

ScintiGround-38K: A Grounded Visual Question Answering Benchmark for Paired Anterior–Posterior Bone Scintigraphy

M. Rifqi Dzaky Azhad

Telkom University <https://orcid.org/0009-0009-8448-9678>

Ema Rachmawati

emarachmawati@telkomuniversity.ac.id

Telkom University <https://orcid.org/0000-0003-3693-7228>

Research Article

Keywords: Bone scintigraphy, Visual question answering, Visual grounding, Medical imaging informatics, Benchmark dataset, Data provenance

Posted Date: August 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10672263/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Whole-body bone scintigraphy (WBBS) acquires paired anterior and posterior views of skeletal tracer uptake, yet few question-answering benchmarks retain the labels and bounding boxes of their upstream data rather than reducing them to free text that cannot be checked against the image. We present ScintiGround-38K, a frozen, provenance-preserving resource for visual question answering (VQA), image localization, and grounded VQA. It contains 38,469 instances, comprising 13,371 VQA, 12,549 image-localization, and 12,549 grounded instances derived from 54,768 source records and 23,967 anterior-posterior image pairs from 2,815 patients. All bounding boxes and answer labels originate in the BS-80K source dataset, and question wording is drawn from fixed rule-based templates. Template matching recovers bone-region boxes in native image coordinates, source hotspot boxes provide lesion locations, and metadata-only rows are excluded from training and evaluation. Every instance retains its source identifier, view, native-image box, and patient-level split assignment, and raw upstream images are not redistributed. Two vision-language models, Qwen3-VL-2B-Instruct and PaliGemma-3B-pt-224, were adapted with low-rank adaptation (LoRA) on an identical multitask training set and evaluated on the held-out test set after model selection on validation data. Raw answer accuracy exceeded 0.90, yet balanced accuracy reached only 0.528 and 0.535 for visual question answering, abnormal-class F1 remained below 0.15, and strict grounded accuracy reached 0.765 and 0.701. High raw accuracy therefore conceals weak minority-class and evidence-level behaviour. ScintiGround-38K is intended for methodological research and has not undergone clinical validation.

1. Introduction

Planar whole-body bone scintigraphy (WBBS) with technetium-99m-labelled diphosphonates remains a widely used, low-cost screening examination for skeletal metastases in prostate, breast, and lung cancer, and continues to be performed in large volumes alongside newer tomographic and positron-emission alternatives [1], [2], [3]. The standard acquisition produces paired anterior and posterior whole-body projections in a single pass, so that a lesion suspected on one projection can be checked against its counterpart [4]. The modality thus combines a stable whole-body layout with focal hotspots that can be localized in the native image geometry. Together, these properties make WBBS a suitable setting in which to examine whether a model can connect an answer to the image evidence required to support it.

Interpreting these studies nevertheless remains demanding. Individual reporting workload in nuclear medicine has grown substantially over the past two decades [5], and inter-reader variability has motivated structured and standardized reporting of scintigraphic findings [6]. Quantitative decision support has a long history in this modality: the automated Bone Scan Index (BSI) platform quantifies the skeletal tumour burden of a whole-body scan [7], has been analytically validated as an imaging biomarker [8], has been proposed as a quantitative clinical-trial endpoint [9], and has been extended to breast cancer with neural-network segmentation [10]. Systematic reviews of deep learning for bone-scan classification [11], comparisons of deep learning with human observers on WBBS [12], and meta-

analyses of artificial-intelligence accuracy for tumour bone metastases [13] show that automated interpretation of this modality is now an active research area.

Medical visual question answering (VQA) is often evaluated only by answer correctness. This can be misleading when answers are predictable from question wording or class frequency rather than image content [14]. Groundable medical VQA addresses this limitation by evaluating the answer and its visual evidence together [15]. This concern is not merely theoretical: surveys of hallucination in medical large language and vision-language models [16] and physician-facing primers written for the nuclear-medicine community [17] both identify ungrounded or fabricated image interpretation as a primary obstacle to trustworthy deployment. For bone scintigraphy, this requires a benchmark that preserves source labels and bounding boxes rather than reducing them to untraceable language labels.

The benchmark presented here is derived from BS-80K, a collection of paired whole-body bone scans annotated with region labels and hotspot boxes [18]. These labels and boxes supply the answers and target locations of the derived classification and image-localization tasks, so that no label originates in generated text. Question wording is taken from the controlled text annotations of the LIBS-160K collection [19], which contributes language but no new spatial or clinical evidence.

WBBS studies and radiology question-answering resources therefore supply complementary components, yet we are not aware of a resource that combines paired WBBS, preserved original-image geometry, and separately measurable answer-only, localization-only, and grounded tasks. Section 2 examines this literature in detail.

This study presents ScintiGround-38K, an evidence-first benchmark derived from BS-80K annotations and fixed language templates. The contributions of this work are as follows.

1. Paired anterior and posterior VQA, image-localization, and grounded VQA tasks are defined with original-image locations across bone-region, lesion, and whole-body levels.
2. Provenance is preserved for every released instance through recorded source links, source hashes, quality exclusions, and patient-level splits, together with reported checks for data integrity and template duplication.
3. A release boundary is defined that distributes derived annotations only, so that raw upstream images remain governed by their original providers.
4. Parameter-efficient baseline protocols are specified for Qwen3-VL and PaliGemma, both trained on the same task set, with differences in model inputs, adapted components, and model selection reported explicitly.
5. Class-balanced and localization metrics are reported that separate raw answer accuracy from minority-class and evidence-level performance.

The remainder of this paper is organized as follows. Section 2 reviews related work on deep learning for bone scintigraphy, medical VQA, grounded question answering, and dataset provenance. Section 3 describes the source evidence, construction, and evaluation protocol. Section 4 reports dataset

characteristics, baseline results, and qualitative examples. Section 5 discusses the findings, limitations, responsible use, and release boundary, and Section 6 concludes the paper.

2. Related Work

2.1. Deep learning for whole-body bone scintigraphy

Research on WBBS has addressed lesion detection and anatomical localization in paired views [20]. Ibrahim et al. studied metastatic bone-disease foci with reader comparison [21], and Hajianfar et al. compared artificial-intelligence analysis of whole-body bone scintigraphy with human observers [12]. These studies address clinically relevant detection, localization, or reader-comparison problems rather than question answering, but they establish the importance of preserving the spatial and clinical context of the source modality.

Earlier work by Huang et al. applied anatomy-aware image processing to computer-aided analysis of whole-body bone scans [22], providing historical context for spatially informed analysis rather than a contemporary vision-language baseline. More recent clinical artificial-intelligence studies of bone scintigraphy and tumour bone metastases add further clinical context [13], [23], but none of them is framed as grounded question answering or offers a benchmark comparator.

The broader deep-learning literature on this modality is dominated by image-level classification and hotspot detection. Convolutional networks have been trained to classify whole-body scans of prostate-cancer patients as metastatic or non-metastatic [24], [25], to assist diagnosis across several primary tumours [26], to perform multiclass classification of whole-body scintigraphic images with attention modules [27], to detect metastasis caused by lung cancer using residual and hybrid-attention designs [28], and to detect chest and pelvic metastases from small datasets [29]. Diagnostic-performance studies in nuclear-medicine journals report reader-level comparisons for whole-body bone scans in prostate cancer [30]. Across this body of work the output is a class label, a heat map, or a detection box; none of these resources exposes a question-answering interface, and none releases paired question, answer, and box triplets whose provenance can be traced back to a specific source annotation.

