ΔREPRESENTATION: GEOMETRY SUPERVISED REPRESENTATION LEARNING OF PHENOTYPES VIA COUNTERFACTUAL REASONING FOR MEDICAL VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical vision-language models (VLMs) have shown increasing potential for radiological image interpretation. Medical VLMs encode radiological images into visual representations that capture both anatomical and phenotypic information for diagnosis. Existing approaches improve pathological phenotype representations through semantic-guided representation alignment. However, pathological phenotypes arise as lesion-specific visual changes superimposed on underlying normal anatomy. Such semantic alignment approaches fail to model the phenotype-specific increment relative to the corresponding normal anatomical representation. To address this gap, we propose **ΔRepresentation**, a visual phenotype representation learning framework based on counterfactual reasoning for medical VLMs. It comprises **BaseAnatomy**, a geometry-supervised representation learning module, and **ΔPhenotype**, a counterfactual incremental representation learning module. BaseAnatomy provides fine-grained geometric supervision through spatial relationships across and within anatomical structures. ΔPhenotype computes the representation increment between lesion representations and their corresponding normal anatomical representations, and supervises increments associated with the same phenotype to cluster in the representation space. Experiments on *ReXGroundingCT* and *LIDC-IDRI* demonstrate that ΔRepresentation effectively structures pathological phenotype representations and improves lesion grounding and phenotype characterization accuracy in medical VLMs. Code is available at https://anonymous.4open.science/r/deltarep-CF6D.

1 INTRODUCTION

Medical vision-language models (VLMs) have been developed to assist clinicians in diagnosis based on radiological images Wu et al. (2025); Xu et al. (2026). Through large-scale image-text pretraining, medical VLMs establish associations between radiological images and clinical texts using vision and text encoders. The vision encoder extracts pathology-related visual information into representations, which are subsequently aligned with semantic representations to support diagnostic reasoning. However, vision encoders trained under this paradigm lack explicit modeling of anatomical structures and pathological visual phenotypes in the visual representation space. Consequently, potentially inaccurate visual representations can be aligned with semantic representations and used as evidence for diagnosis.

To reduce such incorrect associations, existing studies have explored fine-grained representation alignment to connect visual phenotypic representations with clinical semantic representations. TRRG Wang et al. (2025) uses report sentences as semantic supervision and aligns sentence embeddings with image-level visual features through contrastive learning. PRIOR Cheng et al. (2023) further performs local alignment between regional visual features and report sentences through attention mechanisms, establishing finer-grained correspondences between abnormalities and their textual descriptions. MedKLIP Wu et al. (2023) introduces pathology concepts extracted from radiology reports and uses them to guide the learning of disease-specific visual representations. MCA-RG Xing et al. (2025) aligns visual features with pathology concepts and introduces pathology–anatomy matching to emphasize disease-related regions. RadAlign Gu et al. (2025) uses learnable

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.

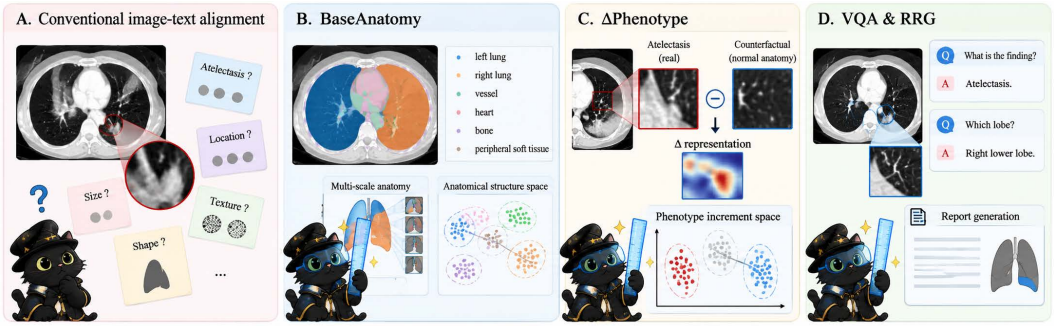

Figure 1: BaseAnatomy structures the anatomical visual representation space, while ΔPhenotype organizes lesion-specific increments relative to counterfactual normal anatomy.

concept tokens to extract attribute-specific visual features and aligns them with textual descriptions of diagnostic criteria through contrastive learning.

However, learning from paired image-text data cannot explicitly establish the incremental relationship between pathological phenotypes and their underlying anatomical structures. Semantic representations derived from text describe the presence and subclasses of pathological phenotypes. They do not explicitly represent the phenotypic changes relative to the corresponding normal anatomy. Moreover, these semantic representations provide limited supervision for separating visually similar phenotypes in the representation space. This limits the vision encoder's ability to learn visual changes associated with pathological phenotypes and distinguish different phenotypes. To establish this relationship directly in the visual representation space, we define a pathological phenotype as the visual change relative to its underlying normal anatomy and construct this phenotype representation by comparing a pathological patch with its corresponding normal counterpart.

To implement this theory, we propose **ΔRepresentation**, a visual representation learning framework for medical VLMs that integrates anatomical geometry supervision with counterfactual phenotype learning as shown in Figure 1. First, **BaseAnatomy** supervises visual embeddings using ideal spatial distributions that preserve geometric relationships between anatomical structures and spatial continuity within each structure. This establishes stable normal anatomical representations as references for learning pathological changes. Building on these references, **ΔPhenotype** uses counterfactual reasoning to learn phenotype-specific increments between the representations of pathological patches and their corresponding normal counterparts. These increments explicitly characterize pathological visual changes relative to normal anatomy. Experiments on *ReXGroundingCT* and *LIDC-IDRI* demonstrate that ΔRepresentation consistently improves lesion grounding and phenotype characterization across medical VLMs. Representation analyses further show that the learned phenotype increments provide more discriminative visual representations for different pathological phenotypes. Our contributions are summarized as follows:

(1) We propose **ΔRepresentation**, a visual representation learning framework for medical VLMs that combines anatomical geometry supervision with counterfactual phenotype learning. It models pathological phenotypes as representation increments relative to normal anatomy.

(2) We develop **BaseAnatomy**, which organizes normal anatomical representations using ideal spatial distributions. It preserves inter-structure geometry and intra-structure spatial continuity to provide stable anatomical references.

(3) We introduce **ΔPhenotype**, which learns phenotype-specific increments through counterfactual reasoning. It progressively refines phenotype clusters as additional samples become available, improving intra-phenotype consistency and inter-phenotype separation.

2 Related Work

2.1 Medical VLMs in Radiology

Medical VLMs adapt pretrained visual and language models to medical interpretation through domain-specific multimodal training. LLaVA-Med combines biomedical figure-caption alignment

with instruction tuning to support medical visual conversation Li et al. (2023). LLaVA-Rad connects pretrained image and language components through a lightweight adapter and specializes the model using radiological image–text pairs Zambrano Chaves et al. (2025). MedGemma incorporates medical image encoder adaptation into a Gemma-based multimodal model Sellergren et al. (2025). Complementing model adaptation, MedRAX integrates chest X-ray analysis tools with multimodal models to address complex clinical queries Fallahpour et al. (2025).

The scope of radiology VLMs also extends to volumetric inputs and anatomically structured interpretation. M3D-LaMed unifies multiple 3D tasks, including report generation, question answering, positioning, and segmentation, within a multimodal instruction-learning framework Bai et al. (2024). Argus studies how encoder pretraining, visual token compression, and model and data scaling affect high-resolution CT report generation Liu et al. (2025). AOR introduces anatomy-centric, region-level reasoning to structure chest X-ray interpretation Li et al. (2026).

