Beyond Fluency: A Clinical Benchmark and Anomaly-Enhanced Baseline for Spine MRI Report Generation

Supplementary Material

6. Lumbar MRI Report Generation

6.1. Ground Truth Reports for SPIDER Dataset

SPIDER provides structured radiological finding matrices rather than free-text reports. To enable report-generation evaluation, we converted these matrices into synthetic reference reports using predefined sentence templates. This yields standardized yet mildly diverse text that faithfully reflects the annotation matrix.

To reduce bias toward any single phrasing, we introduced limited lexical variation in both anatomical localization and finding description. Level information was expressed using simple constructions such as "at {level}" or "{level} shows", while findings were verbalized using a small set of interchangeable formulations, e.g., "disc space narrowing" versus "reduced disc height", and "disc bulge" versus "disc bulging".

Only findings annotated as present were included. In contrast to human-written reports, unaffected structures were not described explicitly. When multiple findings co-occurred at the same intervertebral level, they were merged into a single sentence using conjunctions to improve fluency and readability.

Representative templates included:

- **Disc narrowing:** "At {LEVEL}, there is disc space narrowing." / "{LEVEL} shows reduced disc height."
- **Disc bulging:** "At {LEVEL}, there is a disc bulge." / "{LEVEL} shows disc bulging."
- **Disc herniation:** "At {LEVEL}, there is a focal disc herniation." / "{LEVEL} shows a focal herniated disc."
- **Endplate defects:** "At {LEVEL}, there are endplate defects involving the {upper/lower/both} endplate(s)."
- **Modic changes:** "At {LEVEL}, Modic-type marrow signal changes are present adjacent to the endplates."

Pfirrmann grading was excluded because it is a composite severity score that subsumes lower-level findings and is typically not stated explicitly in clinical narrative reports. Modic changes were encoded as binary present/absent labels, since subtype differentiation cannot be reliably inferred from T1- or T2-weighted images alone. Endplate defects were described explicitly by anatomical location, distinguishing upper, lower, or combined involvement.

For example, a subject with disc narrowing at L2–L3 and L3–L4, and disc bulging at L3–L4 and L4–L5, would yield:

> "L2–L3 and L3–L4 show disc space narrowing. L3–L4 and L4–L5 show disc bulging."

6.2. Model-Specific Prompt Configurations

Different models require different input interfaces, including conversational prompts, task descriptors, and structured metadata fields. To evaluate prompt sensitivity under comparable conditions, we performed a zero-shot prompt-content ablation in which the informational content was held constant across models for each prompt-length setting (short, medium, long), while preserving model-specific formatting requirements.

The exact wording and decomposition of the input followed each model's recommended usage, e.g., system/user role separation for chat-based models and explicit indication or technique fields for structured interfaces. MedGemma, ChatGPT, and VILA-M3 were prompted through conversational inputs, whereas MAIRA-2 and BiomedGPT relied on metadata-style fields rather than role-based prompting. For models that support free-form report generation from task-defining prompts (MedGemma, ChatGPT, and VILA-M3), we additionally constrained the report boundaries to reduce variability unrelated to clinical content. Specifically, prompts requested a single concise paragraph delimited by `[[REPORT_START]]` and `[[REPORT_END]]`, enabling deterministic extraction of the report text while minimizing the effect of model-specific preambles or closing statements on downstream evaluation.

Table 4 summarizes the model-specific input configuration used for each prompt-length setting. This table specifies the prompt-content ablation, detailing the exact information given to each model under the short, medium, and long settings. Table 5 reports the corresponding zero-shot results on LSMRI and SPIDER datasets. Columns correspond to prompt settings (short, medium, long), rows correspond to MRI contrast inputs (T1w, T2w, T1w+T2w), and entries report BERTScore F1, BLEU, METEOR, and ROUGE-L F1, allowing prompt verbosity and image-contrast effects to be compared jointly. Further qualitative examples illustrating model outputs under different prompt-length and contrast configurations are provided in Fig. 7 and Fig. 8. The results of the central slice variation for the best performing models are summarized further in Table 6.

6.3. Fine-tuning Implementation Details

We fine-tuned MedGemma-4B with QLoRA, updating 20.4M of 4.32B parameters (0.47%) while keeping the pre-trained base weights frozen in 4-bit NF4 format and optimizing LoRA adapters in bfloat16 precision. Optimization used AdamW with cosine learning-rate decay, 3% linear

Table 4. Model-specific input configurations across prompt-length settings. Fields marked with ∅ indicate that no value was passed to the model (i.e., the field was omitted or set to `None`), whereas entries labeled "Not specified" indicate that the field was explicitly provided with that value because it is required by the model interface.

Model	Input configuration
Short prompt	
MedGemma	**System role:** You are an expert radiologist. **Prompt:** Describe the findings for the MRI.
ChatGPT	**System role:** You are an expert radiologist. **Prompt:** Describe the findings for the MRI.
MAIRA-2	**Indication:** Not specified. **Technique:** Sagittal MRI of the lumbar spine. **Comparison:** ∅. **Prior frontal image:** ∅. **Prior report:** ∅.
VILA-M3	**Prompt:** Describe the findings for the MRI. **Mode message:** This is a MRI image. **Expert model cards:** No experts available; answer yourself.
BiomedGPT	**Task:** Caption.
Medium prompt	
MedGemma	**System role:** You are an expert radiologist. **Prompt:** Describe the findings for the sagittal lumbar spine MRI.
ChatGPT	**System role:** You are an expert radiologist. **Prompt:** Describe the findings for the sagittal lumbar spine MRI.
MAIRA-2	**Indication:** Lumbar spine degeneration. **Technique:** Sagittal MRI of the lumbar spine. **Comparison:** ∅. **Prior frontal image:** ∅. **Prior report:** ∅.
VILA-M3	**Prompt:** Describe the findings for the sagittal lumbar spine {`MRI Modality`} MRI. **Mode message:** This is a {`MRI Modality`} MRI image. **Expert model cards:** No experts available; answer yourself.
BiomedGPT	**Task:** Caption.
Long prompt	
MedGemma	**System role:** You are an expert radiologist. **Prompt:** Describe the findings and any radiological gradings for this {`MRI Modality`} lumbar spine MRI in a patient with low back pain.
ChatGPT	**System role:** You are an expert radiologist. **Prompt:** Describe the findings and any radiological gradings for this {`MRI Modality`} lumbar spine MRI in a patient with low back pain.
MAIRA-2	**Indication:** Findings and any radiological gradings in low back pain and lumbar spine degeneration. **Technique:** Single sagittal {`MRI Modality`} MRI of the lumbar spine. **Comparison:** ∅. **Prior frontal image:** ∅. **Prior report:** ∅.
VILA-M3	**Prompt:** Describe the findings and any radiological gradings for this {`MRI Modality`} sagittal lumbar spine MRI in a patient with low back pain. **Mode message:** This is a {`MRI Modality`} MRI image. **Expert model cards:** No experts available; answer yourself.
BiomedGPT	**Task:** Caption.

