FROM IMAGE SPACE TO PATIENT SPACE: LEARNING PATIENT-CENTERED COORDINATES IN MEDICAL MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical MLLMs are increasingly moving beyond visual recognition toward more complex medical reasoning, yet they still struggle with spatial-relation reasoning in medical images. Our analyses reveal that medical MLLMs already encode image-space spatial information in the intermediate representations. However, this image-space representation alone is insufficient for clinical spatial reasoning. We therefore propose *Patient-Space Coordinate Learning (PS-Coord)*, a dual-frame training framework that jointly supervises image-space and patient-space coordinates within intermediate hidden states. Across eight backbones spanning 7B–34B parameters, PS-Coord consistently improves CT spatial reasoning, with average gains of 19.67% and 16.17% on MIRP and MSD, respectively. The improvements further generalize to unseen imaging modalities, held-out rotations, and medical visual question answering. Representation analyses show substantially increased accessibility of patient coordinates after training, while counterfactual controls indicate that PS-Coord-trained models rely on the anatomical context of the current image for medical spatial decisions. These results suggest that explicitly connecting image and patient coordinate provides an effective way for enabling spatial reasoning capabilities already latent within medical MLLMs.

1 INTRODUCTION

Medical multimodal large language models (MLLMs) are evolving beyond visual recognition toward more complex medical visual reasoning. Models are increasingly expected to interpret imaging evidence and answer clinically relevant questions (Li et al., 2023; Wu et al., 2025). As medical image understanding moves toward such fine-grained reasoning, anatomical localization becomes a fundamental requirement. The clinical interpretation of a lesion depends not only on its appearance, but also on where it occurs within the patient, including its laterality, organ of origin, and spatial relationship to surrounding structures. Errors in spatial description, such as documented laterality errors in radiology reports, can mischaracterize the underlying abnormality (Sangwaiya et al., 2009; Landau et al., 2015). Reliable medical image reasoning therefore requires medical MLLMs to connect visual evidence to the corresponding location in the patient's anatomical reference frame.

Recent studies explore anatomical grounding via region-aware instruction tuning and inference-time attention refinement (Wang et al., 2025; Liu et al., 2026), while M3D and CT-CHAT add anatomical context through three-dimensional medical vision–language learning (Bai et al., 2024; Hamamci et al., 2026). Yet evaluations show that spatial understanding remains fragile. MIRP finds that models may answer relative-position questions from anatomical priors rather than current-image evidence (Wolf et al., 2025). MedBLINK reports failures in basic spatial perception (Bigverdi et al., 2025), and SpatialMed finds weak reasoning about spatial relations in medical volumes (Trinh et al., 2026). These findings raise a fundamental question in medical visual reasoning: how do medical MLLMs represent and use spatial information, and how are these representations linked to the patient-centered anatomical frame required for clinical interpretation?

As illustrated in Figure 1, we find that medical MLLMs possess an internal image-coordinate representation that encodes spatial information: target locations in the displayed image can be reliably recovered through linear readouts of intermediate hidden states. We examine this structure

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.

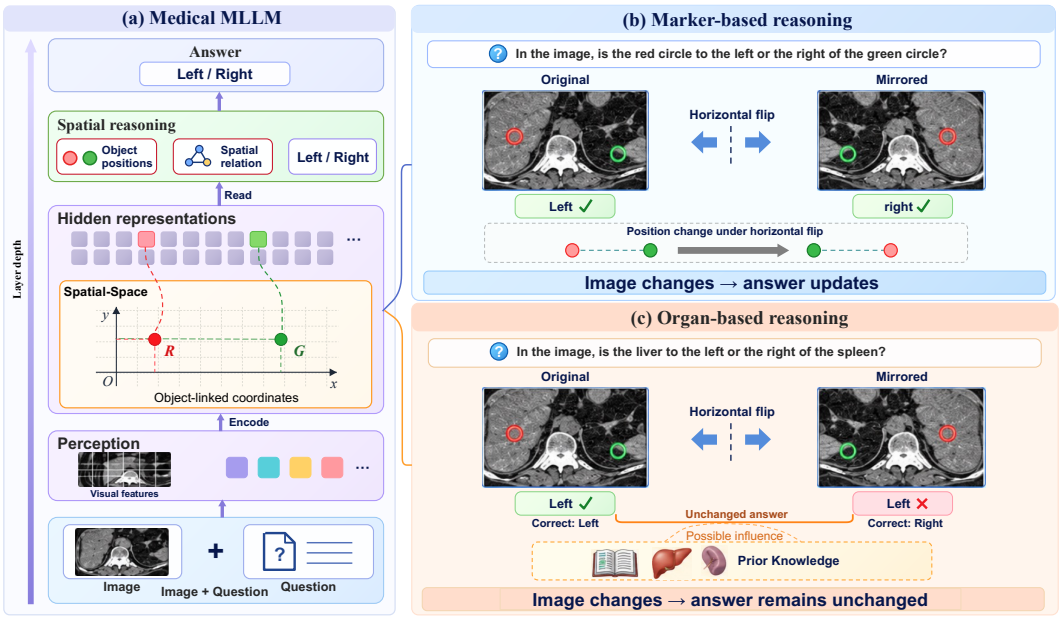

Figure 1: **Image-space representations and their limitation for medical spatial reasoning.** (a) Medical MLLMs encode spatial information in intermediate hidden representations, forming a linearly accessible image-coordinate subspace. (b) For marker-based queries, the answer updates following the image transformation. (c) For organ-based queries, the prediction remain unchanged after the image is transformed, indicating reliance on anatomical priors rather than spatial evidence.

across eight pretrained models ranging from 7B to 34B parameters, including Lingshu (LASA Team et al., 2025), MedVLThinker (Huang et al., 2025), and HuatuoGPT-Vision (Chen et al., 2024), with Qwen2.5-VL (Bai et al., 2025) as a general-purpose comparison. Using layer-wise linear probes with case-disjoint evaluation, we recover the horizontal and vertical coordinates of explicitly marked targets in CT images across rotated and mirrored views, without providing numerical coordinates in the prompts. Image-coordinate correlations peak between layers 19 and 22, revealing a linearly accessible image-coordinate subspace within the models' hidden states, where entity locations are represented internally without being explicitly written in the reasoning text. Our findings establish an important representational foundation for medical spatial reasoning: information about where a target appears in the image is available within the model.

Although these findings show that medical MLLMs contain accessible spatial information, this information does not consistently yield correct answers to medical spatial questions. Controlled comparisons in Figure 1 show that models identify image-relative positions of colored markers and update their answers after horizontal flips. For organs, however, answers may remain unchanged even when the relation reverses. This mismatch suggests that anatomical priors can guide responses without grounding them in current-image evidence. Medical localization also requires locating abnormalities within the patient, not just on the display. We therefore distinguish two reference frames: image coordinates locate targets in the displayed image, whereas patient coordinates locate them relative to the patient's anatomical axes. Rotations and mirroring alter image coordinates but preserve anatomical location. Medical spatial reasoning therefore combines visual grounding and frame interpretation, linking spatial information to patient-centered anatomy.

To address these questions, we propose *Patient-Space Coordinate Learning (PS-Coord)*, a training framework that develops a patient-centered spatial representation within medical MLLMs and connects it to their existing image-coordinate representation. The central idea is to jointly encode where a target appears in the image and where it lies relative to the patient, making both reference frames accessible within the model's hidden states. We construct spatial question-answering examples from CT images with marked targets under diverse rotations and reflections, and combine answer supervision with auxiliary prediction of image and patient coordinates from an intermediate hidden layer. The coordinate targets follow complementary rules: image coordinates track the transformed

display, while patient coordinates preserve the anatomical location of the same target. Learning both alongside spatial answers encourages the model to combine the target's visible position with anatomical context and distinguish changes in presentation of images from changes in patient-relative location. Through this joint training, PS-Coord aims to make the correspondence between the two reference frames available to allow the model's internal spatial representations to support anatomical localization and medical spatial reasoning.

We evaluate PS-Coord on eight backbones from four model families spanning from 7B to 34B parameters. PS-Coord improves performance on both CT spatial-reasoning benchmarks, with average gains of 19.67 and 16.17% on MIRP and MSD. The improvements further generalize to radiographic tasks from modalities never seen during training, as well as to held-out display rotations and medical visual question answering. They are accompanied by increased linear accessibility of patient coordinates in intermediate representations. Counterfactual controls further show that PS-Coord-trained models rely on the anatomical context of the current image, providing evidence that PS-Coord establishes a patient-coordinate for medical spatial reasoning. Our contributions are as follows:

- We identify a linearly accessible image-coordinate subspace in medical MLLMs that is causally associated with spatial predictions.
- We formulate medical spatial reasoning in terms of two reference frames: display-dependent image coordinates and anatomically invariant patient coordinates.
- We propose **PS-Coord**, a dual-frame training framework that jointly supervises image-space and patient-space coordinates from a shared intermediate representation.
- Across eight backbones and six benchmarks, PS-Coord improves medical spatial reasoning, generalizes to unseen modalities, and induces patient-centered spatial representations.

2 Related Work

Medical image understanding and spatial evaluation. Medical MLLMs have expanded from biomedical visual question answering toward anatomical interpretation and clinical reasoning. LLaVA-Med introduces biomedical visual instruction tuning (Li et al., 2023), while Lingshu and HuatuoGPT-Vision develop broader medical understanding through multimodal training (LASA Team et al., 2025; Chen et al., 2024). MedGemma 1.5 extends medical visual capabilities to high-dimensional imaging and localization (Sellergren et al., 2026). Alongside these advances, evaluations examine whether medical answers are grounded in spatial evidence from the current image. MIRP exposes reliance on anatomical priors through relative-position questions under image transformations, separating image-view from anatomical correctness (Wolf et al., 2025). MedBLINK identifies weaknesses in medical visual perception (Bigverdi et al., 2025). These findings motivate examining links between medical knowledge and visual spatial reasoning.

Grounding medical objects and findings. Anatomical grounding connects medical concepts to the structures observed in an image. MedRegA develops region-conditioned medical understanding through MedRegInstruct (Wang et al., 2025), and Lingshu incorporates region information into parts of its medical caption construction (LASA Team et al., 2025). VGRefine improves medical grounding at inference time by identifying visually sensitive attention heads and suppressing attention to irrelevant image regions (Liu et al., 2026). These approaches strengthen the association between medical language and visual evidence. Interpreting a grounded location also requires selecting the appropriate reference frame: an organ's position on the display and its position relative to the patient follow different rules under image transformations.

Three-dimensional medical models and benchmarks. Volumetric medical vision–language learning enriches anatomical context for interpreting organs and abnormalities. M3D combines three-dimensional medical image–language data with models for multiple tasks (Bai et al., 2024). CT-RATE and CT-CHAT support learning and conversational understanding from large chest CT datasets (Hamamci et al., 2026). Among benchmarks, 3D-RAD covers radiological questions on localization and multi-temporal analysis (Gai et al., 2025), while SpatialMed evaluates spatial relationships in medical volumes (Trinh et al., 2026). Together, these studies highlight anatomical context and spatial relationships as key components of medical image understanding.