2.2. Medical VQA datasets

Medical VQA resources provide complementary tasks and annotation designs. VQA-RAD contains clinically generated questions and answers for radiology images [31]. SLAKE adds bilingual, semantically labelled and knowledge-enhanced annotations [32], and PathVQA extends the task to pathology images at scale [33]. The ImageCLEF VQA-Med series standardized shared-task evaluation over modality, plane, organ, and abnormality question categories [34]. More recent resources scale the task further: PMC-VQA constructs a large generative benchmark from the biomedical literature [35], and OmniMedVQA aggregates a comprehensive multi-modality evaluation suite for medical large vision-language models [36]. MIMIC-CXR provides de-identified chest radiographs with free-text reports and is relevant as a source-data and provenance comparator [37]. MedSG-Bench targets grounding across medical image

sequences [38]. Instruction-tuned medical vision-language models such as LLaVA-Med [39] and Med-Flamingo [40] are typically evaluated on these resources. All of these are radiology, pathology, or mixed-modality resources rather than WBBS datasets, so their modality and reporting context limit direct numerical comparison with a bone-scintigraphy benchmark.

2.3. Grounded and localized medical question answering

GEMeX frames chest-radiograph question answering as groundable and explainable, supporting the evaluation of answers together with visual evidence [41]. HEAL-MedVQA localizes the relevant region before answering and reports both components [42]. MIMIC-Ext-CXR-QBA is a credentialed chest-radiograph dataset record with question-box-answer triplets and scene graphs [43]. RadVLM is a multitask conversational radiology model [44] released together with a credentialed instruction-dataset record [45]. These resources serve as methodological comparators; none of them, however, is an openly available WBBS dataset.

Adjacent work grounds language in medical images without an explicit question-answering format. Medical phrase grounding aligns report phrases with image regions [46], ChEX supports interactive localization and region description in chest radiographs [47], MAIRA-2 generates radiology reports in which each finding is tied to a bounding box [48], and Chest ImaGenome supplies scene graphs and anatomical boxes that make localized clinical reasoning measurable [49]. Together these resources establish that localized evidence can be annotated, released, and scored; they also show that the practice is concentrated in chest radiography.

The medical VQA literature also distinguishes answer generation from the localization of visual evidence. Answer-only benchmarks can assess response correctness without testing whether a model identifies the relevant image region. Groundable resources make that evidence measurable, but their chest-radiograph or general-radiology settings are not equivalent to paired whole-body bone scintigraphy. This distinction motivates a benchmark that keeps task design aligned with the geometry and evidence available in its source data.

2.4. Language priors, shortcut learning, and grounding failure

The changing-priors formulation of VQA showed that models can answer correctly by exploiting question-conditioned answer priors rather than image content [14]. Medical adaptations of this critique document modality-preference bias [50], answers produced by guessing rather than evidence [51], and modality-specific confounding that requires causal intervention to remove [52]. Surveys of hallucination in multimodal large language models [53] and of medical hallucination in particular [16] describe the same failure mode at generation time, and a recent systematic study of visual grounding in medical images reports that state-of-the-art medical multimodal models frequently fail to ground predictions in clinically relevant regions even when their answers are correct [15]. These findings motivate two design

decisions adopted here: answers and boxes are scored separately as well as jointly, and every reported answer remains traceable to the source annotation that produced it.

2.5. Dataset documentation and provenance

Documentation practice in machine learning has converged on structured disclosure of dataset composition, collection, and intended use [54] and on comparable reporting for released models [55]. In health care, MINIMAR defines minimum information for medical artificial-intelligence reporting [56], and the Checklist for Artificial Intelligence in Medical Imaging and its 2024 update gives item-level guidance for imaging studies [57], [58]; TRIPOD + AI provides analogous guidance for prediction-model reporting [59]. The FAIR guiding principles frame the stewardship obligations of a released resource [60], and recent multimodal curation experience from the Medical Imaging and Data Resource Center illustrates how interoperability and provenance are maintained in practice [61]. ScintiGround-38K follows these conventions by shipping derived annotations with source identifiers, hashes, exclusion policies, and audit reports rather than an unlabelled convenience sample.

2.6. Positioning of the present work

The relevant literature therefore provides complementary components. WBBS studies address lesion detection, anatomical localization, or human-reader comparison, whereas radiology resources provide question-answer and localized-evidence designs. None provides the exact combination of paired whole-body bone scintigraphy, preserved original-image geometry, and separately measurable answer-only VQA, image-localization-only tasks, and grounded VQA. This is a modality- and annotation-specific distinction, not a claim that earlier work lacks VQA, image localization, or quality control.

3. Materials and Methods

This section describes the source evidence, deterministic task construction, quality assurance, patient-level split governance, baseline protocols, and evaluation rules. Derived dataset instances remain traceable to accepted source evidence rather than to model-generated clinical claims. The procedures are specified before reporting dataset or model results.

3.1. Source evidence and release scope

Throughout this paper, ScintiGround-38K denotes the released collection of derived annotations, whereas WBBS denotes the imaging modality from which those annotations originate; the two are not interchangeable. ScintiGround-38K is a frozen, deterministic dataset containing 38,469 instances for WBBS question answering and image localization. Each instance retains its source label, native-image box, image reference, and patient split. Question wording is drawn from the LIBS-160K template collection [19], while all boxes and labels come from BS-80K [18]. The release contains derived annotations and does not redistribute raw BS-80K images. Metadata-only rows are excluded from training, model selection, and evaluation. The accompanying documentation follows the structured disclosure conventions of datasheets for datasets [54] and the FAIR guiding principles [60], and reporting

of the baseline experiments follows the Checklist for Artificial Intelligence in Medical Imaging [57], [58]. Figure 1 summarizes how the two source datasets and rule-based generation combine into the released instances.

3.2. Dataset construction and task generation

Each source record stores an image reference, patient identifier, view, native-image box, source label, available hotspot boxes, and quality flags. A dataset instance is created only when its source record passes the quality checks. Tasks that need both views use the linked anterior–posterior image pair.

In BS-80K, experienced nuclear-medicine physicians annotated hotspot boxes as normal or abnormal after repeated inspection; the source paper reports annotation accuracy above 95% [18]. Whole-body and region labels remain image-based source labels rather than pathological ground truth. Template matching recovers bone-region and whole-body boxes, but it does not create clinical labels. English templates create the question wording only.

Each dataset instance retains its split, source identifier, label, view, and target level. VQA asks for a yes-or-no answer without a predicted box, image localization asks for one or two boxes without answer scoring, and grounded VQA requires both an answer and one or more target boxes; Table 1 summarizes these task types. Because the question templates are rule-based and repeatable, and because the answers and target locations are supplied by the source annotations, template wording introduces no new clinical or spatial evidence. Figure 2 shows a representative instance for each task type.

Table 1
ScintiGround-38K task types and required outputs

Task type	Question target	Required output	Image evidence	Target levels	Count
VQA	Specified region lesion or whole-body target	Yes or no answer	None	Bone region metastasis whole body	13,371
Image localization	Named bone region or source hotspot	One or two target boxes	Native-image region or hotspot boxes	Bone region metastasis	12,549
Grounded VQA	Specified target and classification	Answer plus one or two target boxes	Native-image region or hotspot boxes	Bone region metastasis	12,549

3.3. Quality assurance and data splits

Patients are assigned to 80%, 10%, and 10% training, validation, and test splits. All records from a patient remain in one split and linked anterior and posterior images remain together. Subject-wise rather than record-wise partitioning is used for two reasons: splitting repeated measurements from the same individual across partitions inflates apparent performance in clinical machine learning, and patient

identity is a documented shortcut confounder in medical imaging models. Table 2 reports the number of source records and image pairs in each split.