Table 5. Zero-shot report generation prompt-content ablation on LSMRI (free-text reports) and SPIDER (structured grading reports). Columns are prompt verbosity settings (Short / Medium / Long) with metrics BERTScore F1 (BERT), BLEU, METEOR, and ROUGE-L F1. Rows are MRI contrast inputs (T1w, T2w, T1w+T2w).

LSMRI — free-text reports

		Short				Medium				Long			
Model	Contrast	BERT	BLEU	METEOR	ROUGE-L	BERT	BLEU	METEOR	ROUGE-L	BERT	BLEU	METEOR	ROUGE-L
ChatGPT-5.0	T1w	0.913	0.005	0.226	0.122	0.911	0.005	0.214	0.117	0.901	0.001	0.137	0.074
	T2w	0.913	0.005	0.217	0.120	0.911	0.005	0.213	0.116	0.901	0.001	0.139	0.071
	T1w+T2w	0.913	0.005	0.232	0.129	0.911	0.004	0.215	0.117	0.901	0.001	0.146	0.073
MAIRA-2	T1w	0.894	0.001	0.064	0.071	0.896	0.001	0.070	0.073	0.897	0.002	0.073	0.079
	T2w	0.895	0.002	0.069	0.069	0.897	0.002	0.075	0.075	0.894	0.001	0.068	0.080
	T1w+T2w	0.890	0.001	0.052	0.085	0.890	0.001	0.053	0.085	0.897	0.001	0.082	0.076
MedGemma-4B	T1w	0.908	0.006	0.127	0.127	0.911	0.007	0.142	0.143	0.911	0.008	0.157	0.141
	T2w	0.909	0.007	0.133	0.130	0.911	0.005	0.149	0.152	0.912	0.006	0.159	0.149
	T1w+T2w	0.908	0.005	0.128	0.130	0.912	0.006	0.153	0.159	0.910	0.008	0.169	0.156
MedGemma-27B	T1w	0.918	0.013	0.249	0.155	0.916	0.011	0.227	0.163	0.911	0.007	0.199	0.117
	T2w	0.917	0.012	0.239	0.150	0.916	0.013	0.230	0.164	0.916	0.008	0.230	0.139
	T1w+T2w	0.911	0.010	0.194	0.138	0.915	0.010	0.227	0.167	0.909	0.009	0.203	0.115
VILA-M3 3B	T1w	0.890	0.001	0.043	0.048	0.890	0.000	0.038	0.045	0.889	0.000	0.055	0.048
	T2w	0.891	0.001	0.045	0.049	0.889	0.000	0.051	0.048	0.889	0.000	0.055	0.048
	T1w+T2w	0.895	0.001	0.056	0.065	0.891	0.001	0.059	0.063	0.889	0.000	0.055	0.048
VILA-M3 8B	T1w	0.890	0.001	0.038	0.045	0.907	0.001	0.033	0.053	0.902	0.001	0.038	0.046
	T2w	0.892	0.001	0.035	0.040	0.908	0.000	0.027	0.045	0.902	0.001	0.041	0.047
	T1w+T2w	0.898	0.001	0.033	0.039	0.894	0.000	0.034	0.041	0.898	0.001	0.057	0.055
VILA-M3 13B	T1w	0.874	0.000	0.024	0.039	0.872	0.001	0.024	0.039	0.887	0.000	0.027	0.041
	T2w	0.874	0.000	0.024	0.039	0.873	0.000	0.023	0.039	0.884	0.001	0.029	0.049
	T1w+T2w	0.880	0.000	0.021	0.043	0.886	0.001	0.026	0.048	0.895	0.000	0.041	0.047

SPIDER - structured grading reports

		Short				Medium				Long			
Model	Contrast	BERT	BLEU	METEOR	ROUGE-L	BERT	BLEU	METEOR	ROUGE-L	BERT	BLEU	METEOR	ROUGE-L
ChatGPT-5.0	T1w	0.922	0.009	0.183	0.164	0.922	0.009	0.191	0.166	0.920	0.010	0.235	0.161
	T2w	0.922	0.008	0.191	0.165	0.922	0.009	0.198	0.169	0.919	0.008	0.230	0.155
	T1w+T2w	0.922	0.004	0.192	0.166	0.919	0.006	0.189	0.159	0.918	0.009	0.235	0.155
MAIRA-2	T1w	0.899	0.007	0.084	0.066	0.905	0.009	0.089	0.073	0.904	0.012	0.091	0.075
	T2w	0.899	0.006	0.085	0.064	0.905	0.009	0.087	0.071	0.903	0.011	0.090	0.072
	T1w+T2w	0.897	0.001	0.066	0.052	0.900	0.006	0.079	0.062	0.903	0.008	0.096	0.069
MedGemma-4B	T1w	0.916	0.004	0.099	0.088	0.918	0.004	0.098	0.093	0.918	0.011	0.113	0.101
	T2w	0.917	0.005	0.102	0.092	0.918	0.004	0.096	0.094	0.920	0.010	0.116	0.111
	T1w+T2w	0.914	0.003	0.088	0.086	0.917	0.003	0.088	0.088	0.916	0.011	0.105	0.100
MedGemma-27B	T1w	0.927	0.029	0.202	0.160	0.925	0.026	0.184	0.151	0.928	0.033	0.226	0.154
	T2w	0.927	0.027	0.196	0.159	0.923	0.024	0.172	0.140	0.931	0.041	0.234	0.183
	T1w+T2w	0.918	0.019	0.145	0.122	0.917	0.015	0.135	0.114	0.920	0.022	0.190	0.135
VILA-M3 3B	T1w	0.898	0.009	0.065	0.036	0.899	0.016	0.080	0.050	0.899	0.017	0.085	0.053
	T2w	0.899	0.010	0.068	0.040	0.899	0.017	0.084	0.053	0.899	0.017	0.086	0.054
	T1w+T2w	0.904	0.004	0.048	0.044	0.899	0.013	0.078	0.055	0.899	0.017	0.086	0.054
VILA-M3 8B	T1w	0.899	0.009	0.059	0.037	0.911	0.000	0.025	0.029	0.906	0.000	0.030	0.023
	T2w	0.900	0.007	0.055	0.034	0.914	0.000	0.027	0.037	0.906	0.000	0.029	0.018
	T1w+T2w	0.900	0.000	0.028	0.015	0.902	0.000	0.025	0.011	0.903	0.001	0.042	0.026
VILA-M3 13B	T1w	0.878	0.000	0.017	0.001	0.884	0.002	0.029	0.034	0.883	0.000	0.023	0.016
	T2w	0.878	0.000	0.017	0.000	0.881	0.000	0.026	0.036	0.887	0.001	0.028	0.035
	T1w+T2w	0.879	0.000	0.016	0.009	0.884	0.000	0.016	0.021	0.884	0.000	0.026	0.036