2.2 Fine-Grained Alignment in Medical VLMs

Fine-grained alignment methods link local image regions to clinically meaningful descriptions. BiRD learns region–language correspondence from grounding instructions constructed from segmentation datasets, while MAIRA-2 associates radiological findings with bounding boxes during report generation Huang et al. (2024); Bannur et al. (2024). Anatomy-aware methods further organize visual inputs around explicit anatomical regions. Reg2RG uses segmentation masks to extract regional features for region-specific report generation, whereas MedRegion-CT combines region-guided tokenization with pseudo-mask supervision and structured lesion attributes Chen et al. (2025); Kyung et al. (2025). These approaches improve local correspondence through regional feature construction, explicit supervision, or region-specific inputs to the language model.

Concept-guided methods further associate visual representations with clinically meaningful semantics. MCA-RG introduces anatomical and pathological concept banks, extracts concept-specific visual features through cross-attention, and combines contrastive learning with feature gating Xing et al. (2025). RadAlign aligns visual concept tokens with textual diagnostic criteria to support disease classification and report generation Gu et al. (2025). Although these methods strengthen regional and concept-specific correspondence, alignment with text-derived anatomical semantics does not explicitly constrain visual embeddings to preserve geometric relationships between anatomical structures or spatial continuity within each structure. Furthermore, phenotypic semantic supervision does not explicitly characterize the counterfactual changes between pathological patches and their normal counterparts, leaving phenotype-specific representation increments unmodeled. These limitations hinder stable anatomical encoding and the learning of phenotype-specific changes relative to normal anatomy, limiting the vision encoder's ability to accurately represent pathological visual phenotypes.

3 Methodology

3.1 Overview

ΔRepresentation consists of two components as shown in Figure 2. **BaseAnatomy** shapes the normal anatomical representation space using ideal spatial distributions, which preserve geometric relationships between anatomical structures and spatial continuity within each structure. **ΔPhenotype** then constructs counterfactual normal references for pathological patches and learns phenotype-specific representation increments by comparing pathological representations with their corresponding normal representations. These increments are progressively refined through phenotype-specific clustering as additional lesion samples become available. BaseAnatomy is trained first, after which its anatomical encoder is frozen to provide stable normal references during ΔPhenotype learning.

3.2 BaseAnatomy: Building Geometry-Supervised Normal Anatomical Representation Space

BaseAnatomy learns an anatomically organized representation space with a visual encoder E_A, using annotation-defined ideal spatial distributions to preserve both relationships across anatomical structures and spatial continuity within each structure. Given a CT image $\mathbf{x}$, E_A extracts ℓ_2-

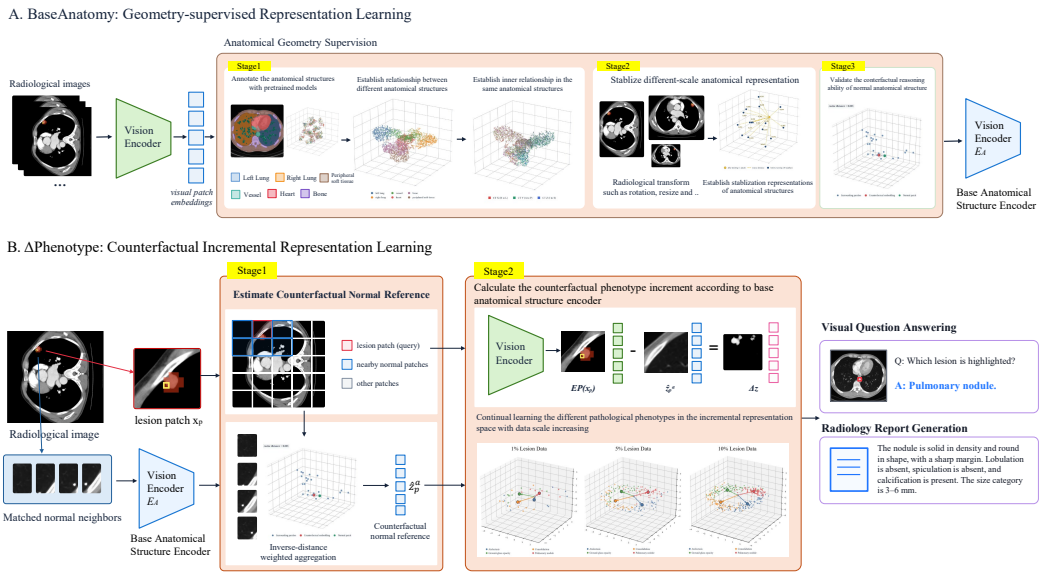

Figure 2: Overview of ΔRepresentation. (a) BaseAnatomy shapes an anatomically structured visual space with ideal geometric supervision that encodes anatomical relationships and spatial continuity. (b) ΔPhenotype constructs counterfactual normal references for lesion patches and learns phenotype-specific visual increments relative to normal anatomy for discriminative phenotype representation learning.

normalized patch representations $\{\mathbf{a}_p\}_{p=1}^{P}$. Each normal patch is associated with an anatomical label c_p and normalized CT coordinates $\mathbf{u}_p$, while slice-position encoding provides its relative position along the volumetric axis. Anatomical annotations define tissue regions, and lesion masks exclude pathological patches from normal anatomical supervision.

Ideal anatomical geometry supervision. We construct the ideal representation space from two complementary spatial relationships. Inter-structure geometry preserves the relative positions between different anatomical structures, while intra-structure geometry preserves the spatial arrangement of patches within the same structure. Specifically, the target distance between normal patches p and q is defined as

$$d_{pq}^{\star} = \sqrt{\alpha\, d_A(c_p, c_q)^2 + (1-\alpha)\|\mathbf{u}_p - \mathbf{u}_q\|_2^2}, \tag{1}$$

where $d_A(c_p, c_q)$ denotes the normalized distance between the centers of their anatomical structures, and α balances inter-structure geometry and intra-structure spatial continuity. For patches from the same anatomical structure, $d_A(c_p, c_q) = 0$, such that their target relationship is determined by their relative spatial positions. This ideal distribution provides direct geometric supervision for anatomical representations without relying on text-derived anatomical semantics.

Geometry-supervised anatomical representation learning. We optimize E_A by matching pairwise cosine distances between patch representations to their corresponding ideal distances using a smooth-L_1 objective. Anatomical compactness encourages patches belonging to the same structure to form coherent distributions, while neighborhood constraints preserve local spatial continuity within each structure. We further apply rotation and scale transformations with matched spatial correspondence, encouraging representations of the same anatomical region to remain consistent across transformed views. Anatomical classes are sampled in a balanced manner to prevent large structures from dominating the learned space. Together, these constraints establish stable normal anatomical representations that serve as references for subsequent phenotype learning. Anatomical annotations and lesion masks are used only as supervision during training.

3.3 ΔPHENOTYPE: LEARNING COUNTERFACTUAL PHENOTYPE INCREMENTS FROM NORMAL ANATOMY

ΔPhenotype learns pathological visual phenotypes as representation increments relative to their underlying normal anatomy. We initialize a frozen anatomical encoder E_A and a trainable phenotype encoder E_P from the same BaseAnatomy checkpoint. E_A preserves the learned anatomical representation space and provides stable normal references, while E_P is optimized to encode phenotype-specific pathological changes. Only the final vision layers and normalization layer of E_P are updated during phenotype learning.After representation learning, only E_P is retained as the VLM vision encoder for downstream inference; E_A and all anatomical/lesion annotations are used only during training.

