A DATASET DETAILS

This section provides a detailed account of the dataset sources and characteristics that constitute the SpineBench benchmark. For each source dataset, we meticulously list its specific name, the imaging modality used, key data dimensions, the primary disease covered and the corresponding number of cases. As presented in Table 4, SpineBench consolidates four currently publicly available spinal disease datasets. These datasets encompass two common imaging modalities in the spinal domain, namely X-ray and Magnetic Resonance Imaging (MRI), and involve 11 prevalent diseases within the spinal field.

Table 4: Detailed information of the datasets included in SpineBench. The "Modality" column specifies the imaging modality used by the dataset. In the "Dim" (Dimension) column, 2D and 3D respectively denote the dimensionality of the original data. The "Disease" column outlines the primary types of spinal diseases contained within the dataset. The "Count" column indicates the number of cases actually adopted in SpineBench.

Challenge / Dataset	Modality	Dim	Disease	Count
BUU Spine Dataset	X-rays	2D	Spondylolisthesis	806
CSXA	X-rays	2D	Cervical Spine Curvature Disorders	4,962
RSNA	MRI	3D	Degenerative Spine Conditions	48,688
VinDr-SpineXR	X-rays	3D	Spinal Bone Lesions	23,202

Currently, publicly available datasets for research in the spinal domain are relatively scarce. In particular, under the influence of data barriers, open-source datasets specifically for spinal disease diagnosis are even more limited. As illustrated in Figure 8, based on previous research [2], we have identified only three additional datasets suitable for spinal disease diagnosis [40, 41, 46] and three for vertebral segmentation [13, 16, 50]. Furthermore, while access links were provided for some other datasets, they are not included in our survey as they were practically inaccessible. Evidently, the paucity of data has become a significant bottleneck hindering the in-depth development of AI applications in the spinal domain. Therefore, further promoting data sharing and openness in the spinal domain, while strictly adhering to research ethics, is crucial to advance technological progress in this field.

B IMAGE SIMILARITY COMPARISON

To optimize the visual similarity of incorrect options for QA pairs in SpineBench, this study ultimately employed the SigLIP2 model to extract image embeddings. In practice, in addition to SigLIP2, we also evaluated the performance of two other models, MedViT [32] and SigLIP [58], in extracting features from spinal disease images. The MedViT model integrates the local receptive field capabilities of Convolutional Neural Networks (CNNs) with the global connectivity of transformers, incorporating an attention mechanism, which can effectively enhance the robustness and generalization ability of medical diagnosis. The SigLIP model, on the other hand, utilizes a Sigmoid binary classification loss function in place of the traditional InfoNCE contrastive learning loss. This approach decouples the loss calculation for each sample from other samples within the batch, thereby greatly simplifying the implementation of

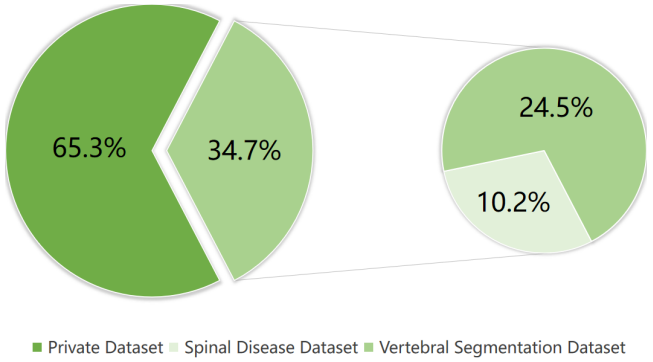


Figure 8: Statistical overview of existing spine datasets. Analysis reveals 17 open-source (5 for disease diagnosis, 12 for vertebra segmentation) and 32 private datasets.

distributed loss and improving computational efficiency, making it a preferred visual encoder for numerous advanced large multi-modal models. SigLIP2 builds upon SigLIP by further integrating techniques such as caption-based pre-training, self-supervised losses (e.g., self-distillation, masked image prediction), and online data filtering, thereby enhancing the encoding efficacy of SigLIP.

