This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Material Diagnosis-Aware Multimodal Retrieval for Paired CBCT Scans and Clinical Notes

Author One[1], Author Two[2], Author Three[2], Author Four[2], and Author Five[1,3]

[1] Affiliation One
[2] Affiliation Two
[3] Affiliation Three
[4] Affiliation Four
[5] Affiliation Five

This Electronic Supplementary Material (ESM) accompanies the paper of the same title, accepted at the ODIN Workshop, MICCAI 2026. Section, table, and equation numbers below are prefixed with "S" and are additional to those in the main paper; references to the main paper's own numbering (e.g. Eq. 1, Table 2) refer to the accepted manuscript.

S1 Cross-Modal Alignment Score — Formal Definition

The alignment score reported in Table 2 and Fig. 2 of the main paper is the mean cosine similarity between paired test-set image and text embeddings:

$$\text{alignment} = \frac{1}{n}\sum_{i=1}^{n} \cos\left(\mathbf{z}_i^{\text{img}}, \mathbf{z}_i^{\text{txt}}\right) \tag{S1}$$

where $\mathbf{z}_i^{\text{img}}$ and $\mathbf{z}_i^{\text{txt}}$ are the ℓ_2-normalized image and text embeddings of test patient i, and n is the number of labeled test patients ($n = 101$).

S2 Diagnosis-Only Trivial Baseline

To isolate how much of the reranked retrieval performance (Table 2, `+RR` rows) is attributable to the diagnosis-overlap reranking step itself, as opposed to the learned multimodal representation, we evaluate a trivial baseline that uses **no embeddings and no learned representation**: candidates are ranked purely by Jaccard diagnosis-set overlap with the query (Eq. 2 in the main paper), with ties broken by a fixed random seed.

Under the diagnosis-overlap relevance definition (Eq. 2), this trivial baseline saturates at a perfect score. This indicates that, given the diagnosis vocabulary size and candidate pool of this dataset, ranking by diagnosis overlap alone is close to sufficient for top-1 correctness. We report this explicitly so that the reranked results in Table 2 (rows marked `+RR`) are interpreted as a setting in which diagnosis codes are already available at query time, rather than as a measure of learned representation quality in isolation. Non-reranked results (rows without `+RR`) remain the primary evidence for the quality of the learned embedding space.

Table S1. Diagnosis-only trivial baseline, $N = 101$ test queries.

Metric	Value
R@1	1.000
R@3	1.000
R@5	1.000
R@10	1.000
MRR	1.000
MAP	1.000
NDCG@5	1.000

Why this saturates: diagnosis-count distribution. Fig. S1 shows the number of diagnosis codes per patient across the full cohort (660 patients). Of the 555 patients carrying at least one diagnosis code, 287 (52%) have exactly one, and 1–2 diagnoses together account for 79% of diagnosed patients. Additionally, several clinically distinct subtypes are merged into a single label for modeling purposes (e.g. acute pulpitis, chronic pulpitis, and pulp necrosis are all mapped to "Pulpitis"; deep caries is mapped to "Caries"). Given a sparse, coarse-grained diagnosis taxonomy and a 245-patient candidate pool, Jaccard overlap between a query and at least one candidate is close to guaranteed regardless of which — or whether any — embedding is used to propose candidates before reranking. We report this distribution for transparency and note that a finer-grained diagnosis taxonomy, or a larger and more diagnostically diverse candidate pool, would likely reduce this ceiling effect.