Counterfactual normal reference and phenotype increment. For each pathological patch, we construct a counterfactual representation of the normal anatomy at the same location. Since BaseAnatomy preserves spatial continuity within anatomical structures, this normal counterpart can be estimated from nearby mask-negative patches belonging to the corresponding anatomical region. For a pathological patch p, we define

$$\hat{\mathbf{a}}_p = \mathrm{norm}\left(\sum_{q\in\mathcal{N}(p)} w_{pq}\mathbf{a}_q\right), \qquad \boldsymbol{\delta}_p = E_P(\mathbf{x})_p - \hat{\mathbf{a}}_p, \tag{2}$$

where $\mathcal{N}(p)$ contains anatomically matched normal neighbors, w_{pq} denotes normalized inverse-distance weights, and $\hat{\mathbf{a}}_p$ represents the estimated normal anatomical representation. The resulting increment $\boldsymbol{\delta}_p$ isolates the pathological visual change relative to this counterfactual normal reference.

Phenotype-specific clustering and continual refinement. We organize the learned increments according to their pathological phenotype labels. Patch-level increments from the same annotated lesion are aggregated into a region-level phenotype representation. Increments associated with the same phenotype are encouraged to remain consistent, while those from different phenotypes are separated in the representation space. Localization supervision further distinguishes lesion increments from hard normal patches, while anatomical preservation constrains normal-patch representations of E_P to remain close to those produced by the frozen E_A. No-lesion samples additionally suppress pathological responses in normal tissues.

Starting from limited lesion samples, these objectives establish an initial organization of phenotype-specific increments. As additional lesion samples become available, their increments are incorporated into the same learning process, continually refining the phenotype distributions while the normal anatomical reference space remains fixed. This enables E_P to progressively improve the encoding and discrimination of pathological visual changes across different anatomical contexts.

4 EXPERIMENTS

4.1 DATASETS AND EVALUATION

We conduct training and evaluation on two public chest CT datasets, **ReXGroundingCT** Baharoon et al. (2026) and **LIDC-IDRI** Armato III et al. (2011), which provide complementary annotations for pulmonary abnormalities. From ReXGroundingCT, we select 2,720 training, 230 development, and 283 test axial CT images covering four target phenotypes: atelectasis, consolidation, ground-glass opacity, and pulmonary nodule. From LIDC-IDRI, we use 681 training, 164 development, and 121 test images; the test set includes 71 nodule-positive and 50 no-lesion cases. All splits are patient-disjoint. For each image, we construct two visual question answering (VQA) questions and one radiology report generation (RRG) prompt. The test set therefore comprises 808 VQA questions (566 from ReXGroundingCT and 242 from LIDC-IDRI) and 404 RRG instances (283 and 121, respectively). For lesion localization, we divide each image into a 4×4 grid and define the reference cell as the one containing the largest annotated lesion area; no-lesion cases have a dedicated *No lesion* answer. We report *Grounding Accuracy* and *Phenotype Accuracy* for both VQA and RRG. Grounding Accuracy measures exact 4×4 grid-cell localization for VQA and the correctness of lesion-location attributes for RRG, while Phenotype Accuracy measures phenotype identification.Appendix A.

Models	Scale	ReXGroundingCT		LIDC-IDRI	
		Grounding Acc.	Phenotype Acc.	Grounding Acc.	Phenotype Acc.
RadFM	14B	10.60	4.95	18.18	2.48
w/ ΔRepresentation	∼14B	**50.53** ↑39.93	**38.87** ↑33.92	**62.81** ↑44.63	**47.11** ↑44.63
LLaVA-Med	7B	43.11	12.37	2.48	5.79
w/ ΔRepresentation	∼7B	**66.78** ↑23.67	**51.94** ↑39.57	**47.11** ↑44.63	**44.63** ↑38.84
Lingshu	7B	77.39	15.19	23.97	42.15
w/ ΔRepresentation	∼7B	**91.17** ↑13.78	**35.69** ↑20.50	**65.29** ↑41.32	**86.78** ↑44.63
Hulu-Med	4B	70.32	26.15	24.79	41.32
w/ ΔRepresentation	∼4B	**86.93** ↑16.61	**39.22** ↑13.07	**60.33** ↑35.54	**85.95** ↑44.63
MedGemma	4B	65.37	16.96	52.89	42.98
w/ ΔRepresentation	∼4B	**79.15** ↑13.78	**35.69** ↑18.73	**73.55** ↑20.66	**85.12** ↑42.14
MedGemma1.5	4B	43.82	15.55	33.06	43.80
w/ RadAlign	∼4B	46.29	14.84	54.55	34.71
w/ MCA-RG	∼4B	46.29	40.64	54.55	36.36
w/ Reg2RG	∼4B	45.94	40.64	47.93	36.36
w/ ΔRepresentation	∼4B	**79.86** ↑36.04	**43.11** ↑27.56	**76.86** ↑43.80	**77.69** ↑33.89

Table 1: Comparison with pretrained medical VLMs and alignment methods on the VQA task.

Models	Scale	ReXGroundingCT		LIDC-IDRI	
		Grounding Acc.	Phenotype Acc.	Grounding Acc.	Phenotype Acc.
RadFM	14B	3.18	4.95	4.96	6.61
w/ ΔRepresentation	∼14B	**14.13** ↑10.95	**15.55** ↑10.60	**34.71** ↑29.75	**37.19** ↑30.58
LLaVA-Med	7B	4.95	41.34	14.05	58.68
w/ ΔRepresentation	∼7B	**19.79** ↑14.84	**57.24** ↑15.90	**52.07** ↑38.02	**94.21** ↑35.53
Lingshu	7B	0.71	3.89	22.31	35.54
w/ ΔRepresentation	∼7B	**12.01** ↑11.30	**20.49** ↑16.60	**48.76** ↑26.45	**74.38** ↑38.84
Hulu-Med	4B	2.47	8.48	38.84	42.15
w/ ΔRepresentation	∼4B	**12.72** ↑10.25	**18.73** ↑10.25	**52.07** ↑13.23	**86.78** ↑44.63
MedGemma	4B	0.35	3.53	13.22	22.31
w/ ΔRepresentation	∼4B	**14.49** ↑14.14	**15.90** ↑12.37	**44.63** ↑31.41	**66.94** ↑44.63
MedGemma1.5	4B	0.71	5.30	29.75	39.67
w/ RadAlign	∼4B	18.02	43.46	28.93	30.58
w/ MCA-RG	∼4B	26.15	51.24	34.71	38.02
w/ Reg2RG	∼4B	26.15	51.24	34.71	38.02
w/ ΔRepresentation	∼4B	**29.33** ↑28.62	**52.65** ↑47.35	**50.41** ↑20.66	**75.21** ↑35.54

Table 2: Comparison with pretrained medical VLMs and alignment methods on the RRG task.

