Learning What To Ask For When: Image Ordering for In-Context Interactive Medical Image Segmentation

Supplementary Material

9. Experimental Details

9.1. Uncertainty Perturbation Scheme

Table 3. Test-time augmentation parameters used for uncertainty estimation.

Category	**Parameter**	**Range**
Gaussian Blur	p	0.3125
	σ_{blur}	[0.1, 0.5]
	kernel size	5
Gaussian Noise	p	0.1875
	σ_{noise}	[0.0, 0.01]
	μ_{noise}	[0.0, 0.02]
Sharpness	p	0.1875
	factor	5.0
Brightness / Contrast	p	0.3125
	brightness	[-0.1, 0.1]
	contrast	[0.7, 1.3]

10. Additional Results: Does order matter?

10.1. IQR for Final Dice and Interactivity Cost on All Datasets

Table 4. Average spread in final Dice Score between the 75th and 25th percentile image orderings across 100 random orderings of 10 images across multiple independent subsets. Values are reported as final Dice $\times 100$ per image with 95% confidence intervals from bootstrapping.

Variant	ACDC	BTCV	BUID	HipXRay	PanDental	SCD	SCR	COBRE	TotalSegmentator	WBC
5P-Pred	$1.38^{+0.14}_{-0.12}$	$1.66^{+0.06}_{-0.05}$	$0.71^{+0.08}_{-0.08}$	$0.88^{+0.07}_{-0.06}$	$3.75^{+0.35}_{-0.32}$	$0.48^{+0.03}_{-0.03}$	$1.11^{+0.04}_{-0.04}$	$1.75^{+0.05}_{-0.04}$	$1.41^{+0.19}_{-0.16}$	$1.07^{+0.07}_{-0.07}$
20P-Pred	$0.33^{+0.07}_{-0.06}$	$0.43^{+0.01}_{-0.02}$	$0.14^{+0.01}_{-0.01}$	$0.11^{+0.01}_{-0.00}$	$0.78^{+0.06}_{-0.06}$	$0.15^{+0.01}_{-0.01}$	$0.15^{+0.00}_{-0.00}$	$0.79^{+0.03}_{-0.03}$	$0.32^{+0.03}_{-0.02}$	$0.16^{+0.01}_{-0.01}$
5P-GT	$1.01^{+0.09}_{-0.08}$	$1.20^{+0.04}_{-0.05}$	$0.59^{+0.06}_{-0.06}$	$0.44^{+0.03}_{-0.03}$	$0.88^{+0.04}_{-0.04}$	$0.40^{+0.03}_{-0.03}$	$0.69^{+0.02}_{-0.02}$	$0.80^{+0.01}_{-0.02}$	$0.86^{+0.02}_{-0.03}$	$0.72^{+0.05}_{-0.06}$
20P-GT	$0.32^{+0.07}_{-0.06}$	$0.35^{+0.00}_{-0.01}$	$0.12^{+0.01}_{-0.01}$	$0.09^{+0.00}_{-0.01}$	$0.45^{+0.05}_{-0.04}$	$0.14^{+0.01}_{-0.01}$	$0.13^{+0.00}_{-0.01}$	$0.47^{+0.01}_{-0.01}$	$0.25^{+0.02}_{-0.01}$	$0.12^{+0.01}_{-0.01}$

Table 5. Average spread in interactivity cost between the 75th and 25th percentile image orderings across 100 random orderings of 10 images across multiple independent subsets. Values are reported as interactions per image with 95% confidence intervals from bootstrapping.

Variant	ACDC	BTCV	BUID	HipXRay	PanDental	SCD	SCR	COBRE	TotalSegmentator	WBC
5P-Pred	$0.22^{+0.02}_{-0.02}$	$0.32^{+0.01}_{-0.01}$	$0.26^{+0.02}_{-0.02}$	$0.37^{+0.04}_{-0.04}$	$0.12^{+0.05}_{-0.05}$	$0.21^{+0.03}_{-0.03}$	$0.29^{+0.01}_{-0.01}$	$0.42^{+0.01}_{-0.01}$	$0.33^{+0.01}_{-0.01}$	$0.39^{+0.04}_{-0.03}$
20P-Pred	$0.43^{+0.05}_{-0.05}$	$0.55^{+0.01}_{-0.02}$	$0.26^{+0.02}_{-0.02}$	$0.46^{+0.04}_{-0.03}$	$0.60^{+0.04}_{-0.05}$	$0.24^{+0.03}_{-0.03}$	$0.61^{+0.02}_{-0.02}$	$0.69^{+0.02}_{-0.02}$	$0.46^{+0.01}_{-0.01}$	$0.81^{+0.04}_{-0.04}$
5P-GT	$0.18^{+0.01}_{-0.01}$	$0.27^{+0.01}_{-0.01}$	$0.22^{+0.01}_{-0.01}$	$0.20^{+0.01}_{-0.01}$	$0.18^{+0.01}_{-0.01}$	$0.13^{+0.02}_{-0.02}$	$0.17^{+0.01}_{-0.01}$	$0.23^{+0.01}_{-0.01}$	$0.26^{+0.01}_{-0.01}$	$0.31^{+0.02}_{-0.02}$
20P-GT	$0.38^{+0.04}_{-0.04}$	$0.45^{+0.01}_{-0.01}$	$0.22^{+0.02}_{-0.02}$	$0.35^{+0.04}_{-0.04}$	$0.26^{+0.01}_{-0.02}$	$0.20^{+0.02}_{-0.02}$	$0.51^{+0.02}_{-0.02}$	$0.46^{+0.02}_{-0.02}$	$0.37^{+0.01}_{-0.01}$	$0.68^{+0.03}_{-0.03}$

10.2. Click Curves for 5P-Pred

Figure 5 shows the average score per click index across all datasets.