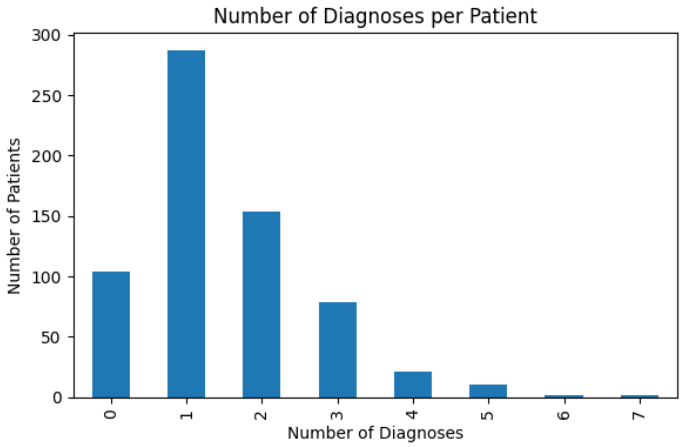


Fig. S1. Number of diagnosis codes per patient, full cohort ($N = 660$). Bars: 0 diagnoses (unlabeled) ≈ 105; 1 ≈ 287; 2 ≈ 154; 3 ≈ 78; 4 ≈ 22; 5 ≈ 11; 6 ≈ 2; 7 ≈ 2.

S3 Lexical Baselines With and Without Diagnosis-Overlap Reranking

We extend the text-only baselines of Table 4 (main paper) by applying the same diagnosis-overlap reranking formula used for our fine-tuned model (combined = $0.6 \cdot \text{dense_score} + 0.4 \cdot \text{Jaccard_overlap}$) to BM25 and TF-IDF, so that lexical baselines are compared under identical conditions to our reranked results.

Table S2. BM25 and TF-IDF, text-to-text retrieval, raw vs. reranked (+RR), $N = 101$.

Method	R@1	R@5	MRR	NDCG@5
BM25	0.812	0.980	0.880	0.942
BM25 +RR	0.812	0.980	0.882	0.944
TF-IDF	0.822	0.960	0.884	0.940
TF-IDF +RR	**0.990**	**1.000**	**0.993**	**0.997**

TF-IDF+RR (R@1 = 0.990) closely approaches our fine-tuned model's reranked text-to-text result (Table 2, T→T +RR: R@1 = 0.980), using only lexical term matching plus diagnosis-overlap reranking. This is consistent with the finding in Sect. S2 above and is offered as further evidence that the diagnosis-overlap reranking step, not the learned embedding, is the dominant contributor to the near-ceiling reranked scores in Table 2.

Note: BM25's raw relevance scores are unbounded, unlike TF-IDF cosine similarity $\in [0, 1]$*; under the fixed-weight reranking formula above this numerically suppresses the diagnosis-overlap term's influence for BM25 specifically, which is why BM25+RR shows negligible change from BM25 alone. Per-query score normalization would be required for a strictly fair BM25+RR comparison; we report the result as obtained and flag this as an implementation detail rather than a substantive finding.*

S4 Zero-Shot Baseline Under Identical Reranking

Table 2 in the main paper applies diagnosis-overlap reranking to the fine-tuned model's text-query modes. For completeness, we report the zero-shot BioMedCLIP baseline (Table 4, main paper) under the identical reranking procedure.

Consistent with Sects. S2 and S3, the reranked baseline matches or exceeds the reranked fine-tuned model on both text-query modes. The clearest and most consistent gain attributable to the diagnosis-aware supervised contrastive fine-tuning objective is in the non-reranked image-to-text mode (img2txt: +0.238 R@1 in this run; +0.109 to +0.119 across independent seeded reproductions, see Sect. S6), which never receives diagnosis-overlap reranking and is therefore a cleaner measure of representation quality.

Table S3. Baseline (zero-shot BioMedCLIP) vs. Finetuned, all four modes, reranking applied identically to both where applicable (txt2img, txt2txt).

Mode	Baseline R@1	Finetuned R@1	Δ R@1
I→I/img2img	0.436	0.416	−0.020
I→T/img2txt	0.257	0.495	+0.238
T→I/txt2img (+RR)	0.970	0.970	0.000
T→T/txt2txt (+RR)	1.000	0.980	−0.020

S5 Informativeness-Weighted vs. Uniform Mean Pooling — Extended Ablation

Table 3 (main paper) compares whole-scan informativeness-weighted pooling (Eq. 1) against FDI lesion-quadrant cropping and a whole-scan/lesion blend, but not against the simpler control of uniform mean pooling over the same selected slices. We report this control here, run twice with independent seeds to assess stability.

Both runs use identical slice selection (same intensity-std threshold, same kept slices, same $N = 20$ subsampling) — only the pooling weight differs: $w_k = s_k / \sum_j s_j$ (informativeness-weighted) vs. $w_k = 1/N$ (uniform).