4.2 Experimental Setup

We conduct a comprehensive evaluation of ΔRepresentation from both downstream performance and module effectiveness. First, we compare ΔRepresentation across six pretrained medical VLMs Wu et al. (2025); Li et al. (2023); Xu et al. (2026); Jiang et al. (2025); Sellergren et al. (2025; 2026) and with existing representation alignment methods Gu et al. (2025); Xing et al. (2025); Chen et al. (2025) on VQA and RRG tasks to evaluate anatomical grounding and pathological phenotype characterization (Tables 1 and 2). For each VLM, ΔRepresentation is initialized from and trained on its native vision encoder, while retaining the corresponding language backbone. We then analyze BaseAnatomy by visualizing anatomical representations before and after training, examining their spatial continuity and cross-scale consistency, and comparing inferred counterfactual normal representations with their corresponding real normal representations (Figs. 3 and 4). For ΔPhenotype, we progressively increase lesion-positive training data from 1% to 100% and evaluate both downstream performance and phenotype representation distributions, assessing whether additional samples continually refine intra-phenotype consistency and inter-phenotype separation (Table 3 and Fig. 5). Finally, we provide an end-to-end case study to illustrate how anatomical representations and phenotype-specific increments support pathological characterization and report generation (Fig. 6).

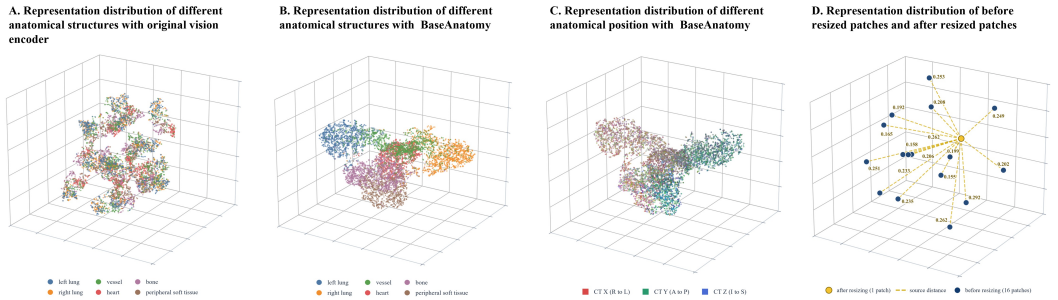

Figure 3: Anatomical representation analysis of BaseAnatomy. (A–B) Representation distributions before and after BaseAnatomy training. (C) Spatial continuity across anatomical positions. (D) Cross-scale consistency of anatomical representations.

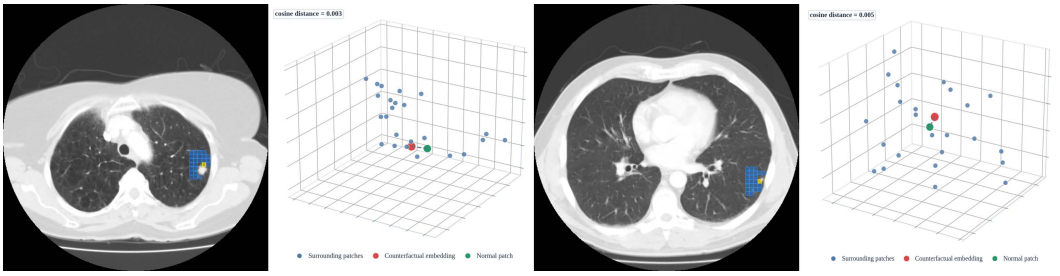

Figure 4: Counterfactual representation analysis. Comparison between inferred counterfactual representations and their corresponding normal representations.

4.3 Comparison with Different Medical VLMs and Alignment Methods

Table 1 presents the comparison on the VQA task. To evaluate the effectiveness and generalizability of ΔRepresentation, we apply it to six pretrained medical VLM backbones on ReXGroundingCT and LIDC-IDRI. ΔRepresentation consistently improves anatomical grounding and pathological phenotype characterization across all evaluated backbones and datasets. For example, with MedGemma1.5, the absolute gains in grounding and phenotype accuracies are 36.04% and 27.56% on ReXGroundingCT, and 43.80% and 33.89% on LIDC-IDRI, respectively. Averaged across the six backbones and two datasets, ΔRepresentation achieves absolute improvements of 31.20% in grounding accuracy and 33.51% in phenotype accuracy, demonstrating its effectiveness across different pretrained medical VLMs. Furthermore, we compare ΔRepresentation with representative visual–text alignment methods, including RadAlign, MCA-RG, and Reg2RG, under the same MedGemma1.5 backbone. ΔRepresentation achieves the highest grounding and phenotype accuracies on both datasets among these methods. The proposed framework combines anatomical supervision based on ideal spatial distributions with counterfactual learning of phenotype-specific representation increments. These results support its effectiveness in improving anatomical localization and pathological phenotype characterization.

Table 2 further reports the comparison on the RRG task. To evaluate the effectiveness of ΔRepresentation for open-ended RRG, we apply it to the same six pretrained medical VLM backbones. ΔRepresentation consistently improves anatomical grounding and pathological phenotype characterization across both datasets, demonstrating that its benefits extend to free-form report generation. For example, with MedGemma1.5, the absolute gains in grounding and phenotype accuracies are 28.62% and 47.35% on ReXGroundingCT, and 20.66% and 35.54% on LIDC-IDRI, respectively. Similar improvements are observed across the other backbones, with average absolute gains of 20.80% in grounding accuracy and 28.57% in phenotype accuracy across all backbone–dataset combinations. Furthermore, compared with RadAlign, MCA-RG, and Reg2RG under the same MedGemma1.5 backbone, ΔRepresentation achieves the highest grounding and phenotype accuracies on both datasets. These results demonstrate the effectiveness of the proposed visual representation learning framework in improving the anatomical and pathological accuracy of generated reports. Detailed phenotype evaluation results are provided in Appendix B.

Lesion Data Scale	ReXGroundingCT (VQA)		ReXGroundingCT (RRG)		LIDC-IDRI (VQA)		LIDC-IDRI (RRG)	
	Grounding Acc.	Phenotype Acc.	Grounding Acc.	Phenotype Acc.	Grounding Acc.	Phenotype Acc.	Grounding Acc.	Phenotype Acc.
1%	54.06	25.80	10.95	15.55	43.80	54.55	40.50	50.41
5%	72.44 ↑18.38	34.63 ↑8.83	10.95 ↑0.00	15.55 ↑0.00	44.63 ↑0.83	62.81 ↑8.26	40.50 ↑0.00	51.24 ↑0.83
10%	75.27 ↑2.83	39.93 ↑5.30	22.26 ↑11.31	31.45 ↑15.90	60.33 ↑15.70	68.60 ↑5.79	43.80 ↑3.30	53.72 ↑2.48
25%	78.09 ↑2.82	40.64 ↑0.71	23.32 ↑1.06	43.11 ↑11.66	73.55 ↑13.22	77.69 ↑9.09	44.63 ↑0.83	69.42 ↑15.70
50%	79.15 ↑1.06	43.11 ↑2.47	28.27 ↑4.95	50.18 ↑7.07	76.03 ↑2.48	77.69 ↑0.00	47.11 ↑2.48	75.21 ↑5.79
100%	**79.86** ↑0.71	**43.11** ↑0.00	**29.33** ↑1.06	**52.65** ↑2.47	**76.86** ↑0.83	**77.69** ↑0.00	**50.41** ↑3.30	**75.21** ↑0.00

Table 3: Performance on ReXGroundingCT and LIDC-IDRI with different proportions of lesion data for the VQA and RRG tasks.