10.3. Best-vs-Worst Permutation Curves on All Datasets

Figure 6 shows the best-versus-worst permutation curves in terms of interactivity cost.
Figure 7 shows the corresponding final Dice curves.

11. Additional Results: Can uncertainty policies beat random?

11.1. Metric Trajectories along Ordering Positions

Figure 8 summarizes the ordering trajectory for final Dice.
Figure 9 summarizes the ordering trajectory for interactivity cost.

11.2. Average Uncertainty Runs vs Random

Figure 10 compares uncertainty and random selection when averaging across all 10 starts for initial Dice (in-context prediction with no interaction).
Figure 11 compares uncertainty and random selection when averaging across all 10 starts for interactivity cost.
Figure 12 compares uncertainty and random selection when averaging across all 10 starts for final Dice.

11.3. Best Uncertainty Runs vs Random

Figure 13 compares uncertainty and random selection when choosing the best start among 10 runs for interactivity cost.
Figure 14 compares uncertainty and random selection when choosing the best start among 10 runs for final Dice.

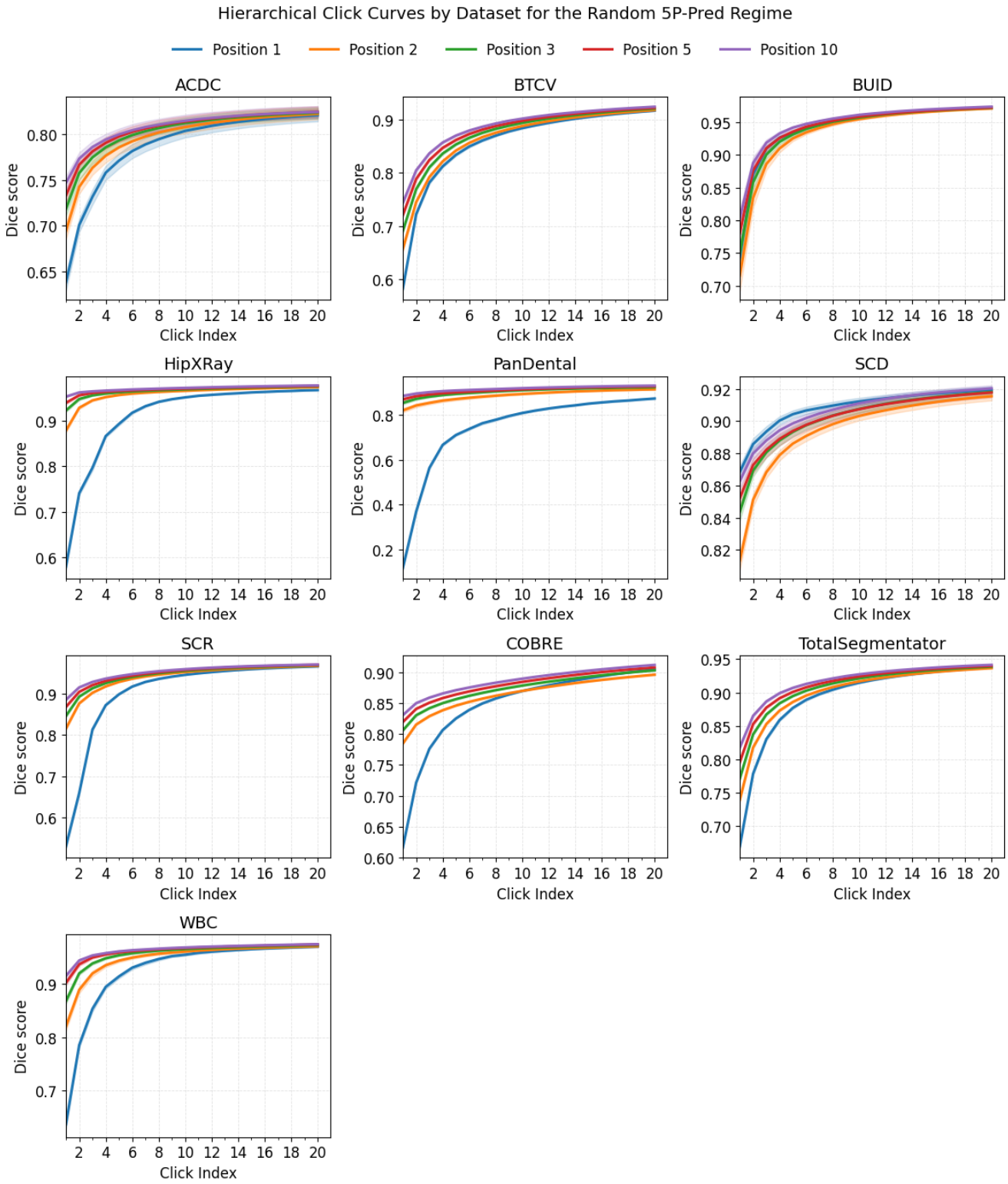

Figure 5. Position-wise click curves under the random 5-click prediction-commit regime.

