lesion-CLIP projection subspaces on the LiLA dataset.

fractions, reaching an accuracy of 70.21% and an AUC of 90.51% under full training data supervision. Notably, although Radiomics leverages privileged 3D masks, it still falls short of Lesion-CLIP. Appendix E.3 further details the robust category-wise performance of Lesion-CLIP on MSWAL.

4.3 Generative Evaluation

To evaluate our model's capability to capture fine-grained phenotypic descriptions beyond discrete categories, we perform free-text lesion attribute generation by aligning the learned visual representations with a lightweight LLM, Qwen3-4B-Instruct-2507 (Yang et al., 2025). The complete attribute generation workflow and detailed experimental setup are provided in Appendices C and D, respectively. As summarized in Table 4, Lesion-CLIP achieves a GREEN score of 45.59%, representing a 15.79% absolute improvement over the second-best method, which highlights its enhanced clinical semantic consistency. In addition, it yields a BLEU-4 of 34.67% and a ROUGE-L of 46.67%, demonstrating robust lexical fidelity. Notably, further visual-side ablations, including the text-only, random-token, and shuffled-image settings, all lead to severe performance degradation, confirming that the generated attribute descriptions are indeed grounded in lesion-specific visual cues. To complement these quantitative findings, qualitative case studies are provided in Appendix E.4.

4.4 Subspace Analysis

We first analyze the geometric structure of the five dimension-specific projection subspaces on the LiLA dataset. As depicted in Fig. 3, t-SNE (Van der Maaten & Hinton, 2008) projections reveal distinct clusters for each semantic dimension. This separation is further substantiated by low linear Centered Kernel Alignment (CKA) (Kornblith et al., 2019) similarity scores between subspaces (mean < 0.2), indicating minimal redundancy in the learned feature spaces. In addition, the principal angles between subspaces remain consistently significant ($> 10^\circ$), demonstrating that Lesion-CLIP maintains distinct representation manifolds even for semantically related dimensions like type and attribute. These geometric properties suggest that Lesion-CLIP effectively avoids inter-dimensional

Table 5: Ablation study on the impact of decoupling levels on zero-shot lesion prediction on the LiLA dataset. Best results are highlighted as **first**, second, and third.

Decoupling Level			Type		Size		Location	
Data	Space	Direction	ACC (%)	AUC (%)	ACC (%)	MAE	ACC (%)	AUC (%)
×	×	×	22.69	69.85	40.78	0.91	34.55	64.91
✓	×	×	58.93	81.48	65.09	0.43	68.73	78.97
✓	✓	×	60.38	82.82	66.98	**0.39**	71.38	79.02
✓	✓	✓	**62.21**	**84.35**	**67.64**	**0.39**	**72.84**	**80.34**

Table 6: Ablation study on the impact of decoupling levels on computation cost and lesion classification under linear probing settings with varying training data ratios on the MSWAL dataset. Best results are highlighted as **first**, second, and third.

Decoupling Level			Computational Cost			1% Data		100% Data	
Data	Space	Direction	Params (M)	FLOPs (G)	VRAM (GB)	ACC (%)	AUC (%)	ACC (%)	AUC (%)
×	×	×	**130.12**	**380.77**	**2.22**	46.50	56.59	57.75	84.71
✓	×	×	**130.12**	**380.77**	**2.22**	49.85	58.16	64.44	87.30
✓	✓	×	131.17	**380.77**	**2.32**	50.46	61.28	69.30	89.62
✓	✓	✓	131.17	**380.77**	**2.32**	**51.67**	**68.28**	**70.21**	**90.51**

collapse, thereby preserving the semantic specificity and independence of each learned dimension. A comprehensive extension of this analysis to the external MSWAL dataset, which demonstrates consistent subspace geometry across diverse clinical domains, is provided in Appendix E.5.

4.5 Ablation Study

To systematically evaluate the architectural choices of Lesion-CLIP, we conduct a progressive ablation study on the levels of decoupling, analyzing zero-shot prediction on LiLA and downstream linear probing on MSWAL. Starting from the baseline configuration (× × ×), introducing data decoupling (✓ × ×) yields a substantial performance improvement across all three evaluated semantic dimensions, as presented in Tables 5 and 6. Specifically, zero-shot type accuracy surges from 22.69% to 58.93%, while the 100% data linear probing accuracy increases by 6.69%, underscoring that decomposing holistic clinical texts into distinct dimensions is fundamental to mitigating semantic sparsity. Incorporating space decoupling (✓✓×) further enhances representation granularity, reducing the zero-shot size MAE to 0.39. Finally, integrating direction decoupling to form the complete Lesion-CLIP framework (✓✓✓) consistently delivers the best performance across all settings. Notably, the complete decoupling design introduces a negligible overall parameter increase of only 1.05M, virtually zero additional visual inference FLOPs and a minimal visual inference peak allocated VRAM increase of 0.10 GB.

5 Conclusion

In this work, we present Lesion-CLIP, a disentangled VLP framework for multi-dimensional lesion understanding. By systematically addressing *what*, *where*, and *how* to align, our method reconfigures medical VLP from coarse-grained matching to fine-grained, lesion-level semantic parsing. To support this framework, we curate LiLA, a pathology-calibrated multi-dimensional 3D CT lesion dataset providing rich, localized cross-modal supervision. Extensive experiments on LiLA and the public MSWAL dataset demonstrate that Lesion-CLIP achieves state-of-the-art performance across both discriminative and generative tasks.

Limitations and Future Work. While Lesion-CLIP achieves promising performance, it currently relies on pre-localized 3D lesion patches and is primarily validated on abdominal CT data. Future work will focus on developing an end-to-end joint analysis framework while expanding our multi-dimensional pretraining to diverse anatomical regions and other imaging modalities.

