GEOMETRY-SUPERVISED VISUAL REPRESENTATION LEARNING FOR MULTI-PHENOTYPE LESION INTERPRETATION IN MEDICAL VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical vision-language models (VLMs) have shown increasing potential for clinical image interpretation. However, these models still struggle to interpret multi-phenotype lesions whose diagnosis requires the joint assessment of multiple pathological phenotypes. Existing vision–language alignment methods produce visual representations that fail to preserve anatomical hierarchies and relationships among phenotypic subclasses. This stems from their reliance on semantic supervision, which lacks geometric constraints to preserve these relationships in the visual embedding space. Moreover, the sparsity of lesion-related anatomical and phenotypic representations makes it difficult for medical VLMs to capture important diagnostic evidence. To address these limitations, we propose **PureVision**, a geometry-supervised visual representation learning framework for multi-phenotype lesion interpretation in medical VLMs. It combines a geometry-supervised representation learning module, **PureEyes**, and an anatomy-guided evidence aggregation module, **PureNeurons**. PureEyes provides geometric supervision through ideal spatial distributions that encode anatomical hierarchies and phenotypic subclass relationships. PureNeurons projects visual representations into the learned latent space, using their positions to selectively aggregate lesion-specific anatomical and phenotypic evidence. Experiments on *LIDC-IDRI*, *CBIS-DDSM*, and *3DReasonKnee* demonstrate that PureVision improves lesion grounding and phenotype characterization in visual question answering and radiology report generation. Code is available at: https://anonymous.4open.science/r/purevision-06C2.

1 INTRODUCTION

Medical vision-language models (VLMs) have been developed to assist clinicians in diagnosing diseases based on medical images Wu et al. (2025). Through large-scale pretraining on medical image–diagnosis pairs, they have learned to diagnose diseases characterized by a single pathological phenotype visible in medical images Li et al. (2023); Sellergren et al. (2025). However, existing medical VLMs exhibit inaccurate lesion grounding and confusion in pathological phenotype descriptions when dealing with complex diseases whose diagnosis requires simultaneous consideration of anatomical structures and multiple pathological phenotypes. We attribute these limitations to the lack of fine-grained visual representations that can effectively distinguish anatomical structures and diverse pathological phenotypes.

To better capture anatomical structures and pathological phenotypes, existing research has explored semantically guided visual representation learning Xing et al. (2025); Gu et al. (2025). For anatomical representations, BiRD Huang et al. (2024) uses segmentation-derived region annotations for referring and grounding instruction tuning, linking anatomical descriptions to localized image regions. VividMed Luo et al. (2025) uses phrase embeddings to prompt a localization module that predicts segmentation masks and bounding boxes, explicitly grounding anatomical concepts in images. For pathological phenotypes, MCA-RG Xing et al. (2025) aligns visual features with pathology concepts and introduces a pathology–anatomy matching loss to emphasize disease-related regions. RadAlign Gu et al. (2025) uses learnable concept tokens to extract attribute-specific visual features and aligns them with textual descriptions of diagnostic criteria through contrastive learning.

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.

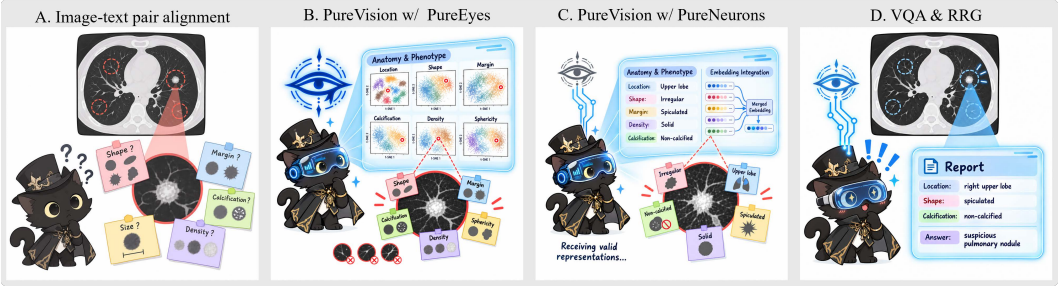

Figure 1: PureEyes structures the visual latent representation space, while PureNeurons captures and delivers lesion-relevant evidence.

However, the distribution of these text embeddings in the semantic representation space deviates from the ideal distribution of visual representations determined by anatomical structures and pathological phenotypes in the vision embedding space. We assume that under an ideal distribution, the geometric relation of representations should align with the spatial correlation of anatomical structures and pathological phenotypes: the close geometric relationship of anatomical structures is reflected as a short distance in the embedding space; the relationship between anatomical structures and their sub-structures should form a hierarchical representation; pathological phenotypes across different severity levels should form a smooth and continuous distribution in the embedding space, reflecting the gradual variation of their visual characteristics. Text embeddings used for supervision lack these geometric attributes, leading to insufficient anatomical and phenotypic information in visual representations. Moreover, pathological regions are typically sparse and account for only a small proportion of all image patches. The pathological information represented by these sparse patches can be interfered with by the large number of normal patches, making it difficult to accurately localize pathological regions and distinguish their phenotypic characteristics.

To bridge these gaps, we propose a representation learning framework for medical VLMs' visual encoders named **PureVision** as shown in Figure 1, which uses ideal spatial distributions to constrain the visual embedding space and dynamically translate these anatomical and phenotypic representations to semantic space. The ideal spatial distribution reflects the actual relationships among anatomical structures and phenotypic subclasses. Firstly, to address the distribution deviation, we develop **PureEyes**, which directly supervises visual encoders using ideal spatial distributions that preserve anatomical hierarchies and relationships among phenotypic subclasses. Furthermore, to effectively capture and deliver sparse lesion information, we develop **PureNeurons**, which dynamically selects lesion-relevant patches and delivers their anatomical and phenotypic evidence through semantic fusion. Experiments are conducted on *LIDC-IDRI*, *CBIS-DDSM*, and *3DReasonKnee*, covering CT, mammography, and MRI of different lesion types. PureVision improves lesion grounding and phenotype characterization across visual question answering (VQA) and radiology report generation (RRG). Our contributions are summarized as follows:

(1) We propose **PureVision**, a representation learning framework for medical VLM visual encoders that combines supervision based on ideal spatial distributions with selective information delivery for lesion grounding and pathological phenotype characterization.

(2) We develop **PureEyes**, a visual representation learning module that directly supervises separate anatomical and phenotypic encoders using ideal distributions to preserve anatomical hierarchies and continuity across ordered phenotypic subclasses.

(3) We introduce **PureNeurons**, an anatomy-guided information delivery module that selectively aggregates lesion-specific anatomical and phenotypic evidence to reduce interference from normal-tissue patches and delivers this evidence as soft semantic tokens to maintain the alignment with semantic embeddings.

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 for complex clinical querying 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). Together, these studies address domain adaptation, volumetric processing, and reasoning organization. The following subsection examines the more specific mechanisms used to associate visual evidence with image regions and medical concepts.

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 region 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). Together, 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 together with contrastive learning and feature gating, whereas RadAlign aligns visual concept tokens with textual diagnostic criteria to guide report generation Xing et al. (2025); Gu et al. (2025). Although these methods demonstrate the value of regional and concept-specific supervision, clinical text or concept embeddings do not necessarily impose an ideal structure on the visual latent space: anatomical hierarchies may remain poorly organized, and ordered phenotype subclasses may not exhibit the expected continuity. Moreover, lesion-specific evidence is often concentrated in only a few image patches and can therefore be diluted by normal-tissue representations. This combination of distribution mismatch and sparse lesion encoding limits the accurate interpretation of fine-grained pathological features after alignment.

3 Methodology

3.1 Overview

PureVision consists of two components as shown in Figure 2. **PureEyes** shapes anatomical and phenotypic visual representation spaces using annotation-defined ideal geometries, which maintain the relationship of anatomical structures and phenotypic subclasses. **PureNeurons** then identifies sparse lesion-relevant patches from the projection of the learned representation space and aligns their anatomical and phenotypic evidence into language-compatible soft tokens for downstream reasoning. The visual encoders are trained first and subsequently frozen during semantic alignment and inference.

