



**Supplemental Figure 10.** Grader and UMAP metric agreement on 100 volume pairs vs absolute delta in UMAP metric for the volumes by experiment setup. 1 grade difference Chew et al. cases exhibit the biggest range of deltas in UMAP metric between each volume pair. This is the easiest experiment setup and graders agree most with the UMAP metric. When the Chew et al. grade is the same for both volumes, the UMAP metric deltas are closer to 0, leading to more disagreement. The longitudinal setup had the smallest UMAP metric absolute delta range, leading to most disagreement.