Quality checks verify unique identifiers, valid bounding-box coordinates, complete image pairs, and patient-level split assignment. Source hashes, quality flags, and identifiers link each instance to its source record. Ambiguous hotspot assignments and lower-precision shoulder boxes remain marked so they can be examined separately or excluded in later analyses.

Table 2 Patient-level data splits					
Split	Patients with source records	Source records	Patients with image pairs	Image pairs	Patient fraction
Train	2,252	43,853	2,135	19,215	0.800
Validation	282	5,489	267	2,403	0.100
Test	281	5,426	261	2,349	0.100
Total	2,815	54,768	2,663	23,967	1.000

3.4. Baseline protocols

Qwen3-VL-2B-Instruct [62] and PaliGemma-3B-pt-224 [63] were both adapted with low-rank adaptation (LoRA) [64]. Qwen3-VL extends the design lineage of Qwen2.5-VL [65], and PaliGemma combines a SigLIP image encoder with a Gemma language decoder [66]; the successor PaliGemma 2 family [67] was not used here so that the released baselines match widely available checkpoints. Both models were trained on the same multitask training set of 30,862 instances, including 30,204 paired-image and 658 single-image instances, covering VQA, image localization, and grounded VQA. Qwen3-VL received the anterior and posterior images natively, while PaliGemma received them composed into one anterior-posterior canvas because it accepts one image input.

Qwen3-VL models were selected using validation results. The PaliGemma run used one final adapter without an epoch-by-epoch validation comparison. Both models were then evaluated on the test set. These differences mean that the results do not support a direct model-family comparison. Figure 3 outlines this shared training procedure.

3.5. Evaluation protocol

Models are trained only on the training split. Prompts, thresholds, and model selection use validation data, while the test set is evaluated after these choices are fixed. Confidence intervals use patient-level bootstrap resampling. VQA and grounded VQA report accuracy, balanced accuracy, macro-F1, and class precision, recall, and F1. Balanced accuracy and macro-F1 are reported as primary summaries because raw accuracy is uninformative under the severe label imbalance of the source data.

Image localization reports mean intersection over union (IoU), box accuracy at IoU ≥ 0.50 , strict localization accuracy, and valid responses. Grounded VQA also reports strict grounded accuracy, which credits a prediction only when both the answer and the box are correct.

Bilingual evaluation understudy (BLEU) [68], metric for evaluation of translation with explicit ordering (METEOR) [69], and recall-oriented understudy for gisting evaluation, longest common subsequence (ROUGE-L) [70] are reported as supplementary lexical-overlap measures for VQA and grounded VQA. They are calculated from standardized saved predictions in task-specific test subsets and do not replace the primary metrics, because template wording can overlap without a correct class or location, a limitation of n-gram overlap that has been documented for medical open-ended question answering.

4. Results

This section reports dataset characteristics and baseline results on the test set. Metrics and prediction examples are taken from the saved results. Each model was evaluated after selection on validation data.

4.1. Dataset characteristics and dataset comparison

ScintiGround-38K contains 38,469 dataset instances in patient-level training, validation, and test splits. Table 3 reports the number of instances in each split. These counts include only instances used by the main benchmark.

Table 3
Dataset instances by split

Split	Dataset instances
Train	30,862
Validation	3,853
Test	3,754
Total	38,469

Table 4 reports the number of source records and anterior–posterior image pairs retained in the release. The source records include 51,803 normal and 2,965 abnormal labels. These counts show the marked class imbalance in the underlying source labels.

Table 4
Source records and image pairs in the release

Measure	Count
Source records	54,768
Patients with source records	2,815
Anterior–posterior image pairs	23,967
Patients with image pairs	2,663

The quality audit found no duplicate or missing identifiers, invalid boxes, or split leakage. The sampling procedure retained all 54,768 source records through at least one selected instance. Exact duplicate-question rates were 0.973 for VQA, 0.965 for image localization, and 0.972 for grounded VQA.

ScintiGround-38K was compared structurally with SLAKE [32], PathVQA [33], MedSG-Bench [38], and HEAL-MedVQA [42]; this comparison concerns dataset design rather than model performance. ScintiGround-38K combines three task types with region and lesion localization, but not whole-body localization. Differences in modality, language, task definition, and annotation policy limit direct numerical comparison.

4.2. Baseline results

Table 5 reports answer results on the test set. On the shared grounded VQA task, balanced accuracy is near 0.530 for both models even though raw accuracy exceeds 0.90. This gap shows that high raw accuracy alone does not establish reliable classification of the minority class.

Table 5
Test-set answer results. Bold values are higher within the same task

Model	Task	Accuracy	Balanced accuracy	Macro-F1
Qwen3-VL-2B-Instruct	VQA	0.886	0.528	0.529
PaliGemma-3B-pt-224	VQA	0.888	0.535	0.541
Qwen3-VL-2B-Instruct	Grounded VQA	0.906	0.539	0.550
PaliGemma-3B-pt-224	Grounded VQA	0.908	0.524	0.524

Lexical-overlap measures were calculated from standardized saved predictions and are supplementary to the primary answer metrics. On the VQA test subset ($n = 1,305$), Qwen3-VL obtained BLEU 0.353, METEOR 0.541, and ROUGE-L 0.458; PaliGemma obtained 0.351, 0.523, and 0.442, respectively. On the grounded VQA test subset ($n = 1,224$), Qwen3-VL obtained BLEU 0.358, METEOR 0.532, and ROUGE-L 0.447; PaliGemma obtained 0.344, 0.516, and 0.435, respectively. These measures do not assess localization quality and do not change the primary classification, localization, or strict grounded results.

Table 6 reports image-localization results on the test set. Strict localization accuracy requires a correct box under the task rule and, for grounded VQA, strict grounded accuracy requires both a correct answer and a correct box.

Table 7 reports class-specific precision, recall, and F1 for the Normal and Abnormal answer labels. It makes the class imbalance visible without treating these source labels as clinical diagnoses.

Table 6
Test-set localization results. Bold values are higher within the same task

Model	Task	Mean IoU	Box accuracy (IoU ≥ 0.50)	Strict localization accuracy	Valid responses
Qwen3-VL-2B-Instruct	Image localization	0.642	0.857	0.827	0.998
PaliGemma-3B-pt-224	Image localization	0.633	0.787	0.762	1.000
Qwen3-VL-2B-Instruct	Grounded VQA	0.640	0.843	0.765	0.999
PaliGemma-3B-pt-224	Grounded VQA	0.629	0.776	0.701	1.000

Table 7
Class-specific test results. Bold values are higher within the same task and class

Model	Task	Normal			Abnormal		
		Precision	Recall	F1	Precision	Recall	F1
Qwen3-VL-2B-Instruct	VQA	0.897	0.985	0.939	0.357	0.071	0.118
PaliGemma-3B-pt-224	VQA	0.899	0.985	0.940	0.414	0.085	0.141
Qwen3-VL-2B-Instruct	Grounded VQA	0.913	0.991	0.951	0.526	0.087	0.149
PaliGemma-3B-pt-224	Grounded VQA	0.910	0.996	0.951	0.600	0.052	0.096

4.3. Qualitative examples

Figure 4 shows two grounded VQA test examples with source boxes and parsed outputs from both models. The vertebral-column example uses the posterior image, and the right-knee example uses the anterior image. They illustrate answer and box parsing without changing the test-set summary metrics.