LLM USE STATEMENT

LLMs were used to assist with grammar and writing polishing. All equations, analyses, and core research contributions are entirely our own.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. Specifically, it was conducted retrospectively in accordance with the principles of the Declaration of Helsinki, utilizing previously obtained images that had been anonymized before conducting analysis, thus posing no risk or harm to patients. An informed consent waiver was applied. Approval was granted by the institutional ethics committee.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, we have provided comprehensive details throughout the main text and appendices. Specifically, the anonymous repository link to our code and data is available in Appendix A. Further information regarding dataset characteristics, lesion attribute generation, and implementation specifications can be found in Appendices B, C, and D, respectively. Additionally, we present extended results in Appendix E.

REFERENCES

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023.

Louis Blankemeier, Ashwin Kumar, Joseph Paul Cohen, Jiaming Liu, Longchao Liu, Dave Van Veen, Syed Jamal Safdar Gardezi, Hongkun Yu, Magdalini Paschali, Zhihong Chen, et al. Merlin: a computed tomography vision–language foundation model and dataset. *Nature*, 652 (8112):1318–1328, 2026.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European Conference on Computer Vision*, pp. 1–21. Springer, 2022.

Weiwei Cao, Jianpeng Zhang, Zhongyi Shui, Sinuo Wang, Zeli Chen, Xi Li, Le Lu, Xianghua Ye, Qi Zhang, Tingbo Liang, et al. Boosting vision semantic density with anatomy normality modeling for medical vision-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23041–23050, 2025.

Rachel Lea Draelos, David Dov, Maciej A Mazurowski, Joseph Y Lo, Ricardo Henao, Geoffrey D Rubin, and Lawrence Carin. Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. *Medical Image Analysis*, 67:101857, 2021.

Dhakshinamoorthy Ganeshan, Phuong-Anh Thi Duong, Linda Probyn, Leon Lenchik, Tatum A McArthur, Michele Retrouvey, Emily H Ghobadi, Stephane L Desouches, David Pastel, and Isaac R Francis. Structured reporting in radiology. *Academic Radiology*, 25(1):66–73, 2018.

Ibrahim Ethem Hamamci, Sezgin Er, Anjany Sekuboyina, Enis Simsar, Alperen Tezcan, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Furkan Almas, Irem Doğan, Muhammed Furkan Dasdelen, et al. Generatect: Text-conditional generation of 3d chest ct volumes. In *European Conference on Computer Vision*, pp. 126–143. Springer, 2024.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, et al. Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature Biomedical Engineering*, pp. 1–19, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical Image Analysis*, 67:101821, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3942–3951, 2021.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529. PMLR, 2019.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 525–536. Springer, 2023.

Zhixiu Lu, Hailong Li, Nehal A Parikh, Jonathan R Dillman, and Lili He. Radclip: Enhancing radiologic image analysis through contrastive language–image pretraining. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Tony CW Mok, Zi Li, Yunhao Bai, Jianpeng Zhang, Wei Liu, Yan-Jie Zhou, Ke Yan, Dakai Jin, Yu Shi, Xiaoli Yin, et al. Modality-agnostic structural image representation learning for deformable multi-modality medical image registration. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11215–11225. IEEE, 2024.

J Martijn Nobel, Koos van Geel, and Simon GF Robben. Structured reporting in radiology: a systematic review to explore its potential. *European Radiology*, 32(4):2837–2854, 2022.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, et al. Green: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 374–390, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Zhongyi Shui, Jianpeng Zhang, Weiwei Cao, Sinuo Wang, Ruizhe Guo, Le Lu, Lin Yang, Xianghua Ye, Tingbo Liang, Qi Zhang, and Ling Zhang. Large-scale and fine-grained vision-language pre-training for enhanced CT image understanding. In *The Thirteenth International Conference on Learning Representations*, 2025.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.

Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer Research*, 77(21):e104, 2017.

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in Neural Information Processing Systems*, 35:33536–33549, 2022a.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3876–3887, 2022b.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.

Zhaodong Wu, Qiaochu Zhao, Ming Hu, Yulong Li, Haochen Xue, Zhengyong Jiang, Angelos Stefanidis, Qiufeng Wang, Imran Razzak, Zongyuan Ge, et al. Mswal: 3d multi-class segmentation of whole abdominal lesions dataset. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 378–388. Springer, 2025.

Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of Medical Imaging*, 5(3):036501–036501, 2018.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1):AIoa2400640, 2025.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pp. 2–25. PMLR, 2022.

Qingqing Zhu, Qiao Jin, Tejas S Mathai, Yin Fang, Zhizheng Wang, Yifan Yang, Maame Sarfo-Gyamfi, Benjamin Hou, Ran Gu, Praveen TS Balamuralikrishna, et al. Ct-bench: A benchmark for multimodal lesion understanding in computed tomography. *arXiv preprint arXiv:2602.14879*, 2026.

A Code and Data Availability

To facilitate scientific reproducibility and future research in multi-dimensional lesion understanding, we commit to open-sourcing our core assets while strictly adhering to patient privacy regulations. Our code and data are available at https://anonymous.4open.science/r/Lesion-CLIP.

B Dataset Details

B.1 Dataset Comparison

As detailed in Table 7, we compare our curated LiLA dataset with representative CT benchmarks across spatial modalities, granularity, and clinical validation. While existing large-scale datasets either focus on 2D slices with bounding boxes, such as DeepLesion (Yan et al., 2018) and CT-Bench (Zhu et al., 2026), or provide full 3D scans only at a coarse, scan-level granularity with high semantic sparsity, like CT-RATE (Hamamci et al., 2026), Rad-ChestCT (Draelos et al., 2021), and MedVL-CT69K (Shui et al., 2025), LiLA bridges this gap by offering a massive collection of 30,134 lesions annotated at a fine-grained, 3D lesion level. Furthermore, unlike segmentation-only 3D benchmarks such as LiTS (Bilic et al., 2023) and MSWAL (Wu et al., 2025), which lack clinical narratives or text-only datasets without spatial boundaries, LiLA integrates both voxel-level 3D lesion masks and detailed text reports, enabling multi-dimensional supervision for lesion understanding. Crucially, while existing medical VLP datasets primarily rely on noisy labels parsed from clinical reports, LiLA distinguishes itself by incorporating pathology calibration, establishing a high-quality resource for multi-dimensional lesion understanding.