3.2 **PureEyes**: Building Ideal-geometry Supervised Visual Representation Space

PureEyes learns complementary anatomical and phenotypic representations with visual encoders E_A and E_P, preserving spatial correspondence while independently optimizing them toward different ideal visual geometries. The fine-tuning process consists of three stages: (1) lesion-centric pheno-

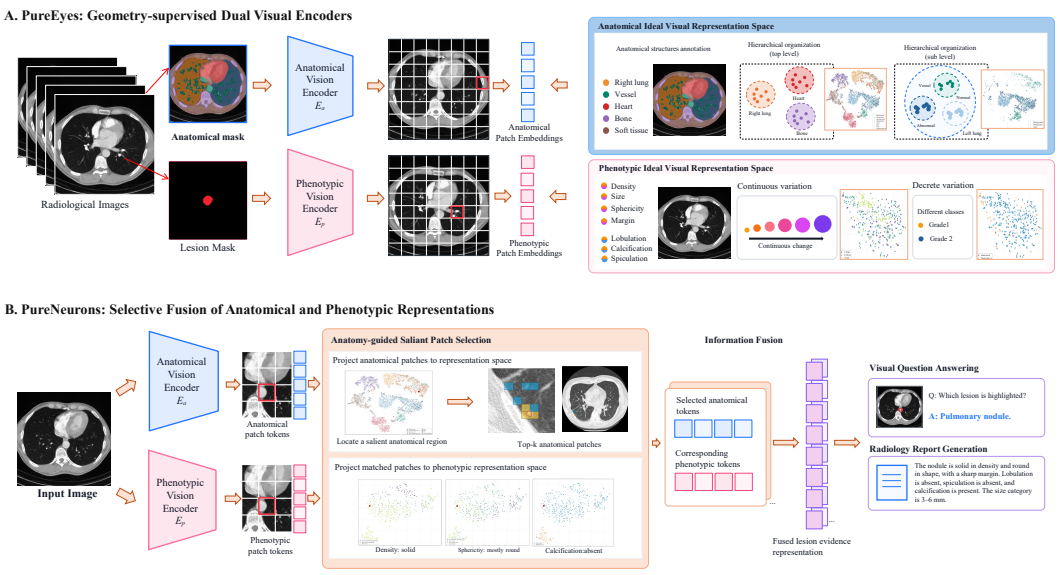

Figure 2: Overview of PureVision. (a) PureEyes shapes anatomical and phenotypic visual spaces with ideal geometries encoding anatomical hierarchies and phenotype relationships. (b) PureNeurons identifies lesion-relevant patches in these spaces and selectively fuses the corresponding anatomical and phenotypic evidence for downstream tasks.

type geometry learning, where isolated lesion patches are organized according to phenotype-specific geometric constraints; (2) context-aware phenotype geometry learning, where the full image is encoded and the lesion-region representations are further constrained under the same ideal geometry; and (3) anatomical geometry learning, where annotation-defined anatomical regions are organized according to their hierarchical relationships.

Phenotypic geometry supervision. Different phenotype types require different geometric structures. For continuous attributes, such as lesion size, normalized measurement differences define the target pairwise distances in the visual space:

$$\mathcal{L}^g_{\text{cont}} = \mathbb{E}_{(n,m)} \left[\rho_\delta \left(d^g_{nm} - \alpha_g \left| \tilde{y}^g_n - \tilde{y}^g_m \right| \right) \right], \tag{1}$$

where d^g_{nm} denotes the distance between normalized phenotype representations, $\tilde{y}^g_n$ is the normalized measurement, α_g controls the target distance scale, and ρ_δ is a robust penalty:

$$\rho_\delta(r) = \begin{cases} \dfrac{r^2}{2\delta}, & |r| < \delta, \\ |r| - \dfrac{\delta}{2}, & \text{otherwise.} \end{cases} \tag{2}$$

This objective matches visual-space distances to measurement-defined geometric distances, preserving continuous phenotypic variation.

For discrete phenotypes, representations of the same category are encouraged to form compact regions, while different categories are separated by a margin. For ordered phenotypes, an additional ranking constraint preserves the relative relationships among grades. Together, these objectives construct phenotype-specific geometries for continuous, ordinal, and categorical attributes; detailed formulations and implementation settings are provided in Appendix C.

We first apply these geometric constraints to representations extracted from isolated lesion patches to establish lesion-centric phenotype spaces. We then encode the complete image and apply the same constraints to lesion-overlapping patch representations, allowing phenotype geometry to incorporate surrounding anatomical context through self-attention. Lesion masks are used only to identify supervised patches during training.

Anatomical geometry supervision. The anatomical branch organizes tissues hierarchically, separating anatomical structures while preserving the relationship between normal and pathological

states within the same anatomical parent. Its objective is

$$\mathcal{L}_A = \lambda_a \mathcal{L}_{\text{anatomy}} + \lambda_p \mathcal{L}_{\text{parent}} + \lambda_h \mathcal{L}_{\text{hierarchy}} + \lambda_d \mathcal{L}_{\text{distill}} + \lambda_c \mathcal{L}_{\text{conf}}. \tag{3}$$

Here, anatomy and parent supervision preserve tissue identity, hierarchical supervision separates normal and lesion states within the same parent region, distillation prevents unnecessary drift of non-lesion features, and confounding-tissue contrast improves discrimination from visually similar normal structures. Anatomical annotations, pseudo-labels, and lesion masks are used only as training supervision.

3.3 PureNeurons: Grounding Sparse Lesion Position and Align Phenotypes to Semantic Representation

PureNeurons captures sparse lesion-specific information from the structured visual spaces learned by PureEyes and delivers it to the language model as dense semantic evidence. It first trains a shared alignment layer to map anatomical and phenotypic patch representations into a common semantic space while preserving their spatial correspondence. Based on the aligned representations, PureNeurons then identifies lesion-relevant regions from anatomical evidence and selectively aggregates the corresponding anatomical and phenotypic information into language-compatible soft semantic representations for the decoder.

Lesion Grounding in the Aligned Space. The alignment layer associates different regions of the anatomical representation space with their corresponding text labels. During inference, each anatomical patch is projected into the same aligned space, where its semantic identity can be inferred from the text label it matches best. To identify lesion-related patches, we compare their strongest match to lesion labels with their strongest match to other anatomical labels:

$$\rho_i = \max_{c \in \mathcal{C}_A^{\text{lesion}}} S_{i,c}^A - \max_{c \in \mathcal{C}_A^{\text{other}}} S_{i,c}^A, \tag{4}$$

where $S_{i,c}^A$ denotes the correspondence between patch i and text label c. A larger ρ_i indicates that the patch is more likely to belong to a lesion region than to normal anatomy.

Probability-weighted semantic fusion. Within each anatomical or phenotype group, patch-level similarities are spatially aggregated and normalized across candidate descriptions. Rather than making a hard subtype prediction, PureNeurons retains all candidates with different weights. Their language-model input embeddings are fused as

$$v_{g,\ell} = \sum_{c \in \mathcal{C}_g} p_{g,c}\, \bar{u}_{g,c,\ell}, \tag{5}$$

where $p_{g,c}$ is the semantic weight of subtype c and $\bar{u}_{g,c,\ell}$ is its length-aligned token embedding.

The resulting anatomical and phenotype soft tokens, together with the estimated lesion location, are inserted after the visual tokens and before the task instruction. The visual tokens are retained, allowing the frozen decoder to integrate explicit lesion-specific evidence with the complete image context for downstream tasks.

