SPINEDIAG: A MULTIMODAL SPINE SCOLIOSIS BENCHMARK FOR RADIATION-FREE DIAGNOSIS

Anonymous authors
Paper under double-blind review

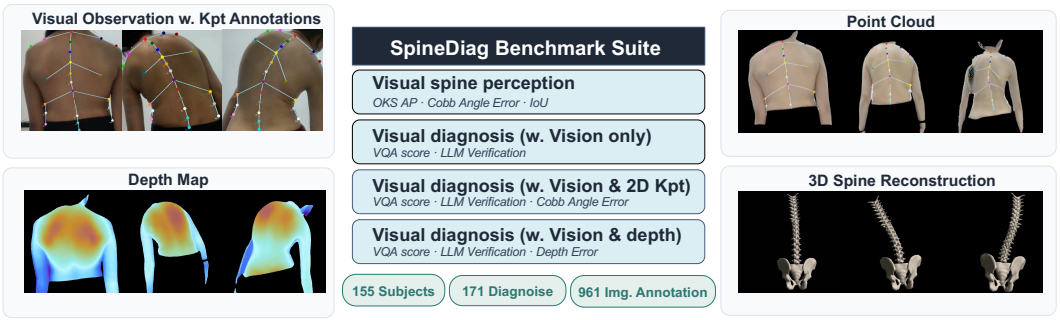


Figure 1: **SpineDiag overview.** Each subject contributes radiation-free data: Visual observations and multi-pose keyframes annotated with 24 anatomical keypoints, depth maps, and body-surface point clouds. Four benchmarks cover visual spine perception, visual diagnosis from vision only, vision with 2D keypoints, or vision with depth. Cobb-angle references are assessed by clinicians on post-visit X-ray or computed from visual observations.

ABSTRACT

Adolescent idiopathic scoliosis (AIS) is a common spinal deformity. However, large-scale diagnosis and repeated follow-up are limited by the radiation exposure of full-spine X-ray. To this end, we build **SpineDiag**, a paired multimodal clinical AIS dataset with 155 subjects and radiation-free modalities (visual observations, depth maps, and point clouds, annotated keypoints, and physician-verified diagnosis descriptions). To build a unified spine representation, we first design the 2D spinal keypoint protocol from medical expertise, then lift this 2D anatomical representation with depth to reconstruct a 3D kinematic spine model in simulation. We further propose a benchmark suite with four tasks: (1) visual spine perception, (2) visual diagnosis, (3) visual diagnosis with 2D keypoints, and (4) visual diagnosis with depth. The results are assessed with keypoint and geometric metrics, a clinically Visual Question Answering (VQA) score, and reported LLM verification diagnosis score. We train and evaluate pose estimators and vision-language models (VLMs) on these tasks. The experimental results expose open problems of current models in radiation-free diagnosis, which needs further investigation and more robust models.

1 INTRODUCTION

Adolescent idiopathic scoliosis (AIS) affects an estimated 5.2% of school-age people (Konieczny et al., 2013). Because the deformity can progress during growth, early screening and regular follow-up are important. Current assessment relies on hands-on examination, such as scoliometer measurement, and radiographic imaging. Contact-free and radiation-free assessment with cameras could make screening and follow-up more accessible and repeatable, which requires models that infer spinal deformity from visual observations of the back, such as RGB images, videos, and depth maps.

In this domain, training and evaluating current VLMs or perception models requires clinical data that pair observations with annotations such as physician diagnoses. However, no public dataset

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.

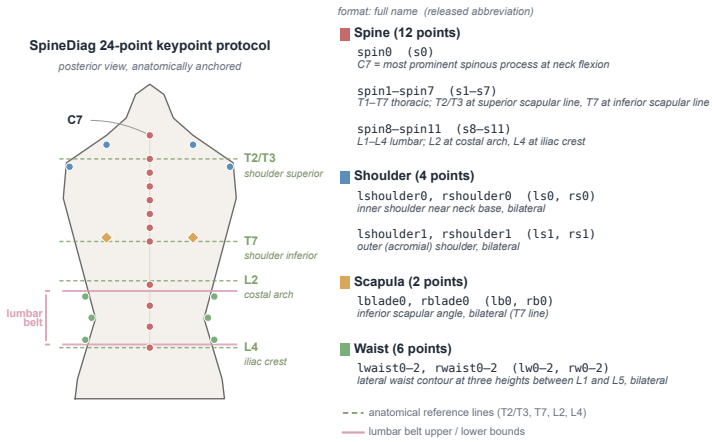


Figure 2: **Anatomical keypoint protocol.** The 12 spinal landmarks are anchored to reference lines, with bilateral shoulder, scapular, and waist points. Schematic; no subject photograph.

currently provides this pairing. Existing resources each miss part of this pairing. Clinical AIS studies estimate curve severity from back photographs (Yang et al., 2019), surface topography (Minotti et al., 2023; Hassan et al., 2025), or 3D ultrasound, which requires probe contact (Zheng et al., 2016). Each uses a single modality, and none provides a public multimodal release. Public spine imaging benchmarks (Wang et al., 2022; Sekuboyina et al., 2021; Deng et al., 2021; Pang et al., 2021) annotate vertebrae on X-ray, CT, or MR, but contain no visual observations of the body surface. Human pose and shape datasets (Ionescu et al., 2014; von Marcard et al., 2018; Patel et al., 2021; Black et al., 2023; Cai et al., 2022) drive modern pose, depth, and body-model research (Loper et al., 2015; Goel et al., 2023; Xia et al., 2025), but mostly capture healthy bodies without clinical labels. Recent spine keypoint datasets, SpineTrack (Ullah Khan et al., 2025) and SIMSPINE (Ullah Khan & Stricker, 2026), annotate spinal landmarks in natural images and multi-view motion capture, but target general human motion without clinical labels. As a result, pathological spinal anatomy remains out of distribution for current vision models.

We introduce **SpineDiag**, a paired multimodal clinical AIS dataset of 155 subjects and 171 clinical visits, contributed by multiple medical centres (Figure 1). Its visual observations comprise standard-position RGB images (posterior, lateral, and Adam's forward-bend) for all visits. Time-of-flight (ToF) depth maps and body-surface point clouds accompany 711 of the 961 keypoint-annotated frames. Five professional annotators label 24 anatomical keypoints under our defined protocol, followed by verification from the spine surgeon. We further lift these keypoints with depth to a 3D kinematic spine model to support medical simulation. Every visit has a Cobb-angle reference and physician-verified diagnosis descriptions over four fields: spinal curvature, shoulders, back, and scapulae. Cobb references are assessed by clinicians on post-visit X-ray or computed from visual annotations. All released data are de-identified and distributed under a data use agreement.

On these data we define four benchmarks (*centre* of Figure 1). *Visual spine perception* predicts the 24 anatomical keypoints from visual observations. It is scored with COCO OKS AP (Lin et al., 2014) and with the angle error and endpoint IoU of a geometric Cobb angle derived from the predicted spinal keypoints. The other three benchmarks answer nine multiple-choice questions over the four diagnosis fields. *Visual diagnosis* (`Diag`) uses visual observations only. *Visual diagnosis with 2D keypoints* adds annotated (`Diag-Kpt2D`) or predicted (`Diag-PredKpt2D`) keypoints. *Visual diagnosis with depth* adds depth as a point cloud (`Diag-Pointcloud`) or as depth-lifted 3D keypoints (`Diag-Kpt3D`). We score diagnosis with a clinically weighted VQA score that assigns 70% of the weight to spinal curvature, so a missed curve costs more than a minor shoulder finding. LLM verification of free-text diagnoses is reported separately, and keypoints predicted during reasoning are further scored by Cobb angle error and depth error. We evaluate four recent pose estimators (Sun et al., 2019; Xu et al., 2022; Yang et al., 2023; 2024) and fine-tune open vision–language models from 4B to 11B parameters (Qwen Team, 2025b; 2026; Li et al., 2024; Grattafiori et al., 2024), all under subject-level splits.

Our experiments reveal three findings. First, pathological visual observation challenges off-the-shelf pose estimators. OKS mAP ranges from 0.263 (ViTPose) to 0.756 (X-Pose), yet X-Pose still yields a 30.8° geometric Cobb error, so keypoint accuracy does not transfer to Cobb accuracy (Sections 4.1 and 5). Second, diagnosis requires domain adaptation. VLMs score at most 0.44 on the weighted VQA before fine-tuning (Appendix F) and 0.69–0.76 after fine-tuning on visual observations alone (Table 6). Third, structured spine representations help more than raw geometry. Annotated 2D keypoints bring Qwen3.5-4B/9B to $0.777/0.785$ and depth-lifted 3D keypoints to $0.793/0.783$, whereas raw point clouds leave Qwen3.5 at 0.695–0.697. Because these settings use different image subsets, we treat the differences as descriptive (Section 3.1).