Table 6. Zero-shot report generation performance under different slice selection configurations on LSMRI and SPIDER. Columns are metrics (BERTScore F1, BLEU, METEOR, ROUGE-L F1). Rows are slice index selected for input with respect to the central index.

			LSMRI Dataset (Free Text Report)				SPIDER Dataset (Structured Grading Report)			
Model	Size	Stack	BERTScore F1	BLEU	METEOR	ROUGE-L F1	BERTScore F1	BLEU	METEOR	ROUGE-L F1
MedGemma	4B	Central Slice	0.911	0.005	0.149	0.152	0.918	0.004	0.094	0.092
		±1	0.912	0.007	0.153	0.153	0.919	0.005	0.099	0.096
		±2	0.911	0.007	0.151	0.151	0.919	0.004	0.101	0.098
	27B	Central Slice	0.916	0.013	0.230	0.164	0.922	0.022	0.164	0.135
		±1	0.916	0.013	0.228	0.163	0.923	0.025	0.169	0.137
		±2	0.915	0.011	0.218	0.159	0.924	0.024	0.176	0.140

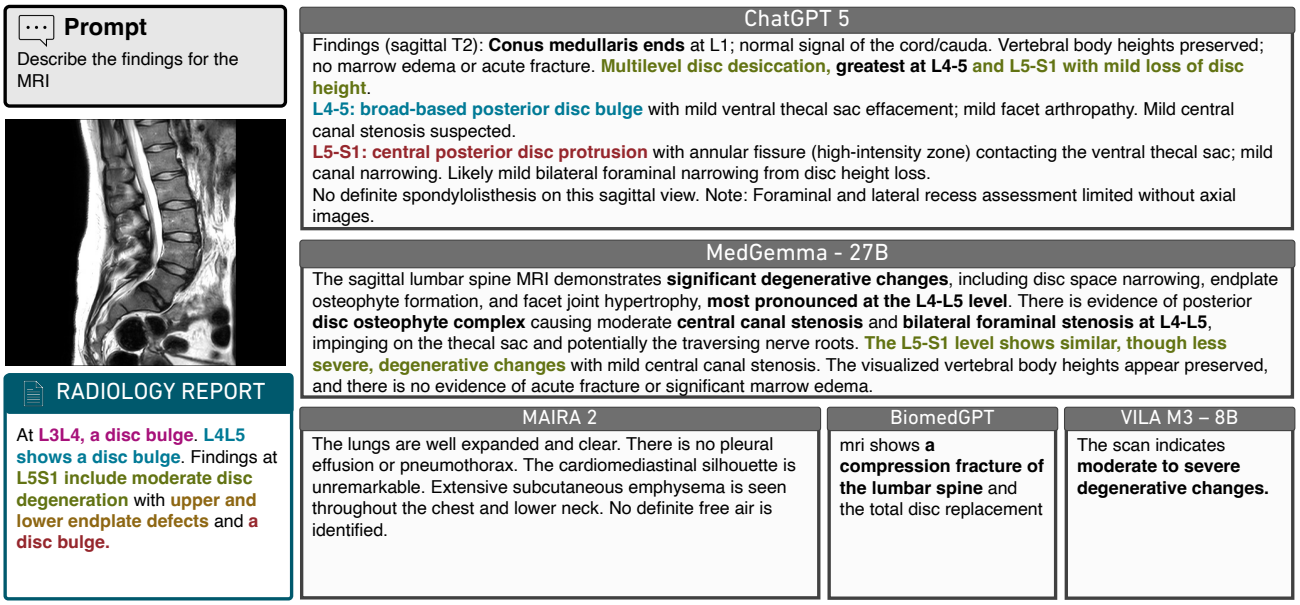


Figure 7. Zero-shot report generation and outputs from identical inputs. Each model (BiomedGPT-Base, ChatGPT-5.0, MAIRA-2, MedGemma-27B, VILA-M3-8B) receives the central sagittal slice, a fixed prompt, and the clinical reference report.

warmup, gradient accumulation of 16 steps, and the selected hyperparameters $lr = 2^{-4}$, $r = 32$, and $\alpha = 64$; models were trained with an 80/20 patient-level train/validation split and selected by best validation loss. Training used SPIDER supervision converted into either grading-derived report text or per-finding level targets, with both T1w and T2w sagittal MRI inputs. Figure 9 shows a representative qualitative example of the fine-tuning strategy.

By *fine-tuning strategy*, we mean both the parameter-efficient adaptation setup and the supervision method used during training. We evaluated several supervision regimes derived from SPIDER, including: separate finding-specific targets (e.g., disc bulging, disc herniation, disc narrowing, endplate defects, Modic changes, Pfirrmann grade, spondylolisthesis), a mixed-finding subset, and full report-style targets. This distinction is important, as the fine-tuned variants differ mainly in their training label space, while the optimization framework remained fixed across experiments.