4 Experiments

4.1 Dataset and Evaluation

We conduct training and evaluation on three public datasets, including **LIDC-IDRI** Armato III et al. (2011), **CBIS-DDSM** Lee et al. (2017), and **3DReasonKnee** Sambara et al. (2025), which collectively cover diverse imaging modalities, anatomical structures, and disease types. LIDC-IDRI Armato III et al. (2011) is a CT dataset for pulmonary nodules, containing 12120 samples. It covers 6 anatomical structures and 7 nodule phenotypes. CBIS-DDSM Lee et al. (2017) is a mammography dataset for breast lesions, containing 2543 test samples. It covers 2 anatomical structures and 3 lesion phenotypes. 3DReasonKnee Sambara et al. (2025) is an annotated knee MRI dataset for musculoskeletal disease assessment based on the OAI dataset Lester (2012), containing 7846 test samples. We select the meniscus injury subset, which covers 5 anatomical structures and 1 lesion

Models	Scale	LIDC-IDRI		CBIS-DDSM		3DReasonKnee	
		Grounding Acc.	Phenotypes Acc.	Grounding Acc.	Phenotypes Acc.	Grounding Acc.	Phenotypes Acc.
RadFM	14B	2.00	5.00	2.00	17.00	5.00	3.00
w/ PureVision	∼14B	**11.00** +9.00	**23.00** +18.00	**23.00** +21.00	**24.00** +7.00	**21.00** +16.00	**24.00** +21.00
LLaVA-Med	7B	14.00	27.79	13.00	30.83	12.00	16.00
w/ PureVision	∼7B	**23.00** +9.00	**44.86** +17.07	**24.5** +11.50	**37.50** +6.67	**32.00** +20.00	**31.00** +15.00
Lingshu	7B	36.00	33.86	30.50	12.00	28.00	27.00
w/ PureVision	∼7B	**42.50** +6.50	**50.57** +16.71	**44.00** +13.50	**35.50** +23.50	**44.00** +16.00	**40.00** +13.00
Hulu-Med	4B	32.50	42.14	30.50	7.67	35.00	26.00
w/ PureVision	∼4B	**47.50** +15.00	**51.14** +9.00	**46.00** +15.50	**35.33** +27.66	**44.00** +9.00	**43.00** +17.00
MedGemma	4B	21.50	34.79	38.50	8.50	25.00	25.00
w/ PureVision	∼4B	**28.00** +6.50	**51.50** +16.71	**43.00** +4.50	**32.00** +23.50	**42.00** +17.00	**31.00** +6.00
MedGemma1.5	4B	11.50	10.50	24.00	10.00	29.00	25.00
w/ RadAlign	∼4B	23.00	32.93	37.50	31.17	36.00	34.00
w/ MCA-RG	∼4B	23.50	29.29	32.50	19.67	30.00	29.00
w/ Reg2RG	∼4B	11.50	23.79	32.00	17.17	33.00	23.00
w/ PureVision	∼4B	**45.50** +34.00	**49.93** +39.43	**41.50** +17.50	**48.67** +38.67	**43.00** +14.00	**50.00** +25.00

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

Models	Scale	LIDC-IDRI		CBIS-DDSM		3DReasonKnee	
		Grounding Acc.	Phenotypes Acc.	Grounding Acc.	Phenotypes Acc.	Grounding Acc.	Phenotypes Acc.
RadFM	14B	0.00	0.00	0.00	5.57	0.00	3.83
w/ PureVision	∼14B	**8.00** +8.00	**9.70** +9.70	**9.50** +9.50	**16.43** +10.86	**10.93** +10.93	**5.46** +1.63
LLaVA-Med	7B	5.00	13.00	0.00	11.41	4.92	8.20
w/ PureVision	∼7B	**28.00** +23.00	**33.10** +20.10	**12.00** +12.00	**27.33** +15.92	**13.66** +8.74	**9.59** +1.39
Lingshu	7B	12.50	39.50	7.00	7.28	23.50	21.80
w/ PureVision	∼7B	**40.00** +27.50	**47.74** +8.24	**20.00** +13.00	**12.81** +5.53	**41.53** +18.03	**34.43** +12.63
Hulu-Med	4B	25.50	37.19	12.00	17.04	24.59	29.02
w/ PureVision	∼4B	**57.50** +32.00	**50.18** +12.99	**24.50** +12.50	**22.18** +5.14	**50.27** +25.68	**45.05** +16.03
MedGemma	4B	15.50	16.49	10.50	16.51	32.24	25.00
w/ PureVision	∼4B	**45.00** +29.50	**45.79** +29.30	**14.50** +4.00	**27.24** +10.73	**48.09** +15.85	**34.03** +9.03
MedGemma1.5	4B	13.50	39.38	10.00	12.19	36.61	25.00
w/ RadAlign	∼4B	15.50	37.37	13.50	25.92	48.63	28.96
w/ MCA-RG	∼4B	11.50	33.08	24.00	18.40	45.36	27.32
w/ Reg2RG	∼4B	13.00	34.26	11.00	21.14	48.09	21.31
w/ PureVision	∼4B	**20.00** +6.50	**44.61** +5.23	**39.50** +29.50	**31.55** +19.36	**50.27** +13.66	**37.77** +12.77

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

phenotype. Based on these datasets, we construct visual question answering (VQA) and radiology report generation (RRG) evaluation sets to assess our method's ability in lesion localization and phenotype characterization, including 2,600 VQA pairs (1,600 for LIDC-IDRI, 800 for CBIS-DDSM, and 200 for 3DReasonKnee) and 583 RRG instances (200 for LIDC-IDRI, 200 for CBIS-DDSM, and 183 for 3DReasonKnee). Specifically, for lesion localization evaluation, we divide each image into a 4×4 grid, requiring the model to identify the grid cell containing the lesion. We use *Grounding Accuracy* and *Phenotype Characterization Accuracy* to evaluate the model's ability in lesion localization and pathological attribute characterization, respectively. Detailed information on benchmark distributions, anatomical structures' annotation, phenotype categories, VQA/RRG construction, and evaluation protocols is provided in Appendix A.

4.2 Experimental Setup

We conduct a comprehensive evaluation of PureVision from both downstream performance and module effectiveness. First, we compare PureVision with multiple 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 existing representation alignment methods Gu et al. (2025); Xing et al. (2025); Chen et al. (2025) on VQA and RRG tasks to evaluate lesion grounding and pathological phenotype characterization (Tables 1 and 2). We then analyze PureEyes on LIDC-IDRI by visualizing anatomical and phenotypic embeddings with t-SNE and measuring anatomy–text similarity, assessing whether the learned representations preserve the intended anatomical structure and phenotype relationships (Figs. 3 and 4). For PureNeurons, we evaluate lesion patch recall and redundant patch rate under different selection

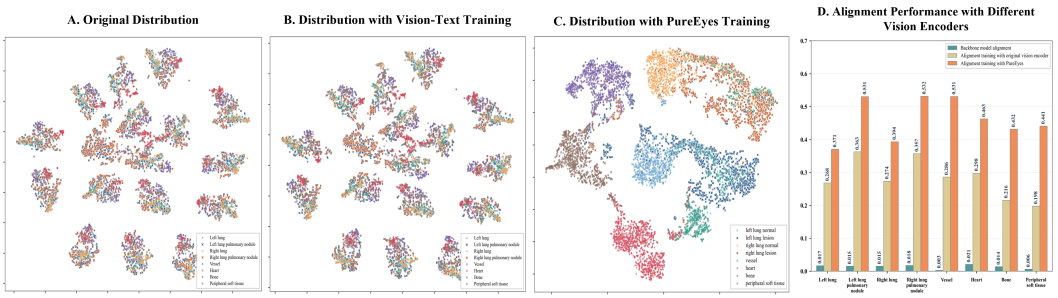

Figure 3: t-SNE visualization of anatomical representations and anatomy–text alignment across different vision encoders on LIDC-IDRI.

budgets, together with ablations of lesion grounding and soft semantic fusion. Finally, we provide qualitative patch-selection examples and an end-to-end case study to illustrate how the learned representations and selected lesion evidence support downstream predictions. Implementation details are provided in Appendix C