5. Discussion

ScintiGround-38K is a paired bone-scintigraphy resource with source labels, controlled questions, and patient-level splits. Its contribution lies in benchmark design and measurement rather than in clinical decision support or in a matched comparison of model families.

5.1. Interpretation and comparison boundary

Severe class imbalance and repeated question templates require balanced, class-specific reporting and non-image control baselines. Rare abnormal ankle examples, lower-precision shoulder boxes, and ambiguous hotspot-to-region assignments limit fine-grained claims. Raw accuracy alone is therefore not an adequate summary.

The pattern observed across the two baselines is itself informative. Both adapted models exceeded 0.90 raw answer accuracy while their balanced accuracy stayed near the 0.50 level expected of an uninformative binary decision, and abnormal-class F1 remained below 0.15. Localization behaved differently: mean IoU near 0.64 and box accuracy between 0.78 and 0.86 indicate that both models learned the stable whole-body layout of the modality, yet strict grounded accuracy fell below the corresponding box accuracy because the answer component failed on the minority class. Answer competence and evidence competence therefore dissociate on this benchmark, which is consistent with reports that medical multimodal models can produce plausible answers without attending to the clinically relevant region [15]. Reporting the two components separately, as done here, is what makes the dissociation visible.

The models differ in image format, adapted components, and model-selection procedure. Qwen3-VL received both images natively, while PaliGemma received a single composed canvas, and only Qwen3-VL used epoch-by-epoch validation selection. The results therefore describe this benchmark and do not establish a model-family comparison.

5.2. Limitations

ScintiGround-38K contains derived annotations and has not undergone clinical validation. Its labels come from specialist image review, but image-based labels can differ from pathological findings [18]. The dataset derives from a single source, has no external test cohort, and does not redistribute raw images.

The source labels are severely imbalanced. Balanced metrics, class-specific results, and confidence intervals are therefore needed when interpreting the results. Small rare-class or region-specific subgroups should not support independent claims.

The question language is controlled rather than open ended. Its finite template bank can create wording-only shortcuts, so non-image control baselines are needed before a score can be interpreted as image-grounded behaviour.

Some hotspot assignments and shoulder boxes have spatial uncertainty. These instances remain marked so that future analyses can examine them separately or exclude them. Model selection uses validation data, and the test set is used only after the protocol is set.

5.3. Responsible use and release

The dataset is intended for reproducible research on question answering and image localization. It is not intended for clinical diagnosis, treatment selection, or patient-care deployment. Future evaluations should use the test set, report class-specific and anatomical-region results, and distinguish source labels from generated model text.

The release includes code, source hashes, policies, derived manifests, audit reports, documentation, and checksums. Raw upstream images and annotations are not redistributed. A persistent identifier, version tag, and licence will accompany the archived release at publication, and access to upstream source images remains governed by the original data providers' terms.

6. Conclusion

ScintiGround-38K is a derived resource for paired anterior and posterior whole-body bone-scintigraphy (WBBS) VQA, image localization, and grounded VQA. Its instances retain source image references, native-image boxes, question templates, and patient-level splits. The resource supports research on answer and image evidence alignment. Test-set VQA balanced accuracy is near 0.530 for both models, while Qwen3-VL attains higher image-localization scores. Abnormal-class F1 remains low. The models differ in image format, adapted components, and model selection, so the results do not establish a model-family comparison. Taken together, the baselines show that raw answer accuracy above 0.90 can coexist with essentially uninformative minority-class behaviour and with strict grounded accuracy well below the corresponding box accuracy. Answer correctness and evidence localization are therefore separable capabilities in this modality, and a benchmark that measures only the former would have reported these systems as strong. Preserving source labels, native-image boxes, and patient-level provenance is what makes the difference measurable.

Three directions follow from these findings. First, non-image and question-only control baselines should be established so that any future score can be attributed to image evidence rather than template wording. Second, imbalance-aware training and evaluation, including region-stratified reporting for the marked shoulder and ambiguous hotspot instances, is required before fine-grained anatomical claims are made. Third, external validation on an independent bone-scintigraphy cohort, and extension of the localization targets to the whole-body level, would test whether the behaviour reported here generalizes beyond a single source dataset. Until such work is completed, ScintiGround-38K should be used as a methodological research resource rather than as evidence of clinical readiness.

References

1. Jambor I, Kuisma A, Ramadan S, Huovinen R, Sandell M, Kajander S, Kemppainen J, Kauppila E, Auren J, Merisaari H, Saunavaara J, Noponen T, Minn H, Aronen HJ, Seppänen M: Prospective evaluation of planar bone scintigraphy, SPECT, SPECT/CT, 18F-NaF PET/CT and whole body 1.5T MRI, including DWI, for the detection of bone metastases in high risk breast and prostate cancer patients: SKELETA clinical trial. *Acta Oncol* 55:59-67, 2016
2. Wang Y, Qiu Y, Yan X: Prostate-specific membrane antigen PET versus [99mTc]Tc-MDP bone scan for diagnosing bone metastasis in prostate cancer: a head-to-head comparative meta-analysis. *Front Med* 11:1451565, 2024
3. Min JW, Um SW, Yim JJ, Yoo CG, Han SK, Shim YS, Kim YW: The Role of Whole-Body FDG PET/CT, Tc 99m MDP Bone Scintigraphy, and Serum Alkaline Phosphatase in Detecting Bone Metastasis in Patients with Newly Diagnosed Lung Cancer. *J Korean Med Sci* 24:275-280, 2009
4. Shafi A, Thorsson O, Edenbrandt L: New Routine for Nuclear Medicine Technologists to Determine When to Add SPECT/CT to a Whole-Body Bone Scan. *J Nucl Med Technol* 42:28-32, 2014
5. Amasha AAH, Kasalak Ö, Glaudemans AWJM, Noordzij W, Dierckx RAJO, Koopmans KP, Kwee TC: Increased individual workload for nuclear medicine physicians over the past years: 2008-2023 data from The Netherlands. *Ann Nucl Med* 39:273-277, 2025
6. Evangelista L, Filippi L: Structured reporting in prostate cancer: the revolution of quality in nuclear medicine scan interpretation. *Eur Radiol* 34:1155-1156, 2023
7. Ulmert D, Kaboteh R, Fox JJ, Savage C, Evans MJ, Lilja H, Abrahamsson PA, Björk T, Gerdtsen A, Bjartell A, Gjertsson P, Höglund P, Lomsky M, Ohlsson M, Richter J, Sadik M, Morris MJ, Scher HI, Sjöstrand K, Yu A, Suurküla M, Edenbrandt L, Larson SM: A Novel Automated Platform for Quantifying the Extent of Skeletal Tumour Involvement in Prostate Cancer Patients Using the Bone Scan Index. *Eur Urol* 62:78-84, 2012
8. Anand A, Morris MJ, Kaboteh R, Båth L, Sadik M, Gjertsson P, Lomsky M, Edenbrandt L, Minarik D, Bjartell A: Analytic Validation of the Automated Bone Scan Index as an Imaging Biomarker to Standardize Quantitative Changes in Bone Scans of Patients with Metastatic Prostate Cancer. *J Nucl Med* 57:41-45, 2016
9. Mota JM, Armstrong AJ, Larson SM, Fox JJ, Morris MJ: Measuring the unmeasurable: automated bone scan index as a quantitative endpoint in prostate cancer clinical trials. *Prostate Cancer Prostatic Dis* 22:522-530, 2019
10. Inaki A, Nakajima K, Wakabayashi H, Mochizuki T, Kinuya S: Fully automated analysis for bone scintigraphy with artificial neural network: usefulness of bone scan index (BSI) in breast cancer. *Ann Nucl Med* 33:755-765, 2019
11. Kao YS, Huang CP, Tsai WW, Yang J: A systematic review for using deep learning in bone scan classification. *Clin Transl Imaging* 11:271-283, 2023
12. Hajianfar G, Sabouri M, Salimi Y, Amini M, Bagheri S, Jenabi E, Hekmat S, Maghsudi M, Mansouri Z, Khateri M, Hosein Jamshidi M, Jafari E, Bitarafan Rajabi A, Assadi M, Oveisi M, Shiri I, Zaidi H: Artificial intelligence-based analysis of whole-body bone scintigraphy: The quest for the optimal

