

Geometry-Guided Identity Correction for Semi-Supervised CBCT Tooth and Root-Canal Segmentation

ODIN 2026

Challenge Track, Paper #2
Satellite event of MICCAI 2026
Strasbourg, 27 Sept 2026

Yukun Du^{1,*} Yude Ji^{1,*} Wei Xu¹ Rui Zhang^{2,†}
¹School of Information Science and Technology ²Clinical Medical College
Shihezi University, Shihezi, Xinjiang, China • zr794121@shzu.edu.cn

*equal contribution †corresponding author

STSR 2026

Task 1

6th

final test score
0.6474 • team djw

The foreground was never the problem. The names were.

Image Dice reached **0.9563** while identity accuracy sat at **0.2518**: the network found the teeth, then gave many of them their **mirror-image FDI class**. A training-free geometric rule on the predictions alone puts the names back — and touches no voxel of the foreground.

Identity accuracy (IA)
0.2518 → 0.5493

official validation, 20 cases

Challenge Score
0.5158 → 0.7207

same predictions, correction only

Image Dice / mIoU / NSD
unchanged

0.9563 • 0.9166 • 0.9934

Inference budget
7.5 GB

worst case 34.3 s • limits 8 GB / 60 s (A10)

1 Task, and where it broke

STSR 2026 Task 1: joint tooth and root-canal delineation in CBCT, **65 classes** (32 FDI teeth, 32 canals, background), from **30 annotated** and **299 unannotated** volumes. Two problems in one:

(a) **Segmentation** — thin, low-contrast structures.

(b) **Identification** — the two-digit FDI class, which encodes **which side of the mouth** the tooth is on.

We solved (a) and failed (b). The failures were not random: they were overwhelmingly **contralateral** — the right tooth, under its mirrored name.

2 Semi-supervised training

One 3D nnU-Net on **24 annotated + 40 unannotated** volumes: deep-supervised Dice + cross-entropy over 65 classes, plus a **detached weak-to-strong consistency** term.

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + w(t) \text{MSE}[\sigma(p_{\theta}(a_s(x_u))), \sigma(\text{sg}(p_{\theta}(a_u(x_u))))]$$

The target runs under `no_grad` — *not* an EMA teacher. Plan `nnUNetPlansBigMem`, patch $160 \times 112 \times 160$, batch 2; submitted checkpoint is the trainer-best model at **epoch 496**, selected on identity metrics rather than loss. It beat the supervised baseline on 7 of 8 internal metrics — though objective and patch size changed together, so the comparison is uncontrolled.

3 Geometry-guided identity correction

Training-free. No second model, no atlas, no ground truth — it reads only the predicted label map and the NIFTI affine.

- Midline.** Map tooth voxels to world coordinates; the patient-specific midline is $x_{\text{mid}} = \text{median}\{x(v)\}$.
- Side test.** In LPS, $x < x_{\text{mid}}$ is right. Extract each tooth class with 6-connectivity; compare each component's mean world- x against the side its FDI class claims.
- Swap.** On conflict apply the contralateral map within the same jaw, $g(k) = k \pm 8$. Ties and components under 5 voxels are left alone.
- Cleanup.** Flag duplicate classes, redundant quadrant coverage, and components isolated from the arch; remove if small, else relabel to the majority neighbouring class.

It renames components, never edits masks — which is exactly why every image-level metric is identical before and after.

Canals deliberately bypass the rule. Their centroids give only weak laterality cues; applying it to canals dropped internal canal instance Dice from **0.4512 to 0.3240**.