4.3 Comparison with Existing Methods and Alignment methods

Table 1 presents the comparison on the VQA task. To demonstrate the effectiveness and generalizability of PureVision across different medical VLMs, we apply PureVision to multiple pretrained medical VLM backbones. PureVision improves lesion grounding and pathological phenotype characterization across all evaluated backbones and datasets. For example, with MedGemma1.5, PureVision improves the grounding accuracy and phenotype characterization accuracy from 11.50%/10.50% to 45.50%/49.93% on LIDC-IDRI, from 24.00%/10.00% to 41.50%/48.67% on CBIS-DDSM, and from 29.00%/25.00% to 43.00%/50.00% on 3DReasonKnee, respectively. Across all evaluated backbones and datasets, PureVision achieves average improvements of 14.19 percentage points (pp) and 18.94 pp in grounding accuracy and phenotype characterization accuracy, respectively, demonstrating its effectiveness in enhancing visual representations of pretrained medical VLMs. Furthermore, to compare PureVision with existing visual-text alignment approaches, we apply representative alignment methods, including RadAlign, MCA-RG, and Reg2RG, under the same MedGemma1.5 backbone. PureVision achieves the highest grounding and phenotype characterization accuracies across all three datasets, demonstrating that the proposed supervision based on the ideal visual semantic distribution and the extraction of sparse lesion-specific information can effectively enhance the model's capability in lesion localization and description. Detailed phenotypes evaluation results are provided in Appendix B.1.

Table 2 further reports the comparison on the RRG task. To demonstrate the effectiveness of PureVision for open-ended radiology report generation, we apply PureVision to different pretrained medical VLM backbones. PureVision improves lesion grounding and pathological phenotype characterization across all evaluated backbones and datasets, with particularly pronounced gains in lesion localization. For example, with MedGemma1.5, PureVision improves the grounding and phenotype accuracies from 13.50/39.38 to 20.00/44.61 on LIDC-IDRI, from 10.00/12.19 to 39.50/31.55 on CBIS-DDSM, and from 36.61/25.00 to 50.27/37.77 on 3DReasonKnee. Similar improvements are observed across other pretrained backbones, demonstrating that the enhanced visual representations learned by PureVision can be effectively transferred to free-form radiology report generation. Furthermore, compared with existing representation alignment methods under the same MedGemma1.5 backbone, PureVision achieves the highest grounding and phenotype characterization accuracies across all three datasets.

4.4 PureEyes Supervised Vision Encoder Performance

Figure 3 compares anatomical patch representations on LIDC-IDRI. The original MedGemma1.5 encoder produces locally clustered but highly entangled representations, and vision–text alignment improves cross-modal similarity without substantially reorganizing this anatomical geometry. In contrast, PureEyes forms compact and well-separated anatomical clusters, with finer subregions for normal lung tissue, pulmonary lesions, and vessels. Normal and abnormal lung patches are also

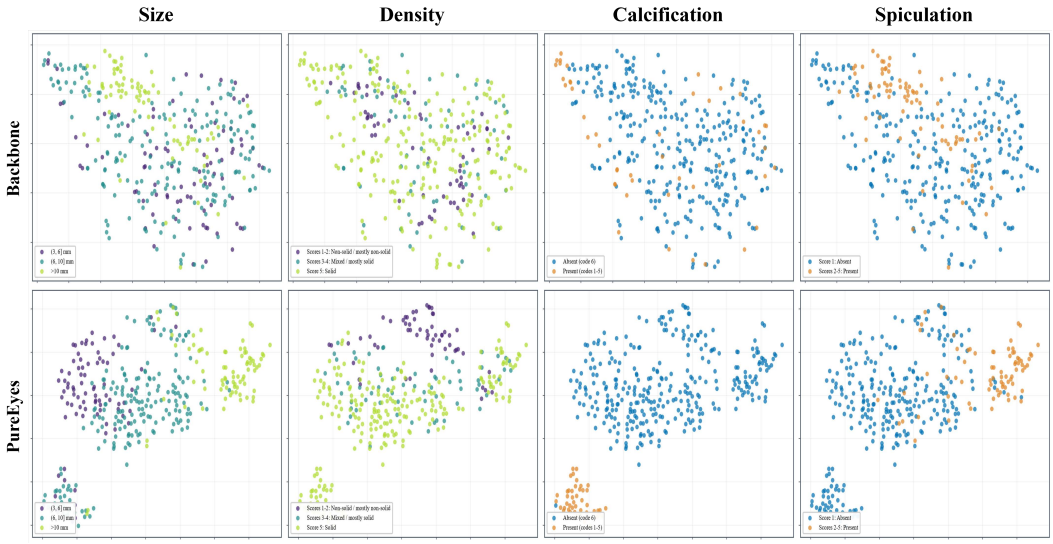

Figure 4: t-SNE visualization of pathological phenotype representations on LIDC-IDRI, comparing the backbone and PureEyes.

Lesion Extraction Efficiency			Lesion Information Delivery Ablation		
Selected Patches (K)	Lesion Patch Recall	Redundant Patch Rate	Method	Grounding Acc.	Phenotypes Acc.
4	60.11	45.23	MedGemma1.5	11.50	10.50
8	67.64	65.81	w/o PureNeurons	32.50	36.57
12	68.47	75.82	w/ lesion grounding	40.00	38.43
16	68.84	81.23	w/ lesion grounding and soft semantic fusion	45.50	49.93

Table 3: PureNeurons ablation study on lesion extraction efficiency and lesion information delivery.

clearly separated (Appendix B.2), while the partial overlap between nodules and vessels reflects their similar appearance in individual CT slices. The rightmost panel further shows improved visual–text alignment across anatomical categories, with left- and right-lung nodule similarities increasing from 0.363/0.357 to 0.531/0.532. These results show that PureEyes learns a more structured anatomical space that better supports semantic alignment.

Figure 4 compares phenotype representations produced by the original MedGemma1.5 vision encoder and the PureEyes-trained encoder. We visualize two continuously or ordinally varying phenotypes, size and density, and two discrete phenotypes, calcification and spiculation. With the original encoder, different size and density levels are scattered without clear ordering, while the discrete categories substantially overlap. In contrast, PureEyes arranges size and density representations along coherent progression trajectories, with adjacent levels remaining close, and organizes calcification and spiculation categories into more compact and distinguishable clusters. These results show that PureEyes captures both gradual phenotype transitions and discrete category boundaries. Visualizations of sphericity, margin, and lobulation are provided in Appendix B.3.

4.5 Lesion Extraction Efficiency and Information Delivery Ablation in PureNeurons

Table 3 evaluates the lesion extraction efficiency and information delivery design of PureNeurons. Increasing the number of selected patches improves lesion patch recall from 60.11% at $K = 4$ to 68.84% at $K = 16$, while the redundant patch rate increases from 45.23% to 81.23%. Notably, the recall improvement becomes limited beyond $K = 8$, indicating that a relatively small number of patches already captures most lesion-related evidence, whereas further increasing the selection budget mainly introduces redundant visual information. The information delivery ablation further shows consistent improvements from the original MedGemma1.5 to the different PureNeurons configurations. Grounding and phenotype accuracies increase from 11.50/10.50 to 32.50/36.57 without PureNeurons, further to 40.00/38.43 with lesion grounding, and finally to 45.50/49.93 with

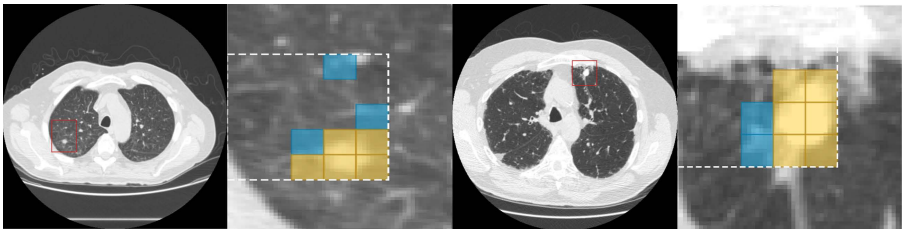

Figure 5: Visualization of lesion patch extraction by PureNeurons on LIDC-IDRI.