3 FROM IMAGE SPACE TO PATIENT SPACE

To understand how medical MLLMs use spatial information to support medical reasoning, we investigate the spatial representations encoded in their hidden states. First, we identify image-coordinate subspaces within the models' hidden representations and characterize their layerwise organization, target binding, and influence on spatial decisions through linear readouts and controlled interventions. Second, we introduce the patient reference frame, which is essential for interpreting anatomical locations in medical diagnosis and reasoning, and formalize its relationship to image coordinates under changes in display orientation.

3.1 IMAGE-COORDINATE SUBSPACES

Medical MLLMs contain a linearly accessible representation of image-space location within their intermediate layers. Given an image and a query about a specified target, intermediate hidden states encode the target's horizontal and vertical coordinates, even though numerical coordinates are absent from the input and are never explicitly produced during reasoning. We refer to the span of these directions as an *image-coordinate subspace*, as shown in Figure 1a.

Let $h_{\ell,t}$ denote the hidden state at layer ℓ and at the token position associated with target t. For each spatial axis $k \in \{x, y\}$, a linear direction $w_{\ell,k}$ produces the coordinate estimate

$$\hat{c}_{\ell,k}(t) = w_{\ell,k}^{\top} h_{\ell,t}. \tag{1}$$

For a pair of targets A and B, the corresponding spatial decision is obtained by comparing their projected coordinates:

$$\hat{r}_{\ell,x}(A,B) = \mathbb{1}[\hat{c}_{\ell,x}(A) < \hat{c}_{\ell,x}(B)]\,, \qquad \hat{r}_{\ell,y}(A,B) = \mathbb{1}[\hat{c}_{\ell,y}(A) < \hat{c}_{\ell,y}(B)]\,. \tag{2}$$

Layerwise organization of spatial information. We characterize how spatial information is distributed across model depth. At each layer, we measure the Pearson correlation between linearly decoded coordinates and ground-truth target positions. Figure 2 presents the resulting profiles across four 7B models, comparing image-coordinate readouts with patient-coordinate readouts and shuffled-label controls. These profiles identify in which image-space information is most readily accessible.

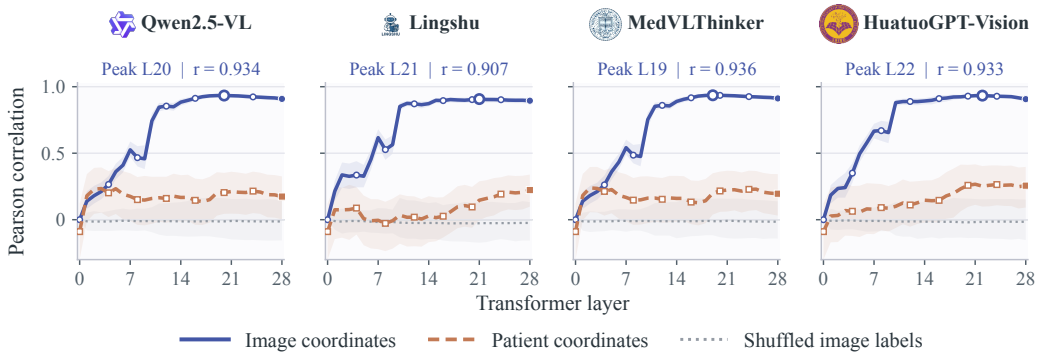

Figure 2: **Layerwise image-coordinate correlations across four 7B multimodal models.** Each panel reports Pearson correlations for image coordinates, patient coordinates, and shuffled labels across Transformer layers. The highlighted peaks correspond to the image-coordinate maxima.

Target-specific spatial binding. We examine whether recovered spatial information binds to the queried target by applying the same coordinate direction to a target's token state and comparing correlations with its own and the other target's positions. The results are shown in Figure 3a.

Accurate and prompt-robust spatial subspace. We evaluate whether the recovered directions support correct spatial judgments on held-out cases and remain effective when prompts change. For Qwen2.5-VL-7B and Lingshu-7B, we decode spatial relations from the target representations at

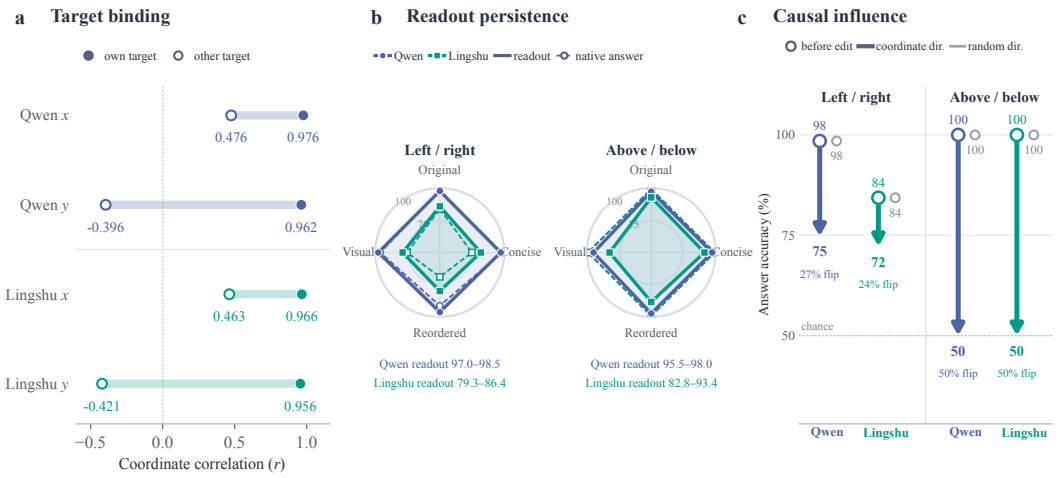

Figure 3: **Target binding, prompt persistence, and causal effects of the image-coordinate subspace.** (a) Spatial information is associated with the queried target rather than the other target. (b) Fixed readout directions accurately recover spatial relations across prompts and held-out cases. (c) Hidden-state edits alter the model's spatial answers.

Layer 19. As shown in Figure 3b, spatial information remains highly recoverable across different contextual settings and on held-out cases, indicating that the image-coordinate subspace is consistently preserved rather than being specific to a particular prompt or example.

Causal influence on spatial decisions. We intervene on target-token hidden states along the recovered spatial directions and measure the effect on the model's native answers across perturbation magnitudes. As shown in Figure 3c, Horizontal and vertical interventions produce substantially larger effects than interventions along random directions. These results show that the recovered directions are not merely readable: manipulating them can causally influence the model's spatial decisions.

3.2 The Patient Reference Frame

Medical reasoning requires interpreting that location relative to the patient's anatomy: a lesion in the *right* lung or the *anterior* mediastinum. The same anatomical position can appear at different screen positions across acquisitions, viewers, and reformats, while its clinical description does not change. We therefore distinguish image coordinates from patient coordinates, which express location relative to the patient's anatomical axes. Recent studies show that medical MLLMs rely on anatomical priors for relative-position questions (Wolf et al., 2025), misjudge image orientation (Bigverdi et al., 2025), and struggle with spatial relations in volumes (Trinh et al., 2026). We therefore relate patient coordinates to the image coordinates recovered in Section 3.1.

Two reference frames for one target. A medical target has both a position in the displayed image and a position in the patient's anatomy. We formalize this distinction on an axial CT slice. Let W and H be the image width and height, $c = (c_x, c_y)$ the body centre, and $p = (p_x, p_y)$ the target position. In the canonical radiological display, the normalized patient coordinate is

$$b = \begin{bmatrix} (c_x - p_x)/(W/2) \\ (c_y - p_y)/(H/2) \end{bmatrix}, \tag{3}$$

where positive b_x denotes patient-right and positive b_y denotes anterior. This coordinate describes the target relative to the patient's anatomical axes, as shown in Figure 4.

A display transformation $g = (\theta, m)$ applies a horizontal reflection M_m and a rotation R_θ:

$$M_m = \begin{bmatrix} (-1)^m & 0 \\ 0 & 1 \end{bmatrix}, \qquad R_\theta = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}. \tag{4}$$

The corresponding normalized image coordinate is

$$i_g = S^{-1} R_\theta M_m (p - c), \qquad S = \max(W, H)/2. \tag{5}$$

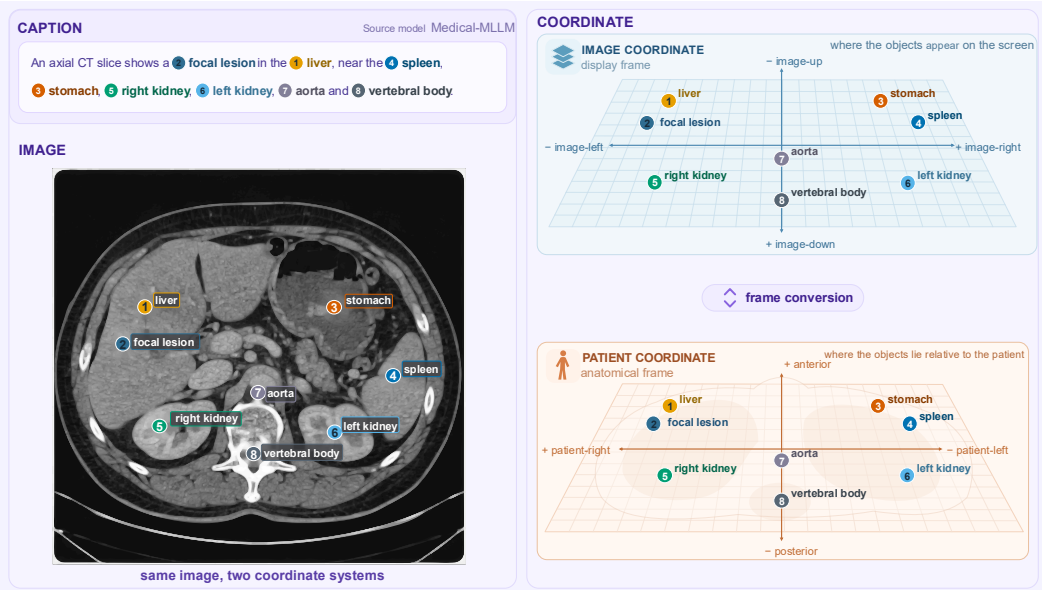

Figure 4: **The same anatomical targets in image and patient coordinates.** Colored labels link entities in an illustrative description to their locations in an axial CT slice and the two coordinate frames. Image and patient coordinates follow the display and anatomical axes, respectively.

Therefore, the patient coordinate b is invariant to how the image is displayed, whereas the image coordinate i_g changes equivariantly with reflections and rotations.

Image space alone is insufficient for medical reasoning. The image-coordinate subspace identified in Section 3.1 makes i_g linearly accessible from intermediate hidden states. However, it cannot determine the corresponding patient-space coordinate because transformation g is not explicitly given in the prompt. Patient-centered reasoning must combine three operations: grounding the named structure in the current image, inferring the display orientation from anatomical cues, and converting the observed image position into the patient frame. As illustrated in Figure 6, a correct patient-side answer alone does not establish that the model has grounded its response in the current image.