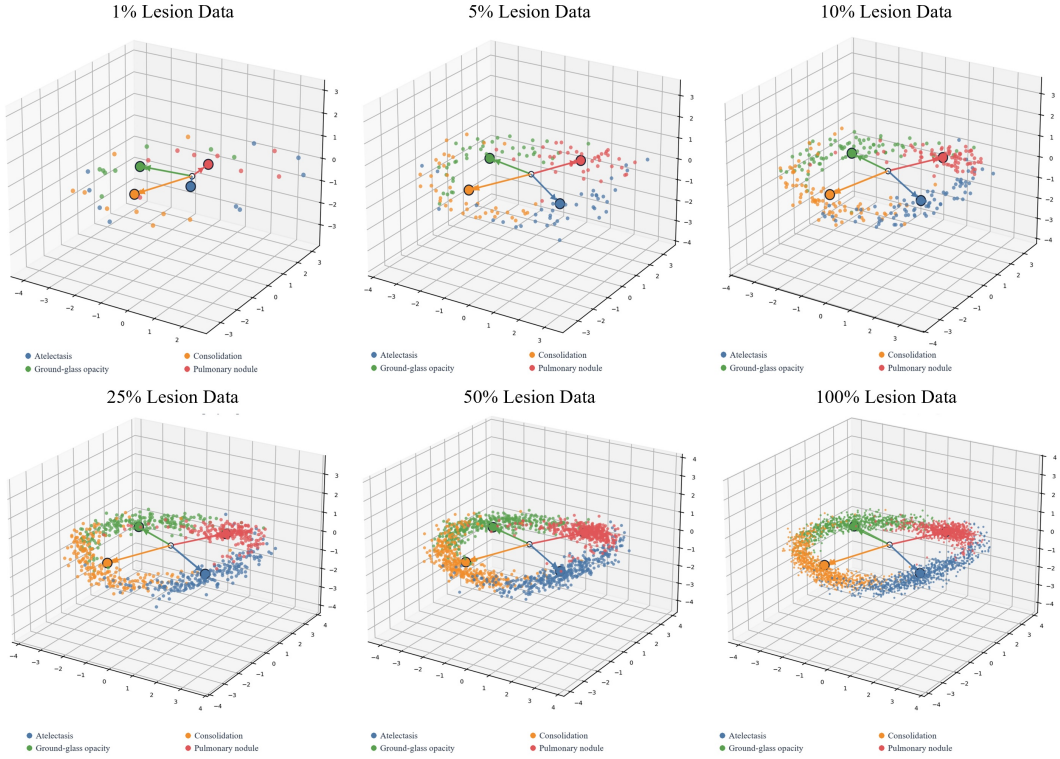

Figure 5: Representation learning performance of visual phenotypes in different lesion data scales.

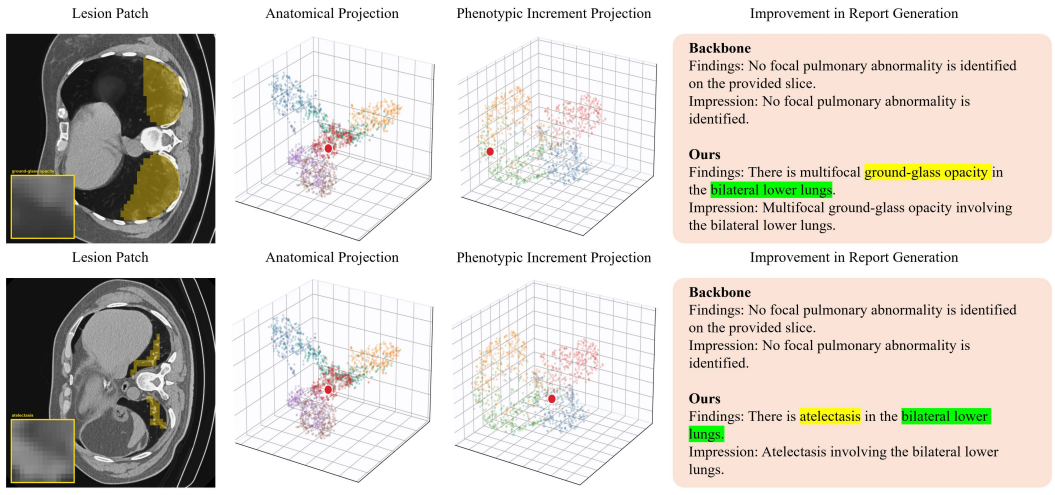

Figure 6: Case study of ΔRepresentation, illustrating anatomical projection by BaseAnatomy, phenotype-increment projection by ΔPhenotype, and their contribution to report generation.

4.4 Spatial Organization and Scale Consistency of Visual Anatomical Representations with BaseAnatomy

Figure 3 compares anatomical patch representations before and after training. With the original vision encoder, patches from different anatomical structures are scattered and substantially intermixed in the t-SNE projection (Fig. 3A). After training, patches form more coherent clusters whose relative arrangement broadly reflects the spatial relationships among anatomical structures (Fig. 3B). Within each structure, patch representations also vary gradually with their positions along the three CT spatial axes (Fig. 3C), indicating that the encoder retains fine-grained location information rather than anatomical identity alone. We further examine whether these representations remain stable across image scales. After downsampling the image, a single patch covers the region occupied by 16 patches at the original scale. Its representation lies near the center of the original patches' representation distribution (Fig. 3D). This cross-scale consistency suggests that learning fine-grained anatomical spatial relationships enables the vision encoder to produce stable anatomical representations despite changes in image scale.

We further validate the counterfactual normal representations constructed for pathological regions. As shown in Fig. 4, the inferred counterfactual embeddings lie close to the corresponding real normal representations in the anatomical representation space, with cosine distances of only 0.003 and 0.005 in the two representative cases. This indicates that BaseAnatomy provides a sufficiently structured anatomical space for estimating the normal representation underlying a pathological patch. The resulting counterfactual representations therefore provide reliable references for calculating phenotype-specific increments in ΔPhenotype.

4.5 Continual Representation Learning Performance of Visual Phenotypes and t-SNE Visualization in Different Data Scale

Table 3 and Fig. 5 examine the effect of lesion-positive training data on phenotype representation learning. As the training scale increases from 1% to 100%, phenotype accuracy rises from 25.80 to 43.11 for ReXGroundingCT VQA and from 15.55 to 52.65 for RRG. The corresponding accuracies on LIDC-IDRI increase from 54.55 to 77.69 and from 50.41 to 75.21. The representation projections show a parallel change: categories that overlap at smaller data scales form increasingly compact and distinguishable clusters as more lesion examples are included. In particular, atelectasis, consolidation, ground-glass opacity, and pulmonary nodules become more clearly separated. These results indicate that the encoder learns discriminative lesion representations and that additional lesion data improve both phenotype characterization and the organization of the learned representation space.

4.6 Case Study

Figure 6 illustrates two representative cases in which the original backbone fails to identify the pulmonary abnormality. For both ground-glass opacity and atelectasis, the lesion patches are first located within the learned anatomical representation space, preserving their anatomical context. Their phenotype increments are then projected into the corresponding phenotype regions, providing discriminative evidence for pathological characterization. With these representations, our model correctly identifies both abnormalities and describes their bilateral lower-lung locations, whereas the original backbone reports no focal pulmonary abnormality. These examples show that ΔRepresentation improves report generation by jointly preserving anatomical information and encoding phenotype-specific visual changes.

5 Conclusion

We introduced **ΔRepresentation**, a visual representation learning framework for pathological phenotype interpretation in medical VLMs. **BaseAnatomy** structures normal anatomical representations with ideal spatial distributions, while **ΔPhenotype** learns phenotype-specific representation increments through counterfactual reasoning. Experiments across two datasets and multiple medical VLM backbones show consistent improvements in lesion grounding and phenotype characterization, demonstrating the effectiveness of anatomy-referenced incremental representation learning for fine-grained medical image interpretation.