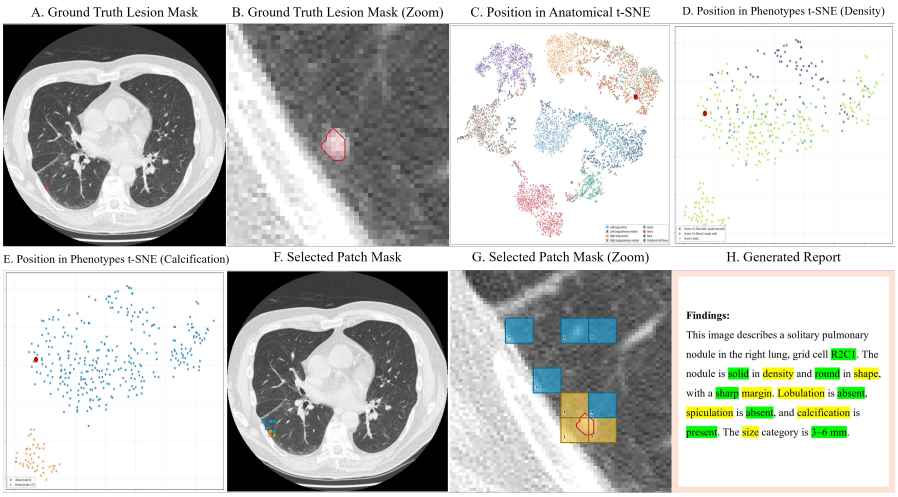

Figure 6: Case study of PureVision, illustrating the visual representations learned by PureEyes, lesion evidence selected by PureNeurons, and the resulting report generation.

lesion grounding and soft semantic fusion. These results demonstrate the importance of both lesion-focused information extraction and soft phenotype fusion for effectively delivering localized pathological evidence to the language model. Representative qualitative cases illustrating the lesion regions selected by PureNeurons are shown in Figure 5

4.6 Case Study

Figure 6 presents an end-to-end case study of PureVision on LIDC-IDRI, illustrating the interaction between PureEyes, PureNeurons, and the downstream report generation task. PureEyes first organizes anatomical and pathological representations into structured latent visual spaces, where the projected embeddings show that the input sample is located within the corresponding anatomical and phenotype distributions. Based on these enhanced visual representations, PureNeurons identifies a compact set of lesion related patches that overlap with the target nodule region, providing localized visual evidence for subsequent reasoning. The selected lesion evidence is then delivered to the decoder together with the original image information, enabling the model to generate a report containing the correct lesion location and multiple pathological phenotype descriptions. This case demonstrates that PureVision can establish structured visual representations, selectively extract lesion relevant evidence, and effectively transfer this information to downstream tasks.

5 Conclusion

We introduced **PureVision**, a geometry-supervised visual representation learning framework for multi-phenotype lesion interpretation in medical VLMs. **PureEyes** structures anatomical and phenotypic representations with ideal geometries, while **PureNeurons** captures and delivers sparse lesion-specific evidence. Experiments across three datasets and multiple imaging modalities show consistent improvements in lesion grounding and phenotype characterization, demonstrating the effectiveness of geometry-aware visual 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/purevision-06C2` 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.

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.

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. Region-aware multimodal large language model via slowfast tokenization and pseudo-mask guidance for 3d ct report generation. *arXiv preprint arXiv:2506.23102*, 2025.

Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4(1):170177, 2017.

Gayle Lester. The osteoarthritis initiative: a nih public–private partnership. *HSS journal*, 8(1): 62–63, 2012.

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.

Lingxiao Luo, Bingda Tang, Xuanzhong Chen, Rong Han, and Ting Chen. Vividmed: Vision language model with versatile visual grounding for medicine. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1800–1821, 2025.

Sraavya Sambara, Sung Eun Kim, Xiaoman Zhang, Luyang Luo, Shreya Johri, Mohammed Baharoon, Du Hyun Ro, and Pranav Rajpurkar. 3dreasonknee: Advancing grounded reasoning in medical vision language models. In *Biocomputing 2026: Proceedings of the Pacific Symposium*, pp. 99–113. World Scientific, 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.

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 VQA and RRG instances from patient-disjoint evaluation splits. For all three datasets, patients used for downstream evaluation are excluded from both representation learning and semantic alignment. Each evaluation image is associated with anatomical masks, a lesion mask, and phenotype labels, while the dataset-provided phenotype descriptions are directly used for RRG evaluation. Grounding questions use a 4×4 image grid, with ground-truth locations determined from the corresponding lesion masks. The benchmark distribution is shown in Fig. 7.

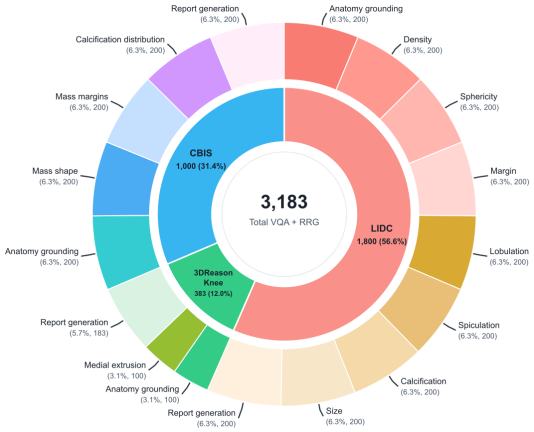

Figure 7: Distribution of benchmark instances by dataset and task.

For LIDC-IDRI, anatomical masks are generated using TotalSegmentator. We select single-nodule images whose nodule masks lie entirely within one grid cell and sample 200 grounding questions, each with one correct cell and three distractors. We additionally sample 200 questions for each of seven phenotypes: density, sphericity, margin, lobulation, spiculation, calcification, and size. Sampling is stratified by answer category within each dimension, yielding 1,600 VQA pairs.

For CBIS-DDSM, breast tissue masks are generated using GrabCut, pectoral muscle masks are generated by a pretrained Attention U-Net, and lesion masks are obtained from the original dataset annotations. We construct 200 grounding questions and 200 questions for each of three phenotypes: mass shape, mass margins, and calcification distribution. All VQA questions contain one correct answer and three distractors, yielding 800 VQA pairs.

For 3DReasonKnee, anatomical masks are obtained primarily from the dataset-provided nnU-Net segmentations. We construct 100 grounding questions and 100 questions assessing medial meniscus medial extrusion grades. All VQA questions contain one correct answer and three distractors, yielding 200 VQA pairs.

A.2 Evaluation Metrics

For VQA, predictions are matched against the reference answers. Grounding Accuracy is the proportion of correctly identified grid cells. Phenotypes Accuracy is computed by first calculating accuracy within each phenotype dimension and then averaging across dimensions:

$$\mathrm{Acc}_{\mathrm{VQA}} = \frac{1}{G} \sum_{g=1}^{G} \frac{N_g^{\mathrm{correct}}}{N_g}, \tag{6}$$

where G is 7 for LIDC-IDRI, 3 for CBIS-DDSM, and 1 for 3DReasonKnee. Missing or invalid answers are counted as incorrect.

For RRG, GPT6-Astra is used as a structured parser to extract the predicted grid location and phenotype labels from each generated report. The parser maps free-form report text to the predefined label space of each dataset. The extracted predictions are then evaluated deterministically against the reference annotations. We use a fixed evaluation prompt and deterministic decoding for all models and datasets. The parser is only allowed to return labels from the predefined candidate set, and missing or invalid outputs are counted as incorrect. Across all three datasets, Grounding Accuracy is the proportion of reports whose predicted grid cell exactly matches the reference cell in the 4×4 grid.

