

1 INTRODUCTION

At a glance: we teach a vision-language model to retrieve clinically similar past CBCT scans and notes, not just matched patients but by softly rewarding shared diagnoses during training. Text-anchored search works reliably; image-anchored and diagnosis-overlap reranking evaluation can make results look better than the embeddings really are.

Clinical Problem

Cone-beam CT (CBCT) is central to dental & maxillofacial diagnosis, but the resulting 3D volumes usually read in isolation from the clinical notes recorded alongside them namely history, chief complaint, diagnosis, treatment. Retrieving clinically similar prior cases, combining 3D imaging and text, would support diagnosis, treatment planning and patient education aiding in better treatment outcomes.

Challenges

- Text-only search ignores radiographic content.
- Image-only search overlooks clinical documentation.
- Standard CLIP-style objectives pair each scan only with its own note, treating every patient as a negative yet diagnostically related patients from different visits never cluster together.

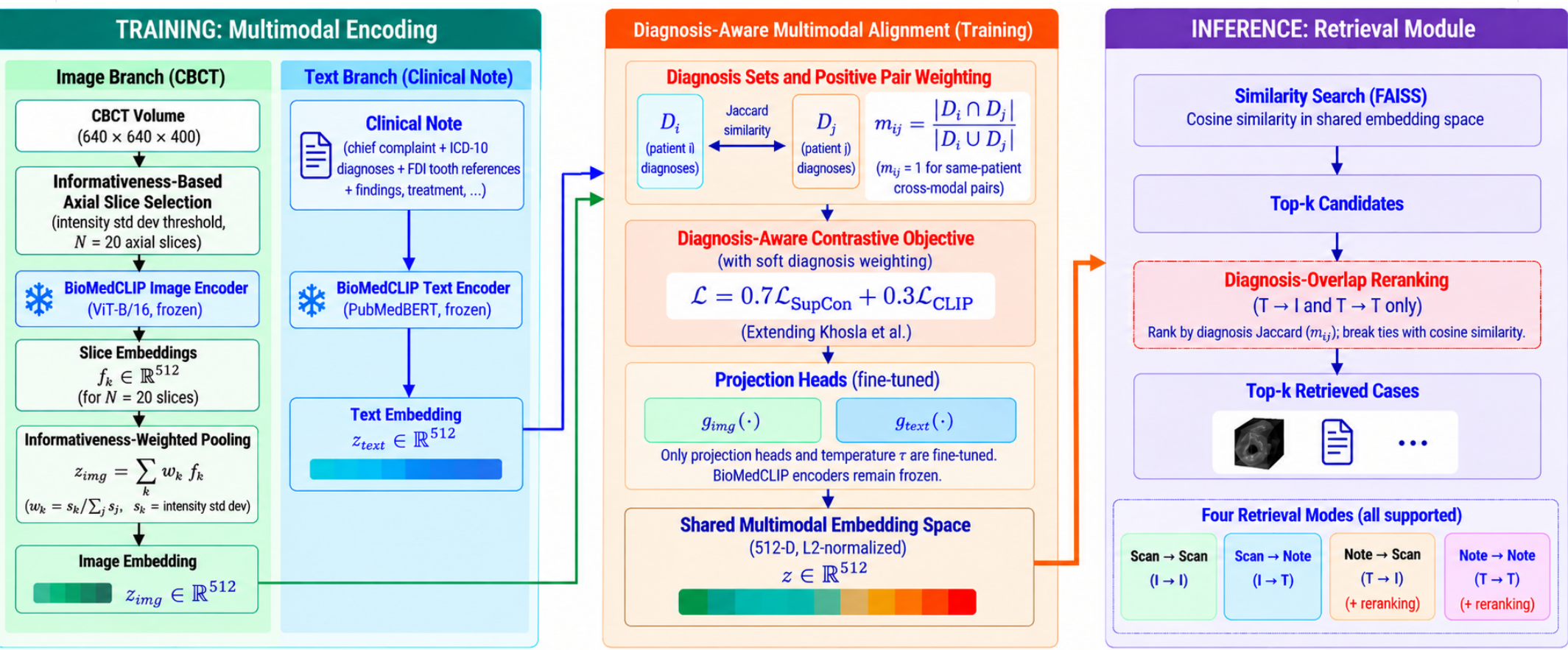
2 OBJECTIVES & CONTRIBUTIONS

Our Contributions

- A multimodal retrieval framework for paired 3D CBCT + clinical notes on a shared BioMedCLIP embedding space, with a standardised benchmark on MMDental (splits, preprocessing, four-mode evaluation).