Contributions.

- **SpineDiag**, a paired multimodal clinical AIS dataset from multiple medical centres, pairing visual observations, depth maps, and point clouds with anatomical keypoints, Cobb references, and physician-verified diagnoses.
- A four-benchmark suite from visual spine perception to diagnosis with keypoints or depth, evaluated by perception metrics, geometric Cobb error, a clinically VQA score, and LLM verification.
- Pose and vision–language baselines under subject-level splits, which expose the gap between keypoint and Cobb accuracy and the value of structured spine representations.

2 THE SPINEDIAG DATASET

SpineDiag supports contact-free and radiation-free spine assessment. For each subject, it pairs visual observations, depth maps, and point clouds with anatomical keypoints, Cobb references, and diagnosis descriptions. Table 1 summarises the release and Table 2 its Cobb severity distribution. We describe acquisition (§2.1), annotation (§2.2), and splits (§2.3); licensing and ethics are detailed in Appendices I and J.

Table 1: **SpineDiag release.** 155 subjects across 171 clinical visits; each row states its counting unit.

Data	Count
Keypoint-annotated frames	961
Depth-bearing frames	711
Diagnosis descriptions (visits)	171
Archived X-ray (subjects)	87
Unique subjects	**155**

Table 2: **Cobb severity distribution.** Visit-level counts by Cobb reference.

Status (*Cobb angle*)	Visits	Proportion
Very mild ($< 10^\circ$)	40	23.4%
Mild (10–20°)	46	26.9%
Moderate (20–40°)	47	27.5%
Severe ($> 40^\circ$)	38	22.2%
Total visits	**171**	**100%**

2.1 ACQUISITION PROTOCOL

Subjects aged 8–16 years were eligible for enrolment; individual ages were not recorded. All visits are recorded in a clinic-side imaging room with controlled lighting. The camera stands 2–3 m behind the subject with the shoulder-blade midpoint on the optical axis. Subjects wear examination shorts, and only same-gender clinicians and the guardian are present during each session.

Visual observations Standard-position RGB images in three postures (posterior, lateral, and Adam's forward-bend), and multi-pose RGB videos of five movements (flexion–extension, left/right lateral bending, and left/right axial rotation), each 30–60 s at 30 FPS.

Depth maps Time-of-flight depth from an *Azure Kinect* sensor (NFOV unbinned, 640×576 at 30 FPS), synchronised with its 3840×2160 RGB stream.

Point clouds Body-surface point clouds reconstructed from aligned RGB and depth with the factory intrinsics and a background-removal step released with our code; each torso-cropped frame keeps up to $\sim 200{,}000$ points.

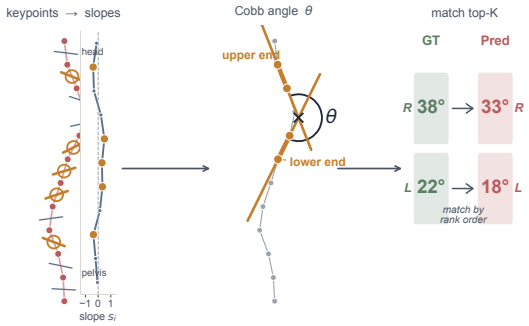


Figure 3: **Geometric Cobb derivation.** Predicted spinal keypoints define slope-change endpoints for a geometric Cobb estimate, which is compared with the available Cobb reference estimate.

Full-spine X-ray (supplementary) Standing posterior–anterior radiographs are archived for $87/155$ subjects whose hospital records were retrievable. We do not annotate them, and we release them only after DICOM de-identification.

Released visual observations are de-identified by mosaicking the head and background, followed by manual verification. **Archived X-ray images and Cobb references are never model inputs** in any benchmark (§3).

2.2 Annotation protocol

2D anatomical keypoints (24 points). We define a 24-point protocol (Figure 2) covering the spine and its biomechanical references: 12 spinal landmarks at C7, T1–T7, and L1–L4 (denoted s_0–s_{11}), 4 shoulder points ($\mathrm{ls}_{0,1}$, $\mathrm{rs}_{0,1}$), 2 inferior-scapular angles (lb_0, rb_0), and 6 lateral-waist points ($\mathrm{lw}_{0..2}$, $\mathrm{rw}_{0..2}$). Reference levels are anchored to palpable / visible anatomy: C7 = most prominent spinous process on neck flexion; T2/T3 = line through superior scapular angles; T7 = line through inferior scapular angles; L4 = line through highest iliac crests; L2 = line through bilateral costal-margin tips. Anchoring annotations to externally palpable structures, rather than imaged vertebrae, makes the protocol reproducible without X-ray. Five professional annotators label the keypoints with a unified interface, followed by review from a spine surgeon. Each frame is annotated independently rather than propagated from neighbouring frames. We further lift the 2D keypoints with depth to a 3D kinematic spine model (Appendix H).

Cobb references. We do not annotate X-ray images. Each Cobb reference is either assessed by a clinician on the post-visit X-ray or computed from visual annotations, so the references have mixed provenance. Visual spine perception compares a geometric Cobb angle derived from predicted spinal keypoints against this reference (§5).

Four-field diagnosis descriptions. Diagnosis descriptions cover *spinal curvature*, *shoulders*, *back*, and *scapulae*, and are evaluated through nine multiple-choice questions: five on spinal curvature, one on shoulders, one on back, and two on scapulae (§3.2). They are extracted from the original medical records with a fixed prompt over Qwen3-32B (Qwen Team, 2025a), and a senior spine surgeon checks and corrects every extracted record. Lenke classification is not provided because the contributing centres did not record it consistently. LLM verification of six-field diagnosis text is reported separately (§4.3).

2.3 Splits and file layout

The release contains 155 subjects across 171 clinical visits, with 961 keypoint-annotated frames (711 of them depth-bearing), diagnosis descriptions for all visits, and archived X-ray for $87/155$ subjects (Table 1). We use **subject-level** 70/10/20 splits with three random seeds, so all records from a subject stay in one partition and frames cannot leak across splits. Files are organised as `subjects/<subject_id>/{rgb/, video/, depth/, pointcloud/,`

`keypoints/, xray/, labels.json}`, where `labels.json` carries the Cobb reference, the Cobb severity stratum (Table 2), the diagnosis description, and X-ray exclusion flags.

3 BENCHMARK PROTOCOL

3.1 BENCHMARK AND SETTING DEFINITIONS

Table 3 lists the inputs, outputs, and metrics of each benchmark and setting. A model may use only the listed inputs, at both training and test time. `Diag-Kpt2D` supplies annotated keypoints as a reference condition, whereas `Diag-PredKpt2D` sees visual observations only and predicts the keypoints in its chain of thought. `Diag-Pointcloud` and `Diag-Kpt3D` require aligned depth at both training and test time and are therefore evaluated on the depth-bearing subset, so score differences across settings are descriptive rather than paired gains. **Archived X-ray images and Cobb references are never model inputs.** All settings use the subject-level splits of §2.3; the seeds and repetitions of each experimental round are given in §4.

Table 3: **SpineDiag benchmarks and settings.** Inputs are available at both training and test time; archived X-ray images and Cobb references are never inputs. LLM verification scores the free-text diagnosis output.