AI USE STATEMENT

In this work, we used generative AI tools for evaluating the report generation task. Specifically, GPT6-Astra was used as a structured parser to extract predicted lesion locations and phenotype labels from generated radiology reports. The extracted predictions were then compared with the reference annotations using the deterministic evaluation metrics described in Appendix A. GPT6-Astra was not used to assign performance scores directly. We did not use generative AI tools for generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, proposing or refining hypotheses, designing research methodology or experiments, implementing methods, assisting with translation, cleaning or reformatting datasets, or supporting qualitative and thematic data analysis. Proof-related uses are not applicable to this work. Additionally, we used generative AI tools for drafting parts of the research paper, improving readability, and suggesting the structure of the manuscript. We reviewed and revised all AI-assisted content, manually checked factual statements and citations against the original sources, and take full responsibility for the final content of this work, including text, claims, evaluations, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not raise ethical concerns that require additional disclosure under the conference guidelines. It involves no new human-subject recruitment or collection of identifiable private information, and all datasets are used in accordance with their original access and usage requirements.

REPRODUCIBILITY STATEMENT

We provide an anonymous code repository at `https://anonymous.4open.science/r/deltarep-CF6D` to support reproduction of the methods and experiments presented in this work.

REFERENCES

Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.

Mohammed Baharoon, Luyang Luo, Michael Moritz, Abhinav Kumar, Sung Eun Kim, Xiaoman Zhang, Miao Zhu, Mahmoud H Alabbad, Maha S Alhazmi, Neel P Mistry, et al. Rexgroundingct: A 3d chest ct dataset for segmentation of findings from free-text reports. *NEJM AI*, 3(7): AIdbp2501220, 2026.

Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

Zhixuan Chen, Yequan Bie, Haibo Jin, and Hao Chen. Large language model with region-guided referring and grounding for ct report generation. *IEEE transactions on Medical Imaging*, 44(8): 3139–3150, 2025.

Pujin Cheng, Li Lin, Junyan Lyu, Yijin Huang, Wenhan Luo, and Xiaoying Tang. Prior: Prototype representation joint learning from medical images and reports. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21304–21314. IEEE, 2023.

Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. Medrax: Medical reasoning agent for chest x-ray. *arXiv preprint arXiv:2502.02673*, 2025.

Difei Gu, Yunhe Gao, Yang Zhou, Mu Zhou, and Dimitris Metaxas. Radalign: Advancing radiology report generation with vision-language concept alignment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 484–494. Springer, 2025.

Xiaoshuang Huang, Haifeng Huang, Lingdong Shen, Yehui Yang, Fangxin Shang, Junwei Liu, and Jia Liu. A refer-and-ground multimodal large language model for biomedicine. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 399–409. Springer, 2024.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

Sunggu Kyung, Jinyoung Seo, Hyunseok Lim, Dongyeong Kim, Hyungbin Park, Jimin Sung, Jihyun Kim, Wooyoung Jo, Yoojin Nam, and Namkug Kim. Medregion-ct: region-focused multimodal llm for comprehensive 3d ct report generation. *arXiv e-prints*, pp. arXiv–2506, 2025.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in neural information processing systems*, 36:28541–28564, 2023.

Qingqiu Li, Zihang Cui, Seongsu Bae, Jilan Xu, Runtian Yuan, Yuejie Zhang, Rui Feng, Quanli Shen, Xiaobo Zhang, Shang Gao, et al. Aor: Anatomical ontology-guided reasoning for medical large multimodal model in chest x-ray interpretation. *Advances in Neural Information Processing Systems*, 38:19331–19371, 2026.

Che Liu, Zhongwei Wan, Yuqi Wang, Hui Shen, Haozhe Wang, Kangyu Zheng, Mi Zhang, and Rossella Arcucci. Argus: benchmarking and enhancing vision-language models for 3d radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16448–16460, 2025.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, et al. Medgemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026.

Yuhao Wang, Yue Sun, Tao Tan, Chao Hao, Yawen Cui, Xinqi Su, Weichen Xie, Linlin Shen, and Zitong Yu. Trrg: towards truthful radiology report generation with cross-modal disease clue enhanced large language models. In *International conference on medical image computing and computer-assisted intervention*, pp. 647–657. Springer, 2025.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training in radiology. *arXiv preprint arXiv:2301.02228*, 2023.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866, 2025.

Qilong Xing, Zikai Song, Youjia Zhang, Na Feng, Junqing Yu, and Wei Yang. Mca-rg: Enhancing llms with medical concept alignment for radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 380–390. Springer, 2025.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: Generalist foundation model for unified multimodal medical understanding and reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, et al. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16(1):3108, 2025.

A Benchmark and Evaluation

A.1 Benchmark Construction

We construct a patient-disjoint benchmark from ReXGroundingCT and LIDC-IDRI. The unit of analysis is an unmarked axial chest CT image, rather than an entire CT volume. The curated benchmark contains 4,199 images: 3,233 from ReXGroundingCT and 966 from LIDC-IDRI. It is divided into 3,401 training, 394 development, and 404 locked test images, with no patient overlap between splits. The ReXGroundingCT training/development/test counts are 2,720/230/283; the corresponding LIDC-IDRI counts are 681/164/121. The ReXGroundingCT test set covers four target phenotypes: atelectasis (35 images), consolidation (43), ground-glass opacity (88), and pulmonary nodule (117). The LIDC-IDRI test set contains 71 nodule-positive images and 50 no-lesion controls.

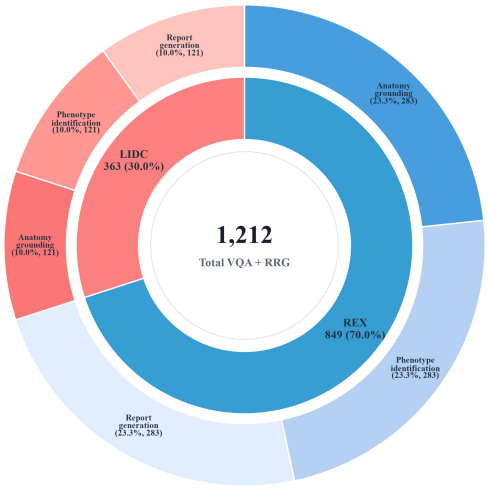

Figure 7: Distribution of benchmark instances by dataset and task.

We construct one lesion-grounding and one phenotype VQA question for each image. For grounding, the image is divided into a 4×4 grid, with rows A–D from top to bottom and columns 1–4 from left to right. For a positive image, the reference answer is the cell containing the largest area of the target lesion mask. Each question presents this cell and three distinct distractors, preferentially including nearby cells without the target lesion. For a no-lesion image, the four options include *No lesion* and three grid cells. The phenotype question specifies the target cell for positive images and asks the model to identify the abnormality. Its positive answer belongs to the four ReXGroundingCT phenotype categories listed above; LIDC-IDRI positive answers are pulmonary nodules, while negative questions include a no-lesion option. Lesion annotations are used to construct reference answers offline and are not provided at test-time inference.

Each image also contributes one radiology report generation (RRG) prompt requesting a concise finding that states the primary abnormality and, when discernible, its laterality and craniocaudal region. For no-lesion images, the appropriate finding indicates the absence of a focal lesion. Unlike the VQA grounding task, RRG does not use or expose the 4×4 grid.For no-lesion images, the appropriate finding indicates the absence of a focal lesion. Unlike the phenotype VQA question, the RRG prompt does not supply a target grid cell. The locked test set therefore comprises 808 VQA questions (566 ReXGroundingCT and 242 LIDC-IDRI) and 404 RRG instances (283 and 121, respectively), for 1,212 tasks in total (Fig. 7).