- A diagnosis-aware cross-modal supervised-contrastive objective with soft, Jaccard-weighted positive supervision, clustering clinically related patients beyond same-patient pairing.
- A rigorous evaluation against lexical, sparse and vision-language baselines, an image-encoding ablation, and an elaborate audit of what the reranked numbers actually measure.

3 FRAMEWORK OVERVIEW

CBCT volumes and clinical notes are encoded into a diagnosis-aware shared embedding space, then indexed with FAISS for four-mode retrieval.



Only the projection heads and temperature are fine-tuned, while the BioMedCLIP encoders remain frozen. Diagnosis-overlap supervision encourages clinically related patients to align across modalities, while preserving same-patient image-text correspondence. The resulting 512-D, L2-normalized shared embedding space is indexed with FAISS to support I→I, I→T, T→I and T→T retrieval. Informativeness-weighted pooling aggregates selected axial-slice embeddings into a patient-level CBCT representation.

Diagnosis-overlap reranking is applied only to text-query modes (T→I and T→T) and is kept separate from the learned embedding representation.

4 DATASET — MMDENTAL

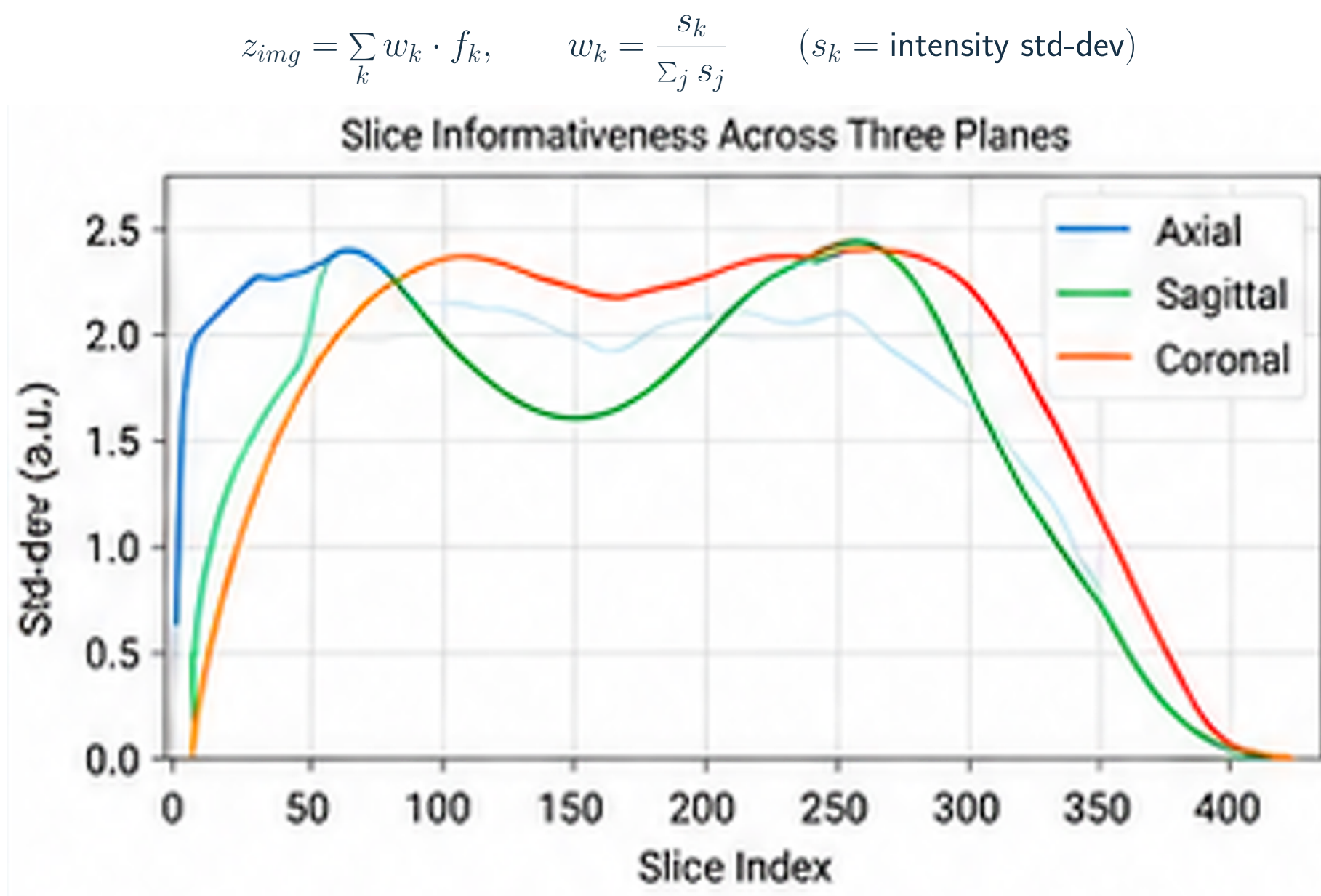
660 total patients
403 CBCT + clinical records
2,124 visit rows (3.22 / patient)
245 / 101 retrieval DB / queries