Benchmark	Setting	Input	Output	Metrics
Visual spine perception	—	Visual observation	2D Kpts.	OKS AP; Cobb angle error; endpoint IoU
Visual diagnosis	`Diag`	Visual observation	VQA	Weighted VQA; LLM verification
Visual diagnosis with 2D keypoints	`Diag-Kpt2D`	+ annotated 2D keypoints	VQA	Weighted VQA; LLM verification
	`Diag-PredKpt2D`	+ predicted 2D keypoints	VQA	Weighted VQA; LLM verification; Cobb angle error
Visual diagnosis with depth	`Diag-Pointcloud`	+ point cloud	VQA	Weighted VQA; LLM verification
	`Diag-Kpt3D`	+ depth-lifted 3D keypoints	VQA	Weighted VQA; LLM verification; depth MAE

3.2 METRICS

Keypoint AP (visual spine perception). We score the 24-point predictions with the COCO Object Keypoint Similarity (OKS) protocol (Lin et al., 2014) computed by `pycocotools`. For a subject with $K{=}24$ keypoints, predicted point distances d_i, an annotated visibility flag v_i, a per-keypoint scale σ_i, and a subject area A,

$$\mathrm{OKS} \;=\; \frac{\sum_{i=1}^{K} \exp\!\left(-d_i^2/\left(2A(2\sigma_i)^2\right)\right)\,\delta(v_i>0)}{\sum_{i=1}^{K}\delta(v_i>0)}, \tag{1}$$

and we report AP averaged over OKS thresholds $\{0.50, 0.55, \ldots, 0.95\}$ together with AP@.5 and AP@.75. Because the SpineDiag test cohort is photographed at a fixed standoff distance, all instances fall in the COCO "large" area bin and AP_M is reported as -1.0 throughout.

Weighted diagnosis VQA (visual diagnosis). The diagnosis VQA benchmark has nine multiple-choice questions in four fields: $Q_{1:5}$ concern spinal curvature, Q_6 shoulders, Q_7 back, and $Q_{8:9}$ scapulae (full questions and answer choices in Appendix E). Each answer is scored by exact correctness, $u_{i,q}=1$ if subject i's answer to question q matches the reference and 0 otherwise. Let $D_{i,k}$ be the mean question correctness in field k. The reported per-subject VQA score is

$$\mathrm{VQA}_i \;=\; 0.7D_{i,\mathrm{spine}}{+}0.1D_{i,\mathrm{shoulders}}{+}0.1D_{i,\mathrm{back}}{+}0.1D_{i,\mathrm{scapulae}}, \qquad \mathrm{VQA} \;=\; \frac{1}{N}\sum_{i=1}^{N}\mathrm{VQA}_i, \tag{2}$$

following the four-field weighting in the experimental record. LLM verification, a Qwen3-32B (Qwen Team, 2025a) four-level rubric over six-field diagnosis text, is another evaluation method (Section 4.3); because the outputs and scoring rules differ, its scores are reported separately and are not pooled or directly ranked against VQA scores.

Cobb angle defination. The Cobb angle is the standard radiographic measure used to quantify the magnitude of spinal curvature in scoliosis. It is measured on a standing spinal radiograph by first identifying the upper and lower end vertebrae, defined as the most tilted vertebrae at the cephalad and caudal ends of the scoliotic curve, respectively. Lines are then drawn along the superior endplate of the upper end vertebra and the inferior endplate of the lower end vertebra; the angle formed by these two lines, or equivalently by their perpendiculars, is defined as the Cobb angle (Langensiepen et al., 2013).

The Cobb angle provides a quantitative measure of the magnitude of coronal spinal curvature and is the standard method for assessing scoliosis. It is widely used in clinical practice to characterize curve magnitude and to monitor changes in spinal curvature over time (Langensiepen et al., 2013).

Geometric Cobb consistency (visual spine perception). We approximate the spine by the predicted spinal keypoints, take the perpendicular bisectors of consecutive segments, and treat sign changes of the slope derivative as the start and end of a Cobb segment (Figure 3, left and middle). Following clinical practice, we keep the two largest reference angles above a threshold $\tau \in \{10^\circ, 20^\circ, 30^\circ\}$ and match each predicted segment to its closest reference segment by rank order (Figure 3, right). Two metrics are reported: a *mean angle error* that penalises sign flips of curvature direction with a fixed 90° cost, and a *mean endpoint IoU* between the start–end intervals of matched segments,

$$\Delta\theta \;=\; \begin{cases} |\theta_{\mathrm{gt}} - \theta_{\mathrm{pred}}| & \text{if directions agree,} \\ 90^\circ & \text{otherwise,} \end{cases} \qquad \mathrm{IoU} \;=\; \frac{\max\big(0,\, \min(e_1, e_2) - \max(s_1, s_2)\big)}{\max(e_1, e_2) - \min(s_1, s_2)}, \tag{3}$$

averaged first within a subject and then over the cohort. We compare predicted geometric Cobb with the available reference estimate, which may originate from a clinician's post-visit X-ray assessment or from visual annotations. The pooled error is not an X-ray-only agreement measure; a Bland–Altman analysis (Bland & Altman, 1986) should distinguish the reference source.

Depth MAE (`Diag-Kpt3D`). For settings that supervise predicted point depth (`3d kpt only`, `diagnosis + 3d kpt` variants), we report a per-keypoint mean absolute depth error $\mathrm{MAE}_{\mathrm{depth}} = \frac{1}{K}\sum_{i=1}^{K} |z_i^{\mathrm{gt}} - z_i^{\mathrm{pred}}|$ on the same OKS-aligned matches.

4 Experiments

We evaluate visual spine perception (Section 4.1) and the three visual diagnosis benchmarks with their five input settings (Section 4.2); the geometric Cobb consistency of predicted keypoints is analysed in Section 5. The reported metrics come from two experimental rounds. The current round fine-tunes Qwen3.5-4B/9B, LLaVA-OneVision-7B, and Llama-3.2-11B-Vision and reports the weighted VQA score of every diagnosis setting (Table 6), together with the keypoint AP and geometric Cobb consistency of the keypoints predicted in `Diag-PredKpt2D` (Table 8). An earlier round fine-tunes Qwen3-VL-4B on a fixed subject-level split. It reports LLM verification of six-field diagnosis text, the Cobb angle error and endpoint IoU of keypoints predicted in the chain of thought, and the depth MAE of predicted 3D keypoints (Section 4.3, Table 7). Pose-estimator OKS AP and the X-Pose geometric Cobb error come from separate pose-baseline runs (Section 4.1). Scores from different rounds use different models, splits, or image subsets, so they are reported separately and not pooled. Implementation details are in Appendix B.

4.1 Visual spine perception

Setup. Models predict the 24 anatomical keypoints (Section 2.2) from a posterior visual observation and are scored with OKS AP (Eq. 1). Training and testing use the 961 keypoint-annotated frames under subject-level splits.

Table 4: **Keypoint OKS mAP.** X-Pose: mean±SD over multiple runs; other entries are single reported values. Best in **bold**.

Method	OKS mAP ↑
ViTPose	0.263
HRNet	0.730
ED-Pose	0.418
X-Pose	$\mathbf{0.756 \pm 0.0097}$

Table 5: **Zero-shot evaluation under six-field LLM verification.** 0^* indicates that no geometrically valid 24-point response was returned. Diagnosis scores are not directly comparable with four-field VQA scores in Table 6.

Model (zero-shot)	Kpt AP ↑	LLM verification ↑
GPT-5.5	0^*	0.586
Gemini-3.1-Pro	0^*	**0.659**
Qwen3.5	0^*	0.493

Table 6: **Diagnosis VQA across input settings.** Mean±SD over multiple runs; −: not run. Settings use different image subsets, so columns are not paired comparisons. Best per column in **bold**.

Model	`Diag` (RGB only)	`Diag-Kpt2D` (+annotated 2D kpts)	`Diag-PredKpt2D` (+predicted 2D kpts)	`Diag-Pointcloud` (+point cloud)	`Diag-Kpt3D` (+3D kpts)
Qwen3.5-4B	0.749 ± 0.0153	0.777 ± 0.0166	0.752 ± 0.0202	0.695 ± 0.0127	$\mathbf{0.793 \pm 0.0131}$
Qwen3.5-9B	$\mathbf{0.757 \pm 0.0148}$	$\mathbf{0.785 \pm 0.0166}$	$\mathbf{0.776 \pm 0.0212}$	$\mathbf{0.697 \pm 0.0124}$	0.783 ± 0.0121
LLaVA-OV-7B	0.694 ± 0.0027	0.689 ± 0.0082	0.693 ± 0.0106	–	0.775 ± 0.0064
Llama-3.2-11B-V	0.687 ± 0.0011	0.694 ± 0.0051	0.699 ± 0.0034	–	0.767 ± 0.0093

No pose baseline uses ground-truth subject boxes at inference; X-Pose is additionally trained on a 1:1 SpineDiag+COCO mixture to test whether COCO data still helps under cohort shift.

Results. Table 4 reports the current OKS-mAP summary. X-Pose is reported as mean±standard deviation (SD) over multiple runs, while the experimental record gives only a single mAP value for each of the other three model families; the earlier backbone-specific single-seed table is retained in Appendix C and is not numerically merged with this summary.