Models (VQA)	Density	Margin	Size	Sphericity	Calcification	Lobulation	Spiculation
RadFM	4.00	3.00	3.00	5.50	7.00	8.50	4.00
w/ PureVision	**11.00** +7.00	**5.00** +2.00	**5.00** +2.00	**33.00** +27.50	**40.00** +33.00	**50.50** +42.00	**16.50** +12.50
LLaVA-Med	17.50	7.50	5.00	39.00	46.50	56.50	22.50
w/ PureVision	**20.50** +3.00	**42.50** +35.00	**27.00** +22.00	**83.00** +44.00	**60.50** +14.00	**60.00** +3.50	**20.50** -2.00
Lingshu	23.50	13.50	10.00	45.50	52.50	63.00	29.00
w/ PureVision	**42.50** +19.00	**27.50** +14.00	**44.50** +34.50	**30.50** -15.00	**80.00** +27.50	**60.00** -3.00	**69.00** +40.00
Hulu-Med	32.00	20.50	34.50	25.00	73.50	50.00	59.50
w/ PureVision	**42.50** +10.50	**27.00** +6.50	**44.00** +9.50	**30.00** +5.00	**84.00** +10.50	**60.50** +10.50	**70.00** +10.50
MedGemma	32.00	19.00	33.50	25.00	22.50	61.50	50.00
w/ PureVision	**42.50** +10.50	**28.00** +9.00	**44.00** +10.50	**31.50** +6.50	**84.00** +61.50	**60.50** -1.00	**70.00** +20.00
MedGemma1.5	25.50	16.00	5.00	5.00	8.00	10.00	4.00
w/ RadAlign	22.50	12.50	9.00	44.50	52.00	62.00	28.00
w/ MCA-RG	27.50	17.50	31.00	25.00	23.00	43.00	38.00
w/ Reg2RG	33.00	9.50	0.00	25.00	12.50	50.00	36.50
w/ PureVision	**34.50** +9.00	**27.00** +11.00	**45.00** +40.00	**29.00** +24.00	**83.50** +75.50	**60.50** +50.50	**70.00** +66.00

Table 4: Detailed pulmonary nodule phenotype accuracy (%) on LIDC-IDRI.

Phenotypes Accuracy is computed using the same formulation across all three datasets: balanced accuracy is calculated within each phenotype dimension and then averaged equally across dimensions:

$$\mathrm{Acc}_{\mathrm{RRG}} = \frac{1}{G}\sum_{g=1}^{G}\frac{1}{|\mathcal{C}_g|}\sum_{c\in\mathcal{C}_g}\frac{N_{g,c}^{\mathrm{correct}}}{N_{g,c}}, \tag{7}$$

where G is 7 for LIDC-IDRI, 3 for CBIS-DDSM, and 1 for 3DReasonKnee. $\mathcal{C}_g$ denotes the reference classes represented among applicable samples in dimension g, $N_{g,c}$ is the number of samples with reference class c, and $N_{g,c}^{\mathrm{correct}}$ counts predictions that exactly match this class. This formulation assigns equal weight to classes within each dimension and to the phenotype dimensions.

B Additional Experimental Results

B.1 Detailed Phenotypes Accuracy

Tables 4 and 5 provide a detailed comparison of phenotype-specific VQA accuracy on LIDC-IDRI and CBIS-DDSM, respectively. PureVision improves most phenotype dimensions across the evaluated pretrained medical VLMs, indicating that the overall gains are broadly distributed across different pathological attributes rather than being dominated by a single phenotype. On LIDC-IDRI, when integrated with MedGemma1.5, PureVision increases the accuracies for density, margin, size, sphericity, calcification, lobulation, and spiculation from 25.50/16.00/5.00/5.00/8.00/10.00/4.00 to 34.50/27.00/45.00/29.00/83.50/60.50/70.00, with particularly pronounced gains in calcification, lobulation, and spiculation. Compared with existing representation alignment methods under the same MedGemma1.5 backbone, PureVision achieves the highest accuracy in five of the seven phenotype dimensions. On CBIS-DDSM, PureVision increases the mass shape, mass margins, and calcification distribution accuracies of MedGemma1.5 from 10.50/12.50/7.00 to 45.00/43.50/57.50 and achieves the highest accuracy across all three attributes among the compared alignment methods. Similar improvements across the remaining backbones further demonstrate the effectiveness of PureVision in fine-grained phenotype understanding across different diseases and imaging modalities. Since 3DReasonKnee contains only one phenotype dimension, namely medial meniscus medial extrusion, we do not provide an additional detailed phenotype-accuracy table for this dataset.

Models (VQA)	Mass Shape	Mass Margins	Calcification Distribution
RadFM	11.00	18.50	21.50
w/ PureVision	**33.50** +22.50	**12.50** -6.00	**26.00** +4.50
LLaVA-Med	33.50	27.50	31.50
w/ PureVision	**33.00** -0.50	**32.00** +4.50	**47.50** +16.00
Lingshu	19.00	4.00	13.00
w/ PureVision	**35.50** +16.50	**18.50** +14.50	**52.50** +39.50
Hulu-Med	13.00	3.50	6.50
w/ PureVision	**31.50** +18.50	**22.50** +19.00	**52.00** +45.50
MedGemma	10.00	12.00	3.50
w/ PureVision	**24.50** +14.50	**20.50** +8.50	**51.00** +47.50
MedGemma1.5	10.50	12.50	7.00
w/ RadAlign	23.50	20.50	49.50
w/ MCA-RG	14.00	16.50	28.50
w/ Reg2RG	12.50	24.50	14.50
w/ PureVision	**45.00** +34.50	**43.50** +31.00	**57.50** +50.50

Table 5: Detailed breast lesion phenotype accuracy (%) on CBIS-DDSM.

B.2 Fine-Grained Distribution of Latent Vision Representation of Pulmonary Patch Representations

To further examine the fine-grained organization of the visual space learned by PureEyes, we isolate several pulmonary patch categories from the LIDC-IDRI t-SNE projection. As shown in Figures 8 and 9, normal and abnormal lung patches form distinguishable subregions within the broader left- and right-lung clusters. In both lungs, most normal patches are concentrated in relatively compact regions, whereas lesion patches occupy different portions of the embedding space, demonstrating that PureEyes can capture pathological variations while preserving their anatomical context. By contrast, pulmonary lesion and vessel patches exhibit partial overlap, which is expected because nodules and cross-sectional vessels may have similar local appearances within a single CT slice. Nevertheless, a substantial proportion of vessel patches form vessel-dominated subclusters that are clearly separated from lesion patches, particularly in the peripheral regions of the embedding space. This observation suggests that the representations learned by PureEyes incorporate not only the local visual appearance of each patch but also contextual information from its surrounding anatomy. Together, these visualizations provide additional qualitative evidence that PureEyes organizes pulmonary patches according to both anatomical identity and fine-grained pathological characteristics.