Operationally, we first randomly sampled 200 images for each spinal disease from the SpineBench to construct a disease image database. Subsequently, we utilized the aforementioned three models to extract feature embeddings from these images, forming corresponding embedding libraries. Then, for each image to be evaluated in SpineBench, its embedding was extracted. By calculating the cosine similarity between this embedding and each embedding in the library, we identified the three most visually similar different disease types in the library, excluding the image itself (if present). Finally, by combining the inherent "Healthy" option and the true disease label of the image, a set of highly confounding options was constructed for each QA pair. Through a comparative analysis of the similarity results generated by the three models, we found that MedViT performed suboptimally to effectively distinguish features of different spinal diseases, while SigLIP2 demonstrated superior performance compared to SigLIP. Consequently, the SigLIP2 model was ultimately selected to extract image feature embeddings for spinal diseases. Partial cosine similarity results obtained after extracting image embeddings and calculating similarities using the three models are presented in Table 6.The first row lists abbreviations for 12 spinal diseases, with their specific meanings detailed in Table 5.

C EVALUATION SETUP

C.1 Evaluation Protocol

This section will detail the evaluation procedure. We evaluate various types of MLLM, including medical-specific models, open-source general-purpose models, and API-based closed-source general-purpose models. For open-source models, we primarily selected versions with a parameter scale of approximately 7B for testing, with their model weights sourced from their respective official Hugging Face repositories. The specific steps of the evaluation are as follows:

Table 5: The abbreviations for 11 spinal diseases.

Name of Disease	Abbreviations
Disc space narrowing	Dsn.
Foraminal stenosis	Fs.
Subarticular stenosis	Sas.
Osteophytes	Ost.
Spinal canal stenosis	Scs.
Spondylolisthesis	Spondy.
Straight cervical vertebrae	Strcv.
Vertebral collapse	Vc.
Cervical kyphosis	Ck.
Cervical lordosis	Cl.
Sigmoid cervical vertebrae	Sigcv.

(1) **Prompt Input**: The predefined prompts, as delineated in Figure 9 and Figure 10, were input into the models under evaluation. The anticipated model output was strictly limited to option identifiers (e.g., "A", "B", etc.). Recognizing that contemporary large models incorporate built-in safety mechanisms designed to prevent the provision of potentially misleading medical judgments, their outputs typically tend to recommend consultation with a professional physician rather than offering direct diagnostic conclusions. To address this, we incorporated a simulated clinical scenario setting within the prompts to encourage the models to provide direct answers to all questions within SpineBench-Sub.
(2) **Instruction Following and Answer Extraction**: If a model failed to output a clear letter-based option identifier as expected, or if its output did not provide clear selection guidance, the Gemini-2.5-Pro model is employed to extract a candidate answer from the raw output.
(3) **Result Determination**: If a valid answer cannot be successfully extracted from the model's output, or if the extracted answer does not perfectly match the ground truth, the instance is judged as incorrect. Any successfully extracted answer is considered the model's final prediction for that question.

This standardized procedure is designed to ensure consistency and comparability in evaluating the instruction-following capabilities and answer accuracy of different models, and it has been specifically adapted to meet the stringent requirements for output format precision in medical tasks.

D MLLMS REASONING EVALUATION

For each MLLM under evaluation, we randomly sample 10 segments of their reasoning processes from their performance in each tasks (spinal disease diagnosis and spinal lesion localization). We subsequently invite collaborating clinical physicians to conduct a professional assessment of these reasoning processes from multiple perspectives, including clinical reasoning plausibility and the application of professional knowledge. The detailed scoring rubrics are presented in Figure 11, respectively.