- deep learning algorithm and comparison with human observer performance. *Z Med Phys* 34:242-257, 2024
13. Tao H, Hui X, Zhang Z, Zhu R, Wang P, Zhou S, Yang K: Accuracy of artificial intelligence in detecting tumor bone metastases: a systematic review and meta-analysis. *BMC Cancer* 25:286, 2025
 14. Agrawal A, Batra D, Parikh D, Kembhavi A: Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 4971-4980, 2018
 15. Liu G, Yu T, Ebrahimkhani S, Shawn LZZ, Ng KP, Cheung NM: How Do Medical MLLMs Fail? A Study on Visual Grounding in Medical Images. *arXiv:2603.14323*, 2026
 16. Zhu Z, Zhang Y, Zhuang X, Zhang F, Wan Z, Chen Y, Long Q, Zheng Y, Wu X: Can We Trust AI Doctors? A Survey of Medical Hallucination in Large Language and Large Vision-Language Models. In: *Findings of the Association for Computational Linguistics: ACL 2025*, pp 6748-6769, 2025
 17. Bradshaw TJ, Tie X, Warner J, Hu J, Li Q, Li X: Large Language Models and Large Multimodal Models in Medical Imaging: A Primer for Physicians. *J Nucl Med* 66:173-182, 2025
 18. Huang Z, Pu X, Tang G, Ping M, Jiang G, Wang M, Wei X, Ren Y: BS-80K: The first large open-access dataset of bone scan images. *Comput Biol Med* 151:106221, 2022
 19. Wei X. LIBS-160K. figshare dataset. Available at <https://figshare.com/articles/dataset/LIBS-160K/28715375>. Accessed 10 August 2026.
 20. Huang K, Huang S, Chen G, Li X, Li S, Liang Y, Gao Y: An end-to-end multi-task system of automatic lesion detection and anatomical localization in whole-body bone scintigraphy by deep learning. *Bioinformatics* 39:btac753, 2023
 21. Ibrahim A, Vaidyanathan A, Primakov S, Belmans F, Bottari F, Refaee T, Lovinfosse P, Jadoul A, Derwael C, Hertel F, Woodruff HC, Zacho HD, Walsh S, Vos W, Occhipinti M, Hanin FX, Lambin P, Mottaghy FM, Hustinx R: Deep learning based identification of bone scintigraphies containing metastatic bone disease foci. *Cancer Imaging* 23:12, 2023
 22. Huang JY, Kao PF, Chen YS: A Set of Image Processing Algorithms for Computer-Aided Diagnosis in Nuclear Medicine Whole Body Bone Scan Images. *IEEE Trans Nucl Sci* 54:514-522, 2007
 23. Uchuwat Y, Ruengchaijatuporn N, Sukprakun C, Vachatimanont S, Chantadisai M, Kingpetch K, Chaiwatanarat T, Tepmongkol S, Buakhao C, Phuangmali K, Sriswasdi S, Rakvongthai Y: Impact of artificial intelligence assistance on bone scintigraphy diagnosis. *Phys Eng Sci Med* 48:1791-1799, 2025
 24. Papandrianos N, Papageorgiou E, Anagnostis A, Papageorgiou K: Bone metastasis classification using whole body images from prostate cancer patients based on convolutional neural networks application. *PLoS One* 15:e0237213, 2020
 25. Papandrianos N, Papageorgiou EI, Anagnostis A: Development of Convolutional Neural Networks to identify bone metastasis for prostate cancer patients in bone scintigraphy. *Ann Nucl Med* 34:824-832, 2020

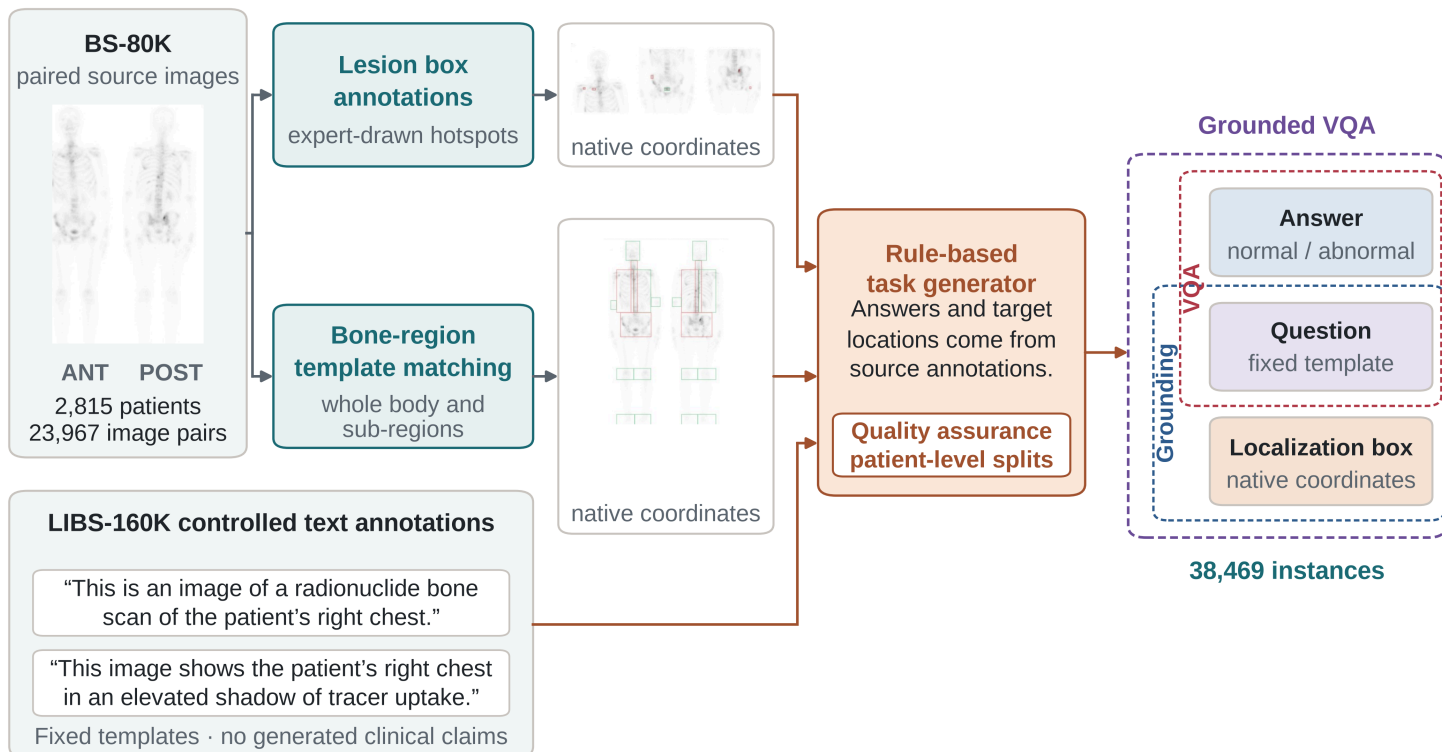
26. Zhao Z, Pi Y, Jiang L, Xiang Y, Wei J, Yang P, Zhang W, Zhong X, Zhou K, Li Y, Li L, Yi Z, Cai H: Deep neural network based artificial intelligence assisted diagnosis of bone scintigraphy for cancer bone metastasis. *Sci Rep* 10:17046, 2020
27. Lin Q, Cao C, Li T, Cao Y, Man Z, Wang H: Multiclass classification of whole-body scintigraphic images using a self-defined convolutional neural network with attention modules. *Med Phys* 48:5782-5793, 2021
28. Guo Y, Lin Q, Zhao S, Li T, Cao Y, Man Z, Zeng X: Automated detection of lung cancer-caused metastasis by classifying scintigraphic images using convolutional neural network with residual connection and hybrid attention mechanism. *Insights Imaging* 13:24, 2022
29. Cheng DC, Liu CC, Hsieh TC, Yen KY, Kao CH: Bone Metastasis Detection in the Chest and Pelvis from a Whole-Body Bone Scan Using Deep Learning and a Small Dataset. *Electronics* 10:1201, 2021
30. Han S, Oh JS, Lee JJ: Diagnostic performance of deep learning models for detecting bone metastasis on whole-body bone scan in prostate cancer. *Eur J Nucl Med Mol Imaging* 49:585-595, 2022
31. Lau JJ, Gayen S, Ben Abacha A, Demner-Fushman D: A dataset of clinically generated visual questions and answers about radiology images. *Sci Data* 5:180251, 2018
32. Liu B, Zhan LM, Xu L, Ma L, Yang Y, Wu XM: SLAKE: A Semantically-Labeled Knowledge-Enhanced Dataset for Medical Visual Question Answering. In: *Proceedings of the IEEE 18th International Symposium on Biomedical Imaging*, pp 1650-1654, 2021
33. He X, Zhang Y, Mou L, Xing E, Xie P: PathVQA: 30000+ Questions for Medical Visual Question Answering. *arXiv:2003.10286*, 2020
34. Ben Abacha A, Hasan SA, Datla VV, Liu J, Demner-Fushman D, Müller H: VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019. In: *CLEF 2019 Working Notes, CEUR Workshop Proceedings vol 2380*, 2019
35. Zhang X, Wu C, Zhao Z, Lin W, Zhang Y, Wang Y, Xie W: Development of a large-scale medical visual question-answering dataset. *Commun Med* 4:277, 2024
36. Hu Y, Li T, Lu Q, Shao W, He J, Qiao Y, Luo P: OmniMedVQA: A New Large-Scale Comprehensive Evaluation Benchmark for Medical LLM. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 22170-22183, 2024
37. Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng CY, Mark RG, Horng S: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data* 6:317, 2019
38. Yue J, Zhang S, Jia Z, Xu H, Han Z, Liu X, Wang G: MedSG-Bench: A Benchmark for Medical Image Sequences Grounding. *arXiv:2505.11852*, 2025
39. Li C, Wong C, Zhang S, Usuyama N, Liu H, Yang J, Naumann T, Poon H, Gao J: LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. In: *Advances in Neural Information Processing Systems* 36, pp 28541-28564, 2023