4 Patient-Space Coordinate Learning (PS-Coord)

The preceding analysis identifies image-space information that is accessible within medical MLLMs but insufficiently connected to patient-centered spatial decisions. We then propose PS-Coord, a new training framework that supplies the missing frame by supervising both coordinates of the same target from a single intermediate representation. As illustrated in Figure 5, two auxiliary heads predict a target's image and patient coordinates from a shared intermediate representation. Their complementary transformation rules encourage the model to distinguish displayed position from anatomical location, while answer supervision connects this distinction to spatial reasoning.

For a medical image containing a specified target, a display transformation g yields a training example (I_g, q, y, n, b, i_g), where q is a spatial question, y is its answer, n denotes the medical-semantic description of the target, and b specifies the target's anatomical location, while i_g specifies its position in the image coordinates. Coordinates and transformation parameters are used only as supervision and are not supplied in the prompt.

Let $h_g^{(\ell)} \in \mathbb{R}^d$ denote the hidden state at the last prompt token of the MLLM f_ϕ for input (I_g, q), where this position has access to the complete multimodal context. Two affine heads predict

$$\widehat{b}_g = W_b h_g^{(\ell)} + a_b, \qquad \widehat{i}_g = W_i h_g^{(\ell)} + a_i, \tag{6}$$

where $W_b, W_i \in \mathbb{R}^{2\times d}$. Supervising both heads at the same token encourages a representation that combines target location with anatomical context.

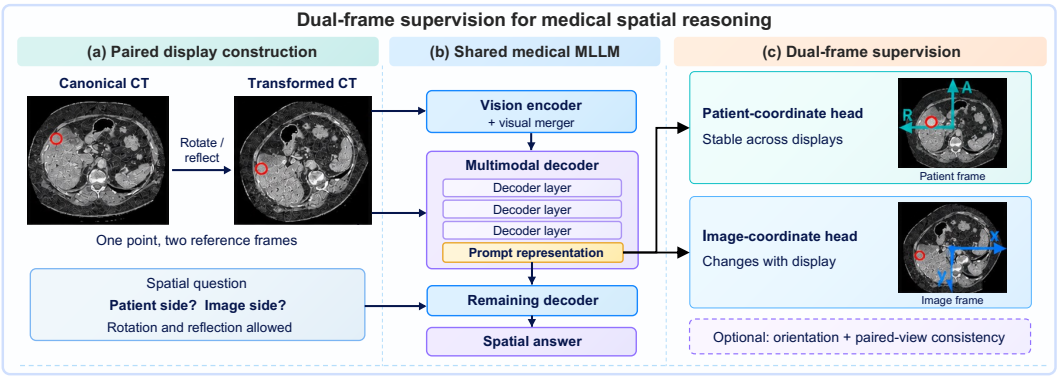

Figure 5: **Dual-frame supervision for medical spatial reasoning.** During training, PS-Coord jointly supervises image- and patient-space coordinates from intermediate representations of canonical and transformed views. Patient coordinates remain invariant across views, whereas image coordinates vary with the display transformation.

Define the mean squared error as $\mathcal{E}(u, v) = \|u - v\|_2^2 / \dim(u)$. The training objective is

$$\mathcal{L}_{\text{PS-Coord}} = \mathcal{L}_{\text{ans}} + \lambda_c \left[\mathcal{E}(\widehat{b}_g, b) + \mathcal{E}(\widehat{i}_g, i_g) \right], \tag{7}$$

$\mathcal{L}_{\text{ans}}$ averages autoregressive cross-entropy over answer tokens, including the end-of-sequence token.

The coordinate targets impose complementary constraints. Across transformed displays, predicting b encourages the model to interpret the target relative to anatomical axes, while predicting i_g preserves its current visual position. The model must therefore learn to transform spatial information between the image-coordinate and patient-coordinate frames.

5 Experiments

5.1 Experimental Setup

Backbone models. We evaluate PS-Coord on Qwen2.5-VL at 7B and 32B (Bai et al., 2025), Lingshu at 7B and 32B (LASA Team et al., 2025), MedVLThinker at 7B and 32B (Huang et al., 2025), and HuatuoGPT-Vision at 7B and 34B (Chen et al., 2024). Qwen2.5-VL provides a general-purpose reference, while the remaining families incorporate medical multimodal training. For each backbone, we compare the original checkpoint with its corresponding PS-Coord checkpoint.

Training data. We construct spatial question-answering examples from axial CT slices derived from TotalSegmentator(Wasserthal et al., 2023). The spatial training set contains 32,000 examples drawn from 1,037 canonical slices across 683 cases. Training displays combine horizontal reflections with rotations sampled on a 15° grid. Questions cover patient laterality, anterior–posterior location, image laterality, displayed back direction, and reflection status. Image and patient coordinates follow the conventions in Section 3.2 and are supplied as auxiliary supervision rather than prompt content. Detailed implementations are provided in the Appendix.

Benchmarks. Six benchmarks in three groups (Table 1): *Spatial reasoning* on CT—MIRP (Wolf et al., 2025), relative-position questions about organs, and MSD, patient-side laterality of a marked lesion (Antonelli et al., 2022); *Cross-modal transfer* to radiographs modal never seen in training—MedBLINK pelvic orientation (Bigverdi et al., 2025) and RSNA laterality of a marked finding (Shih et al., 2019); and *Medical VQA*—VQA-RAD (Lau et al., 2018) and SLAKE (Liu et al., 2021), exact match. Multiple-choice questions are scored by option probability under both option orders.

5.2 Main Results

Table 1 compares each backbone before and after PS-Coord training. Across both CT spatial-reasoning benchmarks, PS-Coord consistently improves performance for all four model families.

		Spatial reasoning		Cross-modal transfer		Medical VQA	
Model	**Variant**	**MIRP**	**MSD**	**MedBLINK**	**RSNA**	**VQA-RAD**	**SLAKE**
Qwen2.5-VL-7B	Base	55.89	46.00	81.00	43.17	**56.76**	52.50
	PS-Coord	**84.78**	**64.33**	**83.25**	**49.67**	55.43	**78.04**
Lingshu-7B	Base	59.00	46.67	65.75	55.67	55.21	74.84
	PS-Coord	**76.67**	**68.00**	**84.00**	**60.67**	**58.98**	**82.66**
MedVLThinker-7B	Base	60.11	48.67	71.75	**45.50**	**57.43**	51.46
	PS-Coord	**81.78**	**63.33**	**78.25**	45.00	56.98	**78.13**
HuatuoGPT-Vision-7B	Base	59.78	47.00	57.00	50.00	**57.43**	52.12
	PS-Coord	**88.11**	**67.67**	**82.25**	**60.33**	56.10	**77.57**
Qwen2.5-VL-32B	Base	72.56	48.00	68.50	50.67	66.74	64.94
	PS-Coord	**82.56**	**61.67**	**85.25**	**51.67**	**78.05**	**80.96**
Lingshu-32B	Base	58.78	51.00	63.50	23.00	66.30	84.17
	PS-Coord	**82.56**	**63.67**	**78.75**	**78.67**	**71.18**	**84.83**
MedVLThinker-32B	Base	73.33	48.00	53.75	**51.33**	66.74	66.64
	PS-Coord	**83.22**	**63.00**	**62.75**	46.67	**76.94**	**80.68**
HuatuoGPT-Vision-34B	Base	47.33	53.33	50.00	49.67	33.04	28.28
	PS-Coord	**64.44**	**66.33**	**50.75**	**56.17**	**49.00**	**68.43**
Avg. gain (pp)		**+19.67**	**+16.17**	**+11.75**	**+9.98**	**+5.38**	**+19.55**

Table 1: **Results on eight backbones and six benchmarks.** Scores are percentages (higher is better). Bold marks the better score in each Base/PS-Coord pair, while teal and coral indicate increases and decreases, respectively. *Avg. gain* is the unweighted mean difference between PS-Coord and Base across eight pairs.

(a) Patient-coordinate correlation

	Peak patient-coordinate r		
Backbone (7B)	Base	PS	Δr
Qwen2.5-VL	0.234	**0.897**	+0.662
Lingshu	0.223	**0.888**	+0.665
MedVLThinker	0.239	**0.884**	+0.644
HuatuoGPT	0.267	**0.900**	+0.633
Avg.	0.241	**0.892**	**+0.651**

(b) Held-out rotations ($90°$, $270°$)

	Patient L/R		Patient A/P	
Backbone (7B)	Base	PS	Base	PS
Qwen2.5-VL	51.3	**97.5**	46.4	**94.5**
Lingshu	47.5	**96.3**	46.9	**93.9**
MedVLThinker	48.9	**98.1**	48.6	**92.5**
HuatuoGPT	53.6	**97.5**	49.1	**93.1**
Avg.	50.3	**97.3**	47.7	**93.5**

Table 2: **Patient-coordinate correlation and held-out rotation generalization. (a)** Peak Pearson correlation between decoded and ground-truth patient coordinates across layers for four 7B backbones. Base and PS denote the pretrained and PS-Coord checkpoints, respectively; Δr is their difference. **(b)** Patient-frame accuracy (%) on cases with $90°$ and $270°$ rotations held out during training. L/R and A/P denote patient left/right and anterior/posterior, respectively.

MIRP evaluates the relative spatial relationship between two named organs, whereas MSD tests the patient-side localization of a marked lesion. Both tasks require the model to utilize the spatial evidence of the current image rather than anatomical priors. The spatial training also improves general medical question answering. PS-Coord improves performance on SLAKE across all eight backbones, while yielding gains of 4.9–16.0% on VQA-RAD for the four larger models.

5.3 Generalization: Unseen Modalities and Unseen Orientations

PS-Coord generalizes well beyond the configurations seen during training, which improves spatial reasoning on both unseen modalities and unseen rotations. On radiographs (Table 1), MedBLINK improves across all eight backbones, while RSNA improves on six of eight. On held-out $90°$ and $270°$ rotations (Table 2b), the four 7B models improve from near chance to 97.3% left/right and 93.5% anterior/posterior accuracy on average. These provide evidence that PS-Coord establishes a patient-centered coordinate representation within the spatial subspace of medical MLLMs rather than memorizing training-specific display configurations.

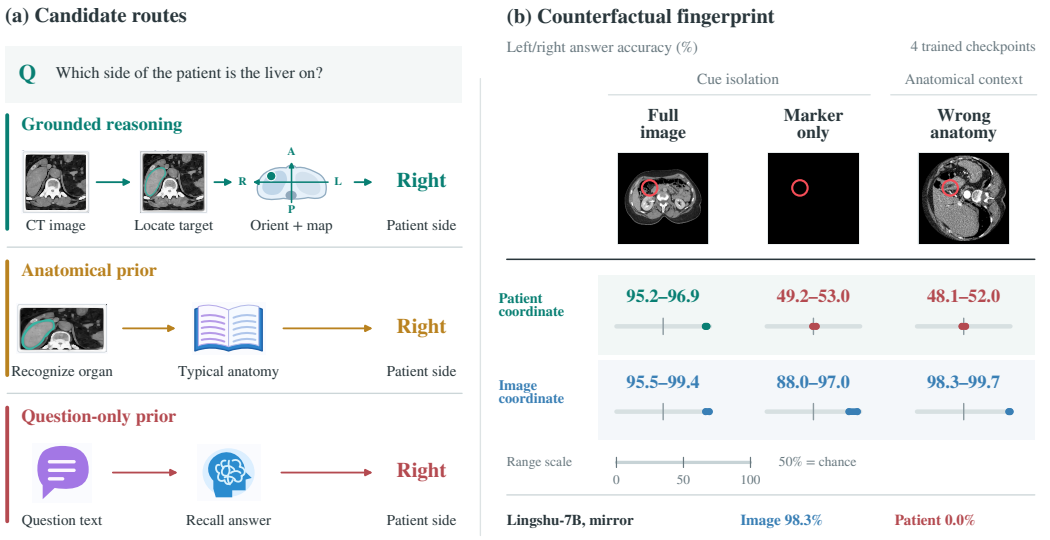

Figure 6: **Reasoning routes and counterfactual tests for patient-coordinate grounding.** (a) Three routes can yield the same patient-side answer. (b) Patient- and image-relative left/right accuracy for four trained 7B checkpoints. Removing or replacing anatomy reduces patient-relative accuracy to chance, whereas background replacement leaves image-relative accuracy high.