Prompt for the spinal disease diagnose task

This is a simulated medical scenario for educational or evaluation purposes. You are to adopt the persona of a board-certified spine surgeon with extensive clinical experience. Your task is to analyze the following multiple-choice question and, drawing upon your simulated expertise, identify the single most accurate answer among the options. Important: Maintain your role throughout. Provide your reasoning for selecting the best option. Do not break character by stating you are an AI or advising consultation with a real physician. The Question is: {question}. The candidate Options are: [A: {optionA_text}, B: {optionB_text}, C: {optionC_text}, D: {optionD_text}, E: {optionE_text}]. After providing your detailed reasoning, you must conclude your entire response with the following exact phrase on a new line: My final option: [X] (where X is the letter: A, B, C, D, or E corresponding to your chosen answer)."
<image>

Figure 9: Prompt used to evaluate MLLMs in spinal disease diagnose task.

Prompt for the spinal lesion localization task

This is a simulated medical scenario for educational or evaluation purposes. You are to adopt the persona of a board-certified spine surgeon with extensive clinical experience. Your task is to analyze the following multiple-choice question and, drawing upon your simulated expertise, identify all appropriate answers among the options, as there may be more than one correct choice. Important: Maintain your role throughout. Provide your reasoning for selecting each of your chosen options. Do not break character by stating you are an AI or advising consultation with a real physician. The Question is: {question}. The candidate Options are: [A: {optionA_text}, B: {optionB_text}, C: {optionC_text}, D: {optionD_text}, E: {optionE_text}]. After providing your detailed reasoning, you must conclude your entire response with the following exact phrase on a new line: My final options: [X, Y, ...] (where X, Y, ... are the letters: A, B, C, D, or E corresponding to your chosen answers, separated by commas if multiple. For example: My final options: [A, C]).
<image>

Figure 10: Prompt used to evaluate MLLMs in spinal lesion localization task.

Scoring Guide for Model Reasoning Process

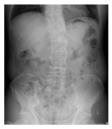
Clinical Rationality of Reasoning: Primarily considers the logicality of the overall thinking process, regardless of whether the answer is correct or not
Focus: How does the reasoning process on the surface of the model perform in terms of clinical logic and validity? Is it consistent with established medical knowledge, diagnostic principles, and anatomical understandings related to spinal pathology?
Scoring Criteria:
1: The reasoning appears clinically illogical, ineffective, or fundamentally flawed.
3: The reasoning has some clinical basis but contains major defects or omissions.
5: The reasoning appears clinically reasonable, logical, and follows appropriate diagnostic pathways.

Recognition and Application of Key Visual Features: Primarily considers whether details in the image are observed
Focus: Does the reasoning process correctly identify and prioritize the most relevant visual evidence in the image (e.g., signal changes, structural deformities, displacement, specific anatomical landmarks) to support its conclusions?
Scoring Criteria:
1: Fails to identify key features or severely focuses on irrelevant details.
3: Identifies some relevant features but may misinterpret or omit key features.
5: Clearly identifies and appropriately uses the most critical visual features for specific diagnostic/localization tasks.

Quality of Spatial Reasoning (especially for localization tasks):
Focus: How does the reasoning process demonstrate an understanding of three-dimensional spatial relationships within the spine (e.g., vertebral levels, anterior-posterior structures, intervertebral foramen space, spinal canal relationships) even when interpreting two-dimensional images?
Scoring Criteria:
1: Demonstrates poor or incorrect spatial understanding.
3: Shows basic spatial awareness but may make errors in precise localization or spatial relationships.
5: Exhibits accurate and nuanced understanding of relevant spinal spatial anatomical structures.

Table 6: Results of cosine similarity calculations for all disease features derived from four images of patients with Spondylolisthesis. Abbreviations: the full terms of all departments are listed in Table 5.