4.2 Visual diagnosis

Settings. The three diagnosis benchmarks comprise five settings: `Diag` for visual diagnosis, `Diag-Kpt2D` and `Diag-PredKpt2D` for visual diagnosis with 2D keypoints, and `Diag-Pointcloud` and `Diag-Kpt3D` for visual diagnosis with depth. Every setting answers the nine questions of Appendix E and is scored with the weighted VQA score (Eq. 2). The settings differ only in the model input. `Diag` uses the RGB image only. `Diag-Kpt2D` adds the annotated 24-point coordinates to the prompt, a reference condition that measures how much diagnosis benefits from accurate keypoints. `Diag-PredKpt2D` also uses the RGB image only, but the model predicts the 2D keypoints in its chain of thought before answering. `Diag-Pointcloud` adds the body-surface point cloud reconstructed from aligned depth. `Diag-Kpt3D` adds keypoints lifted with depth to 3D camera coordinates (x, y, z), where z is the camera-coordinate component rather than a raw depth pixel. The two depth settings require aligned depth at both training and test time and are evaluated on the depth-bearing subset.

Results. Table 6 reports all settings. Within the Qwen3.5 family, the larger model helps in most settings (e.g., $0.749 \rightarrow 0.757$ for `Diag` and $0.752 \rightarrow 0.776$ for `Diag-PredKpt2D`), but not with 3D keypoints (0.793 vs. 0.783), so scale alone gives limited gains. Structured keypoints help most. Supplied coordinates outperform predicted ones (`Diag-Kpt2D` $0.777/0.785$ vs. `Diag-PredKpt2D` $0.752/0.776$ for Qwen3.5-4B/9B), and depth-lifted 3D keypoints match or exceed 2D keypoints ($0.793/0.783$). `Diag-Kpt3D` also lifts LLaVA-OV-7B and Llama-3.2-11B-V to 0.775 and 0.767, whereas 2D keypoints change their scores little. In contrast, raw point clouds lower Qwen3.5-4B/9B to $0.695/0.697$, below their RGB-only scores; with only a few hundred depth-bearing frames, the models likely cannot learn to read unstructured 3D tokens. Because settings use different image subsets, we report these differences descriptively rather than as paired gains.

Per-question analysis. Figure 4 breaks the weighted score down by question. Within spinal curvature, models detect a curve (Q1) and kyphosis (Q5) well but struggle to localise it: Q2 (main curve location) and especially Q3 (thoracic curve) are the hardest, with Q3 ranging from 0.40 to

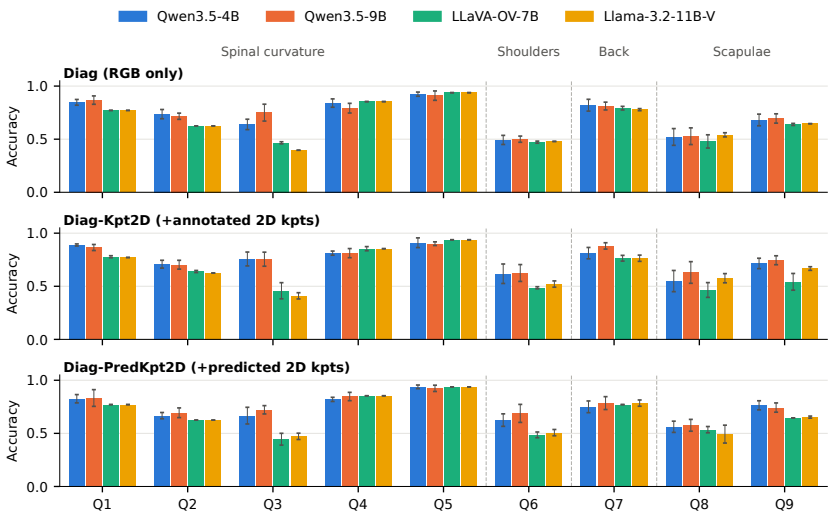


Figure 4: **Per-question diagnosis accuracy** under `Diag`, `Diag-Kpt2D`, and `Diag-PredKpt2D`. Bars show mean±SD over multiple runs; dashed lines separate the four diagnosis fields. Per-question results of `Diag-Kpt3D` and `Diag-Pointcloud` are in Appendix F.

0.75 under `Diag`. For Qwen3.5, the height comparisons Q6 (shoulders) and Q8 (scapulae) score lowest (0.49–0.53 under `Diag`): small asymmetries admit either "level" or "one side higher", and the camera angle can alter apparent heights. Asking whether a back prominence exists (Q7) avoids fine height comparison and scores higher (0.78–0.82). Keypoints mainly help the hard questions: with annotated 2D keypoints, Qwen3.5-4B improves on Q3 from 0.64 to 0.76 and on Q6 from 0.49 to 0.62. In contrast, the accuracies of LLaVA-OV-7B and Llama-3.2-11B-V on Q1, Q2, Q4, and Q5 change by at most 0.014 across the three settings, so keypoints barely change their answers on these questions.

4.3 Diagnosis-text evaluation with six-field LLM verification

Protocol. A Qwen3-VL-4B experiment provides the second diagnosis evaluation method. It generated six-field free-text diagnosis (overall, spine, shoulder, scapula, back, others), scored by Qwen3-32B (Qwen Team, 2025a) on a four-level rubric: correct $= 3$, partially correct $= 2$, incorrect $= 1$, and critical error $= 0$. The six-field score was the sum divided by 18. Its output format, evaluator, model, and experimental split differ from the nine-question VQA evaluation, so their numerical scores are reported separately rather than pooled or directly ranked. This experiment also measured keypoint AP, depth MAE, and geometric-Cobb consistency.

Results. Table 7 reports six settings under the LLM-verification rubric. Within this protocol, 2D-keypoint CoT changes the diagnosis score from 0.712 to 0.757 and the geometric-Cobb error from 23.2° to 6.1°. The 2.9° depth-bearing result is the best fixed-split Cobb value in this analysis, not a VQA score or an X-ray-only agreement result. The point-cloud variant is retained in the table without treating its rubric score as comparable with Table 6.

4.4 Zero-shot comparison under LLM verification

Goal. Evaluate zero-shot keypoint AP and diagnosis using the six-field LLM-verification rubric. These values do not use the four-field VQA metric and must not be ranked against Table 6. Depth-bearing settings were not evaluated in this zero-shot study.

Findings. Under LLM verification (Table 5), zero-shot keypoint AP is 0 for the three reported VLMs, while diagnosis scores range from 0.493 to 0.659. These are diagnosis-text rubric results, not weighted VQA scores.

Table 7: **Six-field diagnosis evaluation with LLM verification (Qwen3-32B).** The diagnosis column uses a four-level rubric (sum over six fields divided by 18), rather than the nine-question VQA metric. † kpt AP for joint variants is measured on the *diagnosis_2d* subset and is not directly comparable to the *2d kpt only* (961-frame) row. Reported values are preserved.

Data Setting	Train Data	Test Task	kpt AP ↑	Depth MAE ↓	diag ↑	Cobb (deg, IoU)
2d kpt only	2d kpt	2d kpt	0.728	—	—	23.216, 0.499
diag only	diagnosis	diagnosis	—	—	0.712	—
diag + pc	diag ∩ depth	diagnosis	—	—	0.733	—
diag + 2d kpt	diag (kpt CoT)	diag (kpt CoT)	0.931†	—	0.757	6.110, 0.735
3d kpt only	2d kpt ∩ depth	same	0.668	144.959	—	—
diag + 3d kpt	diag ∩ depth (3d kpt CoT)	diag (3d kpt CoT)	**0.977**†	**16.214**	**0.785**	**2.870, 0.848**

Table 8: **Geometric Cobb consistency (top-2 segments, 10° threshold).** Lower angle error and higher endpoint IoU are better; reference Cobb values have mixed provenance. Current-round values are mean±SD over multiple runs; earlier-round values come from a single fixed split (Table 7). Best per column within each round in **bold**.