3D CBCT: 640×640×400 voxels, 0.25 mm slices, fixed [-1000, 2500] HU-like window. Records: ICD-10 diagnoses + FDI tooth numbers; FDI codes expanded to explicit anatomy.

Disease prevalence is highly imbalanced (multi-label). Bars show clinically significant dental disease prevalence.

5 REPRESENTATION LEARNING

We encode 3D CBCT volumes and clinical notes into a shared embedding space using BioMedCLIP.



Axial slices carry the most stable signal across the dental-arch range and are pooled for the final embedding.

Clinical note representation

Clinical notes are serialized (chief complaint + ICD-10 diagnoses + FDI-expanded tooth references + findings) and encoded with BioMedCLIP text encoder (PubMedBERT) → $z_{text} \in \mathbb{R}^{512}$.

6 DIAGNOSIS-AWARE ALIGNMENT

Clinical relatedness between two patients is scored by Jaccard overlap of diagnosis sets D_i, D_j (=1 for same-patient cross-modal pairs, preserving the within-patient CLIP signal):

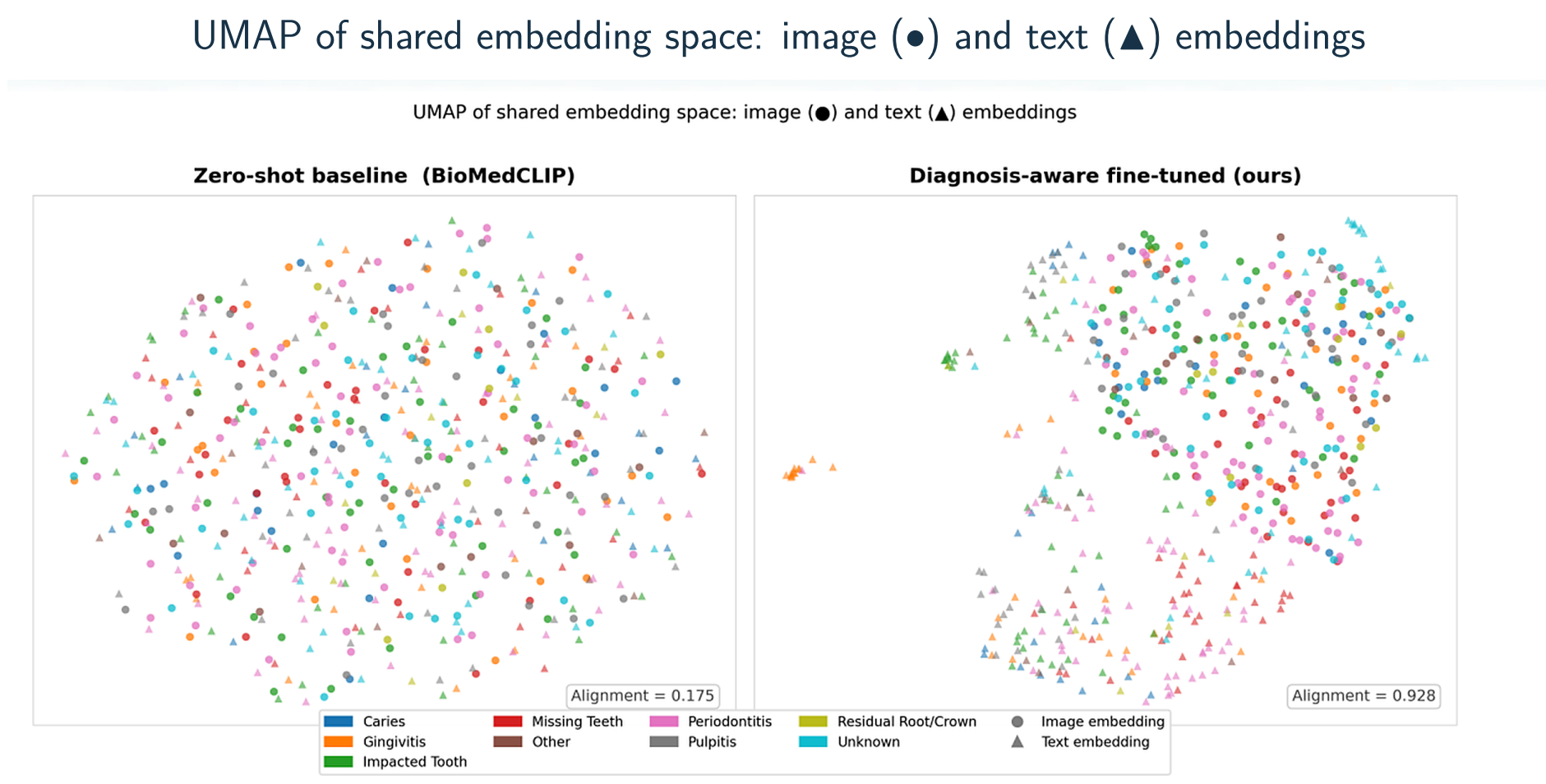
$$m_{ij} = \frac{|D_i \cap D_j|}{|D_i \cup D_j|}$$

This weights positive pairs in a cross-modal supervised-contrastive loss, combined with a cosine-similarity CLIP term:

$$\mathcal{L} = 0.7\mathcal{L}_{\text{SupCon}} + 0.3\mathcal{L}_{\text{CLIP}}$$

Only the diagnostic projection heads and temperature τ are fine-tuned — the BioMedCLIP encoders remain frozen (key for stability with 3D data).

7 EMBEDDING SPACE



Fine-tuning improves cross-modal pairing, not clean diagnostic cluster separation.

ALIGNMENT 0.175 → 0.928

Fine-tuning substantially improves image-text correspondence in the shared embedding space. Paired modalities become more closely aligned while diagnosis categories remain overlapping rather than forming clearly separated clusters.

This suggests that the learned representation primarily improves cross-modal alignment, rather than enforcing diagnosis-specific clustering. The remaining overlap reflects the heterogeneous and multi-label nature of the clinical data.