7. Metric Robustness Analysis for Clinical Application

For the metric robustness analysis, we constructed synthetic report augmentations by applying deterministic, rule-based transformations to the original reference reports. Each augmentation category was designed to isolate a specific lexical or clinical-semantic change while preserving the original report structure whenever possible. All perturbations were applied independently, and each transformed report was compared against its unmodified reference.

7.1. Reference Conditions

Exact Match. The exact match condition was used as the reference corresponding to a normalized score of 100 for each evaluation metric. In this setting, each report was evaluated against itself.

Random Sampling. The random sampling condition was

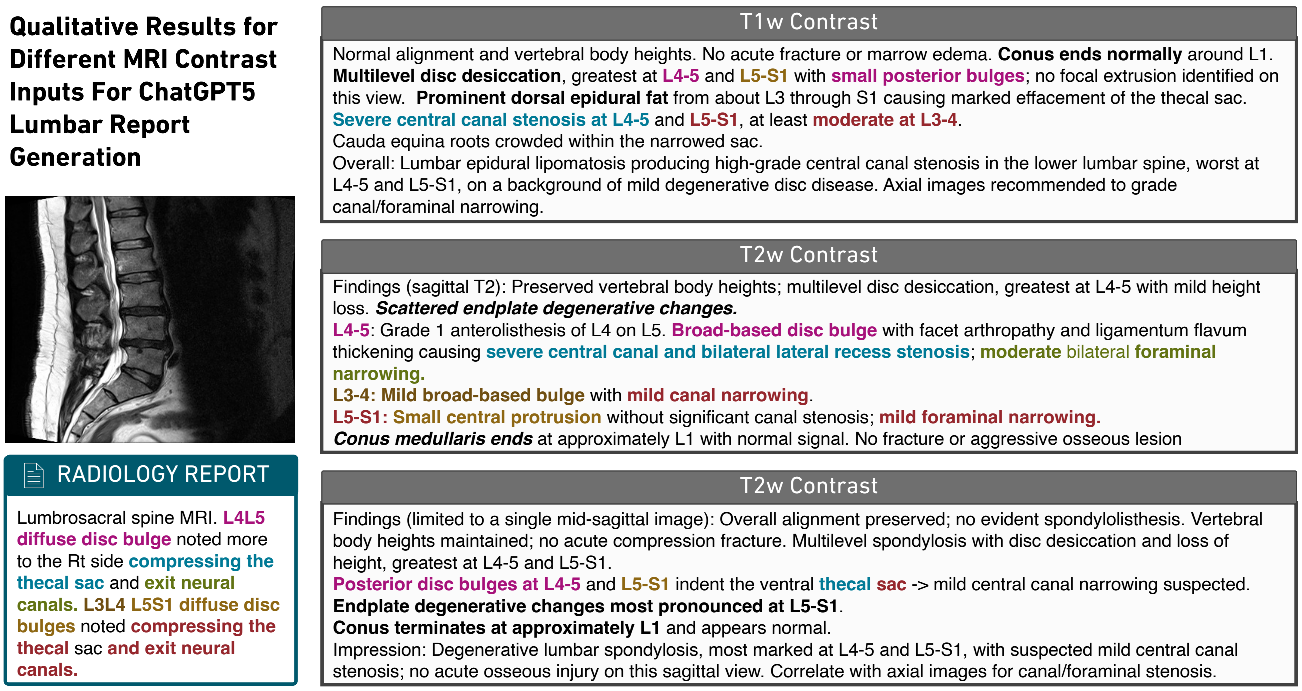


Figure 8. Zero-shot report generation for ChatGPT-5.0 model of the same central slice image for different MRI contrast.

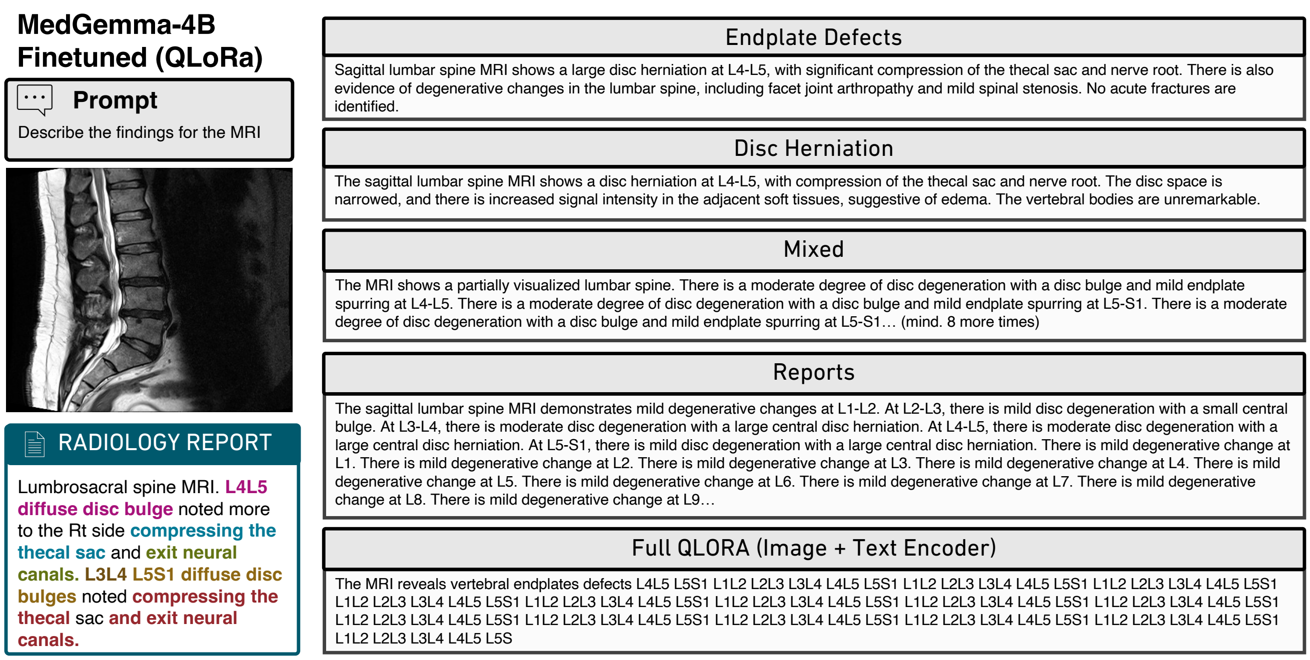


Figure 9. Qualitative comparison of MedGemma-4B outputs after fine-tuning with different SPIDER-derived supervision targets shows that finding-specific, mixed, and report-based supervision produce distinct output structures.

used as the reference corresponding to a normalized score of 0 for each evaluation metric. For this baseline, each report was evaluated against a synthetic report of identical length generated by uniformly sampling words from the empirical word distribution of the full report corpus.