Model	Keypoint source	Kpt AP ↑	Angle error (deg) ↓	Endpoint IoU ↑
Current round				
X-Pose	pose estimator	0.756 ± 0.010	30.8 ± 1.8	0.497 ± 0.014
Qwen3.5-4B	`Diag-PredKpt2D` CoT	$\mathbf{0.830 \pm 0.020}$	21.2 ± 3.2	$\mathbf{0.534 \pm 0.030}$
Qwen3.5-9B	`Diag-PredKpt2D` CoT	0.776 ± 0.021	$\mathbf{20.0 \pm 4.7}$	0.482 ± 0.051
LLaVA-OV-7B	`Diag-PredKpt2D` CoT	0.018 ± 0.015	59.2 ± 4.4	0.373 ± 0.040
Llama-3.2-11B-V	`Diag-PredKpt2D` CoT	0.019 ± 0.021	60.1 ± 1.6	0.404 ± 0.016
Earlier round (Qwen3-VL-4B, fixed split)				
2d kpt only		—	23.216	0.499
diag + 2d kpt		—	6.110	0.735
diag + 3d kpt		—	**2.870**	**0.848**

5 Geometric Cobb consistency

Pipeline. We apply the geometric Cobb pipeline and metrics of Section 3.2 (Figure 3) with the top-2 reference segments above a 10° threshold; the reference may come from clinician assessment of post-visit X-ray or from visual annotations.

Results. In the current round (Table 8, top), keypoints predicted in the `Diag-PredKpt2D` chain of thought by Qwen3.5-4B reach 0.830 keypoint AP and a 21.2° Cobb angle error, and Qwen3.5-9B reaches 20.0°; both are lower than the 30.8° of X-Pose but remain far from clinical precision. X-Pose reaches 0.756 OKS mAP yet a 30.8° Cobb error, so keypoint accuracy alone does not guarantee Cobb accuracy. LLaVA-OV-7B and Llama-3.2-11B-V rarely produce valid keypoints (AP below 0.02) and reach about 60°. The earlier round (Table 8, bottom) is reported at the 10° threshold. The 2D-keypoint CoT setting records 6.1° angle error and 0.735 endpoint IoU, versus 23.2° and 0.499 for keypoints alone. The 2.9° and 0.848 depth-bearing result is the lowest value in this earlier fixed split, not a result from the current four-field VQA metric or proof of X-ray-only agreement. A provenance-stratified Bland–Altman analysis (Bland & Altman, 1986) would be needed before asserting agreement with X-ray-based Cobb estimates.

6 Conclusion

We present SpineDiag, a paired multimodal clinical dataset and benchmark for contact-free and radiation-free spine assessment. Its four benchmarks span visual spine perception and diagnosis with keypoints or depth. Our baselines show that current pose estimators do not yet recover accurate Cobb angles from pathological backs, and that VLMs diagnose better with structured spine representations than with raw geometry. Matched-subset evaluation, clinician validation, and larger cohorts are needed before any clinical use.

AI USE STATEMENT

In this work, we used generative AI tools for cleaning and reformatting the dataset: diagnosis records were extracted from the original electronic medical records by Qwen3-32B (Qwen Team, 2025a) with a fixed prompt, and every extracted record was checked and, when necessary, corrected by a senior spine surgeon (Section 2.2). Qwen3-32B was also used as the LLM verifier for the separately reported six-field diagnosis-text evaluation (Section 4.3); its scores are distinct from the nine-question VQA metric. Among the remaining required-disclosure tasks, generative AI was used only for translation of author-written notes into English; it was not used to generate synthetic data, theoretical frameworks, hypotheses, methodology or experiment designs, method implementations, qualitative analyses, or result interpretations. Additionally, we used generative AI tools for editing for readability, literature search, reference formatting, and LaTeX formatting. We have reviewed all AI-assisted work: every reported number was checked against the experimental records, every added reference was verified through CrossRef or arXiv, and all edited text was reviewed by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves pediatric human subjects and sensitive clinical data. The study was approved by the host institution's research ethics committee, and written informed consent was obtained from both the legal guardian and the minor before any imaging session; the consent covers anonymised non-commercial research use under a data use agreement (DUA) and the right to request data removal before publication (Appendix J). Original RGB and archived X-ray images are available only under a DUA; DICOM headers, burned-in pixel PHI, and free-text clinical records are de-identified before release (Appendix J). The dataset is licensed for non-commercial research use (Appendix I). The cohort comprises pediatric AIS referrals aged 8–16 years by inclusion criterion, without recorded individual ages, and does not establish generalisation to other populations; SpineDiag models should not be used for standalone diagnosis, commercial screening, or individual-level medical decisions without prospective validation and clinician oversight (Appendix G).

REPRODUCIBILITY STATEMENT

The subject-level partitioning and the three shuffles with three repeated VQA runs per shuffle are described in Sections 2.3 and 4. The weighted VQA metric and the LLM-verification rubric are defined separately in Sections 3.2 and 4.3; the geometric Cobb pipeline is detailed in Section 5. Qwen3-VL and specialist training configurations are given in Section B. DUA-controlled corpus access is described in Appendix I, and a project page with data examples is provided in the supplementary material.

REFERENCES

Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8726–8737, 2023.

J. Martin Bland and Douglas G. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476):307–310, 1986. doi: 10.1016/S0140-6736(86)90837-8.

Alexander G. Bruno, Mary L. Bouxsein, and Dennis E. Anderson. Development and validation of a musculoskeletal model of the fully articulated thoracolumbar spine and rib cage. *Journal of Biomechanical Engineering*, 137(8):081003, 2015. doi: 10.1115/1.4030408.

Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. HuMMan: Multi-modal 4D human dataset for versatile sensing and modeling. In *European Conference on Computer Vision (ECCV)*, pp. 557–577, 2022. doi: 10.1007/978-3-031-20071-7_33.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Scott L. Delp, Frank C. Anderson, Allison S. Arnold, Peter Loan, Ayman Habib, Chand T. John, Eran Guendelman, and Darryl G. Thelen. Opensim: open-source software to create and analyze dynamic simulations of movement. *IEEE Transactions on Biomedical Engineering*, 54(11):1940–1950, 2007. doi: 10.1109/TBME.2007.901024.

Yang Deng, Ce Wang, Yuan Hui, Qian Li, Jun Li, Shiwei Luo, Mengke Sun, Quan Quan, Shuxin Yang, You Hao, Pengbo Liu, Honghu Xiao, Chunpeng Zhao, Xinbao Wu, and S. Kevin Zhou. CTSpine1K: A large-scale dataset for spinal vertebrae segmentation in computed tomography. *arXiv preprint arXiv:2105.14711*, 2021. doi: 10.48550/arXiv.2105.14711.

Ben Glocker, Darko Zikic, Ender Konukoglu, David R. Haynor, and Antonio Criminisi. Vertebrae localization in pathological spine CT via dense classification from sparse annotations. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, volume 8150 of *Lecture Notes in Computer Science*, pp. 262–270, Berlin, Heidelberg, 2013. Springer. doi: 10.1007/978-3-642-40763-5_33.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14783–14794, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Mostafa Hassan, Jose Maria Gonzalez Ruiz, Nada Mohamed, Thomaz Nogueira Burke, Qipei Mei, and Lindsey Westover. Ensemble learning of deep CNN models and two stage level prediction of Cobb angle on surface topography in adolescents with idiopathic scoliosis. *Medical Engineering & Physics*, 140:104332, 2025. doi: 10.1016/j.medengphy.2025.104332.

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 (ICLR)*, 2022.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. doi: 10.1109/TPAMI.2013.248.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *Proc. ECCV*, 2024.

Markus Rafael Konieczny, Hüsseyin Senyurt, and Rüdiger Krauspe. Epidemiology of adolescent idiopathic scoliosis. *Journal of Children's Orthopaedics*, 7(1):3–9, 2013. doi: 10.1007/s11832-012-0457-4.

Marta Kinga Łabecka and Magdalena Plandowska. Moiré topography as a screening and diagnostic tool—a systematic review. *PLOS ONE*, 16(12):e0260858, 2021. doi: 10.1371/journal.pone.0260858.

S. Langensiepen, O. Semler, R. Sobottke, O. Fricke, J. Franklin, E. Schönau, and P. Eysel. Measuring procedures to determine the cobb angle in idiopathic scoliosis: a systematic review. *European Spine Journal*, 22(11):2360–2371, 2013. doi: 10.1007/s00586-013-2693-9.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. doi: 10.1145/2816795.2818013.

Martina Minotti, Stefano Negrini, Andrea Cina, Fabio Galbusera, Fabio Zaina, and Tito Bassani. Deep learning prediction of curve severity from rasterstereographic back images in adolescent idiopathic scoliosis. *European Spine Journal*, 33(11):4164–4170, 2023. doi: 10.1007/s00586-023-08052-1.