40. Moor M, Huang Q, Wu S, Yasunaga M, Zakka C, Dalmia Y, Reis EP, Rajpurkar P, Leskovec J: Med-Flamingo: a Multimodal Medical Few-shot Learner. arXiv:2307.15189, 2023
41. Liu B, Zou K, Zhan LM, Lu Z, Dong X, Chen Y, Xie C, Cao J, Wu XM, Fu H: GEMeX: A Large-Scale, Groundable, and Explainable Medical VQA Benchmark for Chest X-Ray Diagnosis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 21310-21320, 2025
42. Nguyen D, Ho MK, Ta H, Nguyen TT, Chen Q, Rav K, Dang QD, Ramchandre S, Phung SL, Liao Z, To MS, Verjans J, Nguyen PL, Phan VMH: Localizing Before Answering: A Benchmark for Grounded Medical Visual Question Answering. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, pp 7670-7678, 2025
43. Müller P, Jungmann F, Kaissis G, Rueckert D: MIMIC-Ext-CXR-QBA: A Structured, Tagged, and Localized Visual Question Answering Dataset with Question-Box-Answer Triplets and Scene Graphs for Chest X-ray Images. PhysioNet, DOI: 10.13026/8qmz-da41, 2025
44. Deperrois N, Matsuo H, Ruipérez-Campillo S, Vandenhirtz M, Laguna S, Ryser A, Fujimoto K, Nishio M, Sutter TM, Vogt JE, Kluckert J, Frauenfelder T, Blüthgen C, Nooralahzadeh F, Krauthammer M: RadVLM: A Multitask Conversational Vision-Language Model for Radiology. arXiv:2502.03333, 2025
45. Deperrois N, Matsuo H, Ruiperez-Campillo S, Vandenhirtz M, Laguna S, Ryser A, Fujimoto K, Nishio M, Sutter T, Vogt J, Kluckert J, Frauenfelder T, Bluethgen C, Nooralahzadeh F, Krauthammer M: RadVLM Instruction Dataset. PhysioNet, DOI: 10.13026/et5g-h222, 2025
46. Chen Z, Zhou Y, Tran A, Zhao J, Wan L, Ooi GSK, Cheng LTE, Thng CH, Xu X, Liu Y, Fu H: Medical Phrase Grounding with Region-Phrase Context Contrastive Alignment. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2023, Lecture Notes in Computer Science, pp 371-381, 2023
47. Müller P, Kaissis G, Rueckert D: ChEX: Interactive Localization and Region Description in Chest X-Rays. In: Computer Vision - ECCV 2024, Lecture Notes in Computer Science, pp 92-111, 2024
48. Bannur S, Bouzid K, Castro DC, Schwaighofer A, Thieme A, Bond-Taylor S, Ilse M, Pérez-García F, Salvatelli V, Sharma H, Meissen F, Ranjit M, Srivastav S, Gong J, Codella NCF, Falck F, Oktay O, Lungren MP, Wetscherek MT, Alvarez-Valle J, Hyland SL: MAIRA-2: Grounded Radiology Report Generation. arXiv:2406.04449, 2024
49. Wu JT, Agu NN, Lourentzou I, Sharma A, Paguio JA, Yao JS, Dee EC, Mitchell W, Kashyap S, Giovannini A, Celi LA, Moradi M: Chest ImaGenome Dataset for Clinical Reasoning. In: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, 2021. arXiv:2108.00316
50. Ye S, Naseem U, Meng M, Feng D, Kim J: A Causal Approach to Mitigate Modality Preference Bias in Medical Visual Question Answering. In: Proceedings of the First International Workshop on Vision-Language Models for Biomedical Applications, pp 13-17, 2024
51. Jiang S, Chen Y, Song S, Zhang Y, Jin Y, Feng Y, Wu J, Liu Z: Knowing or Guessing? Robust Medical Visual Question Answering via Joint Consistency and Contrastive Learning. In: Medical Image