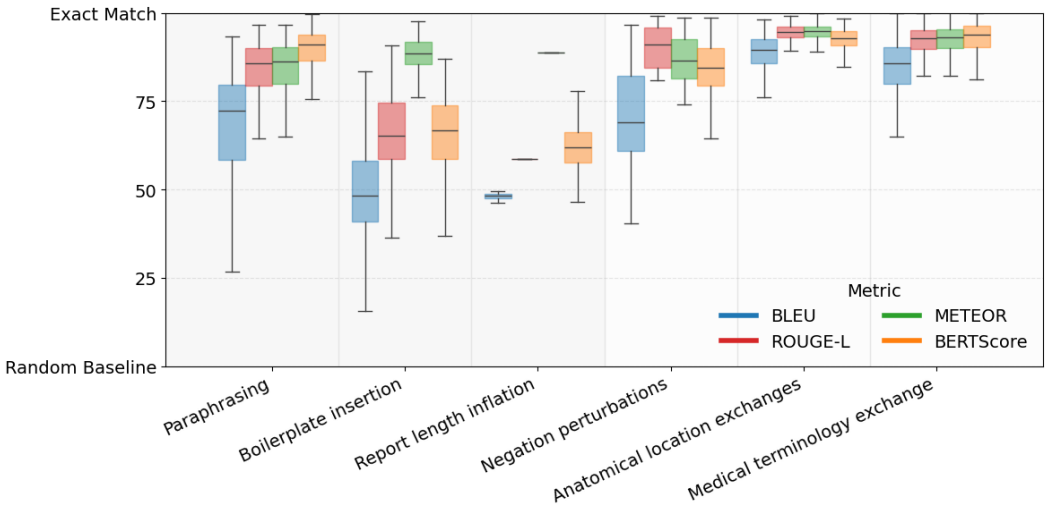


Figure 10. Distribution of normalized report-generation scores under six controlled perturbations to reference reports only. Surface-form changes such as boilerplate insertion and length inflation cause large score shifts, while several clinically important perturbations receive only small penalties.

7.2. Experiment Conditions

Negation Perturbations. Negation perturbations were generated by systematically removing or replacing negation expressions that reverse diagnostic polarity. Common negation tokens such as *no*, *not*, *without*, *neither*, *nor*, *never*, *absence*, and *negative* were removed or substituted using deterministic, rule-based transformations. In cases where simple removal of negation terms resulted in grammatically incorrect or clinically implausible sentences, phrase-level substitutions were applied to ensure syntactic correctness and semantic consistency. These substitutions were defined manually based on frequent patterns observed in the reference reports and were applied using case-insensitive exact matching. Representative examples are shown in Table 7.

Table 7. Representative phrase-level substitutions applied during negation perturbations.

Original phrase	Replacement phrase
no significant thecal sac or nerve root compression noted	significant thecal sac and nerve root compression noted
no disc protrusion or herniation noted	disc protrusion and herniation noted
no disc herniation or protrusion noted	disc protrusion and herniation noted
no evidence of	evidence of
negative for	positive for
no significant	significant

Anatomical Location Exchanges. Anatomical location exchanges were generated by substituting vertebral levels or regional descriptors with plausible but incorrect alternatives while preserving sentence structure. Representative substitutions are listed in Table 8.

Table 8. Representative anatomical location substitutions applied during location exchange perturbations.

Original term	Replacement term
L1–L2	L2–L3
L2–L3	L3–L4
L3–L4	L4–L5
L4–L5	L5–S1
L5–S1	L4–L5
lumbar	thoracic
thoracic	lumbar
sacral	lumbar

Paraphrasing. Paraphrasing perturbations were generated using lexical substitutions that preserve semantic meaning while altering surface form. Representative examples are listed in Table 9.

Medical Terminology Exchange. Medical terminology exchange perturbations were generated by replacing semantically equivalent medical terms commonly used interchangeably in radiology reports. These substitutions preserve diagnostic meaning while altering surface form. Representative examples are listed in Table 10.

Boilerplate Insertion. Boilerplate insertion perturbations

Table 9. Representative paraphrasing substitutions applied during paraphrasing perturbations.

Original term or phrase	Replacement term or phrase
maintained	preserved
noted	observed
small	minor
mild	slight
compatible with	consistent with
various	several
largely	mostly

Table 10. Representative medical terminology substitutions applied during terminology exchange perturbations.

Original term	Replacement term
disc bulge	disc herniation
disc herniation	disc bulge
stenosis	narrowing
narrowing	stenosis
neural foraminal narrowing	foraminal stenosis
spondylolisthesis	vertebral slippage
vertebral slippage	spondylolisthesis

were generated by inserting standardized preambles or closing statements commonly produced by large language models. These insertions do not replace existing content but introduce additional templated structure and verbosity. Representative examples are shown in Table 11.

Table 11. Representative boilerplate phrases inserted during boilerplate insertion perturbations.

Insertion position	Inserted phrase
Prefix	Certainly, here is the report based on the MRI image:
Prefix	Below is the generated report based on the MRI scan:
Suffix	This concludes the MRI report.
Suffix	No additional significant abnormalities are identified.

Report Length Inflation. Report length inflation was generated by duplicating each report in its entirety and concatenating the two copies, resulting in a report containing the same words twice in the original order. This augmentation preserves lexical content and semantic meaning while increasing verbosity and repetition.