5.4 Emergence of Patient Coordinates.

The performance improvements are accompanied by a marked change in the hidden representations. Before training, peak patient-coordinate correlations range from 0.223 to 0.267 across the four 7B backbones. After PS-Coord training, they increase to 0.884–0.900 (Table 2a), with peaks at layers 18–19, consistent with the supervised layers. Image-coordinate readability is preserved and further strengthened over the same layers, increasing from 0.907–0.936 to 0.978–0.985. These results indicate that PS-Coord induces a linearly accessible patient-centered coordinate representation while preserving the image-coordinate information already present in the pretrained models.

5.5 Grounded Medical Spatial Reasoning.

As illustrated in Figure 6a, a correct patient-side answer can arise through multiple routes, including grounded medical spatial reasoning, anatomical priors, and question-level shortcuts. We therefore evaluate four PS-Coord-trained 7B models under controlled counterfactual inputs: full images, marker-only images, and images with the surrounding anatomy replaced. The decrease in patient-coordinate accuracy after anatomical background replacement, while image-coordinate accuracy remains largely unchanged in Figure 6b, suggests that PS-Coord-trained models rely on the anatomical context of the current image, rather than on question text or display position alone.

6 Conclusion

We investigate how medical MLLMs represent spatial information and whether that information maps onto the patient-coordinate that clinical interpretation requires. Our analyses reveal that medical MLLMs have a linearly accessible image-coordinate subspace, but this subspace is not connected to the patient coordinate. PS-Coord addresses this gap by jointly supervising image and patient coordinates within intermediate hidden states. Across eight backbones, PS-Coord training improves medical spatial reasoning, generalizes to held-out modalities and rotations , and yields gains on medical VQA tasks. Increased patient-coordinate readability and dependence on current anatomical context provide complementary evidence for learned patient-coordinate interpretation. These findings suggest that PS-Coord enhances anatomically grounded spatial reasoning in medical MLLMs, leading to improved performance on medical spatial reasoning tasks. Extending this formulation from two-dimensional displays to volumetric anatomy is a natural direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools (ChatGPT and Claude) to polish the English prose of the manuscript, to assist with the implementation of the training and evaluation code and to assist with the analysis and interpretation of the experimental results. We have not used generative AI tools for generating or cleaning data, for research ideation, or for designing experiments. Additionally, we used generative AI tools for LaTeX formatting, editing and debugging code. We have reviewed all AI-assisted work: AI-assisted code was reviewed and tested by the authors, and all reported results were produced by the authors' own experimental runs; AI-assisted interpretations of results were verified by the authors against the raw experimental outputs; and all AI-edited text was checked by the authors for technical accuracy.

REFERENCES

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, Bram van Ginneken, Michel Bilello, Patrick Bilic, Patrick F Christ, Richard K G Do, Marc J Gollub, Stephan H Heckers, Henkjan Huisman, William R Jarnagin, Maureen K McHugo, Sandy Napel, Jennifer S Golia Pernicka, Kawal Rhode, Catalina Tobon-Gomez, Eugene Vorontsov, James A Meakin, Sebastien Ourselin, Manuel Wiesenfarth, Pablo Arbeláez, Byeonguk Bae, Sihong Chen, Laura Daza, Jianjiang Feng, Baochun He, Fabian Isensee, Yuanfeng Ji, Fucang Jia, Ildoo Kim, Klaus Maier-Hein, Dorit Merhof, Akshay Pai, Beomhee Park, Mathias Perslev, Ramin Rezaiifar, Oliver Rippel, Ignacio Sarasua, Wei Shen, Jaemin Son, Christian Wachinger, Liansheng Wang, Yan Wang, Yingda Xia, Daguang Xu, Zhanwei Xu, Yefeng Zheng, Amber L Simpson, Lena Maier-Hein, and M Jorge Cardoso. The Medical Segmentation Decathlon. *Nature communications*, 13(1):4128, 2022. doi: 10.1038/s41467-022-30695-9. URL https://doi.org/10.1038/s41467-022-30695-9.

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, 2024. URL https://arxiv.org/abs/2404.00578.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report, 2025. URL https://arxiv.org/abs/2502.13923.

Mahtab Bigverdi, Wisdom Ikezogwo, Kevin Zhang, Hyewon Jeong, Mingyu Lu, Sungjae Cho, Linda Shapiro, and Ranjay Krishna. MedBLINK: Probing Basic Perception in Multimodal Language Models for Medicine, 2025. URL https://arxiv.org/abs/2508.02951.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. HuatuoGPT-Vision, Towards Injecting Medical Visual Knowledge into Multimodal LLMs at Scale, 2024. URL https://arxiv.org/abs/2406.19280.

Xiaotang Gai, Jiaxiang Liu, Yichen Li, Zijie Meng, Jian Wu, and Zuozhu Liu. 3D-RAD: A Comprehensive 3D Radiology Med-VQA Dataset with Multi-Temporal Analysis and Diverse Diagnostic Tasks. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2506.11147.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, Weicheng Dai, Murong Xu, Hadrien Reynaud, Muhammed Furkan Dasdelen, Bastian Wittmann, Tamaz Amiranashvili, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, Abdullah Alanbay, Anjany Sekuboyina, Berkan Lafci, Ahmet Kaplan, Zhiyong Lu, Malgorzata Polacin, Bernhard Kainz, Christian Bluethgen, Kayhan Batmanghelich, Mehmet Kemal Ozdemir, and Bjoern Menze. Generalist foundation models from a multimodal dataset for 3D computed tomography. *Nature biomedical engineering*, 10(8):1610–1628, 2026. doi: 10.1038/s41551-025-01599-y. URL https://doi.org/10.1038/s41551-025-01599-y.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. MedVLThinker: Simple Baselines for Multimodal Medical Reasoning. In *Machine Learning for Health*, 2025. URL https://arxiv.org/abs/2508.02669.

Elliot Landau, David Hirschorn, Iakovos Koutras, Alexander Malek, and Seleshie Demissie. Preventing errors in laterality. *Journal of digital imaging*, 28(2):240–246, 2015. doi: 10.1007/s10278-014-9738-4. URL https://doi.org/10.1007/s10278-014-9738-4.

LASA Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. Lingshu: A Generalist Foundation Model for Unified Multimodal Medical Understanding and Reasoning, 2025. URL https://arxiv.org/abs/2506.07044.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5:180251, 2018. doi: 10.1038/sdata.2018.251. URL https://doi.org/10.1038/sdata.2018.251.

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. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.00890.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. SLAKE: A Semantically-Labeled Knowledge-Enhanced Dataset for Medical Visual Question Answering. In *IEEE International Symposium on Biomedical Imaging*, 2021. URL https://arxiv.org/abs/2102.09542.

Guimeng Liu, Tianze Yu, Somayeh Ebrahimkhani, Lin Zhi Zheng Shawn, Kok Pin Ng, and Ngai-Man Cheung. How Do Medical MLLMs Fail? A Study on Visual Grounding in Medical Images. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2603.14323.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Minal Jagtiani Sangwaiya, Shyla Saini, Michael A Blake, Keith J Dreyer, and Mannudeep K Kalra. Errare humanum est: frequency of laterality errors in radiology reports. *AJR. American journal of roentgenology*, 192(5):W239–44, 2009. doi: 10.2214/ajr.08.1778. URL https://doi.org/10.2214/ajr.08.1778.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, Liron Yatziv, Tiffany Chen, Bram Sterling, Kenneth Philbrick, Richa Tiwari, Yun Liu, Madhuram Jajoo, Chandrashekar Sankarapu, Swapnil Vispute, Harshad Purandare, Abhishek Bijay Mishra, Sam Schmidgall, Tao Tu, Anil Palepu, Chunjong Park, Tim Strother, Rahul Thapa, Yong Cheng, Preeti Singh, Kat Black, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Joelle Barral, Tris Warkentin, Shravya Shetty, Dale Webster, Sunny Virmani, David F. Steiner, Can Kirmizibayrak, and Daniel Golden. MedGemma 1.5 Technical Report, 2026. URL https://arxiv.org/abs/2604.05081.

George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, Ritu R Gill, Myrna C B Godoy, Stephen Hobbs, Jean Jeudy, Archana Laroia, Palmi N Shah, Dharshan Vummidi, Kavitha Yaddanapudi, and Anouk Stein. Augmenting the National Institutes of Health Chest Radiograph Dataset with Expert Annotations of Possible Pneumonia. *Radiology. Artificial intelligence*, 1(1):e180041, 2019. doi: 10.1148/ryai.2019180041. URL https://doi.org/10.1148/ryai.2019180041.

Quoc-Huy Trinh, Xi Ding, Yang Liu, Zhenyue Qin, Xingjian Li, Gorkem Durak, Halil Ertugrul Aktas, Andrea M. Bejar, Ulas Bagci, and Min Xu. Beyond Medical Diagnostics: How Medical Multimodal Large Language Models Think in Space, 2026. URL https://arxiv.org/abs/2603.13800.

Lehan Wang, Haonan Wang, Honglong Yang, Jiaji Mao, Zehong Yang, Jun Shen, and Xiaomeng Li. Interpretable Bilingual Multimodal Large Language Model for Diverse Biomedical Tasks. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.18387.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. *Radiology. Artificial intelligence*, 5(5):e230024, 2023. doi: 10.1148/ryai.230024. URL https://doi.org/10.1148/ryai.230024.

Daniel Wolf, Heiko Hillenhagen, Billurvan Taskin, Alex Bäuerle, Meinrad Beer, Michael Götz, and Timo Ropinski. Your other Left! Vision-Language Models Fail to Identify Relative Positions in Medical Images. In *Medical Image Computing and Computer Assisted Intervention*, 2025. URL https://arxiv.org/abs/2508.00549.

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. doi: 10.1038/s41467-025-62385-7. URL https://www.nature.com/articles/s41467-025-62385-7.

A BACKBONE MODELS

We evaluate eight checkpoints from four model families, spanning from 7B to 34B parameters. Qwen2.5-VL provides a general-purpose multimodal reference; the other families incorporate medical multimodal training. Each Base/PS-Coord comparison starts from the same public checkpoint. Parameter-count labels identify the released model scales.