A.2 Evaluation Metrics

We score generated answers separately for each dataset, using all N_s test images from source s ($N_{\text{ReXGroundingCT}} = 283$ and $N_{\text{LIDC}} = 121$) as the denominator of every metric. Let g_i denote the reference grid cell or *No lesion*, and let d_i denote the reference phenotype or *no lesion*. For VQA, *Grounding Accuracy* and *Phenotype Accuracy* are exact-match accuracies of the parsed generated answers:

$$\begin{aligned} \text{Acc}^{(s)}_{\text{VQA-G}} &= \frac{1}{N_s} \sum_{i \in \mathcal{T}_s} \mathbf{1}[\hat{g}_i = g_i] \,, \\ \text{Acc}^{(s)}_{\text{VQA-P}} &= \frac{1}{N_s} \sum_{i \in \mathcal{T}_s} \mathbf{1}\left[\hat{d}^{\text{VQA}}_i = d_i\right] \,, \end{aligned} \tag{3}$$

where $\mathcal{T}_s$ is the test set for source s.

For RRG, we use GPT6-Astra as a structured parser to extract the phenotype and lesion-location attributes from each generated finding. Given a generated finding r_i, the parser produces $(\hat{d}_i^{\mathrm{RRG}}, \hat{\ell}_i) = P_{\mathrm{GPT}}(r_i)$, where $\hat{d}_i^{\mathrm{RRG}}$ denotes the extracted phenotype and $\hat{\ell}_i$ denotes the extracted lesion-location attributes, including laterality and craniocaudal region. The extracted predictions are then compared deterministically with the corresponding reference annotations. Missing, ambiguous, or incorrect extracted attributes are treated as incorrect. For no-lesion cases, an explicit statement indicating the absence of a focal lesion is parsed as *No lesion*. We compute Phenotype Accuracy and Grounding Accuracy as

$$\begin{aligned} \mathrm{Acc}^{(s)}_{\mathrm{RRG-P}} &= \frac{1}{N_s} \sum_{i \in T_s} \mathbf{1}\left[\hat{d}_i^{\mathrm{RRG}} = d_i\right], \\ \mathrm{Acc}^{(s)}_{\mathrm{RRG-G}} &= \frac{1}{N_s} \sum_{i \in T_s} \mathbf{1}\left[\hat{\ell}_i = \ell_i\right]. \end{aligned} \tag{4}$$

Here, d_i and ℓ_i denote the reference phenotype and lesion-location attributes, respectively. GPT6-Astra is used only to convert free-form findings into structured predictions; metric scores are computed deterministically from the extracted outputs.

B Disease-wise RRG&VQA Results

Table 4 decomposes the ReXGroundingCT results by phenotype for both VQA and RRG. The four disease subsets contain 35 atelectasis, 43 consolidation, 88 ground-glass opacity, and 117 pulmonary nodule cases. For each task, we report a composite accuracy obtained by pooling the Grounding and Phenotype decisions within the corresponding disease subset. Method rows report the absolute accuracy together with the gain over the corresponding native backbone.

Backbone-level consistency. The disease-wise breakdown shows that the aggregate improvement is not driven by one phenotype or one language backbone. Adding ΔRepresentation improves nearly every disease-metric pair across RadFM, LLaVA-Med, Lingshu, Hulu-Med, MedGemma, and MedGemma1.5, while the remaining cells are maintained. The size of the gain differs across backbones because their native report behavior is highly uneven: some models identify a disease but omit its location, whereas others default to a narrow set of common findings. The consistent direction of change indicates that the visual representation is useful beyond the decoder on which it was originally developed.

MedGemma1.5 comparison. With the MedGemma1.5 text backbone fixed, this block compares different alignment strategies. Under the same MedGemma1.5 backbone, Δ Representation improves disease-wise performance across both VQA and RRG. The gains are observed across atelectasis, consolidation, ground-glass opacity, and pulmonary nodules, rather than being concentrated in a single phenotype. Compared with RadAlign, MCA-RG, and Reg2RG, ΔRepresentation shows stronger VQA performance across the four phenotype subsets and remains consistently improved over the native backbone on RRG.

Residual difficulty by phenotype. The magnitude of improvement varies across phenotypes and backbones, indicating that representation learning does not affect all lesion appearances equally. Nevertheless, the improvements are distributed across all four phenotype subsets rather than being driven by a single disease category. This disease-wise analysis therefore complements the aggregate results in the main paper and shows that the overall gains persist across heterogeneous pathological appearances.

C High-Dimensional Representation Diagnostics

Table 5 provides quantitative diagnostics of the learned representation space in its original high-dimensional form, complementing the qualitative projections in the main paper. We evaluate three

Models	Scale	Atelectasis Phenotype		Consolidation Phenotype		Ground-glass opacity Phenotype		Pulmonary nodule Phenotype	
		VQA Acc.	RRG Acc.	VQA Acc.	RRG Acc.	VQA Acc.	RRG Acc.	VQA Acc.	RRG Acc.
RadFM	14B	7.14	4.29	8.14	4.65	6.82	3.41	8.55	4.27
w/ ΔRepresentation	∼14B	44.29 ↑37.14	17.14 ↑12.86	48.84 ↑40.70	18.60 ↑13.95	36.93 ↑30.11	13.07 ↑9.66	49.15 ↑40.60	14.10 ↑9.83
LLaVA-Med	7B	28.57	24.29	29.07	29.07	26.70	18.75	27.78	23.93
w/ ΔRepresentation	∼7B	55.71 ↑27.14	45.71 ↑21.43	72.09 ↑43.02	45.35 ↑16.28	54.55 ↑27.84	30.68 ↑11.93	59.40 ↑31.62	39.74 ↑15.81
Lingshu	7B	51.43	1.43	56.98	2.33	40.34	2.27	45.30	2.56
w/ ΔRepresentation	∼7B	64.29 ↑12.86	15.71 ↑14.29	69.77 ↑12.79	19.77 ↑17.44	56.82 ↑16.48	14.77 ↑12.50	65.81 ↑20.51	16.24 ↑13.68
Hulu-Med	4B	45.71	4.29	55.81	5.81	45.45	5.11	48.29	5.98
w/ ΔRepresentation	∼4B	57.14 ↑11.43	15.71 ↑11.43	72.09 ↑16.28	18.60 ↑12.79	55.68 ↑10.23	15.34 ↑10.23	67.09 ↑18.80	14.96 ↑8.97
MedGemma	4B	41.43	1.43	45.35	2.33	37.50	1.14	42.31	2.56
w/ ΔRepresentation	∼4B	55.71 ↑14.29	15.71 ↑14.29	62.79 ↑17.44	20.93 ↑18.60	47.16 ↑9.66	14.20 ↑13.07	63.68 ↑21.37	13.68 ↑11.11
MedGemma1.5	4B	30.00	2.86	37.21	3.49	25.57	2.84	29.91	2.99
w/ RadAlign	∼4B	31.43 ↑1.43	31.43 ↑28.57	38.37 ↑1.16	40.70 ↑37.21	26.14 ↑0.57	27.27 ↑24.43	30.77 ↑0.85	29.49 ↑26.50
w/ MCA-RG	∼4B	42.86 ↑12.86	34.29 ↑31.43	51.16 ↑13.95	48.84 ↑45.35	36.93 ↑11.36	28.98 ↑26.14	45.73 ↑15.81	43.59 ↑40.60
w/ Reg2RG	∼4B	41.43 ↑11.43	35.71 ↑32.86	52.33 ↑15.12	46.51 ↑43.02	36.36 ↑10.80	28.98 ↑26.14	45.73 ↑15.81	44.02 ↑41.03
w/ ΔRepresentation	∼4B	**67.14** ↑37.14	**38.57** ↑35.71	**61.63** ↑24.42	**52.33** ↑48.84	**55.11** ↑29.55	**31.25** ↑28.41	**64.53** ↑34.62	**44.87** ↑41.88

Table 4: **Disease-wise VQA and RRG performance on ReXGroundingCT.** Values are composite accuracies (%), obtained by pooling the Grounding and Phenotype decisions within each task. Gains are absolute percentage points over the native backbone.

properties that correspond to the central objectives of ΔRepresentation: the geometric stability of BaseAnatomy, the separability of phenotype increments, and the quality of the estimated counterfactual normal references. We further examine how these representations change under spatial pooling.