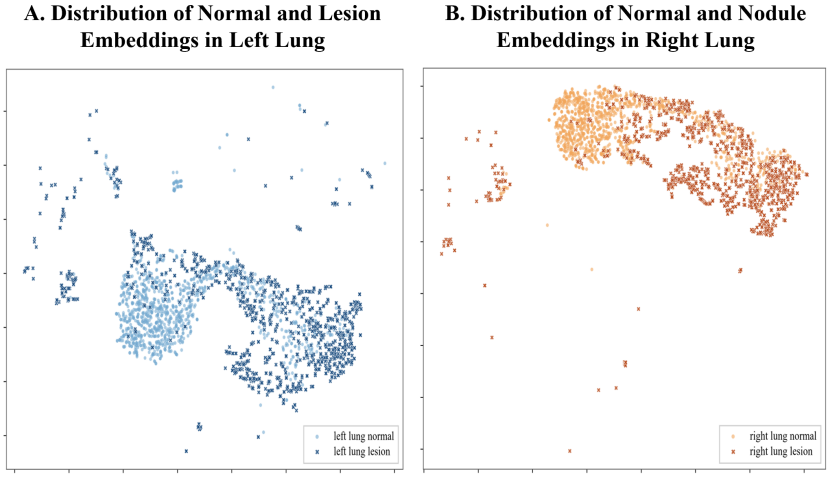

Figure 8: Fine-Grained Separation of Normal and Lesion Lung Representations

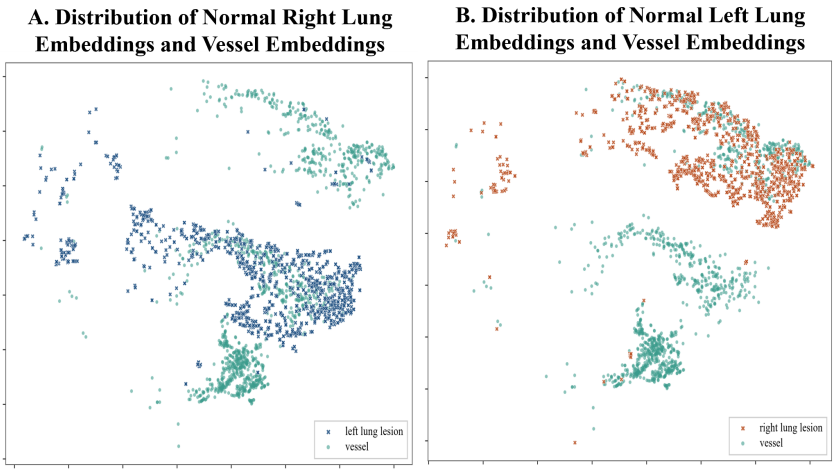

Figure 9: Fine-Grained Organization of Pulmonary Lesion and Vessel Representations

B.3 Additional Phenotypes Distribution

Figure 10 presents additional t-SNE visualizations for the remaining LIDC-IDRI phenotypes, including sphericity, margin, and lobulation. With the original MedGemma1.5 vision encoder, samples assigned different phenotype scores are substantially intermixed, and the latent distributions exhibit little correspondence with the underlying phenotype levels. In contrast, the representations learned with PureEyes show markedly clearer organization. Samples with similar phenotype scores tend to remain close in the embedding space, whereas samples exhibiting larger differences in sphericity, margin, or lobulation are distributed into more distinguishable regions. In particular, the learned distributions preserve gradual transitions between adjacent phenotype levels while increasing the separation between visually distinct manifestations. These results are consistent with the examples reported in the main text and further demonstrate that PureEyes learns a structured phenotype representation space across all evaluated pulmonary nodule characteristics, rather than improving only a small subset of attributes.

C Implementation Details

C.1 Architecture and Training Stages

For every medical VLM backbone, the anatomical encoder E_A and phenotypic encoder E_P are initialized independently from the backbone vision encoder and optimized as separate visual encoders. For the MedGemma-1.5 implementation, both encoders use the complete SigLIP vision tower with 896×896 inputs, patch size 14, a 64×64 patch grid, and 1,152-dimensional visual tokens. PureEyes updates the complete visual towers during representation learning. Afterward, both encoders are frozen, and a shared alignment layer consisting of RMS normalization followed by a bias-free $1{,}152 \rightarrow 2{,}560$ projection is trained to map anatomical and phenotypic patch representations into the language-compatible space. No LoRA or additional residual adapter is used. Training masks are used only to identify supervised patches and are never provided during inference.

C.2 Optimization

All stages use AdamW with weight decay 0.01, 5% linear warm-up followed by cosine decay, gradient clipping at 1.0, and BF16 mixed precision. Phenotypic representation learning uses a learning rate of 5×10^{-7} for 30 epochs with a global batch size of 16. Anatomical representation learning uses a learning rate of 2×10^{-6}, and the semantic alignment stage uses a learning rate of 2×10^{-5} for 30 epochs. During semantic alignment, the visual encoders and language model remain frozen, and only the shared RMS normalization and projection parameters are optimized.

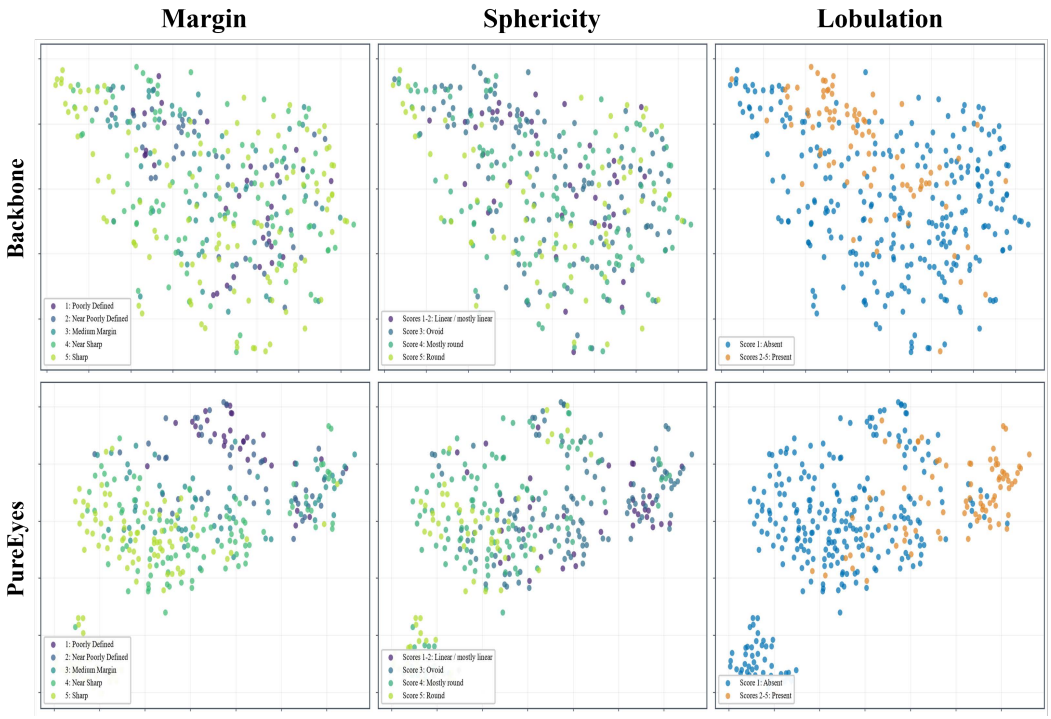

Figure 10: Additional t-SNE visualizations of phenotype representations on LIDC-IDRI.

C.3 Distance definition and robust penalty

All phenotype geometry objectives operate on ℓ_2-normalized representations,

$$\bar{\mathbf{e}}_i^g = \frac{\mathbf{e}_i^g}{\|\mathbf{e}_i^g\|_2}, \qquad d_{ij}^g = \left\|\bar{\mathbf{e}}_i^g - \bar{\mathbf{e}}_j^g\right\|_2, \tag{8}$$

where g denotes a phenotype dimension. We use the Smooth-L_1 penalty

$$\rho_\delta(r) = \begin{cases} \dfrac{r^2}{2\delta}, & |r| < \delta, \\ |r| - \dfrac{\delta}{2}, & \text{otherwise}, \end{cases} \tag{9}$$

with $\delta = 1.0$ throughout. For normalized embeddings, this Euclidean distance satisfies $d_{ij}^2 = 2 - 2\cos(\bar{\mathbf{e}}_i^g, \bar{\mathbf{e}}_j^g)$ and therefore preserves the angular geometry of the pretrained representation while providing a bounded distance space for explicit geometric supervision.

C.4 Continuous phenotype geometry

For phenotypes associated with continuous physical measurements, we supervise visual distances using the corresponding measurement differences. In LIDC-IDRI, lesion size is treated as a continuous phenotype and is represented by its physical diameter in millimeters. The measurement is normalized using statistics computed from the training split,

$$\tilde{y}_i^g = \frac{y_i^g - y_{\min}^g}{y_{\max}^g - y_{\min}^g}, \tag{10}$$

and the target distance between two samples is defined as

$$d_{ij}^{g,*} = \alpha_g \left|\tilde{y}_i^g - \tilde{y}_j^g\right|, \tag{11}$$

where $\alpha_g = 1.5$ matches the maximum target distance used by the discrete phenotype objectives. The continuous geometry loss is

$$\mathcal{L}^g_{\text{cont}} = \frac{1}{|\mathcal{P}_g|} \sum_{(i,j)\in\mathcal{P}_g} \rho_\delta \left(d^g_{ij} - d^{g,*}_{ij} \right). \tag{12}$$

This objective preserves metric information in the phenotype space: lesions with similar physical measurements are encouraged to remain nearby, while larger measurement differences correspond to proportionally larger visual distances.