52. Liu B, Liu L, Ding J, Yang X, Peng W, Liu L: CIMB-MVQA: Causal intervention on modality-specific biases for medical visual question answering. *Med Image Anal* 107:103850, 2026
53. Bai Z, Wang P, Xiao T, He T, Han Z, Zhang Z, Shou MZ: Hallucination of Multimodal Large Language Models: A Survey. *arXiv:2404.18930*, 2024
54. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé H III, Crawford K: Datasheets for datasets. *Commun ACM* 64:86-92, 2021
55. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, Spitzer E, Raji ID, Gebru T: Model Cards for Model Reporting. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp 220-229, 2019
56. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH: MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. *J Am Med Inform Assoc* 27:2011-2015, 2020
57. Mongan J, Moy L, Kahn CE: Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiol Artif Intell* 2:e200029, 2020
58. Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, Kahn CE Jr, for the CLAIM 2024 Update Panel: Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiol Artif Intell* 6:e240300, 2024
59. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, Ghassemi M, Liu X, Reitsma JB, van Smeden M, Boulesteix AL, Camaradou JC, Celi LA, Denaxas S, Denniston AK, Glocker B, Golub RM, Harvey H, Heinze G, Hoffman MM, Kengne AP, Lam E, Lee N, Loder EW, Maier-Hein L, Mateen BA, McCradden MD, Oakden-Rayner L, Ordish J, Parnell R, Rose S, Singh K, Wynants L, Logullo P: TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 385:e078378, 2024
60. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B: The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3:160018, 2016
61. Chen W, Whitney HM, Kahaki S, Meyer C, Li H, Sá RC, Lauderdale D, Napel S, Gersing K, Grossman RL, Giger ML: Multimodal data curation via interoperability: use cases with the Medical Imaging and Data Resource Center. *Sci Data* 12:1340, 2025
62. Qwen Team: Qwen3-VL Technical Report. *arXiv:2511.21631*, 2025

63. Beyer L, Steiner A, Pinto AS, Kolesnikov A, Wang X, Salz D, Neumann M, Alabdulmohsin I, Tschannen M, Bugliarello E, Unterthiner T, Keysers D, Koppula S, Liu F, Grycner A, Gritsenko A, Houlsby N, Kumar M, Rong K, Eisenschlos J, Kabra R, Bauer M, Bošnjak M, Chen X, Minderer M, Voigtlaender P, Bica I, Balazevic I, Puigcerver J, Papalampidi P, Henaff O, Xiong X, Soricut R, Harmsen J, Zhai X: PaliGemma: A versatile 3B VLM for transfer. arXiv:2407.07726, 2024
64. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W: LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685, 2021
65. Bai S, Chen K, Liu X, Wang J, Ge W, Song S, Dang K, Wang P, Wang S, Tang J, Zhong H, Zhu Y, Yang M, Li Z, Wan J, Wang P, Ding W, Fu Z, Xu Y, Ye J, Zhang X, Xie T, Cheng Z, Zhang H, Yang Z, Xu H, Lin J: Qwen2.5-VL Technical Report. arXiv:2502.13923, 2025
66. Gemma Team: Gemma: Open Models Based on Gemini Research and Technology. arXiv:2403.08295, 2024
67. Steiner A, Pinto AS, Tschannen M, Keysers D, Wang X, Bitton Y, Gritsenko A, Minderer M, Sherbondy A, Long S, Qin S, Ingle R, Bugliarello E, Kazemzadeh S, Mesnard T, Alabdulmohsin I, Beyer L, Zhai X: PaliGemma 2: A Family of Versatile VLMs for Transfer. arXiv:2412.03555, 2024
68. Papineni K, Roukos S, Ward T, Zhu WJ: BLEU: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp 311-318, 2002
69. Lavie A, Agarwal A: METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In: Proceedings of the Second Workshop on Statistical Machine Translation, pp 228-231, 2007
70. Lin CY: ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out, pp 74-81, 2004

Figures



Green outlines denote normal labels and red outlines abnormal labels; all source boxes are solid and given in native image coordinates.

Figure 1

Dataset construction. BS-80K supplies paired source images, lesion annotations, and bone-region labels; LIBS-160K supplies controlled text annotations. Rule-based generation combines source labels, native-image boxes, and language templates into task-specific dataset instances, subject to quality assurance and patient-level splitting. VQA pairs a question with an answer, image localization pairs a question with a target box, and grounded VQA requires both. Green outlines denote normal labels, and red outlines denote abnormal labels; all source boxes are solid and given in native image coordinates. The release retains split assignments and excludes metadata-only rows from training and evaluation. Raw BS-80K images remain external to the release

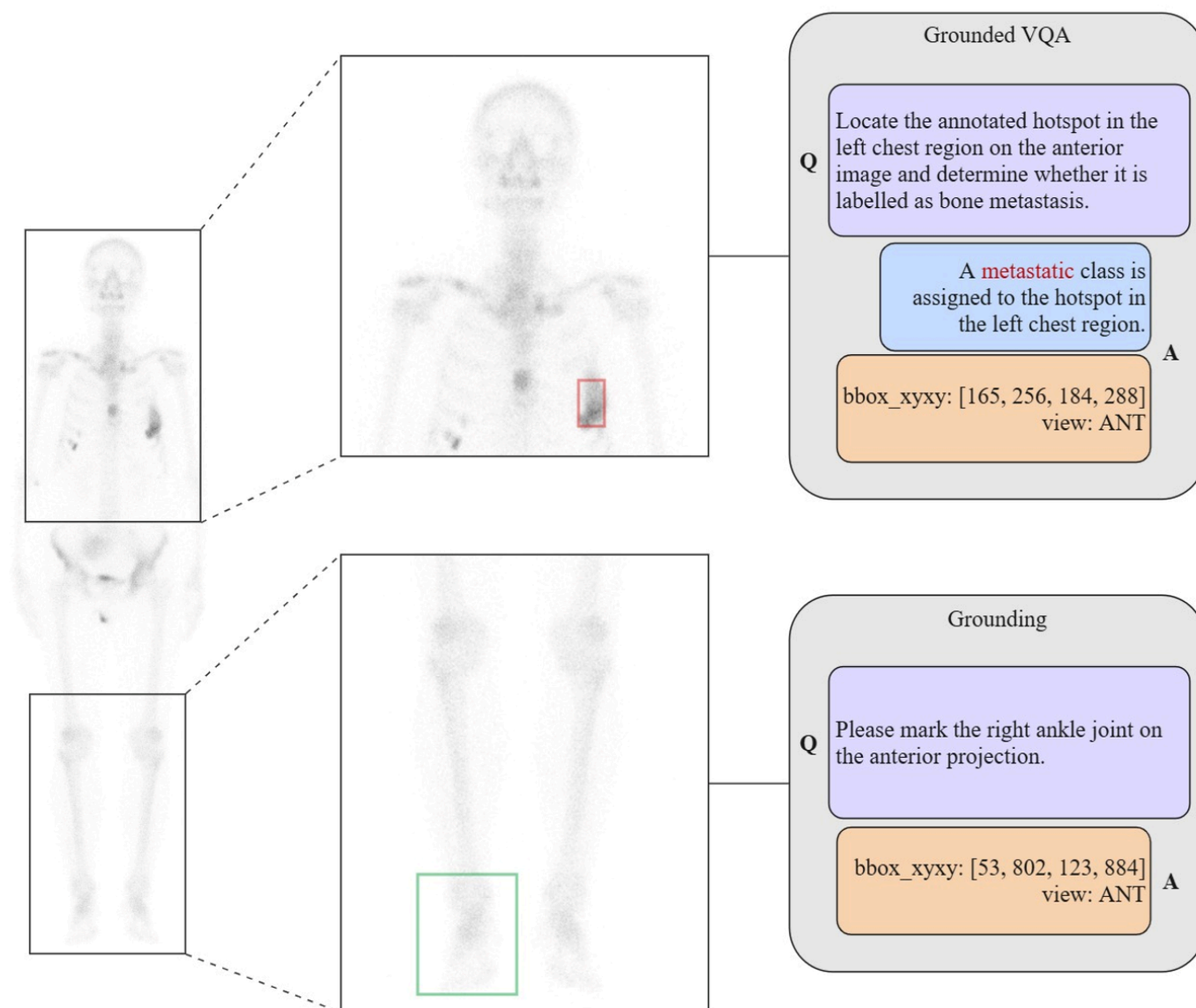


Figure 2

Examples of VQA, image-localization, and grounded VQA instances. Each instance links a question to a source answer, target box, or both. Scintigraphy panels use black uptake on white; green outlines denote normal labels, red outlines denote abnormal labels, and all source boxes are solid