Table 7: Comparison of LiLA with representative CT datasets.

Dataset	Modality	Granularity	#Scans	#Lesions	Annotation	Pathology Calibration
2D Lesion-Level Datasets						
DeepLesion	2D CT	Lesion	34K	34K	Box	×
CT-Bench	2D CT	Lesion	8K	20K	Box & Report	×
3D Scan-Level Datasets						
CT-RATE	3D CT	Scan	26K	-	Report	×
Rad-ChestCT	3D CT	Scan	36K	-	Report	×
MedVL-CT69K	3D CT	Scan	272K	-	Report	×
3D Lesion-Level Datasets						
LiTS	3D CT	Lesion	201	2K	Mask	×
MSWAL	3D CT	Lesion	694	3K*	Mask	×
LiLA	**3D CT**	**Lesion**	**6K**	**30K**	**Mask & Report**	✓

* The number is calculated based on the released subset of data.

B.2 Dataset Statistics

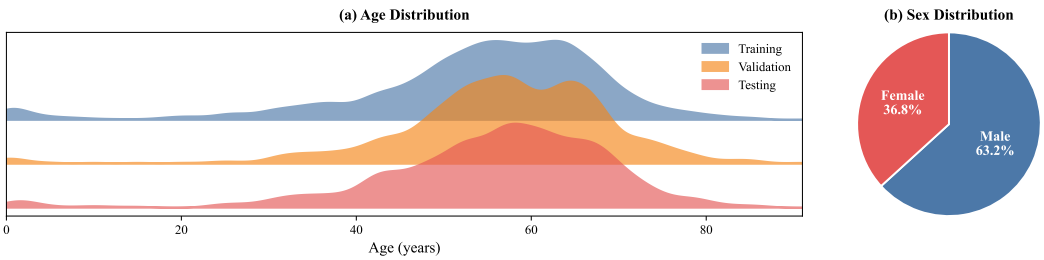


Figure 4: Patient demographic distribution on the LiLA dataset.

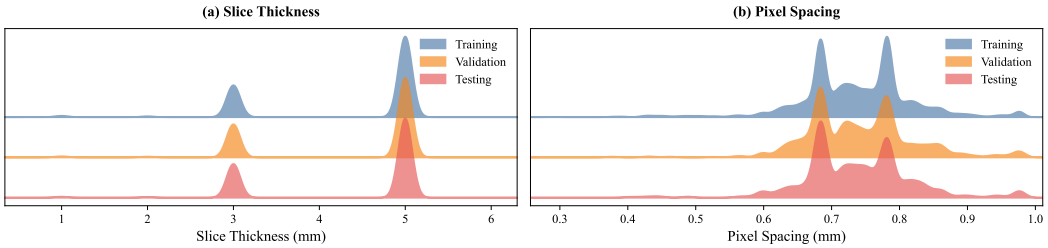


Figure 5: CT acquisition parameter distribution on the LiLA dataset.

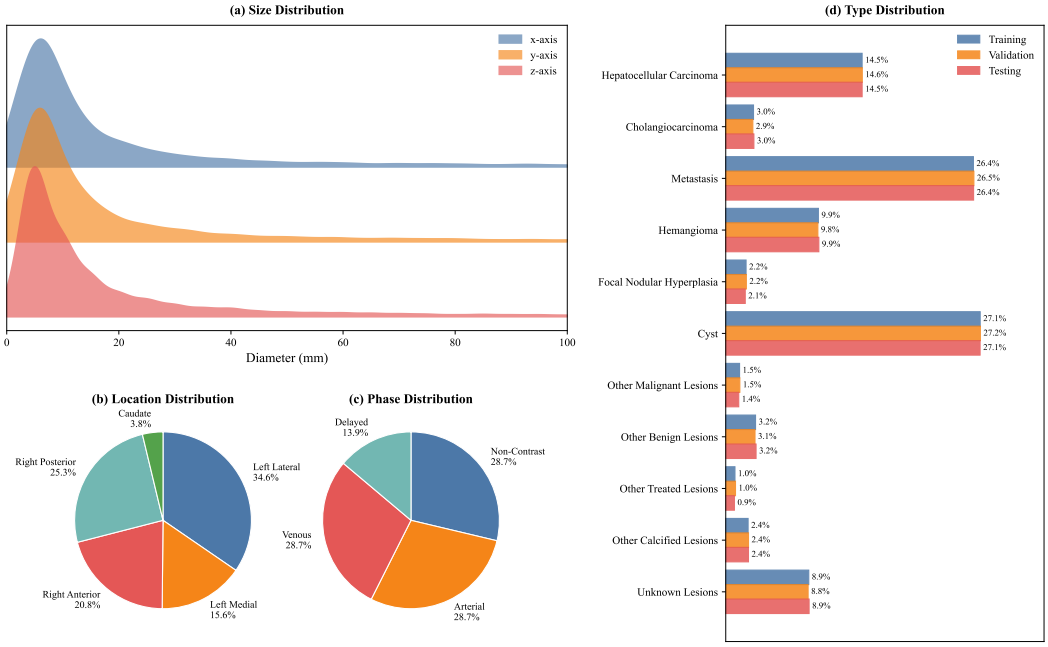


Figure 6: Lesion data distribution on the LiLA dataset.