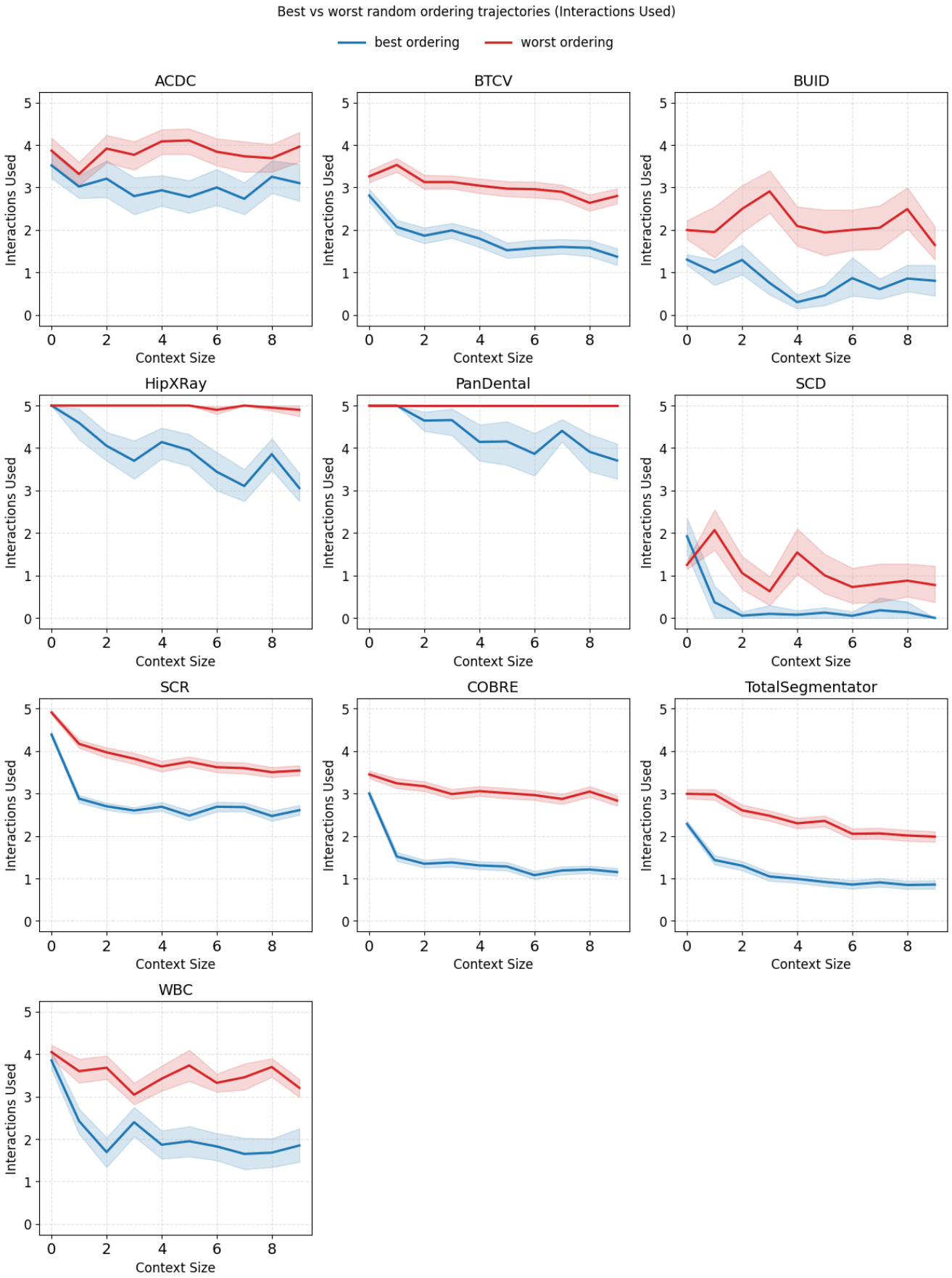

Figure 6. Best-vs-worst permutation curves for interactivity cost on all datasets.