ID	Correct Answer	Dsn.	Fs.	Sas.	Ost.	Scs.	Spondy.	Strcv.	Vc.	Ck.	Cl.	Sigcv.
MedViT												
0	Spondylolisthesis	0.991	0.987	0.973	0.972	0.986	0.944	0.981	0.991	0.931	0.942	0.932
1	Spondylolisthesis	0.994	0.994	0.969	0.975	0.989	0.968	0.983	0.991	0.962	0.965	0.962
2	Spondylolisthesis	0.993	0.993	0.986	0.989	0.993	0.979	0.992	0.992	0.968	0.975	0.970
3	Spondylolisthesis	0.995	0.996	0.984	0.990	0.995	0.996	0.990	0.996	0.995	0.996	0.995
SigLIP												
0	Spondylolisthesis	0.957	0.952	0.798	0.958	0.850	0.951	0.902	0.960	0.897	0.908	0.903
1	Spondylolisthesis	0.955	0.951	0.808	0.953	0.855	0.955	0.916	0.959	0.918	0.919	0.918
2	Spondylolisthesis	0.966	0.959	0.804	0.966	0.851	0.970	0.881	0.965	0.872	0.883	0.875
3	Spondylolisthesis	0.963	0.958	0.806	0.972	0.858	0.962	0.884	0.971	0.886	0.887	0.890
SigLIP2												
0	Spondylolisthesis	0.966	0.939	0.867	0.964	0.892	0.973	0.940	0.969	0.937	0.944	0.940
1	Spondylolisthesis	0.967	0.921	0.840	0.969	0.864	0.979	0.924	0.964	0.919	0.921	0.923
2	Spondylolisthesis	0.961	0.912	0.852	0.956	0.882	0.965	0.916	0.963	0.903	0.913	0.911
3	Spondylolisthesis	0.975	0.926	0.850	0.977	0.881	0.978	0.921	0.972	0.914	0.917	0.919



Question

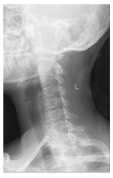
Which spinal disorder is this patient suffering from?

A: "Disc space narrowing",
B: "Osteophytes",
C: "Vertebral collapse",
D: "Healthy",
E: "Spondylolisthesis"

Disease Misdiagnosis

Let's carefully analyze the spinal radiograph provided. On inspection of the lumbar spine in this AP (anteroposterior) view, observe the following:- The vertebral bodies are well aligned, without significant anterior or posterior translation, making spondylolisthesis (option E) unlikely. - There is no visible vertebral body collapse (option C), as all vertebral heights appear maintained. - I do not see any prominent osteophyte formation (option B) along the anterior, lateral, or posterior vertebral body margins; the bony outlines are smooth. - The disc spaces are preserved; there is no clear evidence of narrowed intervertebral disc spaces (option A), which would typically appear as a reduction in the gap between vertebral bodies. Given the above findings, there is an absence of pathological radiographic signs. Thus, based on this image alone, there are no features of degenerative or structural spinal pathology visible. My final option: [D]

Figure 14: An example of disease misdiagnosis case.



Question

What type of spinal disease does this patient have?

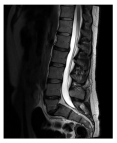
A: "Osteophytes",
B: "Disc space narrowing",
C: "Healthy",
D: "Foraminal stenosis",
E: "Straight cervical vertebrae"

Healthy Patient Misdiagnosis

To analyze this cervical spine lateral X-ray as a board-certified spine surgeon, I will focus on the following:1. **Alignment**: The cervical spine normally has a lordotic (gentle backward) curve. On this image, the cervical vertebrae are aligned almost in a straight line, lacking the expected lordotic curvature.2. **Disc Spaces**: I do not see significant narrowing of the intervertebral disc spaces; the height appears preserved across the levels.3. **Vertebral Bodies and Osteophytes**: There is no obvious evidence of bone spur (osteophyte) formation along the vertebral body margins.4. **Neural Foramina**: The foramen cannot be evaluated well on this lateral view, so foraminal stenosis cannot be definitively diagnosed here.5. **General Bone and Soft Tissue Health**: No obvious pathological fractures or significant soft tissue abnormality observed.**Analysis of Options:**- **A: Osteophytes**: Not clearly present in this radiograph.- **B: Disc space narrowing**: The disc spaces are maintained.- **C: Healthy**: The absence of the normal cervical lordosis suggests an abnormality.- **D: Foraminal stenosis**: No clear evidence, and not well evaluated on lateral films.- **E: Straight cervical vertebrae**: The cervical spine is indeed straight, lacking the natural lordotic curve. This is sometimes referred to as cervical straightening, which can be caused by muscle spasm (often from pain or trauma) or poor positioning but is still an abnormal finding.Based on the image, the most accurate answer is E: Straight cervical vertebrae (loss of cervical lordosis).My final option: [E]