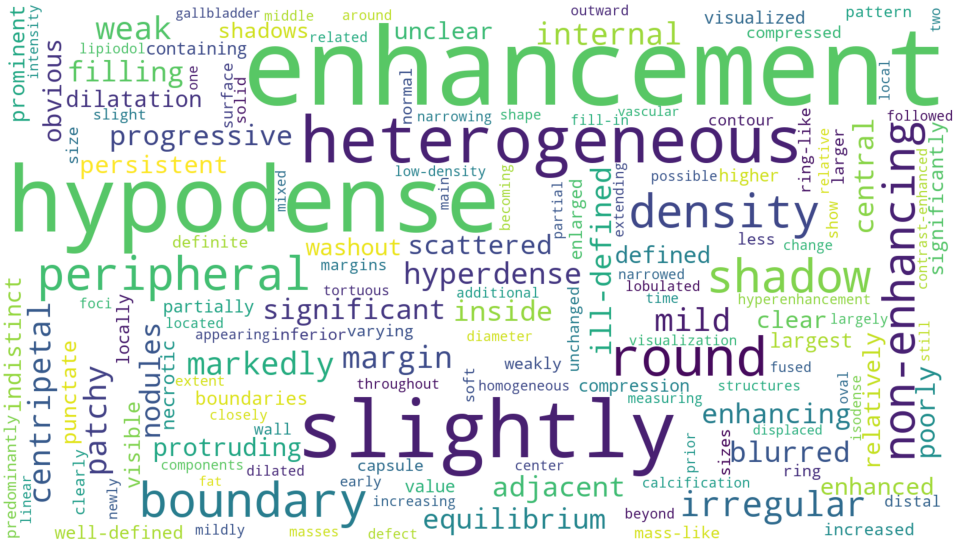


Figure 7: Lesion attribute word cloud on the LiLA dataset.

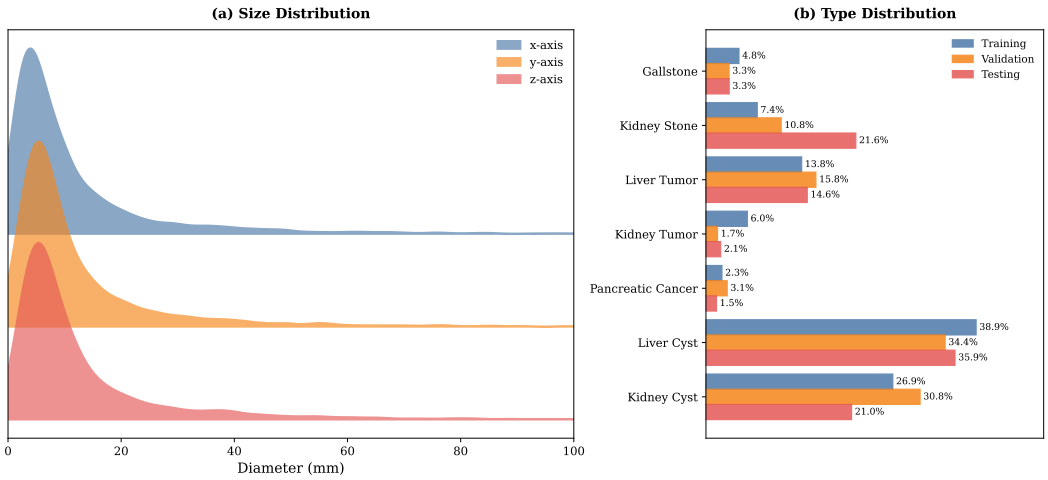


Figure 8: Lesion data distribution on the MSWAL dataset.

B.3 Prompt Design for Text Grounding

System Prompt

```
You are a medical imaging text grounding expert. Your task is to
    ground a target lesion in the {findings} of a radiology report
    using the provided {type}, {size}, and auxiliary context from
    {impressions}, and then return two fields: {location} and
    {attribute}.

Input fields:
- {type}: the curated label of the target lesion.
- {size}: a size clue used to identify the target lesion when
    multiple lesions are described.
- {impressions}: auxiliary context for interpreting the report.
- {findings}: the only source text for determining {location} and
    extracting {attribute}.

Location vocabulary:
Map the target lesion to one category in the fixed location
    vocabulary {vocabulary} for the corresponding anatomy. The
    vocabulary is anatomy-specific and includes "other regions".
- {vocabulary}: [...]

Instructions:
1. Identify the target lesion. Use {type} as the primary target
    label. Use {size} only as an auxiliary clue to match the target
    lesion.

2. Use {findings} as the extraction source. {impressions} may be
    used for context, but do not copy {attribute} from the
    {impressions} section.

3. Determine {location}. Use the anatomical location phrase of the
    target lesion to assign one category from the fixed location
    vocabulary {vocabulary}.

4. Extract {attribute}. {attribute} should describe imaging
    appearance, including density or signal intensity, morphology,
```

```
    margin, enhancement pattern, internal components, treatment-
    related appearance, and relevant adjacent-structure involvement.
- Do not include anatomical location words.
- Do not include explicit size values.
- Do not include reporting fillers such as "seen", "noted", "
    visualized", "also seen", or equivalent connectors unless they
    are necessary for grammaticality.
- Retain tightly associated perilesional findings, such as adjacent
     invasion, ductal or vascular changes, or treatment material,
    when they are part of the target lesion description.

5. Handle missing information. If the target lesion location cannot
     be determined from {findings} or mapped to {vocabulary}, return
     "other regions" for {location}. If the target lesion attribute
    cannot be matched in {findings}, return an empty string for
    {attribute}.

6. Return only {location} and {attribute}, without additional
    explanations.
```

User Prompt

```
Please process the following lesion information entry:

Type: {type}
Size: {size}
Impressions: {impressions}
Findings: {findings}

Return {location} and {attribute} according to the system
    instruction.
```

B.4 Prompt Templates for Vision-Language Pretraining

Table 8: Implemented prompt templates for multi-dimensional VLP.

Dimension	Prompt Template
Type	This is a CT patch of {type}.
Size	This is a CT patch of a lesion {size}.
Location	This is a CT patch of a lesion located in {location}.
Phase	This is a CT patch of a lesion in the {phase} phase.
Attribute	This is a CT patch of a lesion characterized by {attribute}.