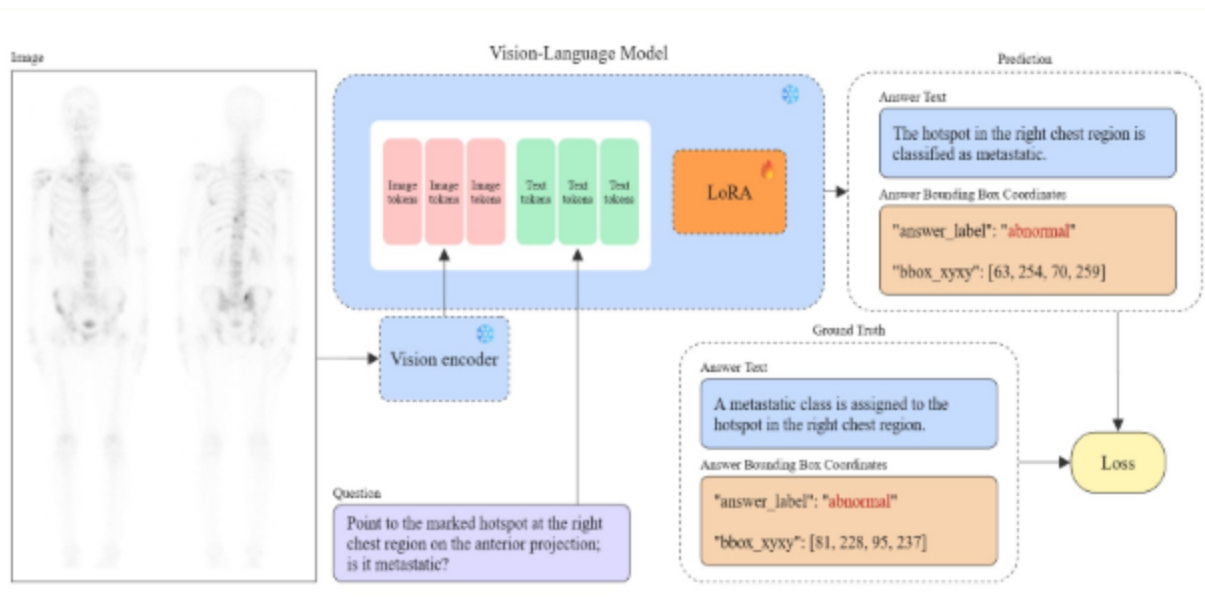
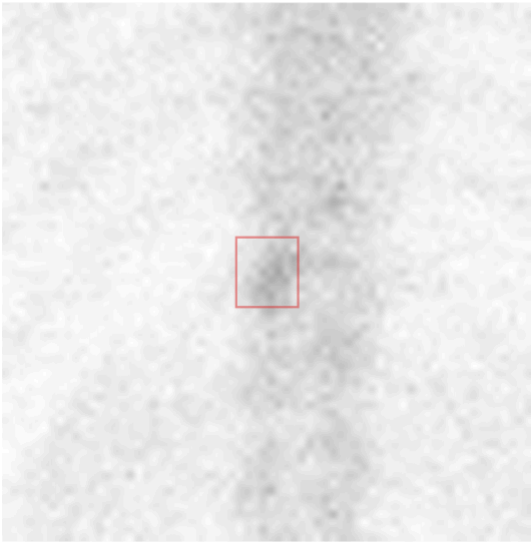


Figure 3

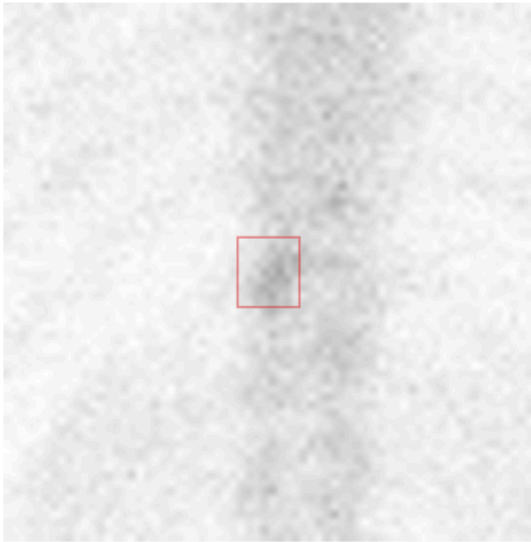
Vision-language-model training. Paired scintigraphy images and a question are encoded as image and text tokens by frozen model components. Trainable low-rank adapters update the model response, which is rendered as an answer label and, for image-localization tasks, a native-image bounding box

Qwen3-VL-2B-Instruct
vertebra hotspot (metastasis), Posterior view
Abnormal



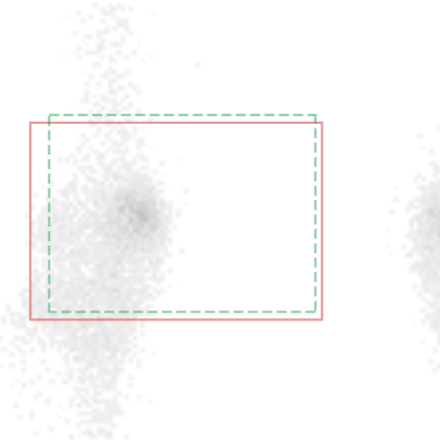
Ground truth: Abnormal
Predicted: Abnormal
Box IoU: 0.0000

PaliGemma-3B
vertebra hotspot (metastasis), Posterior view
Abnormal



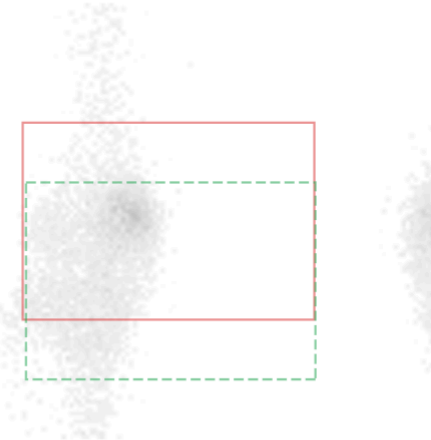
Ground truth: Abnormal
Predicted: Normal
Box IoU: 0.0000

Qwen3-VL-2B-Instruct
right knee (bone region), Anterior view
Abnormal



Ground truth: Abnormal
Predicted: Normal
Box IoU: 0.8388

PaliGemma-3B
right knee (bone region), Anterior view
Abnormal



Ground truth: Abnormal
Predicted: Normal
Box IoU: 0.5328

— abnormal or positive — normal or negative — source target - - - prediction

Figure 4

Grounded-VQA test examples. Each example shows the image evidence and parsed prediction for both models. Scintigraphy panels use black uptake on white. Solid outlines are source targets and dashed outlines are parsed model outputs; green denotes normal labels and red denotes abnormal labels