7.3. Metric Behavior on Perturbed Reference Reports

To isolate metric behavior from model performance, this analysis was performed on the reference reports alone. Each experiment compares an original report with a deterministically perturbed version of the same report, without using model-generated text. Figure 10 summarizes the resulting score distributions across perturbation types and highlights the differing sensitivity of standard metrics to lexical versus clinically meaningful changes.

7.4. IVD-level Binary Classification of Degenerative Findings

We derived per-level classification targets from the structured radiological grading matrices provided in SPIDER and LumbarDISC. For each finding, the target output consists of the set of intervertebral disc (IVD) levels at which the abnormality is present. Pfirrmann grading and vertebral levels outside the T12–S1 range were excluded from this task. The same label-generation procedure was applied to LumbarDISC, restricted to spinal canal stenosis, which was the only finding used in our per-level evaluation.

Given the radiological gradings (Tab. 12), we generate one target string per finding by enumerating all positive IVD levels. For example:

- **Disc narrowing**: `T12L1 L1L2 L2L3 L3L4 L5S1`
- **Spondylolisthesis**: `None`
- **Disc bulging**: `L2L3 L3L4 L4L5 L5S1`

Table 12. Example SPIDER radiological grading matrix. Binary values denote absence (0) or presence (1), except endplate defects, which encode location: 0 = absent, 1 = upper endplate only, 2 = lower endplate only, 3 = both endplates.

Finding	T12–L1	L1–L2	L2–L3	L3–L4	L4–L5	L5–S1
Disc narrowing	1	1	1	1	0	1
Spondylolisthesis	0	0	0	0	0	0
Endplate defects	2	0	0	0	0	0
Disc bulging	0	0	1	1	1	1
Disc herniation	0	0	0	0	0	0
Modic changes	0	0	0	0	0	0

Each finding-specific label was independently predicted using the following standardized prompt:

```
Target finding: {FINDING}.
Valid spinal levels to choose from (use ONLY
these, and only if clearly present in the image):
{LEVELS}.


OUTPUT RULES (STRICT):
- If the target finding is present: output
ONLY the matching levels from the list above,
separated by single spaces (e.g., L3L4 L4L5).
- If the target finding is NOT present (or is
ambiguous/uncertain): output EXACTLY None.
- Use UPPERCASE exactly as shown. No extra text,
punctuation, labels, or newlines. One line only.


Now examine the MRI and output either the
space-separated levels or None.
```

8. Anomaly-Guided Report Generation

8.1. Weakly-Supervised Heatmap Generation

Because pixel-level pathology annotations are unavailable for lumbar spine MRI at scale, we derive weakly supervised spatial targets from the structured per-IVD grading labels provided by the SPIDER dataset. For each annotated finding and each positive IVD level, we place an anatomically motivated coarse region within the corresponding channel of the six-channel heatmap tensor $\mathbf{H} \in \mathbb{R}^{6\times H\times W}$, following the expected spatial distribution of that pathology. Specifically:

- **Disc bulge / herniation:** shallow posterior caps at the posterior disc margin, clipped to retain only posterior-facing support, modeling spinal canal protrusion.
- **Spinal canal stenosis:** larger posterior regions extending toward the expected canal location.
- **Disc narrowing:** contracted quadrilateral regions spanning the entire disc space.
- **Endplate defects:** thin bands along the superior or inferior endplate of adjacent vertebral bodies, coded by location (upper, lower, or both).
- **Modic changes:** trapezoidal regions extending from the endplates into the subchondral vertebral marrow.
- **Spondylolisthesis:** a continuous posterior displacement trace across the disc space and adjacent vertebral bodies.

When multiple findings co-occur at the same IVD level, their finding heatmaps are combined by a per-pixel maximum to prevent artificial intensity accumulation. All targets are modeled as soft spatial distributions using anisotropic Gaussian smoothing aligned with the estimated disc orientation, allowing for imprecise landmark estimates from SpineNetV2 [44]. Per-finding and per-channel normalization balances supervision across findings with different spatial extents and disc levels with different prevalences.

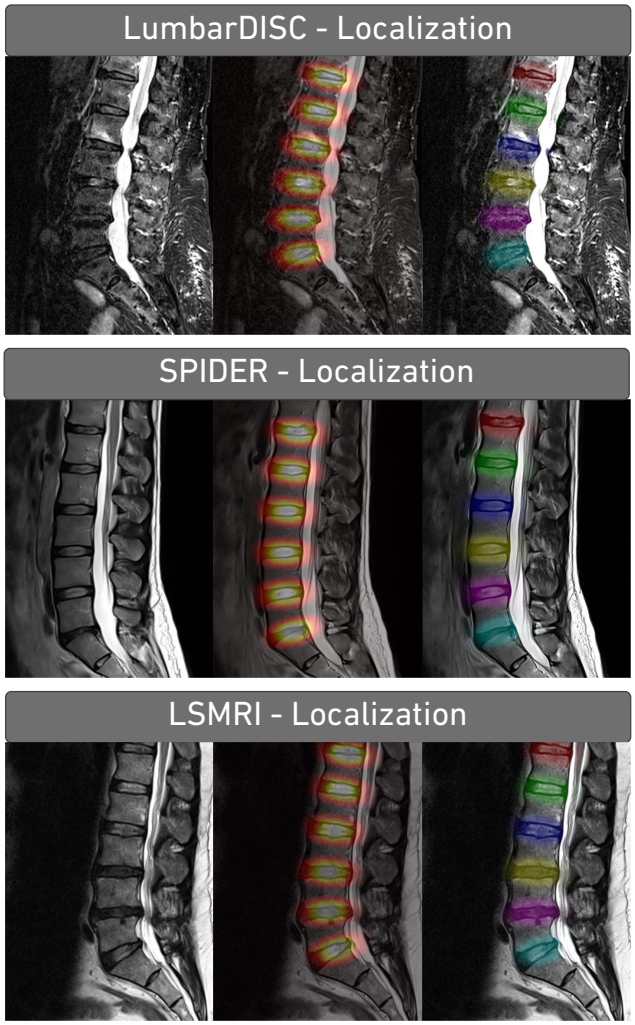


Figure 11. **Localization heatmaps test examples of each dataset.** *Left*: central sagittal T2-weighted MRI slice. *Center*: ground truth six-channel localization heatmap. *Right*: predicted alpha-blended overlay of localization heatmaps.