C Details of Lesion Attribute Generation

As illustrated in Fig. 9, to generate free-text lesion attribute descriptions, we derive visual representations of multi-phase lesion patches from each VLP method using its frozen vision encoder and inherent projector. These representations are then aligned with a lightweight LLM, Qwen3-4B-Instruct-2507, via a two-layer Multi-Layer Perceptron (MLP). Specifically, the aligned visual representations are mapped into the input embedding space of the LLM to serve as soft visual tokens. Combined with a clinically oriented instruction, the LLM is prompted to generate lesion attribute descriptions. For cases with missing phases, the corresponding phase prefixes and their aligned soft visual tokens are simply omitted from the input prompt. During fine-tuning, we apply Low-Rank Adaptation (LoRA) (Hu et al., 2022) under a Causal Language Modeling (CLM) loss. The CLM

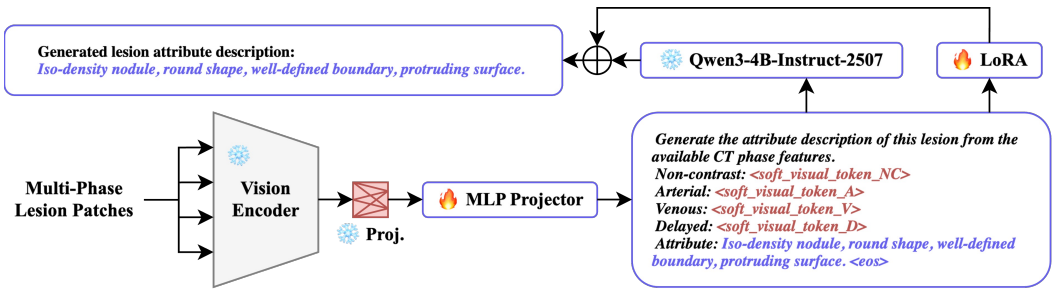


Figure 9: Schematic of LLM-driven lesion attribute generation. The red font highlights the phase-specific soft visual tokens aligned from the multi-phase lesion patches, while the purple font denotes the target lesion attribute descriptions used for loss optimization.

loss can be written as:

$$\mathcal{L}_{\text{CLM}} = -\sum_{t=1}^{T} \log P_\theta(y_t \mid y_{<t}, \mathbf{V}, \mathbf{X}_{\text{inst}}), \tag{7}$$

where θ denotes the trainable parameters including the MLP projector and LoRA adapters, y_t is the t-th token of the target lesion attribute description, $y_{<t}$ represents the previously generated tokens, $\mathbf{V}$ is the sequence of aligned soft visual tokens, and $\mathbf{X}_{\text{inst}}$ represents the clinical instruction tokens.

D Implementation Details

All experiments were performed using PyTorch 2.5.0 on an NVIDIA A800 GPU under Ubuntu 22.04. CT intensity values were clipped to $[-1000, 1000]$ Hounsfield units and linearly normalized to $[-1, 1]$ as network input. Model selection was consistently determined by peak validation performance. For all LLM-related processes, including text grounding and attribute generation, we consistently employed greedy decoding with the temperature set to 0 to ensure deterministic and reproducible outputs.

Lesion-CLIP was trained for 50 epochs with a batch size of 16, using the AdamW optimizer with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-2}, which took approximately 55.7 hours. CT-ViT (Hamamci et al., 2024) was adopted as the foundation for the vision encoder architecture. Specifically, the 3D ViT encoder was trained from scratch using a $4 \times 4 \times 4$ patch size and comprised 4 spatial and 4 temporal Transformer layers. Conversely, the text encoder was initialized with the pretrained weights of BiomedVLP-CXR-BERT-specialized (Boecking et al., 2022) and kept frozen during training. For the projection layers, we set the feature sizes defined in Sec. 3.2 to $D_v = 512$, $D_t = 512$, and $D = 512$. All projectors are implemented as bias-free linear layers using orthogonal initialization with a gain of 1.0. The learnable log scale parameter for each dimension was initialized to $\theta_m = 2.659$, corresponding to the standard CLIP temperature of 0.07 (Radford et al., 2021). Mixed-precision training with FP16 autocasting was employed to accelerate computation.

For linear probing, a single linear classifier was trained on the image latent features extracted from the frozen pretrained Lesion-CLIP model. Specifically, it was trained using the AdamW optimizer with a batch size of 16, a learning rate of 1×10^{-2}, and a weight decay of 1×10^{-2} over 500 epochs.

For attribute generation, an MLP projector was designed to map the visual representations to the input space of Qwen3-4B-Instruct-2507. To match the hidden size of the LLM, this projector was configured with a hidden size of 2048 and an output size of 2560, with its learnable scaling factor initialized to 0.1 to balance visual and text magnitudes. The LLM was fine-tuned using LoRA with a rank of $r = 8$, $\alpha = 16$, and a dropout rate of 0.05. This optimization was performed for 5 epochs with a batch size of 8, using the AdamW optimizer with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-4} under BF16 precision.

For all competing methods, input patches were resized to match each pretrained encoder's native resolution. For 2D CLIP and BiomedCLIP, we extracted the central axial slice and replicated it into a three-channel format.

E DETAILED RESULTS

E.1 COMPARISON WITH FULLY SUPERVISED MODELS

To complement Table 1, Fig. 10 provides a class-by-class breakdown of Lesion-CLIP's zero-shot performance against fully supervised models across all 11 lesion types in LiLA. Notably, while zero-shot accuracy fluctuates across categories, Lesion-CLIP demonstrates a consistent advantage in AUC, outperforming CT-ViT and 3D DenseNet-121 in all categories, and 3D ResNet-50 in 9 out of 11, indicating robust ranking capability.