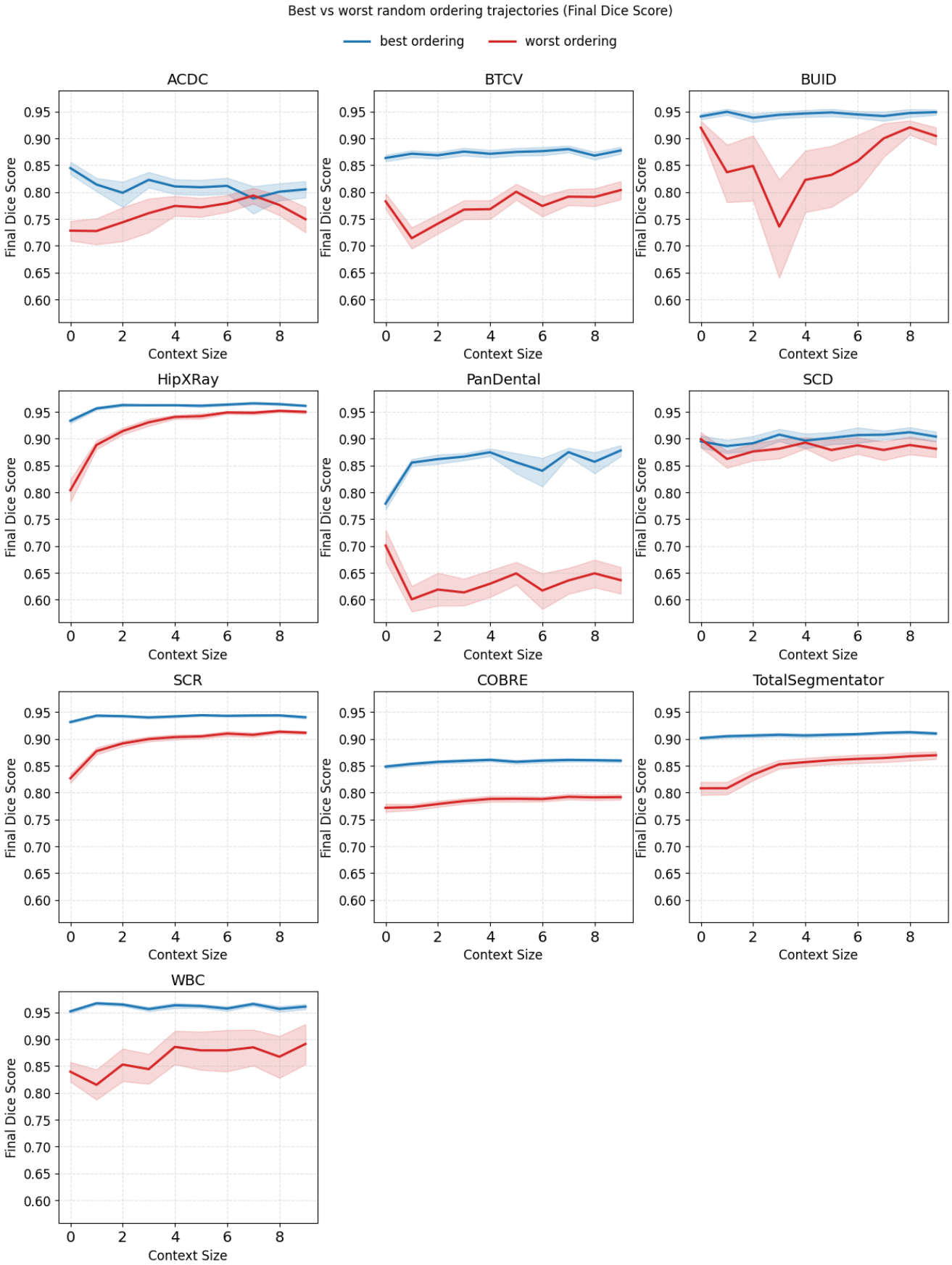

Figure 7. Best-vs-worst permutation curves for final Dice on all datasets.