8.2. Anomaly Detection Training

Anomaly detection training is done in two phases, localization and anomaly detection. In the first phase, the model is pre-trained exclusively on disc-level *localization* heatmaps $\mathbf{H}^{\text{loc}}$ (coarse quadrilateral disc regions) on 85% of LumbarDISC and 30% of LSMRI to establish level-consistent disc localization. In the second phase, supervision is gradually shifted toward *anomaly* heatmaps $\mathbf{H}^{\text{AD}}$ via linear interpolation:

$$\mathbf{H}^{t} = (1-t)\,\mathbf{H}^{\text{loc}} + t\,\mathbf{H}^{\text{AD}}, \quad t \in [0,1], \tag{4}$$

where t increases monotonically over training epochs. A channel-exclusivity penalty $\mathcal{L}_{\text{IVD}}$ is ramped up concurrently to discourage co-activation across IVD channels.

8.2.1. Training data splits

Localization Pre-Training. The U-Net is pre-trained using disc-level localization heatmaps $\mathbf{H}^{\text{loc}}$ on 85% of LumbarDISC ($n \approx 1{,}679$) and 30% of LSMRI ($n \approx 155$), for 351 epochs with a fixed learning rate of 10^{-3}. **Anomaly**

Fine-Tuning. The model is fine-tuned using pathology-specific heatmaps $\mathbf{H}^{\text{AD}}$ derived from SPIDER ($n = 218$, 80/10/10 patient-level train/val/test split) for 501 epochs, with the supervision schedule of Eq. (2) ramping t linearly from 0 to 1 over the first 1,000 training steps. **Evaluation..**

The 10% SPIDER test split (held out from both stages) is used for all anomaly detection evaluation. The 30% LSMRI subset used in pre-training is disjoint from the 70% LSMRI held-out split used to evaluate downstream report generation, ensuring no data leakage between the anomaly detector and the report generation assessment.

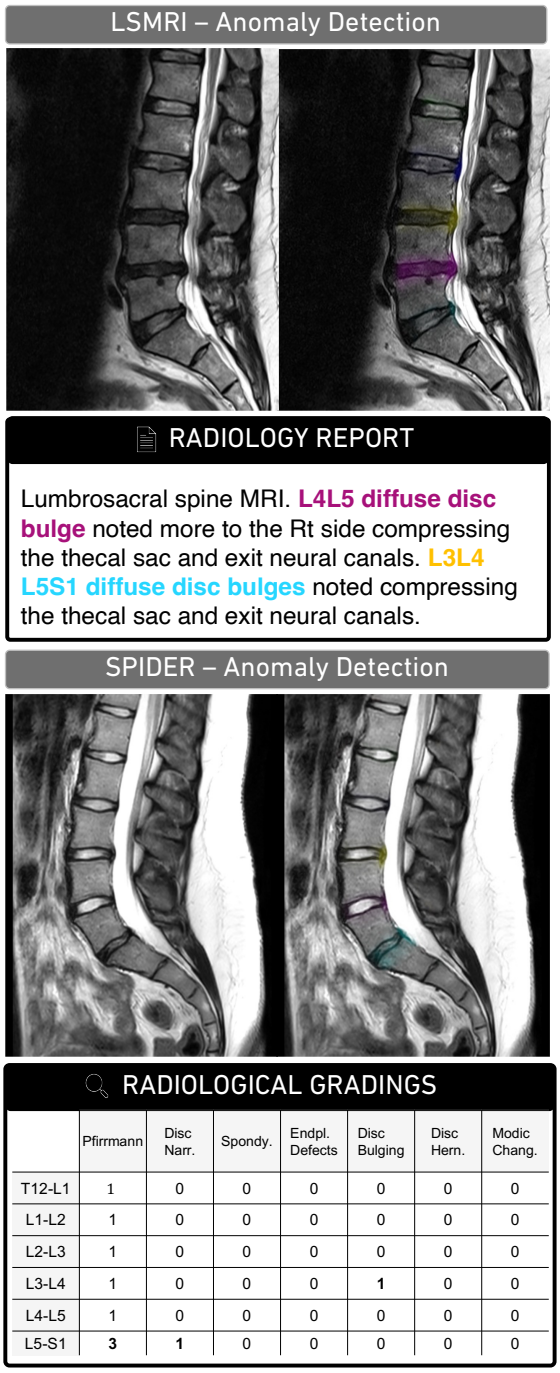
Data Augmentations. The medium augmentation regime was selected after a three-way ablation (low/medium/high), as it reduced the intensity-distribution mismatch between best- and worst-performing validation cases while avoiding the variance increase and performance degradation observed under aggressive augmentation (overall DICE: low 90.6, medium 90.2, high 87.8 on LSMRI/LumbarDISC localization evaluation).

8.2.2. Heatmaps Regression Results

Table 13. Heatmap accuracy by dataset and training stage (medium augmentation, mean $\pm$ SD). DICE after thresholding at 0.1; RMSE on $[0, 1]$-normalised continuous values.

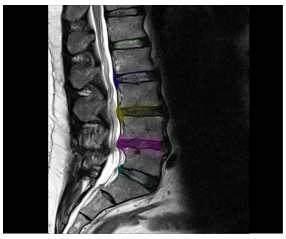
Dataset	Stage	DICE	RMSE
LumbarDISC + LSMRI	Localization	90.2 ± 1.0	0.6 ± 0.2
SPIDER (test, pos. pix)	Anomaly	82.3 ± 2.0	0.4 ± 0.1

Table 13 summarizes the numerical heatmap results in datasets and training stages. The drop in DICE between Phase 1 and Phase 2 (90.2 $\rightarrow$ 82.3 overall) reflects the transition from broad disc-region targets to fine-grained pathology-specific subregions restricting evaluation to positive ground-truth pixels. Performance is lowest at upper levels (T12–L1, DICE 51.8 ± 9.2), consistent with a lower prevalence of pathology and a lower density of supervision at those levels in the SPIDER data set. Figure 11 show qualitative localization heatmaps and Figure 12 shows heatmap overlay examples alongside the corresponding ground-truth annotations (clinical report for LSMRI; radiological grading for SPIDER), confirming anatomically consistent localization across IVD levels.