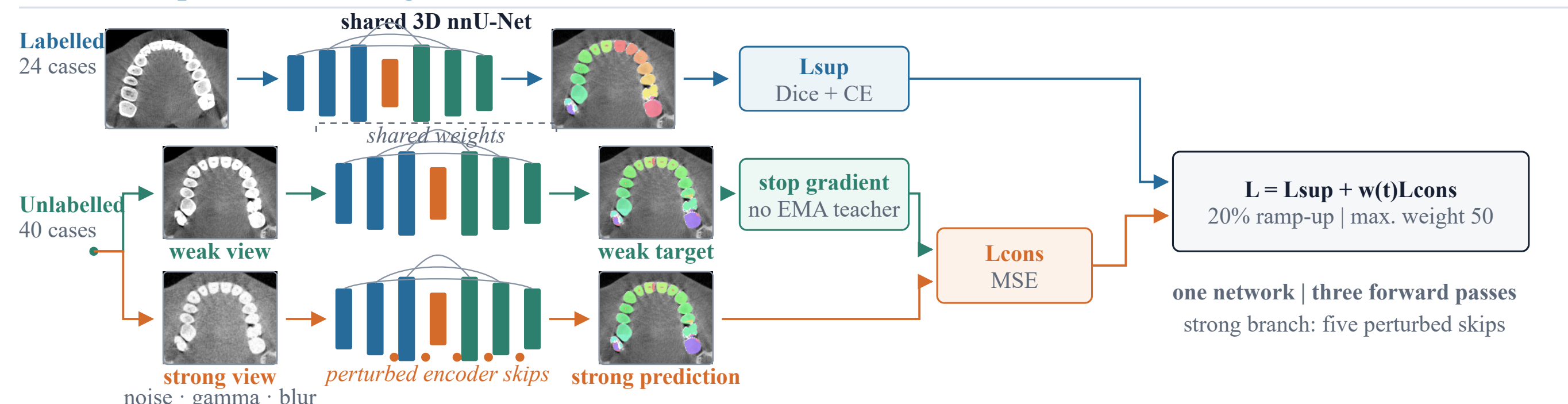
4 A confound we did not control

FDI classes encode laterality — yet the submitted training **mirrored on the left–right axis** twice over: nnU-Net's default `mirror_axes` keeps all three axes without remapping tooth ids, and the weak/strong views flip independently, coinciding on only **1 step in 8** while the loss compares them voxel-wise with no inverse transform.

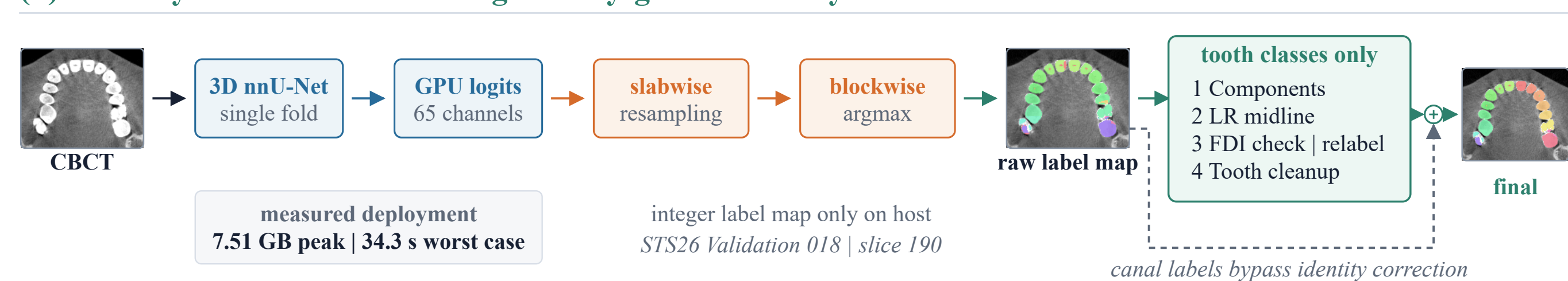
Both push the network toward *mirror-invariant* class scores — the opposite of what FDI needs. The ToothFairy2 winner reports that disabling mirroring on that axis alone “drastically improves results”. We could not run the ablation before the deadline, so we report this as a **plausible aggravating factor, not a measured cause**.