Anatomical representation stability. BaseAnatomy shows consistently high agreement under rotation, scaling, and translation, with Pearson correlations of 0.9742, 0.9751, and 0.9750, respectively. These results indicate that the learned anatomical representations preserve their relative organization under moderate spatial transformations rather than depending on a specific image configuration. This quantitatively supports the cross-transformation consistency observed in the anatomical representation analysis.

Phenotype organization and counterfactual normal references. The learned phenotype increments achieve a cosine silhouette score of 0.3607 and a nearest-prototype accuracy of 75.46% on the ReX four-phenotype evaluation set. These measurements show that phenotype-specific increments exhibit meaningful separation in the original representation space, consistent with the progressively clearer phenotype organization visualized in Fig. 5. For counterfactual normal estimation, the normal-ring representations remain highly compact, with a compactness score of 0.9882. The mean and median cosine distances between abnormal representations and their estimated normal references are 0.0233 and 0.0210, respectively. Together, these results indicate that BaseAnatomy provides a locally coherent anatomical space from which stable normal references can be estimated for pathological patches.

Sensitivity to spatial pooling. We additionally evaluate representation behavior before and after exact 4×4 spatial pooling. Fixed 5-NN accuracy increases from 60.98% to 64.23%, while macro accuracy increases from 58.98% to 60.83%. In contrast, the cosine silhouette and own-centroid cosine decrease modestly after pooling. This pattern suggests that pooling preserves, and in the local classification metrics slightly improves, discriminative neighborhood structure while introducing a small reduction in within-class geometric compactness. Therefore, the pooled representations are not strictly invariant to spatial aggregation, but retain the anatomical and phenotype-related organization required by the downstream visual-language interface.

Overall, the high-dimensional measurements support the qualitative observations in the main paper from three complementary perspectives. BaseAnatomy maintains stable anatomical geometry under spatial transformations, ΔPhenotype produces separable phenotype-specific increments, and the resulting representations remain informative after the spatial pooling used for downstream VLM alignment.

(a) Absolute representation quality

Component	Evaluation cohort	Metric	Score
BaseAnatomy	ReX supervised pool	Rotation Pearson $r \uparrow$	0.9742
		Scaling Pearson $r \uparrow$	0.9751
		Translation Pearson $r \uparrow$	0.9750
ΔPhenotypes	ReX four-phenotype set, $n = 379$	Cosine silhouette $\uparrow$	0.3607
		Nearest-prototype accuracy $\uparrow$	75.46%
Normal reference	LIDC validation, 232 cases / 1,120 lesion patches	Normal-ring compactness $\uparrow$	0.9882
		Abnormal-to-normal cosine distance (mean)	0.0233
		Abnormal-to-normal cosine distance (median)	0.0210

(b) Scale sensitivity

Metric	Evaluation cohort	$1\times$	$4\times$	Δ
Fixed 5-NN accuracy $\uparrow$	ReX validation, $n = 123$	60.98%	**64.23%**	+3.25 pp
Macro accuracy $\uparrow$	ReX validation, $n = 123$	58.98%	**60.83%**	+1.85 pp
Cosine silhouette $\uparrow$	ReX public test, $n = 55$	**0.5397**	0.5157	-0.0240
Own-centroid cosine $\uparrow$	ReX public test, $n = 55$	**0.9424**	0.9203	-0.0221

Table 5: High-dimensional representation diagnostics. Panel (a) reports absolute representation-quality measurements. Panel (b) isolates the effect of exact 4×4 spatial pooling. Here, $\Delta = 4\times - 1\times$ and pp denotes percentage points.

D Implementation Details

Backbone initialization and input. For each evaluated medical VLM, the anatomical encoder E_A is initialized from its native vision tower, and ΔRepresentation is trained separately for each backbone. After BaseAnatomy training, E_A is frozen and the phenotype encoder E_P is initialized from the same BaseAnatomy checkpoint, following Sec. 3.3. The numerical settings below describe the MedGemma 1.5–4B configuration used for the detailed representation analyses; architecture-dependent dimensions follow the corresponding native vision tower for the other VLMs. For MedGemma 1.5–4B, each axial CT slice is resized to 896×896, normalized to $[-1, 1]$, and divided into a 64×64 grid with patch size 14. The current slice index and total slice count are used to encode the relative craniocaudal position, and patch embeddings are ℓ_2-normalized.

BaseAnatomy training. BaseAnatomy follows the geometry-supervised learning objective described in Sec. 3.2. We set $\alpha = 0.65$ in Eq. (1) and sample anatomical classes in a balanced manner. Rotation and scale transformations are applied with matched spatial correspondence, using rotations in $[-10^\circ, 10^\circ]$ and isotropic scaling in $[0.92, 1.08]$. Anatomical annotations define normal tissue regions, while lesion masks exclude pathological patches from anatomical supervision. No additional representation-learning objectives beyond those described in Sec. 3.2 are introduced.

ΔPhenotype training. After BaseAnatomy converges, E_A remains frozen and E_P is optimized to encode phenotype-specific pathological changes as described in Sec. 3.3. For each pathological patch, its counterfactual normal reference is estimated from 8–48 anatomically matched, mask-negative neighbors within grid radii 4–10 using normalized inverse-distance weighting. Phenotype-specific consistency and separation, lesion localization, anatomical preservation, and no-lesion suppression follow the objectives described in Sec. 3.3. No additional phenotype-learning objectives are introduced.

Optimization. Both representation-learning stages use AdamW in BF16 with weight decay 0.01, cosine learning-rate decay, and gradient clipping at 1.0. Checkpoints are selected only on the patient-disjoint development split.

Vision–language alignment and inference. After representation learning, only E_P is retained as the VLM vision encoder. For downstream VQA and RRG, the trained vision encoder and text decoder are frozen and only the corresponding multimodal projector is optimized. For the MedGemma 1.5–4B configuration, each 4×4 block of the 64×64 visual grid is average-pooled, yielding 256 visual tokens. Alignment uses batch size two, gradient accumulation over two steps, a learning rate of 1.2×10^{-5}, and two epochs. At inference, the model receives only the unmarked CT slice, slice-position information, and text prompt. Anatomical masks, lesion masks, and counterfactual normal-reference estimation are used only during representation learning and are not provided at inference time.

Compute. Representation learning uses one or two NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs, while alignment and evaluation additionally use an NVIDIA RTX A6000 with 48 GB memory.