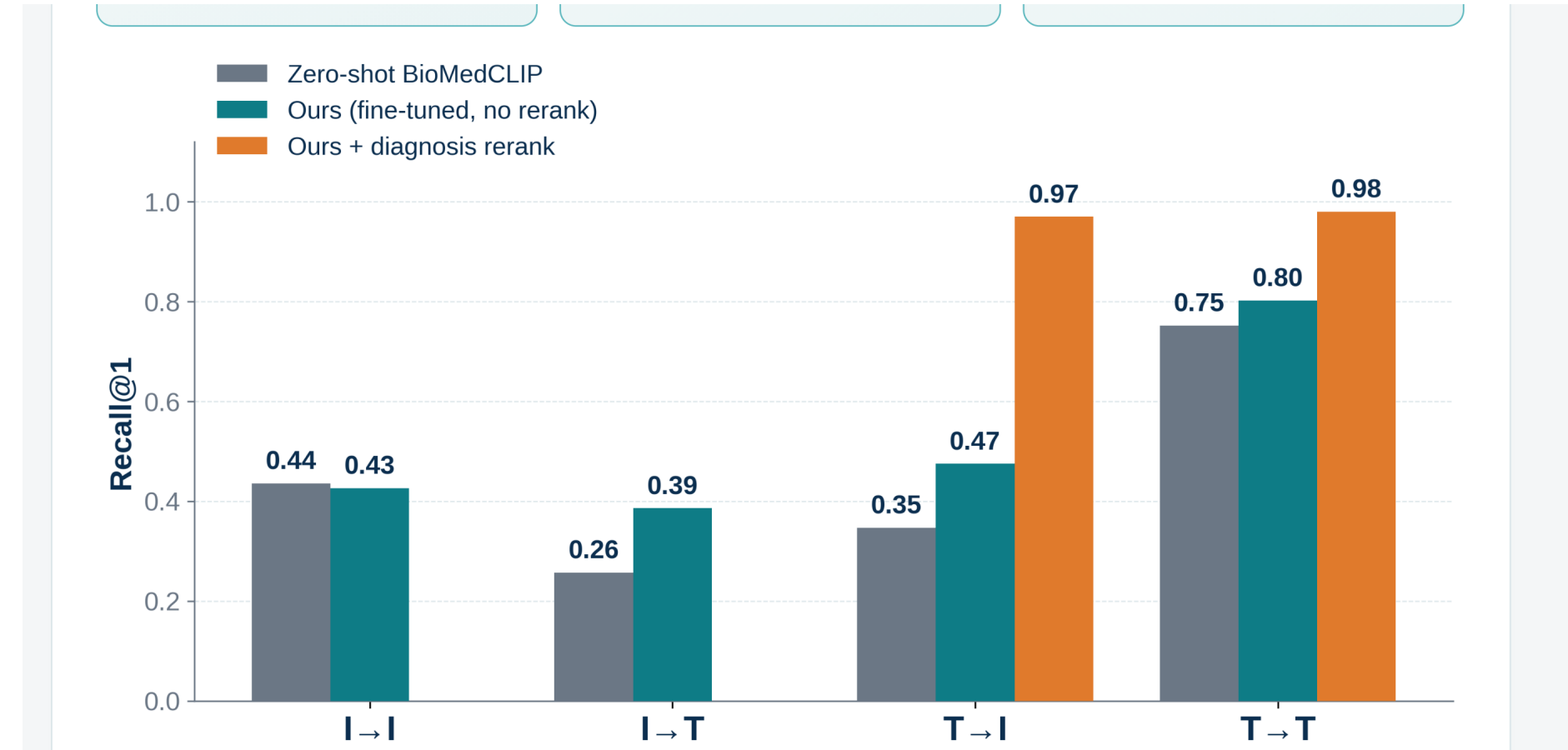
8 RESULTS

Per reviewer guidance, non-reranked results (teal) are our primary evidence of representation quality. Diagnosis-overlap reranking (orange) is reported for completeness but shares information with the relevance definition — see Critical Analysis.