Qwen2.5-VL(Bai et al., 2025). Qwen2.5-VL is a general-purpose vision–language model with a visual encoder and a language model supporting multimodal instruction following. We use its 7B and 32B instruction checkpoints to assess whether patient-frame supervision benefits a model without the medical specialization of the other families.

Lingshu(LASA Team et al., 2025). Lingshu is a medical multimodal model family developed for medical image understanding and reasoning. Its 7B and 32B checkpoints provide medically adapted backbones at the same nominal scales as Qwen2.5-VL.

MedVLThinker(Huang et al., 2025). MedVLThinker studies multimodal medical reasoning, including reinforcement-learning-based adaptation . We use the released 7B and 32B `RL_m23k` checkpoints. This identity matters because a family name alone does not specify the training stage of the evaluated model.

HuatuoGPT-Vision(Chen et al., 2024). HuatuoGPT-Vision targets broad medical visual knowledge and medical question answering. The selected 7B checkpoint uses a Qwen2.5-VL implementation. The 34B checkpoint instead uses its native CLIP visual tower, multimodal projector, and Yi/LLaVA language interface. We preserve this native architecture rather than replace it with Qwen modules. Consequently, absolute layer indices should not be interpreted as identical relative depths across scales or architectures for HuatuoGPT-Vision-34B.

Backbone	Public checkpoint identifier	Blocks
Qwen2.5-VL-7B	`Qwen/Qwen2.5-VL-7B-Instruct`	28
Lingshu-7B	`lingshu-medical-mllm/Lingshu-7B`	28
MedVLThinker-7B	`UCSC-VLAA/MedVLThinker-7B-RL_m23k`	28
HuatuoGPT-Vision-7B	`FreedomIntelligence/HuatuoGPT-Vision-7B-Qwen2.5VL`	28
Qwen2.5-VL-32B	`Qwen/Qwen2.5-VL-32B-Instruct`	64
Lingshu-32B	`lingshu-medical-mllm/Lingshu-32B`	64
MedVLThinker-32B	`UCSC-VLAA/MedVLThinker-32B-RL_m23k`	64
HuatuoGPT-Vision-34B	`FreedomIntelligence/HuatuoGPT-Vision-34B-hf`	60

Table 3: **Evaluated backbone checkpoints.** Blocks denotes the number of language Transformer blocks. Checkpoint identifiers disambiguate model versions; processor and chat-template choices remain part of the evaluation protocol.

Model interface. Each checkpoint uses its corresponding processor and tokenizer. Image content is placed in the user message, followed by the question and the generation prefix. For processors exposing dynamic-resolution limits, minimum and maximum image areas are 128^2 and 512^2 pixels. These are processor bounds, not a claim that every source image is resized to a single square resolution. The native HuatuoGPT-Vision-34B image interface is retained. Its historical newline-template issue is discussed in Appendix D.

B DATA SOURCES, CONSTRUCTION, AND SEPARATION

B.1 TRAINING SOURCES AND SAMPLE COMPOSITION

TotalSegmentator(Wasserthal et al., 2023). TotalSegmentator supplies CT volumes and anatomical segmentations. We use canonical axial slices and their organ annotations to construct spatial questions with explicit target markers. Segmentations support slice filtering, source auditing, and

Training question family	Rows
Patient left/right	8,558
Patient anterior/posterior	6,394
Image left/right	4,959
Displayed back direction	6,025
Reflection status	6,064
Spatial subtotal	32,000

Table 4: **Training-data composition.** Counts report the full training pool. Checkpoints trained with a 24,000-input budget use a shuffled subset of this pool.

anatomical identification; the spatial prompts do not supply segmentation masks or numeric coordinates. The source-traceable training collection contains 32,000 spatial question–display examples derived from 1,037 canonical slices across 683 cases.

B.2 Slice filtering and image rendering

Source CT volumes and segmentation volumes are reoriented to the closest canonical RAS axis ordering with NiBabel. Axial slices are windowed at level 40 HU and width 400 HU, clipped to $[0, 255]$, and stored as 8-bit PNGs. Transposition and reversal of both displayed axes produce the radiological convention used here: patient-right at image-left and anterior at image-top. Source manifests retain slice and case identities. The coordinate labels use the normalized source-image frame, rather than the physical NIfTI affine coordinates.

The body-center routine thresholds the source grayscale image at intensity > 20, selects the largest connected component, and uses its unweighted centroid $c = (c_x, c_y)$. Candidate marker pixels remain after six binary-erosion iterations; a slice is retained only if more than 50 candidates remain. To select slices with identifiable orientation, the anatomical filter requires a spinal-cord centroid below the body center and a lateral-organ annotation with area at least 100 pixels and $|x_{\text{organ}} - c_x| \geq 0.1W$.

Rendering uses the largest component above intensity 20, dilated by eight pixels and intersected with a circular field of view of radius

$$r_{\text{FOV}} = \max(8,\ 0.97 \min(c_x, W - 1 - c_x, c_y, H - 1 - c_y) - 1)\,. \tag{8}$$

Marker centers lie within $0.85 r_{\text{FOV}}$. The square canvas is centered on the body; its half-width is the ceiling of the maximum distance from c to a source-image corner. Pixels outside the rendering mask are black. The training marker is a red circle with radius $\max(4, \lfloor 0.035W \rfloor)$ and a two-pixel outline. Horizontal reflection precedes bilinear counterclockwise rotation. This masking and centering reduce access to the CT table and rectangular acquisition boundaries as orientation cues.

In the training samples, family probabilities are 0.30 for patient left/right, 0.25 for anterior/posterior, and 0.15 each for back direction, reflection, and image left/right. Patient-left/right examples require $|p_x - c_x| \geq 0.08W$; anterior/posterior examples require $|p_y - c_y| \geq 0.08H$; image-left/right examples require transformed $|d_x| \geq 0.05W$.

Angles are sampled on a 15° grid, excluding

$$\{75, 90, 105, 255, 270, 285\}^\circ.$$

The retained angles are

$$\{0, 15, 30, 45, 60, 120, 135, 150, 165, 180, 195, 210, 225, 240, 300, 315, 330, 345\}^\circ,$$

with reflection probability 0.5. The principal held-out rotation results evaluate 90° and 270°.

B.3 Training prompts

Every spatial question begins with: *This is an axial CT slice with a red circle marker drawn on it. The image may be rotated or mirror-reflected.* The question suffixes are:

1. *Is the red circle on the patient's left side or the patient's right side? Answer with one word: left or right.*

2. *Is the red circle in the anterior (front) half or the posterior (back) half of the patient's body? Answer with one word: anterior or posterior.*
3. *In which direction of the image is the patient's back? Answer with one of: top, top-right, right, bottom-right, bottom, bottom-left, left, top-left.*
4. *Is this image mirror-reflected (a left-right flip of the patient), regardless of rotation? Answer with one word: yes or no.*
5. *In the image, is the red circle in the left half or the right half of the image? Answer with one word: left or right.*

Coordinates and the sampled transformation parameters are auxiliary targets and are absent from the input text.

B.4 Training–evaluation separation

The audited candidate training pool contains 6,307 slices from 1,054 cases. The principal held-out pool contains 2,193 slices from 93 cases; a separate chest pool contains 110 slices from 38 cases. Neither evaluation pool shares a case or an exactly duplicated original PNG with the training pool.

C Training Implementation and Hyperparameters

C.1 Objective and supervised token

For a spatial example, the two affine heads read the state at the final prompt token before the answer. The supervised layer is returned as `hidden_states[18]`, corresponding to the output after the eighteenth language block. Prompt length is computed before answer concatenation, so the readout position is `prompt_length-1`. The heads predict the two-dimensional image and patient coordinates, and the loss is

$$\mathcal{L} = \mathcal{L}_{\text{ans}} + \lambda_c \left[\frac{\widehat{b} - b_2^2}{2} + \frac{\widehat{i} - i_2^2}{2} \right], \qquad \lambda_c = 0.5. \tag{9}$$

Answer cross-entropy averages over answer tokens, including the end-of-sequence token; prompt labels are masked. For VQA examples, only $\mathcal{L}_{\text{ans}}$ is active. The two active affine heads contain $4d + 4$ parameters.

C.2 Optimization

The language model is adapted with LoRA (Hu et al., 2022), while the visual encoder and multimodal projection modules are fully updated. Model weights use bfloat16, and auxiliary heads use float32. Table 5 gives the shared settings.

LoRA targets the language-model projections `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`, with no trained LoRA bias. Gradient checkpointing is enabled. Qwen-compatible implementations save the full visual module; the native 34B implementation saves its visual tower and multimodal projector separately. At inference, the adapted language decoder produces answers without consuming the auxiliary heads' predictions.

The four 7B main checkpoints process 32,000 inputs and perform 5,000 optimizer updates, with seed 0. The four larger main checkpoints process 24,000 inputs and perform 3,000 updates, with seed 0. Within each backbone, objective variants share data, shuffle order, seed, and optimizer budget. This matches training exposure, not FLOPs across different architectures.

C.3 Training procedure and checkpoint handling

1. Initialize the backbone, language adapters, trainable visual modules, and float32 auxiliary heads. Set the recorded Python and PyTorch seed and shuffle training rows.
2. Load an image and tokenize the question using the model-specific chat template and generation prefix. Record prompt length, append the answer and end-of-sequence token, and mask prompt labels.

Setting	Value
LoRA rank / scaling / dropout	16 / 32 / 0.05
Language-adapter peak learning rate	10^{-4}
Visual-module peak learning rate	10^{-5}
Coordinate-head peak learning rate	10^{-3}
Optimizer	AdamW; weight decay 0; otherwise PyTorch defaults
Scheduler	OneCycleLR; initial rising fraction 0.1
Microbatch / gradient accumulation	1 input / 8 inputs
Model gradient clipping	Global norm 1; separately stored heads excluded
Supervised layer / token	Hidden-state index 18 / final prompt token
Coordinate loss weight	$\lambda_c = 0.5$
Image area bounds, where supported	128^2–512^2 pixels
Precision	bfloat16 model; float32 heads

Table 5: **Shared training configuration.** The learning-rate, layer, rank, and optimizer settings are fixed configurations. Only the coordinate-loss weight has the systematic sensitivity sweep reported in Appendix H.

3. Run the multimodal forward pass. Compute answer cross-entropy and, on spatial rows, the two coordinate errors from the last prompt state at layer 18.
4. Divide the loss by the accumulation factor 8 and backpropagate. Every eight inputs, clip the trainable model gradients to norm 1, update AdamW (Loshchilov & Hutter, 2019), advance the one-cycle scheduler, and clear gradients.
5. Stop at the recorded input budget and save the language adapter, updated visual modules, and coordinate heads. Preserve the model configuration and source/data hashes used by the run.

Training environment The training environment includes Python 3.12, CUDA 13.0, PyTorch 2.14.0, Transformers 5.17.0, PEFT 0.20.0, Accelerate 1.15.0, NumPy 2.5.3, SciPy 1.18.1, Pillow 12.3.0, and `qwen-vl-utils` 0.0.14.

