Table R1: Quantitative comparison to more image-based SOR and video SOD methods. Our method consistently outperforms them by a large margin. Due to time limitations, more comparisons on other datasets will be included in the final version if possible.

Method		RVSOD		DAVSOD	
		$SA-SOR$	MAE	$SA-SOR$	MAE
SOR	Fang (2022)	0.350	0.0984	-	-
	Liu (2022)	0.563	0.0728	-	-
	PSR (2023)	0.405	0.074	-	-
	SeqRank (2024)	0.512	0.0761	-	-
VSOD	DCFNet (2021)	-	0.1180	-	-
	SCOTCH (2023)	-	0.1230	-	-
VSOR	Lin (2022)	0.560	0.0745	-	-
	Ours	**0.603**	**0.0698**	-	-

Table R2: Quantitative comparison by varying the bounding box sizes. We find that further enlarging the local context size fails to achieve improvement.

Expand Scale	$SA-SOR$	
	RVSOD	DAVSOD
1×	0.6092	0.2421
2×	**0.629**	**0.246**
3×	0.5900	0.2374
4×	0.5684	0.2393

Figure R1: When the motion of an instance is too large (the instance in the red box) to accurately capture the trajectory, the saliency of the instance is often high due to the high feature contrast between them.

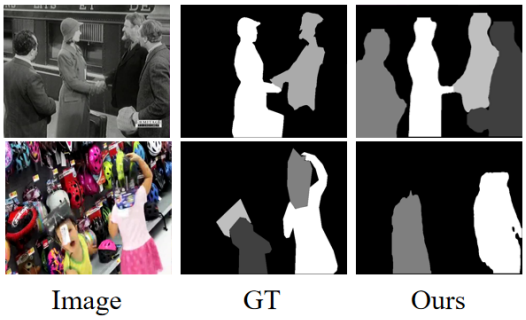

Figure R2: Failed cases. The frames are sourced from the RVSOD and DAVSOD datasets respectively. Missed or redundant detection of objects often leads to unsatisfactory SOR results.

Figure R4: The frames from the DAVSOD dataset, which contains many low-quality samples with blurry appearances. Our detector is affected by a large number of non-salient instances, making it challenging to accurately detect all significant objects. Additionally, some frames exhibit motion blur.

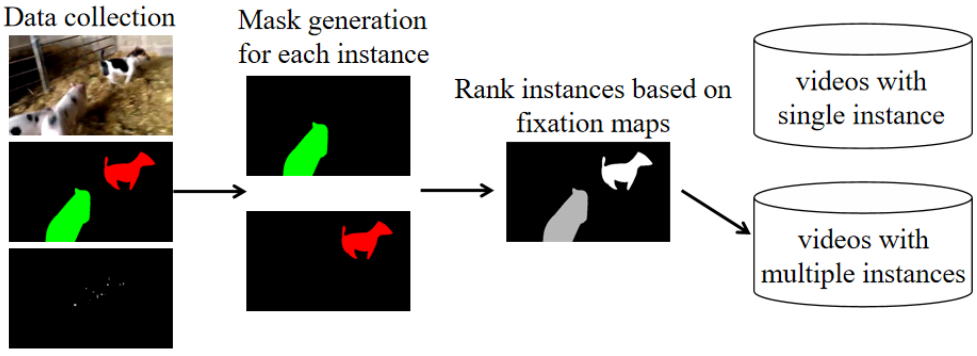

Figure R3: Flowchat for dataset collection and annotation.