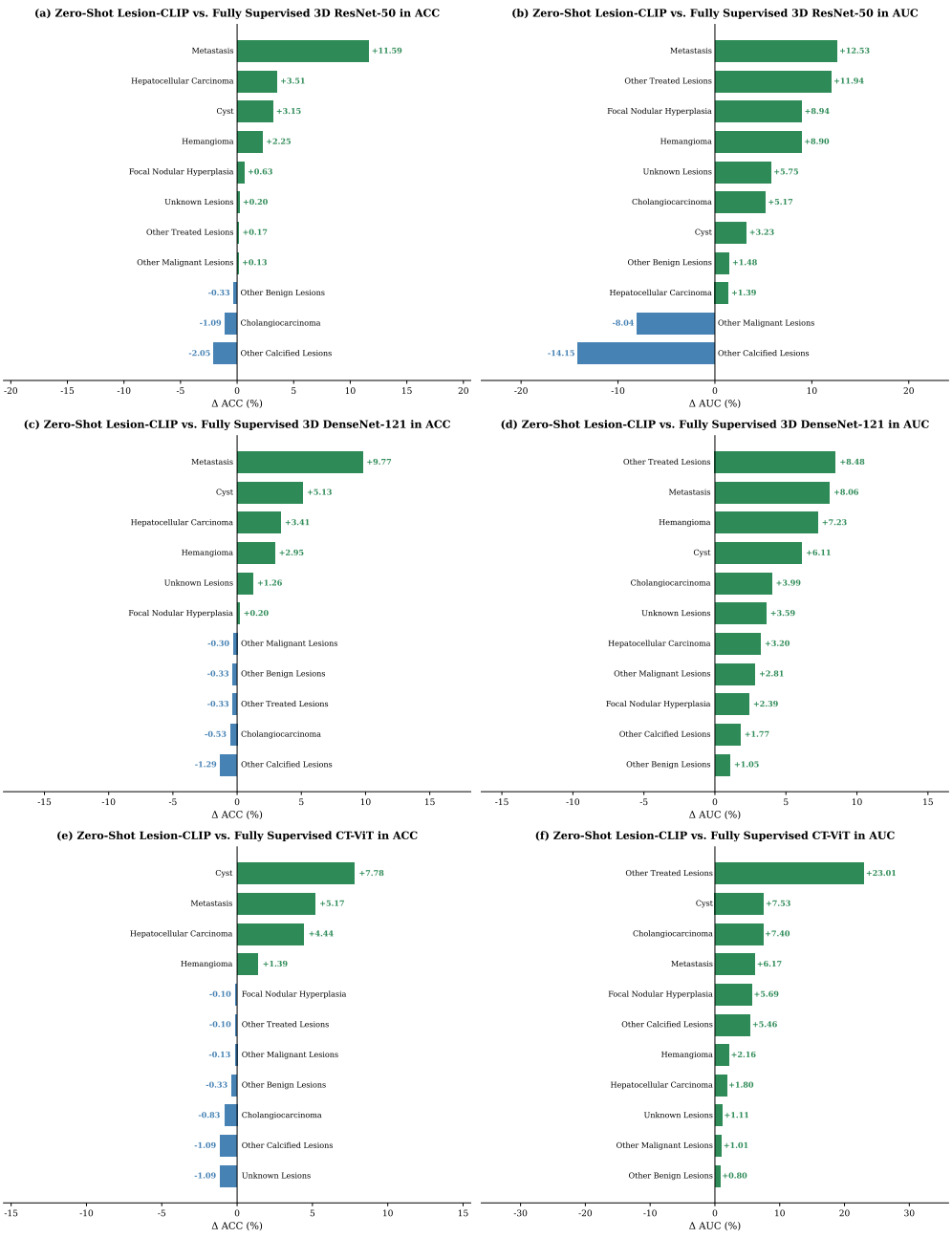


Figure 10: Zero-shot Lesion-CLIP is competitive with fully supervised models on the LiLA dataset.

E.2 Per-Phase Performance

To further validate the multi-phase robustness of Lesion-CLIP, we evaluate its zero-shot lesion classification performance across individual imaging phases on LiLA. As presented in Fig. 11, the model achieves consistently stable accuracy and AUC across non-contrast (60.16% / 82.44%), arterial (58.31% / 82.89%), and venous (61.78% / 84.12%) phases, with a moderate decrease in the delayed (53.68% / 78.48%) phase due to its smaller sample size. Notably, as 3D masks were originally annotated on venous-phase images and propagated via registration, the peak performance on the venous phase is partly attributable to its superior annotation fidelity.

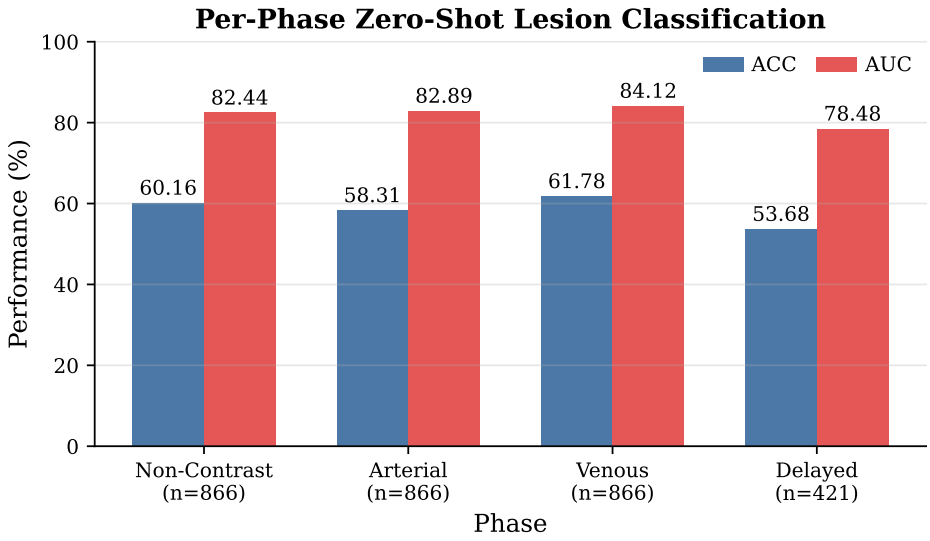


Figure 11: Zero-shot lesion classification performance of Lesion-CLIP across different imaging phases on the LiLA dataset.