D Benchmark Definitions and Evaluation Protocols

D.1 Benchmarks

MIRP (Wolf et al., 2025). MIRP tests relative-position reasoning in medical images and the influence of anatomical priors. The main comparison uses 300 underlying questions with number, letter, and dot presentations, yielding 900 evaluated presentations. These interfaces repeat questions and are not 900 independent cases. The provenance audit associates the underlying questions with 122 case groups.

Medical Segmentation Decathlon (MSD) (Antonelli et al., 2022). MSD provides segmented medical images across multiple anatomical tasks. We construct patient-side questions from tumor annotations. The main side subset contains 300 examples from 254 cases. The side label refers to the patient and remains unchanged when the display is transformed.

MedBLINK(Bigverdi et al., 2025). MedBLINK evaluates basic medical visual perception. The main table uses the pelvic orientation subset, with 200 images and two A/B option orders per image. The options are correct orientation and upside down. Accuracy averages over the two orders within image. The archived chest-orientation subset is a separate task. Exact-image grouping supports image-level accounting but does not establish that all images are from distinct patients.

RSNA(Shih et al., 2019). The RSNA pneumonia dataset provides chest radiographs with expert finding annotations. The main laterality evaluation contains 600 images and reports `image_orig`: whether the pneumonia opacity is in the left or right half of the image.

The subset retains cases whose positive bounding boxes all lie entirely within one image half and balances the two sides. The source DICOM pixel intensities are min–max normalized to 8-bit

Benchmark	Inputs	Main score	Reference or task
MIRP	900	Option accuracy	Image-relative organ relation; three interfaces
MSD	300	Option accuracy	Patient-side location of a marked tumor
MedBLINK	200	Two-order accuracy	Pelvic radiograph orientation; 400 prompts
RSNA	600	Option accuracy	Image-relative finding laterality
VQA-RAD	451	Normalized EM	Medical VQA; official question split
SLAKE	1,061	Normalized EM	Medical VQA; official question split

Table 6: **Main benchmark protocols.** Inputs are benchmark items or presentations, not independent patients. Option scoring and free generation are defined below.

grayscale and resized to 512×512. No extra target circle or box is drawn on the evaluated image. Thus this task is not the same marker-assisted setup as MSD.

VQA-RAD and SLAKE (Lau et al., 2018; Liu et al., 2021). VQA-RAD contains clinically generated radiology questions, while SLAKE combines medical visual questions with semantic and knowledge annotations. We evaluate 451 VQA-RAD and 1,061 SLAKE questions using normalized exact match. These are general medical VQA tasks, not direct measurements of coordinate prediction.

D.2 Answer scoring

For an answer option a, let $V(a)$ contain the first token IDs of the literal, space-prefixed, capitalized, and space-prefixed capitalized spellings. The common option scorer is

$$s(a) = \max_{v \in V(a)} \log p_\phi(v \mid I, q). \tag{10}$$

The selected answer maximizes $s(a)$. This is a maximum over first-token alternatives, not a full answer-string likelihood. The main MIRP and MSD comparison uses the frozen questions and their native options; RSNA compares left/right options. MedBLINK compare A/B and evaluate both option orders.

VQA generation is greedy, with a maximum of 24 new tokens. Normalization lowercases and trims text, removes parenthesized text, replaces English and Chinese punctuation with spaces, and collapses whitespace. It also maps the exact aliases x-ray/xray to a common form and normalizes the evaluator's exact yes/no synonym sets. No generative judge or semantic-similarity metric is used. The recheck of length-limited outputs retains the 24-token result as the main score rather than selecting the better result from different caps.

Historical HuatuoGPT-Vision-34B interface. The main table retains the archived chat template, including a documented newline-formatting issue for HuatuoGPT-Vision-34B. Base and adapted checkpoints are identified by their actual model states; later template repairs constitute a separate evaluation setting.

E Geometry of Image and Patient Coordinates

We derive the coordinate relationships used to construct supervision. These are properties of the coordinate labels and their transformations.

E.1 Conventions and forward and inverse maps

Let $W, H > 0$ denote the source image dimensions, $c \in \mathbb{R}^2$ the body center, and $p \in \mathbb{R}^2$ a target position in a canonical axial CT display. Pixel coordinates increase rightward and downward. Define

$$D = (W/2, H/2), \qquad S = \max(W, H)/2, \qquad u = p - c, \qquad b = -D^{-1}u. \tag{11}$$

Positive b_x denotes patient-right and positive b_y denotes anterior. A display transformation first reflects horizontally, if $m = 1$, and then rotates counterclockwise by θ:

$$Q_g = R_\theta M_m, \quad R_\theta = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}, \quad M_m = \begin{bmatrix} (-1)^m & 0 \\ 0 & 1 \end{bmatrix}. \tag{12}$$

The sign convention for R_θ accounts for the downward-positive image axis. Coordinates are body-centered offsets; translation to the square rendering canvas is removed before normalization.

[Exact frame conversion] For a fixed source image and target,

$$i_g = -S^{-1}Q_g D b, \qquad b = -SD^{-1}Q_g^\top i_g. \tag{13}$$

Proof. Substituting $u = -Db$ into $i_g = S^{-1}Q_g u$ gives the forward map. Since both R_θ and M_m are orthogonal, $Q_g^\top Q_g = I$. Multiplying the forward map by $-SD^{-1}Q_g^\top$ gives the inverse. □

The distinction between D and S matters for rectangular images. Patient coordinates normalize the two source axes separately, whereas image coordinates use one isotropic scale. A pure rotation of the normalized patient vector is therefore sufficient only when $W = H$. These coordinates describe a normalized two-dimensional anatomical frame.

[Display invariance, equivariance, and composition] For any two displays g, h of the same source and target, let $T_{h\leftarrow g} = Q_h Q_g^\top$. Then

$$b_h = b_g = b, \qquad i_h = T_{h\leftarrow g} i_g, \qquad i_{h2} = i_{g2}. \tag{14}$$

Writing $g = (\theta_g, m_g)$, composition satisfies

$$Q_g Q_h = Q_{(\theta_g + (-1)^{m_g}\theta_h,\ m_g \oplus m_h)}. \tag{15}$$

Proof. The canonical displacement u is unchanged by rendering. Eliminating u between $i_g = S^{-1}Q_g u$ and $i_h = S^{-1}Q_h u$ yields the second equality, and orthogonality gives norm preservation. Finally, $M_m R_\theta = R_{(-1)^m\theta} M_m$ and $M_{m_g} M_{m_h} = M_{m_g \oplus m_h}$ give the composition law. □

These identities apply to analytic coordinates. The training targets are computed from the analytic map rather than re-estimated from the rendered marker pixels.

E.2 Why image location does not identify patient location

[Ambiguity without a reference frame] If admissible observations include both an unreflected display and a horizontally reflected display with unknown g, image coordinate i alone does not uniquely determine patient coordinate b whenever $i_x \neq 0$.

Proof. Take $Q_0 = I$ and $Q_1 = M_1$. For any admissible i sufficiently near the origin with $i_x \neq 0$, the two pairs

$$(b_0, Q_0) = (-SD^{-1}i, I), \qquad (b_1, Q_1) = (-SD^{-1}M_1 i, M_1)$$

produce the same i by Equation (13). Their patient horizontal coordinates have opposite signs. Hence no single function of i alone can return the correct patient coordinate for both pairs. □

For an artificial balanced distribution containing these two indistinguishable coordinate observations, a classifier using only i has at most 50% patient-side accuracy. It explains why preserving an image-coordinate readout does not by itself supply a patient reference frame. Anatomical symmetry, incomplete fields of view, or atypical anatomy can also make orientation ambiguous even when the full image is available.

E.3 What dual-coordinate supervision constrains

Write the errors of the two heads as $e_b(g) = \widehat{b}_g - b$ and $e_i(g) = \widehat{i}_g - i_g$. All norms below are Euclidean.

[Prediction consistency from coordinate accuracy] For two displays of the same source target,

$$\widehat{b}_h - \widehat{b}_g \leq e_b(h) + e_b(g), \tag{16}$$

$$\widehat{i}_h - T_{h\leftarrow g}\widehat{i}_g \leq e_i(h) + e_i(g), \tag{17}$$

$$\widehat{i}_g + S^{-1}Q_g D\widehat{b}_g \leq e_i(g) + e_b(g). \tag{18}$$

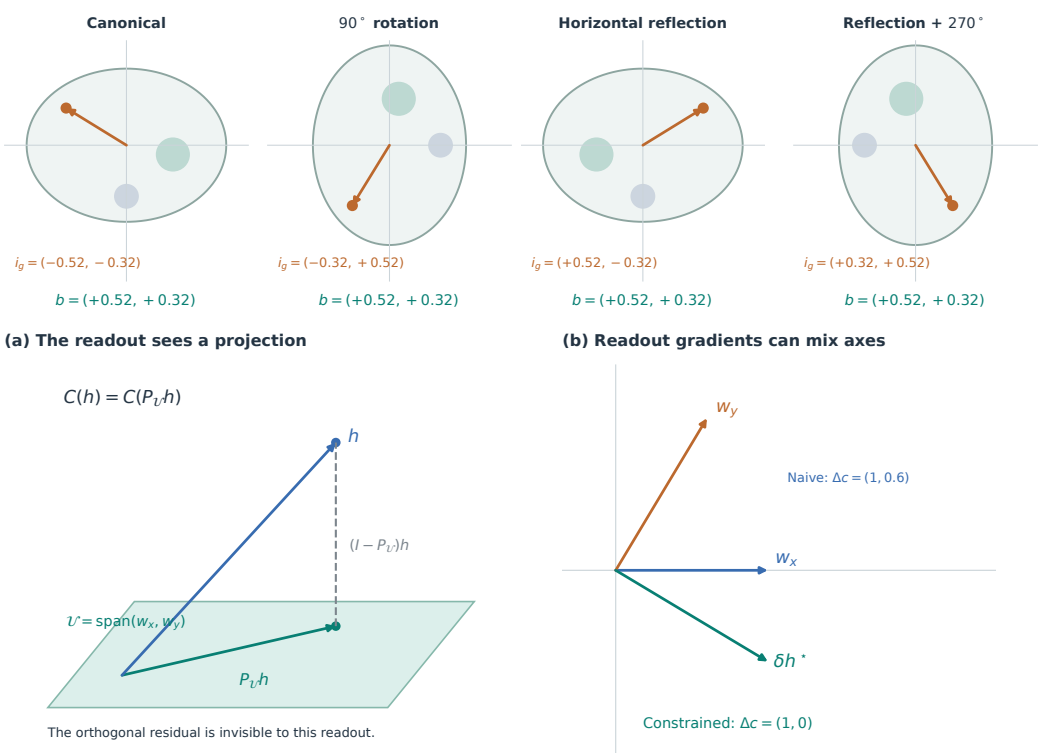

Figure 7: **Coordinate transformations and the geometry of a linear readout.** Top: The patient coordinate is unchanged, while the image-coordinate vector follows the display transformation. Bottom left: a hidden state decomposes into a component in the readout subspace and an orthogonal component invisible to that readout. Bottom right: correlated readout directions can cause cross-axis changes

Proof. Subtract the corresponding exact identities from Proposition E.1. The first two residuals become $e_b(h) - e_b(g)$ and $e_i(h) - T_{h\leftarrow g}e_i(g)$. Apply the triangle inequality and $T_{h\leftarrow g_2} = 1$. For the third, Equation (13) cancels the true coordinates, leaving $e_i(g) + S^{-1}Q_g De_b(g)$. Since $D_2 = S$ and $Q_{g_2} = 1$, the same inequality gives the result. □

For the two-dimensional mean squared error $\mathcal{E}(\widehat{c}, c) = \widehat{c} - c^2/2$, each error norm is $\sqrt{2\mathcal{E}}$. Thus, accurate dual-coordinate predictions imply approximate patient invariance, image equivariance, and agreement between the predicted frames.

