Supplementary Materials for

"A Novel Tracking Framework for Devices in X-ray Leveraging Supplementary Cue-Driven Self-Supervised Features"

Table 1: **Ablations.** We explore (a) the effect of reconstruction and weak-label prediction to understand which representation space is dominant for Catheter tip tracking with manual initialization. (b) The feature crop size (appearance tokens (ϕ)) is crucial since larger crop can lead to lack of precision, whereas small crops can risk losing the context due to incorrect model predictions.

(a) Pretraining Reconstruction (α) and weak-label (β) loss weights

α	β	RMSE		
		mean	std	max
0.75	1	1.53	1.09	10.55
1	1	1.69	1.06	6.13
1	0.75	1.65	1.13	6.99
1	0.5	**1.21**	**0.68**	**4.04**
1	0.25	1.24	0.74	4.55
1	0	1.33	0.81	4.99

(b) Effect of different appearance token size ($|\phi|$) for HiFT. For token size ($s \times s$), effective pixel level size is ($s \times 16, s \times 16$).

$\|\phi\|$	mean	std	max
4×4	1.48	1.18	7.49
3×3	**1.21**	**0.68**	4.04
2×2	1.26	0.73	**4.01**

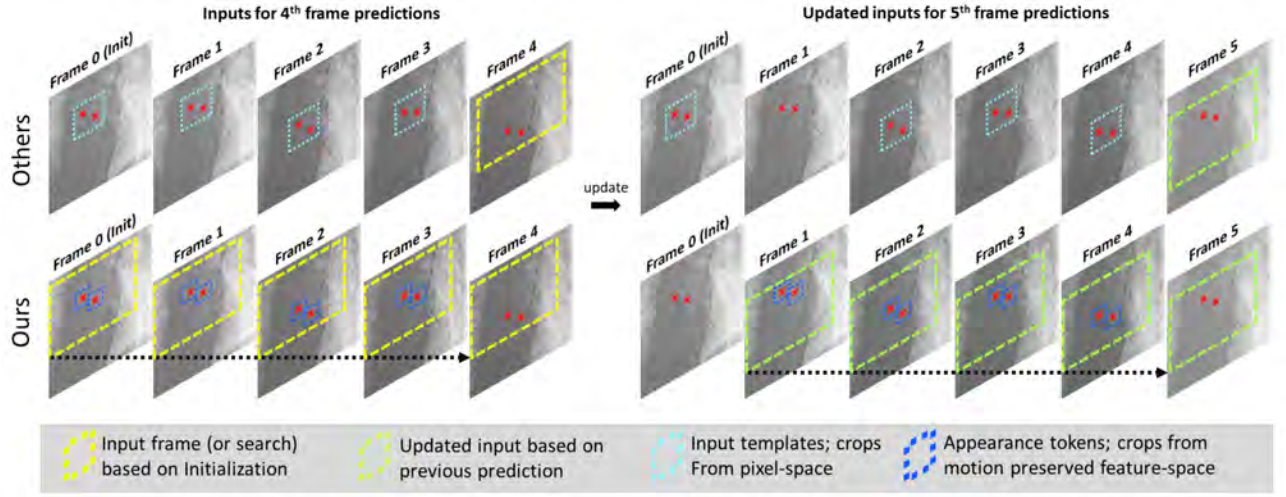
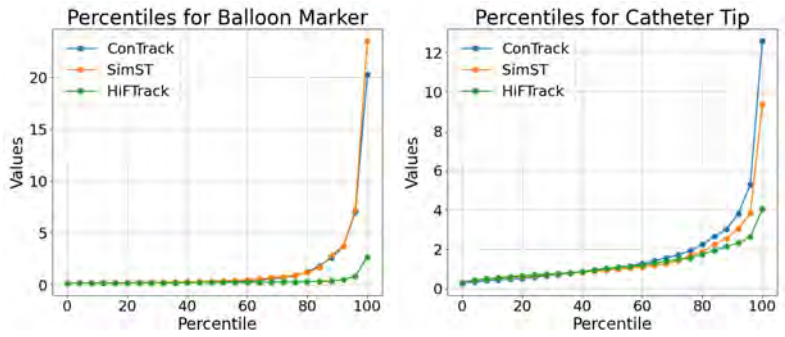


Figure 1: **Approach used by traditional tracking frameworks vs Ours:** Other frameworks use template crops and a search crop, relying on spatial correlation between them to perform tracking. In contrast, our approach first extracts spatio-temporal features from 5 consecutive frames using symmetrical cropping and spatial-correlation is enforced only in the motion-preserved feature space. We update the crops for the current frame and also past frames with same crop parameters. Frames without dotted windows in the figure are not used as model inputs.