E.3 Per-Class Performance

Fig. 12 illustrates the per-class zero-shot AUC performance across the 11 lesion types on the LiLA dataset. Lesion-CLIP achieves a macro-averaged AUC of 84.35%, with performance remaining reasonably balanced across classes. Specifically, 8 of the 11 types exceed an AUC of 80%, and even the lowest-performing class remains close to 70%.

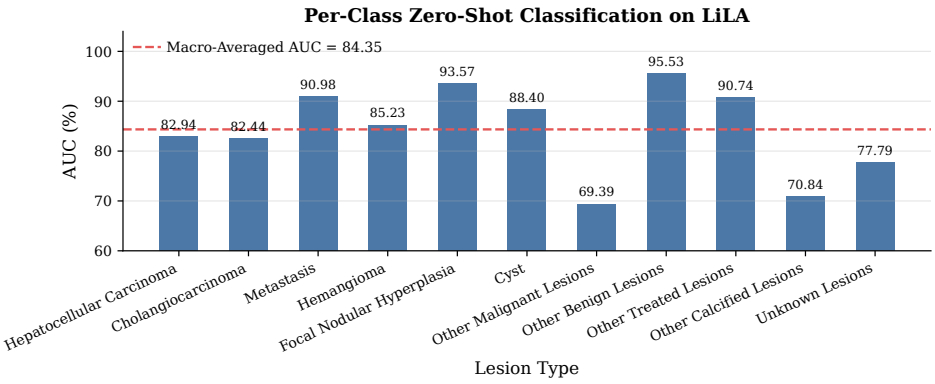


Figure 12: Per-class zero-shot lesion classification performance of Lesion-CLIP on the LiLA dataset.

To further investigate whether Lesion-CLIP generalizes to unseen anatomical domains and captures fine-grained features rather than relying on organ-level shortcuts, we break down the linear probing performance by class on MSWAL in Fig. 13. Although pretrained solely on the liver-centric data, Lesion-CLIP demonstrates strong cross-organ feature transferability, achieving a 98.70% AUC on pancreatic cancer and 90.42% AUC on kidney cysts. Crucially, the model's resistance to anatomical shortcut learning is validated by its precise intra-organ discrimination in multi-pathology organs. For instance, it successfully differentiates kidney stones (90.68% AUC), kidney cysts (90.42% AUC), and kidney tumors (80.35% AUC) within the renal environment. Since these renal lesions share identical anatomical backgrounds, their consistently strong AUC scores indicate that the frozen representations capture lesion-specific features rather than merely reflect host organ characteristics.

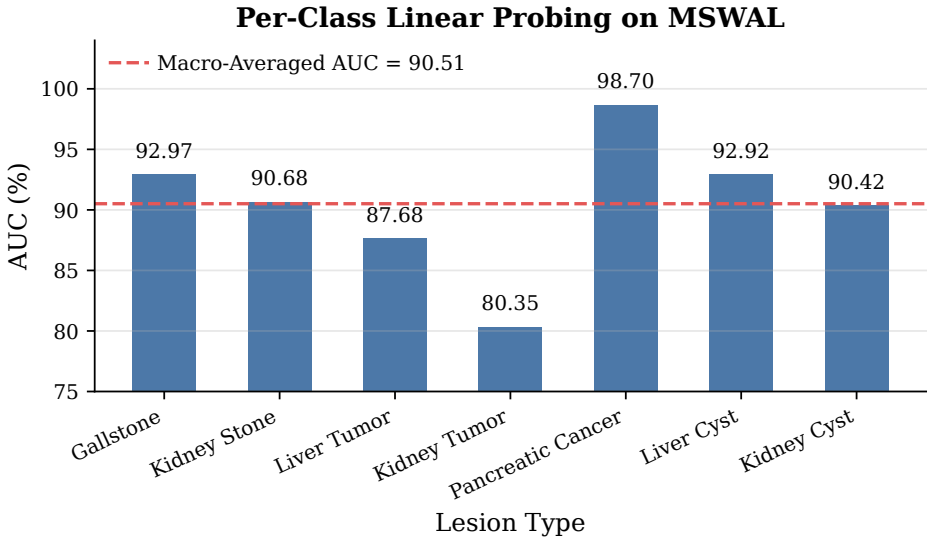


Figure 13: Per-class lesion classification performance of Lesion-CLIP under full-data linear probing on the MSWAL dataset.

E.4 Case Studies

To qualitatively evaluate Lesion-CLIP, we analyze its clinical efficacy across both discriminative and generative tasks. First, in the zero-shot lesion classification task presented in Fig. 14, our model demonstrates robust similarity-ranking capability. In the illustrated success cases, the ground-truth categories are correctly ranked first with clear similarity margins. Even in the presented failure cases, the true type is still ranked as the runner-up with a marginal similarity gap, and the highly ranked candidates remain clinically plausible and aligned with the ground-truth benign or malignant status. Second, in the lesion attribute generation task illustrated in Fig. 15, Lesion-CLIP captures and describes multi-phase phenotypic characteristics. As depicted in the representative cases, the model identifies temporal enhancement patterns and translates them into descriptive text, while faithfully reflecting the absence of contrast enhancement on corresponding non-enhancing cases.