5 Pipeline

(a) Semi-supervised training



(b) Memory-aware inference and geometry-guided identity correction



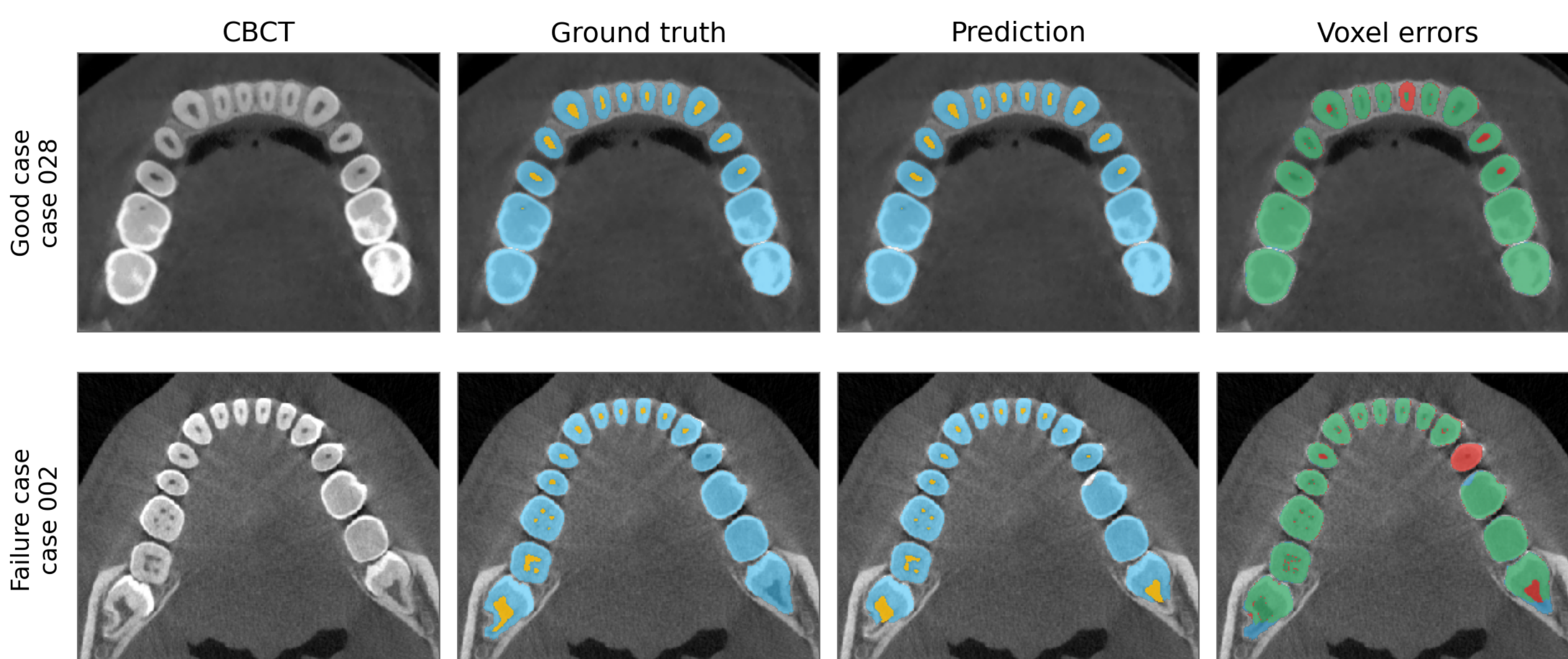
Supervised Dice–CE with weak-to-strong consistency; slabwise export gives the raw 65-class map; laterality correction and cleanup act on teeth while canals bypass. The Validation 018 slice shown has **6,120 relabelled voxels** and **unchanged foreground**.

6 Official results

Validation (20 cases)	Dice _{img}	Dice _{inst}	NSD _{inst}	IA	Score
raw prediction	0.9563	0.4372	0.5570	0.2518	0.5158
+ correction	0.9563	0.6774	0.7694	0.5493	0.7207
Final test, team djw (rank 6)					0.6474

Read this honestly. Both validation rows are stored predictions with tooth-only correction, *no* cleanup, `tile_step_size=0.5`; the submitted container adds cleanup and step 1.0. Internally those cost 0.0045 IA — far too little to explain the gap to 0.6474, and no component-level test metrics were released. That row measures **the correction's effect**, not the container's expected test score.

7 Qualitative



Submitted configuration. Columns: CBCT, ground truth, prediction, voxel errors. Top — good case (028; tooth/canal image Dice 0.9666/0.9162). Bottom — canal failure (002; 0.9320/0.4649). Teeth blue, canals yellow; green correct, cyan missed, red wrong or extra.

8 What is still broken

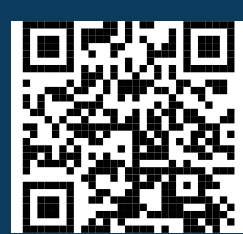
Canals. Where missed canal voxels go: **63.4%** to the contralateral canal, 29.1% to a tooth class, 6.3% to another canal, and only **1.2% to background**. Almost nothing is lost — it is an *identity* failure, not a detection failure.

So why not inherit the parent tooth's identity? We audited all **603** reference canal instances in the 30 annotated volumes. Only **53.9%** lie inside the tooth their class names; **43.8%** lie inside the *contralateral* same-slot tooth. **No fixed offset rule does better.** Canal identity needs canal-specific cues, not tooth geometry.

Other limits. The midline assumes a roughly symmetric dentition; a partly edentulous jaw could shift it enough to relabel correct components, and our six held-out cases contain no such extreme. Single split, no cross-validation; 40 of 299 unannotated scans under a screening rule we cannot reproduce.

9 Takeaways

- Score overlap and identity separately.** A 0.9563 Dice hid a 0.2518 IA; one number would have concealed the whole failure mode.
- Some identity errors are geometric, not semantic.** A median-midline rule over predictions more than doubled official IA at zero training cost.
- Check augmentation against your label space.** Mirroring is free accuracy for foreground and actively harmful when the class id encodes a side.
- Root canals remain open.** Their identity is not recoverable from the parent tooth. That is the next target.



Code

Full source: github.com/EdmundJi/stsr2026-djw • Contact: zr794121@shzu.edu.cn

Challenge data are available from the STSR organizers under their data-use terms; the data subset and checkpoint are not redistributed. Developed within the ODIN 2026 challenge cluster (STS, ToothFairy, CBCT-reporting, Bite2Text). AI-assisted coding and language tools were used; study design, method choice, execution, interpretation and all numbers are the authors' own. No competing interests.