+0.129 R@1 gain, image→text (Table 2)
0.257 → 0.386, non-reranked

0.475 T→I R@1, non-reranked
up from 0.347 zero-shot

0.928 cross-modal alignment
(0.175 → 0.928, embedding-only)

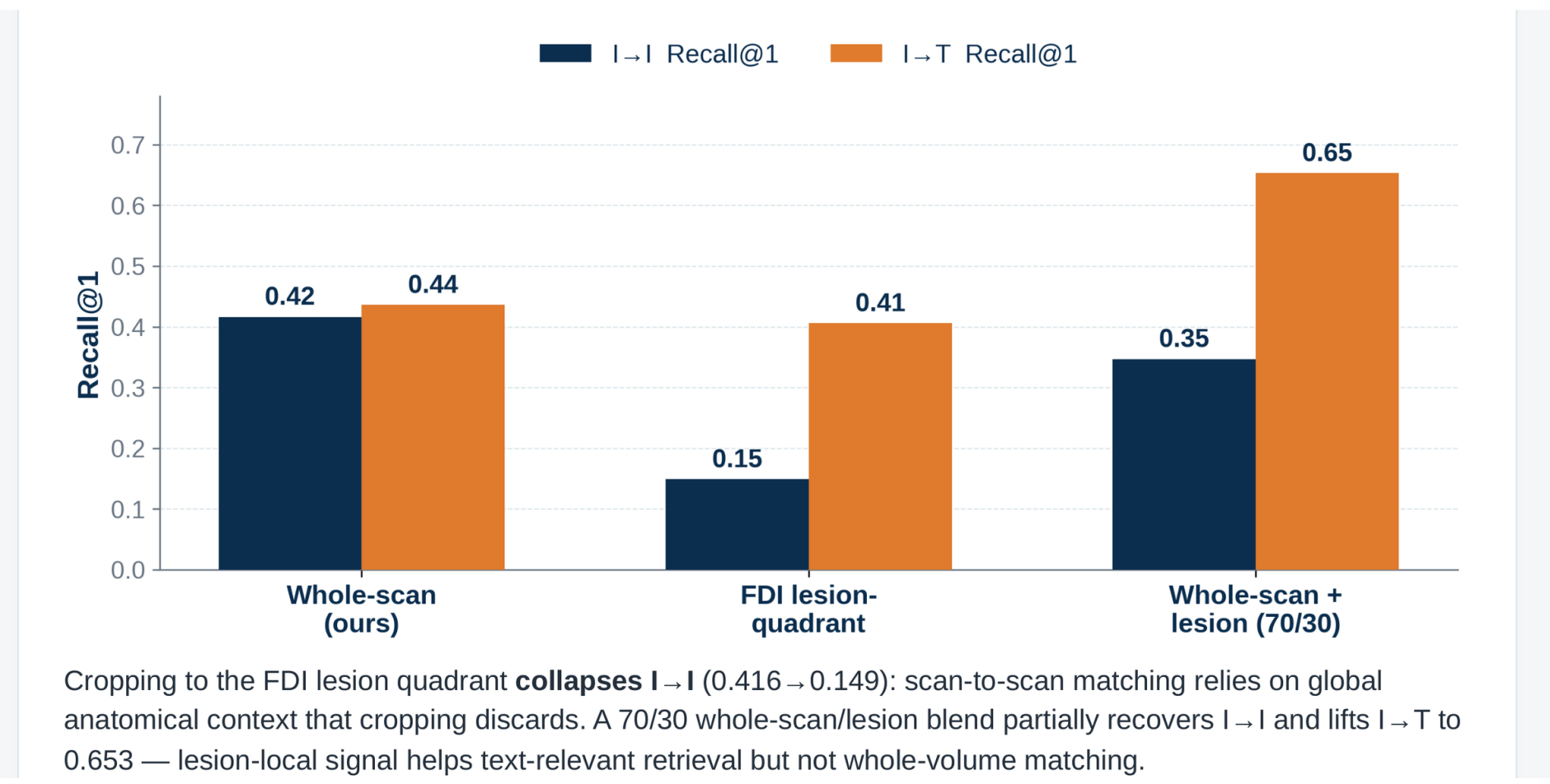


What the three bars mean: grey = zero-shot BioMedCLIP before any fine-tuning (untrained baseline). Teal = our model after diagnosis-aware fine-tuning, no reranking. Orange = that same fine-tuned model, plus diagnosis-overlap text-reranking at retrieval time. Reranking only works for text queries, so it's only defined for T→I/T→T — that's why I→I/I→T have no orange bar.