F Mathematical Foundations of the Spatial Readout Subspace

F.1 Definition and Hidden-State Decomposition

Fix a model, layer, token-selection rule, and fitted affine coordinate readout. For $h \in \mathbb{R}^d$, write

$$C(h) = A^\top h + a, \qquad A = [w_x\ w_y] \in \mathbb{R}^{d\times 2}, \quad a \in \mathbb{R}^2. \tag{19}$$

Here A uses column readout vectors; it is the transpose of the head matrix convention used in the method. The intercept permits centering and does not affect the subspace.

[Readout subspace] The coordinate readout subspace is $\mathcal{U} = (A)$, with dimension $(A) \le 2$. The nullspace $\ker(A^\top)$ contains changes to the hidden state that leave this particular coordinate readout unchanged.

[Sufficiency for the fitted coordinate readout] Let † denote the Moore–Penrose pseudoinverse and define

$$P_{\mathcal{U}} = A(A^\top A)^\dagger A^\top. \tag{20}$$

Then $P_{\mathcal{U}}$ is the orthogonal projector onto $\mathcal{U}$, and

$$h = P_{\mathcal{U}}h + (I - P_{\mathcal{U}})h, \qquad C(h) = C(P_{\mathcal{U}}h). \tag{21}$$

Proof. Take the compact singular value decomposition $A = U_r\Sigma_r V_r^\top$. Substitution gives $P_{\mathcal{U}} = U_r U_r^\top$, which is symmetric, idempotent, and has image (A). Since $A^\top(I - U_r U_r^\top) = 0$, the second component makes no contribution to $A^\top h$. □

This proposition is an algebraic consequence of a linear readout. Its empirical significance comes from prediction on held-out cases, target-specific binding, prompt transfer, and intervention controls. It does not prove that all spatial information lies in two dimensions or that the rest of the hidden state is spatially irrelevant to other readouts or to a nonlinear decoder.

F.2 Relative decisions and coordinate error

For target states h_A, h_B and a shared axis readout w_k, the decoded difference is

$$\widehat{d}_k = C_k(h_B) - C_k(h_A) = w_k^\top(h_B - h_A). \tag{22}$$

The affine intercept cancels, and exchanging A, B negates the difference. A fixed comparison rule therefore produces antisymmetric relations except at ties.

[A sufficient margin for a correct relation] Let $d_k = c_k(B) - c_k(A)$ be the true coordinate difference, and e_A, e_B the two scalar prediction errors. If

$$|d_k| > |e_A| + |e_B|, \tag{23}$$

the decoded and true differences have the same sign.

Proof. $\widehat{d}_k = d_k + e_B - e_A$, so $|\widehat{d}_k - d_k| \le |e_A| + |e_B| < |d_k|$. This perturbation cannot cross zero. □

High correlation alone is insufficient to establish accurate relations: correlation is insensitive to positive rescaling and offsets, and errors near a decision boundary can change the answer. We therefore measure held-out relation accuracy directly rather than infer it from Pearson correlation.

F.3 Readout directions and minimum-norm edits

A readout vector specifies a linear functional. It is not automatically an isolated causal control axis. In particular, the edit $\delta h = \alpha w_x$ changes the decoded coordinates by

$$\delta C = \alpha \begin{bmatrix} {w_x}^2 \\ w_y^\top w_x \end{bmatrix}. \tag{24}$$

The vertical coordinate is preserved only if the two readout vectors are orthogonal, or if the dose is zero.

[Minimum-norm edit with a prescribed decoded change] If A has full column rank, the unique minimum-Euclidean-norm solution to $A^\top \delta h = \delta c$ is

$$\delta h^\star = A(A^\top A)^{-1}\delta c. \tag{25}$$

For rank-deficient A, a feasible $\delta c \in \mathrm{range}(A^\top)$ has minimum-norm solution $\delta h^\star = (A^\top)^\dagger \delta c$.

Proof. For full rank, substituting Equation (25) into the constraint gives δc. Every other feasible solution can be written as $\delta h^\star + z$, with $z \in \ker(A^\top) = \mathcal{U}^\perp$. Orthogonality gives $\delta h^\star + z^2 = \delta {h^\star}^2 + z^2$, minimized uniquely at $z = 0$. The pseudoinverse gives the same decomposition on the attainable output subspace when the rank is deficient. □

Let $m(h)$ be a downstream answer margin with all other intervention-site states fixed. Where differentiable,

$$m(h + \delta h) - m(h) = \nabla_h m(h)^\top \delta h + o(\delta h). \tag{26}$$

A nonzero decoded-coordinate change does not ensure a nonzero native-answer change: the downstream decoder may be insensitive to that direction.

F.4 Standardization, basis choice, and dual-frame training

If ridge regression is fitted to $z = F^{-1}(h-\mu)$, where $F = (s_1, \ldots, s_d)$ contains training-fold feature scales, a fitted standardized-space vector β corresponds to

$$w = F^{-1}\beta, \qquad a_{\text{raw}} = a_{\text{standardized}} - w^\top \mu. \tag{27}$$

Directions must be converted to the activation coordinates of the intervention before comparing Euclidean norms or editing states. More generally, under an invertible reparameterization $h' = Bh$, the equivalent readout is $A' = B^{-\top}A$. Linear recoverability is preserved, but angles and minimum Euclidean norms are not invariant to arbitrary changes of basis. The paper's subspace and intervention geometry refer to the recorded model activation basis.

For the jointly trained heads, define $\mathcal{U}_b = (W_b^\top)$ and $\mathcal{U}_i = (W_i^\top)$. Their combined span has dimension at most four, but they may overlap and need not be orthogonal. For two-dimensional MSE and a fixed training example,

$$\nabla_h \mathcal{L}_{\text{coord}} = W_b^\top(\widehat{b} - b) + W_i^\top(\widehat{i} - i). \tag{28}$$

Thus the coordinate-loss gradient at the supervised state lies in the current combined readout span. The heads and encoder both evolve during optimization, and answer gradients add other directions. This local identity motivates accessible coordinate representations without imposing a fixed four-dimensional module or proving convergence to a particular spatial basis.

G Representation Probes, Binding, and Interventions

The unified probe experiment uses 40 training-disjoint CT cases, eight displays per case, two swapped marker-color assignments, and clean/marker-only conditions: $1{,}280$ inputs per checkpoint. Prompts do not include organ names, numeric coordinates, or transformation angles. We retain the final prompt token to get the primary layer profiles.

The probe prompt is *This axial CT may be rotated or reflected for display. Locate the {color} circle in the displayed image and in the patient's anatomical frame.*

At each layer, five-fold case-grouped ridge regression predicts the two image and two patient coordinates. Features are standardized using training-fold statistics, and the primary ridge penalty is 1. Penalties 0.1 and 10 provide readout sensitivity checks; they do not change the trained MLLM. The frame-level score is the mean of the two axis-specific Pearson correlations computed from out-of-fold predictions. Axis-specific R^2 and mean absolute error are retained in the accompanying numeric data.

Specifically, if $Z_{\text{tr}}, Z_{\text{te}}$ are standardized features of dimension d, the implementation uses $K = Z_{\text{tr}}Z_{\text{tr}}^\top/d$ and predicts

$$\widehat{Y}_{\text{te}} = \frac{Z_{\text{te}}Z_{\text{tr}}^\top}{d}(K + \alpha I)^{-1}(Y_{\text{tr}} - \overline{Y}_{\text{tr}}) + \overline{Y}_{\text{tr}}. \tag{29}$$

The reported α refers to this normalized kernel; it is equivalent to penalty αd for an unnormalized primal sum-of-squares ridge objective. Training-fold feature standard deviations below 10^{-6} are replaced by 1. All views, colors, and conditions of a case remain in its fixed, quadrant-stratified fold. Constant-target or constant-prediction correlations are recorded as zero by the implementation.

Before training, the four 7B image-coordinate peaks occur at layers 19–22, whereas the larger-model peaks occur at layers 32–47. After training, patient-coordinate peaks increase across all eight backbones.

G.1 Gradient-derived directions and fixed-direction prompt transfer

The prompt-transfer and intervention experiments use gradient-derived directions, which are distinct from the ridge coefficients above. For relation task k, define the native answer-logit margin m_k and the gradients at the two target-name states as $G_{A,k}$ and $G_{B,k}$. On fitting cases,

$$\widetilde{v}_k = \frac{1}{N_{\text{fit}}}\sum_{j \in \text{fit}} \frac{G_{A,k}^{(j)} - G_{B,k}^{(j)}}{2}, \qquad v_k = \frac{\widetilde{v}_k}{\widetilde{v}_k + 10^{-9}}. \tag{30}$$

Backbone	Image Base	Image PS	Patient Base	Patient PS
Qwen2.5-VL-7B	0.934 (20)	0.983 (18)	0.234 (3)	0.897 (18)
Lingshu-7B	0.907 (21)	0.978 (19)	0.223 (28)	0.888 (19)
MedVLThinker-7B	0.936 (19)	0.985 (19)	0.239 (2)	0.884 (18)
HuatuoGPT-Vision-7B	0.933 (22)	0.985 (19)	0.267 (21)	0.900 (18)
Qwen2.5-VL-32B	0.898 (47)	0.945 (48)	0.168 (1)	0.604 (64)
Lingshu-32B	0.908 (46)	0.963 (44)	0.278 (61)	0.740 (52)
MedVLThinker-32B	0.905 (47)	0.955 (48)	0.168 (1)	0.552 (63)
HuatuoGPT-Vision-34B	0.911 (32)	0.982 (35)	0.342 (58)	0.799 (43)

Table 7: **Peak coordinate correlations across eight backbones.** Each entry gives mean-axis Pearson r and its descriptive peak layer in parentheses. Peaks are selected independently for each frame and checkpoint; these are not same-layer contrasts.

Backbone	λ_c	Patient side	MIRP	MSD
Qwen2.5-VL-7B	0	88.50	83.10	60.00
	0.1	88.62	83.81	61.33
	0.25	88.62	85.48	61.33
	0.5	90.88	86.00	63.67
	1	90.88	85.90	63.00
	2	88.50	84.52	63.33
Lingshu-7B	0	90.00	83.24	66.67
	0.1	89.62	80.29	66.67
	0.25	91.12	83.10	67.67
	0.5	92.62	83.29	70.00
	1	92.00	81.71	67.33
	2	90.88	81.81	66.33

Table 8: **Completed coordinate-loss-weight sweep.** Scores are percentages under the matched-budget ablation protocols. $\lambda_c = 0$ is answer-only training, and 0.5 is the main configuration.