E.5 Subspace Analysis

To assess the generalizability of our disentangled representations beyond the LiLA domain, we extend the subspace geometry analysis to the external MSWAL dataset. As depicted in Fig. 16, the five projection subspaces on MSWAL exhibit topological characteristics remarkably consistent with those observed on LiLA. Specifically, t-SNE visualizations maintain clear inter-dimensional separation. Except for a clinically expected correlation of 0.60 between type and attribute, pairwise linear CKA similarities remain low (mean < 0.3), suggesting preserved semantic specificity. Furthermore, all principal angles remain consistently above 10°, preserving the distinct representation manifolds. This cross-dataset geometric stability indicates that Lesion-CLIP captures robust structural priors of multi-dimensional lesion semantics.

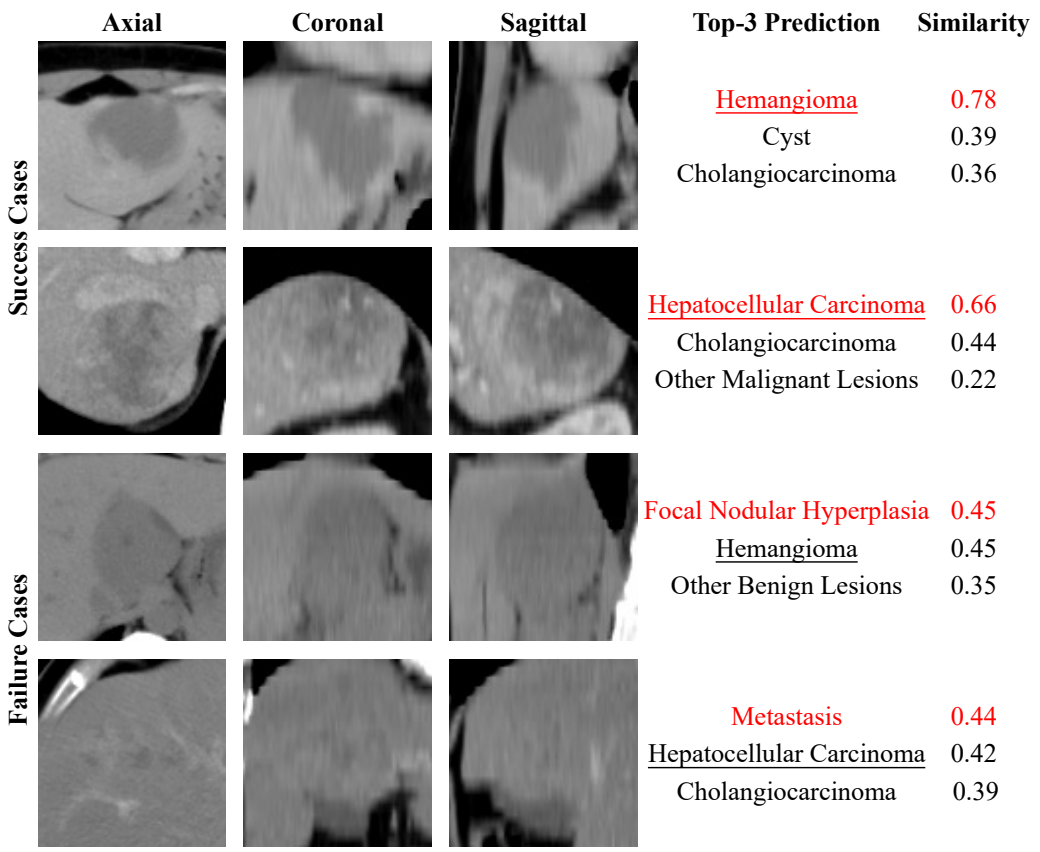


Figure 14: Case studies of Lesion-CLIP for zero-shot lesion classification on the LiLA dataset. The red font indicates the model's top-1 prediction, and the underline represents the ground truth.

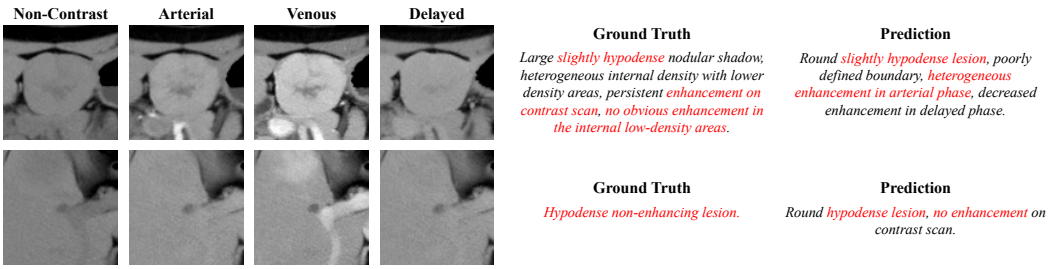


Figure 15: Case studies of Lesion-CLIP for lesion attribute generation on the LiLA dataset. The red font indicates parts where the model's prediction is highly consistent with the ground truth.

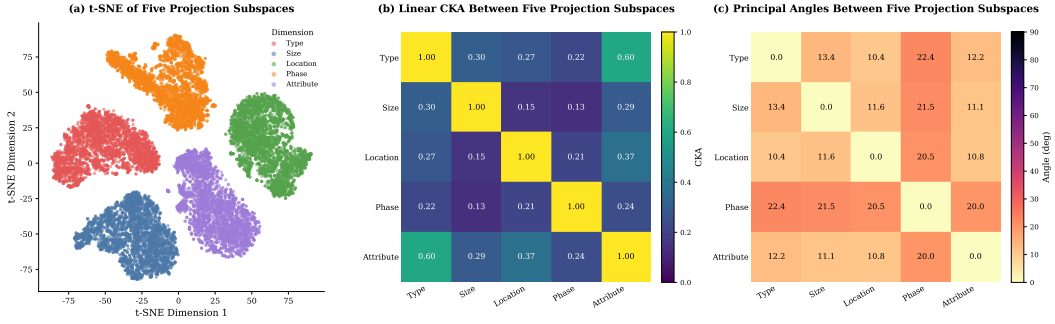


Figure 16: Geometry of the five Lesion-CLIP projection subspaces on the MSWAL dataset.