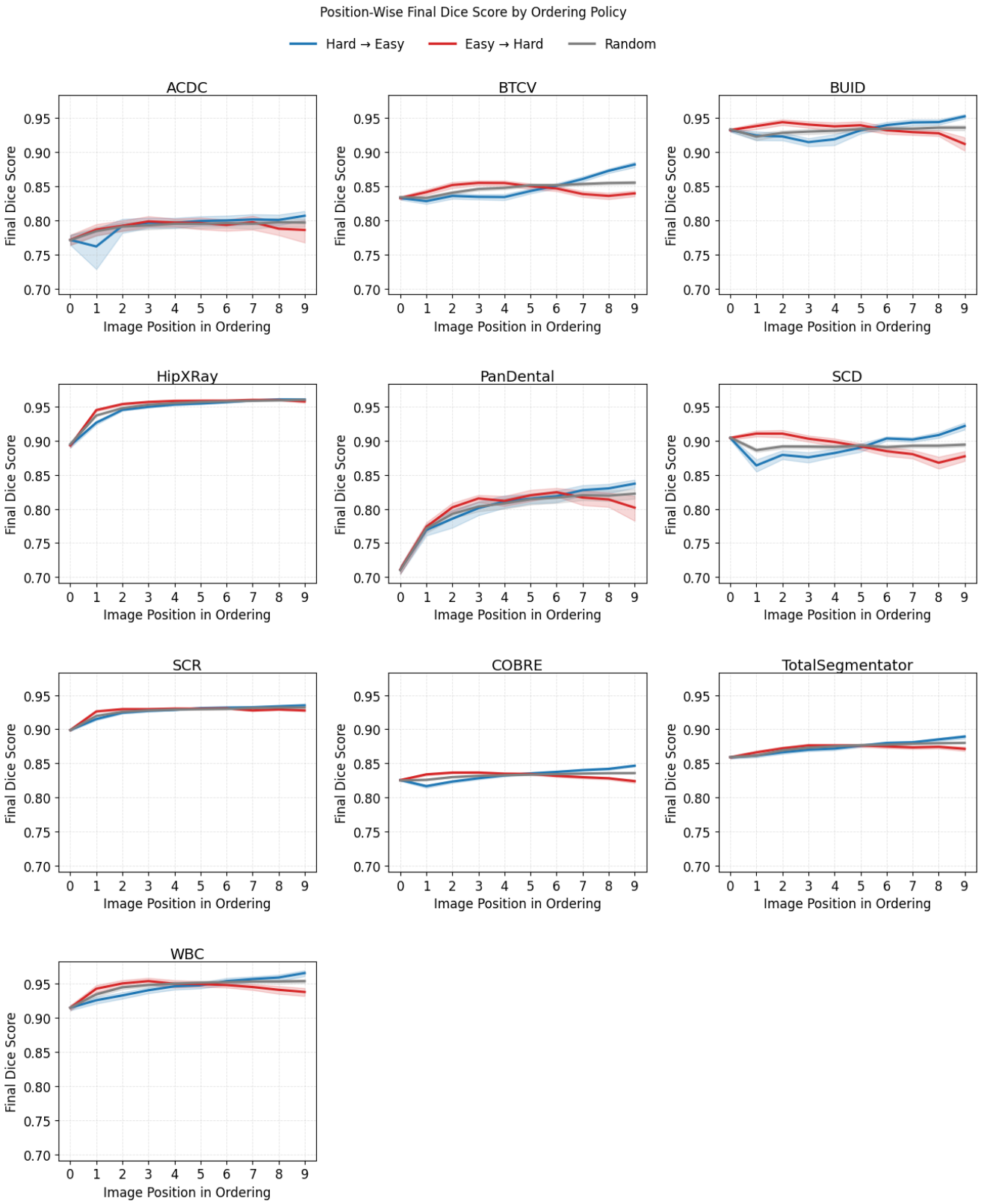

Figure 8. Final Dice trajectories across position indices

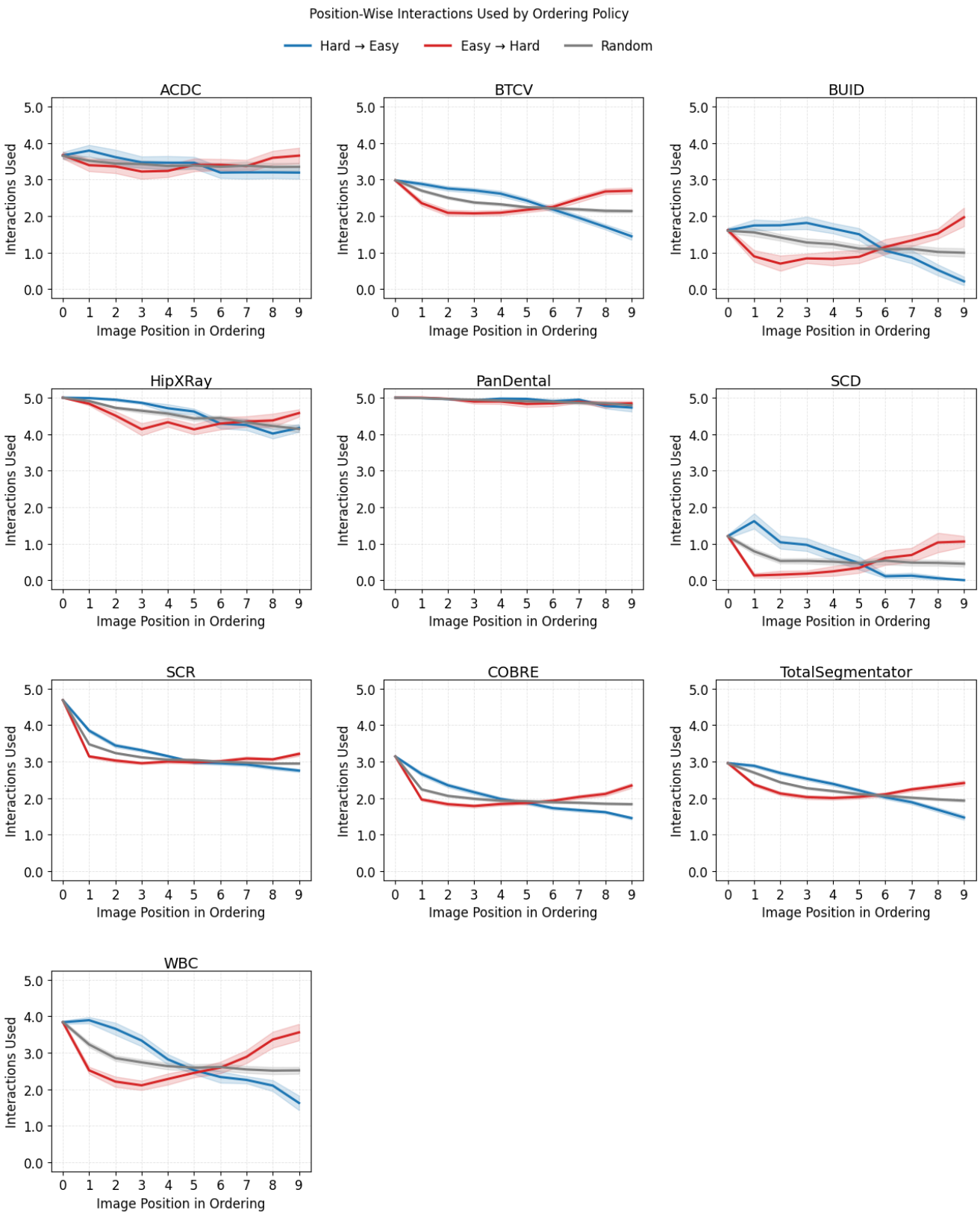

Figure 9. Interaction cost trajectories across position indices

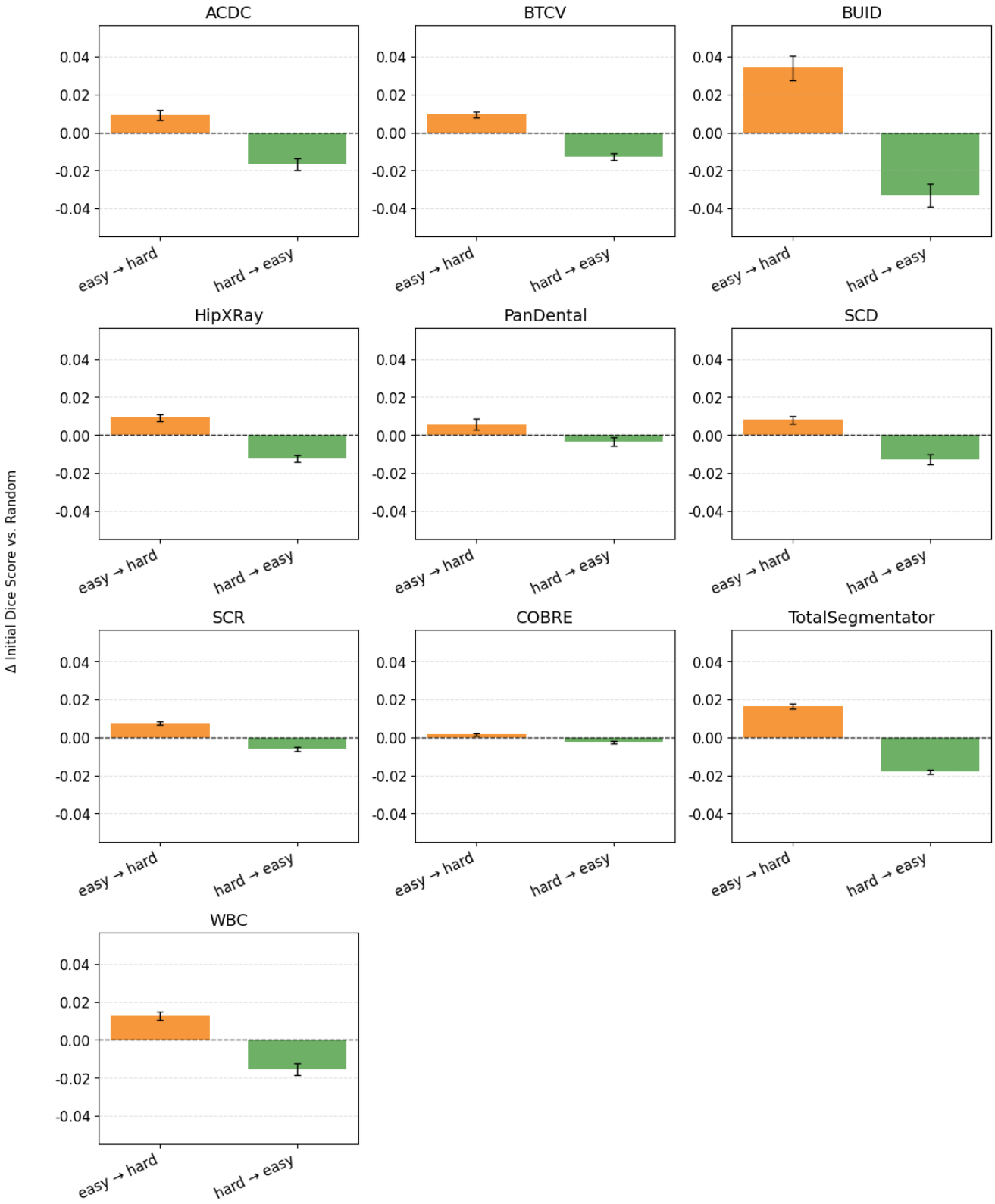