A separate reproducibility run found a larger I→T gain (+0.238); across its own seeded reruns the gain was +0.109–0.119 R@1 (Supp. S3/S4/S6). The Table 2 figure above is our primary, citable number.

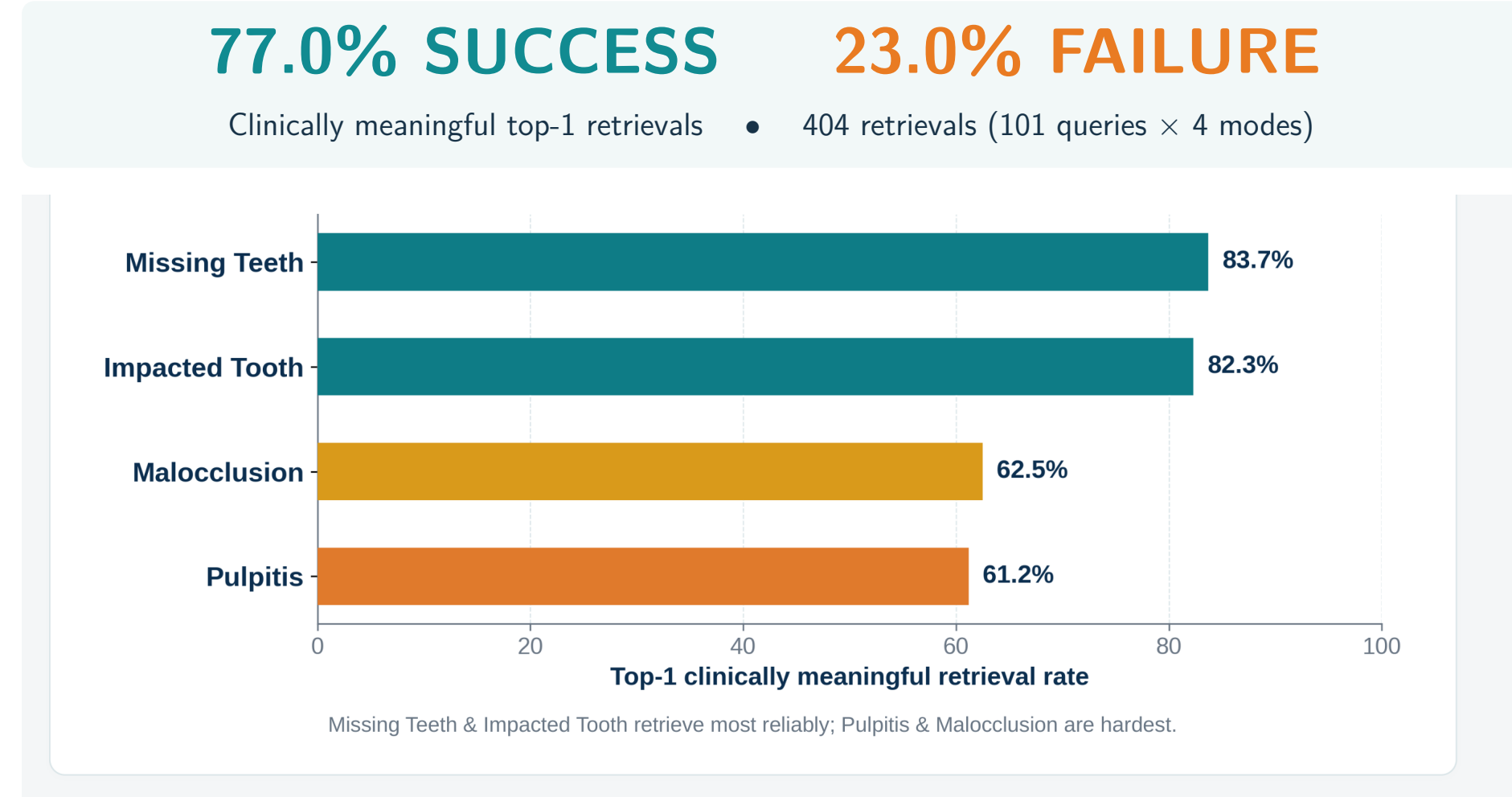
On 0.928 alignment: this is mean cosine similarity of paired embeddings (Supp. Eq. S1) — fixed once fine-tuning ends, before any retrieval happens. Reranking reorders retrieved candidates; it cannot change an embedding's cosine similarity. So 0.928 is neither “with” nor “without” reranking it isn't applicable here. Diagnosis information does shape it, but via the training loss (disclosed), not a retrieval-time shortcut.