Shumao Pang, Chunlan Pang, Lei Zhao, Yangfan Chen, Zhihai Su, Yujia Zhou, Meiyan Huang, Wei Yang, Hai Lu, and Qianjin Feng. SpineParseNet: Spine parsing for volumetric MR image by a two-stage segmentation framework with semantic image representation. *IEEE Transactions on Medical Imaging*, 40(1):262–273, 2021. doi: 10.1109/TMI.2020.3025087.

Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13468–13478, 2021.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025b.

Qwen Team. Qwen3.5 model card. `https://huggingface.co/Qwen/Qwen3.5-4B`, 2026.

Anjany Sekuboyina, Malek E. Husseini, Amirhossein Bayat, Maximilian Löffler, Hans Liebl, Hongwei Li, Giles Tetteh, Jan Kukačka, Christian Payer, Darko Štern, Martin Urschler, Maodong Chen, Dalong Cheng, Nikolas Lessmann, Yujin Hu, Tianfu Wang, Dong Yang, Daguang Xu, Markus Rempfler, Björn H. Menze, and Jan S. Kirschke. VerSe: A vertebrae labelling and segmentation benchmark for multi-detector CT images. *Medical Image Analysis*, 73:102166, 2021. doi: 10.1016/j.media.2021.102166.

Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3D human pose and shape estimation in the wild. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2020.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5693–5703, 2019.

Muhammad Saif Ullah Khan and Didier Stricker. SIMSPINE: A biomechanics-aware simulation framework for 3D spine motion annotation and benchmarking. *arXiv preprint arXiv:2602.20792*, 2026.

Muhammad Saif Ullah Khan, Stephan Krauß, and Didier Stricker. Towards unconstrained 2D pose estimation of the human spine. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 6162–6171, 2025. doi: 10.1109/cvprw67362.2025.00614.

Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conference on Computer Vision (ECCV)*, pp. 614–631, 2018. doi: 10.1007/978-3-030-01249-6_37.

Liansheng Wang, Cong Xie, Yi Lin, Hong-Yu Zhou, Kailin Chen, Dalong Cheng, Florian Dubost, Benjamin Collery, Bidur Khanal, Bishesh Khanal, Rong Tao, Shangliang Xu, Upasana Upadhyay Bharadwaj, Zhusi Zhong, Jie Li, Shuxin Wang, and Shuo Li. Evaluation and comparison of accurate automated spinal curvature estimation algorithms with spinal anterior-posterior X-ray images: The AASCE2019 challenge. *Medical Image Analysis*, 72:102115, 2022. doi: 10.1016/j.media.2021.102115.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Hongbo Wu, Chris Bailey, Parham Rasoulinejad, and Shuo Li. Automatic landmark estimation for adolescent idiopathic scoliosis assessment using BoostNet. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*, Lecture Notes in Computer Science, pp. 127–135. Springer International Publishing, 2017. doi: 10.1007/978-3-319-66182-7_15.

Hongbo Wu, Chris Bailey, Parham Rasoulinejad, and Shuo Li. Automated comprehensive adolescent idiopathic scoliosis assessment using MVC-Net. *Medical Image Analysis*, 48:1–11, 2018. doi: 10.1016/j.media.2018.05.005.

Yan Xia, Xiaowei Zhou, Etienne Vouga, Qixing Huang, and Georgios Pavlakos. Reconstructing humans with a biomechanically accurate skeleton. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. Oral.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Jie Yang, Ailing Zeng, Shilong Liu, Feng Li, Ruimao Zhang, and Lei Zhang. Explicit box detection unifies end-to-end multi-person pose estimation. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Jie Yang, Ailing Zeng, Ruimao Zhang, and Lei Zhang. X-Pose: Detecting any keypoints. In *European Conference on Computer Vision (ECCV)*, 2024.

Junlin Yang, Kai Zhang, Hengwei Fan, Zifang Huang, Yifan Xiang, Jingfan Yang, Lin He, Lei Zhang, Yahan Yang, Ruiyang Li, Yi Zhu, Chuan Chen, Fan Liu, Haoqing Yang, Yaolong Deng, Weiqing Tan, Nali Deng, Xuexiang Yu, Xiaoling Xuan, Xiaofeng Xie, Xiyang Liu, and Haotian Lin. Development and validation of deep learning algorithms for scoliosis screening using back images. *Communications Biology*, 2(1):390, 2019. doi: 10.1038/s42003-019-0635-8.

Yifan Yao, Wenjun Yu, Yongbin Gao, Jiuqing Dong, Qiangqiang Xiao, Bo Huang, and Zhicai Shi. W-Transformer: Accurate Cobb angles estimation by using a transformer-based hybrid structure. *Medical Physics*, 49(5):3246–3262, 2022. doi: 10.1002/mp.15561.

Yong-Ping Zheng, Timothy Tin-Yan Lee, Kelly Ka-Lee Lai, Benjamin Hon-Kei Yip, Guang-Quan Zhou, Wei-Wei Jiang, James Chung-Wai Cheung, Man-Sang Wong, Bobby King-Wah Ng, Jack Chun-Yiu Cheng, and Tsz-Ping Lam. A reliability and validity study for Scolioscan: a radiation-free scoliosis assessment system using 3D ultrasound imaging. *Scoliosis and Spinal Disorders*, 11(1):13, 2016. doi: 10.1186/s13013-016-0074-y.

A Related work

Clinical AIS surface vs. X-ray pairing. Clinical AIS screening from non-radiographic surface signals has matured along three threads. (i) Bare-back photograph CNNs pioneered by Yang et al. (2019) reach specialist-level severity grading on 2D RGB alone. (ii) Radiation-free 3D ultrasound (Zheng et al., 2016) and surface topography or Moiré rasterstereography (Łabecka & Plandowska, 2021; Minotti et al., 2023; Hassan et al., 2025) regress Cobb angle from depth-like body-surface measurements, but each is single-modality and rarely paired with X-ray on the same subject at scale. (iii) X-ray-side CNN cascades (Wu et al., 2018; Yao et al., 2022) treat radiographs as the ground-truth pipeline. SpineDiag pairs the RGB images, videos, depth maps, and point clouds of each subject with Cobb reference estimates from clinician post-visit X-ray assessment or visual annotations, alongside diagnosis labels (and additionally archives standing posterior–anterior X-ray as a supplementary modality for a subset of subjects), enabling a controlled comparison of these previously siloed input modalities under one protocol.

Public spine imaging datasets and benchmarks. Public spine imaging benchmarks have driven much of the recent progress in automated vertebra analysis, but they are uniformly single-modality and rarely subject-aligned across modalities. SpineWeb-era resources such as the pathological-CT collection of Glocker et al. (2013), the multi-detector CT benchmark VerSe (Sekuboyina et al., 2021), and the large-scale CTSpine1K (Deng et al., 2021) target vertebra localisation and segmentation in CT, while SpineParseNet (Pang et al., 2021) provides MR vertebra and disc parsing. For AIS specifically, BoostNet (Wu et al., 2017) and the AASCE2019 challenge (Wang et al., 2022) supply spinal AP X-rays with landmarks and Cobb-angle ground truth. None of these resources include surface modalities (RGB, depth, or point cloud), and none pair comparable Cobb reference estimates with surface imaging on the same subject. SpineDiag contributes per-subject multimodal data paired with Cobb references of mixed provenance, enabling radiation-free screening benchmarks that prior corpora cannot support.

Computer-vision human-shape datasets and recovery. Computer-vision research on human shape and pose has been driven by datasets that overwhelmingly capture healthy, often athletic adults moving in controlled studios or natural scenes. Studio-grade benchmarks such as Human3.6M (Ionescu et al., 2014) and HuMMan (Cai et al., 2022), in-the-wild captures like 3DPW (von Marcard et al., 2018) and MS-COCO (Lin et al., 2014), and large-scale synthetic corpora AGORA (Patel et al., 2021), BEDLAM (Black et al., 2023) and SSP-3D (Sengupta et al., 2020) all build their ground truth on the SMPL body prior (Loper et al., 2015) (with biomechanical extensions like HSMR (Xia et al., 2025)), which is fitted to non-pathological anatomy. Modern recovery models such as 4DHumans (Goel et al., 2023) inherit this distribution. Crucially, none provide comparable Cobb references paired with surface imaging, so adolescent idiopathic scoliosis remains an out-of-distribution regime for these models.