C.5 Ordinal phenotype geometry

For ordinal phenotypes, including density, sphericity, margin, lobulation, and spiculation in LIDC-IDRI and medial meniscus extrusion grade in 3DReasonKnee, the target pairwise distance is determined by the separation between their ordered labels. For an attribute with C_g levels,

$$d^{g,*}_{ij} = 1.5 \frac{|y^g_i - y^g_j|}{C_g - 1}, \tag{13}$$

and the pairwise objective is

$$\mathcal{L}^g_{\text{pair}} = \frac{1}{|\mathcal{P}_g|} \sum_{(i,j)\in\mathcal{P}_g} \rho_\delta \left(d^g_{ij} - d^{g,*}_{ij} \right). \tag{14}$$

We further impose an ordinal ranking constraint. For anchor i, let $\bar{d}_{i,c}$ denote its mean distance to samples belonging to category c. For two categories c_{near} and c_{far} satisfying $\Delta_{\text{near}} < \Delta_{\text{far}}$, where $\Delta_c = |y^g_i - c|$, we define

$$\mathcal{L}^g_{\text{rank}} = \frac{1}{|\mathcal{R}_g|} \sum \left[\bar{d}_{i,c_{\text{near}}} - \bar{d}_{i,c_{\text{far}}} + m_o \left(\Delta_{\text{far}} - \Delta_{\text{near}} \right) \right]_+, \tag{15}$$

with $m_o = 0.1$. The final ordinal objective is

$$\mathcal{L}^g_{\text{ord}} = \mathcal{L}^g_{\text{pair}} + \mathcal{L}^g_{\text{rank}}. \tag{16}$$

C.6 Categorical phenotype geometry

For categorical phenotypes without an ordinal relationship, the target distance is zero for samples from the same category and 1.5 for samples from different categories:

$$d^{g,*}_{ij} = \begin{cases} 0, & y^g_i = y^g_j, \\ 1.5, & y^g_i \neq y^g_j. \end{cases} \tag{17}$$

The corresponding objective is

$$\mathcal{L}^g_{\text{cat}} = \frac{1}{|\mathcal{P}_g|} \sum_{(i,j)\in\mathcal{P}_g} w^g_{ij} \rho_\delta \left(d^g_{ij} - d^{g,*}_{ij} \right), \tag{18}$$

where w^g_{ij} compensates for category imbalance. Calcification is treated as categorical in LIDC-IDRI. For CBIS-DDSM, mass shape, mass margins, and calcification distribution are treated as categorical phenotypes.

For all phenotype objectives, each anchor is paired with samples from the current batch together with a memory bank containing the most recent 1,024 detached lesion representations. Self-pairs are excluded. The seven LIDC-IDRI phenotype dimensions use normalized weights of 0.125, 0.125, 0.125, 0.25, 0.1875, 0.09375, and 0.09375 for margin, calcification, density, sphericity, size, lobulation, and spiculation, respectively.

C.7 ANATOMICAL GEOMETRY OBJECTIVES

The anatomical encoder is optimized with

$$\mathcal{L}_A = 4\mathcal{L}_{\text{anatomy}} + 0.55\mathcal{L}_{\text{parent}} + 4\mathcal{L}_{\text{hierarchy}} + 10\mathcal{L}_{\text{distill}} + \mathcal{L}_{\text{conf}}. \tag{19}$$

$\mathcal{L}_{\text{anatomy}}$ organizes the anatomical categories according to a regular-simplex target geometry, while $\mathcal{L}_{\text{parent}}$ preserves the geometry of their higher-level anatomical parent regions. $\mathcal{L}_{\text{hierarchy}}$ further separates normal and lesion states within the same anatomical parent. To avoid unnecessary deformation of normal anatomical representations, $\mathcal{L}_{\text{distill}}$ minimizes the cosine distance between the learned encoder and a frozen pretrained teacher on non-lesion patches.

C.8 CONFOUNDING-TISSUE SEPARATION

Some normal anatomical structures exhibit local appearances similar to lesions and can therefore occupy neighboring regions in the pretrained representation space. We explicitly increase their separation using a teacher-relative confounding-tissue objective. Let s_{ij}^A and s_{ij}^T denote the cosine similarities between a target anatomical patch and a confounding-tissue patch under the learned anatomical encoder and frozen teacher, respectively. We define

$$\mathcal{L}_{\text{sep}} = \frac{1}{|\mathcal{Q}_{\text{conf}}|} \sum_{(i,j)\in\mathcal{Q}_{\text{conf}}} \left[s_{ij}^A - s_{ij}^T + m_c\right]_+, \tag{20}$$

where $m_c = 0.01$. Thus, the learned representation is required to reduce the similarity of a confounding pair by at least m_c relative to the pretrained teacher, rather than forcing unrelated anatomical representations to move arbitrarily far apart.

We additionally apply supervised contrastive regularization to improve the compactness of the confounding anatomical categories:

$$\mathcal{L}_{\text{conf}} = \mathcal{L}_{\text{sep}} + 0.05\,\mathcal{L}_{\text{SupCon}}, \tag{21}$$

with temperature 0.1. For LIDC-IDRI, pulmonary vessels are treated as the primary confounding tissue for left- and right-lung regions because vessel cross-sections can exhibit local appearances similar to pulmonary nodules.

C.9 SEMANTIC ALIGNMENT AND LESION SELECTION

After PureEyes training, both visual encoders are frozen. Anatomical and phenotypic patch representations share the same alignment layer. Text targets are encoded by the frozen language model, and the aligner is trained using temperature-scaled cosine classification with temperature 0.07 together with a cosine attraction term to the corresponding text representation. At inference time, the anatomical correspondence between patch i and label c is

$$S_{i,c}^A = \cos\left(G(\mathbf{h}_i^A),\, \mathbf{t}_c^A\right). \tag{22}$$

The lesion score is

$$\rho_i = \max_{c\in\mathcal{C}_{\text{lesion}}^A} S_{i,c}^A - \max_{c\in\mathcal{C}_{\text{other}}^A} S_{i,c}^A. \tag{23}$$

PureNeurons first identifies the most lesion-consistent local 5×5 neighborhood and then selects the top $K = 8$ patches within this region. Anatomical and phenotypic representations are matched using their shared spatial patch indices.

C.10 SOFT SEMANTIC FUSION

For phenotype group g, similarities over the selected patches are first aggregated as

$$s_{g,c} = \sum_{i\in\mathcal{R}} w_i S_{i,c}^g, \tag{24}$$

where w_i is obtained from a softmax over the selected lesion scores. The aggregated similarities are standardized within each phenotype group and converted into subtype probabilities:

$$p_{g,c} = \text{softmax}_c\left(\frac{(s_{g,c} - \mu_g)/\sigma_g}{\tau_s}\right), \qquad \tau_s = 0.125. \tag{25}$$

The soft semantic representation is then constructed as

$$\mathbf{v}_{g,\ell} = \sum_{c \in \mathcal{C}_g} p_{g,c}\, \bar{\mathbf{u}}_{g,c,\ell}, \tag{26}$$

where shorter candidate sequences are extended by repeating their final token embedding to obtain a common length within each phenotype group. The fused anatomical and phenotype tokens are inserted after the visual tokens and before the task instruction.

C.11 Computational Resources

All experiments were conducted on NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs. Phenotypic representation learning used two GPUs with distributed data parallel training, while anatomical representation learning and semantic alignment were each performed on a single GPU. Downstream inference and evaluation used two GPUs. All stages used BF16 mixed precision where applicable.