Figure 10. Uncertainty vs. random when averaging across all 10 starts for initial Dice.

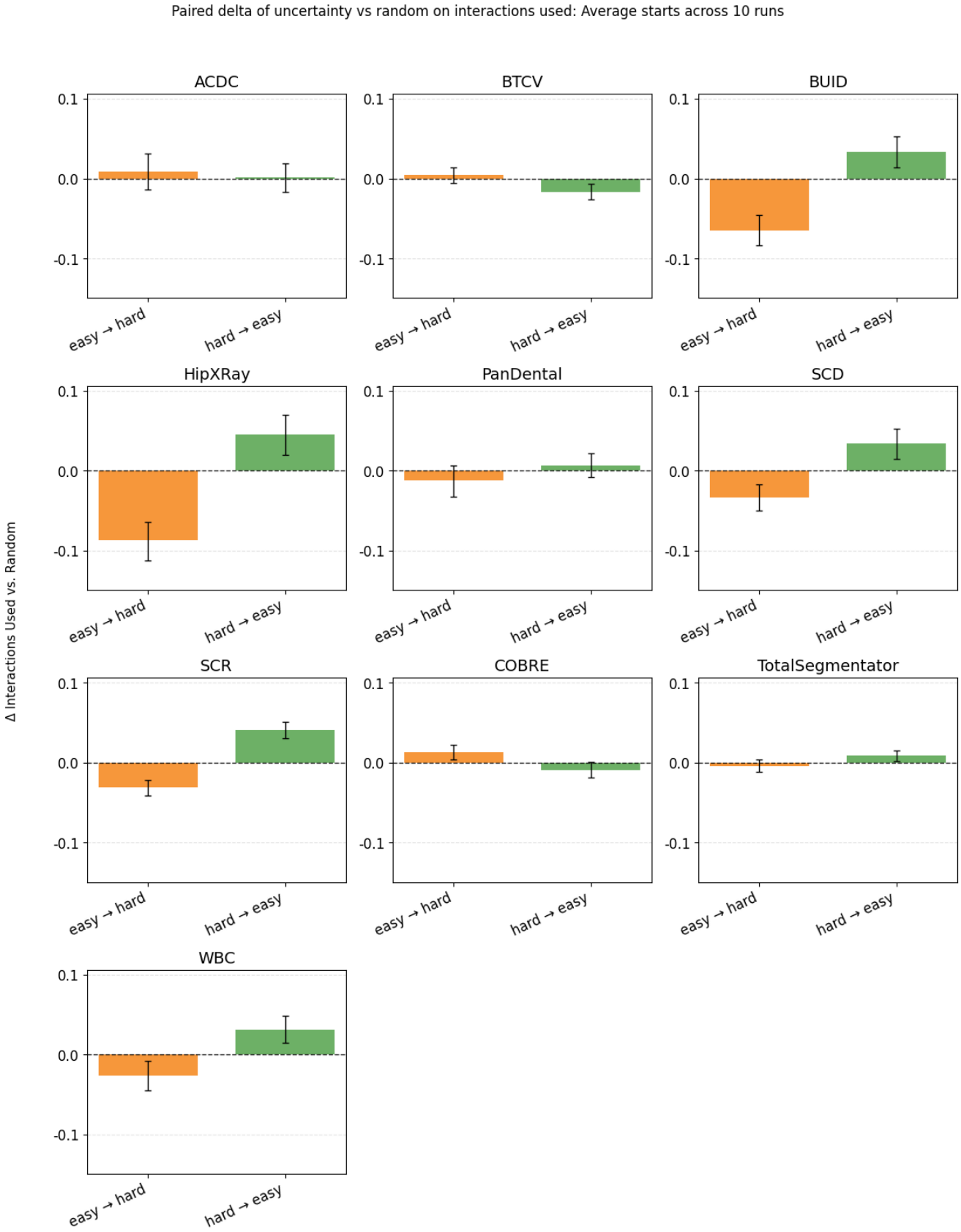

Figure 11. Uncertainty vs. random when averaging across all 10 starts for interactivity cost.