The left/right task supplies v_x and the above/below task supplies v_y. Subtracting the two target gradients extracts the average antisymmetric sensitivity of the relation margin.

We keep the directions fixed and evaluate four prompts per task: original wording, a concise paraphrase, reversed order of the referents in the sentence, and an explicit image-grounding instruction. For left/right, these are:

1. *This is an axial chest CT slice with two coloured circle markers drawn on it. In the image, is the {A} circle to the left or to the right of the {B} circle? Answer with one word: left or right.*
2. *Look at the marked CT image. Is the {A} marker left or right of the {B} marker? Reply only: left or right.*
3. *Relative to the {B} circle, is the {A} circle on the left or on the right? Respond with one word.*
4. *Inspect the displayed image rather than anatomical conventions. Is marker {A} horizontally left or right of marker {B}? Answer only left or right.*

H Loss-Weight Sensitivity

We evaluate $\lambda_c \in \{0, 0.1, 0.25, 0.5, 1, 2\}$ for Qwen2.5-VL-7B and Lingshu-7B under the same $24{,}000$-input budget. The $\lambda_c = 0$ and 0.5 arms reuse the matched ablation checkpoints; the other four values are additional completed runs. Data, seed, and example order are held fixed within each backbone. Zero corresponds to answer-only training, and 0.5 is the main configuration.

Across the tested weights, each reported metric remains within $2.4\,pp$ of the default for Qwen2.5-VL-7B and $3.7\,pp$ for Lingshu-7B. $\lambda_c = 0.5$ provides the strongest performance across the evaluated metrics.

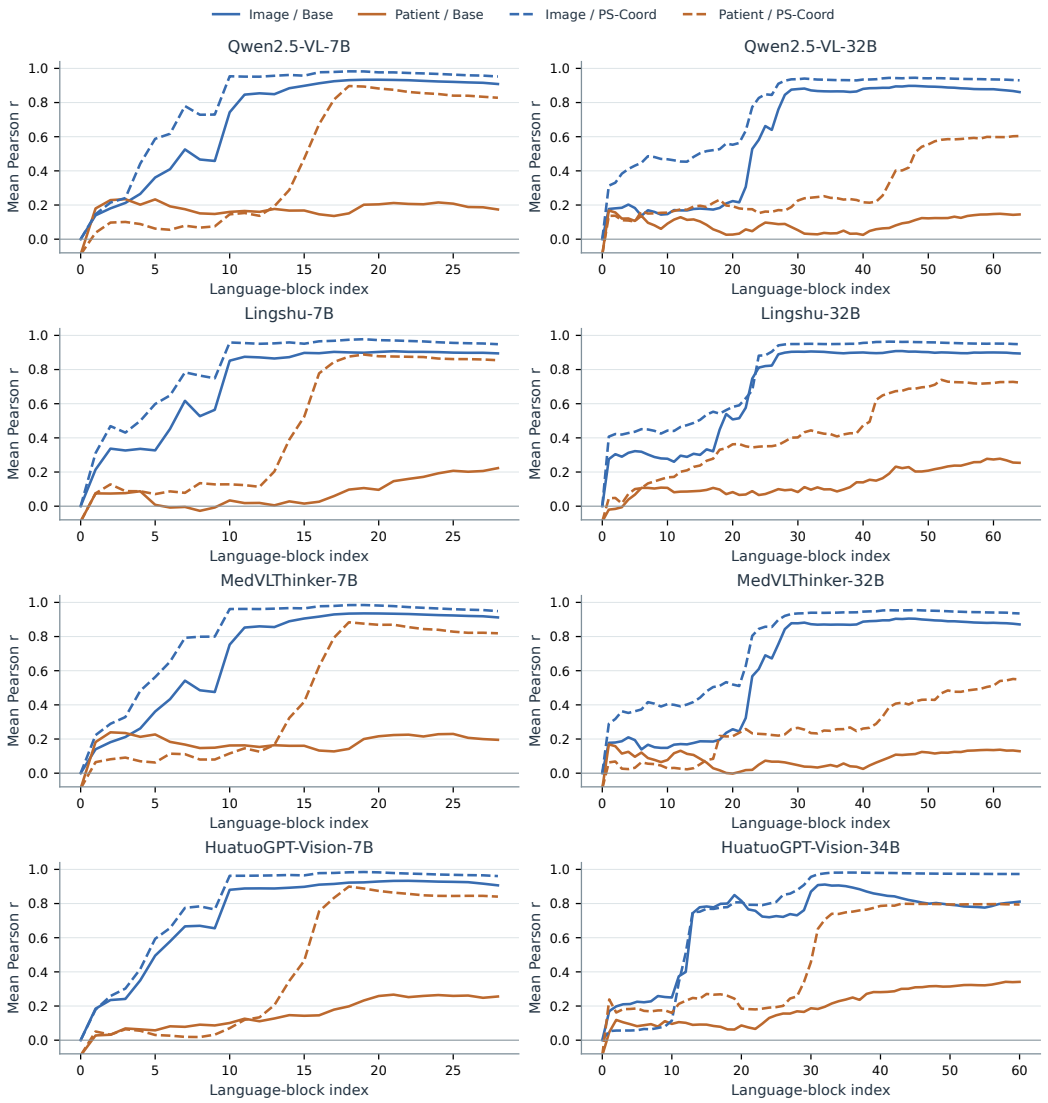

Figure 8: **Coordinate readability across all eight backbones.** Mean out-of-fold Pearson correlation over the two axes of each frame, from the final prompt token. Color distinguishes image and patient coordinates; solid and dashed lines distinguish Base and PS-Coord.

Backbone	LR Base	LR PS	AP Base	AP PS
Qwen2.5-VL-7B	51.25	97.50	46.41	94.53
Lingshu-7B	47.50	96.25	46.88	93.91
MedVLThinker-7B	48.91	98.12	48.59	92.50
HuatuoGPT-Vision-7B	53.59	97.50	49.06	93.12
Qwen2.5-VL-32B	52.03	89.06	51.88	88.44
Lingshu-32B	51.87	93.91	52.19	87.81
MedVLThinker-32B	50.47	92.50	52.34	92.19
HuatuoGPT-Vision-34B	52.50	70.94	44.84	87.50

Table 9: **Held-out** 90°/270° **patient-frame accuracy (%).** Forty cases, with both reflections, two prompt wordings, and both option orders. LR denotes left/right; AP denotes anterior/posterior.

Backbone	P-full	P-markers	P-replaced	I-full	I-markers	I-replaced
Qwen2.5-VL-7B	96.88	52.97	52.03	99.06	92.03	98.91
Lingshu-7B	95.16	49.22	48.12	98.59	97.03	99.06
MedVLThinker-7B	96.88	52.34	50.94	95.47	87.97	98.28
HuatuoGPT-Vision-7B	96.41	52.34	51.72	99.38	95.31	99.69
Qwen2.5-VL-32B	92.81	52.19	48.75	99.53	93.91	99.38
Lingshu-32B	96.25	51.72	49.53	98.59	94.69	99.38
MedVLThinker-32B	91.88	50.94	50.31	99.53	95.47	99.69
HuatuoGPT-Vision-34B	82.97	48.59	52.97	99.84	87.81	100.00

Table 10: **Anatomical-context controls across all trained backbones.** P denotes patient frame and I image frame. Replaced-anatomy images retain the original labels for diagnostic comparison.

Backbone	G-mean	G-all8	LR-mean	LR-all8	AP-mean	AP-all8
Qwen2.5-VL-7B	65.62	23.75	96.80	90.00	94.92	80.62
Lingshu-7B	66.56	38.12	95.08	78.75	93.52	77.50
MedVLThinker-7B	62.11	24.38	97.11	89.38	92.89	75.63
HuatuoGPT-Vision-7B	61.41	30.63	95.47	84.38	93.91	76.88
Qwen2.5-VL-32B	60.78	19.38	90.62	63.75	91.72	68.12
Lingshu-32B	71.41	32.50	94.69	72.50	92.11	72.50
MedVLThinker-32B	58.98	16.88	90.70	63.13	94.45	70.00
HuatuoGPT-Vision-34B	64.92	25.00	77.73	50.62	91.09	65.62

Table 11: **Mean and all-eight-correct rates (%) for trained checkpoints.** G denotes choosing the marker associated with a named organ. LR/AP denote patient axes. Each all8 score requires all eight displays to be correct for a target, wording, and option order, then averages these groups.

I Held-out rotation generalization

The 90° and 270° results average both reflections, two wordings, and both option orders. Table 9 includes all eight backbones. The four 7B checkpoints achieve average patient left/right and anterior/posterior accuracies of 97.3% and 93.5%, respectively, after training. The base models are near chance on those two metrics. The larger-model results are also retained rather than inferred from the 7B mean.

J Removing or replacing the anatomical background

We compare three conditions while preserving the queried marker positions. Full images retain the original CT. Marker-only images place the same markers on a black background. Anatomy-replacement images use the next case in the frozen case list as a donor background.

For the four trained 7B checkpoints under the common first wording, full-image patient-left/right accuracy is 95.2–96.9%. It falls to 49.2–53.0% with markers alone and 48.1–52.0% with replaced anatomy. Image-left/right accuracy remains 98.3–99.7% after anatomy replacement. This dissociation supports dependence on current anatomical context for patient-frame judgments.

K Correctness across displays and anatomical identity

A patient-side answer should remain constant across displays, but a constant incorrect answer does not establish patient-frame reasoning. We therefore also report all-eight-correct rates: for each target, wording, and option order, every one of the eight displayed predictions must be correct. Anatomical identity is measured separately by choosing which of two markers indicates a named organ. It is a marker-assisted recognition task, not free localization.

These stricter summaries use the same predictions as the single-view analysis. They help separate reliable patient-side consistency from image-side following and organ recognition. The gap between high patient-side accuracy and lower organ-identification accuracy also shows why learning a patient frame should not be equated with solving all anatomical grounding tasks.

L SCOPE, LIMITATIONS, AND REPRODUCIBILITY

Anatomical and clinical scope. The learned targets describe two-dimensional, body-centered coordinates on selected axial CT slices. The slice filter favors anatomy with recognizable orientation cues and does not test all clinical variations, such as situs inversus, severe resection, truncated fields of view, oblique planes, or acquisition artifacts.

Representation and causal interpretation. A successful linear probe shows recoverable information under a specified layer, token, and evaluation distribution. A successful intervention shows sensitivity to an edit under its specified dose and token scope. Neither alone demonstrates that the coordinate subspace uniquely mediates downstream medical reasoning. Our background controls support dependence on current anatomical context. The mathematical propositions specify what follows from a linear readout and coordinate labels, rather than assert an unobserved neural mechanism.

Compute and reproducibility. The accompanying supplement contains training data, configurations, and benchmark manifests. Reproducing training additionally requires the upstream model and dataset permissions, the exact processor/chat interface, model and adapter revisions, saved visual modules, and the recorded training code.

Data stewardship. The work uses existing medical datasets rather than newly recruited participants. Access, redistribution, and derived-data release remain subject to each dataset's terms. Model outputs and transformed image composites are research artifacts and should not be interpreted as clinical annotations for new patients.