Surface spine keypoint datasets. Closest to our keypoint benchmark, SpineTrack (Ullah Khan et al., 2025) annotates nine spinal keypoints on real-world and synthetic RGB images for unconstrained 2D spine pose estimation, and SIMSPINE (Ullah Khan & Stricker, 2026) augments indoor multi-camera motion capture with vertebra-level 3D spinal keypoints derived from musculoskeletal modelling. Both target general human motion and provide no clinical labels such as Cobb angles or diagnoses, whereas SpineDiag annotates pathological backs and pairs them with clinical references.

B Implementation details

The six-field LLM-verification experiment uses Qwen3-VL-4B-Instruct (Qwen Team, 2025b) fine-tuned with the Qwen-VL-Series-Finetune pipeline using LoRA (Hu et al., 2022) (rank r=32, α=64, dropout 0.05) on q/k/v/o/gate/up/down_proj; vision-tower adapters are enabled while lm_head and embed_tokens are kept frozen. Training uses four GPUs with per-device batch 1 and gradient accumulation 4. We force the SDPA attention kernel because FlashAttention-2 (Dao et al., 2022) produced CUDA errors on Qwen3-VL. Supervision uses chain-of-thought outputs (Wei et al., 2022) (`<think>...</think>` followed by a JSON block). These implementation details do not describe the separate Qwen3.5/LLaVA/Llama diagnosis-VQA runs in Table 6. Specialist baselines—

Table 9: **Comparison of SpineDiag against representative AIS clinical, public spine imaging, and computer-vision human-shape datasets.** "Per-subj. multimodal" = the same enrolled subject contributes surface modalities linked to reported reference labels; "Reference labels" lists the reported target labels. SpineDiag's Cobb references come from clinician post-visit X-ray assessment or visual annotations; standing PA X-ray images themselves are additionally archived as a supplementary modality for $87/155$ subjects (* in the table).

Dataset / Work	Modality	Reference labels	AIS-specific	Per-subj. multimodal	Public
AIS clinical surface vs. X-ray					
Yang et al. (2019)	Back RGB	X-ray	✓	×	×
Scolioscan (Zheng et al., 2016)	3D ultrasound	US-Cobb	✓	×	×
Moiré topo. (Łabecka & Plandowska, 2021)	Structured light	—	✓	×	×
Rasterstereo. (Minotti et al., 2023; Hassan et al., 2025)	Surface topo.	X-ray Cobb	✓	×	×
Public spine imaging benchmarks					
SpineWeb (Glocker et al., 2013)	X-ray / CT	Vertebra labels	partial	×	✓
AASCE2019 (Wang et al., 2022)	AP X-ray	Cobb	✓	×	✓
VerSe (Sekuboyina et al., 2021)	CT	Vertebra seg.	×	×	✓
CTSpine1K (Deng et al., 2021)	CT	Vertebra seg.	×	×	✓
MRSpineSeg (Pang et al., 2021)	MR	Vertebra/disc seg.	×	×	✓
CV human-shape / pose / SMPL					
Human3.6M (Ionescu et al., 2014)	RGB / video	3D pose (SMPL)	×	×	✓
3DPW (von Marcard et al., 2018)	RGB + IMU	SMPL	×	partial	✓
AGORA (Patel et al., 2021)	Synthetic RGB	SMPL-X	×	partial	✓
BEDLAM (Black et al., 2023)	Synth. RGB / video	SMPL-X	×	partial	✓
SSP-3D (Sengupta et al., 2020)	Synthetic RGB	SMPL	×	×	✓
HuMMan (Cai et al., 2022)	RGB-D + IMU + LiDAR	SMPL	×	✓	✓
Surface spine keypoint datasets					
SpineTrack (Ullah Khan et al., 2025)	RGB (real + synthetic)	2D spine keypoints	×	×	✓
SIMSPINE (Ullah Khan & Stricker, 2026)	Multi-view RGB	3D spine keypoints	×	×	✓
SpineDiag (ours)	**RGB+video+depth+PC+kpt+X-ray***	**Cobb ref.+Dx (4 fields)**	✓	✓	✓

Table 10: **Earlier single-seed RGB→2D-keypoint baselines.** Original backbone-specific values; not pooled with the current summary. Best per column in **bold**.

Method	Backbone	Pretrained on	Train set	AP	AP@.5	AP@.75	AP_M	AP_L
ViTPose	ViT-Base	COCO+AIC+MPII	SpineDiag	0.259	0.651	0.258	−1.0	0.259
ViTPose	ViT-Large	COCO+AIC+MPII	SpineDiag	0.263	0.795	0.050	−1.0	0.263
HRNet	w32	COCO 256×192	SpineDiag	0.730	**1.000**	0.815	−1.0	0.730
HRNet	w48	COCO 256×192	SpineDiag	**0.754**	**1.000**	0.836	−1.0	**0.754**
ED-Pose	ResNet-50	COCO	SpineDiag	0.418	0.677	0.562	−1.0	0.418
X-Pose	Swin-T	UniPose	SpineDiag	0.739	**1.000**	**0.938**	−1.0	0.739
X-Pose	Swin-T	UniPose	SpineDiag+COCO	0.735	**1.000**	0.876	−1.0	0.735
Qwen3-VL-4B (LoRA, CoT)	—	Qwen3-VL-4B-Instruct	SpineDiag (`2d kpt`)	0.728	**1.000**	0.873	−1.0	0.728

HRNet (Sun et al., 2019), ViTPose (Xu et al., 2022), ED-Pose (Yang et al., 2023) and X-Pose (Yang et al., 2024)—follow their reference recipes and use no ground-truth subject boxes at inference.

C Earlier keypoint baseline details

The following backbone-specific, single-seed AP results were reported before the current model-family mAP summary in Table 4. The experimental record does not establish a row-by-row mapping between those earlier backbones and the current aggregate entries, so we preserve rather than overwrite the original values.

The same multi-run X-Pose record reports geometric-Cobb mean angle error $(30.818 \pm 1.8245)^\circ$ and endpoint IoU 0.497 ± 0.0137, with $\pm$ denoting SD; these are separate from the earlier X-Pose single-seed AP rows above.

D Subject detection sub-task

As a by-product of visual spine perception, end-to-end keypoint detectors also produce subject bounding boxes. We score them with the standard single-class COCO box-AP protocol (Lin et al., 2014) on the same `2d kpt` test split, reporting AP, AP@.5, AP@.75 and recall $AR_{1,10,100}$. Table 11 reports detection AP and AR.

E DIAGNOSIS VQA QUESTIONS

All visual diagnosis settings ask the nine multiple-choice questions below. The prompt instructs the model to answer from visible evidence only, to use the patient's anatomical left and right, and to return one answer letter per question in a fixed JSON format.

Spinal curvature.

Q1. Is there any obvious spinal curvature? (A) Yes; (B) No.

Q2. Where is the main spinal curve located? (A) Thoracic; (B) Lumbar; (C) Thoracolumbar; (D) None.

Q3. Does the thoracic spine have a lateral curve? (A) No; (B) Yes, curving to the left; (C) Yes, curving to the right.

Q4. Does the lumbar spine have a lateral curve? (A) No; (B) Yes, curving to the left; (C) Yes, curving to the right.

Q5. Is there any abnormal kyphosis? (A) Yes; (B) No.

Shoulders.

Q6. Are the shoulders at the same height? (A) Yes; (B) No, the right shoulder is higher than the left; (C) No, the left shoulder is higher than the right; (D) Unable to observe.

Back.

Q7. Is there any back prominence? (A) No; (B) Yes, left-sided; (C) Yes, right-sided; (D) Yes, bilateral.

Scapulae.

Q8. Are the scapulae at the same height? (A) Yes; (B) No, the left scapula is higher than the right; (C) No, the right scapula is higher than the left; (D) Unable to observe.

Q9. Is there any scapular prominence? (A) No; (B) Yes, left; (C) Yes, right; (D) Yes, bilateral; (E) Unable to observe.

F ADDITIONAL FOUR-FIELD VQA SETTINGS

The current experimental record also evaluates two settings outside Table 6: `ZeroShotDiag` before fine-tuning, and `Diag-Divided` with separate conversations for the four fields. The fine-tuned entries report mean±SD over multiple runs, the data designs can contain different images, so differences across columns are not paired effects.