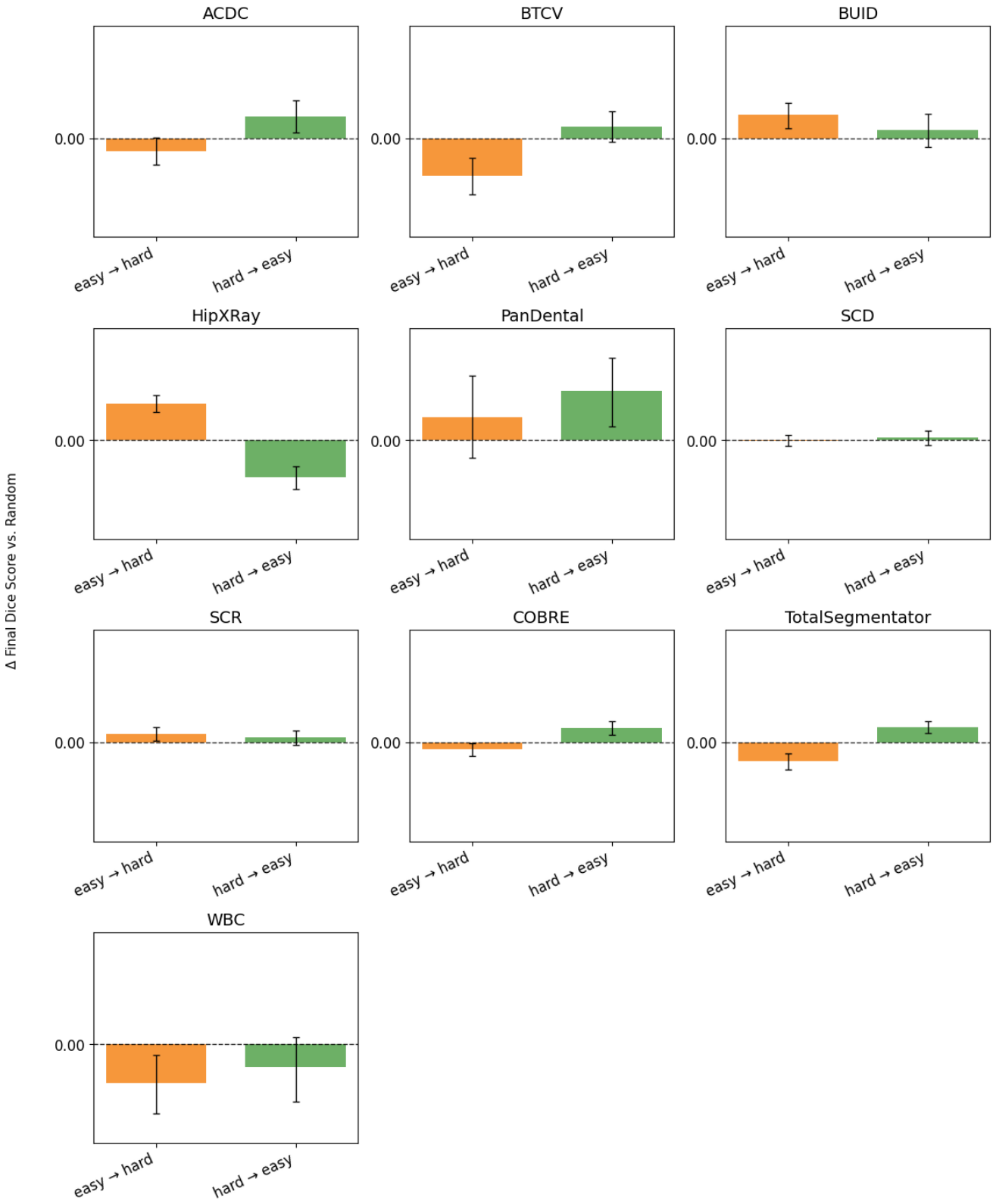

Figure 12. Uncertainty vs. random when averaging across all 10 starts for final Dice.