Figure 15: An example of healthy patient misdiagnosis case.

E CASE STUDY



Question

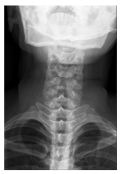
Can you specify the type of spinal disease this patient is experiencing?

A: "Spinal canal stenosis",
B: "Subarticular stenosis",
C: "Healthy",
D: "Osteophytes",
E: "Neural foraminal narrowing"

Correct Disease Diagnose

Based on the provided MRI image, the spinal canal appears narrowed, which is indicative of spinal canal stenosis. This condition involves the narrowing of the spinal canal, which can lead to compression of the spinal cord or nerve roots, resulting in symptoms such as pain, numbness, or weakness in the limbs. The narrowing is most prominent in the lumbar region, which is consistent with the image provided. The other options do not align with the findings visible in the MRI. Subarticular stenosis (B) typically involves the narrowing of the joint space, which is not evident here. Osteophytes (D) are bone spurs that can cause narrowing but are not the primary feature in this image. Neural foraminal narrowing (E) involves the narrowing of the openings through which nerves exit the spinal column, which is not the main issue in this case. The spinal cord and nerve roots appear to be within normal limits, ruling out a healthy state (C).Therefore, the most accurate diagnosis based on the MRI is spinal canal stenosis. My final option: A

Figure 12: An example of correct disease diagnosis case.



Question

What is the nature of this patient's spinal disease?

A: "Cervical kyphosis",
B: "Disc space narrowing",
C: "Healthy",
D: "Foraminal stenosis",
E: "Osteophytes"

Healthy Patient Diagnose

To address this question as a spine surgeon, I will carefully analyze the provided cervical spine X-ray (AP view). First, let's consider each of the answer choices:A: Cervical Kyphosis – This refers to an abnormal reversal of the normal cervical lordotic curvature, best visualized on a lateral X-ray. The provided image is AP, which is not sufficient for curve assessment. B: Disc Space Narrowing – This condition implies reduction in the normal height of the intervertebral disc spaces between vertebral bodies. C: Healthy – Implies the absence of pathological changes. D: Foraminal Stenosis – Narrowing of the neural foramina, which is generally assessed in oblique or lateral views, not AP.E: Osteophytes – Bone spurs, seen as bony outgrowths from the margins of the vertebral bodies. On close inspection of this AP cervical X-ray:- The vertebral body heights appear preserved and symmetric. - There are no obvious abnormal bony growths or outgrowths (osteophytes) at the margins of the vertebral bodies. - The disc spaces, as much as can be evaluated on an AP film, seem preserved and symmetric with no clear evidence of narrowing. - There is no apparent misalignment or evidence of significant sclerosis or lytic lesions. - Assessment for foraminal stenosis is not possible on an AP view, as the foramina are not visualized in this projection. - The AP view does not provide information on cervical curvature, so assessment for kyphosis isn't accurate. In summary, there are no radiographic signs of abnormality in disc space, osteophytes, or other bony pathology evident in this AP X-ray. Given that none of the listed pathologies are visible and the appearance is within normal limits, the image best represents a healthy cervical spine. My final option: [C]

Figure 13: An example of healthy patient diagnosis case.