Table S4. Informativeness-weighted ("std") vs. uniform mean pooling, R@1, two independently seeded full retrains, with bootstrap 95% confidence intervals (2000 resamples, $N = 101$).

Mode	Run A: std / uniform	Run A 95% CI (diff)	Run B: std / uniform	Run B 95% CI (diff)
I→I/img2img	0.416 / 0.406	$(-0.020, 0.040)$	0.426 / 0.416	$(-0.020, 0.050)$
I→T/img2txt	0.366 / 0.376	$(-0.040, 0.020)$	0.396 / 0.366	$(-0.040, 0.099)$
T→I/txt2img (+RR)	0.970 / 0.960	$(0.000, 0.030)$	0.990 / 0.960	$(0.000, 0.069)$
T→T/txt2txt (+RR)	0.980 / 0.980	$(0.000, 0.000)$	0.980 / 0.980	$(-0.030, 0.030)$

In both independent runs, no retrieval mode shows a statistically significant difference between the two pooling schemes (all 95% CIs on the paired difference include zero). We conclude that informativeness weighting's contribution to retrieval quality, isolated from slice selection, is not distinguishable from sampling noise at this dataset size ($N = 101$), and revise our interpretation of this component of the method accordingly relative to the main paper's original framing.

S6 Reproducibility Notes

Seed variance. We observed that fixing a training seed (Python `random`, NumPy, and PyTorch, including `torch.backends.cudnn.deterministic = True`) reduces but does not fully eliminate run-to-run variance in our current training pipeline: two runs with an identical seed produced R@1 differing by up to 0.03 on some modes (Table S4, Run A vs. Run B). We attribute residual variance to sources not covered by the seeding above — most plausibly hash-randomized iteration order over Python `set` objects used in diagnosis-label positive-pair construction, and/or non-deterministic CUDA reduction kernels not addressed by `cudnn.deterministic` alone. We report results across two independent runs throughout this supplement rather than a single point estimate, and recommend `PYTHONHASHSEED` fixation and `torch.use_deterministic_algorithms(True)` for readers seeking exact reproduction.

S7 Dataset Provenance and CBCT Acquisition Details

For transparency, we summarize the CBCT acquisition details of MMDental [1] relevant to interpreting our results, drawn directly from the dataset's data descriptor.

Single-center collection. All 660 patients were recruited at a single site, Hangzhou Dental Hospital. As the dataset's own authors note, this limits generalizability of any model trained on it, including ours, to other patient populations, imaging protocols, or institutions. This is a limitation of the underlying data resource, not specific to our method, and is already stated in the main paper's Discussion (Sect. 6).

Two CBCT scanners Scans were acquired on two devices (HiRes 3D-Plus; Oral and Maxillofacial CBCT unit) sharing identical imaging geometry ($640\times640\times400$ voxels, 0.25 mm slice thickness), though exposure settings (kV/mA) varied by machine and patient demographic. Since our informativeness score (Eq. 1, main paper) is based on intensity standard deviation within a fixed Hounsfield-like window ($[-1000, 2500]$), we checked whether this exposure variation could confound slice selection. We verified that (i) scan geometry and orientation required no additional alignment across scanners, and (ii) intensity characteristics within our fixed window are consistent across both machines. We found no evidence of a measurable confound from acquisition exposure in our pipeline.

Machine-translated clinical text. Clinical notes in MMDental were originally recorded in Chinese and machine-translated for the released dataset (main paper, Sect. 3). This is an existing, previously-acknowledged limitation of our text pipeline and is unrelated to the CBCT acquisition points above; we restate it here alongside the imaging provenance details for a single, consolidated transparency section.

References

1. Wang, C., Zhang, Y., Wu, C., Liu, J., Huang, X., Wu, L., Wang, Y., Feng, X., Lu, Y., Wang, Y.: MMDental – a multimodal dataset of tooth CBCT images with expert medical records. Scientific Data **12**, 1172 (2025). https://doi.org/10.1038/s41597-025-05398-7