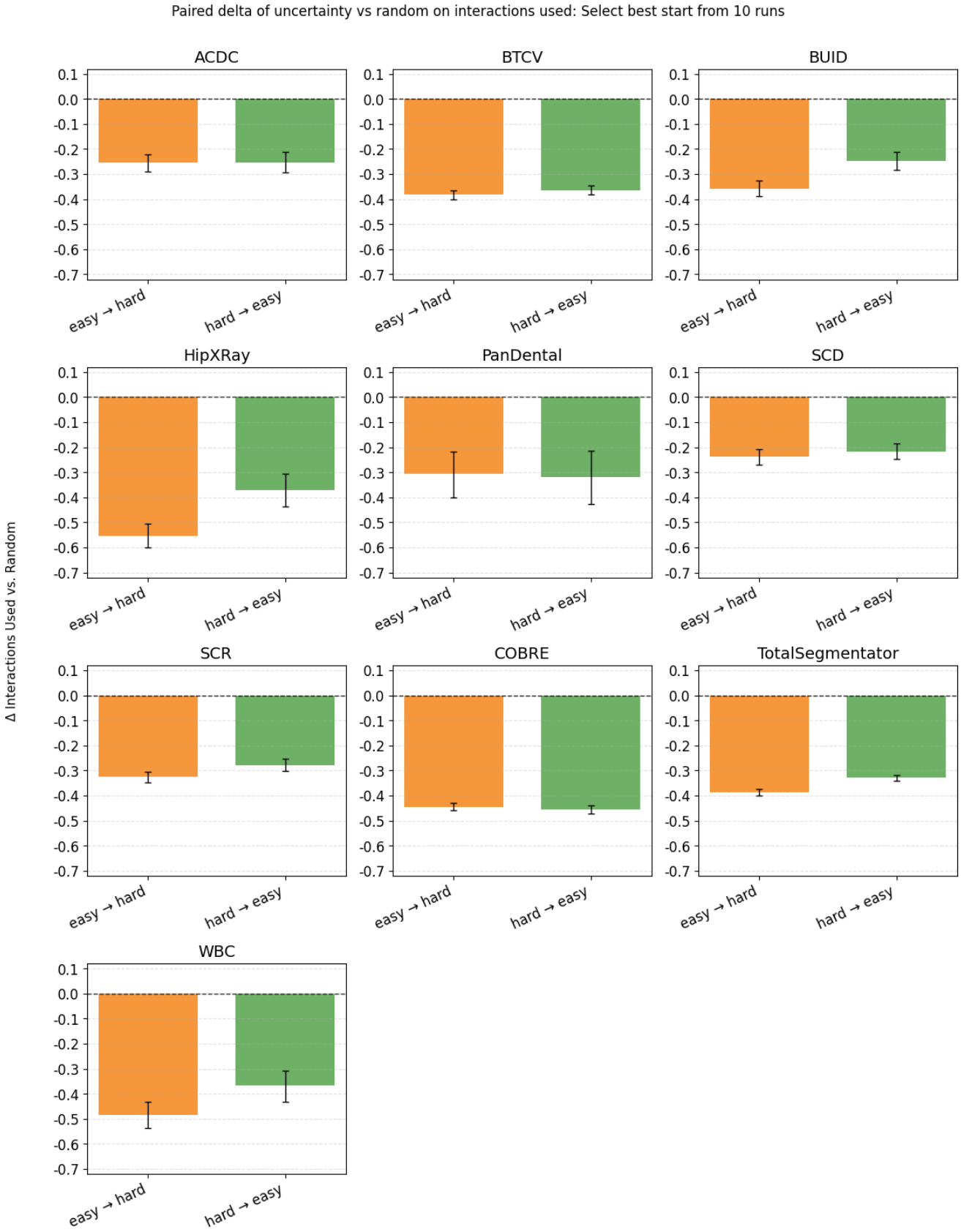

Figure 13. Uncertainty vs. random when selecting the best start among 10 runs for interactivity cost.

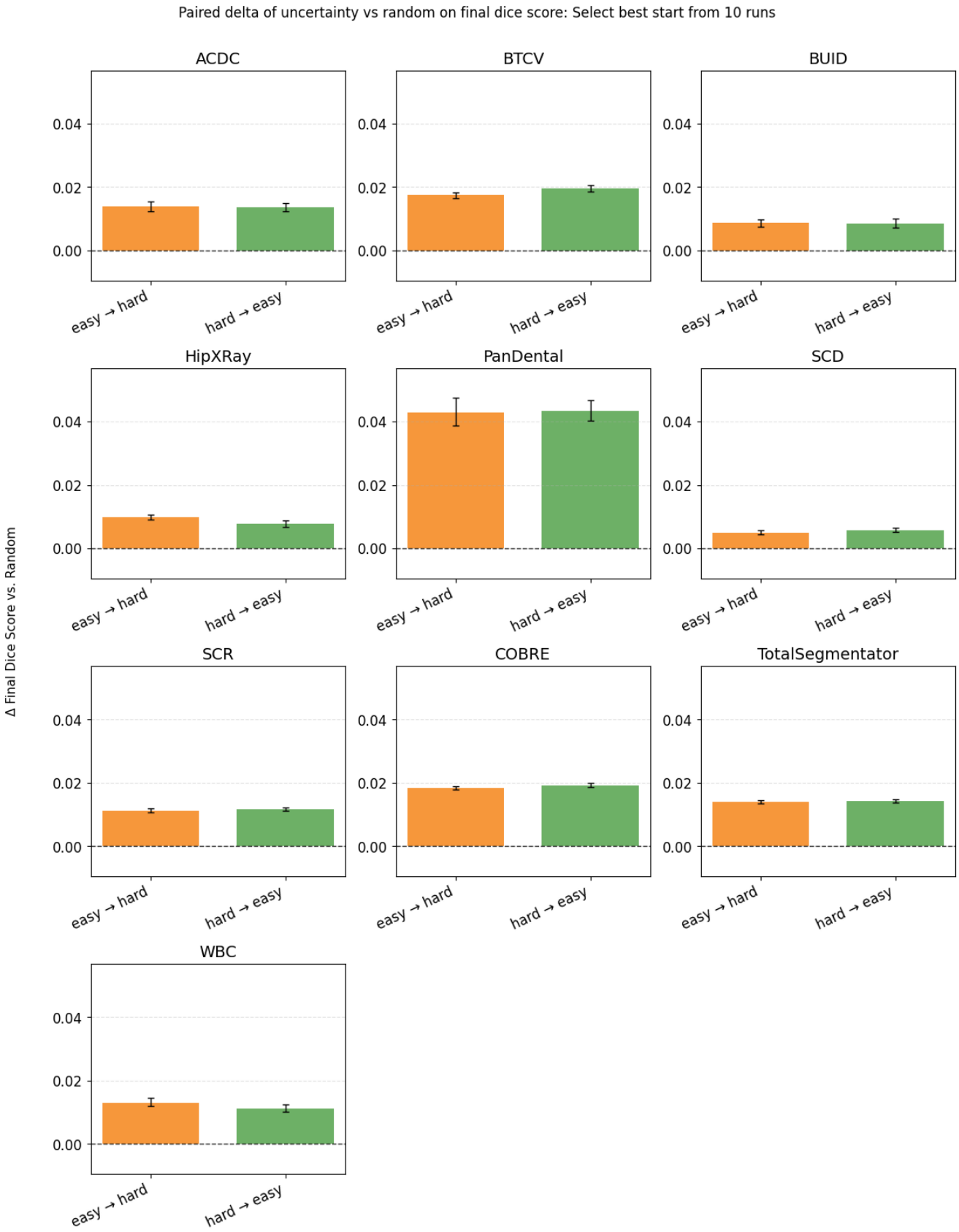

Figure 14. Uncertainty vs. random when selecting the best start among 10 runs for final Dice.