Answering all nine questions in one conversation outperforms separate per-field conversations for Qwen3.5-4B and Qwen3.5-9B ($0.749/0.757$ vs. $0.726/0.746$), suggesting that the fields share visual evidence. The weaker models show the opposite trend (LLaVA-OV-7B 0.707 vs. 0.694; Llama-3.2-11B-V 0.692 vs. 0.687; Qwen3.5-2B 0.722 vs. 0.697), consistent with their difficulty in producing a full reasoning trace.

Table 11: **Subject detection on SpineDiag** (single-class COCO protocol). Best per column in **bold**.

Method	AP ↑	AP@.5	AP@.75	AP_M	AP_L	AR_1	AR_{10}	AR_{100}
ED-Pose	0.666	0.769	0.733	−1.0	0.666	**0.900**	**0.911**	0.922
X-Pose	**0.784**	0.954	**0.841**	−1.0	**0.784**	0.888	0.908	**0.923**
Qwen3-VL-4B (LoRA)	0.741	**0.972**	0.733	−1.0	0.741	0.830	0.830	0.830

Table 12: **Additional four-field VQA results.** Values are transcribed from the experimental record. These columns complement, rather than replace, the results in Table 6. Best per column in **bold**.

Model	Zero-shot	Divided
Qwen3.5-2B	0.334	0.722 ± 0.0314
Qwen3.5-4B	0.334	0.726 ± 0.0282
Qwen3.5-9B	0.358	$\mathbf{0.746 \pm 0.0206}$
LLaVA-OneVision-7B	**0.437**	0.707 ± 0.0082
Llama-3.2-11B-Vision	0.000	0.692 ± 0.0041

Table 13: **Per-question accuracy of the depth-bearing settings.** Mean over multiple runs on the depth-bearing subset.

Model	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
`Diag-Kpt3D`									
Qwen3.5-4B	0.961	0.905	0.674	0.921	0.929	0.389	0.802	0.476	0.714
Qwen3.5-9B	0.937	0.889	0.651	0.953	0.929	0.286	0.818	0.452	0.794
LLaVA-OV-7B	0.929	0.881	0.659	0.953	0.929	0.342	0.738	0.365	0.809
Llama-3.2-11B-V	0.929	0.881	0.611	0.953	0.929	0.310	0.706	0.452	0.809
`Diag-Pointcloud`									
Qwen3.5-4B	1.000	0.667	0.648	0.667	1.000	0.185	0.833	0.056	0.667
Qwen3.5-9B	1.000	0.667	0.667	0.667	1.000	0.333	0.593	0.222	0.667

The experimental record additionally lists tool-assisted Codex scores of 0.531 (GPT-5.6-Sol) and 0.644 (GPT-6-Astra) using all available data. Because this setting differs in data access and tool use, we do not rank these two values against Table 6. Qwen3.5-2B is excluded from the main-text results because it did not reliably emit the requested reasoning trace after fine-tuning. For completeness, its weighted VQA scores are 0.697 ± 0.0304 (`Diag`), 0.646 ± 0.0522 (`Diag-Kpt2D`), 0.713 ± 0.0096 (`Diag-PredKpt2D`), 0.697 ± 0.0147 (`Diag-Pointcloud`), and 0.718 ± 0.0607 (`Diag-Kpt3D`).

Table 13 lists the per-question accuracies of the two depth-bearing settings, which are evaluated on the depth-bearing subset and are therefore not drawn in Figure 4.

G Limitations and open problems

Cohort scale and external validity. SpineDiag is still a modest clinical cohort: the current release contains $N=155$ subjects, with continued enrolment. The data come from a multi-centre pediatric AIS collaboration with an 8–16-year inclusion range, but individual ages were not recorded, so we cannot report the age distribution. The present population does not establish generalisation to adults, non-AIS spinal deformities, other countries, or other clinical systems. Table 2 reports primary-curve Cobb-angle stratification, but the cohort remains too small to establish subgroup generalisation. We therefore treat SpineDiag as a benchmark for radiation-free AIS research, not as evidence of population-wide screening readiness.

Access and privacy. The strongest privacy risks come from pediatric back RGB, X-ray images, and free-text clinical records. The RGB / X-ray-bearing corpus therefore requires controlled access under a data-use agreement. This access model protects participants but makes the benchmark less frictionless than fully public vision datasets. It also means that claims depending on raw RGB, archived X-ray, or original clinical text must be checked under the same reviewer-safe access pathway described in the dataset card.

Evaluation maturity. The four-field VQA tables report mean$\pm$SD over multiple runs, but the image subsets differ across input designs, so cross-column differences are not paired estimates. These run-level SDs do not substitute for uncertainty across independent clinical cohorts; released split IDs and per-question outputs are needed for audit. Specialist-keypoint and six-field LLM-verification re-

sults retain their single-seed or fixed-split limitations and cannot be merged with the VQA table. Blinded surgeon evaluation and a diagnosis-label audit remain important before clinical interpretation.

Modeling limits. Cobb reference estimates have two possible sources; pooled angle error is not an X-ray-only agreement measure. The earlier geometric-Cobb results are fragile across subject splits, and their best depth-bearing fixed-split value does not establish generalisation or RGB-only clinical accuracy. `Diag-Pointcloud` and `Diag-Kpt3D` require depth at both training and test time and cover only the depth-bearing subset. Controlled multimodal fusion and uncertainty calibration remain open problems before any model can be considered for clinical triage.

Training-recipe negatives. Reinforcement-learning post-training with GRPO (Shao et al., 2024) did not improve over plain supervised fine-tuning in our setting—OKS regressed and per-step time grew prohibitive on Qwen3-VL—so we retain SFT as the main recipe and leave reward-shaping investigations to future work. More broadly, SpineDiag models should not be used for standalone diagnosis, commercial screening, or individual-level medical decisions without prospective validation and clinician oversight.

H From 2D keypoints to a 3D kinematic spine model

Annotated or tracked 2D keypoints are combined with the aligned depth map of each frame to reconstruct 3D keypoints. These 3D keypoints drive an OpenSim spinal model (Delp et al., 2007), where inverse kinematics (IK) is solved to estimate the spinal pose and morphology at each frame; on video, keypoints are tracked across frames with CoTracker (Karaev et al., 2024). The IK problem described above was solved using the Thoracolumbar Spine and Rib Cage Model (Bruno et al., 2015) in OpenSim. For the present analysis, only the thoracic vertebrae, lumbar vertebrae, and sacrum were retained from the original model. Here, T denotes a thoracic vertebra and L denotes a lumbar vertebra, with the following number indicating the vertebral level (e.g., T1 denotes the first thoracic vertebra and L1 denotes the first lumbar vertebra). This mapping retargets the 12 spinal keypoints onto the vertebrae of the model for inverse kinematics only; it does not redefine the annotated levels in Section 2.2. Because the model has no C7 vertebra, s0 is mapped to T1. The spinal keypoints were retargeted as follows: T1 $\leftarrow$ s0, T2 $\leftarrow$ s1, T4 $\leftarrow$ s2, T6 $\leftarrow$ s3, T8 $\leftarrow$ s4, T10 $\leftarrow$ s5, T12 $\leftarrow$ s6, L1 $\leftarrow$ s7, L2 $\leftarrow$ s8, L3 $\leftarrow$ s9, L4 $\leftarrow$ s10, and L5 $\leftarrow$ s11. The resulting vertebral correspondences were then used as the target positions for solving the IK problem.

I License, hosting and maintenance

SpineDiag is licensed under *CC BY-NC 4.0* for non-commercial research use, with access to the complete sensitive corpus governed by a *Data Use Agreement (DUA)*. We commit to maintaining the dataset (issue tracking and at least one annotation correction round) for two years post-publication, with a contact e-mail listed on the dataset card.

J Ethics

The study was approved by the host institution's research ethics committee; the approval is valid through 2027-02-03 with annual renewal, and the committee name and approval number are withheld for double-blind review. Written informed consent was obtained from both the legal guardian and the minor before any imaging session; consent forms include explicit clauses on (i) anonymised release for non-commercial research, (ii) third-party academic re-use under DUA, and (iii) the right to request data removal at any time before publication. Because pediatric back imagery and clinical records are sensitive, the corpus is available only under DUA-controlled access. All DICOM headers are exhaustively scrubbed before release using a custom `pydicom`-based pipeline that removes every PHI tag listed in the DICOM *Confidentiality Profile* (Basic Profile). Burned-in PHI in pixel data is checked manually and masked when present. Free-text EHR snippets used for diagnosis extraction are scrubbed by an additional regex-based PHI redactor. The Ethics statement summarises this paragraph.