(a) (b)

Figure 2: Percentile plot of errors for (a) Balloon Markers and (b) Catheter Tip. The results show the reduction in failures for HiFT indicating high stability and robustness.

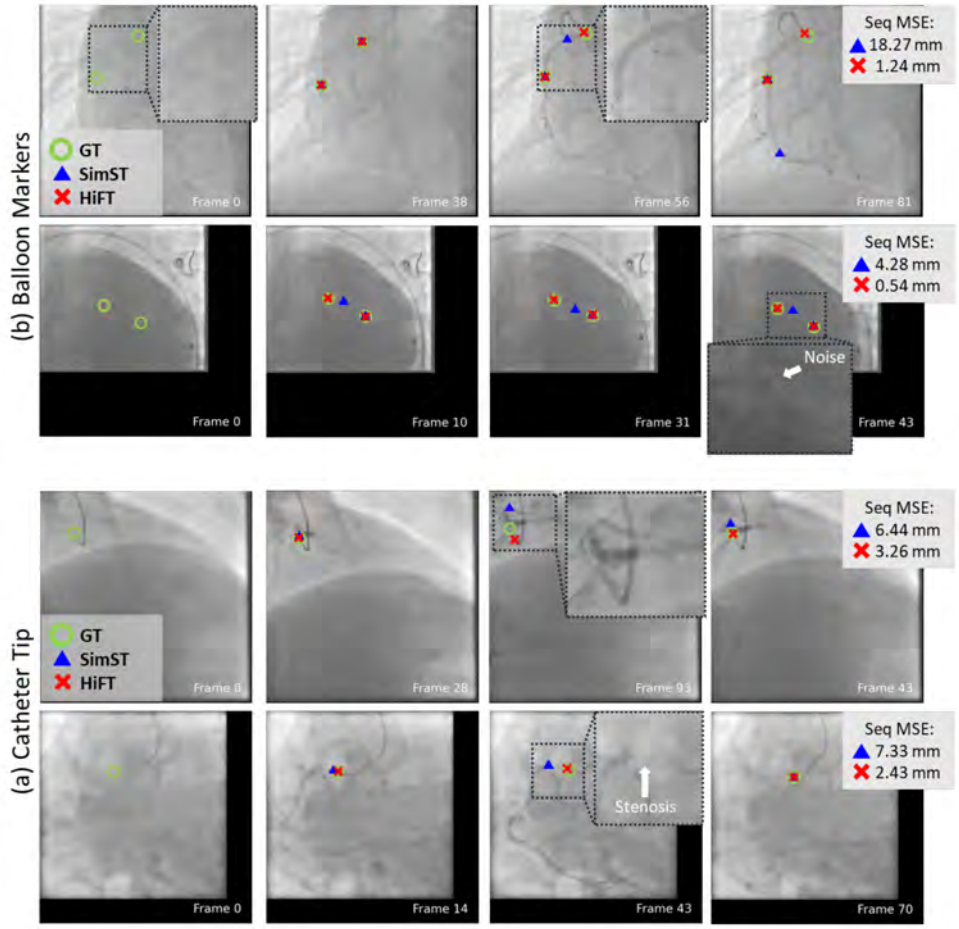


Figure 3: Qualitative examples of tracking performance of SimST and HiFT for (a) Catheter Tip and (b) Balloon Marker. Balloon Markers are small, making them vulnerable to noise and distractions. Tracking is challenging during occlusion for balloon markers and catheter tip. Abnormalities such as stenosis can lead to inconsistent contrast uptake, confusing the network to detect the catheter tip. Our tracker achieves fairly good performance in all scenarios.