	Pfirrmann	Disc Narr.	Spondy.	Endpl. Defects	Disc Bulging	Disc Hern.	Modic Chang.
T12-L1	1	0	0	0	0	0	0
L1-L2	1	0	0	0	0	0	0
L2-L3	1	0	0	0	0	0	0
L3-L4	1	0	0	0	**1**	0	0
L4-L5	1	0	0	0	0	0	0
L5-S1	**3**	**1**	0	0	0	0	0

Figure 12. **Qualitative anomaly detection results on LSMRI (top) and SPIDER (bottom).** Predicted heatmaps overlaid on the central sagittal T2w slice with corresponding ground-truth annotations. *LSMRI:* diffuse disc bulges at L3–L4 through L5–S1 (reference report); activation concentrates at lower lumbar levels. *SPIDER:* disc bulging at L3–L4 and disc narrowing (Pfirrmann 3) at L5–S1 (grading table); heatmap activates selectively at those levels, sparing upper IVDs.



Model
ChatGPT-5.0
Prompt
Describe the findings for the MRI
Input
+Image
+ Anomaly Heatmap

REFERECE RADIOLOGY REPORT

Lumbrosacral spine MRI. **L4L5 diffuse disc bulge** noted more to the Rt side **compressing the thecal sac** and **exit neural canals. L3L4 L5S1 diffuse disc bulges** noted **compressing the thecal** sac **and exit neural canals.**

IMAGE ONLY

Findings (sagittal T2): Preserved vertebral body heights; multilevel disc desiccation, greatest at L4-5 with mild height loss. ***Scattered endplate degenerative changes.***
L4-5: Grade 1 anterolisthesis of L4 on L5. **Broad-based disc bulge** with facet arthropathy and ligamentum flavum thickening causing **severe central canal and bilateral lateral recess stenosis**; **moderate** bilateral **foraminal narrowing.**
L3-4: Mild broad-based bulge with **mild canal narrowing**.
L5-S1: Small central protrusion without significant canal stenosis; **mild foraminal narrowing.**
Conus medullaris ends at approximately L1 with normal signal. No fracture or aggressive osseous lesion

IMAGE + ANOMALY DETECTION HEATMAP

Normal conus terminating at L1. Preserved lumbar lordosis with minimal retrolisthesis of L4 on L5. **Multilevel degenerative disc disease** with scattered endplate irregularity/Schmorl nodes. L5-S1: advanced disc desiccation and **height loss** with adjacent **Modic change**; shallow posterior disc osteophyte complex without significant canal compromise. **L4-L5**: moderate disc height loss with **broad-based posterior bulge** producing **moderate central canal stenosis**. **L3-L4**: mild degeneration with **small posterior bulge** causing **mild canal narrowing**. **L2-L3**: tiny posterior protrusion without stenosis. No acute fracture or high-grade spondylolisthesis.

Figure 13. **Qualitative report generation results with and without anomaly detection (AD) heatmap input (ChatGPT-5.0, LSMRI).** The same sagittal T2w slice is shown with (*img,+,AD*) and without (*img only*) the predicted heatmap, alongside the clinical reference. With the heatmap, the model correctly identifies Grade 1 anterolisthesis at L4–L5, a broad-based disc bulge with severe central canal stenosis, and additional findings at L3–L4 and L5–S1, closely matching the reference. Without it, the report is clinically plausible but omits the anterolisthesis and underspecifies the severity and distribution of stenosis

8.3. Anomaly-Guided Report Generation Results

This section provides extended evaluation of the anomaly detection (AD) module and its integration into report generation. The AD module is trained with weak supervision from structured per-IVD grading labels (SPIDER), producing six-channel spatial heatmaps that encode anatomically plausible pathological regions without dense annotation. These heatmaps serve dual purposes: as standalone interpretability outputs for direct clinical inspection, and as auxiliary visual input to ground VLM report generation spatially.

Three input conditions are evaluated: the central sagittal slice alone (*img*), the slice augmented with the predicted heatmap overlay (*img + AD*), and an *AD-guided* condition in which a structured finding summary extracted from the heatmap is passed as textual context — introduced for models lacking native multi-image conditioning (MedGemma) to preserve the grounding signal within a single-image interface. Table 14 reports the impact of heatmap integration on zero-shot report generation across all evaluated models. ChatGPT-5.0 shows consistent improvements across all metrics; MedGemma benefits more reliably from *AD-guided* than from direct overlay; and VILA-M3 shows mixed but generally positive trends at larger model sizes. Figure 13 provides a qualitative example illustrating how heatmap overlay promotes more anatomically specific, level-grounded reports compared to the image-only baseline.

Table 14. Impact of anomaly detection (AD) heatmap overlay on zero-shot report generation on the held-out 70% LSMRI test split Models are evaluated under three input conditions: sagittal slice alone (*img*), slice augmented with the predicted disc-level heatmap overlay (*img + AD*), and AD-derived structured finding summary passed as textual context (*AD-guided*), the latter introduced for models without native multi-image conditioning. Performance is reported using BERTScore F1, BLEU, ROUGE-L, and METEOR (higher is better on 0-100 normalized sxale).

Model	Size	Input	BERTScore	METEOR	BLEU	ROUGE-L
ChatGPT	5.0	img	91.14	21.58	5.45	11.84
		img + AD	**91.60**	**23.89**	**13.41**	**23.89**
VILA-M3	3B	img	**88.89**	**5.09**	0.39	**4.72**
		img + AD	88.43	2.91	**0.71**	4.48
	8B	img	**90.78**	2.79	0.47	4.61
		img + AD	90.05	**6.49**	**0.67**	**6.92**
	13B	img	87.25	2.35	0.19	3.84
		img + AD	**88.39**	**5.41**	**0.23**	**4.09**
MedGemma	4B	img	91.09	14.76	5.03	15.12
		img + AD	90.50	9.75	2.90	12.44
		AD-guided	**91.69**	**18.28**	**12.50**	**17.27**
	27B	img	91.58	**22.59**	**11.86**	16.32
		img + AD	91.05	16.96	6.76	13.37
		AD-guided	**91.66**	20.11	10.96	**16.41**