Supplementary Material

S1. Where the ordinal loss helps. Grade 0 accounts for ∼78% of the test patches, so an aggregate score says little about the rarer, clinically relevant grades (Fig. S1). Cross-entropy treats the grades as unordered categories and drifts towards the frequent ones: on BE it labels most grade-2 patches as grade 1 and recovers only 31 of the 61 most severe patches, against 46 for the ordinal loss. The ordinal loss keeps its predictions near the true grade, so its mistakes are smaller rather than fewer: the mean absolute error drops from 0.376 to 0.313 on BE and from 0.290 to 0.255 on JSN, which is exactly what QWK rewards (Table 1). On JSN the two objectives are equivalent overall (0.626 vs. 0.627 QWK) and differ only in how they trade rare grades against frequent ones.

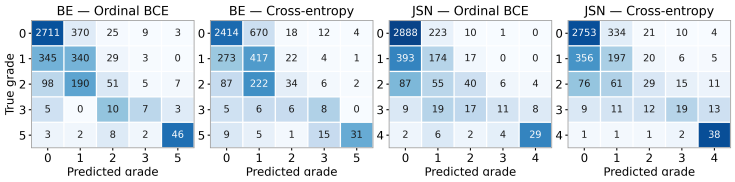

Fig. S1. Confusion matrices on the test set at 100% labels (seed 42, the only seed with cross-entropy checkpoints). Rows: reference grade; columns: prediction; cells: patch counts; shading: fraction of the row.

S2. Why does joint-type conditioning hurt? Figure S2 shows two severe (grade 5) BE patches that the baseline scores correctly and the joint-conditioned model does not. Both models look at the same eroded bone margin but the conditioned model predicts the grade that is most frequent for that joint: the identifier acts as a prior on the score rather than as anatomical context. Accuracy barely changes while severe errors grow from 4.7% to 5.3% – similarly many mistakes, but larger ones.

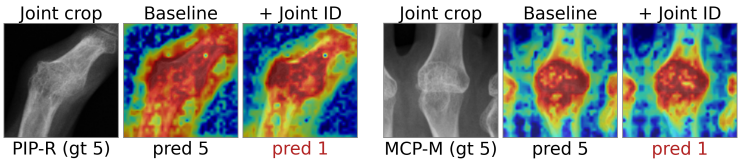

Fig. S2. Grad-CAM, two grade-5 BE cases: same attention, collapsed grade.

S3. Does the multi-task result depend on λ? We repeated the BE/JSN multi-task experiment for $\lambda \in \{0.25, 0.5, 1, 2, 4\}$ at 25/50/100% labels with three seeds each, selecting λ on validation; $\lambda = 0.25$ was the best value at every label fraction. On test (Table S1) λ acts as a trade-off knob: lowering it favours BE, raising it favours JSN, and no value is best for both. All settings stay within the seed-to-seed variability of the single-task models (up to ±0.06 QWK on BE), so the conclusion of Table 3 does not depend on λ.

Table S1. Multi-task test QWK at 100% labels vs. λ (three seeds).

	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 2$	$\lambda = 4$	Single-task
BE	0.560	0.550	0.546	0.525	0.489	0.556
JSN	0.633	0.631	0.625	0.642	0.610	0.607