9 IMAGE-ENCODING ABLATION



Cropping to the FDI lesion quadrant **collapses** I→I (0.416 → 0.149); scan-to-scan matching relies on global anatomical context that cropping discards. A 70/30 whole-scan/lesion blend partially recovers I→I and lifts I→T to 0.653 — lesion-local signal helps text-relevant retrieval but not whole-volume matching.

10 CLINICAL ERROR ANALYSIS



Missing Teeth & Impacted Tooth retrieve most reliably; Pulpitis & Malocclusion are hardest.

Retrieval performance varies substantially across diagnosis categories. Missing Teeth and Impacted Tooth are retrieved most reliably, whereas Pulpitis and Malocclusion remain the most challenging, reflecting overlapping radiographic presentations.

Overall, 77.0% of the 404 top-1 retrievals are clinically meaningful, while 23.0% are true failures. This highlights both the clinical potential of diagnosis-aware retrieval and the remaining difficulty of distinguishing visually overlapping dental conditions.

DISCUSSION & LIMITATIONS

- Diagnosis-aware supervision clearly helps **text-anchored** retrieval and cross-modal alignment; **image-anchored** (I→I) gains are flat (0.436→0.426) MMDental lacks localized lesion annotations, and many-to-one tooth references (mean 3.1/diagnosis) dilute diagnosis-specific signal in whole-volume embeddings.
- Single-center cohort (Hangzhou Dental Hospital) limits cross-population generalization; clinical notes are machine-translated from Chinese.
- Two CBCT scanners with different exposure settings, checked for confounds in the fixed-window intensity pipeline; none found.
- Diagnosis-grounded evaluation (245/101 patients) is underpowered for rare diagnoses (e.g. Pericoronitis).
- Uncertainty is quantified via bootstrap 95% CIs for the pooling ablation only; headline Recall@1/MRR/NDCCG remain single-run estimates and small differences (esp. I→I) should be read cautiously.

CONCLUSION & FUTURE WORK

- To our knowledge, the first diagnosis-aware, dual-encoder retrieval framework built specifically for pairing 3D CBCT scans with clinical notes in the dental domain** cross-modal retrieval has been explored for other modalities (chest X-ray, cardiac records), but not previously for CBCT+notes. Released as a reusable, standardized benchmark protocol on MMDental.
- Reliable gains concentrate in text-anchored retrieval and cross-modal alignment; image-anchored retrieval needs localized supervision (lesion masks), not just better pooling or cropping.
- Future work: multi-center validation (e.g. ToothFairy4), finer-grained / hierarchical diagnosis taxonomy, lesion-level annotation, per-query normalization for fairer lexical+rerank comparisons, and connecting retrieval with ontology-grounded structured reporting (Lumetti et al., MICCAI 2026).



29TH INTERNATIONAL CONFERENCE ON MEDICAL IMAGE COMPUTING & COMPUTER ASSISTED INTERVENTION
Sept. 27- Oct. 1, 2026 • Strasbourg Convention Center, FRANCE



ODIN WORKSHOP
odin-workshop.org

